LAGRANGIAN DUALITY

MARKUS GRASMAIR

1. Introduction and Motivation

Until now, we have mostly considered gradient based optimisation methods for both constrained and unconstrained optimisation problems. The only exceptions occurred, when we discussed the Nelder–Mead method near the start of the lecture, and then again for the active set method for the solution of quadratic programmes. All other methods, however, required the function we want to minimise to be continuously differentiable, so that we could make use of its gradient.

There exist a large number of applications, though, which require the solution of *non-smooth* optimisation problems. One example is the problem of sparse regression, where one tries to find a sparse solution to an under-determined regression problem

$$Ax \approx b,$$

where $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$ with $d \gg n$. One approach is *Lasso regression* (or ℓ^1-regression), which can be written as

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

where

$$\|x\|_1 = \sum_{i=1}^{d} |x_i|$$

is the 1-norm, and $\alpha > 0$ is a regularisation parameter. A more general, related idea is *elastic net regression*, requiring the solution of

$$\min_{x\in\mathbb{R}^d} \frac{1}{2}\|Ax-b\|_2^2 + \alpha\|x\|_1 + \frac{\beta}{2}\|x\|_2^2 \tag{1}$$

with regularisation parameters $\alpha > 0$ and $\beta > 0$. (Put differently, the elastic net is some weighted mean of Lasso and ridge regression.)

The idea behind these regression methods is that they promote *sparsity* in the solutions in the sense that it can be expected (and proven) that a large number of components of any solution of (1) is zero. Numerically, though, the optimisation problem (1) is more challenging than standard ridge regression, because the 1-norm is non-differentiable at all points $x \in \mathbb{R}^d$ where at least one of the components is 0 — which includes all the expected solutions. Still, one might be tempted to try to apply a gradient based method like gradient descent, because the functional, though non-smooth overall, still has a well-defined gradient in *almost all* points.

In order to see that this might be a bad idea, we consider the much simpler 2-dimensional problem

$$\min_{(x,y)\in\mathbb{R}^2} f(x,y) \qquad \text{where} \qquad f(x,y) = \frac{1}{4}x^4 + \frac{1}{2}x^2 + |y|.$$

Date: 13th April 2023.

The function f is differentiable for $y \neq 0$ with

$$\nabla f(x,y) = \begin{pmatrix} x^3 + x \\ \operatorname{sgn}(y) \end{pmatrix}.$$

However, if we apply the gradient descent method to the minimisation of this function f, the iterates converge almost randomly to some point on the x-axis, though not necessarily to the actual solution $(0,0)$ of the optimisation problem.

The goal of this note is to provide a glimpse into a different approach to the solution of such optimisation problems, which is based on the concept of *duality*. In order to introduce this approach, we have to discuss constrained optimisation problems once more. In particular, we will investigate the Lagrangian of a constrained optimisation problem in more detail.

2. Primal and dual problems

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

$$\min_x f(x) \qquad \text{subject to } \begin{cases} c_i(x) = 0, & i \in \mathcal{E}, \\ c_i(x) \geq 0, & i \in \mathcal{I}. \end{cases} \tag{2}$$

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 (2) that does not require any differentiability of the involved functions. 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 2.1. *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

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

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 (2). 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 (2) is equivalent to solving the unconstrained problem

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

or, explicitly, the *primal problem*

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

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

Definition 2.2. The *Lagrangian dual* of (P) is the optimisation problem

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

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

Remark 2.3. 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 2.4. Consider the linear programme

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

The corresponding Lagrangian is

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

Thus the dual objective function is

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

Thus we can write the dual problem as

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

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

Example 2.5. 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

$$q(\lambda) = -\lambda - \bigl(c^T x_\lambda + \lambda \|x_\lambda\|^2\bigr) = -\lambda - \frac{\|c\|^2}{2\lambda}$$

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

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

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.

3. Weak duality

We will now study the relationship between the primal and the dual problem. We have already seen in Example 2.5 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 3.1. 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 3.2. *Let X and Y be non-empty sets and let $h\colon X\times Y\to\mathbb{R}\cup\{\pm\infty\}$. Then*

$$\sup_{y\in Y}\inf_{x\in X} h(x,y) \leq \inf_{x\in X}\sup_{y\in Y} h(x,y). \tag{3}$$

Proof. Assume to the contrary that (3) 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 (3) holds. □

Lemma 3.3 (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 3.2. □

The next result shows that the duality gap can be strictly positive in some cases.

Example 3.4. 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

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

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

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

4. 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 somehow 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 4.1. 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.

Proposition 4.2. *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 4.3. 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 3.4: 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 4.1, 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 4.4. *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 4.2. □

5. 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 5.1. 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 5.2. 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 5.3 (strong duality for convex programmes)**.** *Assume that f 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 [1, Thm. 11.15]. Note that the second part of the Theorem is an immediate consequence of Theorem 4.4, once it has been established that the duality gap is zero. □

Example 5.4. We consider again the linear programme discussed in Example 2.4, 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 5.3 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 5.3 is applicable and it follows that the dual problem (L') has a solution λ^*. In addition, it can be shown (see Remark 5.5 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

$$\begin{aligned} A^T \lambda &= c, \\ Ax &\geq b, \\ \lambda &\geq 0, \\ \lambda^T (Ax - b) &= 0. \end{aligned} \tag{4}$$

Conversely, if (x^*, λ^*) solve the system (4), 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 5.5. 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 5.4.

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 5.3 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. [2, 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.

6. Dual methods for non-smooth optimisation

We will now apply the idea of Lagrangian duality to the solution of non-smooth (convex) optimisation problems.

6.1. **ℓ^1-regression.** To start with, we consider the simple ℓ^1-based regression problem

$$\min_x \frac{1}{2}\|x-b\|_2^2 + \alpha\|x\|_1, \tag{P_1}$$

where $b \in \mathbb{R}^d$ is a given vector and $\alpha > 0$. An equivalent reformulation, which allows us to apply Lagrangian duality, is the problem

$$\min_{x,y} \frac{1}{2}\|x-b\|_2^2 + \alpha\|y\|_1 \qquad \text{s.t. } x = y. \tag{$\hat{P}_1$}$$

This problem has the Lagrangian

$$\mathcal{L}(x,y;\lambda) = \frac{1}{2}\|x-b\|_2^2 + \alpha\|y\|_1 - \langle \lambda, x-y\rangle.$$

We obtain therefore the dual objective function

$$q(\lambda) = \min_{x,y\in\mathbb{R}^d} \Big(\frac{1}{2}\|x-b\|_2^2 + \alpha\|y\|_1 - \langle \lambda, x-y\rangle\Big).$$

The optimisation problem on the right has the nice property that the variables x and y can be completely separated from each other. (In fact, this is one of the main reasons why the idea of complicating the problem by doubling the number of variables makes sense at all.) Therefore we can write

$$q(\lambda) = q_1(\lambda) + q_2(\lambda)$$

with

$$\begin{aligned} q_1(\lambda) &= \min_{x\in\mathbb{R}^d} \Big(\frac{1}{2}\|x-b\|_2^2 - \langle \lambda, x\rangle\Big), \\ q_2(\lambda) &= \min_{y\in\mathbb{R}^d} \big(\alpha\|y\|_1 + \langle \lambda, y\rangle\big). \end{aligned}$$

The first optimisation problem is a quadratic problem that can be easily solved by computing the gradient and setting it to 0. We thus obtain that the solution is

$$x_\lambda = \operatorname*{arg\,min}_{x\in\mathbb{R}^d} \Big(\frac{1}{2}\|x-b\|_2^2 - \langle \lambda, x\rangle\Big) = b + \lambda,$$

and the corresponding function value is

$$q_1(\lambda) = \frac{1}{2}\|x_\lambda - b\|_2^2 - \langle \lambda, x_\lambda\rangle = -\frac{1}{2}\|\lambda\|_2^2 - \langle \lambda, b\rangle.$$

For the second term, we obtain

$$q_2(\lambda) = \min_{y\in\mathbb{R}^d} \Big(\sum_{i=1}^d \big(\alpha|y_i| + \lambda_i y_i\big)\Big) = \sum_{i=1}^d \min_{y_i\in\mathbb{R}} \big(\alpha|y_i| + \lambda_i y_i\big). \tag{5}$$

We therefore have to analyse the optimisation problem

$$\min_{t\in\mathbb{R}} \alpha|t| + \lambda_i t.$$

For the solution of this problem, we have two possibilities:

- If $|\lambda_i| \le \alpha$, then
$$\alpha|t| + \lambda_i t \ge \big(\alpha - |\lambda_i|\big)|t| \ge 0$$
for all t, and
$$\min_{t\in\mathbb{R}} \alpha|t| + \lambda_i t = 0,$$
which we obtain by setting $t = 0$.
- If $|\lambda_i| > \alpha$, then the term $\alpha|t| + \lambda_i t$ can be made arbitrarily small by letting t tend to $-\infty$ if $\lambda_i > 0$, and to $+\infty$ if $\lambda_i < 0$. Thus
$$\inf_{t\in\mathbb{R}} \alpha|t| + \lambda_i t = -\infty.$$

Inserting this result in (5), we obtain that

$$q_2(\lambda) = \sum_{i=1}^{d} \min_{y_i \in \mathbb{R}} \big(\alpha|y_i| + \lambda_i y_i\big) = \begin{cases} -\infty & \text{if } |\lambda_i| > \alpha \text{ for any } i, \\ 0 & \text{if } |\lambda_i| \le \alpha \text{ for all } i. \end{cases}$$

This can be shortened to

$$q_2(\lambda) = \begin{cases} -\infty & \text{if } \|\lambda\|_\infty > \alpha, \\ 0 & \text{if } \|\lambda\|_\infty \le \alpha, \end{cases}$$

where $\|\lambda\|_\infty := \max_i |\lambda_i|$ denotes the maximum norm of λ.

Collecting all these results, we obtain finally that

$$q(\lambda) = q_1(\lambda) + q_2(\lambda) = \begin{cases} -\infty & \text{if } \|\lambda\|_\infty > \alpha, \\ -\dfrac{1}{2}\|\lambda\|_2^2 - \langle \lambda, b\rangle & \text{if } \|\lambda\|_\infty \le \alpha. \end{cases}$$

Thus the dual problem can be written as the constrained optimisation problem

$$(D_1) \qquad \max_{\lambda} -\frac{1}{2}\|\lambda\|_2^2 - \langle \lambda, b\rangle \qquad \text{s.t. } \|\lambda\|_\infty \le \alpha.$$

The primal problem $(\hat{P}_1)$ is convex and the constraints are linear and feasible (the pair $(x, y) = (0, 0)$ is obviously feasible) and thus Theorem 5.3 is applicable. Since the primal objective function is coercive, it follows that there exists a primal solution, and consequently also a primal-dual solution (x^*, y^*, λ^*).

Moreover, if λ^* solves the dual problem (D_1), then (x^*, y^*) is a solution of

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

(this follows from the fact that (x^*, y^*, λ^*) is a saddle point of the Lagrangian). As we have seen above during the computation of the dual objective function, the x-coordinate of the solution of this problem is $b + \lambda$. Thus we conclude that

$$x^* = b + \lambda^*,$$

where λ^* solves the dual problem

$$(D_1) \qquad \max_{\lambda} -\frac{1}{2}\|\lambda\|_2^2 - \langle \lambda, b\rangle \qquad \text{s.t. } \|\lambda\|_\infty \le \alpha.$$

It thus remains to solve (D_1). Here we note that both the objective function *and the constraint* can be separated into the different components of the vector λ. That is, the component λ_i^* of the solution of (D_1) solves the problem

$$\max_{\lambda_i} -\frac{1}{2}\lambda_i^2 - \lambda_i b_i \qquad \text{s.t. } |\lambda_i| \le \alpha.$$

The solution of this constrained optimisation problem in one variable can easily be computed as

$$\lambda_i^* = \begin{cases} \alpha & \text{if } b_i < -\alpha, \\ -b_i & \text{if } |b_i| \le \alpha, \\ -\alpha & \text{if } b_i > \alpha. \end{cases}$$

Therefore,

$$x_i^* = b_i + \lambda_i^* = \begin{cases} b_i + \alpha & \text{if } b_i < -\alpha, \\ 0 & \text{if } |b_i| \le \alpha, \\ b_i - \alpha & \text{if } b_i > \alpha. \end{cases}$$

Remark 6.1. The *soft-thresholding operator* with parameter α is defined as the function $S_\alpha \colon \mathbb{R} \to \mathbb{R}$,

$$S_\alpha(t) = \begin{cases} t + \alpha & \text{if } t < -\alpha, \\ 0 & \text{if } |t| \leq \alpha, \\ t - \alpha & \text{if } t > \alpha. \end{cases}$$

Using this operator, the solution of the problem $(\hat{P}_1)$ can be written as

$$x_i^* = S_\alpha(b_i),$$

or, more succintely, as

$$x^* = S_\alpha(b),$$

where the action of S_α on the vector b is interpreted componentwise.

6.2. **Elastic net regression.** We now consider the more complicated case of elastic net regression

$$\min_{x \in \mathbb{R}^d} \frac{1}{2}\|Ax - b\|_2^2 + \alpha\|x\|_1 + \frac{\beta}{2}\|x\|_2^2. \tag{P_e}$$

We use the same approach as before, and rewrite this as the constrained problem

$$\min_{x,\, y \in \mathbb{R}^d} \frac{1}{2}\|Ax - b\|_2^2 + \frac{\beta}{2}\|x\|_2^2 + \alpha\|y\|_1 \qquad \text{s.t. } x = y. \tag{$\hat{P}_e$}$$

Here we obtain the Lagrangian

$$\mathcal{L}(x, y, \lambda) = \frac{1}{2}\|Ax - b\|_2^2 + \frac{\beta}{2}\|x\|_2^2 + \alpha\|y\|_1 - \langle \lambda, x - y \rangle$$

and the dual objective function

$$q(\lambda) = \min_{x,\, y \in \mathbb{R}^d} \Big(\frac{1}{2}\|Ax - b\|_2^2 + \frac{\beta}{2}\|x\|_2^2 + \alpha\|y\|_1 - \langle \lambda, x - y \rangle \Big).$$

Again, this problem separates into the two problems

$$\begin{aligned} q_1(\lambda) &= \min_{x \in \mathbb{R}^d} \Big(\frac{1}{2}\|Ax - b\|_2^2 + \frac{\beta}{2}\|x\|_2^2 - \langle \lambda, x \rangle \Big) \\ q_2(\lambda) &= \min_{y \in \mathbb{R}^d} \big(\alpha\|y\|_1 + \langle \lambda, y \rangle \big), \end{aligned}$$

with

$$q(\lambda) = q_1(\lambda) + q_2(\lambda).$$

The second problem is precisely the same as before, and we therefore obtain that

$$q_2(\lambda) = \begin{cases} -\infty & \text{if } \|\lambda\|_\infty > \alpha, \\ 0 & \text{if } \|\lambda\|_\infty \leq \alpha, \end{cases}$$

which eventually will result in a constraint for the dual optimisation problem.

The optimisation problem for q_1 is again quadratic and thus can be solved by computing the gradient and setting it to zero. This results in the solution

$$x_\lambda = \operatorname*{arg\,min}_{x \in \mathbb{R}^d} \Big(\frac{1}{2}\|Ax - b\|_2^2 + \frac{\beta}{2}\|x\|_2^2 - \langle \lambda, x \rangle \Big) = (\beta \operatorname{Id} + A^T A)^{-1}(A^T b + \lambda).$$

We therefore obtain the dual optimisation problem

$$\max_{\lambda} \Big(\frac{1}{2}\|Ax_\lambda - b\|_2^2 + \frac{\beta}{2}\|x_\lambda\|_2^2 - \langle \lambda, x_\lambda \rangle \Big) \qquad \text{s.t. } \|\lambda\|_\infty \leq \alpha, \tag{D_e}$$

where x_λ is given as

$$x_\lambda = (\beta \operatorname{Id} + A^T A)^{-1}(A^T b + \lambda). \tag{6}$$

In principle, it would also be possible to insert the definition of x_λ directly into (D_e) and then simplify the result, in order to obtain an explicit expression in terms of λ.

In contrast to the case of the (dual) ℓ_1 problem (D_1), it is no longer possible to provide a simple analytic solution formula for (D_e). The reason for this is the presence of the matrix A, which couples the different components of x_λ (and thus λ). Instead, it is necessary to solve this problem numerically. For this, there are several options available. For instance, it is possible to use an active set method, as the constraint $\|\lambda\|_\infty \leq \alpha$ can also be written as the set of linear constraints $-\alpha \leq \lambda_i \leq \alpha$ for all $1 \leq i \leq d$. Instead, we will discuss, how a projected gradient ascent method can be used.[1]

In order to define the gradient ascent method, it is necessary to compute the gradient $\nabla q_1(\lambda)$. At first glance, this appears to necessitate an explicit expression of q_1 in terms of λ after all. However, it turns out that we can circumvent this computation. In order to do so, we define the function $g\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$,

$$g(x,\lambda) = \frac{1}{2}\|Ax-b\|_2^2 + \frac{\beta}{2}\|x\|_2^2 - \langle \lambda, x\rangle.$$

Then

$$x_\lambda = \arg\min_x g(x,\lambda) \qquad \text{and} \qquad q_1(\lambda) = g(x_\lambda,\lambda).$$

In particular, we have that

$$\nabla_x g(x_\lambda,\lambda) = 0.$$

Thus an application of the chain rule implies that

$$\nabla q_1(\lambda) = \nabla g(x_\lambda,\lambda) = (D_\lambda x_\lambda)^T \nabla_x g(x_\lambda,\lambda) + \nabla_\lambda g(x_\lambda,\lambda) = -x_\lambda.$$

A gradient ascent step for the function q_1 with step length $\tau > 0$ thus simply becomes

$$\lambda + \tau \nabla q_1(\lambda) = \lambda - \tau x_\lambda,$$

where x_λ is defined in (6).

Moreover, the projection π_Ω on the feasible set $\Omega := \big\{\lambda \in \mathbb{R}^d : \|\lambda\|_\infty \leq \alpha\big\}$ can be easily computed by "cutting off" all components of λ that are larger than α or smaller than $-\alpha$. Explicitly, we have that the i-th component of $\pi_\Omega(\lambda)$ is given as

$$\big(\pi_\Omega(\lambda)\big) = \min\big\{\alpha, \max\{-\alpha, \lambda_i\}\big\}.$$

We thus obtain the dual projected gradient ascent algorithm summarised as Algorithm 1. Provided that τ is chosen sufficiently small, this algorithm will converge to the solution of (P_e) (essentially, this follows from our analysis of the projected gradient descent method from exercise sheet 5).

Initialisation: Choose a sufficiently small step length $0 < \tau$ and $\lambda \in \mathbb{R}^d$;
Set $x \leftarrow (\beta \operatorname{Id} + A^T A)^{-1}(A^T b + \lambda)$;
while *not converged* **do**
  $\lambda \leftarrow \lambda - \tau x_\lambda$;
  $\lambda_i \leftarrow \min\big\{\alpha, \max\{-\alpha, \lambda_i\}\big\}$ for $1 \leq i \leq d$;
  $x \leftarrow (\beta \operatorname{Id} + A^T A)^{-1}(A^T b + \lambda)$;
end

Algorithm 1: Dual projected gradient asscent method for the solution of (P_e).

[1] Note here that we want to solve a maximisation problem, not a minimisation problem. Thus we need to use steps in gradient direction, instead of the negative gradient direction.

References

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

[2] 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