LAGRANGIAN DUALITY

MARKUS GRASMAIR

1. Primal and dual problems

Assume that we are given a constrained optimisation problem of the form

$$\min_x f(x) \qquad \text{subject to } \begin{cases} c_i(x) = 0, & i \in \mathcal{E}, \\ c_i(x) \geq 0, & i \in \mathcal{I}. \end{cases} \tag{1}$$

We have seen earlier that, given some constraint qualification, the KKT conditions are a necessary optimality condition for this constrained problem, provided that the involved functions f and c_i are C^1, and some constraint qualification holds. With the help of the Lagrangian $\mathcal{L}\colon \mathbb{R}^d \times \mathbb{R}^{\mathcal{E}\cup\mathcal{I}}$,

$$\mathcal{L}(x, \lambda) = f(x) - \sum_{i \in \mathcal{E}\cup\mathcal{I}} \lambda_i c_i(x),$$

these can be written as

$$\begin{aligned} \nabla_x \mathcal{L}(x^*, \lambda^*) &= 0, \\ c_i(x^*) &= 0, \quad i \in \mathcal{E}, \\ c_i(x^*) &\geq 0, \quad i \in \mathcal{I}, \\ \lambda_i^* &\geq 0, \quad i \in \mathcal{I}, \\ \lambda_i^* c_i(x^*) &= 0, \quad i \in \mathcal{I}. \end{aligned}$$

In particular, the first line states that a (local) solution x^* of the constrained problem needs to be a critical point of the Lagrangian, and the second and third line state that a solution needs to be admissible. We have also seen that the Hessian of the Lagrangian with respect to the x-variable can be used to formulate second order necessary and sufficient (local) optimality conditions.

We will now develop a new interpretation of the relation between the Lagrangian and the constrained optimisation problem (1) that is not based on first order optimality conditions but rather on global properties. For that, we first consider what happens, if we try to maximise the Lagrangian, for fixed $x \in \mathbb{R}^d$, with respect to the (admissible) Lagrange parameters.

Lemma 1.1. *Define the function* $p\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$,

$$p(x) := \max_{\substack{\lambda \in \mathbb{R}^{\mathcal{E}\cup\mathcal{I}} \\ \lambda_i \geq 0,\ i \in \mathcal{I}}} \mathcal{L}(x, \lambda).$$

Then

$$p(x) = \begin{cases} f(x) & \text{if } c_i(x) = 0,\ i \in \mathcal{E},\ \text{and } c_i(x) \geq 0,\ i \in \mathcal{I}, \\ +\infty & \text{else.} \end{cases}$$

Proof. Assume first that $c_i(x) = 0$, $i \in \mathcal{E}$, and $c_i(x) \geq 0$, $i \in \mathcal{I}$. Then

$$\mathcal{L}(x, \lambda) = f(x) - \sum_{i \in \mathcal{I}} \lambda_i c_i(x) \leq f(x)$$

Date: 17th April 2024.

for all $\lambda_i \geq 0$, $i \in \mathcal{I}$, as the products $\lambda_i c_i(x)$, $i \in \mathcal{I}$, are necessarily non-negative. Moreover, we obtain equality by choosing $\lambda_i = 0$ for all $i \in \mathcal{I}$. This proves that $p(x) = f(x)$, if x satisfies all the equality and inequality constraints.

On the other hand, if any of the equality constraints is not satisfied, say $c_i(x) > 0$ for some $i \in \mathcal{E}$, then $\mathcal{L}(x, \lambda)$ can be made arbitrarily large by letting λ_i tend to $-\infty$. Similarly, if $c_i(x) < 0$ for any $i \in \mathcal{E} \cup \mathcal{I}$, then again $\mathcal{L}(x, \lambda)$ can be made arbitrarily large by letting λ_i tend to $+\infty$. Thus $p(x) = +\infty$, if any of the constraints fails to hold. □

Now assume that we want to solve the constrained problem (1). Then this is actually equivalent to solving the *unconstrained* problem

$$\min_{x \in \mathbb{R}^d} p(x),$$

because the fact that $p(x) = +\infty$ for all infeasible points effectively restricts the minimisation of p to the feasible set, on which p and f coincide.

This shows that solving the constrained optimisation problem (1) is equivalent to solving the unconstrained problem

$$\min_{x \in \mathbb{R}^d} p(x),$$

or, explicitly, the *primal problem*

$$\min_{x \in \mathbb{R}^d} \max_{\substack{\lambda \in \mathbb{R}^{\mathcal{E} \cup \mathcal{I}} \\ \lambda_i \geq 0,\ i \in \mathcal{I}}} \mathcal{L}(x, \lambda). \tag{P}$$

Now one defines the *dual problem* by exchanging the order of the minimum and the maximum in (P):

Definition 1.2. The *Lagrangian dual* of (P) is the optimisation problem

$$\max_{\substack{\lambda \in \mathbb{R}^{\mathcal{E} \cup \mathcal{I}} \\ \lambda_i \geq 0,\ i \in \mathcal{I}}} \min_{x \in \mathbb{R}^d} \mathcal{L}(x, \lambda). \tag{D}$$

More precisely, we first define a function $q \colon \mathbb{R}^{\mathcal{E} \cup \mathcal{I}} \to \mathbb{R} \cup \{-\infty\}$ by setting

$$q(\lambda) := \min_{x \in \mathbb{R}^d} \mathcal{L}(x, \lambda),$$

and then maximise q with respect to λ subject to the constraint that the Lagrange parameters for the inequality constraints are non-negative, that is, we solve the problem

$$\max_{\substack{\lambda \in \mathbb{R}^{\mathcal{E} \cup \mathcal{I}} \\ \lambda_i \geq 0,\ i \in \mathcal{I}}} q(\lambda).$$

Note that the minimum in the definition of q is taken over *all* $x \in \mathbb{R}^d$ irrespective of the constraints.

Remark 1.3. If one wants to be accurate, one should always read the minima and maxima in the definitions of the primal and the dual problem as infima and suprema, respectively, as it is not clear that these optimisation problems actually have solutions. Indeed, in the case of the primal formulation, the maximisation problem with respect to λ has a solution if and only if x is feasible.

Example 1.4. Consider the linear programme

$$\min_x \langle c, x \rangle \qquad \text{s.t. } Ax \geq b. \tag{L}$$

The corresponding Lagrangian is

$$\mathcal{L}(x, \lambda) = \langle c, x \rangle - \langle \lambda, Ax - b \rangle.$$

Thus the dual objective function is

$$q(\lambda) = \min_x \big(\langle c, x\rangle - \langle \lambda, Ax - b\rangle\big) = \langle \lambda, b\rangle + \min_x \langle c - A^T\lambda, x\rangle = \begin{cases} \langle \lambda, b\rangle & \text{if } A^T\lambda = c, \\ -\infty & \text{if } A^T\lambda \neq c. \end{cases}$$

Thus we can write the dual problem as

$$(L') \qquad \max_\lambda \langle b, \lambda\rangle \qquad \text{s.t. } \begin{cases} \lambda \geq 0, \\ A^T\lambda = c. \end{cases}$$

Note that we have again a linear programme, but the roles of the objective and constraint are reversed.

Example 1.5. Consider the quadratic programme

$$(Q) \qquad \min_x \frac{1}{2}\langle x, Gx\rangle - \langle c, x\rangle \qquad \text{s.t. } Ax \geq b,$$

where $G \in \mathbb{R}^{d\times d}$ is symmetric and positive definite.

Here the Lagrangian is

$$\mathcal{L}(x, \lambda) = \frac{1}{2}\langle x, Gx\rangle - \langle c, x\rangle - \langle \lambda, Ax - b\rangle.$$

The dual objective function is therefore

$$q(\lambda) = \min_x \Big(\frac{1}{2}\langle x, Gx\rangle - \langle c, x\rangle - \langle \lambda, Ax - b\rangle\Big).$$

The optimisation problem we have to solve here is strictly convex (and quadratic), and thus the solution is uniquely determined by the first order optimality condition

$$Gx - c - A^T\lambda = 0,$$

which yields

$$x = G^{-1}(c + A^T\lambda).$$

Next we note that

$$\frac{1}{2}\langle x, Gx\rangle - \langle c, x\rangle - \langle \lambda, Ax - b\rangle = \langle \lambda, b\rangle - \frac{1}{2}\langle x, Gx\rangle + \langle x, Gx - c - A^T\lambda\rangle.$$

Inserting the optimality condition and the value of x, we thus obtain that

$$q(\lambda) = \langle \lambda, b\rangle - \frac{1}{2}\big\langle G^{-1}(c + A^T\lambda), c + A^T\lambda\big\rangle.$$

The dual problem is thus the quadratic programme

$$(Q') \qquad \max_\lambda \Big(\langle \lambda, b\rangle - \frac{1}{2}\big\langle G^{-1}(c + A^T\lambda), c + A^T\lambda\big\rangle\Big) \qquad \text{s.t. } \lambda \geq 0.$$

2. Weak duality

We will now study the relationship between the primal and the dual problem. We have already seen in Example 2.4 that it can happen that these problems have the same optimal values. This relationship does not hold for all problems, as we see in a later example, but we can show that the primal problem is always larger or equal to the dual problem.

Definition 2.1. By

$$d := \min_{x\in\mathbb{R}^d} p(x) \; - \max_{\substack{\lambda\in\mathbb{R}^{\mathcal{E}\cup\mathcal{I}} \\ \lambda_i\geq 0,\, i\in\mathcal{I}}} q(\lambda)$$

we denote the duality gap for the primal dual pair (P) and (D).

We will next show that the duality gap is always non-negative. That is, all the values that the primal problem admits are larger or equal than all the values that the dual problem admits. This will be a consequence of the following lemma, which is an important result in itself.

Lemma 2.2. *Let X and Y be non-empty sets and let $h\colon X \times Y \to \mathbb{R} \cup \{\pm\infty\}$. Then*

$$\sup_{y\in Y} \inf_{x\in X} h(x,y) \leq \inf_{x\in X} \sup_{y\in Y} h(x,y). \tag{2}$$

Proof. Assume to the contrary that (2) does not hold. That is,

$$\sup_{y\in Y} \inf_{x\in X} h(x,y) > \inf_{x\in X} \sup_{y\in Y} h(x,y).$$

Then there exists some $\varepsilon > 0$ and some $\hat{y} \in Y$ such that

$$\inf_{x\in X} h(x,\hat{y}) > \inf_{x\in X} \sup_{y\in Y} h(x,y) + \varepsilon.$$

Since $\sup_{y\in Y} h(x,y) \geq h(x,\hat{y})$ for all $x \in X$, it follows that

$$\inf_{x\in X} h(x,\hat{y}) > \inf_{x\in X} \sup_{y\in Y} h(x,y) + \varepsilon \geq \inf_{x\in X} h(x,\hat{y}) + \varepsilon,$$

which is an obvious contradiction. Thus (2) holds. □

Lemma 2.3 (Weak duality)**.** *Let d be the duality gap for (P) and (D). Then $d \geq 0$.*

Proof. This is an immediate consequence of the definition of the primal and the dual problem and Lemma 2.2. □

We now now consider two concrete, simple examples, where the dual problem can be found and solved exactly. In the first example, we have a convex problem on a convex set. In this case, it turns out that the duality gap is zero. In the second example, however, where neither the function nor the feasible set is convex, the duality gap turns out to be strictly positive.

Example 2.4. Let $c \in \mathbb{R}^d \setminus \{0\}$ and consider the optimisation problem

$$\min_x c^T x \qquad \text{s.t. } \|x\|^2 \leq 1.$$

The solution of this problem is

$$x^* = -\frac{c}{\|c\|},$$

and the corresponding Lagrange multiplier is

$$\lambda^* = \frac{\|c\|}{2}.$$

Now we will compute the dual problem and its solution. First we note that the Lagrangian of this problem is the function $\mathcal{L}\colon \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$,

$$\mathcal{L}(x,\lambda) = c^T x - \lambda(1 - \|x\|^2).$$

Thus the function $q\colon \mathbb{R} \to \mathbb{R} \cup \{-\infty\}$ is

$$q(\lambda) = \inf_{x\in\mathbb{R}^d} \bigl(c^T x - \lambda(1 - \|x\|^2)\bigr) = -\lambda - \inf_{x\in\mathbb{R}^d} \bigl(c^T x + \lambda\|x\|^2\bigr).$$

For $\lambda \leq 0$, the term $c^T x + \lambda\|x\|^2$ is unbounded below. Else, it is coercive and has a unique global minimum

$$x_\lambda = -\frac{c}{2\lambda}.$$

Thus

$$q(\lambda) = -\lambda - \left(c^T x_\lambda + \lambda \|x_\lambda\|^2\right) = -\lambda - \frac{\|c\|^2}{4\lambda}$$

if $\lambda > 0$, and $q(\lambda) = -\infty$ else. Now consider the dual optimisation problem

$$\max_{\lambda \geq 0} q(\lambda) = \max_{\lambda > 0} -\lambda - \frac{\|c\|^2}{4\lambda}.$$

A short computation shows that this problem has the unique solution

$$\hat{\lambda} = \frac{\|c\|}{2},$$

which was precisely the Lagrange multiplier for the primal problem.

Moreover, it is easy to verify that the optimal function values for the primal and dual problem are the same.

Example 2.5. Consider the optimisation problem

$$\min_x -\frac{1}{1+x^2} \qquad \text{s.t. } x^2 \geq 1.$$

The obvious solutions to this problem are the points $x = \pm 1$, where the value of the objective function is $-1/2$.

Now we compute the dual of this problem: The Lagrangian is

$$\mathcal{L}(x, \lambda) = -\frac{1}{1+x^2} - \lambda(x^2 - 1).$$

For $\lambda > 0$, the term $-\lambda x^2$ dominates the Lagrangian and we have

$$q(\lambda) = \inf_x \mathcal{L}(x, \lambda) = -\infty.$$

On the other hand, for $\lambda = 0$ we have

$$q(0) = \inf_x \mathcal{L}(x, 0) = \inf_x -\frac{1}{1+x^2} = -1.$$

Finally, for $\lambda < 0$ we have

$$q(\lambda) = \inf_x \mathcal{L}(x, \lambda) = \inf_x \Big(-\frac{1}{1+x^2} - \lambda(x^2 - 1)\Big) = -1 + \lambda,$$

as the infimum is always attained at $x = 0$. Thus

$$q(\lambda) = \begin{cases} -\infty & \text{if } \lambda > 0, \\ -1 + \lambda & \text{if } \lambda \leq 0. \end{cases}$$

Since the dual problem is a maximisation problem, the function value $-\infty$ for $\lambda > 0$ effectively serves as a constraint $\lambda \leq 0$. In addition, we have the constraint $\lambda \geq 0$ from the fact that we have an inequality constraint. Thus we obtain the dual problem

$$\max_\lambda -1 + \lambda \qquad \text{s.t. } \lambda = 0$$

with the (only possible) solution $\lambda = 0$ and an objective value of -1.

Consequently, we have a (non-zero) duality gap

$$d = \min_{x \in \mathbb{R}} p(x) \; - \max_{\lambda \geq 0} q(\lambda) = -\frac{1}{2} - (-1) = \frac{1}{2}.$$

3. Strong duality

The most important situation is that where the duality gap is equal to zero, as in this case the dual problem can be used for solving the original (*primal*) problem. In order to arrive at such results, we have to introduce the notion of saddle points. Note that the definition below is somewhat different from the standard notion of saddle points used in basic calculus classes in that we are interested in *global* optimality properties with respect to the different variables.

Definition 3.1. The point $(x^*, \lambda^*) \in \mathbb{R}^d \times \mathbb{R}^{\mathcal{E}\cup\mathcal{I}}$ with $\lambda_i^* \geq 0$, $i \in \mathcal{I}$, is a *saddle point* of the Lagrangian, if (or a *primal–dual solution* of (P)), if

$$\mathcal{L}(x^*, \lambda) \leq \mathcal{L}(x^*, \lambda^*) \leq \mathcal{L}(x, \lambda^*)$$

for all $(x, \lambda) \in \mathbb{R}^d \times \mathbb{R}^{\mathcal{E}\cup\mathcal{I}}$ with $\lambda_i \geq 0$, $i \in \mathcal{I}$.

That is, a saddle point is a maximiser with respect to the (feasible) dual variables and a minimiser with respect to the primal variables.

In the following result, we show that saddle points of the Lagrangian are *primal-dual* solutions of the constrainted optimisation problem. That is, the x-coordinates are solutions of the primal problem and the λ-coordinates are solutions of the dual problem. In addition, the λ-coordinates are actually Lagrange multipliers for the constrained problem, provided that the involved functions are sufficiently regular.

Proposition 3.2. *Assume that (x^*, λ^*) is a saddle point of the Lagrangian and that $\mathcal{L}(x^*, \lambda^*) \in \mathbb{R}$. Then x^* is a solution of (P), λ^* is a solution of (D), and the complementarity conditions $\lambda_i^* c_i(x^*) = 0$, $i \in \mathcal{E} \cup \mathcal{I}$ hold. If the functions f and c_i, $i \in \mathcal{E} \cup \mathcal{I}$, are $\mathcal{C}^1$, then x^* is a KKT point with Lagrange multiplier λ^*.*

Proof. We note first that, for every λ with $\lambda_i \geq 0$, $i \in \mathcal{I}$, we have

$$q(\lambda) = \min_x \mathcal{L}(x, \lambda) \leq \mathcal{L}(x^*, \lambda) \leq \mathcal{L}(x^*, \lambda^*) = \min_x \mathcal{L}(x, \lambda^*) = q(\lambda^*).$$

Here the third and fourth relation are consequences of the assumption that (x^*, λ^*) is a saddle point. This shows that λ^* solves (D).

Similarly,

$$p(x) = \max_{\substack{\lambda \in \mathbb{R}^{\mathcal{E}\cup\mathcal{I}} \\ \lambda_i \geq 0,\, i \in \mathcal{I}}} \mathcal{L}(x, \lambda) \geq \mathcal{L}(x, \lambda^*) \geq \mathcal{L}(x^*, \lambda^*) = \max_{\substack{\lambda \in \mathbb{R}^{\mathcal{E}\cup\mathcal{I}} \\ \lambda_i \geq 0,\, i \in \mathcal{I}}} \mathcal{L}(x^*, \lambda) = p(x^*),$$

showing that x^* solves (P). In particular, since $\mathcal{L}(x^*, \lambda^*)$ is finite, this implies that x^* is a feasible point.

Now assume that the complementarity condition does not hold. Since x^* is feasible, this implies that there exists $i \in \mathcal{I}$ such that $c_i(x^*) > 0$ and $\lambda_i^* > 0$. In this case, however, replacing λ_i^* with $\hat{\lambda}_i := 0$ increases the value of the Lagrangian (without changing x^*). This is a contradiction to the assumption that (x^*, λ^*) is a saddle point (again, this uses the assumption that $\mathcal{L}(x^*, \lambda^*)$ is finite).

Finally, if the functions f and c_i, $i \in \mathcal{E}\cup\mathcal{I}$, are $\mathcal{C}^1$, then the fact that x^* minimises $\mathcal{L}(\cdot, \lambda^*)$ implies that the first order optimality condition

$$\nabla_x \mathcal{L}(x^*, \lambda^*) = 0$$

holds. As a consequence, all KKT conditions are satisfied. □

Remark 3.3. Note that the converse in general does not hold. That is, if (x^*, λ^*) is a KKT point, it is not necessarily a saddle point of the Lagrangian. Indeed, it is by no means guaranteed that the Lagrangian has any saddle points at all.

This can be seen in the problem discussed in Example 2.5: Here the points $x^* = \pm 1$ are KKT points (and global solutions) with Lagrange multipliers $\lambda^* = 1/4$.

However, the points $(x^*, \lambda^*) = (\pm 1, 1/4)$ are not saddle points of the Lagrangian in the sense of Definition 3.1, as

$$\mathcal{L}(0, 1/4) = -1 < -\frac{1}{2} = \mathcal{L}(\pm 1, 1/4).$$

Also, the Lagrange multiplier $\lambda^* = 1/4$ does not solve the dual problem, which is only finite for $\lambda = 0$.

Theorem 3.4. *Assume that x^* is a solution of (P), λ^* is a solution of (D), and that the duality gap is zero. Then (x^*, λ^*) is a saddle point of the Lagrangian.*

In particular, the complementarity conditions hold and x^* is a KKT point with Lagrange multiplier λ^* provided that the functions f and c_i are $\mathcal{C}^1$.

Proof. Since the duality gap is zero and x^* and λ^* solve the primal and dual problems, respectively, we have that

$$p(x^*) = q(\lambda^*)$$

Thus we have for every x that

$$\mathcal{L}(x, \lambda^*) \geq \min_{\hat{x}} \mathcal{L}(\hat{x}, \lambda^*) = q(\lambda^*) = p(x^*) = \max_{\substack{\lambda \in \mathbb{R}^{\mathcal{E} \cup \mathcal{I}} \\ \lambda_i \geq 0,\, i \in \mathcal{I}}} \mathcal{L}(x^*, \lambda) \geq \mathcal{L}(x^*, \lambda^*).$$

Similarly we have for every λ with $\lambda_i \geq 0$, $i \in \mathcal{I}$, that

$$\mathcal{L}(x^*, \lambda) \leq \max_{\substack{\hat{\lambda} \in \mathbb{R}^{\mathcal{E} \cup \mathcal{I}} \\ \hat{\lambda}_i \geq 0,\ i \in \mathcal{I}}} \mathcal{L}(x^*, \hat{\lambda}) = p(x^*) = q(\lambda^*) = \min_x \mathcal{L}(x, \lambda^*) \leq \mathcal{L}(x^*, \lambda^*).$$

This shows that (x^*, λ^*) is a saddle point of the Lagrangian.

The other assertions follow from Proposition 3.2. □

4. Duality and convex programming

In the following, we will discuss the application of duality theory to convex programmes, that is, convex optimisation problems with concave and linear inequality constraints, and linear equality constraints. More precisely, we assume that we are given constraints of the form

$$\begin{aligned} c_i(x) &\geq 0, \quad i \in \mathcal{I}, \\ Ax &\geq b, \\ Cx &= d, \end{aligned}$$

where the functions $c_i \colon \mathbb{R}^d \to \mathbb{R}$ are concave, $A \in \mathbb{R}^{m \times d}$, $C \in \mathbb{R}^{\ell \times d}$ are matrices, and $b \in \mathbb{R}^m$, $d \in \mathbb{R}^\ell$ are vectors. The inequalities $Ax \geq b$ are understood componentwise.

Definition 4.1. We say that *Slater's constraint qualification* is satisfied, if there exists $x \in \mathbb{R}^d$ with $Ax \geq b$, $Cx = d$, and $c_i(x) > 0$ for all $i \in \mathcal{I}$.

Remark 4.2. In the specific situation of only linear inequality and equality constraints, Slater's constraint qualification is equivalent to the feasibility of the constraints. In the case of additional non-linear (but concave) constraints, the condition is somehow stonger.

Theorem 4.3 (strong duality for convex programmes)**.** *Assume that $f \colon \mathbb{R}^d \to \mathbb{R}$ is convex and that Slater's constraint qualification holds. Assume morever that*

$$\inf_x p(x) > -\infty$$

(that is, the primal problem is bounded). Then the dual problem has a solution λ^* and the duality gap is zero.

If in addition the primal problem has a solution x^*, then (x^*, λ^*) is a saddle point of the Lagrangian. If moreover the functions f and c_i are $\mathcal{C}^1$, then x^* is a KKT point with Lagrange multiplier λ^*.

Proof. See [2, Thm. 11.15]. Note that the second part of the Theorem is an immediate consequence of Theorem 3.4, once it has been established that the duality gap is zero. □

Example 4.4. We consider again the linear programme discussed in Example 1.4, that is, the programme

$$(L) \qquad \min_x c^T x \qquad \text{s.t. } Ax \geq b,$$

with corresponding dual programme

$$(L') \qquad \max_\lambda b^T \lambda \qquad \text{s.t. } \begin{cases} \lambda \geq 0, \\ A^T \lambda = c. \end{cases}$$

Now we will apply the results of Theorem 4.3 to this situation: To that end, we note first that the objective function is convex (since it is linear), and that we only have linear constraints. As a consequence, Slater's constraint qualification is satisfied if and only if the problem is *primal feasible*, that is, there exists a point $x \in \mathbb{R}^d$ satisfying the primal constraints $Ax \geq b$. Now assume that the problem is primal feasible and *bounded*, that is,

$$\inf_{Ax \geq b} c^T x > -\infty.$$

Then Theorem 4.3 is applicable and it follows that the dual problem (L') has a solution λ^*. In addition, it can be shown (see Remark 4.5 below) that in such a situation, the primal problem (L) admits a solution x^* as well.

Thus the primal-dual pair (x^*, λ^*) satisfies the KKT conditions, which in this case can be written as

$$\begin{aligned} A^T \lambda &= c, \\ Ax &\geq b, \\ \lambda &\geq 0, \\ \lambda^T (Ax - b) &= 0. \end{aligned} \tag{3}$$

Conversely, if (x^*, λ^*) solve the system (3), then x^* solves (L), λ^* solves (L'), and (since the duality gap is zero) $c^T x^* = b^T \lambda^*$.

In addition, if (x, λ) is any *primal-dual feasible* pair, that is, if $Ax \geq b$, $A^T \lambda = c$, and $\lambda \geq 0$, then $c^T x \geq b^T \lambda$. If, actually, $c^T x = b^T \lambda$, then (x, λ) is a primal-dual solution.

Remark 4.5. We consider again the linear programme

$$(L) \qquad \min_x c^T x \qquad \text{s.t. } Ax \geq b$$

with dual

$$(L') \qquad \max_\lambda b^T \lambda \qquad \text{s.t. } \begin{cases} \lambda \geq 0, \\ A^T \lambda = c, \end{cases}$$

from Example 4.4.

Since the dual is again a linear programme, we can try to compute its dual (the *double-dual* of (L)), and expect it to be a linear programme again. The Lagrangian of the dual programme is

$$\mathcal{L}'(\lambda; y, s) = b^T \lambda - y^T (A^T \lambda - c) - \lambda^T s,$$

and thus we obtain the double-dual problem (note that we have a maximisation problem, and thus the Lagrange parameters for the inequality constraints have to be non-positive!)

$$\min_{\substack{y,\,s\\ s\leq 0}} \max_{\lambda} \big(b^T\lambda - y^T(A^T\lambda - c) - s^T\lambda\big).$$

This can be rewritten as the linear programme

$$\min_{y,s} c^T y \qquad \text{s.t.} \begin{cases} s \leq 0, \\ Ay + s = b. \end{cases}$$

The Lagrange parameter s in this problem can now be interpreted as a slack variable, and we see that this double-dual is equivalent to the problem

$$\min_{y} c^T y \qquad \text{s.t. } Ay \geq b,$$

which is again the primal problem.

Thus we have shown that, apart from possible slack variables, the double-dual of a linear programme is again the primal programme. In particular, we can apply Theorem 4.3 to the *dual programme*, and conclude in particular that the primal problem has a solution provided that the dual programme is feasible and bounded. This, however, is guaranteed if the primal programme is feasible and bounded, since in this case the dual programme actually has a solution. In addition, if the primal problem is unbounded, then it follows from weak duality that the value of the dual problem is $-\infty$. This is only possible, if the dual problem is infeasible. Similarly, if the dual problem is unbounded (above), then weak duality implies that the value of the primal problem is $+\infty$, which implies that the primal problem is infeasible.

Thus we obtain the following results (cf. [3, Thm. 13.1]):

- If the primal (or the dual) problem is feasible and bounded, then there exists a primal-dual solution.
- If the primal problem is unbounded, then the dual problem is infeasible.
- If the dual problem is unbounded, then the primal problem is infeasible.

5. Legendre–Fenchel transform

Assume that $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a function and that we want to solve a problem of the form

$$\min_{x} f(x) \qquad \text{s.t. } Ax = b. \tag{4}$$

The Lagrangian of this problem is

$$\mathcal{L}(x,\lambda) = f(x) - \langle \lambda, Ax - b\rangle.$$

For the dual objective function, we thus obtain the expression

$$\begin{aligned} q(\lambda) = \inf_x \big(f(x) - \langle \lambda, Ax - b\rangle\big) &= \langle \lambda, b\rangle + \inf_x \big(f(x) - \langle \lambda, Ax\rangle\big) \\ &= \langle \lambda, b\rangle - \sup_x \big(\langle A^T\lambda, x\rangle - f(x)\big). \end{aligned}$$

We will now discuss the last expression in more detail.

Definition 5.1. Assume that $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a function. Its *Legendre–Fenchel transform* is the function $f^*\colon \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ defined by

$$f^*(y) := \sup_{x\in\mathbb{R}^d} \big(\langle y, x\rangle - f(x)\big).$$

Remark 5.2. Assume that $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a function and that there exists some $x_0 \in \mathbb{R}^d$ such that $f(x_0) < +\infty$. (Such functions are called *proper*.) Then we have

$$f^*(y) = \sup_{x \in \mathbb{R}^d} \big(\langle y, x \rangle - f(x) \big) \geq \langle y, x_0 \rangle - f(x_0) > -\infty$$

for all $y \in \mathbb{R}^d$. Thus the only possibilty for f^* to attain the value $-\infty$ is if $f(x) = +\infty$ for all x; in this case, we obtain $f^*(y) = -\infty$ for all y. Since this case is not particularly interesting for optimisation, we can in practice always assume that f^* is again a function from $\mathbb{R}^d$ to $\mathbb{R} \cup \{+\infty\}$. Note, though, that it is possible that f^* attains the value $+\infty$, even if f is only finite valued.

Lemma 5.3. *The Legendre–Fenchel transform f^* of any proper function $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is convex and lower semi-continuous.*

Proof. For every fixed $x \in \mathbb{R}^d$, the function

$$y \mapsto \langle y, x \rangle - f(x)$$

is linear, and thus in particular continuous and convex. Thus f^* is the supremum of continuous and convex functions, and thus itself lower semi-continuous and convex. □

Lemma 5.4. *Assume that $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper function. Then*

$$f(x) + f^*(y) \geq \langle x, y \rangle$$

for all x, $y \in \mathbb{R}^d$. Moreover, if f is differentiable at $x \in \mathbb{R}^d$ and

$$f(x) + f^*(y) = \langle x, y \rangle,$$

then

$$\nabla f(x) = y.$$

Proof. Let x, $y \in \mathbb{R}^d$ be fixed and define the function $h\colon \mathbb{R}^d \to \mathbb{R} \cup \{-\infty\}$,

$$h(z) = \langle y, z \rangle - f(z).$$

By definition, we have that

$$f^*(y) = \sup_{z \in \mathbb{R}^d} h(z) \geq h(x) = \langle y, x \rangle - f(x),$$

which implies that

$$f(x) + f^*(y) \geq \langle y, x \rangle.$$

Now assume that this is actually an equality. Then the supremum of h is attained at x. Now, if f is differentiable at x, this implies that

$$0 = \nabla h(x) = y - \nabla f(x),$$

which concludes the proof. □

Assume now that $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a (proper) function with Legendre–Fenchel dual $f^*\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. Then we can compute the Legendre–Fenchel dual of f^*, which is the function $(f^*)^*\colon \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ defined as

$$(f^*)^*(x) := \sup_{y \in \mathbb{R}^d} \langle x, y \rangle - f^*(y).$$

Theorem 5.5. *Assume that $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper function. Then*

$$(f^*)^*(x) \leq f(x)$$

for all $x \in \mathbb{R}^d$. Moreover, if f is lower semi-continuous and convex, then

$$(f^*)^*(x) = f(x)$$

for all $x \in \mathbb{R}^d$.

Proof. See [1, Thm. 4.2.1]. □

Corollary 5.6. *Assume that $f\colon \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is proper, lower semi-continuous, and convex and that x, $y \in \mathbb{R}^d$ are such that*

$$f(x) + f^*(y) = \langle x, y\rangle. \tag{5}$$

If f^* is differentiable at y, then

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

Proof. Since f is lower semi-continuous and convex, it follows that $f = (f^*)^*$. Thus (5) implies that

$$(f^*)^*(x) + f^*(y) = \langle x, y\rangle.$$

Applying Lemma 5.4 to the function f^*, we thus obtain that

$$\nabla f^*(y) = x,$$

which proves the assertion. □

Example 5.7. We compute the Legendre–Fenchel dual explicitly in two simple examples. Consider first the function $f\colon \mathbb{R} \to \mathbb{R}$,

$$f(x) = x^2.$$

Then

$$f^*(y) = \sup_{x\in\mathbb{R}}(xy - x^2).$$

That is, in order to find f^*, we have to maximise, for fixed y, the function

$$h(x) := xy - x^2.$$

Since this is a strictly concave and differentiable function (as $f(x) = x^2$ is convex), the maximiser of h is the unique point x where $h'(x) = 0$. We have

$$h'(x) = y - 2x,$$

and thus $h'(x) = 0$ if and only if $x = y/2$. Thus we obtain that

$$f^*(y) = \sup_{x\in\mathbb{R}} h(x) = h(y/2) = \frac{y^2}{2} - \frac{y^2}{4} = \frac{y^2}{4}.$$

Now we consider the function $g\colon \mathbb{R} \to \mathbb{R}$,

$$g(x) = |x|.$$

Then

$$g^*(y) = \sup_{x\in\mathbb{R}}(xy - |x|).$$

Since g is not differentiable (though still convex), we have to argue a bit differently for the computation of g^*. Denote again by

$$h(x) := xy - |x|$$

the function to be maximised. Assume first that $y > 1$. Then

$$h(x) = xy - |x| \geq xy - x = x(y-1).$$

In particular, $\lim_{x\to\infty} h(x) = +\infty$, and thus

$$g^*(y) = \sup_{x\in\mathbb{R}} h(x) = +\infty \qquad \text{if } y > 1.$$

Similarly, if $y < -1$, then

$$h(x) = xy - |x| \geq xy + x = x(y+1)$$

and thus $\lim_{x\to-\infty} h(x) = +\infty$. Therefore

$$g^*(y) = \sup_{x\in\mathbb{R}} h(x) = +\infty \qquad \text{if } y < -1.$$

Finally, if $|y| \leq 1$, we can estimate

$$h(x) = xy - |x| \leq |x|(|y| - 1) \leq 0.$$

Since $h(0) = 0$, it follows that

$$g^*(y) = \sup_{x \in \mathbb{R}} h(x) = 0 \qquad \text{if } -1 \leq y \leq 1.$$

In total, we obtain that

$$g^*(y) = \begin{cases} +\infty & \text{if } |y| > 1, \\ 0 & \text{if } |y| \leq 1. \end{cases}$$

Example 5.8. Here are some important examples of convex functions together with their Legendre–Fenchel transform.[1] More examples can be found in [1, p. 50]. Note that all of the functions below are convex and lower semi-continuous, which implies that $(f^*)^* = f$.

- Let $1 < p < \infty$ and denote by

  $$p_* := \frac{p}{p-1}$$

  the *Hölder conjugate* of p. Consider the function $f \colon \mathbb{R}^d \to \mathbb{R}$,

  $$f(x) = \frac{1}{p}\|x\|_p^p = \frac{1}{p}\sum_{i=1}^d |x_i|^p.$$

  Then

  $$f^*(y) = \frac{1}{p_*}\|y\|_{p_*}^{p_*} = \frac{1}{p_*}\sum_{i=1}^d |y_i|^{p_*}.$$

- Define $f \colon \mathbb{R}^d \to \mathbb{R}$,

  $$f(x) = \|x\|_1 = \sum_{i=1}^d |x_i|.$$

  Then

  $$f^*(y) = \begin{cases} 0 & \text{if } \|y\|_\infty \leq 1, \\ +\infty & \text{if } \|y\|_\infty > 1. \end{cases}$$

- Define $f \colon \mathbb{R}^d \to \mathbb{R}$,

  $$f(x) = \sqrt{1 + \|x\|_2^2}.$$

  Then

  $$f^*(y) = \begin{cases} -\sqrt{1 - \|y\|_2^2} & \text{if } \|y\|_2 \leq 1, \\ +\infty & \text{if } \|y\|_2 > 1. \end{cases}$$

- Define $f \colon \mathbb{R} \to \mathbb{R}$,

  $$f(x) = e^x.$$

  Then

  $$f^*(y) = \begin{cases} y \log y - y & \text{if } y > 0, \\ 0 & \text{if } y = 0, \\ +\infty & \text{if } y < 0. \end{cases}$$

[1] As an exercise, you can (and should) verify that the transforms are correct.

- Define $f\colon \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$,

$$f(x) = \begin{cases} -\log(x) & \text{if } x > 0, \\ +\infty & \text{if } x \leq 0. \end{cases}$$

  Then

$$f^*(y) = \begin{cases} -\log(-y) - 1 & \text{if } y < 0, \\ +\infty & \text{if } y \geq 0. \end{cases}$$

In addition, we have the following calculation rules for the Legendre–Fenchel transform:

Lemma 5.9. *Assume that $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a function.*

- *Assume that $A \in \mathbb{R}^{d\times d}$ is a non-singular matrix and define $g\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, $g(x) := f(Ax)$. Then*

$$g^*(y) = f^*(A^{-T}y),$$

  where $A^{-T} = (A^{-1})^T = (A^T)^{-1}$. In particular, if $a \in \mathbb{R}$, $a \neq 0$, and $g(x) = f(ax)$, then $g^*(y) = f^*(y/a)$.
- *Assume that $b \in \mathbb{R}^d$ and define $h\colon \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$, $h(x) := f(x+b)$. Then*

$$h^*(y) = f^*(y) - \langle y, b\rangle.$$

- *Assume that $\lambda > 0$ and define $k\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, $k(x) = \lambda f(x)$. Then*

$$k^*(y) = \lambda f^*(y/\lambda).$$

- *Assume that*

$$f(x) = \sum_{i=1}^{d} f_i(x_i)$$

  for some functions $f_i\colon \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$. Then

$$f^*(y) = \sum_{i=1}^{d} f_i^*(y_i).$$

Proof. We have that

$$\begin{aligned} g^*(y) &= \sup_{x\in\mathbb{R}^d} \big(\langle x, y\rangle - g(x)\big) = \sup_{x\in\mathbb{R}^d} \big(\langle x, y\rangle - f(Ax)\big) \\ &= \sup_{x\in\mathbb{R}^d} \big(\langle A^{-1}Ax, y\rangle - f(Ax)\big) = \sup_{\tilde{x}\in\mathbb{R}^d} \big(\langle A^{-1}\tilde{x}, y\rangle - f(\tilde{x})\big) \\ &= \sup_{\tilde{x}\in\mathbb{R}^d} \big(\langle \tilde{x}, A^{-T}y\rangle - f(\tilde{x})\big) = f^*(A^{-T}y). \end{aligned}$$

Next we compute

$$\begin{aligned} h^*(y) &= \sup_{x\in\mathbb{R}^d} \big(\langle x, y\rangle - h(x)\big) = \sup_{x\in\mathbb{R}^d} \big(\langle x, y\rangle - f(x+b)\big) \\ &= \sup_{x\in\mathbb{R}^d} \big(\langle x+b, y\rangle - \langle b, y\rangle - f(x+b)\big) \\ &= \sup_{\tilde{x}\in\mathbb{R}^d} \big(\langle \tilde{x}, y\rangle - f(\tilde{x})\big) - \langle b, y\rangle = f^*(y) - \langle b, y\rangle. \end{aligned}$$

Then we have

$$\begin{aligned} k^*(y) &= \sup_{x\in\mathbb{R}^d} \big(\langle x, y\rangle - k(x)\big) = \sup_{x\in\mathbb{R}^d} \big(\langle x, y\rangle - \lambda f(x)\big) \\ &= \lambda \sup_{x\in\mathbb{R}^d} \big(\langle x, y/\lambda\rangle - f(x)\big) = \lambda f^*(y/\lambda). \end{aligned}$$

Finally, assume that $f(x) = \sum_{i=1}^d f_i(x_i)$. Then

$$\begin{aligned} f^*(y) &= \sup_{x \in \mathbb{R}^d} \bigl(\langle x, y \rangle - f(x) \bigr) = \sup_{x \in \mathbb{R}^d} \Bigl(\sum_{i=1}^d x_i y_i - \sum_{i=1}^d f_i(x_i) \Bigr) \\ &= \sup_{x \in \mathbb{R}^d} \sum_{i=1}^d \bigl(x_i y_i - f_i(x_i) \bigr) = \sum_{i=1}^d \sup_{x_i \in \mathbb{R}} \bigl(x_i y_i - f_i(x_i) \bigr) = \sum_{i=1}^d f_i^*(y_i). \end{aligned}$$

□

6. Dual methods for constrained optimisation

We now return to the problem (4), that is, the constrained optimisation problem

$$\min_x f(x) \qquad \text{s.t. } Ax = b.$$

Using the Legendre–Fenchel transform of f, we can write the dual objective for this problem as

$$q(\lambda) = \langle \lambda, b \rangle - f^*(A^T \lambda).$$

The dual problem to (4) thus becomes

$$\max_\lambda \bigl(\langle \lambda, b \rangle - f^*(A^T \lambda) \bigr).$$

Proposition 6.1. *Assume that $f \colon \mathbb{R}^d \to \mathbb{R}$ is lower semi-continuous and convex and that $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$ are such that $b \in \operatorname{Ran}(A)$. Assume moreover that the problem*

$$\min_{x \in \mathbb{R}^d} f(x) \qquad \textit{s.t. } Ax = b$$

admits a solution x^*.

Then the problem

$$\max_{\lambda \in \mathbb{R}^m} \bigl(\langle \lambda, b \rangle - f^*(A^T \lambda) \bigr)$$

has a solution λ^* and we have

$$f(x^*) + f^*(A^T \lambda^*) = \langle A^T \lambda^*, x \rangle.$$

Moreover, if f is differentiable at x^*, then

$$\nabla f(x^*) = A^T \lambda^*,$$

and if f^* is differentiable at $A^T \lambda^*$, then

$$\nabla f^*(A^T \lambda^*) = x^*.$$

Proof. The condition $b \in \operatorname{Ran}(A)$ implies that there exists $x \in \mathbb{R}^d$ such that $Ax = b$, which is, in this case, nothing more than Slater's constraint qualification. Since f is convex, we can apply Theorem 4.3. In particular, we obtain that the dual problem

$$\max_{\lambda \in \mathbb{R}^m} \bigl(\langle \lambda, b \rangle - f^*(A^T \lambda) \bigr)$$

has a solution λ^* and that the duality gap is 0. This implies that

$$f(x^*) = \inf_{Ax = b} f(x) = \sup_\lambda \bigl(\langle \lambda, b \rangle - f^*(A^T \lambda) \bigr) = \langle \lambda^*, b \rangle - f^*(A^T \lambda^*).$$

Since x^* solves the primal problem, it is in particular feasible, that is, $Ax^* = b$. Thus we have

$$f(x^*) + f^*(A^T \lambda^*) = \langle \lambda^*, b \rangle = \langle \lambda^*, Ax^* \rangle = \langle A^T \lambda^*, x^* \rangle.$$

Now assume that f is differentiable at x^*. Then Lemma 5.4 implies that

$$\nabla f(x^*) = A^T \lambda^*.$$

Similarly, if f^* is differentiable at $A^T\lambda^*$, then Corollary 5.6 implies that

$$\nabla f^*(A^T\lambda^*) = x^*,$$

which concludes the proof. □

Proposition 6.1 shows that we can use the dual problem in order to find the solution of the primal problem. In particular, if we can ensure that f^* is differentiable at the dual solution λ^*, we can recover the primal solution by simply evaluating $\nabla f^*(A^T\lambda^*)$.

Example 6.2. Let $\alpha_1, \alpha_2 > 0$. The *elastic net* (with parameters α_1 and α_2) is the function $f\colon \mathbb{R}^d \to \mathbb{R}$,

$$f(x) := \alpha_1\|x\|_1 + \frac{\alpha_2}{2}\|x\|_2^2 = \sum_{i=1}^{d}\Big(\alpha_1|x_i| + \frac{\alpha_2}{2}x_i^2\Big).$$

Assume now that $A \in \mathbb{R}^{m\times d}$ is some matrix and that $b \in \mathbb{R}^m$ is a vector satisfying $b \in \operatorname{Ran} A$. We consider the optimisation problem

$$\min_{x\in\mathbb{R}^d} f(x) \qquad \text{s.t. } Ax = b. \tag{6}$$

Note that the function f is non-smooth, and thus we cannot (or rather: should not) apply the methods for constrained optimisation that we previously discussed, like the Augmented Lagrangian method or sequential quadratic programming. The function f is convex, however, and thus we can apply the theory developed above. Thus, instead of trying to solve (6) directly, we instead consider the dual problem

$$\max_{\lambda\in\mathbb{R}^m}\big(\langle\lambda, b\rangle - f^*(A^T\lambda)\big),$$

or the equivalent minimisation problem

$$\min_{\lambda\in\mathbb{R}^m}\big(f^*(A^T\lambda) - \langle\lambda, b\rangle\big).$$

For this, we need to compute the Legendre–Fenchel transform of f. We see that we can write

$$f(x) = \sum_{i=1}^{d} g(x_i) \qquad \text{where} \qquad g(t) := \alpha_1|t| + \frac{\alpha_2}{2}t^2.$$

From Lemma 5.9 we obtain therefore that

$$f^*(y) = \sum_{i=1}^{d} g^*(y_i)$$

and it remains to compute g^*. For $s \in \mathbb{R}$ we have

$$g^*(s) = \sup_{t\in\mathbb{R}}\big(st - g(t)\big).$$

Assume therefore that $s \in \mathbb{R}$ is fixed and denote

$$h(t) := st - g(t) = st - \alpha_1|t| - \frac{\alpha_2}{2}t^2.$$

The function h is strictly concave, continuous, and satisfies $\lim_{t\to\pm\infty} h(t) = -\infty$. Therefore it attains a unique maximiser $t^* \in \mathbb{R}$. In order to find t^*, we have to be slightly careful, as the function h is not differentiable in the point $t = 0$. Since this is the only point where the concave function h is not smooth and we have already shown that a unique maximiser exists, we can conclude that we have two possibilities: Either there exists a point $t \neq 0$ such that $h'(t) = 0$, in which case this point is the maximiser of h, or there exists no such point, in which case the maximiser is $t^* = 0$.

For $t \neq 0$ we have

$$h'(t) = s - \alpha_1 \operatorname{sgn}(t) - \alpha_2 t.$$

For $t > 0$ the equation $h'(t) = 0$ is equivalent to

$$0 = s - \alpha_1 - \alpha_2 t$$

or

$$t = \frac{1}{\alpha_2}(s - \alpha_1).$$

Thus, if

$$t^* := \frac{1}{\alpha_2}(s - \alpha_1) > 0,$$

or equivalently, if

$$s > \alpha_1,$$

then

$$t^* = \frac{1}{\alpha_2}(s - \alpha_1).$$

In this case (recall that we have assumed that $s > \alpha_1$)

$$\begin{aligned} g^*(s) = h(t^*) &= st^* - \alpha_1 |t^*| - \frac{\alpha_2}{2}(t^*)^2 \\ &= \frac{1}{\alpha_2} s(s - \alpha_1) - \frac{1}{\alpha_2}|s - \alpha_1| - \frac{\alpha_2}{2}\frac{1}{\alpha_2^2}(s - \alpha_1)^2 = \frac{1}{2\alpha_2}(s - \alpha_1)^2. \end{aligned}$$

Similarly, if $t < 0$, the equation $h'(t) = 0$ is equivalent to

$$0 = s + \alpha_1 - \alpha_2 t$$

or

$$t = \frac{1}{\alpha_2}(s + \alpha_1).$$

Therefore, if

$$s < -\alpha_1,$$

then

$$t^* = \frac{1}{\alpha_2}(s + \alpha_1)$$

and

$$g^*(s) = h(t^*) = st^* - \alpha_1 |t^*| - \frac{\alpha_2}{2}(t^*)^2 = \frac{1}{2\alpha_2}(s + \alpha_1)^2.$$

In total, we thus obtain that

$$g^*(s) = \begin{cases} \dfrac{1}{2\alpha_2}(s - \alpha_1)^2 & \text{if } s > \alpha_1, \\ 0 & \text{if } -\alpha_1 \leq s \leq \alpha_1, \\ \dfrac{1}{2\alpha_2}(s + \alpha_1)^2 & \text{if } s < -\alpha_1. \end{cases}$$

Define now the *soft-thresholding* operator with parameter $\tau > 0$ as

$$S_\tau(s) := \begin{cases} s - \tau & \text{if } s > \tau, \\ 0 & \text{if } -\tau \leq s \leq \tau, \\ s + \tau & \text{if } s < -\tau. \end{cases}$$

Then we can write g^* as

$$g^*(s) = \frac{1}{2c} S_1(s)^2.$$

We note here that g^* is differentiable on the whole of $\mathbb{R}$ with

$$(g^*)'(s) = \frac{1}{\alpha_2} S_{\alpha_1}(s).$$

Slightly abusing notation, we now define the vector valued soft-thresholding operator $S_\tau\colon \mathbb{R}^d \to \mathbb{R}^d$

$$S_\tau(y) := \big(S_\tau(y_1), \dots, S_\tau(y_d)\big)$$

by applying the standard soft-thresholding operator S_α componentwise to the vector y. Then we obtain that

$$f^*(y) = \sum_{i=1}^d g^*(y_i) = \sum_{i=1}^d \frac{1}{2\alpha_2} S_{\alpha_1}(y_i)^2 = \frac{1}{2\alpha_2}\|S_{\alpha_1}(y)\|_2^2.$$

Moreover,

$$\nabla f^*(y) = \frac{1}{\alpha_2} S_{\alpha_1}(y).$$

Thus the dual problem is equivalent to the unconstrained, smooth, convex optimisation problem

$$\min_{\lambda\in\mathbb{R}^m} R(\lambda) \qquad \text{with } R(\lambda) := \frac{1}{2\alpha_2} S_{\alpha_1}(A^T\lambda)^2 - \langle \lambda, b\rangle, \tag{7}$$

and if $\lambda^* \in \mathbb{R}^m$ solves (7), then

$$x^* := \nabla f^*(A^T\lambda^*) = \frac{1}{\alpha_2} S_{\alpha_1}(A^T\lambda^*)$$

solves (6). Since f^* and thus also R are continuously differentiable, we can use a gradient based method like gradient descent or a quasi-Newton method for the solution of (7); moreover, the gradient of R easily computes as

$$\nabla R(\lambda) = A\nabla f(A^T\lambda) - b = \frac{1}{\alpha_2} A S_{\alpha_1}(A^T\lambda^*) - b.$$

It is important to note, though, that S_{α_1} is not differentiable, and thus R is not twice differentiable. Therefore we should not try to use Newton's method.

7. Sums of convex functions

In many important applications one needs to solve optimisation problems of the form

$$\min_x \big(f(x) + g(Ax)\big), \tag{8}$$

where $f\colon \mathbb{R}^d \to \mathbb{R}$ and $g\colon \mathbb{R}^m \to \mathbb{R}$ are convex functions, and $A \in \mathbb{R}^{m\times d}$ is a matrix. Examples include different regression methods, which typically result in an optimisation problem of the form

$$\min_x f(x) + \frac{1}{2}\|Ax - b\|_2^2,$$

where f is a regularisation term that encodes statistical prior knowledge about the solution.

Such problems can also be treated with convex duality after some reformulations. We start by rewriting the problem (8) as the constrained problem

$$\min_{(x,y)\in\mathbb{R}^d\times\mathbb{R}^m} \big(f(x) + g(y)\big) \qquad \text{s.t. } Ax = y. \tag{9}$$

This is a convex optimisation problem with linear constraints (in the unknowns x and y). Moreover, we see that Slater's constraint qualification is trivially satisfied by the pair $(x, y) = (0, 0)$. Thus we can apply the previously developed duality theory.

The Lagrangian of the problem (9) is the function $\mathcal{L}\colon \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$,

$$\mathcal{L}(x, y, \lambda) = f(x) + g(y) - \langle \lambda, Ax - y\rangle.$$

The dual objective function is therefore

$$q(\lambda) = \inf_{x,y}\big(f(x) + g(y) - \langle \lambda, Ax - y\rangle\big) = \inf_{x,y}\big(f(x) - \langle \lambda, Ax\rangle + g(y) + \langle \lambda, y\rangle\big).$$

We now note that we have two independent optimisation problems, one with respect to x, one with respect to y, with no coupling between the two variables. Thus we can solve the two problems separately and obtain that

$$\begin{aligned} q(\lambda) &= \inf_{x}\big(f(x) - \langle \lambda, Ax\rangle\big) + \inf_{y}\big(g(y) + \langle \lambda, y\rangle\big) \\ &= -\sup_{x}\big(\langle A^T\lambda, x\rangle - f(x)\big) - \sup_{y}\big(\langle -\lambda, y\rangle - g(y)\big) = -f^*(A^T\lambda) - g^*(-\lambda). \end{aligned}$$

The dual problem is therefore

$$\max_{\lambda\in\mathbb{R}^m}\big(-f^*(A^T\lambda) - g^*(-\lambda)\big),$$

or the equivalent minimisation problem

$$\min_{\lambda\in\mathbb{R}^m}\big(f^*(A^T\lambda) + g^*(-\lambda)\big).$$

Theorem 7.1. *Assume that $f\colon \mathbb{R}^d \to \mathbb{R}$ and $g\colon \mathbb{R}^m \to \mathbb{R}$ are convex and lower semi-continuous and that $A \in \mathbb{R}^{m\times d}$. Assume moreover that the problem*

$$\min_{x\in\mathbb{R}^d}\big(f(x) + g(Ax)\big) \tag{10}$$

has a solution x^*. Then the problem

$$\min_{\lambda\in\mathbb{R}^m}\big(f^*(A^T\lambda) + g^*(-\lambda)\big) \tag{11}$$

has a solution λ^* and we have that

$$f(x^*) + g(Ax^*) + f^*(A^T\lambda^*) + g^*(-\lambda^*) = 0.$$

Moreover, if g is differentiable at Ax^*, then

$$\lambda^* = -\nabla g(Ax^*),$$

and if f^* is differentiable at $A^T\lambda^*$, then

$$x^* = \nabla f^*(A^T\lambda^*).$$

Proof. After rewriting the problem (10) in the form (9), we can apply Theorem 4.3. From this we obtain that the problem (11) has a solution λ^* and that

$$f(x^*) + g(Ax^*) = -f^*(A^T\lambda^*) - g^*(-\lambda^*),$$

the right hand side being the optimal value of the dual problem. Now we can write

$$\begin{aligned} f(x^*) + f^*(A^T\lambda^*) + g(Ax^*) + g^*(-\lambda^*) = 0 &= \langle x^*, A^T\lambda^*\rangle - \langle x^*, A^T\lambda^*\rangle \\ &= \langle x^*, A^T\lambda^*\rangle + \langle Ax^*, -\lambda^*\rangle. \end{aligned} \tag{12}$$

From Lemma 5.4 we obtain that

$$\begin{aligned} f(x^*) + f^*(A^T\lambda^*) &\geq \langle x^*, A^T\lambda^*\rangle, \\ g(Ax^*) + g^*(-\lambda^*) &\geq \langle Ax^*, -\lambda^*\rangle. \end{aligned}$$

Thus the only possibility how we can have an equality in (12) is that, actually,

$$\begin{aligned} f(x^*) + f^*(A^T\lambda^*) &= \langle x^*, A^T\lambda^*\rangle, \\ g(Ax^*) + g^*(-\lambda^*) &= \langle Ax^*, -\lambda^*\rangle. \end{aligned}$$

Now Lemma 5.4 and Corollary 5.6 imply that

$$\begin{aligned} \nabla f(x^*) = A^T\lambda^*, &\quad \nabla f^*(A^T\lambda^*) = x^*, \\ \nabla g(Ax^*) = -\lambda^*, &\quad \nabla g^*(-\lambda^*) = Ax^*, \end{aligned}$$

whenever the gradients actually exist. □

Example 7.2. We now consider *elastic net regression*, which is the minimisation problem

$$\min_{x\in\mathbb{R}^d} \frac{1}{2}\|Ax-b\|_2^2 + \alpha_1\|x\|_1 + \frac{\alpha_2}{2}\|x\|_2^2.$$

This problem fits into our model with

$$f(x) = \alpha_1\|x\|_1 + \frac{\alpha_2}{2}\|x\|_2^2$$

and

$$g(y) = \frac{1}{2}\|y-b\|_2^2.$$

Moreover, as computed earlier, we have that

$$f^*(p) = \frac{1}{2\alpha_2}\|S_{\alpha_1}(p)\|_2^2,$$

where S_{α_1} is the soft-thresholding operator with parameter α_1, and

$$g^*(q) = \frac{1}{2}\|q\|_2^2 + \langle q, b\rangle.$$

As a consequence, the dual problem (written as minimisation problem) reads

$$\min_{\lambda\in\mathbb{R}^m} \big(f^*(A^T\lambda) + g^*(-\lambda)\big),$$

or, explicitly,

$$\min_{\lambda\in\mathbb{R}^m} \Big(\frac{1}{2\alpha_2}\|S_{\alpha_1}(A^T\lambda)\|_2^2 + \frac{1}{2}\|\lambda\|_2^2 - \langle\lambda, b\rangle\Big).$$

This is an unconstrained, convex optimisation problem with a once differentiable objective function, which can be solved by any gradient based method. Moreover, if λ^* is a solution of the dual problem, then

$$x^* = \nabla f^*(A^T\lambda^*) = \frac{1}{\alpha_2}S_{\alpha_1}(A^T\lambda^*).$$

References

[1] J. M. Borwein and A. S. Lewis. *Convex analysis and nonlinear optimization*, volume 3 of *CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC*. Springer-Verlag, New York, 2000. Theory and examples.

[2] Osman Güler. *Foundations of optimization*, volume 258 of *Graduate Texts in Mathematics*. Springer, New York, 2010.

[3] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.

Trondheim, Norway

Email address: markus.grasmair@gmail.no