MSE2122 - Nonlinear Optimization Lecture Notes VIII

Fernando Dias (based on previous version by Fabricio Oliveira)

October 18, 2023

Abstract

In this lecture, we look into the notion of Lagrangian duality, which consists of a theoretical framework that allows deriving not only optimality conditions for constrained optimisation problems, but also solution methods build upon results from duality theory. First, we consider Lagrangian duality under the notion of relaxation, and show how primal and dual solutions are related. Then we describe the notion of strong duality, a key property that connect Lagrangian duality and optimality conditions for constrained problems. Lastly, we discuss the subgradient method, one of the most widespread methods for solving constrained optimisation problems using Lagrangian duals.

Contents

1 The concept of relaxation

The idea of using relaxations is central in several constrained optimisation methods. Generally, it consists of techniques that remove constraints from the problem to allow for a version, i.e., a *relaxation*, that is simpler to solve and/or can provide information to be used for solving the original problem.

A classic example of the employment of relaxations for solving constrained problems is the branch-and-bound method that uses linear (continuous) relaxations of integer problems to guide the search for optimal solutions that are also integers. However, there are several other settings in which relaxations are purposely derived to lead to problems with a convenient structure that can be exploited.

Let $f : \mathbb{R}^n \to \mathbb{R}$ and $S \subseteq \mathbb{R}^n$. Consider the following problem:

$$(P): \ \min. \ \{f(x) : x \in S\}$$

Definition 1.1 provides the conditions for P_R to be a *relaxation* of P, where:

$$(P_R): \ \min. \ \{f_R(x) : x \in S_R\}$$

with $f_R : \mathbb{R}^n \to \mathbb{R}$, $S_R \subseteq \mathbb{R}^n$.

Definition 1.1. Relaxation

P_R is a relaxation of P if and only if:
1. $f_R(x) \leq f(x)$, for all $x \in S$;
2. $S \subseteq S_R$.

In specific, P_R is said to be a relaxation for P if $f_R(x)$ bounds $f(x)$ from below (in a minimisation setting) for all $x \in S$ and the enlarged feasible region S_R contains S.

The motivation for using relaxations arises from the possibility of finding a solution to the original problem P by solving P_R. Such a strategy would only make sense if P_R possess some attractive property or feature that we can use in our favour to, e.g., improve solution times or create separability that can be further exploited using parallelised computation (which we will discuss in more detail in the upcoming lectures). Theorem 1.2 presents the technical result that allows for using relaxations for solving P.

Theorem 1.2. Relaxation theorem

Let us define:

$$(P): \ \min. \ \{f(x) : x \in S\} \quad \text{and} \quad (P_R): \min. \ \{f_R(x) : x \in S_R\}$$

If P_R is a relaxation of P, then the following hold:

1. if P_R is infeasible, so is P;
2. if $\overline{x}_R$ is an optimal solution to P_R such that $\overline{x}_R \in S$ and $f_R(\overline{x}_R) = f(\overline{x}_R)$, then $\overline{x}_R$ is optimal to P as well.

Proof. Result (1) follows since $S \subseteq S_R$. To show (2), notice that $f(\overline{x}_R) = f_R(\overline{x}_R) \leq f_R(x) \leq f(x)$ for all $x \in S$.

😎

The first part of the proof is a consequence of $S \subset S_R$, meaning that if $x \notin S$, then $x \notin S_R$. The second part combines the optimality of $\overline{x}_R$ (first inequality) and the definition of a relaxation (second inequality) to derive the optimality condition of $\overline{x}_R$ for P, which is $f(\overline{x}_R) \leq f(\overline{x})$ for all $x \in S$.

2 Lagrangian dual problems

Lagrangian duality is the body of theory supporting the use of *Lagrangian relaxations* to solve constrained optimisation problems. In what follows, we refer to the relaxation obtained using Lagrangian

duality as the *(Lagrangian) dual* problem. Consequently, we refer to the original problem as the *primal* problem.

Let $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$, $h : \mathbb{R}^n \to \mathbb{R}^l$, and assume that $X \subseteq \mathbb{R}^n$ is an open set. Then, consider P defined as:

$$(P): \quad \text{min.} \quad f(x)$$
$$\text{subject to: } g(x) \leq 0$$
$$h(x) = 0$$
$$x \in X.$$

For a given set of *dual variables* $(u, v) \in \mathbb{R}^{m+l}$ with $u \geq 0$, the *Lagrangian relaxation* (or *Lagrangian dual function*) of P is:

$$(D): \quad \theta(u, v) = \inf_{x \in X} \phi(x, u, v)$$

where:

$$\phi(x, u, v) := f(x) + u^\top g(x) + v^\top h(x)$$

is the *Lagrangian function.*

Notice that the Lagrangian dual function $\theta(u, v)$ has a built-in optimisation problem in x, meaning that evaluating $\theta(u, v)$ still requires solving an optimisation problem, which amounts to finding the minimiser $\overline{x}$ for $\phi(x, u, v)$, given (u, v).

2.1 Weak and strong duality

Weak and robust duality are, to some extent, consequences of Theorem 1.2 and the fact that the Lagrangian relaxation is indeed a relaxation of P. We start with the equivalent to Definition 1.1, referred to as *weak duality*.

> **Theorem 2.1. Weak Lagrangian duality**
>
> Let x be a feasible solution to P, and let (u, v) be such that $u \geq 0$, i.e., feasible for D. Then $\theta(u, v) \leq f(x)$.

Proof. From feasibility, $u \geq 0$, $g(x) \leq 0$ and $h(x) = 0$. Thus, we have that
$$\theta(u, v) = \inf_{x \in X} \{ f(x) + u^\top g(x) + v^\top h(x) \} \leq f(x) + u^\top g(x) + v^\top h(x) \leq f(x).$$
Which completes the proof.

😎

The proof uses the fact that the infimal of the Lagrangian function $\phi(x, u, v)$, and in fact, any value for $\phi(x, u, v)$ for all primal feasible x and dual feasible $u \geq 0$ (a condition for the Lagrangian relaxation to be indeed a relaxation) bounds to $f(x)$. This arises from observing $g(x) \leq 0$ for a feasible x.

The *Lagrangian dual problem* is the problem used to obtain the best possible relaxation bound $\theta(u, v)$ for $f(x)$, in light of Theorem 2.1. This can be achieved by optimising $\theta(u, v)$ in the space of the dual variables (u, v), that is:

$$(D): \quad \theta(u, v) = \inf_{x \in X} \phi(x, u, v).$$

The use of Lagrangian dual problems is an alternative for dealing with constrained optimisation problems, as they allow to convert of the constrained primal into a (typically) unconstrained dual that is potentially easier to handle or present exploitable properties that can benefit specialised algorithms, such as separability.

Employing Lagrangian relaxations to solve optimisation problems is possible due to the following important results, which are posed as corollaries of Theorem 2.1.

> **Corollary 2.2** (Weak Lagrangian duality II).
> $$\sup_{u,v}\{\theta(u,v) : u \geq 0\} \leq \inf_x\{f(x) : g(x) \leq 0, h(x) = 0, x \in X\}.$$

> **Proof.** We have $\theta(u,v) \leq f(x)$ for any feasible x and (u,v), thus implying: $\sup_{u,v}\{\theta(u,v) : u \geq 0\} \leq \inf_x\{f(x) : g(x) \leq 0, h(x) = 0, x \in X\}$ 😎

> **Corollary 2.3** (Strong Lagrangian duality). If $f(\overline{x}) = \theta(\overline{u}, \overline{v})$, $\overline{u} \geq 0$, and $\overline{x} \in \{x \in X : g(x) \leq 0, h(x) = 0\}$, then $\overline{x}$ and $(\overline{u}, \overline{v})$ are optimal solutions to P and D, respectively.

> **Proof.** Use part (2) of Theorem 1.2 with D being a Lagrangian relaxation. 😎

Notice that Corollary 2.3 implies that if the optimal solution value of the primal and the dual problems match, the respective primal and dual solutions are optimal. However, to use Lagrangian relaxations to solve constrained optimisation problems, we need the opposite clause also to hold, which is called *strong duality* and, unfortunately, does not always hold.

2.1.1 Geometric interpretation of Lagrangian duality

Let us focus on a graphical interpretation of Lagrangian dual problems to investigate the cases in which strong duality can hold. For that, let us first define some auxiliary elements.

For the sake of simplicity, consider $(P) : \min.\ \{f(x) : g(x) \leq 0, x \in X\}$ with $f : \mathbb{R}^n \mapsto \mathbb{R}$, a single constraint $g : \mathbb{R}^n \mapsto \mathbb{R}$ and $X \subseteq \mathbb{R}^n$ an open set.

Let us define the mapping $G = \{(y, z) : y = g(x), z = f(x), x \in X\}$, which consists of a mapping of points $x \in X$ to the (y, z)-space obtained using $(f(x), g(x))$. In this setting, solving P means finding a point with minimum ordinate z for which $y \leq 0$. Figure 1 illustrates this setting.
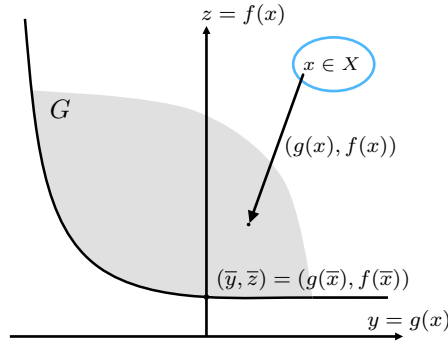

Figure 1: Illustration of the mapping G, in which solving P amounts to finding the lowermost point on the vertical axis (the ordinate) still contained within G.

Now, assume that $u \geq 0$ is given. The Lagrangian function is given by:

$$\theta(u) = \min_x\{f(x) + ug(x) : x \in X\},$$

, which can be represented by a hyperplane of the form $z = \alpha - uy$. Therefore, optimising the Lagrangian dual problem $(D) : \sup_u\{\theta(u)\}$ consists of finding the slope $-u$ that would achieve the maximum intercept on the ordinate z while being a supporting hyperplane for G. Figure 2 illustrates this effect. Notice that, in this case, the optimal values of the primal and dual problems coincide.

The *perturbation function* $v(y) = \min.\ \{f(x) : g(x) \leq y, x \in X\}$ is an analytical tool that plays an important role in understanding when strong duality holds, which, in essence, is the underlying reason why the optimal values of the primal and dual problems coincide.

Specifically, notice that $v(y)$ is the greatest monotone nonincreasing lower envelope of G. Moreover, the reason why $f(\overline{x}) = \theta(\overline{u})$ is related to the convexity of $v(y)$, which implies that:
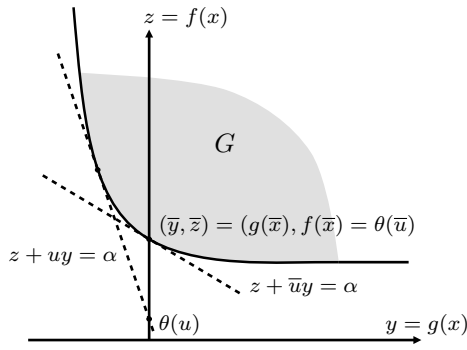
Figure 2: Solving the Lagrangian dual problem is the same as finding the coefficient u such that $z = \alpha - uy$ is a supporting hyperplane of G with the uppermost intercept α. Notice that, for $\overline{u}$, the hyperplane supports G at the same point that solves P.

$$v(y) \geq v(0) - \overline{u}y \text{ for all } y \in \mathbb{R}.$$

Notice that this is a consequence of Theorem 12 from Lecture 2 (that states that convex sets have supporting hyperplanes for all points on their boundary) and Theorem 5 in Lecture 3 (that convex functions have convex epigraphs)

A *duality gap* exists when the perturbation function $v(y)$ does not have supporting hyperplanes within all its domains, which is otherwise the case when $v(y)$ is convex. Figure 3 illustrates a case in which $v(y)$ is not convex and therefore $\theta(\overline{u}) < f(\overline{x})$.
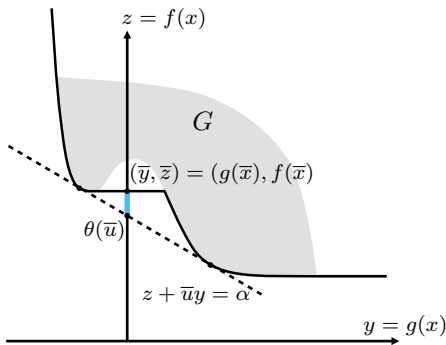

Figure 3: An example in which the perturbation function $v(y)$ is not convex. Notice the consequent mismatch between the intercept of the supporting hyperplane and the lowermost point on the ordinate still contained in G.

Let us illustrate the above with two numerical examples. First, consider the following problem:

$$(P): \quad \text{min.} \quad x_1^2 + x_2^2$$
$$x_1 + x_2 \geq 4$$
$$x_1, x_2 \geq 0.$$

The Lagrangian dual function is given by:

$$(D): \quad \theta(u) = \inf\{x_1^2 + x_2^2 + u(-x_1 - x_2 + 4) : x_1, x_2 \geq 0\}$$
$$= \inf\{x_1^2 - ux_1 : x_1 \geq 0\} + \inf\{x_2^2 - ux_2 : x_2 \geq 0\} + 4u$$
$$= \begin{cases} -1/2u^2 + 4u, \text{ if } u \geq 0 \\ -4u, \quad 1 + 4u^2 \text{ if } u < 0. \end{cases}$$

Figures 4a and 4b provide a graphical representation of the primal problem P and dual problem D. As can be seen, both problems have as optimal value $f(\overline{x}_1, \overline{x}_2) = \theta(\overline{u}) = 8$, with the optimal solution $\overline{x} = (2, 2)$ for P and $\overline{u} = 4$ for D.
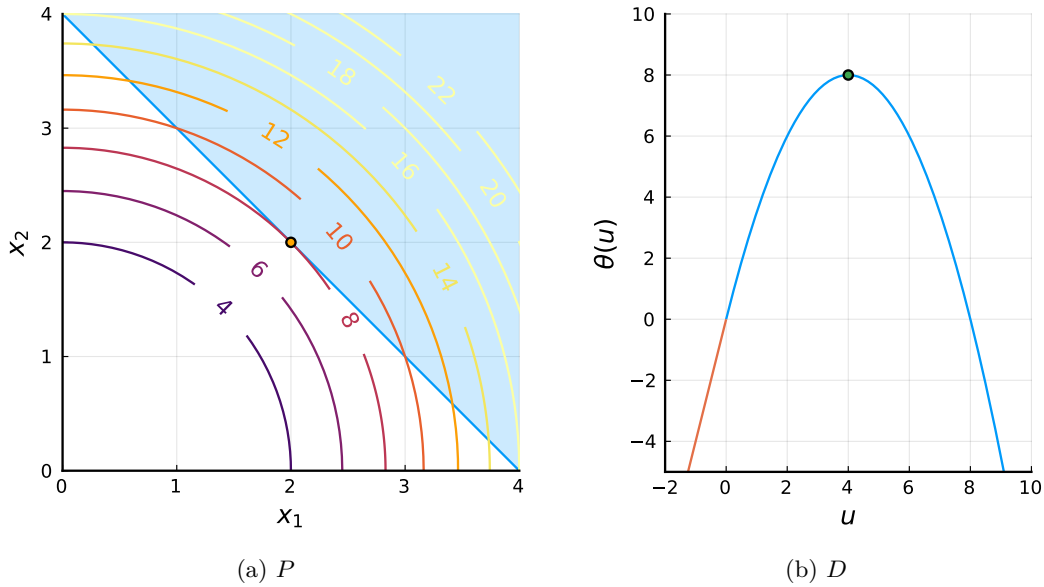
(a) P

(b) D

Figure 4: The primal problem P as a constrained optimisation problem, and the dual problem D, as an unconstrained optimisation problem. Notice how the Lagrangian dual function is discontinuous due to the implicit minimisation in x of $\theta(u) = \inf_{x \in X} \phi(x, u)$.
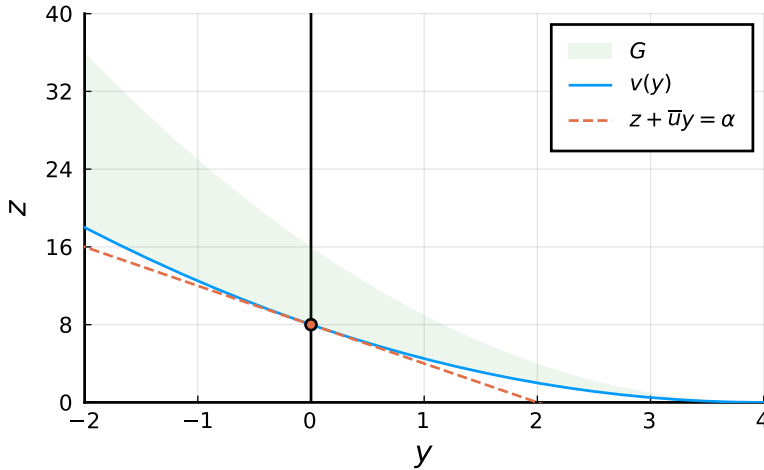

Figure 5: The G mapping for the first example.

To draw the (g, f) map of X, we proceed as follows. First, notice that:
$$v(y) = \min. \ \{x_1^2 + x_2^2 : -x_1 - x_2 + 4 \geq y\}$$
which shows that $(x_1, x_2) = (0, 0)$ if $y > 4$. For $y \leq 4$, $v(y)$ can be equivalently rewritten as:

$$v(y) = \min. \ \{x_1^2 + x_2^2 : -x_1 - x_2 + 4 = y\}.$$
Let $h(x) = -x_1 - x_2 + 4$ and $f(x) = x_1^2 + x_2^2$. Now, the optimality conditions for $\overline{x}$ to be an optimum for v are such that:

$$\nabla f(\overline{x}) + u \nabla h(\overline{x}) = 0 \Rightarrow \begin{cases} 2x_1 - u = 0 \\ 2x_2 - u = 0 \end{cases} \Rightarrow \overline{x}_1 = \overline{x}_2 = u/2.$$

From the definition of $h(x)$, we see that $u = 4 - y$, and thus $\overline{x} = (\frac{4-y}{2}, \frac{4-y}{2})$, which, substituting in $f(x)$ gives $v(y) = (4 - y)^2/2$. Note that $v(y) \geq v(0) - \overline{u}y$ holds for all $y \in \mathbb{R}$, that is, $v(y)$ is convex. Also, notice that the supporting hyperplane is exactly $z = 8 - 4y$.

Now, let us consider a second example: the feasible set is not convex, so the mapping G will not be convex either. For that, consider the problem:

6

$$(P): \quad \text{min.} \quad -2x_1 + x_2$$
$$\text{subject to: } x_1 + x_2 = 3$$
$$x_1, x_2 \in X.$$

Where $X = \{(0,0),(0,4),(4,4),(4,0),(1,2),(2,1)\}$. The optimal point $\bar{x} = (2,1)$. The Lagrangian dual function is given by:

$$\theta(v) = \min\{(-2x_1 + x_2) + v(x_1 + x_2 - 3) : (x_1, x_2) \in X\}$$

$$= \begin{cases} -4 + 5v, & \text{if } v \leq -1 \\ -8 + v, 5 & \text{if } -1 \leq v \leq 2 \\ -3v, +p_1 & \text{if } v \geq 2. \end{cases}$$

Figure 6a provides a graphical representation of the problem. Notice that to obtain the Lagrangian dual function, one must simply take the lowermost segments of the hyperplanes obtained when considering each $x \in X$, which leads to a piecewise concave function, as represented in Figure 6b.
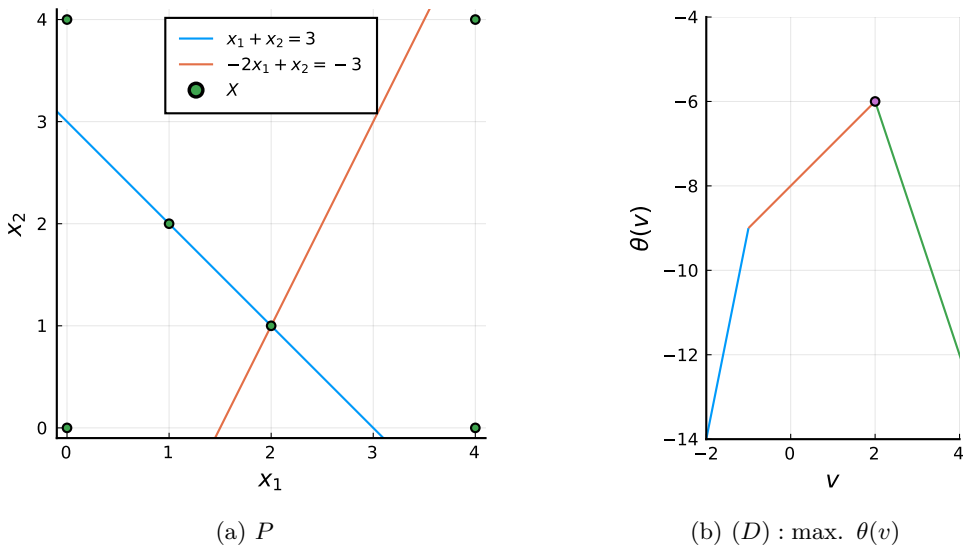

(a) P

(b) $(D): \text{max. } \theta(v)$

Figure 6: The primal problem P as a constrained optimisation problem and the dual problem D. Notice how the Lagrangian dual function is concave and piecewise linear, despite the nonconvex nature of P.

Similarly to the previous example, we can plot the G mapping, which in this case consists of the points $x \in X$ mapped as $(h(x), f(x))$, with $h(x) = x_1 + x_2 - 3$ and $f(x) = -2x_1 + x_2$. Notice that $v(y)$ in this case is discontinuous, represented by the three lowermost points. $v(y)$ does not have a supporting hyperplane at the minimum of P, which illustrates the existence of a duality gap, as stated by the fact that $-3 = f(\bar{x}) > \theta(\bar{v}) = -6$.
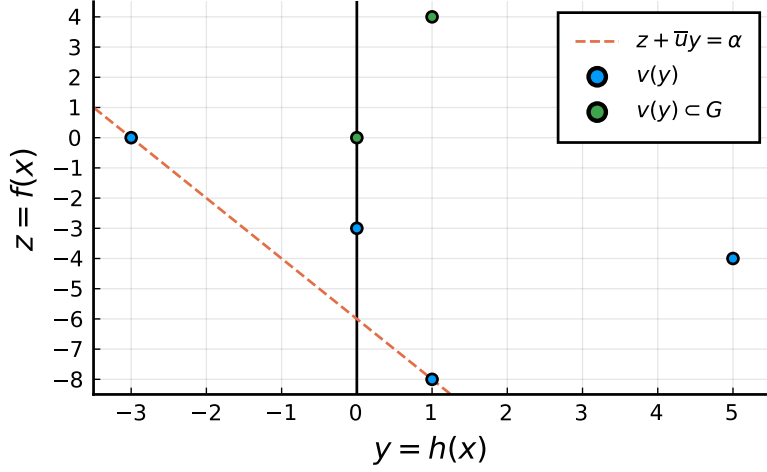
Figure 7: The G mapping for the second example. The blue dots represent the perturbation function $v(y)$, which is not convex and thus cannot be supported everywhere. Notice the duality gap represented by the difference between the intercept of $z = -6 - 2y$ and the optimal value of P at $(0,-3)$.

2.1.2 Strong duality

From the previous graphical interpretation and related examples, it becomes clear that there is a strong tie between strong duality and the convexity of P. This is formally described in Theorem 2.4.

> **Theorem 2.4. Strong duality**
>
> Let $X \subseteq \mathbb{R}^n$ be a nonempty convex set. Moreover, let $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ be convex functions, and let $h : \mathbb{R}^n \to \mathbb{R}^l$ be an affine function: $h(x) = Ax - b$. Suppose that Slater's constraint qualification holds. Then:
>
> $$\inf_x \{f(x) : g(x) \leq 0, h(x) = 0, x \in X\} = \sup_{u,v} \{\theta(u,v) : u \geq 0\},$$
>
> where $\theta(u,v) = \inf_{x \in X} \{f(x) + u^\top g(x) + v^\top h(x)\}$ is the Lagrangian function. Furthermore, if $\inf_x \{f(x) : g(x) \leq 0, h(x) = 0, x \in X\}$ is finite and achieved at $\overline{x}$, then $\mathbf{l}\sup_{u,v} \{\theta(u,v) : u \geq 0\}$ is achieved at $(\overline{u}, \overline{v})$ with $\overline{u} \geq 0$ and $\overline{u}^\top g(\overline{x}) = 0$.

The proof for the strong duality theorem follows the following outline:

1. Let $\gamma = \inf_x \{f(x) : g(x) \leq 0, h(x) = 0, x \in X\}$. Suppose that $-\infty < \gamma < \infty$, hence finite (for unbounded problems, $f(x) = -\infty$ implies $\theta(u,v) = -\infty$ since $\theta(u,v) \leq f(x)$ from Theorem 2.1; the right-hand side holds by assumption of the existence of a feasible point from Slater's constraint qualification).

2. Formulate the inconsistent system:

$$f(x) - \gamma < 0, \quad g(x) \leq 0, \quad h(x) = 0, \quad x \in X.$$

3. Use the separation theorem (or a variant form of Farkas theorem) to show that $(\overline{u}_0, \overline{u}, \overline{v})$ with $\overline{u}_0 > 0$ and $\overline{u} \geq 0$ exists such that, after scaling using $\overline{u}_0$ one obtains $\theta(\overline{u}, \overline{v}) := f(x) + \overline{u}^\top g(x) + \overline{v}^\top h(x) \geq \gamma$, $x \in X$, which requires the assumption of Slater's constraint qualification.

4. From weak duality (Theorem 2.1), we have that $\theta(\overline{u}, \overline{v}) \leq \gamma$, which combined with the above, yields $\theta(\overline{u}, \overline{v}) = \gamma$.

5. Finally, an optimal $\overline{x}$ solving the primal problem implies that $g(\overline{x}) \leq 0$, $h(\overline{x}) = 0$, $\overline{x} \in X$, and $f(x) = \gamma$. From 3, we have $\overline{u}^\top g(\overline{x}) \geq 0$. As $g(\overline{x}) \leq 0$ and $\overline{u} \geq 0$, $\overline{u}^\top g(\overline{x}) \geq 0 = 0$.

The proof uses a variant of the Farkas theorem that states the existence of a solution for the system $u_0(f(x) - \gamma) + u^\top g(x) \geq 0, x \in X$ with $(u_0 u, v) \neq 0$, what can be shown to be the case if Slater's constraint qualification holds. This, combined with weak duality stated in Theorem 2.1, yields strong duality.

2.2 Employing Lagrangian duality for solving optimisation problems

Weak duality can be used to derive a stopping criterion for solution methods that generate both primal and dual feasible solutions, also known as primal-dual pairs. Such methods are typically referred to as primal-dual methods, the primal-dual interior point method (which we will discuss in detail in an upcoming lecture) and perhaps the most widely known.

For feasible x and (u, v), one can bound how suboptimal $f(x)$ is, by noticing that:

$$f(x) - f(\overline{x}) \leq f(x) - \theta(u, v),$$

which is a consequence of $f(\overline{x}) \geq \theta(u, v)$ (i.e., weak duality). We say that x is ϵ-optimal, with $\epsilon = f(x) - \theta(u, v)$.

In essence, (u, v) is a certificate of (sub-)optimality of x, as (u, v) proves that x is ϵ-optimal. Moreover, in case strong duality holds, under the conditions of Theorem 6, one can expect ϵ to converge to zero.

To see how this is the case, observe the following. First, as can be seen in Theorem 2.4, a consequence of strong duality is that complementarity conditions $\overline{u}^\top g(\overline{x}) \geq 0 = 0$ hold for an optimal primal-dual pair $(\overline{x}, (\overline{u}, \overline{v}))$. Secondly, notice that, by definition, $\overline{x}$ and $(\overline{u}, \overline{v})$ are primal and dual feasible, respectively.

The last component missing is to notice that, if $\overline{x}$ is a minimiser for $\phi(x, \overline{u}, \overline{v}) = f(x) + \overline{u}^\top g(x) + \overline{v}^\top h(x)$, then we must have:

$$\nabla f(\overline{x}) + \sum_{i=1}^{m} u_i \nabla g_i(\overline{x}) + \sum_{i=1}^{l} v_i \nabla h_i(\overline{x}) = 0.$$

Combining the above, one can see that we have listed all of the KKT optimality conditions, which under the assumptions of Theorem 2.4 are known to be necessary and sufficient for global optimality. In this case, any primal-dual pair for which the objective function values match will automatically be a point satisfying the KKT conditions and, therefore, globally optimal. This provides an alternative avenue to search for optimal solutions, relying on Lagrangian dual problems.

2.3 Saddle point optimality and KKT conditions

An alternative perspective for establishing necessary and sufficient conditions for strong duality involves identifying saddle points for the Lagrangian dual problem.

Let us first define saddle points in the context of Lagrangian duality. Let:

$$(P): \ \min. \ \{f(x) : g(x) \leq 0, h(x) = 0, x \in X\}.$$

Let us define the Lagrangian function $\phi(x, u, v) = f(x) + u^\top g(x) + v^\top h(x)$. A solution $(\overline{x}, \overline{u}, \overline{v})$ is called a *saddle point* if $\overline{x} \in X$, $\overline{u} \geq 0$, and:

$$\phi(\overline{x}, u, v) \leq \phi(\overline{x}, \overline{u}, \overline{v}) \leq \phi(x, \overline{u}, \overline{v})$$

for all $x \in X$ and (u, v) such that $u \geq 0$.

Notice that this definition implies that:

- $\overline{x}$ minimises $\phi(x, u, v)$ when (u, v) is fixed at $(\overline{u}, \overline{v})$;
- $(\overline{u}, \overline{v})$ maximises $\phi(x, u, v)$ when x is fixed at $\overline{x}$.

This insight allows for the development of methods that can alternatively solve the Lagrangian dual problem in the space of primal variables x and dual variables (u, v) in a block-coordinate descent fashion.

Theorem 2.5 establishes the relationship between the existence of saddle points for Lagrangian dual problems and zero duality gaps.

> **Theorem 2.5. Saddle point optimality and zero duality gap**
>
> A solution $(\overline{x}, \overline{u}, \overline{v})$ with $\overline{x} \in X$ and $\overline{u} \geq 0$ is a saddle point for the Lagrangian function $\phi(x, u, v) = f(x) + u^\top g(x) + v^\top h(x)$ if and only if:
>
> 1. $\phi(\overline{x}, \overline{u}, \overline{v}) = \min. \{\phi(x, \overline{u}, \overline{v}) : x \in X\}$
> 2. $g(\overline{x}) \leq 0$, $h(\overline{x}) = 0$, and
> 3. $\overline{u}^\top g(\overline{x}) = 0$
>
> Moreover, $(\overline{x}, \overline{u}, \overline{v})$ is a saddle point if and only if $\overline{x}$ and $(\overline{u}, \overline{v})$ are optimal solutions for the primal (P) and dual (D) problems, respectively, with $f(\overline{x}) = \theta(\overline{u}, \overline{v})$.

From Theorem 2.5, it becomes clear that there is a strong connection between the existence of saddle points and the KKT conditions for optimality. Figure 8 illustrates the existence of a saddle point and the related zero optimality gap.
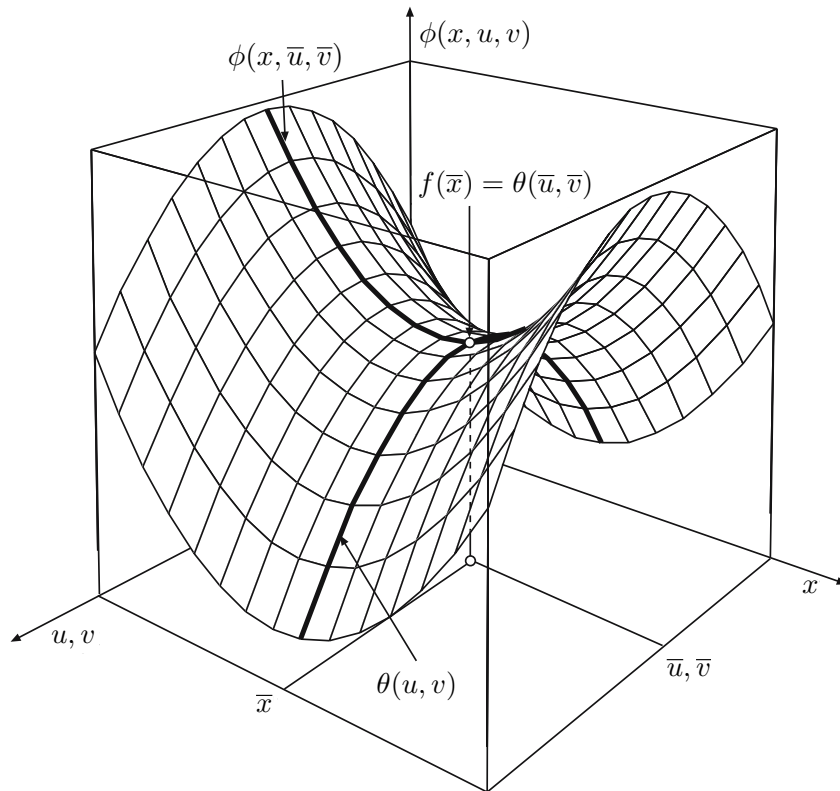

Figure 8: Illustration of a saddle point for the Lagrangian dual problem

3 Properties of Lagrangian functions

Lagrangian duals are a useful framework for devising solution methods for constrained optimisation problems if solving the dual problem can be done efficiently or exposes some exploitable structure.

One important property that Lagrangian dual functions present is that they are *concave piecewise linear* in the dual multipliers. Moreover, they are continuous and thus have subgradients everywhere. However, they are typically not differentiable, requiring the employment of a *nonsmooth optimisation* method to be appropriately solved. Theorem 3.1 establishes the concavity of the Lagrangian dual function.

> **Theorem 3.1. Concavity of Lagrangian dual functions**
>
> Let $X \subseteq \mathbb{R}^n$ be a nonempty compact set, and let $f : \mathbb{R}^n \to \mathbb{R}$ and $\beta : \mathbb{R}^n \to \mathbb{R}^{m+l}$, with $w^\top \beta(x) = \binom{u}{v}^\top \binom{g(x)}{h(x)}$ be continuous. Then $\theta(w) = \inf_x \{f(x) + w^\top \beta(x) : x \in X\}$ is concave in $\mathbb{R}^{m+l}$

Proof. Since f and β are continuous and X is compact, θ is finite on R^{m+l}. Let w_1, $w_2 \in \mathbb{R}^{m+l}$, and let $\lambda \in (0,1)$. We have:

$$
\begin{aligned}
\theta[\lambda w_1 + (1-\lambda)w_2] &= \inf_x\{f(x) + [\lambda w_1 + (1-\lambda)w_2]^\top \beta(x) : x \in X\} \\
&= \inf_x\{\lambda[f(x) + w_1^\top \beta(X)] + (1-\lambda)[f(x) + w_2^\top \beta(x)] : x \in X\} \\
&\geq \lambda \inf_x\{f(x) + w_1^\top \beta(x) : x \in X\} + (1-\lambda)\inf_x\{f(x) + w_2^\top \beta(x) : x \in X\} \\
&= \lambda\theta(w_1) + (1-\lambda)\theta(w_2).
\end{aligned}
$$
😎

The proof uses the fact that the Lagrangian function $\theta(w)$ is the infimum of affine functions in w and therefore concave. An alternative approach to show the concavity of the Lagrangian function is to show that it has subgradients everywhere. This is established in Theorem 3.2.

> **Theorem 3.2. Lagrangian Dual Subgradient**
>
> Let $X \subset \mathbb{R}^n$ be a nonempty compact set, and let $f : \mathbb{R}^n \mapsto \mathbb{R}$ and $\beta : \mathbb{R}^n \mapsto \mathbb{R}^{m+l}$, with $w^\top \beta(x) = \binom{u}{v}^\top \binom{g(x)}{h(x)}$ be continuous. If $\overline{x} \in X(\overline{w}) = \{x \in X : x = \mathrm{argmin}\{f(x) + w^\top \beta(x)\}\}$, then $\beta(\overline{x})$ is a subgradient of $\theta(\overline{w})$.

Proof. Since f and β are continuous and X is compact, $X(\overline{w}) \neq \emptyset$ for any $w \in \mathbb{R}^{m+l}$. Now, let $\overline{w} \in \mathbb{R}^{m+l}$ and $\overline{x} \in X(\overline{w})$. Then:

$$
\begin{aligned}
\theta(w) &= \inf\{f(x) + w^\top \beta(x) : x \in X\} \\
&\leq f(\overline{x}) + w^\top \beta(\overline{x}) \\
&= f(\overline{x}) + (w - \overline{w})^\top \beta(\overline{x}) + \overline{w}^\top \beta(\overline{x}) \\
&= \theta(\overline{w}) + (w - \overline{w})^\top \beta(\overline{x}).
\end{aligned}
$$
😎

Theorem 3.2 can be used to derive a simple optimisation method for Lagrangian functions using subgradient information that is easily available from the term $\beta(w)$.

3.1 The subgradient method

One challenging aspect concerning the solution of Lagrangian dual functions is that they are often not differentiable. This requires an adaptation of the gradient method to consider subgradient information instead.

The challenge with using subgradients (instead of gradients) is that subgradients are not guaranteed to be descent directions (as opposed to gradients being the steepest descent direction under adequate norms). Nevertheless, for suitable step size choices, convergence can be observed. Figure 9 illustrates that subgradients are not necessarily descent directions.
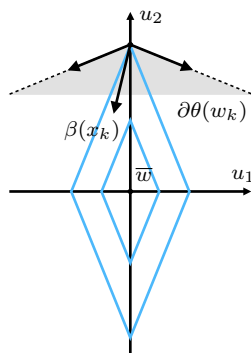

Figure 9: One possible subgradient $\beta(x_k)$ that is a descent direction for suitable step size. Notice that other subgradients without descent direction are available within the subdifferential $\partial\theta(w_k)$.

Algorithm 1 summarises the subgradient method. Notice that the stopping criterion emulates the

optimality condition $0 \in \partial \theta(w_k)$. However, in practice, one also enforces more heuristically driven criteria, such as a maximum number of iterations or a given number without observable improvement on the $\theta(w)$ value.

Algorithm 1 Subgradient method

1: **initialise.** tolerance $\epsilon > 0$, initial point w_0, iteration count $k = 0$.
2: **while** $||\beta(x_k)||_2 > \epsilon$ **do**
3:      $x_k \leftarrow \mathrm{argmin}_x \{\theta(w_k) = \inf_x \{f(x) + w_k^\top \beta(x)\}\}$
4:      $LB_k = \max\{LB_k, \theta(w_k)\}$
5:      update λ_k
6:      $w_{k+1} = w_k + \lambda_k \beta(x_k)$.
7:      $k \leftarrow k + 1$.
8: **end while**
9: **return** $LB_k = \theta(w_k)$.

One critical aspect of the subgradient method is the step size update described in Step 5 of Algorithm 1. Theoretical convergence is guaranteed if Step 5 generates a sequence $\{\lambda_k\}$ such that: $\sum_{k=0}^{\infty} \lambda_k \to \infty$ and $\lim_{k \to \infty} \lambda_k = 0$. However, discrepant performance can be observed for distinct parametrisation of the method.

The classical step update rule employed for the subgradient method is known as the Polyak rule, which gives

$$\lambda_{k+1} = \frac{\alpha_k (LB_k - \theta(w_k))}{||\beta(x_k)||^2}$$

with $\alpha_k \in (0, 2)$ and LB_k being the best-available lower-estimate of $\theta(\overline{w})$. The following result inspires this rule.

> **Proposition 3.3. Improving step size**
>
> f w_k is not optimal, then, for all optimal dual solutions $\overline{w}$, we have
> $$||w_{k+1} - \overline{w}|| < ||w_k - \overline{w}||$$
> for all step sizes λ_k such that
> $$0 < \lambda_k < \frac{2(\theta(\overline{w}) - \theta(w_k))}{||\beta(x_k)||^2}.$$

Proof. We have that $||w_{k+1} - \overline{w}||^2 = ||w_k + \lambda_k \beta(x_k) - \overline{w}||^2 =$
$$||w_k - \overline{w}||^2 - 2\lambda_k (\overline{w} - w_k)^\top \beta(x_k) + (\lambda_k)^2 ||\beta(x_k)||^2.$$
By the subgradient inequality: $\theta(\overline{w}) - \theta(w_k) \le (\overline{w} - w_k)^\top \beta_k$. Thus
$$||w_{k+1} - \overline{w}||^2 \le ||w_k - \overline{w}||^2 - 2\lambda_k (\theta(\overline{w}) - \theta(w_k))^\top \beta(x_k) + (\lambda_k)^2 ||\beta(x_k)||^2.$$
Parametrising the last two terms by $\gamma_k = \frac{\lambda_k ||\beta(x_k)||^2}{\theta(\overline{w}) - \theta(w_k)}$ leads to

$$||w_{k+1} - \overline{w}||^2 \le ||w_k - \overline{w}||^2 - \frac{\gamma_k (2 - \gamma_k)(\theta(\overline{w}) - \theta(w_k))^2}{||\beta(x_k)||^2}.$$

Notice that if $0 < \lambda_k < \frac{2(\theta(\overline{w}) - \theta(w_k))}{||\beta(x_k)||^2}$ then $0 < \gamma_k < 2$ and, thus, $||w_{k+1} - \overline{w}|| < ||w_k - \overline{w}||$. 😎

In practice, since $\theta(\overline{w})$ is not known, it must be replaced by a proxy LB_k, which is chosen to be a lower bound on $\theta(\overline{w})$ to satisfy the subgradient inequality still. The α_k is then reduced from the nominal value 2 to correct for the estimation error of term $\theta(\overline{w}) - \theta(w_k)$.