Seminar

Optimization in Communications and Signal Processing

Ferienakademie 2009, Sarntal

Dual Problem Algorithms

Manuel Stein

1 Introduction

In the previous chapter on *weak and strong duality*, we have seen that given any nonlinear programming problem, referred to as the *primal problem* with equality and inequality constraints

$$
\begin{array}{ll}
\text{Minimize} & f(\boldsymbol{x}) \\
\text{subject to} & \boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0} \\
& \boldsymbol{h}(\boldsymbol{x}) = \mathbf{0} \\
& \boldsymbol{x} \in \mathcal{X},
\end{array}
$$

$$
\text{with } f : \mathbb{R}^n \to \mathbb{R},\ \boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m,\ \boldsymbol{h} : \mathbb{R}^n \to \mathbb{R}^l \text{ and } \mathcal{X} \subseteq \mathbb{R}^n,
$$

many other, closely related, nonlinear programming problems exist. Our analysis has focused on the so called *Lagrangian dual problem*, given by

$$
\begin{array}{ll}
\text{Maximize} & \Theta(\boldsymbol{u}, \boldsymbol{v}) \\
\text{subject to} & \boldsymbol{u} \geq \mathbf{0},
\end{array}
$$

$$
\text{with } \Theta(\boldsymbol{u}, \boldsymbol{v}) = \inf \{ f(\boldsymbol{x}) + \boldsymbol{u}^\mathsf{T} \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{v}^\mathsf{T} \boldsymbol{h}(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X} \}.
$$

This kind of *dualization* accommodates the constraints of the primal problem linearly into the objective function of the dual problem by the use of *Lagrangian multipliers* $\boldsymbol{u}$ and $\boldsymbol{v}$. Note that given a primal nonlinear programming problem, several Lagrangian dual functions can be formulated, depending on which constraints are handled by $\boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}$ and $\boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}$ and which ones are incorporated into the feasible set $\mathcal{X}$. Thus, dependent on the right design choice, dualization is a flexible and powerful mathematical tool.

Duality It has been shown that under suitable convexity assumptions (e.g. Slater's constraint qualifications) *strong duality* holds, meaning that the objective value of the primal and dual problem are equal at optimality $f(\boldsymbol{x}^\star) = \Theta(\boldsymbol{u}^\star, \boldsymbol{v}^\star)$. Without convexity assumptions *weak duality* reasons that the dual objective $\Theta(\boldsymbol{u}, \boldsymbol{v})$ is always a lower bound to the optimal primal objective $f(\boldsymbol{x}^\star)$. The primal-dual interrelation allows to solve the primal problem in a roundabout way by finding a solution to one of its dual counterparts. As the dual function $\Theta(\boldsymbol{u}, \boldsymbol{v})$ is not explicitly available, this work-around comes at the cost of having to solve a minimization problem with respect to $\boldsymbol{x} \in \mathcal{X}$ at each point where $\Theta(\boldsymbol{u}, \boldsymbol{v}) = \inf \{ f(\boldsymbol{x}) + \boldsymbol{u}^\mathsf{T} \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{v}^\mathsf{T} \boldsymbol{h}(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X} \}$ is evaluated. For completeness and as a reminder, we give the two important duality theorems without proofs [1].

Theorem 1.1 (Weak Duality) *Let $\boldsymbol{x}$ be a feasible solution to the primal problem ($\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}$ and $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$) and $(\boldsymbol{u}, \boldsymbol{v})$ be a feasible solution to the dual problem ($\boldsymbol{u} \leq \boldsymbol{0}$), then*

$$f(\boldsymbol{x}) \geq \Theta(\boldsymbol{u}, \boldsymbol{v}).$$

Theorem 1.2 (Strong Duality) *Let $\mathcal{X}$ be a nonempty convex set in $\mathbb{R}^n$, let $f : \mathbb{R}^n \to \mathbb{R}$ and $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$ be convex and $\boldsymbol{h} : \mathbb{R}^n \to \mathbb{R}^l$ affine ($\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}$). If there exists an $\hat{\boldsymbol{x}} \in \mathcal{X}$ such that $\boldsymbol{g}(\hat{\boldsymbol{x}}) < \boldsymbol{0}$ and $\boldsymbol{h}(\hat{\boldsymbol{x}}) = \boldsymbol{0}$ and $\boldsymbol{0} \in \operatorname{int} \boldsymbol{h}(\mathcal{X})$, where $\boldsymbol{h}(\mathcal{X}) = \{\boldsymbol{h}(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\}$, then*

$$\inf \{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}, \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}, \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}\} = \sup \{\Theta(\boldsymbol{u}, \boldsymbol{v}) : \boldsymbol{u} \geq \boldsymbol{0}\}$$

Simplifications In order to simplify the notation on the following pages the equality constraints $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$ are dropped and the general primal problem is rewritten as

$$\begin{array}{ll} \text{Minimize} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0} \\ & \boldsymbol{x} \in \mathcal{X}, \end{array}$$

with $f : \mathbb{R}^n \to \mathbb{R}$ and $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$.

This notation comes without loss of generality as each scalar equality constraint of the form $h(\boldsymbol{x}) = 0$ can be reformulated by using the inequality constraints $h(\boldsymbol{x}) \leq 0$ and $-h(\boldsymbol{x}) \leq 0$. Therefore, equality constraints could easily be incorporated into $\boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}$. Note that in practical implementation such a reformulation can cause numerical problems.

Consequently, the considered dual problem in simplified notation is given by

$$\begin{array}{ll} \text{Maximize} & \Theta(\boldsymbol{u}) = \inf \{f(\boldsymbol{x}) + \boldsymbol{u}^{\mathsf{T}}\boldsymbol{g}(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\} \\ \text{subject to} & \boldsymbol{u} \geq \boldsymbol{0} \end{array}$$

with $f : \mathbb{R}^n \to \mathbb{R}$ and $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$.

Since the primal problem can also be a maximization problem (in the remainder, if not announced differently, only primal minimization problems are considered) the notation is repeated for that kind of form.

$$\begin{array}{ll} \text{Maximize} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0} \\ & \boldsymbol{x} \in \mathcal{X}, \end{array}$$

with $f : \mathbb{R}^n \to \mathbb{R}$ and $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$.

The according dual problem becomes

$$\begin{array}{ll} \text{Minimize} & \Theta(\boldsymbol{u}) = \sup\{f(\boldsymbol{x}) - \boldsymbol{u}^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\} \\ \text{subject to} & \boldsymbol{u} \geq \boldsymbol{0} \end{array}$$

with $f : \mathbb{R}^n \to \mathbb{R}$ and $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$.

Note that the sign of the Lagrangian multipliers has changed from plus to minus and the dual problem has turned from a max-min to a min-max problem.

Notation Finally, for proofs and further discussions, the set of optimal solutions to the Lagrangian dual subproblem $\inf\{f(\boldsymbol{x}) + \boldsymbol{u}^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\}$ is introduced

$$\mathcal{X}(\boldsymbol{u}) = \{\boldsymbol{x} : \boldsymbol{x} \text{ minimizes } f(\boldsymbol{x}) + \boldsymbol{u}^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}) \text{ with } \boldsymbol{x} \in \mathcal{X}\}.$$

Outline In Section 2 we motivate this chapter. Section 3 inspects the dual function and its properties thoroughly. We discuss concavity and differentiability of the dual function. Then we outline the concept of subgradients and the subdifferential and derive ascent directions. Subsequently, Section 4 and 5 study viable strategies to solve the dual problem by means of algorithms. In particular, we introduce and analyze a method based on subgradients (Section 4) as well as the cutting plane algorithm (Section 5), which uses a piecewise linear approximation of the dual function. At the end of each method, we consider the problem of recovering the primal variables. Finally, we apply duality in the context of primal recovery for convex problems and the cutting plane method.

2 Dual Problem - Motivation and Difficulties

Before starting our analysis we want to motivate this section by giving reasons for solving the primal optimization problem indirectly by one of its Lagrangian dual problems. Not each argument might be clear or straightforward at the moment, but should become comprehensible during the discussion on the next pages.

- The dual problem is always concave.
- For the former reason each local maximum is a global maximum.
- The dual problem may have simpler constraints than the primal problem.
- The dual problem may have smaller dimension than the primal problem.
- Weak duality guarantees the dual objective $\Theta(\boldsymbol{u})$ (at any point $\boldsymbol{u}$) to be a lower bound for the optimal primal objective $f(\boldsymbol{x}^\star)$. This can be useful in certain applications, e.g. branch-and-bound.
- With the exact solution $\boldsymbol{u}^\star$ of the dual problem and strong duality, all optimal solutions $\boldsymbol{x}^\star$ can be found by minimizing $L(\boldsymbol{x}, \boldsymbol{u}^\star)$ with respect to $\boldsymbol{x} \in \mathcal{X}$.
- If the dual is solved approximately by a near-optimal multiplier $\tilde{\boldsymbol{u}}$ and $\boldsymbol{x}_{\tilde{u}}$ minimizes $L(\boldsymbol{x}, \tilde{\boldsymbol{u}})$ with respect to $\boldsymbol{x} \in \mathcal{X}$, then under strong duality $\boldsymbol{x}_{\tilde{u}}$ also solves the perturbed primal problem with changed inequality constraints $\boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{g}(\boldsymbol{x}_{\tilde{u}})$. Moderate constraint violations may be acceptable in practical scenarios.

Although these benefits make maximization of the dual problem an attractive undertaking, there are also problems concerning the solution of the dual problem.

- Evaluation of the dual function $\Theta(\boldsymbol{u})$ at any $\boldsymbol{u}$ requires a minimization of $L(\boldsymbol{x}, \boldsymbol{u})$ with respect to the unconstrained set $\boldsymbol{x} \in \mathcal{X}$. This makes the method intractable if the minimization of $L(\boldsymbol{x}, \boldsymbol{u})$ is not possible in closed form or is in general not simple. Note that the initial primal minimization problem is replaced by a problem of maximizing a minimum, a so called max-min problem.
- The dual function is not differentiable for many types of problems.
- There may be additional minimizers of $L(\boldsymbol{x}, \boldsymbol{u}^\star)$ which are not elements of the primal-feasible set $\mathcal{X}$. This complicates the recovery of primal-feasible solutions after having solved the dual problem with an optimal multiplier $\boldsymbol{u}^\star$.

3 Properties of the Dual Function

In order to introduce methods for solving the dual problem we have to study the basics of the Lagrangian dual function. We state some important properties and theorems concerning the dual function which have strong impact on the appropriate choice of problem solving methods presented subsequently.

Concavity of the Dual Function The following theorem states that the Lagrangian dual function is concave, even under moderate restrictions on the functions $f(\boldsymbol{x})$ and $\boldsymbol{g}(\boldsymbol{x})$. This ensures that any local maximum solves the dual problem globally.

Theorem 3.1 (Dual Concavity) *Let $\mathcal{X}$ be a nonempty and compact set in $\mathbb{R}^n$. Let $f : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$ be continuous functions. Then the dual function*

$$\Theta(\boldsymbol{u}) = \inf \{f(\boldsymbol{x}) + \boldsymbol{u}^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\}$$

is concave.

Proof Since the functions $f(\boldsymbol{x})$ and $\boldsymbol{g}(\boldsymbol{x})$ are assumed to be continuous and the set $\mathcal{X}$ is compact, the dual objective $\Theta(\boldsymbol{u})$ takes a finite value at every finite point $\boldsymbol{u}$. For the proof assume two Lagrangian multipliers $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbb{R}^m$ and a scalar $\lambda \in [0, 1]$.

$$\begin{aligned}\Theta(\lambda\boldsymbol{u}_1 + (1-\lambda)\boldsymbol{u}_2) &= \inf \{f(\boldsymbol{x}) + (\lambda\boldsymbol{u}_1 + (1-\lambda)\boldsymbol{u}_2)^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\} \\ &= \inf \{\lambda(f(\boldsymbol{x}) + \boldsymbol{u}_1^\mathsf{T}\boldsymbol{g}(\boldsymbol{x})) + (1-\lambda)(f(\boldsymbol{x}) + \boldsymbol{u}_2^\mathsf{T}\boldsymbol{g}(\boldsymbol{x})) : \boldsymbol{x} \in \mathcal{X}\} \\ &\geq \lambda \inf \{f(\boldsymbol{x}) + \boldsymbol{u}_1^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\} + (1-\lambda) \inf \{f(\boldsymbol{x}) + \boldsymbol{u}_2^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\} \\ &= \lambda\Theta(\boldsymbol{u}_1) + (1-\lambda)\Theta(\boldsymbol{u}_2)\end{aligned}$$

This is exactly the definition of a concave function. ∎

Note that the proof holds for all $\boldsymbol{u} \in \mathbb{R}^m$ without any non-negativity restrictions.

Dual Differentiability Lets assume for now, that the dual function is differentiable and the gradient $\nabla\Theta(\boldsymbol{u})$ is available at any point $\boldsymbol{u}$. A first approach in order to solve the dual function would be to start at a suitable starting point $\boldsymbol{u}_0$. The gradient $\nabla\Theta(\boldsymbol{u}_0)$ is an ascent direction if it is nonzero. Note that if it is zero, we have already found the maximum. Nevertheless, it is reasonable, in a first iteration $k = 1$, to go from $\boldsymbol{u}_0$ into the direction $\boldsymbol{d}_1 = \nabla\Theta(\boldsymbol{u}_0)/\,||\nabla\Theta(\boldsymbol{u}_0)|| \neq \boldsymbol{0}$ in order to improve the dual objective $\Theta(\boldsymbol{u}_0) \to \Theta(\boldsymbol{u}_0 + s_1\boldsymbol{d}_1)$. The right step-size s_k in each iteration k could be determined by a line search on $\Theta(\boldsymbol{u})$ over the direction $\boldsymbol{d}_k$, i.e. maximizing

$\Theta(\boldsymbol{u}_{k-1} + s_k \boldsymbol{d}_k)$ over $s_k \in \mathbb{R}_+$. Observe that the new point $\boldsymbol{u}_{k-1} + s_k \boldsymbol{d}_k$ is not allowed to violate the non-negativity constraints $\boldsymbol{u} \geq \boldsymbol{0}$. We will see how to handle this problem later.

It should be clear now that differentiability is an useful property for dual solving methods and algorithms. The following theorem states under which conditions the dual function is differentiable. A proof is given in the appendix.

Theorem 3.2 (Dual Differentiability) *Let $\mathcal{X} \subset \mathbb{R}^n$ be a nonempty compact set. Let $f : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$ be continuous functions. Let $\bar{\boldsymbol{u}} \in \mathbb{R}^m$ be a certain multiplier and $\mathcal{X}(\bar{\boldsymbol{u}})$ be the singleton $\{\bar{\boldsymbol{x}}\}$. Then $\Theta(\boldsymbol{u})$ is differentiable at $\bar{\boldsymbol{u}}$ with gradient $\nabla\Theta(\bar{\boldsymbol{u}}) = \boldsymbol{g}(\bar{\boldsymbol{x}})$.*

Subdifferential and Subgradients As already mentioned in the motivation section, most optimization problems don't have a differentiable dual function and consequently the gradient $\nabla\Theta(\boldsymbol{u})$ may not exist. In order to deal with these kinds of problems, the concept of subdifferential and subgradients is introduced. These are crucial for finding the solution of the dual function as they allow characterization of ascent directions which we will study later on.

Definition *(Subgradient of a Concave Function)* Let the function $f : \mathbb{R}^n \to \mathbb{R}$ be concave. Then $\boldsymbol{\xi}$ is called a *subgradient* of f at $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ if

$$f(\boldsymbol{x}) \leq f(\bar{\boldsymbol{x}}) + \boldsymbol{\xi}^{\mathsf{T}}(\boldsymbol{x} - \bar{\boldsymbol{x}}) \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^n . \tag{1}$$

In other words: A concave function is always upper bounded by a linear hyperplane with normal vector $\boldsymbol{\xi}$ touching $f(\boldsymbol{x})$ at $\bar{\boldsymbol{x}}$. For completeness, we repeat the definition for convex functions.

Definition *(Subgradient of a Convex Function)* Let the function $f : \mathbb{R}^n \to \mathbb{R}$ be concave. Then $\boldsymbol{\xi}$ is called a *subgradient* of f at $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ if

$$f(\boldsymbol{x}) \geq f(\bar{\boldsymbol{x}}) + \boldsymbol{\xi}^{\mathsf{T}}(\boldsymbol{x} - \bar{\boldsymbol{x}}) \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^n .$$

Definition *(Subdifferential)* The collection of all subgradients of $f(\boldsymbol{x})$ at $\boldsymbol{x}$ is called the subdifferential $\partial f(\boldsymbol{x})$.

Theorem 3.3 (Convexity of the Subdifferential) *The subdifferential $\partial f(\boldsymbol{x})$ of a convex or concave function $f(\boldsymbol{x})$ at $\boldsymbol{x}$ is a convex set.*

Proof We multiply (1) with α and $(1-\alpha)$

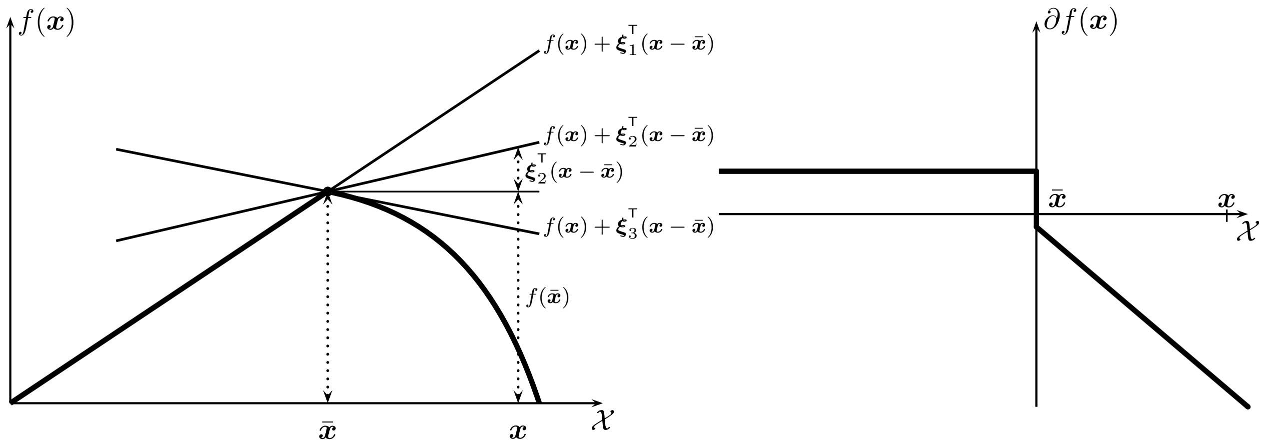
$$\alpha f(\boldsymbol{x}) \leq \alpha f(\bar{\boldsymbol{x}}) + \alpha \boldsymbol{\xi}_1^{\mathsf{T}}(\boldsymbol{x} - \bar{\boldsymbol{x}}) \tag{2}$$

$$(1-\alpha) f(\boldsymbol{x}) \leq (1-\alpha) f(\bar{\boldsymbol{x}}) + (1-\alpha)\boldsymbol{\xi}_2^{\mathsf{T}}(\boldsymbol{x} - \bar{\boldsymbol{x}}). \tag{3}$$

By adding (2) and (3), we obtain

$$f(\boldsymbol{x}) \leq f(\bar{\boldsymbol{x}}) + (\alpha \boldsymbol{\xi}_1^{\mathsf{T}} + (1-\alpha)\boldsymbol{\xi}_2^{\mathsf{T}})(\boldsymbol{x} - \bar{\boldsymbol{x}}).$$

Consequently, with $\alpha \in [0,1]$, $\boldsymbol{\xi}_3 = (\alpha \boldsymbol{\xi}_1 + (1-\alpha)\boldsymbol{\xi}_2)$ is also a subgradient and therefore element of the subdifferential $\partial f(\boldsymbol{x})$. ∎



(a) Subgradients of $f(\boldsymbol{x})$ at $\bar{\boldsymbol{x}}$ (b) Subdifferential $\partial f(\boldsymbol{x})$ of $f(\boldsymbol{x})$

The next theorem states how to obtain a subgradient of the dual function at $\bar{\boldsymbol{u}}$.

Theorem 3.4 (Subgradient of the Dual Function) *Let $\mathcal{X}$ be a nonempty compact set in $\mathbb{R}^n$. Let the functions $f : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$ be continuous, so that for any $\bar{\boldsymbol{u}} \in \mathbb{R}^m$, the set of minimizers for the dual subproblem $\mathcal{X}(\bar{\boldsymbol{u}})$ is not empty. If $\boldsymbol{x}_{\bar{u}} \in \mathcal{X}(\bar{\boldsymbol{u}})$, then $\boldsymbol{g}(\boldsymbol{x}_{\bar{u}})$ is a subgradient of $\Theta(\boldsymbol{u})$ at $\bar{\boldsymbol{u}}$.*

Proof To proof the theorem, assume a certain multiplier $\bar{\boldsymbol{u}}$ and let $\boldsymbol{x}_{\bar{u}} \in \mathcal{X}(\bar{\boldsymbol{u}})$ be one of the minimizers of $L(\boldsymbol{x}, \bar{\boldsymbol{u}})$. Then it holds true, that

$$\begin{aligned}\Theta(\boldsymbol{u}) &\leq f(\boldsymbol{x}_{\bar{u}}) + \boldsymbol{u}^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}_{\bar{u}}) \\ &= f(\boldsymbol{x}_{\bar{u}}) + \boldsymbol{u}^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}_{\bar{u}}) + \bar{\boldsymbol{u}}^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}_{\bar{u}}) - \bar{\boldsymbol{u}}^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}_{\bar{u}}) \\ &= f(\boldsymbol{x}_{\bar{u}}) + \bar{\boldsymbol{u}}^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}_{\bar{u}}) + (\boldsymbol{u} - \bar{\boldsymbol{u}})^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}_{\bar{u}}) \\ &= \Theta(\bar{\boldsymbol{u}}) + (\boldsymbol{u} - \bar{\boldsymbol{u}})^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}_{\bar{u}}).\end{aligned}$$

With the definition for subgradients of a concave function and the result above it follows that $\boldsymbol{g}(\boldsymbol{x}_{\bar{u}})$ is a subgradient of $\Theta(\boldsymbol{u})$ at $\bar{\boldsymbol{u}}$ and the proof is complete. ∎

Note that this is true for all $\boldsymbol{u} \in \mathbb{R}^m$, no matter if $\boldsymbol{u} \geq \boldsymbol{0}$, for which there exists a minimizer $\boldsymbol{x}_u$ to the Lagrangian dual subproblem $\inf\{L(\boldsymbol{x}, \boldsymbol{u}) : \boldsymbol{x} \in \mathcal{X}\}$.

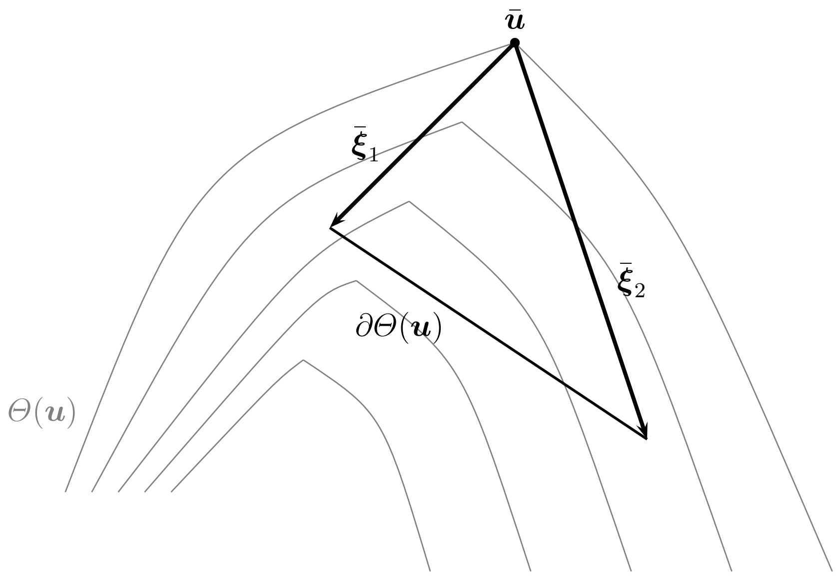


Figure 1: Subgradients and Subdifferential of $\Theta(\boldsymbol{u})$ at $\bar{\boldsymbol{u}}$

Dual Directional Derivatives and Ascent Directions In order to develop methods to improve the dual objective, it is essential to have an idea how *ascent directions* can be derived from the subdifferential $\partial\Theta(\boldsymbol{u})$, the set of all subgradients at $\boldsymbol{u}$.

Definition *(Ascent Direction)* A vector $\boldsymbol{d}$ is called an ascent direction of $\Theta(\boldsymbol{u})$ at $\bar{\boldsymbol{u}}$ if there exists a $\delta > 0$ such that

$$\Theta(\bar{\boldsymbol{u}} + \lambda\boldsymbol{d}) > \Theta(\bar{\boldsymbol{u}}), \quad \forall\lambda \in (0, \delta).$$

As we are interested in improving the dual objective, i.e., going into a direction with positive rate of change, we will need the directional derivative of the dual function. Let us first revisit the general definition of the directional derivative.

Definition *(Directional Derivative)* The directional derivative of a function $f : \mathbb{R}^n \to \mathbb{R}$ at $\bar{\boldsymbol{x}}$ along the vector $\boldsymbol{d}$, denoted by $f'(\bar{\boldsymbol{x}}, \boldsymbol{d})$, is given by the limit

$$f'(\bar{\boldsymbol{x}}, \boldsymbol{d}) = \lim_{\lambda \to 0^+} \frac{f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) - f(\bar{\boldsymbol{x}})}{\lambda}$$

if it exists.

Existence of the Directional Derivative As the dual function need not be differentiable, the existence of the directional derivative seems not trivial. Although it may not be central for the discussion, the next lemma clears the situation and guarantees the general existence of the directional derivative for concave functions. A proof can be found in the appendix.

Lemma 3.5 (Existence of the Directional Derivative) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a concave function. Consider any $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ and a nonzero direction $\boldsymbol{d} \in \mathbb{R}^n$. Then the directional derivative $f'(\bar{\boldsymbol{x}}; \boldsymbol{d})$ of f at $\bar{\boldsymbol{x}}$ exists.*

Dual Directional Derivative In order to derive the directional derivative of the dual function $\Theta(\boldsymbol{u})$ from the subdifferential, the following theorem (for a proof see the appendix) is given:

Theorem 3.6 (Directional Derivative of the Dual Function) *Let $\mathcal{X}$ be a nonempty and compact set in $\mathbb{R}^n$. Let $f : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$ be continuous functions. Let $\bar{\boldsymbol{u}}, \boldsymbol{d} \in \mathbb{R}^m$ and $\partial\Theta(\bar{\boldsymbol{u}})$ be the subdifferential of $\Theta(\boldsymbol{u})$ at $\bar{\boldsymbol{u}}$. Then the directional derivative of $\Theta(\boldsymbol{u})$ at $\bar{\boldsymbol{u}}$ in the direction $\boldsymbol{d}$ satisfies*

$$\Theta'(\bar{\boldsymbol{u}}; \boldsymbol{d}) = \inf\{\boldsymbol{d}^\mathsf{T}\boldsymbol{\xi} : \boldsymbol{\xi} \in \partial\Theta(\bar{\boldsymbol{u}})\}.$$

Using this result, it follows that a vector $\boldsymbol{d}$ is an ascent direction of $\Theta(\boldsymbol{u})$ at $\boldsymbol{u}$, if and only if $\inf\{\boldsymbol{d}^\mathsf{T}\boldsymbol{\xi} : \boldsymbol{\xi} \in \partial\Theta(\boldsymbol{u})\} > 0$, that is, if and only if

$$\boldsymbol{d}^\mathsf{T}\boldsymbol{\xi} \geq \delta \quad \text{for each } \boldsymbol{\xi} \in \partial\Theta(\boldsymbol{u})$$

holds true for some $\delta > 0$.

Corollary 3.7 *This shows that each subgradient* $\boldsymbol{\xi}$ *makes an angle of less than 90 degrees with all ascent directions* $\boldsymbol{d}_{ascent}$.

This result will be very important for the subgradient method that are going to discuss in a moment.

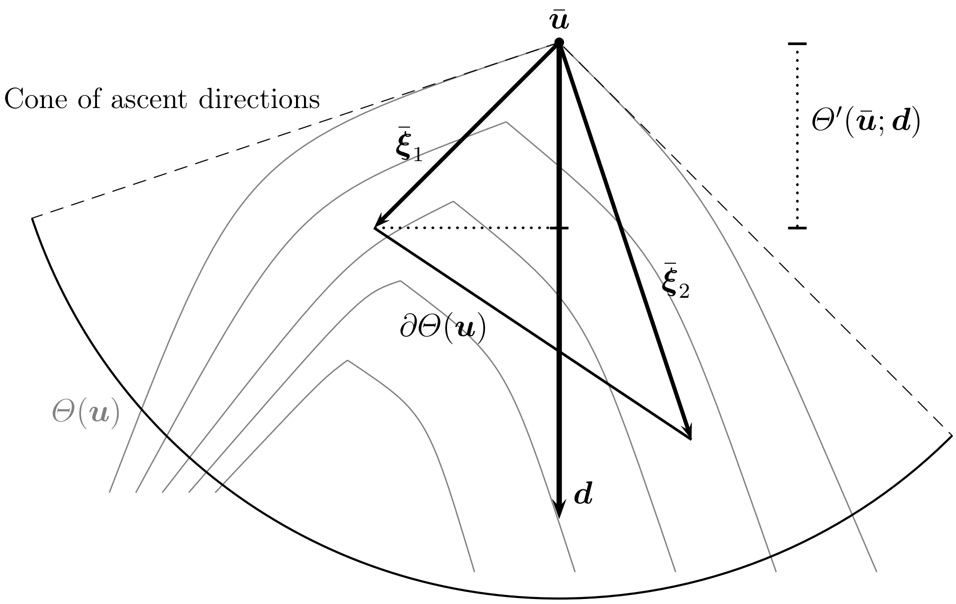


Figure 2: Ascent Directions and Directional Derivative $\Theta'(\bar{\boldsymbol{u}};\boldsymbol{d})$ of $\Theta(\boldsymbol{u})$ at $\bar{\boldsymbol{u}}$

Steepest Ascent Direction In order to solve the dual problem in a simple and fast way, it would be good to know the direction of *steepest ascent* at each point $\boldsymbol{u}$.

Definition A vector $\bar{\boldsymbol{d}}$ is called a direction of steepest ascent of $\Theta(\boldsymbol{u})$ at $\boldsymbol{u}$ if

$$\Theta'(\boldsymbol{u},\bar{\boldsymbol{d}}) = \max_{||\boldsymbol{d}||\leq 1} \Theta'(\boldsymbol{u};\boldsymbol{d}). \tag{4}$$

In order to derive the direction of steepest ascent, we perform the maximization of (4)

$$\begin{aligned}\max_{||\boldsymbol{d}||\leq 1} \Theta'(\boldsymbol{u},\boldsymbol{d}) &= \max_{||\boldsymbol{d}||\leq 1} \inf\{\boldsymbol{d}^{\mathsf{T}}\boldsymbol{\xi} : \boldsymbol{\xi}\in\partial\Theta(\boldsymbol{u})\}\\ &\leq \inf\{\max_{||\boldsymbol{d}||\leq 1} \boldsymbol{d}^{\mathsf{T}}\boldsymbol{\xi} : \boldsymbol{\xi}\in\partial\Theta(\boldsymbol{u})\}\\ &= \inf\{||\boldsymbol{\xi}|| : \boldsymbol{\xi}\in\partial\Theta(\boldsymbol{u})\}\\ &= ||\bar{\boldsymbol{\xi}}||\end{aligned}$$

where $\bar{\boldsymbol{\xi}}$ is the subgradient out of the subdifferential $\partial\Theta(\boldsymbol{u})$ with the smallest Euclidean norm. This result already contains the proof of the next theorem.

Theorem 3.8 (Steepest Ascent of the Dual Function) *Let $\mathcal{X}$ be a nonempty and compact set in $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$ continuous. The direction of steepest ascent $\bar{\boldsymbol{d}}$ of $\Theta(\boldsymbol{u})$ at $\bar{\boldsymbol{u}}$ is given as*

$$\bar{\boldsymbol{d}} = \begin{cases} \mathbf{0} & \text{if } \bar{\boldsymbol{\xi}} = \mathbf{0} \\ \frac{\bar{\boldsymbol{\xi}}}{||\bar{\boldsymbol{\xi}}||} & \text{if } \bar{\boldsymbol{\xi}} \neq \mathbf{0}. \end{cases}$$

The fact that the whole set of subgradients is needed in order to determine the direction with highest positive rate of change, makes it in general intractable to formulate algorithms that replace the missing gradient by the direction of steepest ascent.

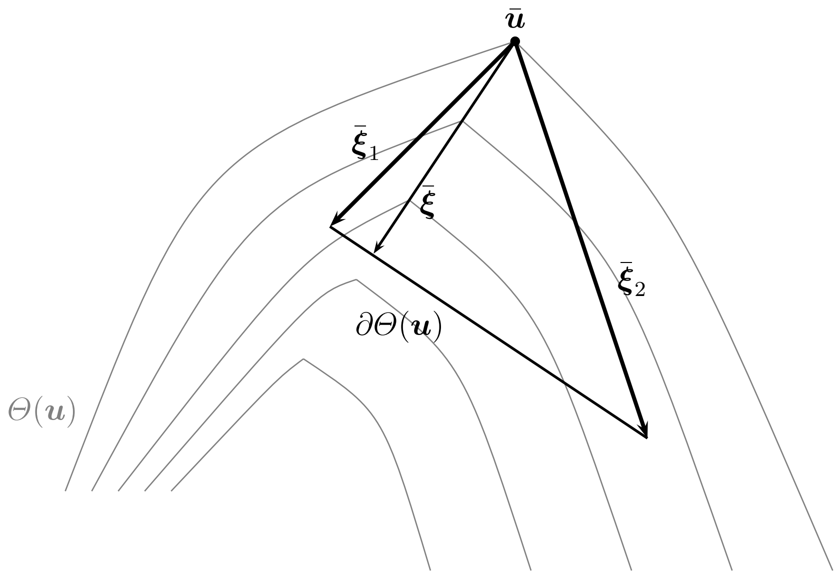


Figure 3: Direction of Steepest Ascent $\bar{\boldsymbol{\xi}}$ of $\Theta(\boldsymbol{u})$ at $\bar{\boldsymbol{u}}$

4 Subgradient Methods

As for some interesting cases the dual problem is not differentiable, we are going to discuss a method to solve the dual problem based on subgradients. It is important to observe that by minimizing $L(\boldsymbol{x}, \bar{\boldsymbol{u}})$ for a certain Lagrangian multiplier $\bar{\boldsymbol{u}}$ with $\boldsymbol{x}_{\bar{u}} \in \mathcal{X}(\bar{\boldsymbol{u}})$ yields a subgradient $\boldsymbol{g}(\boldsymbol{x}_{\bar{u}})$ (Theorem 3.4) without additional effort. As the subgradient $\boldsymbol{g}(\boldsymbol{x}_{\bar{u}})$ makes an angle less than 90 degrees with all ascent direction it need not be an ascent direction, but it holds that a small move from $\bar{\boldsymbol{u}}$ along any subgradient at $\bar{\boldsymbol{u}}$ will decrease the distance to the set of maximizers $\mathcal{U}^\star$ of the dual function $\Theta(\boldsymbol{u})$.

General Method The simplest type of subgradient method is given by

$$\boldsymbol{u}_{k+1} = \boldsymbol{\Pi}_+(\boldsymbol{u}_k + s_k \boldsymbol{g}(\boldsymbol{x}_{u_k}))$$

where the stepsize s_k is a positive scalar, $\boldsymbol{x}_{u_k}$ is the minimizer of $L(\boldsymbol{x}, \boldsymbol{u}_k)$ and $\boldsymbol{\Pi}_+(\boldsymbol{u})$ is the projection onto the set of feasible multipliers $\boldsymbol{u} \geq \boldsymbol{0}$.

Definition The projection $\boldsymbol{\Pi}_{\mathcal{S}}(\boldsymbol{u})$ from $\boldsymbol{u} \in \mathbb{R}^n$ onto a general set $\mathcal{S} \subseteq \mathbb{R}^n$ is defined as

$$\boldsymbol{\Pi}_{\mathcal{S}}(\boldsymbol{u}) = \{\boldsymbol{y} : \boldsymbol{y} \text{ minimizes } ||\boldsymbol{y} - \boldsymbol{u}|| : \boldsymbol{y} \in \mathcal{S}\}.$$

The projection $\boldsymbol{\Pi}_+(\boldsymbol{u})$ onto the set of vectors with non-negative entries $\boldsymbol{u} \geq \boldsymbol{0}$ is performed easily by modifying the entries $[\boldsymbol{u}]_i$ of $\boldsymbol{u}$ in the following way

$$[\boldsymbol{u}]_i = \begin{cases} [\boldsymbol{u}]_i & \text{if } [\boldsymbol{u}]_i \geq 0 \\ 0 & \text{if } [\boldsymbol{u}]_i < 0. \end{cases}$$

Note that the new multiplier $\boldsymbol{u}_{k+1}$ may not improve the dual objective for all values of the stepsize s, i.e., at some iteration k, it may be the case that

$$\Theta(\boldsymbol{u}_{k+1}) < \Theta(\boldsymbol{u}_k) \qquad \forall s_k > 0.$$

Nevertheless, the distance to an optimal solution will be reduced if the stepsize s_k is sufficiently small.

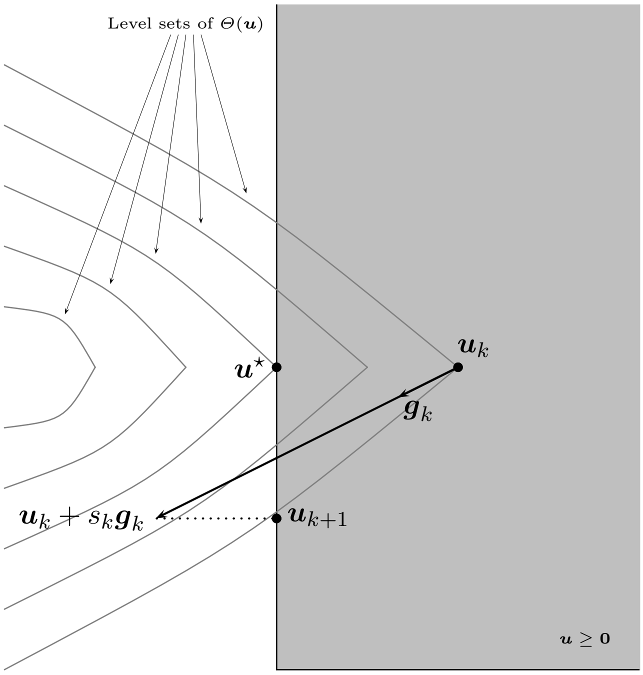


Figure 4: General Subgradient Method

Adequate Stepsize Rule We derive a rule for the stepsize selection, that guarantees the reduction of the distance to an optimal solution in each iteration. Our goal can be formulated as

$$||\boldsymbol{u}_{k+1} - \boldsymbol{u}^\star|| < ||\boldsymbol{u}_k - \boldsymbol{u}^\star|| \,. \tag{5}$$

Using the following equation that holds with equality

$$||\boldsymbol{u}_k + s_k\boldsymbol{g}_k - \boldsymbol{u}^\star||^2 = ||\boldsymbol{u}_k - \boldsymbol{u}^\star||^2 - 2s_k(\boldsymbol{u}^\star - \boldsymbol{u}_k)^\mathsf{T}\boldsymbol{g}_k + (s_k)^2\,||\boldsymbol{g}_k||^2$$

and the subgradient inequality

$$(\boldsymbol{u}^\star - \boldsymbol{u}_k)^\mathsf{T}\boldsymbol{g}_k \geq \Theta(\boldsymbol{u}^\star) - \Theta(\boldsymbol{u}_k),$$

we obtain

$$||\boldsymbol{u}_k + s_k\boldsymbol{g}_k - \boldsymbol{u}^\star||^2 \leq ||\boldsymbol{u}_k - \boldsymbol{u}^\star||^2 - 2s_k(\Theta(\boldsymbol{u}^\star) - \Theta(\boldsymbol{u}_k)) + (s_k)^2\,||\boldsymbol{g}_k||^2 \,. \tag{6}$$

In order to satisfy (5), the sum of the last two terms in (6) has to be negative, i.e.,

$$(s_k)^2 \left|\left|\boldsymbol{g}_k\right|\right|^2 < 2s_k(\Theta(\boldsymbol{u}^\star) - \Theta(\boldsymbol{u}_k)).$$

Therefore, the range of appropriate stepsizes with respect to our goal of reducing distance to the optimal solution is

$$0 < s_k < \frac{2(\Theta(\boldsymbol{u}^\star) - \Theta(\boldsymbol{u}_k))}{\left|\left|\boldsymbol{g}_k\right|\right|^2}$$

and we can state the following theorem without further proof:

Theorem 4.1 (Range of Adequate Stepsize) *If $\boldsymbol{u}_k$ is not optimal, then for every dual optimal solution $\boldsymbol{u}^\star$ we have*

$$||\boldsymbol{u}_{k+1} - \boldsymbol{u}^\star|| < ||\boldsymbol{u}_k - \boldsymbol{u}^\star||$$

for all stepsizes s_k in the range of

$$0 < s_k < \frac{2(\Theta(\boldsymbol{u}^\star) - \Theta(\boldsymbol{u}_k))}{\left|\left|\boldsymbol{g}_k\right|\right|^2}. \tag{7}$$

One suggestion that can be derived from the above theorem is to use a stepsize according to

$$s_k = \frac{(\Theta(\boldsymbol{u}^\star) - \Theta(\boldsymbol{u}_k))}{\left|\left|\boldsymbol{g}_k\right|\right|^2}.$$

This rule selects the stepsize to be in the middle of the range (7). It is obvious that the stepsize has to be chosen depending on the difference to the optimal dual objective and the norm of the gradient in the current iteration. Unfortunately, the optimal dual objective is generally unknown and, therefore, we have to use an appropriate estimate of $\Theta(\boldsymbol{u}^\star)$ or select the stepsize by using a simpler scheme.

Constant Stepsize Rule The first practical stepsize rule which is discussed is a constant stepsize rule, where we select s_k to be the same for all iterations, i.e., $s_k = s$, $\forall k$, with some $s > 0$. In order to remove $||\boldsymbol{g}_k||$ in (7) we assume that the norm $||\boldsymbol{g}(\boldsymbol{x})||$ is bounded over the set $\boldsymbol{x} \in \mathcal{X}$ by some constant C, i.e., $||\boldsymbol{g}_k|| \leq C$, $\forall k$. This is reasonable under the assumption that

$\boldsymbol{g}(\boldsymbol{x})$ is a continuous function and the set $\mathcal{X}$ is compact. Then in the sense of (7) it holds, that the distance to the optimal dual solution $\boldsymbol{u}^\star$ decreases, if

$$0 < s_k < \frac{2(\Theta(\boldsymbol{u}^\star) - \Theta(\boldsymbol{u}_k))}{C^2}, \qquad \forall k, \tag{8}$$

since

$$0 \leq \frac{2(\Theta(\boldsymbol{u}^\star) - \Theta(\boldsymbol{u}_k))}{C^2} \leq \frac{2(\Theta(\boldsymbol{u}^\star) - \Theta(\boldsymbol{u}_k))}{||\boldsymbol{g}_k||^2}, \qquad \forall k.$$

With the reformulation of (8), it is equivalently to say that the distance to $\boldsymbol{u}^\star$ decreases if $\boldsymbol{u}_k$ belongs to the level set

$$\{\boldsymbol{u} : \Theta(\boldsymbol{u}) < \Theta(\boldsymbol{u}^\star) - \frac{sC^2}{2}\}.$$

As a result, the coefficient $\frac{sC^2}{2}$ determines how close we can get to the optimal dual solution with the suggested constant stepsize rule. A solution arbitrarily close to the optimum is obtained by choosing s to be sufficiently small with $s > 0$. To complete the section, the convergence theorem for the constant stepsize method is given:

Theorem 4.2 (Constant Stepsize Rule - Convergence) *Let $\{\boldsymbol{u}_k\}$ be the sequence generated by the subgradient method with constant stepsize $s > 0$.*

If $\Theta(\boldsymbol{u}^\star) = \infty$, then

$$\lim_{k\to\infty} \sup \Theta(\boldsymbol{u}_k) = \Theta(\boldsymbol{u}^\star).$$

If $\Theta(\boldsymbol{u}^\star) < \infty$, then

$$\lim_{k\to\infty} \sup \Theta(\boldsymbol{u}_k) \geq \Theta(\boldsymbol{u}^\star) - \frac{sC^2}{2}.$$

Proof see [2, p. 473].

Diminishing Stepsize Rule Attaining an accurate solution with a constant stepsize, requires to choose s very small. Unfortunately, a small stepsize will result in a slow progress of the method and so it is common to start with moderate stepsize values s_k, which are successively reduced to a small $s > 0$. One possibility is to apply the *diminishing stepsize rule*, where the stepsizes s_k satisfy

$$\lim_{k\to\infty} s_k = 0, \qquad \sum_{k=0}^{\infty} s_k = \infty. \tag{9}$$

This choice allows a fast initial progress, the ability to travel infinitely far for a solution and attain convergence by a small stepsize s_k after some iterations. An example of such a stepsize that satisfies (9) is the *modified harmonic stepsize rule*

$$s_k = \frac{\beta}{k+\gamma} \qquad \text{with } 0 < \beta, \gamma < \infty. \tag{10}$$

The next theorem guarantees convergence of the subgradient method with diminishing stepsize:

Theorem 4.3 (Diminishing Stepsize Rule - Convergence) *Assume that the set of optimal dual solutions $\mathcal{U}^\star$ is not empty and bounded. The stepsize s_k is such that*

$$\lim_{k\to\infty} s_k = 0, \qquad \sum_{k=0}^{\infty} s_k = \infty.$$

Then for the sequence $\{\boldsymbol{u}_k\}$, generated by the subgradient method with a diminishing stepsize rule, we have

$$\lim_{k\to\infty} d(\boldsymbol{u}_k, \mathcal{U}^\star) = 0, \qquad \lim_{k\to\infty} \Theta(\boldsymbol{u}_k) = \Theta(\boldsymbol{u}^\star).$$

Proof see [2, p. 479].

Primary Recovery A possibility to recover the variables of the primal problem while solving the dual problem with the modified harmonic stepsize rule (10), is to average over all intermediate solutions $\boldsymbol{x}_k$. Therefore, the averaged subproblem solution in each iteration k is obtained by

$$\hat{\boldsymbol{x}}_k = \frac{1}{k}\sum_{i=0}^{k-1} \boldsymbol{x}_i.$$

Theorem 4.4 (Convergence of $\hat{\boldsymbol{x}}_k$ to the Primal Solution Set) *Suppose that the set* $\{\boldsymbol{x} \in \mathcal{X} \,|\boldsymbol{g}(\boldsymbol{x}) < \boldsymbol{0}\}$ *is nonempty, let the subgradient method with modified harmonic stepsize be applied to the dual problem and let* $\mathcal{X}^\star$ *be the solution set of the primal problem, then*

$$\lim_{k\to\infty} d(\hat{\boldsymbol{x}}_k, \mathcal{X}^\star) = 0.$$

Proof see [3, Proof Theorem 2].

5 Cutting-Plane Method

The Master Problem A second method for solving the dual problem is called cutting-plane algorithm. In each iteration k, a piecewise linear approximation of the dual function is maximized. The approximation simplifies the dual problem to a linear program with a finite number of constraints (solvable with the simplex method) referred to as the *master program*

$$\begin{array}{lll} \text{Maximize} & z & \\ \text{subject to} & z \leq f(\boldsymbol{x}_i) + \boldsymbol{u}^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}_i) & i = 0, \ldots, k-1 \\ & \boldsymbol{u} \geq \boldsymbol{0}. & \end{array}$$

where the points $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{k-1}$ are points in $\mathcal{X}$. Note that without restriction to a finite number of points, the problem would be an optimization problem with infinite number of constraints and intractable.

The Subproblem Let $(z_k, \boldsymbol{u}_k)$ be the optimal solution to the master problem. In order to check if the algorithm can be terminated, consider the following problem, referred to as the *subproblem*,

$$\begin{array}{ll} \text{Minimize} & f(\boldsymbol{x}) + \boldsymbol{u}_k^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in \mathcal{X}. \end{array}$$

Let $\boldsymbol{x}_k \in \mathcal{X}(\boldsymbol{u}_k)$ be the appropriate optimal minimizer of the subproblem, so that

$$\Theta(\boldsymbol{u}_k) = f(\boldsymbol{x}_k) + \boldsymbol{u}_k^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}_k).$$

If $z_k = \Theta(\boldsymbol{u}_k)$, then the maximizer $\boldsymbol{u}_k$ is an optimal solution to the Lagrangian dual problem, and the method terminates. Otherwise, for $\boldsymbol{u} = \boldsymbol{u}_k$ there exists a $\boldsymbol{x} = \boldsymbol{x}_k$ that violates the definition of the dual function $\Theta(\boldsymbol{u})$ being the infimum of $L(\boldsymbol{x}, \boldsymbol{u})$. Thus, the dual problem was oversimplified in the master program. To strengthen the representation, the constraint

$$z \leq f(\boldsymbol{x}_k) + \boldsymbol{u}^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}_k) \tag{11}$$

is added to the master-program by considering the point $\boldsymbol{x}_k \in \mathcal{X}(\boldsymbol{u}_k)$. Subsequently the program is solved again with the extended set of constraints. Obviously, the new constraint (11) does not allow the current point $(z_k, \boldsymbol{u}_k)$ to be a solution anymore. Hence, this point is now cutted away in the following iterations.

Graphical Interpretation The cutting plane method can be interpreted as a tangential approximation technique. By definition we have

$$\Theta(\boldsymbol{u}) \le f(\boldsymbol{x}) + \boldsymbol{u}^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}) \qquad \text{for } \boldsymbol{x} \in \mathcal{X}$$

and the hyperplane

$$\{(\boldsymbol{u}, z) : \boldsymbol{u} \in \mathbb{R}^m, z = f(\boldsymbol{x}) + \boldsymbol{u}^\mathsf{T}\boldsymbol{g}(\boldsymbol{x})\},$$

for any fixed $\boldsymbol{x} \in \mathcal{X}$, is an upper bound for the function $\Theta(\boldsymbol{u})$. Therefore, the master program at iteration k can be interpreted as the problem

$$\begin{array}{ll} \text{Maximize} & \hat{\Theta}_k(\boldsymbol{u}) \\ \text{subject to} & \boldsymbol{u} \ge \boldsymbol{0}, \end{array}$$

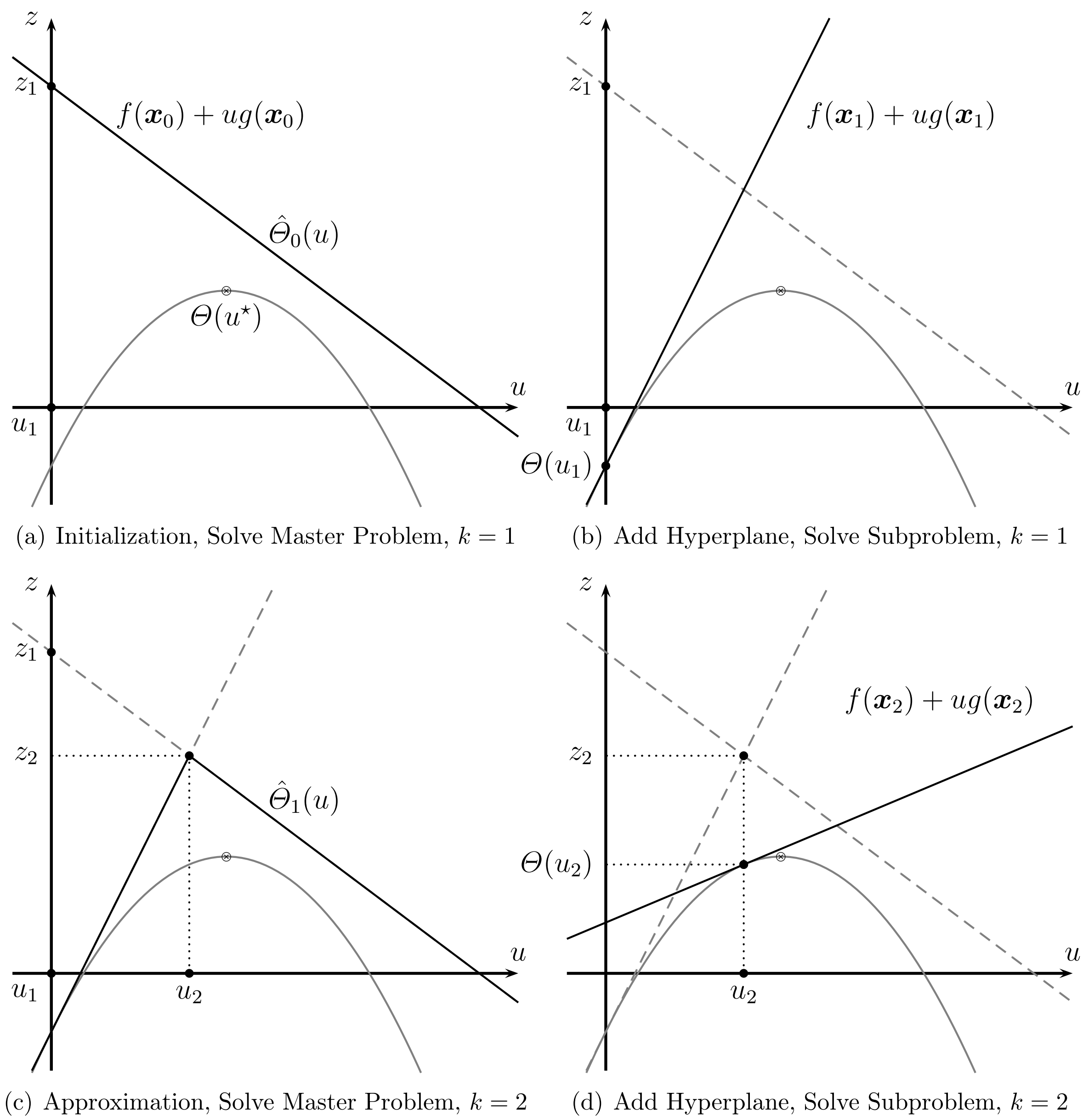
$$\text{with } \hat{\Theta}_k(\boldsymbol{u}) = \min\{f(\boldsymbol{x}_i) + \boldsymbol{u}^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}_i) : i = 0, \ldots, k-1\}.$$

The function $\hat{\Theta}_k(\boldsymbol{u})$ is a piecewise linear function that provides an outer approximation for $\Theta(\boldsymbol{u})$ by considering k bounding hyperplanes. Therefore, in any iteration the function $\hat{\Theta}_k(\boldsymbol{u})$ upper bounds $\Theta(\boldsymbol{u})$ at all points, i.e. $\Theta(\boldsymbol{u}) \le \hat{\Theta}_k(\boldsymbol{u})$, $\forall \boldsymbol{u}$, $\forall k$. Each new constraint

$$z \le f(\boldsymbol{x}_k) + \boldsymbol{u}^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}_k)$$

that is added to the master problem results in a better approximation of $\Theta(\boldsymbol{u})$. The next theorem states that the cutting plane method converges. A proof is found in the appendix.

Theorem 5.1 (Convergence Cutting Plane Method) *If the sequence $\{\boldsymbol{g}_k\}$ is a bounded sequence, then every limit point of the sequence $\{\boldsymbol{u}_k\}$ generated by the cutting plane method is a dual optimal solution.*



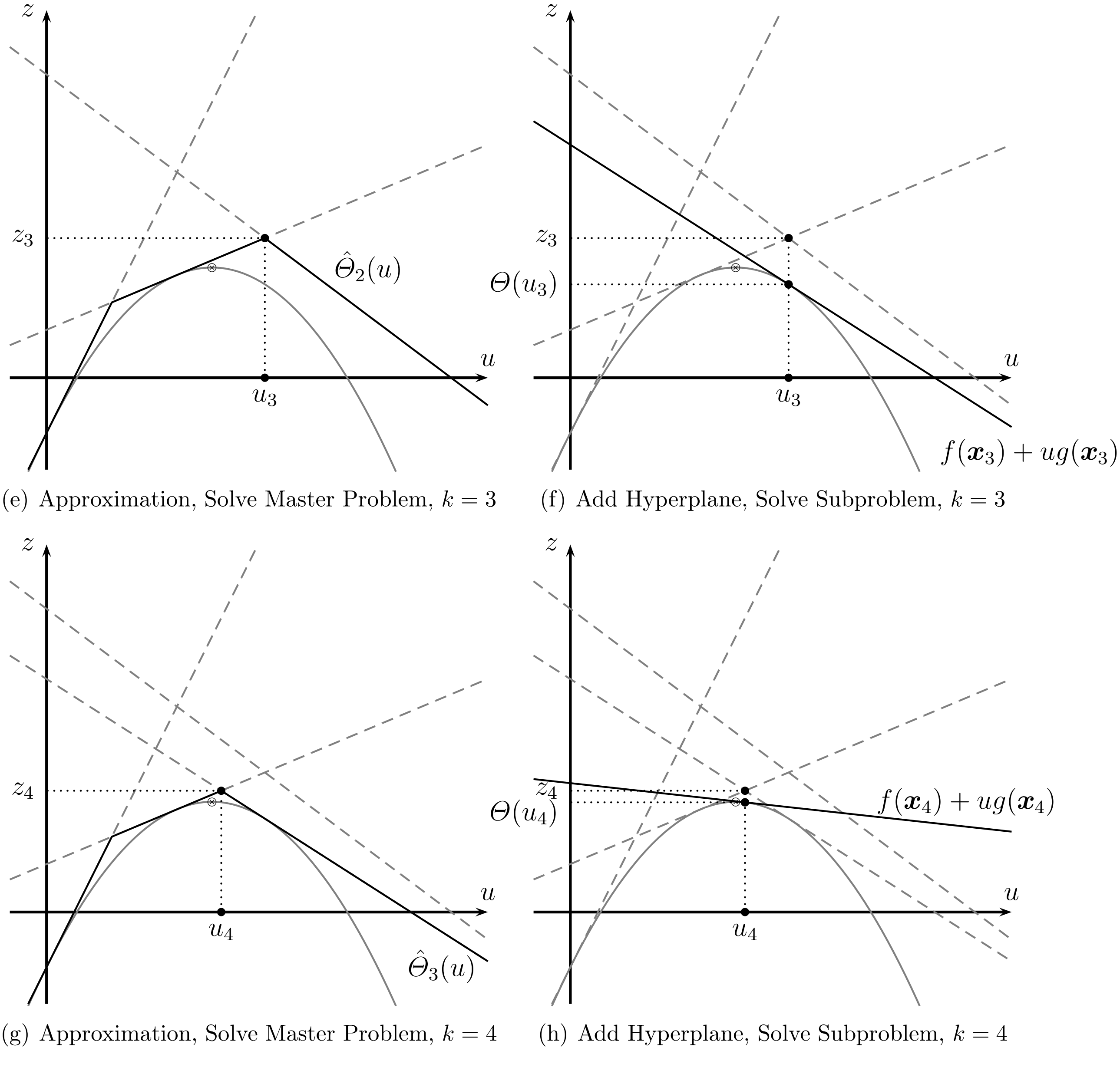
(a) Initialization, Solve Master Problem, $k = 1$

(b) Add Hyperplane, Solve Subproblem, $k = 1$

(c) Approximation, Solve Master Problem, $k = 2$

(d) Add Hyperplane, Solve Subproblem, $k = 2$

Figure 5: Cutting Plane - Graphical Example



(e) Approximation, Solve Master Problem, $k = 3$

(f) Add Hyperplane, Solve Subproblem, $k = 3$

(g) Approximation, Solve Master Problem, $k = 4$

(h) Add Hyperplane, Solve Subproblem, $k = 4$

Figure 5: Cutting Plane - Graphical Example

Stopping Criteria We observe that while running the cutting plane method we produce different sequences of primal and dual variables. These can be used to define a stopping criteria and to recover the primal solution. By repeatedly solving the master problem we get the sequence $\{z_k\}$, where z_k is the maximum value of the outer linear approximation to the dual function in the k-th iteration. This sequence is a monotone decreasing upper bound of the optimal dual objective $\Theta(\boldsymbol{u}^\star)$ and we can state

$$\Theta(\boldsymbol{u}^\star) \leq z_k, \qquad \forall k. \tag{12}$$

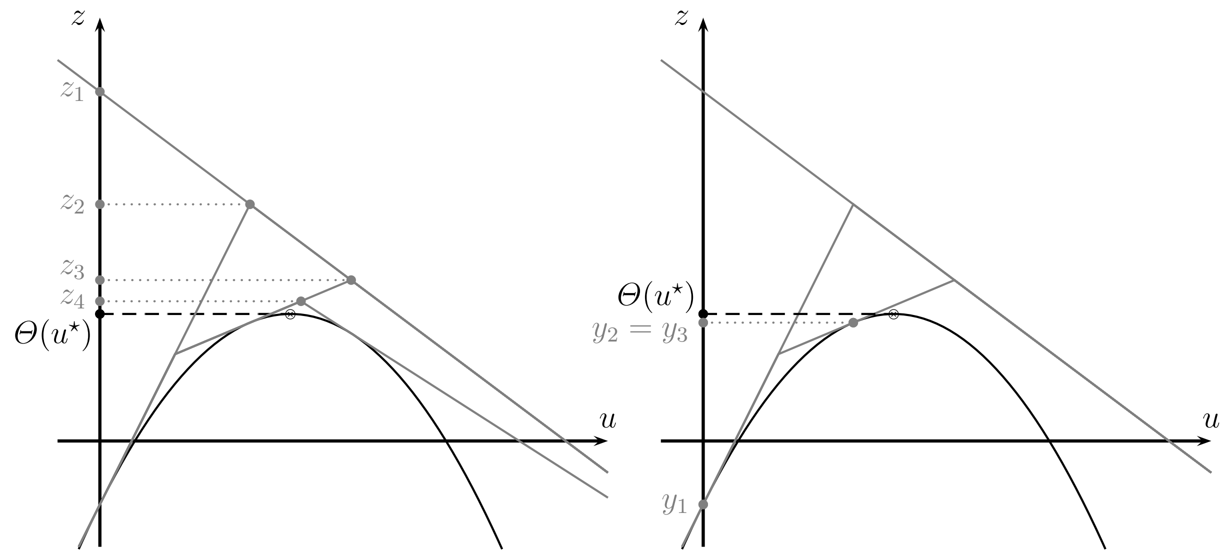
With the solutions to all the subproblems solved after the k-th iteration, the evaluations of the dual function for all $\boldsymbol{u}_i$ with $i = 1 \dots k$, we obtain the sequence $\{y_k\}$ defined by

$$y_k = \max_{1 \leq i \leq k} \Theta(\boldsymbol{u}_i).$$

This sequence increases monotonously and gives a lower bound for the optimal dual objective $\Theta(\boldsymbol{u}^\star)$. Consequently, with (12) we obtain,

$$y_k \leq \Theta(\boldsymbol{u}^\star) \leq z_k \qquad \forall k.$$

With the two sequences $\{z_k\}$ and $\{y_k\}$, which are upper and lower bounds for the optimal dual



(a) The Sequence z_k, $k = 4$ (b) The Sequence y_k, $k = 3$

Figure 6: Sequences produced by the Cutting Plane Method

objective, we can define a stopping criteria with respect to ϵ_k defined as

$$\epsilon_k = |\Theta(\boldsymbol{u}^\star) - \Theta(\boldsymbol{u}_k)| \leq |z_k - y_k|, \qquad \forall k.$$

Primary Recovery in the Convex Case We have discussed a method to solve the dual problem with a simple approximation which is successively refined. The results provide us a well defined and tunable stopping criteria. Until now we have just considered solutions to the dual problem, the optimal dual variables $\boldsymbol{u}^\star$. But this is not yet the solution to the initial primal problem. The last pages of this chapter will deal with the problem of obtaining the primal variables $\boldsymbol{x}^\star$ out of the cutting plane method, for convex problems with *strong duality.* For these problems the bounds on the optimal dual objective apply also to the optimal primal objective

$$y_k \leq \Theta(\boldsymbol{u}^\star) = f(\boldsymbol{x}^\star) \leq z_k, \qquad \forall k.$$

First consider the sequence of primal variables that occurred during the algorithm. We have approximated the dual function by a piecewise linear function

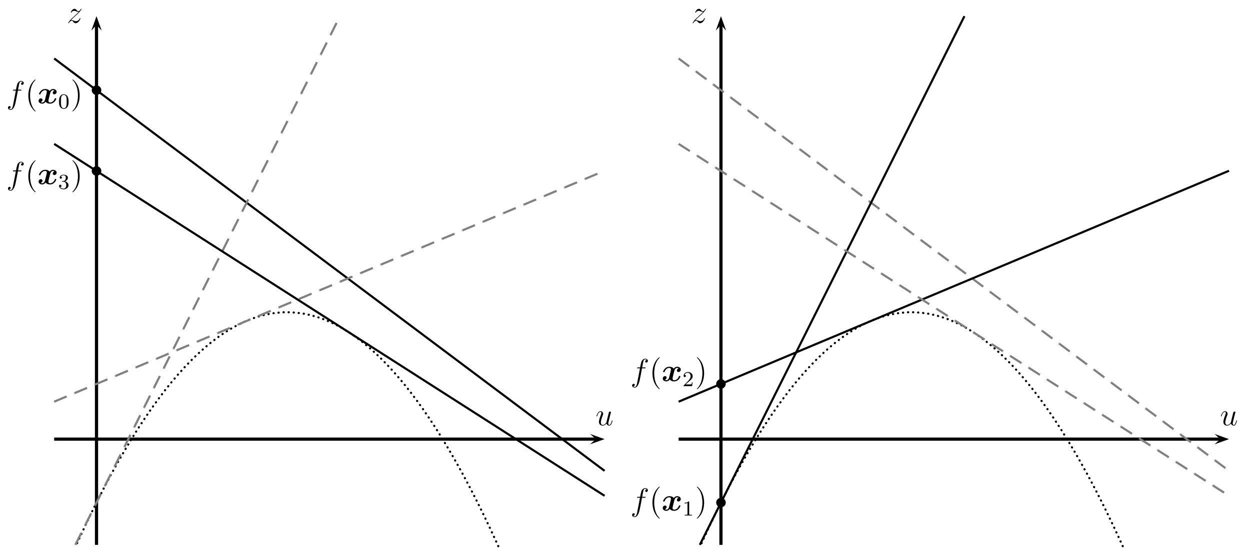
$$\hat{\Theta}_k(\boldsymbol{u}) = \min\{f(\boldsymbol{x}_i) + \boldsymbol{u}^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}_i) : i = 0, \ldots, k-1\}.$$

Each individual linear function $f(\boldsymbol{x}_i) + \boldsymbol{u}^\mathsf{T}\boldsymbol{g}(\boldsymbol{x}_i)$ is generated by a known primal variable $\boldsymbol{x}_i$. Some of these are primal feasible $\boldsymbol{g}(\boldsymbol{x}_i) \leq \boldsymbol{0}$ but far from optimal, others are primal infeasible $\boldsymbol{g}(\boldsymbol{x}_i) > \boldsymbol{0}$ but have a much better objective then the optimal primal objective. Since the primal set $\mathcal{X}$ is convex it seems reasonable to find the best convex combination of all obtained primal variables $\sum_{i=0}^k \lambda_i \boldsymbol{x}_i$ with $\sum_{i=0}^k \lambda_i = 1$, $\lambda_i \geq 0$, $\forall i$ and minimize the primal object $f(\sum_{i=0}^k \lambda_i \boldsymbol{x}_i)$ with respect to $\boldsymbol{\lambda}$ under the constraint $\boldsymbol{g}(\sum_{i=0}^k \lambda_i \boldsymbol{x}_i) \leq \boldsymbol{0}$. Unfortunately, this is very similar to the original primal problem and thus yields no simplification. However, since the primal objective function is convex, we know that

$$f\Big(\sum_{i=0}^k \lambda_i \boldsymbol{x}_i\Big) \leq \sum_{i=0}^k \lambda_i f(\boldsymbol{x}_i).$$

The minimization of $\sum_{i=0}^k \lambda_i f(\boldsymbol{x}_i)$ will provide a minimizer $\hat{\boldsymbol{x}}^\star = \sum_{i=0}^k \lambda_i^\star \boldsymbol{x}_i$ for the upper bound. The same argument holds for the convex constraint function $\boldsymbol{g}(\boldsymbol{x})$ for which

$$\boldsymbol{g}\Big(\sum_{i=0}^k \lambda_i \boldsymbol{x}_i\Big) \leq \sum_{i=0}^k \lambda_i \boldsymbol{g}(\boldsymbol{x}_i).$$



(a) Primal Feasible Variables, $\boldsymbol{g}(\boldsymbol{x}_i) \leq \mathbf{0}$ (b) Primal Infeasible Variables, $\boldsymbol{g}(\boldsymbol{x}_i) > \mathbf{0}$

Figure 7: Primal Variables and Objectives produced by the Cutting Plane Method, $k = 4$

Satisfying the constraint $\sum_{i=0}^{k} \lambda_i \boldsymbol{g}(\boldsymbol{x}_i) \leq \mathbf{0}$ will guarantee $\boldsymbol{g}(\sum_{i=0}^{k} \lambda_i \boldsymbol{x}_i) \leq \mathbf{0}$. Minimizing the low complexity linear program $\sum_{i=0}^{k} \lambda_i f(\boldsymbol{x}_i)$ constrained by $\sum_{i=0}^{k} \lambda_i \boldsymbol{g}(\boldsymbol{x}_i) \leq \mathbf{0}$ will give an upper bounded primal feasible solution to a simplified version of the primal recovery problem. In summary the *simplified primary recovery problem* can be defined by

$$\begin{array}{ll} \text{Minimize} & \sum_{i=0}^{k} \lambda_i f(\boldsymbol{x}_i) \\ \text{subject to} & \sum_{i=0}^{k} \lambda_i \boldsymbol{g}(\boldsymbol{x}_i) \leq \mathbf{0} \\ & \sum_{i=0}^{k} \lambda_i = 1 \\ & \lambda_i \geq 0 \ \forall i. \end{array}$$

Theorem 5.2 (Primal Feasible Solutions for the Convex Case) *Let $\mathcal{X}$ be a nonempty convex set in $\mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$ and $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$ convex functions and $\boldsymbol{x}_0$ an initial feasible solution of the primal problem. Suppose that the points $\boldsymbol{x}_{u,i} \in \mathcal{X}(\boldsymbol{u}_i)$ for $i = 1, \ldots, k$ are generated by any algorithm for solving the dual problem. Let $\lambda_i^\star$ for $i = 1, \ldots, k$ be an optimal solution to the simplified primal recovery problem. Then $\hat{\boldsymbol{x}}_k^\star = \sum_{i=0}^{k} \lambda_i^\star \boldsymbol{x}_i$ with $\sum_{i=0}^{k} \lambda_i^\star = 1$ is a feasible solution to the primal problem.*

Proof Since $\mathcal{X}$ is convex and $\boldsymbol{x}_i \in \mathcal{X}$ for each i, it follows that $\boldsymbol{x}_k^\star \in \mathcal{X}$. Since $\boldsymbol{g}$ is convex it can be deduced that $\boldsymbol{g}(\boldsymbol{x}_k^\star) \leq \mathbf{0}$. Thus, $\boldsymbol{x}_k^\star$ is a feasible solution to the primal problem.

Note that this theorem only guarantees primal feasibility of the solution, irrespective of any considerations regarding accuracy or convergence.

Dual of the Master Problem As the solution to the simplified primary recovery problem misses a link to our upper and lower bound of the optimal dual and primal objective, i.e. the sequences $\{z_k\}$ and $\{y_k\}$, we consider again the master problem of the cutting plane method

$$\begin{array}{ll} \text{Maximize} & z \\ \text{subject to} & z - f(\boldsymbol{x}_i) - \boldsymbol{u}^\mathsf{T} \boldsymbol{g}(\boldsymbol{x}_i) \leq 0, \quad i = 0, \dots, k \\ & \boldsymbol{u} \geq \boldsymbol{0}. \end{array}$$

The associated Lagrangian dual function is given by

$$\begin{aligned} \Theta(\boldsymbol{\lambda}) &= \max_{z, \boldsymbol{u} \geq \boldsymbol{0}} L(z, \boldsymbol{u}, \boldsymbol{\lambda}) = \max_{z, \boldsymbol{u} \geq \boldsymbol{0}} \Big(z - \sum_{i=0}^{k} \lambda_i \big(z - f(\boldsymbol{x}_i) - \boldsymbol{u}^\mathsf{T} \boldsymbol{g}(\boldsymbol{x}_i) \big) \Big) \\ &= \max_{z, \boldsymbol{u} \geq \boldsymbol{0}} \Big(\sum_{i=0}^{k} \lambda_i f(\boldsymbol{x}_i) + \big(1 - \sum_{i=0}^{k} \lambda_i \big) z + \boldsymbol{u}^\mathsf{T} \big(\sum_{i=0}^{k} \lambda_i \boldsymbol{g}(\boldsymbol{x}_i) \big) \Big). \end{aligned}$$

Note that here the primal problem is a maximization problem. Looking at the dual function, we have to impose constraints in order to get a finite objective value of the dual function and therefore a valid solution to the dual problem. As we can choose z without restriction, $1 - \sum_{i=0}^{k} \lambda_i = 0$. Similarly, $\sum_{i=0}^{k} \lambda_i \boldsymbol{g}(\boldsymbol{x}_i) \leq \boldsymbol{0}$, since otherwise we can choose $\boldsymbol{u}$ to be such that the supremum of $L(z, \boldsymbol{u}, \boldsymbol{\lambda})$ tends to infinity. Note that the maximizing $\boldsymbol{u}$ will be the all zero vector as long as $\sum_{i=0}^{k} \lambda_i \boldsymbol{g}(\boldsymbol{x}_i) < \boldsymbol{0}$. For the set of $\boldsymbol{\lambda}$ that satisfy these constraints, the dual function simplifies drastically to

$$\Theta(\boldsymbol{\lambda}) = \max_{z, \boldsymbol{u} \geq \boldsymbol{0}} L(z, \boldsymbol{u}, \boldsymbol{\lambda}) = \sum_{i=0}^{k} \lambda_i f(\boldsymbol{x}_i)$$

and we can formulate the dual problem of minimizing $\Theta(\boldsymbol{\lambda})$ as follows

$$\begin{array}{ll} \text{Minimize} & \sum_{i=0}^{k} \lambda_i f(\boldsymbol{x}_i) \\ \text{subject to} & \sum_{i=0}^{k} \lambda_i \boldsymbol{g}(\boldsymbol{x}_i) \leq \boldsymbol{0} \\ & \sum_{i=0}^{k} \lambda_i = 1 \\ & \lambda_i \geq 0 \ \forall i. \end{array}$$

Surprisingly, this is exactly the same problem as the simplified primary recovery problem. This connects the bounds we obtained from the cutting plane method to the solution of the simplified primary recovery problem. Since the simplified primary recovery problem is dual to the cutting plane master problem

$$y_k \leq \Theta(\boldsymbol{u}^\star) = f(\boldsymbol{x}^\star) \leq f(\hat{\boldsymbol{x}}^\star) \leq \sum_{i=0}^{k} \lambda_i^\star f(\boldsymbol{x}_i) = z_k, \qquad \forall k.$$

As a result, the solution to our simplified primary recovery problem will always lie inside the bounds y_k and z_k, which can be made arbitrarily close by increasing the number of iterations. Knowing that modern LP solvers provide the primal and also the dual solution, allows to obtain a primal solution without additional computational cost each time the master problem has been solved.

References

[1] M. S. Bazaraa, H. D. Sherali and C. M. Shetty: Nonlinear Programming: Theory and Algorithms – Third Edition, Wiley-Interscience, 2006.

[2] D. P. Bertsekas, A. Nedic and A. E. Ozdaglar: Convex Analysis and Optimization, Athena Scientific, 2003

[3] T. Larsson, M. Patriksson and A. Strömberg, Ergodic, Primal Convergence in Dual Subgradient Schemes for Convex Programming, 1997

Appendix: Proofs

Proof (Theorem 3.2) For the proof assume that for any multiplier $\boldsymbol{u}$ there exists a $\boldsymbol{x}_u \in \mathcal{X}(\boldsymbol{u})$ which minimizes the Lagrangian $L(\boldsymbol{x}, \boldsymbol{u})$. So the following equations hold by the definition of the dual function

$$\begin{aligned} \Theta(\boldsymbol{u}) &= f(\boldsymbol{x}_u) + \boldsymbol{u}^\mathsf{T} \boldsymbol{g}(\boldsymbol{x}_u) \\ \Theta(\bar{\boldsymbol{u}}) &= f(\boldsymbol{x}_{\bar{u}}) + \bar{\boldsymbol{u}}^\mathsf{T} \boldsymbol{g}(\boldsymbol{x}_{\bar{u}}). \end{aligned}$$

The next two equations hold with inequality, due to the fact that we do not explicitly state that the optimal minimizer $\boldsymbol{x}$ is used

$$\begin{aligned} \Theta(\bar{\boldsymbol{u}}) &\leq f(\boldsymbol{x}_u) + \bar{\boldsymbol{u}}^\mathsf{T} \boldsymbol{g}(\boldsymbol{x}_u) \\ \Theta(\boldsymbol{u}) &\leq f(\boldsymbol{x}_{\bar{u}}) + \boldsymbol{u}^\mathsf{T} \boldsymbol{g}(\boldsymbol{x}_{\bar{u}}). \end{aligned}$$

Using the four equations above we can see that

$$\begin{aligned} \Theta(\boldsymbol{u}) - \Theta(\bar{\boldsymbol{u}}) &\leq f(\boldsymbol{x}_{\bar{u}}) + \boldsymbol{u}^\mathsf{T} \boldsymbol{g}(\boldsymbol{x}_{\bar{u}}) - f(\boldsymbol{x}_{\bar{u}}) - \bar{\boldsymbol{u}}^\mathsf{T} \boldsymbol{g}(\boldsymbol{x}_{\bar{u}}) = (\boldsymbol{u} - \bar{\boldsymbol{u}})^\mathsf{T} \boldsymbol{g}(\boldsymbol{x}_{\bar{u}}) \\ \Theta(\bar{\boldsymbol{u}}) - \Theta(\boldsymbol{u}) &\leq f(\boldsymbol{x}_u) + \bar{\boldsymbol{u}}^\mathsf{T} \boldsymbol{g}(\boldsymbol{x}_u) - f(\boldsymbol{x}_u) - \boldsymbol{u}^\mathsf{T} \boldsymbol{g}(\boldsymbol{x}_u) = (\bar{\boldsymbol{u}} - \boldsymbol{u})^\mathsf{T} \boldsymbol{g}(\boldsymbol{x}_u) \end{aligned}$$

while the last equation implies, that

$$\Theta(\boldsymbol{u}) - \Theta(\bar{\boldsymbol{u}}) \geq (\boldsymbol{u} - \bar{\boldsymbol{u}})^\mathsf{T} \boldsymbol{g}(\boldsymbol{x}_u).$$

Using the Schwartz inequality, which states $\left|\boldsymbol{x}^\mathsf{T} \boldsymbol{y}\right| \leq ||\boldsymbol{x}||\,||\boldsymbol{y}||$, it follows that

$$\begin{aligned} 0 &\geq \Theta(\boldsymbol{u}) - \Theta(\bar{\boldsymbol{u}}) - (\boldsymbol{u} - \bar{\boldsymbol{u}})^\mathsf{T} \boldsymbol{g}(\boldsymbol{x}_{\bar{u}}) \\ &\geq (\boldsymbol{u} - \bar{\boldsymbol{u}})^\mathsf{T} \boldsymbol{g}(\boldsymbol{x}_u) - (\boldsymbol{u} - \bar{\boldsymbol{u}})^\mathsf{T} \boldsymbol{g}(\boldsymbol{x}_{\bar{u}}) \\ &= (\boldsymbol{u} - \bar{\boldsymbol{u}})^\mathsf{T} (\boldsymbol{g}(\boldsymbol{x}_u) - \boldsymbol{g}(\boldsymbol{x}_{\bar{u}})) \\ &\geq -\,||\boldsymbol{u} - \bar{\boldsymbol{u}}||\,||\boldsymbol{g}(\boldsymbol{x}_u) - \boldsymbol{g}(\boldsymbol{x}_{\bar{u}})||\,. \end{aligned}$$

This can be rewritten as

$$0 \geq \frac{\Theta(\boldsymbol{u}) - \Theta(\bar{\boldsymbol{u}}) - (\boldsymbol{u} - \bar{\boldsymbol{u}})^\mathsf{T} \boldsymbol{g}(\boldsymbol{x}_{\bar{u}})}{||\boldsymbol{u} - \bar{\boldsymbol{u}}||} \geq -\,||\boldsymbol{g}(\boldsymbol{x}_u) - \boldsymbol{g}(\boldsymbol{x}_{\bar{u}})||$$

Only if $\mathcal{X}(\bar{\boldsymbol{u}})$ is the singleton $\{\boldsymbol{x}_{\bar{u}}\}$, it can be shown that for $\boldsymbol{u} \to \bar{\boldsymbol{u}}$ also $\boldsymbol{x}_u \to \boldsymbol{x}_{\bar{u}}$. It follows by the continuity of $\boldsymbol{g}(\boldsymbol{x})$ that for $\boldsymbol{x}_u \to \boldsymbol{x}_{\bar{u}}$ also $\boldsymbol{g}(\boldsymbol{x}_u) \to \boldsymbol{g}(\boldsymbol{x}_{\bar{u}})$, and we have

$$\lim_{\boldsymbol{u}\to\bar{\boldsymbol{u}}} \frac{\Theta(\boldsymbol{u}) - \Theta(\bar{\boldsymbol{u}}) - (\boldsymbol{u} - \bar{\boldsymbol{u}})^{\mathsf{T}} \boldsymbol{g}(\boldsymbol{x}_{\bar{u}})}{||\boldsymbol{u} - \bar{\boldsymbol{u}}||} = 0.$$

Lets remember that in general a function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be differentiable at $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ if there exist a vector $\nabla f(\bar{\boldsymbol{x}})$ (the gradient vector) and a function $\alpha : \mathbb{R}^n \to \mathbb{R}$ such that

$$f(\boldsymbol{x}) = f(\bar{\boldsymbol{x}}) + \nabla f(\bar{\boldsymbol{x}})^{\mathsf{T}}(\boldsymbol{x} - \bar{\boldsymbol{x}}) + ||\boldsymbol{x} - \bar{\boldsymbol{x}}||\, \alpha(\bar{\boldsymbol{x}}, \boldsymbol{x} - \bar{\boldsymbol{x}}) \quad \text{for each } \boldsymbol{x} \in \mathbb{R}^n$$

with

$$\lim_{\boldsymbol{x}\to\bar{\boldsymbol{x}}} \alpha(\bar{\boldsymbol{x}}, \boldsymbol{x} - \bar{\boldsymbol{x}}) = 0.$$

With this definition and the above result it is revealed that $\Theta(\boldsymbol{u})$ is differentiable at $\bar{\boldsymbol{u}}$ with gradient $\nabla\Theta(\bar{\boldsymbol{u}}) = \boldsymbol{g}(\boldsymbol{x}_{\bar{u}})$. ∎

Proof (Lemma 3.5) For the proof assume $\lambda_2 > \lambda_1 > 0$ and let $\bar{\boldsymbol{x}}$ be a certain point in $\mathbb{R}^n$. Assume the direction $\boldsymbol{d} \in \mathbb{R}^n$ to be non-zero, in accordance with the lemma. By the concavity of f it holds, that

$$\begin{aligned} f(\bar{\boldsymbol{x}} + \lambda_1 \boldsymbol{d}) &= f\left[\frac{\lambda_1}{\lambda_2}(\bar{\boldsymbol{x}} + \lambda_2 \boldsymbol{d}) + \left(1 - \frac{\lambda_1}{\lambda_2}\right)\bar{\boldsymbol{x}}\right] \\ &\geq \frac{\lambda_1}{\lambda_2} f(\bar{\boldsymbol{x}} + \lambda_2 \boldsymbol{d}) + \left(1 - \frac{\lambda_1}{\lambda_2}\right) f(\bar{\boldsymbol{x}}). \end{aligned}$$

This implies

$$\frac{f(\bar{\boldsymbol{x}} + \lambda_1 \boldsymbol{d}) - f(\bar{\boldsymbol{x}})}{\lambda_1} \geq \frac{f(\bar{\boldsymbol{x}} + \lambda_2 \boldsymbol{d}) - f(\bar{\boldsymbol{x}})}{\lambda_2}.$$

Thus, the quotient $[f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) - f(\bar{\boldsymbol{x}})]/\lambda$ increases monotonously as $\lambda \to 0^+$. Given any $\lambda \geq 0$ and again considering the concavity of f, we have

$$\begin{aligned} f(\bar{\boldsymbol{x}}) &= f\left[\frac{\lambda}{1+\lambda}(\bar{\boldsymbol{x}} - \boldsymbol{d}) + \frac{1}{1+\lambda}(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d})\right] \\ &\geq \frac{\lambda}{1+\lambda} f(\bar{\boldsymbol{x}} - \boldsymbol{d}) + \frac{1}{1+\lambda} f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) \end{aligned}$$

and

$$\begin{aligned} (1+\lambda) f(\bar{\boldsymbol{x}}) &\geq \lambda f(\bar{\boldsymbol{x}} - \boldsymbol{d}) + f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) \\ \lambda(f(\bar{\boldsymbol{x}}) - f(\bar{\boldsymbol{x}} - \boldsymbol{d})) &\geq f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) - f(\bar{\boldsymbol{x}}) \\ f(\bar{\boldsymbol{x}}) - f(\bar{\boldsymbol{x}} - \boldsymbol{d}) &\geq \frac{f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) - f(\bar{\boldsymbol{x}})}{\lambda}. \end{aligned}$$

Therefore, $f(\bar{\boldsymbol{x}}) - f(\bar{\boldsymbol{x}} - \boldsymbol{d})$ is a constant upper bound on the monotone increasing quotient $[f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) - f(\bar{\boldsymbol{x}})]/\lambda$. Consequently, the limit $\lambda \to 0^+$ exists. ∎

Proof (Theorem 3.6) For the first part of the proof assume $\boldsymbol{x}_{\bar{u}+\lambda d} \in \mathcal{X}(\bar{\boldsymbol{u}} + \lambda \boldsymbol{d})$, i.e., $\boldsymbol{x}_{\bar{u}+\lambda d}$ is a minimizer of $L(\boldsymbol{x}, \bar{\boldsymbol{u}} + \lambda \boldsymbol{d})$. Then the following equation holds by the definition of the dual function

$$\begin{aligned}\Theta(\bar{\boldsymbol{u}} + \lambda \boldsymbol{d}) - \Theta(\bar{\boldsymbol{u}}) &= f(\boldsymbol{x}_{\bar{u}+\lambda d}) + (\bar{\boldsymbol{u}} + \lambda \boldsymbol{d})^{\mathsf{T}} \boldsymbol{g}(\boldsymbol{x}_{\bar{u}+\lambda d}) - \Theta(\bar{\boldsymbol{u}}) \\ &= f(\boldsymbol{x}_{\bar{u}+\lambda d}) + \bar{\boldsymbol{u}}^{\mathsf{T}} \boldsymbol{g}(\boldsymbol{x}_{\bar{u}+\lambda d}) - \Theta(\bar{\boldsymbol{u}}) + \lambda \boldsymbol{d}^{\mathsf{T}} \boldsymbol{g}(\boldsymbol{x}_{\bar{u}+\lambda d}).\end{aligned}$$

As $\boldsymbol{x}_{\bar{u}+\lambda d}$ need not to be a minimizer of $L(\boldsymbol{x}, \bar{\boldsymbol{u}})$, i.e., maybe $\boldsymbol{x}_{\bar{u}+\lambda d} \notin \mathcal{X}(\bar{\boldsymbol{u}})$,

$$\begin{aligned}\Theta(\bar{\boldsymbol{u}}) &\leq f(\boldsymbol{x}_{\bar{u}+\lambda d}) + \bar{\boldsymbol{u}}^{\mathsf{T}} \boldsymbol{g}(\boldsymbol{x}_{\bar{u}+\lambda d}) \\ 0 &\leq f(\boldsymbol{x}_{\bar{u}+\lambda d}) + \bar{\boldsymbol{u}}^{\mathsf{T}} \boldsymbol{g}(\boldsymbol{x}_{\bar{u}+\lambda d}) - \Theta(\bar{\boldsymbol{u}})\end{aligned}$$

and consequently

$$\Theta(\bar{\boldsymbol{u}} + \lambda \boldsymbol{d}) - \Theta(\bar{\boldsymbol{u}}) \geq \lambda \boldsymbol{d}^{\mathsf{T}} \boldsymbol{g}(\boldsymbol{x}_{\bar{u}+\lambda d}).$$

Dividing by λ, taking the positive limit towards zero, $\lambda \to 0^+$, and observing that therefore $\boldsymbol{x}_{\bar{u}+\lambda d} \to \boldsymbol{x}_{\bar{u}} \in \mathcal{X}(\bar{\boldsymbol{u}})$, it holds that

$$\Theta'(\bar{\boldsymbol{u}}; \boldsymbol{d}) = \lim_{\lambda \to 0^+} \frac{\Theta(\bar{\boldsymbol{u}} + \lambda \boldsymbol{d}) - \Theta(\bar{\boldsymbol{u}})}{\lambda} \geq \boldsymbol{d}^{\mathsf{T}} \boldsymbol{g}(\boldsymbol{x}_{\bar{u}}).$$

As $\boldsymbol{g}(\boldsymbol{x}_{\bar{u}}) \in \partial\Theta(\bar{\boldsymbol{u}})$, the former equation completes the first part of the proof implying that

$$\Theta'(\bar{\boldsymbol{u}}; \boldsymbol{d}) \geq \inf \{\boldsymbol{d}^{\mathsf{T}} \boldsymbol{\xi} : \boldsymbol{\xi} \in \partial\Theta(\bar{\boldsymbol{u}})\}. \tag{13}$$

Lets go for the second part: For any subgradient $\boldsymbol{\xi}$ at $\bar{\boldsymbol{u}}$, it follows by the definition of the subgradient for a concave function (1), that

$$\begin{aligned}\Theta(\bar{\boldsymbol{u}} + \lambda \boldsymbol{d}) &\leq \Theta(\bar{\boldsymbol{u}}) + (\bar{\boldsymbol{u}} + \lambda \boldsymbol{d} - \bar{\boldsymbol{u}})^{\mathsf{T}} \boldsymbol{\xi} = \Theta(\bar{\boldsymbol{u}}) + \lambda \boldsymbol{d}^{\mathsf{T}} \boldsymbol{\xi} \\ \Theta(\bar{\boldsymbol{u}} + \lambda \boldsymbol{d}) - \Theta(\bar{\boldsymbol{u}}) &\leq \lambda \boldsymbol{d}^{\mathsf{T}} \boldsymbol{\xi}.\end{aligned}$$

So again dividing by λ and taking the positive limit towards zero $\lambda \to 0^+$

$$\lim_{\lambda \to 0^+} \frac{\Theta(\bar{\boldsymbol{u}} + \lambda \boldsymbol{d}) - \Theta(\bar{\boldsymbol{u}})}{\lambda} \leq \boldsymbol{d}^{\mathsf{T}} \boldsymbol{\xi}$$

it is obtained that $\Theta'(\bar{\boldsymbol{u}}; \boldsymbol{d}) \leq \boldsymbol{d}^{\mathsf{T}} \boldsymbol{\xi}$. This holds true for any $\boldsymbol{\xi} \in \partial\Theta(\bar{\boldsymbol{u}})$ and accordingly we end up with

$$\Theta'(\bar{\boldsymbol{u}}; \boldsymbol{d}) \leq \inf \{\boldsymbol{d}^{\mathsf{T}} \boldsymbol{\xi} : \boldsymbol{\xi} \in \partial\Theta(\bar{\boldsymbol{u}})\}. \tag{14}$$

The results (13) and (14) can only hold at the same time without contradiction, if

$$\Theta'(\bar{\boldsymbol{u}}; \boldsymbol{d}) = \inf\{\boldsymbol{d}^{\mathsf{T}} \boldsymbol{\xi} : \ \boldsymbol{\xi} \in \partial\Theta(\bar{\boldsymbol{u}})\}$$

and the proof is complete. ∎

Proof (Theorem 5.1) For the proof notice that $\boldsymbol{u}_k$ maximizes $\hat{\Theta}_k(\boldsymbol{u})$. Since $\boldsymbol{g}(\boldsymbol{x}_{u_i})$ is a subgradient of $\Theta(\boldsymbol{u})$ at $\boldsymbol{u}_i$ for all i , it holds that

$$\Theta(\boldsymbol{u}) \leq \Theta(\boldsymbol{u}_i) + \boldsymbol{g}(\boldsymbol{x}_{u_i})^{\mathsf{T}}(\boldsymbol{u} - \boldsymbol{u}_i), \qquad \forall \boldsymbol{u} \geq \boldsymbol{0},$$

and by the definition of $\hat{\Theta}_k(\boldsymbol{u})$

$$\Theta(\boldsymbol{u}) \leq \hat{\Theta}_k(\boldsymbol{u}) \leq \hat{\Theta}_k(\boldsymbol{u}_k). \tag{15}$$

Assume that the sequence $\{\boldsymbol{u}_k\}$ converges to $\bar{\boldsymbol{u}} \geq \boldsymbol{0}$. Therefore, we obtain

$$\Theta(\bar{\boldsymbol{u}}) \leq \hat{\Theta}_k(\bar{\boldsymbol{u}}) \leq \hat{\Theta}_k(\boldsymbol{u}_k) \leq \Theta(\boldsymbol{u}_i) + \boldsymbol{g}(\boldsymbol{x}_{u_i})^{\mathsf{T}}(\boldsymbol{u}_k - \boldsymbol{u}_i) \tag{16}$$

for all k and $i < k$. Letting $k \to \infty$ and $i \to \infty$

$$\lim_{k\to\infty} \sup\{\hat{\Theta}_k(\boldsymbol{u}_k) : k\} \leq \lim_{i\to\infty, k\to\infty} \sup\{\Theta(\boldsymbol{u}_i) + \boldsymbol{g}(\boldsymbol{x}_{u_i})^{\mathsf{T}}(\boldsymbol{u}_k - \boldsymbol{u}_i) : i, k\}.$$

Assuming that the sequence $\{\boldsymbol{g}(\boldsymbol{x}_{u_k})\}$ is bounded

$$\lim_{i\to\infty, k\to\infty} \boldsymbol{g}(\boldsymbol{x}_{u_i})^{\mathsf{T}}(\boldsymbol{u}_k - \boldsymbol{u}_i) = 0,$$

implying that

$$\lim_{i\to\infty} \sup\{\Theta(\boldsymbol{u}_i) : i\} \geq \lim_{k\to\infty} \sup\{\hat{\Theta}_k(\boldsymbol{u}_k) : k\}.$$

With (16) it follows, that

$$\lim_{k\to\infty} \inf\{\hat{\Theta}_k(\boldsymbol{u}_k) : k\} \geq \Theta(\bar{\boldsymbol{u}}).$$

By the upper-semicontinuity of $\Theta(\boldsymbol{u})$, it holds that

$$\Theta(\bar{\boldsymbol{u}}) \geq \lim_{i\to\infty} \sup\{\Theta(\boldsymbol{u}_i) : i\}.$$

Consequently,

$$\lim_{i\to\infty} \sup\{\Theta(\boldsymbol{u}_i) : i\} \geq \lim_{k\to\infty} \sup\{\hat{\Theta}_k(\boldsymbol{u}_k) : k\} \geq \lim_{k\to\infty} \inf\{\hat{\Theta}_k(\boldsymbol{u}_k) : k\} \geq \lim_{i\to\infty} \sup\{\Theta(\boldsymbol{u}_i) : i\}.$$

This shows that

$$\lim_{k\to\infty} \hat{\Theta}_k(\boldsymbol{u}_k) = \Theta(\bar{\boldsymbol{u}}).$$

Combining this with (15), we end up with

$$\Theta(\bar{\boldsymbol{u}}) \geq \Theta(\boldsymbol{u}), \qquad \forall \boldsymbol{u}.$$

This shows that $\bar{\boldsymbol{u}}$ must be a dual optimal solution. ∎