ESCUELA POLITÉCNICA NACIONAL

DEPARTAMENTO DE MATEMÁTICA

BILEVEL IMAGING LEARNING WITH TOTAL VARIATION REGULARIZATION: OPTIMALITY CONDITIONS AND TRUST-REGION SOLUTION ALGORITHMS

TRABAJO PREVIO A LA OBTENCIÓN DEL TÍTULO DE DOCTOR EN MATEMÁTICA APLICADA

TESIS

DAVID ALEJANDRO VILLACÍS PROAÑO
david.villacis01@epn.edu.ec

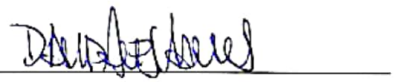
Director: JUAN CARLOS DE LOS REYES PHD
juan.delosreyes@epn.edu.ec

QUITO, MARZO 2022

DECLARACIÓN

Yo, DAVID ALEJANDRO VILLACÍS PROAÑO declaro bajo juramento que el trabajo aquí escrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual, correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su reglamento y por la normatividad institucional vigente.



David Alejandro Villacís Proaño

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por DAVID ALEJANDRO VILLACÍS PROAÑO, bajo mi supervisión.

David Alejandro Villacís Proaño

Juan Carlos De los Reyes PHD
Director de Tesis

ACKNOWLEDGMENTS

This project would not have been possible without the support of many people. Many thanks to my advisor, Juan Carlos De los Reyes, who read my numerous revisions, helped make some sense of the confusion, and guided me through this challenging journey with patience. I'm proud of, and grateful for, my time working with Juan Carlos. Also thanks to my colleagues working at the Research Center on Mathematical Modeling-ModeMat, Luis Miguel Torres, Pedro Merino, Tuomo Valkonen, and the director of the doctoral program Sergio Gonzalez, whom I had the privilege to work with and provided thoughtful conversations and support throughout all these years. Furthermore, I'm grateful to the Mathematics Department at Escuela Politécnica Nacional, especially Diego Recalde and Miguel Yangari for their invaluable help.

Finally, I thank Sofía López, Pablo Velarde, Evelyn Cueva, Cristhian Nuñez, Kateryn Herrera, Paula Castro, and Myrian Guanoluiza for accompanying me and sincerely supporting me in all my ups and downs.

This thesis was developed under the financial support of the following projects: *Métodos Multimalla para la Resolución Numérica de Problemas de Optimización no Suave y Aplicaciones a la Ingeniería, EPN-PIGR 19-02* and the project *Sistema de Pronóstico del Tiempo para todo el Territorio Ecuatoriano: Modelización Numérica y Asimilación de Datos, INAMHI-MODEMAT-EPN 20160041V.*

To Luis, Marianela, Diana, Abigail and Tomás. You are all my reasons.

Contents

List of Figures

List of Tables

List of Algorithms

Abstract

We address the problem of optimal scale-dependent parameter learning in total variation image denoising. Such problems are formulated as bilevel optimization instances with total variation denoising problems as lower-level constraints. For the bilevel problem, we can derive M-stationarity conditions after characterizing the corresponding Mordukhovich generalized normal cone and verifying suitable constraint qualification conditions. We also derive B-stationarity conditions, after investigating the Lipschitz continuity and directional differentiability of the lower-level solution operator. A characterization of the Bouligand subdifferential of the solution mapping, by means of a properly defined linear system, is provided as well. Based on this characterization, we propose a two-phase non-smooth trust-region algorithm for the numerical solution of the bilevel problem and test it computationally for two particular experimental settings.

Chapter 1

Introduction

Computational imaging methods aim to estimate a good-quality image from noisy, incomplete, or indirect measurements. For example, image denoising and image deconvolution try to recover a clean version from a noisy and blurry input image. Image inpainting tries to complete missing information from an image. The damage to an image can be caused by different sources, such as poor lighting conditions, problems in the transmission media, floating-point rounding errors in the analog-to-digital (AD) conversion, or image compression.

Image reconstruction applications are often ill-posed; furthermore, said problems are a type of inverse problem [29]. Therefore, using regularization techniques for inverse problems [79], existing reconstruction methods make different assumptions about the characteristics of the recovered image. Consequently, specific regularizers apply a priori information based on observed properties of the desired output image, such as a tendency to have smooth regions with sharp edges or a form of sparsity on the image gradients, i.e., total variation, see [71].

A critical factor in the quality of the reconstruction obtained by image processing methods is the choice of the parameters in the imaging model. Indeed, a poor choice of parameters can lead to a very bad reconstruction. Furthermore, several questions are of interest in this regard, e.g., what makes a set of "good" parameters? How many parameters should be learned? How can we learn these "good" parameters?

While we can apply a *model-free* approach such as a grid search or random search for finding these optimal parameters, these strategies do not scale with a large number of parameters because of the exponential growth of the grid with the size of the parameters, as reported in [7]. Alternatively, using a *model-based* approach where we define a loss function that describes a "good" parameter makes an assumption of the parameter landscape, allows for the use of optimization techniques to characterize optimal parameters and derive numerical methods for finding them. Moreover, the main benefit

of these optimization-based strategies is that they allow the use of a large number of parameters; indeed, it is the case for most imaging applications.

Data-driven approaches use suitable training data to define a loss function based on an image quality metric, i.e., it measures the quality of the reconstruction with respect to a ground-truth image. This measure is further used to determine the correct model parameters and the structure of the optimal regularizer. Some applications that use this methodology include the work by Tappen et al. [70, 78] where the authors learn the parameters of different Markov random field models. In [60] a learning approach was proposed for learning sparse analysis priors using a smooth version of a l_1 model. Furthermore, applications in the context of sparse coding and dictionary learning are described in [52, 58, 88] and for learning parameters in support vector machines in [46].

Furthermore, these data-driven methods often outperform traditional methods and are gaining popularity partly because of the increased availability of training data and computational resources [36].

Bilevel learning fits into the model-based and data-driven parameter search strategy. Moreover, bilevel programming addresses the problem of optimal parameter search systematically. Indeed, this methodology allows for a precise characterization of the learned parameters and insights into its structure. Bilevel learning methods are so named because they involve two levels of optimization: an upper-level loss function dependent on training data that defines a goal or measure of goodness for the learnable parameters and a lower-level cost function that uses the learnable parameters. By considering a training set (f_k, u_k^{true}) for $k = 1, \dots, p$ containing pairs of damaged and ground-truth information respectively; J as the upper-level loss function and $\mathcal{E}$, the lower level cost function, the bilevel learning paradigm formulates the following optimization problem for finding the optimal parameter θ

$$\min_{\theta} \quad \sum_{k=1}^{P} J(u_k, u_k^{\text{true}}) \tag{1.1a}$$

$$\text{s.t.} \quad u_k \in \arg\min_{u \in \mathbb{R}^n} \mathcal{E}(u; \theta, f_k). \tag{1.1b}$$

Lately, bilevel optimization techniques have had a strong presence in the machine learning community. Particularly in hyperparameter optimization [2], neural architecture search [87], feature learning [30], and sparsity-enforcing regression [62] to name a few. Furthermore, bilevel optimization has been used in the context of neural network training, e.g., in [89], the authors propose a new family of recurrent neural networks with good training stability in the presence of vanishing gradients that is formulated using a stochastic bilevel optimization problem.

When considering a bilevel learning problem with lower-level cost as a variational

imaging model, the seminal work in [25] analyzes a bilevel learning strategy for finding parameters to the corresponding noise model. Likewise, the authors in [48] address the problem of parameter selection using finite-dimensional space for imaging models using Tikhonov and l_1 regularizer. Mixed noise models were studied in [12]. The authors in [11] propose a dynamic sampling technique when dealing with large-size training sets. Furthermore, extensions of the parameter properties with a scale-dependent structure were addressed in [9, 81]. Now, applications for dealing with image segmentation models were found in [64], learning the sample pattern for magnetic resonance imaging (MRI) [75], and non-local models in [19].

This work will focus on bilevel learning problems where the lower-level problem is an image denoising model involving the total variation regularizer. Moreover, we will consider a bilevel parameter learning problem using a finite-dimensional space. Mainly, we are interested in obtaining optimality conditions and devising a numerical algorithm for finding solutions.

The non-differentiability of the total variation regularizer presents challenges when used within the lower level of a bilevel optimization problem. The main challenge of this problem is to characterize optimality conditions for the bilevel problem since it fails to satisfy classical constraint qualification conditions, see section 3.4. Therefore, the requirements for the existence of Lagrange multipliers are not met. This phenomenon leads to alternative notions of stationarity, as detailed in section 2.6 and the references therein. Indeed, we may find stationarity conditions based on different approaches. In particular, assuming Bouligand calculus, we may characterize a Bouligand (B-) stationarity system; Clarke's nonsmooth analysis leads to a Clarke (C-) stationarity system, and Mordukhovich generalized differential calculus applied to a generalized normal cone generates a Mordukhovich (M-) stationarity system that can be proven to be sharper than C-stationarity.

A traditional approach for dealing with this problem is to replace the non-smooth term with a smooth approximation. Then, the solution map for the lower-level problem presents a Gateaux differentiable solution map for which we can derive an optimality system. Furthermore, using an asymptotic analysis of the smooth optimality system, it is possible to retrieve a characterization for Clarke (C-) stationary points. Moreover, we can use the smooth optimality system to solve the problem numerically; indeed, quasi-Newton [11] and Newton semi-smooth [48] methods have been proposed for this task.

This thesis aims to characterize **sharper** stationary points using two different approaches. First, we will explore a generalized mathematical problem with equilibrium constraints (GMPEC) reformulation and verify a suitable constraint qualification condition to characterize Mordukhovich (M-) stationary points, keeping it as a purely

theoretical result. Second, we investigate the differentiability properties of the solution map to obtain a characterization of Bouligand (B-) stationary points. This result will allow us to describe the linear elements of the Bouligand subdifferential of the solution map that will be used for implementing a non-smooth trust region algorithm designed for Lipschitz-continuous functions [18] to obtain a numerical solution. This algorithm is numerically easier to solve and takes fewer iterations when compared to a classical regularized gradient obtained by smoothing the problem. Additionally, we will explore the performance of different parameter structures such as scalar, scale-dependent, and patch-dependent parameters; then their generalization capabilities when dealing with a test set.

The structure of this thesis is as follows. Chapter 2 will set up the notation and some preliminary results that will be used in later chapters. Next, in Chapter 3 we will introduce the bilevel learning problem for learning parameters in imaging denoising models. After that, we will split the discussion into two topics: the lower-level problem's solution uniqueness, existence, and numerical treatment using a matrix-free implementation of the gradient operator and a review of the smooth version of the bilevel problem and its optimality conditions. We finalize this chapter by showing the challenges of this problem to satisfy classical constraint qualification conditions.

Then, this work will distinguish two different bilevel learning problem settings: the problem for learning parameters appearing in the data fidelity term in Chapter 4 and parameters appearing in the regularization term in Chapter 5. In both chapters, we will start by reformulating the bilevel problem as a generalized MPEC, justify suitable constraint qualification conditions and characterize M-stationary points. Furthermore, we will study the properties of the solution operator for the lower-level problem regarding its Bouligand differentiability, which will help us characterize B-stationary points. Finally, with the use of a proper characterization of the linear elements of the Bouligand subdifferential of the solution map, we will design and implement a non-smooth trust-region algorithm to learn optimal patch-based parameters using the CelebA faces dataset [50].

Chapter 2

Preliminaries

We denote the set of all linear mappings between $\mathbb{R}^n$ and $\mathbb{R}^m$ as $\boldsymbol{L}(\mathbb{R}^n, \mathbb{R}^m)$. The scalar product of two vectors $v, w \in \mathbb{R}^n$ will be noted as $\langle v, w \rangle$ and its corresponding Euclidean norm as $\|\cdot\|$.

2.1 Orthogonal Complements

Let $V \subseteq \mathbb{R}^n$ be a finite-dimensional space endowed with an inner product $\langle \cdot, \cdot \rangle$ and its corresponding Euclidean norm $\|\cdot\|$.

Definition 2.1. *Let S be a non-empty subset of the space V. We define $S^\perp$ to be the set of all vectors in V that are orthogonal to every vector in S; that is,*

$$S^\perp = \{x \in V \ \ \langle x, y \rangle = 0, \ \forall y \in S\}.$$

Moreover, the set $S^\perp$ is called the ***orthogonal complement*** *of S.*

Furthermore, we will denote the set of all linear combinations of the elements on S as the span(S). Using the definition, it can be seen that $S^\perp$ is a subspace of V and $V^\perp = \{0\}$ for any space V.

In the following theorems, we will present some important properties of the orthogonal complement; for a more rigorous review, we refer the reader to [31].

Theorem 2.1. *Let W be a subspace of V. Then $(W^\perp)^\perp = W$.*

Theorem 2.2. *[31, pg. 355] Let W be a subspace of the space V, and let $y \in V$. Then there exist unique vectors $u \in W$ and $z \in W^\perp$ such that $y = u + z$. Thanks to this property, it holds*

$$dim(V) = dim(W) + dim(W^\perp).$$

Theorem 2.3. *[31, pg. 355] Let W_1 and W_2 be subspaces of a finite-dimensional inner product space. Then $(W_1 + W_2)^\perp = W_1^\perp \cap W_2^\perp$ and $(W_1 \cap W_2)^\perp = W_1^\perp + W_2^\perp$.*

Now, for a given matrix $A \in \mathbb{R}^{m\times n}$ we define the subspace generated by its columns as range(A)

$$\text{range}(A) := \{b \in \mathbb{R}^m : Ax = b \text{ for some } x \in \mathbb{R}^n\}.$$

Likewise, the space generated by its rows is range$(A^\top)$. Moreover, the set of vectors that satisfy $Ax = 0$ will be noted as $\ker(A)$ and is defined as the nullspace of A

$$\ker(A) := \{x \in \mathbb{R}^n : Ax = 0\}.$$

Theorem 2.4. *[77, pg. 162] Let $A \in \mathbb{R}^{m\times n}$. Then, the nullspace of a matrix is the orthogonal complement of its row space, i.e., $\ker(A) = \text{range}(A^\top)^\perp$ and $\ker(A^\top) = \text{range}(A)^\perp$.*

2.2 Convex Analysis

A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if it satisfies the following property

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y),\ \forall x, y \in \mathbb{R}^n,\ t \in [0,1].$$

It is **strictly convex** if the above inequality is strict whenever $x \neq y$ and $t \in (0,1)$. Furthermore a function is **lower semi-continuous** if, for all $x \in \mathbb{R}^n$, if $x_n \to x$, then

$$f(x) \leq \liminf_{n\to\infty} f(x_n).$$

Moreover, we will say that f is **proper** if $f(x) < \infty$ for at least one $x \in \mathbb{R}^n$.

Now, given a convex, lower semi-continuous function $f : \mathbb{R}^n \to \mathbb{R}$, we recall that its **subgradient** at a point x is defined as the set

$$\partial f(x) := \{p \in \mathbb{R}^n :\ f(y) \geq f(x) + \langle p, y - x\rangle,\ \forall y \in \mathbb{R}^n\}.$$

It is worth noting that in the case f is differentiable, the subgradient is a singleton containing the gradient of the function, i.e., $\partial f(x) = \{\nabla f(x)\}$. The function is **strongly convex** if in addition, for $x, y \in \mathbb{R}^n$ and $p \in \partial f(x)$, we have

$$f(y) \geq f(x) + \langle p, y - x\rangle + \frac{c}{2}\|y - x\|^2.$$

Definition 2.2. *Let $\mathcal{C} \subseteq \mathbb{R}^n$ be an arbitrary cone. Then $\mathcal{C}^* := \{v \in \mathbb{R}^n :\ \langle v, d\rangle \geq 0,\ \forall d \in \mathcal{C}\}$ denotes de dual cone of $\mathcal{C}$.*

2.2.1 Fenchel Conjugate

For any function $f : \mathbb{R}^n \to \mathbb{R}$ one can associate its **Fenchel Conjugate** as follows

$$f^*(y) = \sup_{x \in \mathbb{R}^n} \langle y, x \rangle - f(x), \tag{2.1}$$

which can be seen as the supremum of linear continuous functions, and therefore, convex and lower semi-continuous. Furthermore, we define the **biconjugate** as f^{**} as the conjugate of the conjugate of f. This function is the largest convex, lower semi-continuous function below f; in particular, when f is already convex and lower semi-continuous, we have $f^{**} = f$.

Furthermore, using the definition (2.1) and assuming f is a convex function, we see that x realizes the supremum in (2.1) if and only if $y \in \partial f(x)$ and we have $f(x)+f^*(y) = \langle y, x \rangle$. Also, it follows that $x \in \partial f^*(y)$, from where we deduce the **Legendre-Fenchel identity**

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

2.2.2 Proximal Map

If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, proper and lower semi-continuous, then, for any x, there is a unique minimizer y^* to the strongly convex problem

$$\min_{y \in \mathbb{R}^n} \quad g(y) := f(y) + \frac{1}{2\tau}\|y - x\|^2.$$

Moreover, if g is strongly convex, also satisfies

$$f(y) + \frac{1}{2\tau}\|y - x\|^2 \geq f(y^*) + \frac{1}{2\tau}\|y^* - x\| + \frac{1}{2\tau}\|y - y^*\|^2, \forall y \in \mathbb{R}^n$$

Finally, we define the **proximal map** of the function f at x as $y^* := \text{prox}_{\tau f}(x)$.

2.2.3 Fenchel Duality

An essential notion in convex programming is indeed convex duality. This notion transforms convex problems into other problems which sometimes have a nicer structure that can be exploited. Let us consider the following optimization problem

$$\min_x f(x) + g(\mathbb{K}x),$$

where $f : \mathbb{R}^n \supseteq \Omega_2 \to \mathbb{R}$ and $g : \mathbb{R}^n \supseteq \Omega_1 \to \mathbb{R}$ are convex, lower semi-continuous functions and $\mathbb{K} : \Omega_1 \to \Omega_2$ is a bounded linear operator. Then, since in this case, we

have $f = f^{**}$, the following equality holds true

$$\min_{x\in\Omega_1} f(x) + g(\mathbb{K}x) = \min_{x\in\Omega_1} \sup_{y\in\Omega_2} \langle y, \mathbb{K}x\rangle - g^*(y) + f(x).$$

Assuming that there exists a point x such that $\mathbb{K}x$ is in the relative interior of the domain of g and x in the relative interior of the domain of f [28, Chapter 3, Theorem. 4.2], we are able to formulate its **Fenchel-Rockafellar** dual problem [67, Section 31] as follows

$$\begin{aligned}\min_{x\in\Omega_2} f(x) + g(\mathbb{K}x) &= \min_{x\in\Omega_2} \sup_{y\in\Omega_1} \langle y, \mathbb{K}x\rangle - g^*(y) + f(x),\\ &= \max_{y\in\Omega_1} \inf_{x\in\Omega_2} \langle y, \mathbb{K}x\rangle - g^*(y) + f(x),\\ &= \max_{y\in\Omega_1} -g^*(y) - f^*(-\mathbb{K}^*y).\end{aligned}$$

Under the assumptions described above, we know it has at least a solution y^*. Naming x^* as the solution for the primal problem, then the tuple (x^*, y^*) is a saddle-point for the primal-dual formulation, i.e., for any $(x, y) \in \Omega_1 \times \Omega_2$ we have

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

where $\mathcal{L}(x, y) := \langle y, \mathbb{K}x\rangle - g^*(y) + f(x)$ denotes the **Lagrangian**. This implies that we can characterize extremality conditions for the primal-dual problem as

$$\begin{aligned}g^*(y) - g^*(y^*) &\geq \langle \mathbb{K}^*x^*, y - y^*\rangle,\\ f(x) - f(x^*) &\geq \langle -\mathbb{K}y^*, x - x^*\rangle,\end{aligned}$$

and these conditions can be equivalently written as the following inclusions

$$\begin{aligned}\mathbb{K}x^* &\in \partial g^*(y^*),\\ -\mathbb{K}^*y^* &\in \partial f(x^*).\end{aligned}$$

2.3 Nonsmooth Analysis

A function $F : \mathbb{R}^n \to \mathbb{R}^m$ is Fréchet differentiable at $x \in \mathbb{R}^n$ if there exists $A \in \boldsymbol{L}(\mathbb{R}^n, \mathbb{R}^m)$ such that

$$\lim_{h\to 0} \frac{\|F(x+h) - F(x) - A(h)\|}{\|h\|} = 0,$$

and we write $F'(x) = A$. We say that F is Fréchet differentiable on $\mathbb{R}^n$ if F is Fréchet differentible at each $x \in \mathbb{R}^n$. We often refer to Fréchet differentiable simply as differentiable.

A function $F : \mathbb{R}^n \to \mathbb{R}^m$ is directionally differentiable at $x \in \mathbb{R}^n$ if the following limit exists for all $d \in \mathbb{R}^n$

$$F'(x; d) = \lim_{t \to 0} \frac{F(x + td) - F(x)}{t}.$$

Now, considering a more general setting, let us take the function $F : \mathbb{R}^n \to \mathbb{R}^m$ to be a locally Lipschitz function, i.e., for any $x \in \mathbb{R}^n$ there exist constants $\epsilon = \epsilon(x) > 0$ and $L = L(x) > 0$ such that

$$\|F(x_1) - F(x_2)\| \leq L\|x_1 - x_2\|, \forall x_1, x_2 \in B(x, \epsilon).$$

We will denote by D_F the set of $x \in \mathbb{R}^n$ where F is differentiable. The following theorem shows a property holding for all locally Lipschitz functions regarding their differentiability

Theorem 2.5 (Rademacher)**.** *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a locally Lipschitz continuous function in F. Then, F is differentiable almost everywhere in $\mathbb{R}^n$, i.e., the set $\mathbb{R}^n \backslash D_F$ has a null measure.*

As a consequence of theorem 2.5, we know that for a locally Lipschitz continuous function, we can approximate each $x \in \mathbb{R}^n$ by a sequence $\{x_k\} \subset D_F$ such that $x_k \to x$. Moreover, we may take a sequence of derivatives at each x_k. The set of all such limits is known as the Bouligand subdifferential at x.

Definition 2.3 (Bouligand Subdifferential)**.** *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ locally Lipschitz continuous and $x \in \mathbb{R}^n$ arbitrary but fixed. Then the set*

$$\partial_B F(x) := \{H \in \mathbb{R}^{m \times n} : \exists \{x_k\} \subseteq D_F,\ x_k \to x \text{ and } F'(x_k) \to H\}, \tag{2.2}$$

*is known as the **Bouligand subdifferential** of F at x.*

Definition 2.4 (Clarke Subdifferential)**.** *The Clarke subdifferential at $x \in \mathbb{R}^n$ is defined as the closure of the convex hull of the Bouligand subdifferential, i.e.,*

$$\partial F(x) = cl(conv(\partial_B F(x)))$$

Theorem 2.6. *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ locally Lipschitz continuous and $x \in \mathbb{R}^n$ arbitrary but fixed. Then*

1. *$\partial_B F(x)$ is non-empty and compact.*

2. *$\partial F(x)$ is non-empty, convex and compact.*

Theorem 2.7. *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ locally Lipschitz continuous, then*

1. *The map $x \to \partial_B F(x)$ is closed, i.e., for any $x_k \to x$, $H_k \to H \in \mathbb{R}^{m\times n}$ with $H_k \in \partial_B F(x_k)$ we have $H \in \partial_B F(x)$*

2. *The map $x \to \partial F(x)$ is closed.*

We say that $F : \mathbb{R}^n \to \mathbb{R}^m$ is Bouligand (B-) differentiable at $x \in \mathbb{R}^n$ if F is directionally differentiable and locally Lipschitz continuous. Moreover, we say that F is Bouligand differentiable in $\mathbb{R}^n$ if it is in every $x \in \mathbb{R}^n$.

Lemma 2.1. *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be Bouligand differentiable, then*

$$\lim_{d_k \to 0} \frac{F(x + d_k) - F(x) - F'(x; d_k)}{\|d_k\|} = 0.$$

Moreover, when dealing with a composition of Bouligand differentiable functions, we have a chain rule

Theorem 2.8. *Let $\Omega_1 \subset \mathbb{R}^n$, $\Omega_2 \subset \mathbb{R}^p$, $F : \Omega_1 \to \mathbb{R}^m$ and $G : \Omega_2 \to \mathbb{R}^n$ two B-differentiable functions at the point $x_0 \in \Omega_2$ and $f(x_0) \in \Omega_1$ respectively. Then the composite function $F \circ G$ is B-differentiable at x_0 and*

$$(F \circ G)'(x_0; d) = F'(G(x_0); G'(x_0; d)).$$

Furthermore, we know B-differentiable functions satisfy the following properties:

1. $(\alpha F + \beta G)'(x_0; d) = \alpha F'(x_0; d) + \beta G'(x_0; d)$.

2. $(FG)'(x_0; d) = F(x_0)G'(x_0; d) + G(x_0)F'(x_0; d)$.

3. If $G(x_0) \neq 0$, then

$$\left(\frac{F}{G}\right)'(x_0; d) = \frac{1}{G(x_0)^2}(G(x_0)F'(x_0; d) - F(x_0)G'(x_0; d)).$$

Regarding the stationarity conditions for optimization problems involving a B-differentiable scalar-valued function $f : \mathbb{R}^n \to \mathbb{R}$

$$\min_{x \in \mathbb{R}^n} \quad f(x), \tag{2.3}$$

the following stationarity condition can be used to characterize a locally optimal solution.

Definition 2.5 (B-stationarity)**.** *Let $x^* \in \mathbb{R}^n$ be a local optimal solution of* (2.3)*, then it satisfies the following necessary optimality condition*

$$f'(x^*; x - x^*) \geq 0, \ \forall x \in \mathbb{R}^n. \tag{2.4}$$

A point x^* satisfying (2.4) *is called* Bouligand (B-) stationary.

2.4 Variational Geometry

Let us consider a set $C \subset \mathbb{R}^n$, representing a set of constraints related to an optimization problem. The boundary of C in this scenario plays a crucial role in characterizing a solution. However, in a general setting, C may have a boundary with many curvilinear facets, edges, and corners. This lack of smoothness prevents the use of standard methods of geometric analysis.

In this section, we are interested in associating with each point of a set C certain cones of tangent and normal vectors, which generalize the tangent and normal subspaces in differential geometry. For a more rigorous review, we refer the reader to [68].

A first fundamental cone is the so-called tangent cone and is defined using a limiting process on the difference quotients as follows

Definition 2.6 (Tangent cone)**.** *A vector d is said to be tangent to C at a point x if there are a feasible direction $\{x_k\}$ approaching x and a sequence of positive scalars $\{t_k\}$ with $t_k \to 0$ such that*

$$\lim_{k \to \infty} \frac{x_k - x}{t_k} = d.$$

The set of all tangent vectors to C at x is called ***tangent cone*** *and is noted by $T_C(x)$.*

In the literature, there exist several equivalent formulations of this cone. In particular, for this work, we will make use of one based on the **distance function**.

Definition 2.7. *The distance of a vector x to a set C is defined as*

$$\operatorname{dist}(x, C) := \inf_{y \in C} \|x - y\|$$

Using this function, [45, Theorem 4.1.12] showed that a vector v is **tangent** to C

at x if the following limit holds true

$$\lim_{t\to 0} \frac{\operatorname{dist}(x+tv, C)}{t} = 0. \tag{2.5}$$

The notion of a normal vector can be seen as a counterpart of tangency, and a first generalization for the normal cone is defined as follows

Definition 2.8 (Fréchet normal cone)**.** *Let C be non-empty and closed, and $x^* \in C$. The Fréchet normal cone is*

$$N_C^F(x^*) := \{w : \langle w, x - x^* \rangle \leq o(\|x - x^*\|),\ \forall x \in C\}.$$

Remark 2.1. *Alternatively, the Fréchet normal cone can be written as the polar cone to $T_C(x^*)$, i.e., $N_C^F(x^*) = [T_C(x^*)]^\circ$.*

One difficulty with the Fréchet normal cone is that it is not outer semicontinuous; see [68, Def. 5.4]. Indeed, we obtain the less "irregular" Mordukhovich normal cone by taking the following limiting process

Definition 2.9 (Mordukhovich normal cone)**.** *Let C be non-empty and closed, and $x^* \in C$ be given. The limiting/Mordukhovich normal cone to C at x^* is*

$$N_C^M(x^*) := \left\{ \lim_{k\to\infty} w_k : \exists \{x_k\} : \lim_{k\to\infty} x_k = x^*,\ w_k \in N_C^F(x_k) \right\}.$$

Remark 2.2. *In the case of C being convex, the Fréchet and Mordukhovich normal cones are equivalent to the classical normal cone from convex analysis.*

Proposition 2.1 (Limits of normal vectors)**.** *Let $\{x_k\} \subset C$ be a sequence such that $x_k \to x^*$, $v_k \in N_C^M(x_k)$ and $v_k \to v$, then $v \in N_C^M(x^*)$. In other words, the set-valued mapping $N_C^M : x \to N_C^M(x)$ is outer semicontinuous at x^* relative to C.*

2.5 Optimality Conditions for Standard Nonlinear Programs

Let us start this discussion by laying out some basic concepts for optimality in classical nonlinear programs, for a more in-depth review we refer the reader to [8, 55]. Let us

consider the following constrained optimization problem

$$\min_{x \in \mathbb{R}^n} \quad f(x) \tag{2.6a}$$

$$\text{s.t.} \quad h_i(x) = 0,\ i \in \mathcal{E}, \tag{2.6b}$$

$$g_i(x) \leq 0,\ i \in \mathcal{I} \tag{2.6c}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $h_i : \mathbb{R}^n \to \mathbb{R}$, $g_i : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable functions on a subset of $\mathbb{R}^n$ and $\mathcal{I}$ and $\mathcal{E}$ are two finite sets of indices. We define the **feasible set** Ω to be the set of all points x satisfying the constraints; that is

$$\Omega := \{x \in \mathbb{R}^n :\ h_i(x) = 0, i \in \mathcal{E},\ g_i(x) \leq 0, i \in \mathcal{I}\}.$$

Given a feasible point x, we call $\{x_k\}$ a **feasible sequence** approaching x if $x_k \in \Omega$ for all k sufficiently large and $x_k \to x$. Now, we characterize a local solution of (2.6) as a point x at which all feasible sequences approaching x have the property that $f(x_k) \geq f(x)$ for all k sufficiently large. Indeed, it is of particular importance to characterize the directions in which any step from a feasible point x remains feasible. Such directions are characterized through the tangent cone, see definition 2.6.

Definition 2.10. *The active set $\mathcal{A}(x)$ at any feasible x consists of the equality constraint indices together with the indices of the inequality constraints i for which $g_i(x) = 0$; that is,*

$$\mathcal{A}(x) = \mathcal{E} \cup \{i \in \mathcal{I} : g_i(x) = 0\}.$$

Likewise, we define the linearized feasible direction set as follows

$$F(x) := \left\{ d : \begin{cases} \langle d, \nabla h_i(x) \rangle = 0, \forall i \in \mathcal{E}, \\ \langle d, \nabla g_i(x) \rangle \geq 0, \forall i \in \mathcal{A}(x) \cup \mathcal{I}. \end{cases} \right\}$$

Constraint qualifications are conditions under which the linearized feasible set $F(x)$ is equal, to the tangent cone $T_\Omega(x)$. Indeed, these sets are not necessarily equal, given that the set F is built by linearizing an algebraic description of Ω at x, and the tangent cone relies completely on the geometry of the constraint set Ω. The most widely used constraint qualification conditions are the following ones

Definition 2.11. *Let $x^* \in \Omega$ be a feasible point for problem (2.6). Then*

1. *the Abadie constraint qualification (ACQ) holds at x^* if the linearized feasible set coincides with the tangent cone, e.g., $F(x^*) = T_\Omega(x^*)$.*

2. *the Guinard constraint qualification (GCQ) holds at x^* if the dual of both the*

linearized feasible set and the tangent cones coincide, e.g., $F(x^*)^* = T_\Omega(x^*)^*$.

Moreover, since the constraint functions in (2.6) are differentiable, these constraint qualification conditions can be satisfied by verifying the following conditions.

Definition 2.12 (LICQ)**.** *If the gradient vectors* $\nabla h_i(x^*)$ *and all active constraint gradients* $\nabla g_i(x^*)$, $i \in \mathcal{A}(x^*)$ *are linearly independent, then the* Linear Independence Constraint Qualification *(LICQ) holds.*

Definition 2.13 (MFCQ)**.** *The* Mangasarian-Fromovitz Constraint Qualification *(MFCQ) holds at a point* x^* *if the gradient vectors* $\nabla h_i(x^*)$ *for* $i \in \mathcal{E}$ *are linearly independent and there exist a vector* $d \in \mathbb{R}^n$ *such that* $\langle \nabla g_i(x^*), d\rangle < 0$ *for all* $i \in \mathcal{A}(x^*)$ *and* $\langle \nabla h_i(x^*), d\rangle = 0$ *for all* $i \in \mathcal{E}$.

Furthermore, the following implications can be proved

$$LICQ \Rightarrow MFCQ \Rightarrow ACQ \Rightarrow GCQ$$

Let $x^* \in \mathbb{R}^n$ be a local solution for (2.6), and assuming that any of the constraint qualification conditions mentioned above holds. Then, there exist Lagrange multiplier vectors (λ^*, μ^*), with components λ_i^* for $i \in \mathcal{E}$ and μ_i^* for $i \in \mathcal{I}$ such that the following KKT-condition [44, 47] holds for this constrained optimization problem at (x^*, λ^*, μ^*)

$$\nabla f(x^*) + \sum_{i\in\mathcal{E}} \lambda_i^* \nabla h_i(x) + \sum_{i\in\mathcal{I}} \mu_i^* \nabla g_i(x) = 0, \tag{2.7a}$$

$$h_i(x^*) = 0, \qquad \forall i \in \mathcal{E}, \tag{2.7b}$$

$$g_i(x^*) \geq 0, \qquad \forall i \in \mathcal{I}, \tag{2.7c}$$

$$\mu_i^* \geq 0, \qquad \forall i \in \mathcal{I}, \tag{2.7d}$$

$$\mu_i^* g_i(x^*) = 0, \qquad \forall i \in \mathcal{I}. \tag{2.7e}$$

Condition (2.7d) is known as the *sign condition*, while (2.7e) is called the *complementary slackness condition*. These complementarity conditions imply that either constraint i is active or $\lambda_i^* = 0$, or possibly both. In particular, the Lagrange multipliers corresponding to the inactive inequality constraints are zero; consequently, we can omit the terms for indices $i \notin \mathcal{A}(x^*)$ from (2.7a) and rewrite it as follows

$$\nabla f(x^*) + \sum_{i\in\mathcal{E}} \lambda_i^* \nabla h_i(x) + \sum_{i\in\mathcal{A}} \mu_i^* \nabla g_i(x) = 0$$

A special case of complementarity is known as **strict complementarity** and it is defined as follows.

Definition 2.14 (Strict Complementarity)**.** *Given a local solution* x^* *of* (2.6) *and a*

vector λ^* satisfying (2.7), we say that the strict complementarity condition holds if exactly one of λ_i^* and $g_i(x^*)$ is zero for each index $i \in \mathcal{I}$.

2.6 Optimality Conditions for Problems with Variational Inequalities of the Second Kind Constraints

Let us consider a type of optimality problem where, among the constraints, there arises a **variational inequality of the second kind**

$$\min_{x \in \mathbb{R}^n} \quad f(x, y) \tag{2.8a}$$

$$\text{s.t.} \quad \langle F(x, y), v - y\rangle + j(v) - j(y) \geq 0, \ \forall v \in \mathbb{R}^n, \tag{2.8b}$$

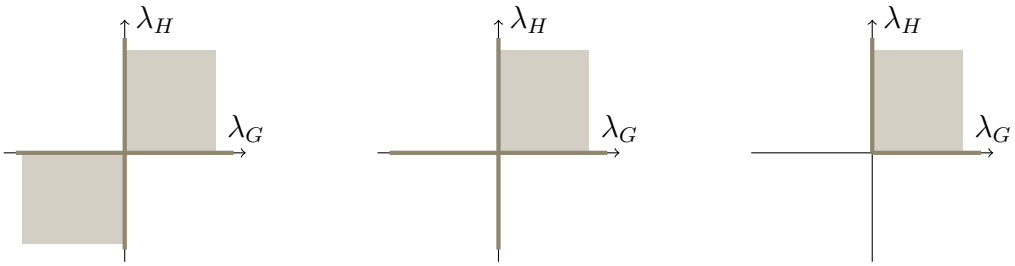
where F maps $\mathbb{R}^n \times \mathbb{R}^m$ into $\mathbb{R}^m$, j is a non-differentiable convex continuous function that maps $\mathbb{R}^m$ into $\mathbb{R}^n$. Due to the non-differentiability of this function, it is impossible to transform (2.8) into the classical Mathematical Program with Equilibrium Constraints (MPEC) setting.

The characterization of optimality systems for this problem remains an active area of research. The structure of the constraints in this problem fails to satisfy classical constraint qualification conditions such as MFCQ, ACQ. This phenomenon leads us to a scenario where the KKT conditions are not necessary optimality conditions and can lead to solutions that are not stationary points. In this scenario, even under the lack of a constraint qualification condition, it is possible to characterize a primal optimality condition to characterize a necessary optimality condition for a local optima of (2.8) known as Bouligand stationarity, see definition 2.5.

It is not hard to see that (2.8) is a generalization of an MPEC problem by taking j as the null function and restricting $y \in \mathbb{R}^n_+$. Indeed, the constraints of the problem now read

$$0 \leq F(x, y) \perp y \geq 0. \tag{2.9}$$

Even in this simplified scenario, there do not exist KKT stationary points since the complementarity structure appearing in the problem's constraints fails to satisfy classical constraint qualification from non-linear programming. Here, several weaker stationarity notions can be derived by making use of tailor-made constraint qualification conditions based on the relaxation of the complementarity structure (2.9). Indeed, several notions of stationarity arise: Mordukhovich stationarity and Clarke stationarity, to name the most used. Considering the Lagrange multipliers corresponding to the inequality constraints in (2.9), the **degenerate set** corresponds to the following index



Clarke stationarity | Mordukhovich stationarity | Strong stationarity

Figure 2.1: Signs of the Lagrange multipliers for indices $i \in \mathcal{B}(x, y)$.

set

$$\mathcal{B}(x, y) := \{i : F_i(x, y) = 0, y_i = 0\}.$$

This set is particularly important since different stationarity notions characterize the multipliers within it differently. It is worth mentioning that when this set is empty, we say that the vector (x^*, y^*) satisfies the **strict complementarity** condition.

In Figure 2.1, we can see the differences in the properties for the Lagrange multipliers within the degenerate set, according to the stationarity criteria. Meaning that different stationarity systems give different kinds of information regarding the multipliers. Moreover, it can be seen that strong stationarity is the one that provides the sharpest characterization, followed by Mordukhovich and Clarke, respectively.

For problem (2.8), fewer stationarity results are available. Indeed, some weak results can be found in [3, 6], and very general conditions can be obtained. The work by De Los Reyes [21] considers a variational inequality of the second kind as a constraint with the non-differentiable term $j(v) = \sum_{i=1}^{n} \|(\mathbb{K}v)_i\|$, with $\mathbb{K}$ being a bounded and linear operator. When exploiting the problem's non-differentiability structure, a C-stationarity system was obtained using a tailored regularization approach. These results were then extended to image processing in [25]. Furthermore, in [22], the authors consider the non-differentiable term as $j(v) = \|v\|_1$ and further characterize stationarity points by investigating the differentiability properties of the solution map. Moreover, for the problem at hand, Bouligand and Strong stationarity conditions were obtained.

Now, let us consider different approach for dealing with the non-differentiable term, Outrata in [59] presents an approach where the variational inequality involved in (2.8) can be reformulated as a **generalized equation**. This methodology allows the use of tools from variational analysis to derive stationarity conditions. In this sense, let us consider the following Generalized Mathematical Program with Equilibrium Con-

straints (GMPEC):

$$\min \quad f(x,y) \tag{2.10a}$$

$$\text{s.t.} \quad 0 \in F_1(x,y) + Q(F_2(x,y)), \tag{2.10b}$$

$$(x,y) \in \omega \tag{2.10c}$$

where $F_1 : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ and $F_2 : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^l$ are two continuously differentiable functions, $\omega \subset \mathbb{R}^n \times \mathbb{R}^m$ closed non-empty and $Q : \mathbb{R}^l \rightrightarrows \mathbb{R}^m$ is a multifunction with closed graph. Indeed, problem (2.8) can be casted in this form, where Q corresponds to the convex subdifferential of J, $F_1(x,y) = \nabla F(x,y)$ and $F_2(x,y) = y$.

The author make use of an exact penalization of the equilibrium constraint to derive a constraint qualification condition and consequently a result for M-stationarity for the problem.

Theorem 2.9 (Outrata [59]). *Let (x^*, y^*) be a local solution of (2.10) and suppose the following constraint qualification*

$$\left.\begin{bmatrix} \nabla_x F_2(x^*,y^*)^\top & -\nabla_x F_1(x^*,y^*)^\top \\ \nabla_y F_2(x^*,y^*)^\top & -\nabla_y F_1(x^*,y^*)^\top \end{bmatrix} \begin{bmatrix} w \\ z \end{bmatrix} \in -N^M_\omega(x^*,y^*), \right\} \text{implies} \begin{cases} w = 0, \\ z = 0 \end{cases} \tag{2.11}$$
$$(w,z) \in N^M_{\operatorname{gph} Q}(F_2(x^*,y^*), -F_1(x^*,y^*))$$

holds true. Then there exists a pair $(\xi, \eta) \in \partial f(x^*, y^*)$, a pair $(\gamma, \delta) \in N^M_\omega(x^*, y^*)$, and a KKT pair $(w^*, z^*) \in N^M_{\operatorname{gph} Q}(F_2(x^*, y^*), -F_1(x^*, y^*))$ such that

$$\begin{aligned} 0 &= \xi + \nabla_x F_2(x^*,y^*)^\top w^* - \nabla_x F_1(x^*,y^*)^\top z^* + \gamma, \\ 0 &= \eta + \nabla_y F_2(x^*,y^*)^\top w^* - \nabla_y F_1(x^*,y^*)^\top z^* + \delta. \end{aligned}$$

where $N^M_{\operatorname{gph} Q}(F_2(x^*, y^*), -F_1(x^*, y^*))$ stands for Mordukhovich normal cone to the graph of Q at $(F_2(x^*, y^*), -F_1(x^*, y^*))$.

2.7 Trust Region Methods

Let us consider a class of algorithms for finding a local solution of the problem

$$\min_{x \in \mathbb{R}^n} \quad f(x) \tag{2.12a}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a real-valued twice-continuously differentiable function. Classical trust-region methods implement an iterative numerical procedure in which the objective function $f(x)$ is approximated in a suitable neighborhood of the current iterate, named

trust region, by a model which is easier to handle than $f(x)$.

By notating the sequence of iterates generated by the algorithm by $\{x_k\}$, at each iterate x_k, we first define a model $m_k(x)$ that approximates the objective function within a suitable neighborhood of x_k. For this type of smooth problem the quadratic model is widely used

$$m_k(x_k + s) = m_k(x_k) + \langle g_k, s\rangle + \frac{1}{2}\langle s, B_k s\rangle, \tag{2.13}$$

where $m_k(x_k) = f(x_k)$, $g_k = \nabla_x f(x_k)$ and B_k is a symmetric approximation to the hessian matrix $\nabla_{xx} f(x_k)$.

Said *trust region* is the set of all points

$$B_{\Delta_k} := \{x \in \mathbb{R}^n : \ \|x - x_k\|_k \leq \Delta_k\},$$

where Δ_k is called the *trust-region radius*, and $\|\cdot\|_k$ is an iteration-dependent norm. Given this model and trust region, we next seek a *trial step* s_k and a *trial point* $x_k + s_k$ with the aim of reducing the model while satisfying the bound $\|s_k\|_k \leq \Delta_k$. The step is then accepted if the quality of the decrease predicted *pred* by the model is "good" when compared with the decrease in the objective function *ared*. A pseudo-code describing the basic computation steps in a trust-region method is depicted in Algorithm 2.1.

Algorithm 2.1 Basic Trust Region Algorithm

1: Chose initial point x_0, initial trust region radius Δ_0 and $tol > 0$.
2: Choose $0 \leq \eta_1 \leq \eta_2 < 1$, $0 < \gamma_1 \leq 1 \leq \gamma_2$
3: **while** Stopping criteria not met **do**
4: Choose $\|\cdot\|_k$ and define the model m_k.
5: Compute a step s_k that "sufficiently" reduces the model m_k and such that $x_k + s_k \in B_{\Delta_k}$.
6: Compute $f(x_k + s_k)$ and define

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}.$$

If $\rho_k \geq \eta_1$, then define $x_{k+1} = x_k + s_k$; otherwise $x_{k+1} = x_k$.
7: Update the trust region radius

$$\Delta_{k+1} = \begin{cases} \gamma_2 \Delta_k, & \text{if } \rho_k \geq \eta_2, \\ \gamma_1 \Delta_k, & \text{if } \rho_k \leq \eta_1, \\ \Delta_k, & \text{else.} \end{cases}$$

8: **end while**

A crucial point in the algorithm presented is the determination of step 5 in algorithm 2.1. Since this is another optimization problem and possibly computationally

expensive to solve, a suitable approximation that provides descent guarantees is often preferred. Indeed, one of the most straightforward strategies for reducing the model within the trust region is to examine the model's behavior along the steepest descent direction $-g_k$ within the trust region. Then, we define the *Cauchy point* as follows

$$x_k^C = \underset{t\geq 0,\ x_k - tg_k \in B_{\Delta_k}}{\arg\min} m_k(x_k - tg_k).$$

Furthermore, when assuming a quadratic model, its minimizer has a closed form.

Now, when dealing with non-differentiable optimization problems, i.e., $f(x)$ is continuous but not necessarily differentiable, we cannot make use of the gradient in the trust-region model. Even though trust-region methods have been investigated for nonsmooth optimization of locally Lipschitz continuous functions in [1], they usually rely on the hypothesis that the cost function has to be *regular*.

Definition 2.15. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz continous at $x^* \in \mathbb{R}^n$. Then f is called* regular *at x^* if f is directionally differentiable at x^* and we have*

$$f^\circ(x^*; d) = f'(x^*; d),\ \forall d \in \mathbb{R}^n,$$

where $f^\circ(x^*; \cdot)$ is the generalized directional derivative *of f at x^*, defined as*

$$f^\circ(x; h) = \limsup_{y\to x,\ t\to 0} \frac{f(y+th) - f(y)}{t}.$$

Indeed, in the case of a locally Lipschitz continuous and regular function f, in [1] we see that the trust-region problem setting presented in (2.13) cannot be used due to the nonexistence of $\nabla f(x_k)$. A classical methodology for adapting the trust region for the nonsmooth setting, as detailed in the seminal work by [63], is to change the model function as follows

$$m_k(x_k + s) = m(x_k) + \phi(x_k, s) + \frac{1}{2}\langle s, B_k s\rangle,$$

where $\phi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a given iteration function. Here, the idea is to let $\phi(x, s)$ and B_k carry certain first-order and second-order information of f respectively, although the first and second-order information of f may not exist in general.

Assuming f and ϕ to be regular, and taking ϕ with the following properties

1. For all $x \in \mathbb{R}^n$, $\phi(x, 0) = 0$ and $\phi(x, \cdot)$ is lower semi-continuous.
2. For any convergent subsequence $\{x_k\}$, if $s_k \to 0$, then

$$f(x_k + s_k) - f(x_k) \leq \phi(x_k, s_k) + o(\|s_k\|),$$

it can be proven that the nonsmooth trust-region algorithm converges to a C-stationary point. Furthermore, assuming that $\partial f(x_k)$ is known, ϕ can be chosen as follows

$$\phi(x, s) = \max_{g \in \partial f(x)} \langle g, s \rangle.$$

Now, for a general nonsmooth problem, this kind of model function does not behave well in practice. In particular, there exist pathological examples where the sequence generated by the algorithm converges to a point that is not stationary in any sense, i.e., neither Clarke- nor Bouligand-stationary, see [18, Lemma 2.15]. The main reason for failure in this example is the lack of neighborhood information. Thus, the authors in [18] consider a generalization of the model function that incorporates information about the objective function in a neighborhood of the current iterate. An example for this generalization can be stated as follows

$$\phi(x, \Delta, s) := \max_{g \in \mathcal{U}(x,\Delta)} \langle g, s \rangle, \quad \text{with} \quad \mathcal{U}(x, \Delta) := \bigcup_{\xi \in B_\Delta(x)} \partial f(\xi). \tag{2.14}$$

Furthermore, given the complexity of evaluating this function, the authors propose a two-phase algorithm that switches the model used according to a threshold radius $\Delta_t > 0$. Using ϕ it is possible to define a stationarity measure

$$\Psi(x, \Delta) := -\min_{\|s\| \leq 1} \phi(x, \Delta, s) \geq 0.$$

By taking appropriate assumptions on the model function, the authors proved the convergence of the sequence of iterates using this algorithm to a C-stationary point as detailed in the following proposition.

Proposition 2.2. *Assuming that the matrices B_k in algorithm 2.2 satisfy*

$$\|B_k\| \leq C_B, \ \forall k \in \mathbb{N}$$

with a constant $C_B > 0$, and this algorithm does not terminate in finitely many steps. Let $\{x_k\}$ be the sequence of iterates generated by algorithm 2.2, then:

1. $0 \in \partial f(\bar{x})$.
2. *If $\{x_k\}$ admits an accumulation point, then the sequence of function values $\{f(x_k)\}$ converges to some $\bar{f} \in \mathbb{R}$.*
3. *Every accumulation point is C-stationary.*

The details of the model switching mechanism are depicted in algorithm 2.2.

Algorithm 2.2 Generic Nonsmooth Trust Region Algorithm

Chose initial point x_0, initial trust region radius Δ_0 and $tol > 0$.
Choose $0 \leq \eta_1 \leq \eta_2 < 1$, $0 < \gamma_1 \leq 1 \leq \gamma_2$
Choose $\|\cdot\|_k$, a subgradient $g_k \in \partial f(x_k)$ and a matrix $B_k \in \mathbb{R}^{n\times n}_{sym}$.
while Stopping criteria not met **do**
  if $\Delta_k \geq \Delta_t$ **then**
    Define the model

$$m_k(x_k + s) := m_k(x_k) + \langle g_k, s\rangle + \frac{1}{2}\langle s, B_k s\rangle$$

    Compute a step s_k that "sufficiently" reduces the model m_k and such that $x_k + s_k \in B_{\Delta_k}$.
    Compute $f(x_k + s_k)$ and define

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}.$$

  else
    Define the model

$$\tilde{m}_k(x_k + s) := m_k(x_k) + \phi(x_k, \Delta_k, s) + \frac{1}{2}\langle s, B_k s\rangle$$

    Compute a step s_k that "sufficiently" reduces the model $\tilde{m}_k$ and such that $x_k + s_k \in B_{\Delta_k}$.
    Compute $f(x_k + s_k)$ and define

$$\rho_k = \begin{cases} \frac{f(x_k)-f(x_k+s_k)}{\tilde{m}_k(x_k)-\tilde{m}_k(x_k+s_k)} & \Psi(x,\Delta) > \|g_k\|\Delta_k, \\ 0 & \Psi(x,\Delta) \leq \|g_k\|\Delta_k. \end{cases}.$$

  end if
  If $\rho_k \geq \eta_1$, then define $x_{k+1} = x_k + s_k$; otherwise $x_{k+1} = x_k$.
  Update the trust region radius

$$\Delta_{k+1} = \begin{cases} \gamma_2\Delta_k, & \text{if } \rho_k \geq \eta_2, \\ \gamma_1\Delta_k, & \text{if } \rho_k \leq \eta_1, \\ \Delta_k, & \text{else.} \end{cases}$$

end while

In the problems presented in this work, we will be dealing with positiveness constraints for the optimization problem, i.e., the optimization problem in this setting reads

$$\min_{x \in \mathbb{R}^n} \quad f(x) \tag{2.15a}$$

$$\text{s.t.} \quad x \geq 0. \tag{2.15b}$$

Problem (2.15) can be adapted to the trust-region methodology by changing the choice for the trust-region norm. In particular, for this application, we use the l_∞ ball for the local approximation of the objective function f. Moreover, we consider only steps that maintain the current iteration x_k positive. Consequently, the trust-region subproblem for an iteration x_k now reads

$$\begin{aligned} \min_{s \in \mathbb{R}^n} \quad & f(x_k) + \langle g, s \rangle + \frac{1}{2}\langle s, Bs \rangle \\ \text{s.t.} \quad & \|s\|_\infty \leq \Delta, \\ & x_k + s \geq 0. \end{aligned} \tag{2.16}$$

Indeed, this problem corresponds to a classical trust-region sub-problem with additional positivity constraints. Such constraints have been studied before in [84, 86]. The main idea is to reformulate the problem by taking advantage of the l_∞ norm used for the ball at the point x_k

$$\min_{s \in \mathbb{R}^n} \quad f(x_k) + \langle g, s \rangle + \frac{1}{2}\langle s, Bs \rangle \tag{2.17a}$$

$$\text{s.t.} \quad \max(-(x_k)_j, -\Delta) \leq s_j \leq \Delta, \ \forall j = 1, \ldots, n. \tag{2.17b}$$

Again, for performance purposes, it is desirable to solve this problem approximately in such a way that we can guarantee a descent in the cost function. With that goal in mind, we will make use of a dogleg strategy that takes into account a Newton step s_N and a Cauchy step s_C. In the context of this constrained problem, let us take $\tilde{B}_\Delta = B_\Delta \cap \mathbb{R}^n_+$ and distinguish the following three cases

1. $s_N \in \tilde{B}_\Delta$,
2. $s_C \in \tilde{B}_\Delta$ and $s_N \notin \tilde{B}_\Delta$,
3. $s_C \notin \tilde{B}_\Delta$ and $s_N \notin \tilde{B}_\Delta$.

For case 1 we take the Newton step; for case 2 a dogleg strategy for box constraints is used; for case 3 we make use of a scaled Cauchy direction as described in Algorithm 2.3.

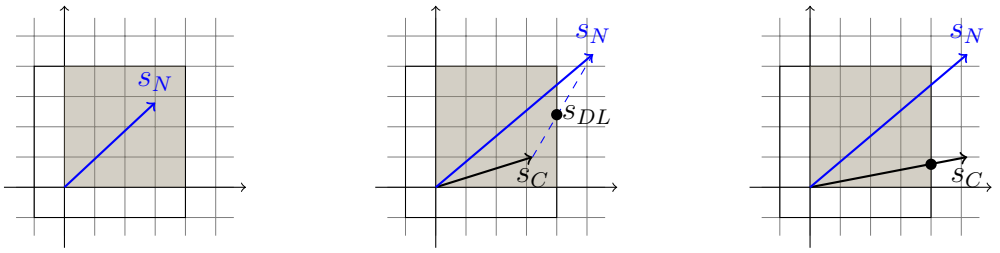


Figure 2.2: Two-dimensional example for the three different possible cases when approximating the trust-region subproblem using a dogleg strategy with l_∞ norm and positivity constraints.

Algorithm 2.3 Dogleg Step for Box Constraints

Calculate Newton's step by solving the linear system $B_k s_N = -g_k$.
if $s_N \in \tilde{B}_\Delta$ **then**
  return s_N
end if
Calculate $s_C = -\frac{\|g_k\|^2}{g_k^T * B_k * g_k} g_k$
if $s_C \in \tilde{B}_\Delta$ **then**
  Find the maximum t such that $s_C + t(s_N - s_C) \in \tilde{B}_\Delta$.
  return $s_{DL} = s_C + t(s_N - s_C)$
end if
Find the maximum t such that $t * s_C / \|s_C\|$ remains in $\tilde{B}_\Delta$.
return $t * \frac{s_C}{\|s_C\|}$

Chapter 3

Bilevel Parameter Learning

3.1 Bilevel Parameter Learning in Imaging

Computational Imaging methods aim to recover a good-quality image[1] from noisy, incomplete, or degraded images. In general, images degrade due to poor imaging conditions or problems in the storage device or the communication channel, to name a few. A frequentist model used to analyze this phenomenon reads as follows

$$f = A(u) + \zeta, \tag{3.1}$$

where $u \in \mathbb{R}^n$ is the original image, $f \in \mathbb{R}^m$ is the observed degraded image, $\zeta \in \mathbb{R}^m$ is the noise contained in the observed image, and $A : \mathbb{R}^n \to \mathbb{R}^m$ is a possible non-linear forward operator that models the acquisition process. In most imaging models, A is rank-deficient, leading to an ill-posed inverse problem. Therefore, non-uniqueness of the solution or instability of the direct inversion of such an operator motivates different solution techniques.

A classical way to solve such inverse problem is to make use of a variational "energy" formulation. Using this methodology we can state the solution of (3.1) as the solution of the following optimization problem

$$u^* := \arg\min_u \mathcal{E}(u; \theta, f), \tag{3.2}$$

where $u^* \in \mathbb{R}^n$ is the reconstructed image, $f \in \mathbb{R}^m$ is the degraded image and θ is a parameter in the cost function. Particularly, the choice of this parameter has a crucial impact on the solution. Indeed, Figure 3.1 show different quality reconstruction

[1] Throughout this work, we will consider images as $n_1 \times n_2$ grayscale pixel grids and, in particular, we will work with a *vectorized* version that considers images as vectors of length $n = n_1 n_2$ arranged according to a row-major ordering.

of a variational denoising model for different parameter choices. There have been numerous approaches for choosing θ, such as cross-validation [76], generalized cross-validation [35], the discrepancy principle [61] and Bayesian methods [72], among others.

This thesis will focus on *Bilevel Parameter Learning* techniques for finding optimal parameters that can be used for a specific image denoising task. This is a supervised learning methodology where we consider a training dataset of P pairs (u_k^{true}, f_k), for $k = 1, \ldots, P$, where each u_k^{true} corresponds to ground-truth data and f_k to the corresponding corrupted one. To obtain the optimal parameter θ, we consider the following class of *bilevel optimization* problems:

$$\min_{\theta} \quad \sum_{k=1}^{P} J(u_k, u_k^{\text{true}}) \tag{3.3a}$$

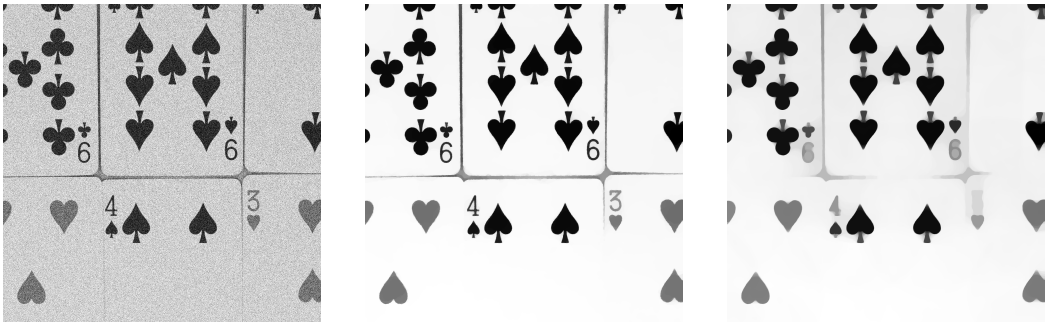
$$\text{s.t.} \quad u_k \in \arg\min_{u \in \mathbb{R}^n} \mathcal{E}(u; \theta, f_k), \tag{3.3b}$$

where the upper level problem handles the optimal parameter loss function J, while the lower-level problem corresponds to the image denoising model of interest.

In recent years, the combination of existing training sets with a bilevel optimization framework has been developed for variational image restoration, both from variational and nonsmooth analysis perspectives. Bilevel techniques for optimal parameter selection of variational models were proposed in the seminal works [25, 48]. The variational models considered in [25] were based on the total variation (TV) seminorm, and different noise statistics were taken into account. Thereafter, apart of total variation denoising problems, the bilevel learning framework has been used for higher-order total variation models [23, 24], blind image deconvolution [42], image segmentation [57, 64], mixed noise models [12], nonlocal models [19], learning the sample pattern for magnetic resonance imaging (MRI) [75] and kernel parameter estimation for support vector machines [46]. The bilevel methodology has also been extended in [13, 20, 37, 81] for learning different optimal scale-dependent total variation (TV) and generalized total variation (TGV) parameters.

Even though all previous approaches derived optimality conditions for the bilevel problem and proposed numerical methods for finding these parameters, they use a local regularization procedure to overcome the lack of differentiability in the lower level problem. This non-smoothness comes from using non-differentiable regularizers such as total-variation or generalized total variation in the denoising problem.

In [42], we find an approach that avoids using said regularization. In this case, the authors use a bilevel learning problem to determine optimal point spread functions in blind deconvolution. The strategy described consists of reformulating the lower level problem as a set-valued equation, leading to a reformulation of the bilevel problem as



Low Optimal High

Figure 3.1: TV reconstruction results for different parameter values.

a generalized problem with equilibrium constraints (GMPEC). Then, by making use of the Robinson strong regularity condition, it is possible to prove the existence of a Lipschitz continuous solution operator. Furthermore, this property allows the use of tools from variational geometry to propose a Mordukhovich (M-) and a Clarke (C-) optimality system.

Another approach to deal with non-differentiable lower-level problems is a novel method proposed by Ochs and coworkers [56], where instead of applying a smooth approximation of the lower level problem, they propose a method based on differentiating the iterations of a non-linear primal-dual algorithm. Moreover, an extension to this idea is presented in [57] for suitable non-linear proximal distance functions that lead to a differentiable algorithm while the minimization problem remains nonsmooth. Alternatively, the work by Ehrhardt and coworkers [27] presents a methodology for solving the lower level problem, which is a numerical challenge; by making use of inexact derivative-free optimization techniques, the methodology allows for an inexact solution to be used in the context of learning optimal parameters of variational imaging models. A similar approach based on randomized Itoh-Abe methods for general bilevel problems is presented in [65].

Furthermore, the authors in [38] propose a different technique for the analysis of the bilevel problem with a total generalized variation regularizer in the lower level problem. Instead of having the primal form of the lower level problem, the authors use its Fenchel predual version. This reformulation yields a bilevel problem that depends on the dual variables, yielding a more amenable structure for the constraints of the reformulated bilevel problem. A similar approach is found in [39, 40] for total variation based models. Even though this reformulation avoids the non-differentiability problems in the original version, the problem is not necessarily easier to solve, requiring a Moreau-Yosida regularization on this pre-dual problem.

3.2 Image Reconstruction Quality Metric

An essential component of the bilevel problem (3.3) is the loss function J, which models the quality of the reconstruction when compared to the original image provided in the dataset. One classic approach is to compute the difference between a ground truth image u^{true} and its reconstruction u using a Mean Squared Error (MSE) criteria

$$J(u, u^{\text{true}}) = MSE(u, u^{\text{true}}) := \frac{1}{2}\|u - u^{\text{true}}\|_2^2,$$

which is closely related to the Peak Signal-to-Noise (PSNR) ratio quality measure

$$PSNR(u, u^{\text{true}}) := 10 \log_{10}(255^2 / MSE(u, u^{\text{true}})).$$

Even though the imaging community uses this measure widely due to its low computational complexity, it depends strongly on the image intensity scaling. Furthermore, PSNR does not necessarily coincide with a human visual response to the image quality.

In [85] the authors exploit a known property of human visual perception. Under the assumption that human visual systems are highly adapted for extracting structural information from a scene, they propose a more reliable quality measure based on the degradation of structural information in a distorted image. This metric, known as Structural Similarity Index (SSIM), is calculated as follows

$$J(u, u^{\text{true}}) = SSIM(u, u^{\text{true}}) = l(u, u^{\text{true}}) c(u, u^{\text{true}}) s(u, u^{\text{true}}),$$

where

$$\begin{aligned} l(u, u^{\text{true}}) &= \frac{2\mu_u \mu_{u^{\text{true}}} + C_1}{\mu_u^2 + \mu_{u^{\text{true}}}^2 + C_1}, \\ c(u, u^{\text{true}}) &= \frac{2\sigma_u \sigma_{u^{\text{true}}} + C_2}{\sigma_u^2 + \sigma_{u^{\text{true}}}^2 + C_2}, \\ s(u, u^{\text{true}}) &= \frac{2\sigma_{u u^{\text{true}}} + C_3}{\sigma_u + \sigma_{u^{\text{true}}} + C_3}, \end{aligned}$$

and μ_u and σ_u correspond to the mean luminance and the standard deviation of the image u respectively. The use of this quality measure in the bilevel optimization context is, however, restrictive due to its non-smoothness and non-convexity.

The authors in [24], guided by the idea of preserving edge information on restored images, propose a novel image quality metric aimed at prioritizing jump preservation:

$$J(u, u^{\text{true}}) := \sum_{j=1}^{m} \|\mathbb{K}(u - u^{\text{true}})_j\|_\epsilon,$$

where $\mathbb{K}$ is a discretization of the gradient operator and $\|\cdot\|_\epsilon$ is a Huber regularization of the Euclidean norm. This metric is differentiable and convex, making it amenable when used within an optimization framework. Furthermore, when used as a cost function in bilevel learning, the reconstructed images attain a higher SSIM. Since SSIM better captures the visual quality of a reconstructed image than PSNR, it is recommended for this task as an alternative for using the non-convex SSIM metric directly, which leads to additional numerical challenges.

3.3 Lower Level Denoising Problem

Recalling the lower level problem in (3.3), when addressing the image denoising case, we can state it as the following optimization problem

$$u^* \in \arg\min_u \mathcal{E}(u; \lambda, \alpha) := \mathcal{F}(\lambda, u) + \mathcal{R}(\alpha, \mathbb{K}u), \tag{3.4}$$

where $u^* \in \mathbb{R}^n$ is the reconstructed image, $\mathbb{K}$ a bounded linear operator, $\mathcal{F}$ is the *data fidelity* and $\mathcal{R}$ is a *regularization* term. Furthermore, λ and α are parameters associated to the data fidelity term and the regularization term, respectively. The data fidelity term is usually modeled based on the statistical estimates or a noise model coming from the physics behind the acquisition of the image, while the regularization term promotes certain features which are known *a-priori* about the image. According to [71], a normally distributed noise model in a corrupted image f corresponds to the following data fidelity term

$$\mathcal{F}(\lambda, u; f) = \frac{\lambda}{2}\|u - f\|^2.$$

In the case of a poisson noise distribution, this term was studied in [49, 73] and it corresponds to $\mathcal{F}(\lambda, u; f) = \lambda \sum_{j=1}^n u_i - f_i \log(u_i)$. In [53] the author studied impulse noise degraded images and proposed the following non-smooth fidelity term $\mathcal{F}(\lambda, u; f) = \lambda\|u - f\|_1$. Figure 3.2 shows a comparison of the reconstruction obtained using a l_2 and a l_1 data fidelity term in an image corrupted by impulse noise.

Original Noisy l_2 Data Term l_1 Data Term

Figure 3.2: Reconstructed images using a Total Variation regularizer and l_2 and l_1 data terms obtained from a original image contaminated with impulse noise.

Regarding the regularization term, its choice is also critical for the quality of the reconstruction. To illustrate these phenomena, let us consider a classical Tikhonov regularizer [79] in the following problem

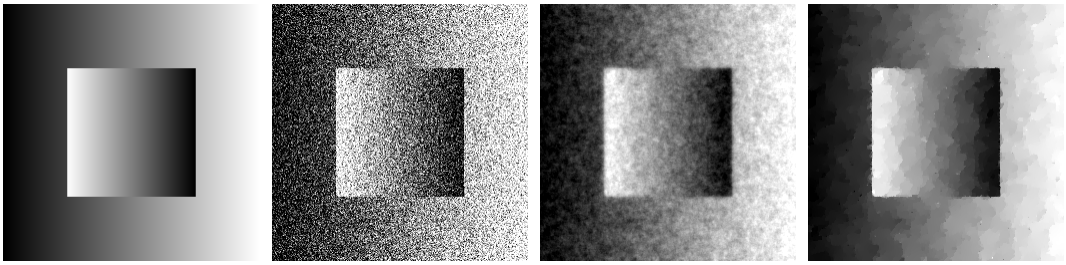
$$\mathcal{E}(\lambda, u) := \frac{\lambda}{2}\|u - f\|^2 + \frac{1}{2}\sum_{j=1}^{m} \|(\mathbb{K}u)_j\|^2, \tag{3.5}$$

where $\|\cdot\|$ is the Euclidean norm and $\mathbb{K} : \mathbb{R}^n \to \mathbb{R}^{m\times 2}$ is the discrete gradient operator with respect to directions in x and y, i.e., $\mathbb{K}u = (\mathbb{K}_x u, \mathbb{K}_y u)$, here $\mathbb{K}_x : \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbb{K}_y : \mathbb{R}^n \to \mathbb{R}^m$ correspond to the discrete partial derivative with respect to the horizontal and vertical direction, respectively. Moreover, the dimensions of the discrete partial derivative depend on the choice of the discretization at hand. For instance, when considering a centered finite differences discretization, the pixels at the image border cannot be calculated. If we avoid using a border condition, we must exclude such pixels from the calculation. It results in a smaller size grid size for the discretization.

As depicted in Figure 3.3, the solution obtained is not desirable, precisely since the regularizer involved has very strong isotropic smoothing properties which leads to a loss of edge information in the reconstructed image. To preserve the edge information as much as possible, Rudin, Osher and Fatemi [71] proposed the use of the *isotropic total variation (TV)* of the image as a regularization term, leading to its famous ROF image denoising model

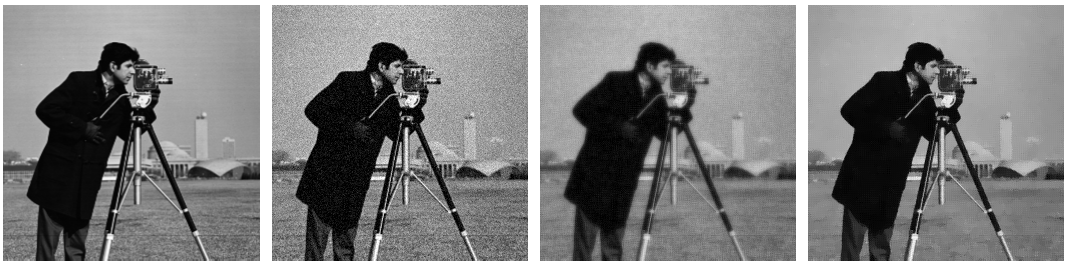
$$\mathcal{E}(\lambda, u) := \frac{\lambda}{2}\|u - f\|^2 + \sum_{j=1}^{m} \|(\mathbb{K}u)_j\|. \tag{3.6}$$

Additionally, the total variation regularizer promotes solutions close to a piecewise constant image, see Figure 3.3. Assuming that a crucial property in visual image quality assessment is the separation of objects in a scene; as a result, having an image reconstruction with sharp edges is highly desirable.



Original Noisy Tikhonov Reg. TV Reg.

Figure 3.4: Reconstructed images for gaussian denoising using a Tikhonov regularizer and a TV regularizer. Eventhough, the reconstruction in the case of the TV regulatizer presents sharp edges, a piecewise reconstruction promotes an artifact called the *staircasing* effect.



Original Noisy Tikhonov Reg. TV Reg.

Figure 3.3: Reconstructed images for gaussian denoising using a Tikhonov regularizer and a TV regularizer.

The main drawback of such a regularization procedure becomes apparent as soon as it is applied to images that do not only consist of constant intensity regions and jumps, but also possess more complicated structures, like smooth intensity variations or textures. This well-known artifact introduced by TV regularization in this case is called stair-casing [66], see Figure 3.4. Several high-order regularizers have been proposed to deal with this problem [9, 14]. Furthermore, let us note that, alternatively, problem (3.6) can be written as

$$\mathcal{E}(u;\alpha) := \frac{1}{2}\|u-f\|^2 + \alpha \sum_{j=1}^{m} \|(\mathbb{K}u)_j\|. \tag{3.7}$$

By considering a twice continuously differentiable data fidelity term and the TV regularizer, we will define the following family of denoising problems

$$\min_{u\in\mathbb{R}^n} \mathcal{E}(u;\lambda,\alpha) := \mathcal{F}(\lambda,u;f) + \alpha \sum_{j=1}^{m} \|(\mathbb{K}u)_j\|, \tag{3.8}$$

So far, the model parameters λ and α were assumed to be scalars, affecting all

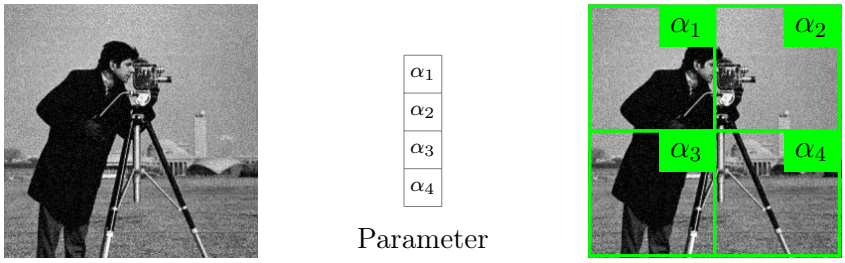


Figure 3.5: Mapping of a patch parameter.

pixels in the reconstructed image equally. This assumption does not necessarily hold in practice. For instance, the effect of perspective in natural images changes the size of objects according to their spatial relationship with respect to the viewer. In this scenario, a source of noise will affect each object depicted in the image differently.

Therefore, to address this issue it is necessary to consider a *scale-dependent* parameter, meaning, we now will consider $\lambda \in \mathbb{R}^n_+$ and $\alpha \in \mathbb{R}^m_+$. This extension to the model allows us to take one scalar value for each pixel of the image model and regularizer. Indeed, a general family of scale-dependent problems looks as follows

$$\min_{u \in \mathbb{R}^n} \mathcal{E}(u; \lambda, \alpha) := \mathcal{F}(\lambda, u; f) + \sum_{j=1}^{m} \alpha_j \|(\mathbb{K}u)_j\|. \tag{3.9}$$

In [26] the authors showed that learned scale-dependent parameters are known to overfit the training dataset, i.e., the quality of the reconstruction obtained using a denoising algorithm with the learned parameters drops when used to denoise an image not used in the training set (validation set). Therefore, there is a need for an intermediate approximation; consequently, a further generalization for *patch-dependent* parameters is necessary. Let us consider $\lambda \in \mathbb{R}^r$ and $\alpha \in \mathbb{R}^p$, with $r << n$ and $p << m$, then we define the patch operators $P : \mathbb{R}^p \mapsto \mathbb{R}^m$ and $Q : \mathbb{R}^r \mapsto \mathbb{R}^n$ that assigns each component of λ and α for a patch in the image, respectively. A graphical description of the mapping of these operators within the image can be found in Figure 3.5. Finally, using these patch operators, the corresponding variational denoising model reads

$$\mathcal{E}(u; \lambda, \alpha) := \mathcal{F}(Q(\lambda), u; f) + \sum_{j=1}^{m} P(\alpha)_j \|(\mathbb{K}u)_j\|. \tag{3.10}$$

Hereafter, without loss of generality, we will study the case of a scale-dependent parameter. However, we can obtain the scalar and the patch-dependent model as particular cases.

Now, regarding the bilevel problem (3.3), when the lower level problem has a closed-

form solution, one can replace this solution in the upper-level problem. In this case, we can reformulate the bilevel problem as a single-level problem, and one can use classical single-level optimization methods to minimize the upper-level loss. Even though this is not the case for the lower-level problems presented in this work, we can prove the existence of a unique minimizer, as well as a necessary optimality condition for the lower-level problem as detailed in Theorem 3.1.

Theorem 3.1. *Let $\mathcal{F}$ be strongly convex with respect to u with $\lambda > 0$, then the optimization problem* (3.9) *has a unique solution u^*. Moreover, a necessary and sufficient condition for the lower level problem is given by the following variational inequality of the second kind*

$$\langle \nabla_u \mathcal{F}(\lambda, u^*), v - u^* \rangle + \sum_{j=1}^{m} \alpha_j \|(\mathbb{K}v)_j\| - \sum_{j=1}^{m} \alpha_j \|(\mathbb{K}u^*)_j\| \geq 0, \ \forall v \in \mathbb{R}^n. \tag{3.11}$$

Proof. To obtain the existence of a unique minimizer, let us recall our assumption for $\mathcal{F}$ to be strongly convex. This property, along with the convexity of the total variation seminorm, yields a strongly convex lower-level optimization problem. Consequently, this problem has a unique minimizer. Now, to obtain the necessary and (due to convexity) sufficient condition, let us take the minimizer u^*

$$\mathcal{F}(\lambda, u^*) + \sum_{j=1}^{m} \alpha_j \|(\mathbb{K}u^*)_j\| \leq \mathcal{F}(\lambda, w) + \sum_{j=1}^{m} \alpha_j \|(\mathbb{K}w)_j\|, \ \forall w \in \mathbb{R}^n.$$

Taking $v \in \mathbb{R}^n$, $w = u^* + t(v - u^*)$ and $t \in \mathbb{R}$ sufficiently small, it yields

$$\begin{aligned} 0 &\leq \mathcal{F}(\lambda, u^* + t(v - u^*)) - \mathcal{F}(\lambda, u^*) + \sum_{j=1}^{m} \alpha_j \|(\mathbb{K}(u_j^* + t(v - u^*)))_j\| - \sum_{j=1}^{m} \alpha_j \|(\mathbb{K}u^*)_j\|, \\ &\leq \mathcal{F}(\lambda, u^* + t(v - u^*)) - \mathcal{F}(\lambda, u^*) + t \sum_{j=1}^{m} \alpha_j \|(\mathbb{K}v)_j\| - t \sum_{j=1}^{m} \alpha_j \|(\mathbb{K}u^*)_j\|, \end{aligned}$$

where we used the convexity of the total variation. Using to the differentiability of $\mathcal{F}$, we get the result by dividing both terms by t and taking the limit $t \to 0$. □

Using the Fenchel duality techniques described in Section 2.2.3, and by introducing a dual variable q we can rewrite the variational inequality (3.11) as follows

$$\nabla_u \mathcal{F}(\lambda, u) + \mathbb{K}^\top q = 0 \tag{3.12a}$$

$$\langle q_j, (\mathbb{K}u)_j \rangle - \alpha_j \|(\mathbb{K}u)_j\| = 0, \qquad \forall j = 1, \ldots, m \tag{3.12b}$$

$$\|q_j\| - \alpha_j \leq 0, \qquad \forall j = 1, \ldots, m. \tag{3.12c}$$

Since the lower-level problems associated with variational image denoising do not present a closed-form solution, we need to find a numerical approximation of its solution. For the separable structure of the problem, consisting of data and a regularization term, several authors have presented different iterative schemes for approximating a solution. It is worth mentioning that this manuscript will focus on models using the isotropic total variation regularizer described in (3.6).

Even though the total variation regularizer is convex, lower semicontinuous, and proper, we cannot use classical smooth optimization techniques due to its non-smoothness. Moreover, in this general model, the data fidelity term can also present non-smoothness depending on the assumptions on the noise models, i.e., impulse noise [53].

A popular numerical alternative for solving this problem is replacing the non-differentiable term with a sufficiently smooth function. As a consequence, fast second-order methods, i.e., methods where both gradient and hessian information are used to define a descent direction, may be devised for the solution of the regularized problems. Indeed, Newton and semi-smooth Newton methods, along with globalization strategies, have been used to solve image restoration models (see, e.g., [25, 41]).

A first approach for dealing with this non-smoothness is to use a direct approach based on sub-gradient methods. Although this approach appears to be the most natural, it comes with the drawback of a slow convergence rate typical for these methods [4, Theorem 8.13]. Now, when assuming a smooth data fidelity term $\mathcal{F}$ and the fact that in this work, the regularizer $\mathcal{R}$ is a simple convex lower semicontinuous function, *forward-backward* methods may be applied. Here, considering a dual variable $q \in \mathbb{R}^{m\times 2}$, the convex dual reformulation of the dual problem reads

$$\min_{q\in\mathbb{R}^{m\times 2}} \mathcal{F}^*(-\mathbb{K}^* q) + \mathcal{R}^*(q).$$

Here, at each iteration, a gradient descent step on $\mathcal{F}^*$ and a proximal step on $\mathcal{R}^*$ are performed. The resulting algorithm behaves robustly and gets faster as the smoothness properties of $\mathcal{F}$ improve [4, Theorem 10.21]. Moreover, accelerated versions of this scheme (like the FISTA algorithm [5]) became quite popular.

Alternatively, using the primal and dual variables in a saddle point formulation of the lower-level problem, the appearing structure may be numerically exploited, namely,

$$\min_{u\in\mathbb{R}^n} \sup_{q\in\mathbb{R}^{m\times 2}} \langle q, \mathbb{K}u\rangle_{\mathbb{R}^{m\times 2}} + \mathcal{F}(u) - \mathcal{R}^*(q).$$

Moreover, for the problem used in this work, the saddle point formulation reads

$$\min_u \max_q \langle q, \mathbb{K}u\rangle_{\mathbb{R}^{m\times 2}} + \mathcal{F}(\lambda, u) - \delta_{\{\|\cdot\|_{2,\infty}\leq\alpha\}}(q),$$

where, using the property that $\mathcal{R}$ is a norm, $\delta_{\{\|\cdot\|_{2,\infty}\leq\alpha\}}$ is the indicator function of the polar ball

$$\delta_{\{\|\cdot\|_{2,\infty}\leq\alpha\}}(q) = \begin{cases} 0 & \text{if } \|q_j\|_2 \leq \alpha_j \ \forall j, \\ +\infty & \text{else.} \end{cases}$$

The strategy considers an alternate update, where we perform a proximal descent on the primal variable u and a proximal ascent step in the dual variable q as follows

$$\begin{aligned} u_{k+1} &= \text{prox}_{\tau\mathcal{F}}(u_k - \tau\mathbb{K}^* q_k), \\ q_{k+1} &= \text{prox}_{\sigma\mathcal{R}^*}(q_k + \sigma\mathbb{K}u_{k+1}). \end{aligned}$$

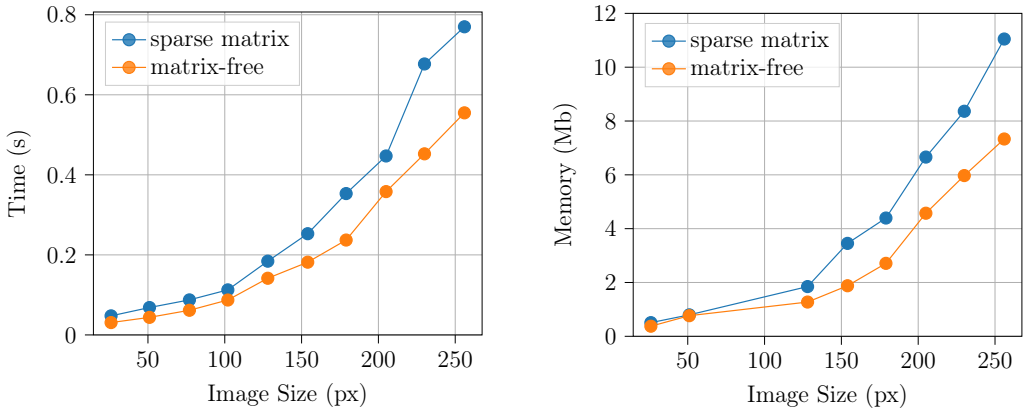
This procedure, called *primal-dual hybrid gradient PDHG*, can further be speed up by considering a relaxation step (see, e.g., [15]). These primal-dual update steps are well-suited for parallel computation, making these methods practical for high-resolution image denoising [82]. Related popular primal-dual methods are the well-known Douglas-Rachford and the Chambolle-Pock algorithms. An extension to non-linear operators $\mathbb{K}$ can be found in [80].

For the denoising problem we are using for the lower level problem, the term $\text{prox}_{\sigma\mathcal{R}^*}$ corresponds to the component-wise orthogonal projection onto l_2-balls with radius α. Therefore, with these components, we can derive a *primal-dual hybrid gradient modified* method (PDHGM) for solving (3.9), see Algorithm 3.1.

Algorithm 3.1 PDHGM for Variational Image Denoising

1: Set up initial value for primal and dual variables (u_0, q_0).
2: Initialize steps $\tau, \sigma \geq 0$.
3: **for** $k \geq 0$ **do**
4: $\quad u_{k+1} = \text{prox}_{\tau\mathcal{F}}(u_k - \tau\mathbb{K}^* q_k)$
5: $\quad q_{k+1} = \text{proj}_{\|\cdot\|\leq\alpha}(q_k + \sigma\mathbb{K}(2u_{k+1} - u_k))$
6: **end for**
7: **return** (u_k, q_k).

All the methods mentioned above rely on the type of implementation of the operator $\mathbb{K}$. The main numerical challenges are due to its size, since this operator scales its memory requirements with the size of the input image. Therefore, special attention must be provided when dealing with high-resolution images. The most naive implementation considers a matrix data structure for the linear operator $\mathbb{K}$; this choice results very restrictive in high-resolution scenarios. Consequently, other data structures are worth exploring. In particular, several numerical programming languages provide a *sparse* matrix representation that allows for efficient memory management when dealing with large matrices. Lately, more memory-efficient numerical implementation can be obtained using matrix-free operator implementations. These operators still repre-



(a) Time Benchmark **(b)** Memory Benchmark

Figure 3.6: Performance benchmark comparison between sparse operator implementation vs matrix free implementation on solving the ROF denoising model using the Chambolle-Pock algorithm.

sent a matrix and can be treated similarly but do not rely on the explicit creation of a dense (or sparse) matrix itself. Conversely, the forward and adjoint operators are represented by small pieces of code that mimic the effect of the matrix on a vector or another matrix.

In Figure 3.6 a comparison benchmark between the two different implementation paradigms is shown. Each paradigm's corresponding version of the gradient operator is used in a primal-dual algorithm that solves a gaussian image denoising model of increasing image size. Furthermore, each model was run 20 times and logged via the python `timeit.timeit` function for the time benchmarks and `memory_profiler` for the memory benchmark. Both tests were run on a MacBookPro 3,2 GHz Intel Core i5 with 16 GB 1600 Mhz DDR3 RAM. Moreover, NumPy and SciPy `scipy.sparse.csr_matrix` as well as PyLops operator.

3.4 Failure of Standard Constraint Qualification Conditions

A key goal in studying an optimization problem is the derivation of optimality conditions; since they allow a proper characterization of stationary points. Therefore, Lagrange multiplier's existence theorems are usually proved on the basis of so-called constraint qualification conditions [54]. Next, we will show that in the case of problem

$$\min_{(\lambda,\alpha)} \quad J(u(\lambda,\alpha);u^{\text{true}}) \tag{3.13a}$$

$$\text{s.t.} \quad u(\lambda,\alpha) \in \arg\min_{u\in\mathbb{R}^n} \mathcal{F}(\lambda,u) + \sum_{j=1}^{m} \alpha_j \|(\mathbb{K}u)_j\|, \tag{3.13b}$$

the situation is not standard at all and classical optimization theory typically fails.

Even though the primal-dual reformulation transforms problem (3.13) into a constrained non-linear optimization one, the difficulties related to the non-smoothness remain in the constraints. One may try to circumvent this by considering a smooth reformulation of the restrictions in order to use standard nonlinear programming techniques. One possibility consists in rewriting the constraints in (3.13) using the Fenchel primal-dual reformulation (3.12) in the equivalent differentiable form

$$\begin{aligned}
\min \quad & J(u,u^{\text{true}}) \\
\text{s.t.} \quad & \nabla_u \mathcal{F}(\lambda,u) + \mathbb{K}^\top q = 0, \\
& \langle q_j, (\mathbb{K}u)_j\rangle^2 - \alpha_j^2 \|(\mathbb{K}u)_j\|^2 = 0, \quad \forall i = 1,\dots,m, \\
& -\langle q_j, (\mathbb{K}u)_j\rangle \le 0, \quad \forall i = 1,\dots,m, \\
& \|q_j\|^2 - \alpha_j^2 \le 0, \quad \forall i = 1,\dots,m, \\
& -\lambda_j \le 0, \quad \forall i = 1,\dots,n,, \\
& -\alpha_j \le 0, \quad \forall i = 1,\dots,m,
\end{aligned}$$

and try to apply nonlinear programming results. Considering a toy example where $\mathcal{F}(\lambda,u) = \frac{\lambda}{2}\|u-f\|^2$, $u \in \mathbb{R}^2$, $\alpha \in \mathbb{R}$, $\lambda \in \mathbb{R}$, $q \in \mathbb{R}^2$ and $\mathbb{K} : \mathbb{R}^2 \to \mathbb{R}^2$ is defined by

$$\mathbb{K} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$$

we may indeed analyze case-by-case and verify whether a standard constraint qualification has a chance to hold. To verify either the *Linear Independence Constraint Qualification Condition (LICQ)* or the *Mangasarian-Fromowitz Constraint Qualification Condition (MFCQ)*, see Section 2.5, we have to analyze the rank of the matrix

formed by the gradients of the equality constraints, which is given by:

$$\nabla h(u,q,\lambda,\alpha) := \begin{pmatrix} \lambda & 0 & 2(u_1-u_2)(q_1^2-\alpha^2) & 0 \\ 0 & \lambda & -2(u_1-u_2)(q_1^2-\alpha^2) & 2u_2(q_2^2-\alpha^2) \\ 1 & -1 & 2q_1(u_1-u_2)^2 & 0 \\ 0 & 1 & 0 & 2q_2u_2^2 \\ u_1-f_1 & u_2-f_2 & 0 & 0 \\ 0 & 0 & 2\alpha(u_1-u_2)^2 & 2\alpha u_2^2 \end{pmatrix} \tag{3.14}$$

We then obtain the following cases:

$(\mathbb{K}\boldsymbol{u})_1 = \mathbf{0}, (\mathbb{K}\boldsymbol{u})_2 \neq \mathbf{0}$: In this case we know that $u_1 - u_2 = 0$. Consequently,

$$\nabla h_3(u,q,\lambda,\alpha) = (0,0,0,0,0,0)^\top.$$

Therefore, the columns of $\nabla h(u,q,\lambda,\alpha)$ are not linearly independent, and neither LICQ nor MFCQ hold.

$(\mathbb{K}\boldsymbol{u})_1 \neq \mathbf{0}, (\mathbb{K}\boldsymbol{u})_2 = \mathbf{0}$: Similar than the previous case, we reach to the same violation of linear independence, with $\nabla h_4(u,q,\lambda,\alpha)$ are equal to zero.

$(\mathbb{K}\boldsymbol{u})_1 = \mathbf{0}, (\mathbb{K}\boldsymbol{u})_2 = \mathbf{0}$: For this case, both columns $\nabla h_3(u,q,\lambda,\alpha)$ and $\nabla h_4(u,q,\lambda,\alpha)$ are equal to zero, failing to fulfill the linear independence requirement for LICQ or MFCQ.

$(\mathbb{K}\boldsymbol{u})_1 \neq \mathbf{0}, (\mathbb{K}\boldsymbol{u})_2 \neq \mathbf{0}$: In this case $|q_i| = \alpha,\ i = 1,2$ and we obtain

$$\nabla h_3(u,q,\lambda,\alpha) = (0,0,2q_1(u_1-u_2)^2,0,0,2\alpha(u_1-u_2)^2)^\top;$$

furthermore, $\nabla h_4(u,q,\lambda,\alpha) = (0,0,0,2q_2u_2^2,0,2\alpha u_2^2)^\top$. The linear independence may be satisfied in this case and existence of Lagrange multipliers may have a chance to be justified. This is, however, a case with scarce practical relevance. In the imaging setting, it would be related to completely smooth images (with no edges).

This toy example illustrates that even in a simplified case, the requirements for the existence of Lagrange multipliers are not met. Consequently, the main focus of this thesis will be to find suitable constraint qualification conditions that guarantee the existence of Lagrange multipliers for the Total Variation Bilevel learning problem (3.3), as well as define appropriate stationarity conditions and its numerical solution.

3.5 Smooth Bilevel Parameter Learning

To finish our review for bilevel parameter learning models, we will address the existence of optimal parameters and the optimality conditions for the regularized version of said problems. These results will be of particular importance when dealing with the numerics; namely, the two-phase trust-region algorithm detailed in Section 2.7, as it will make use of a regularized version of the bilevel learning problem in one of the phases. Therefore, let us consider a regularized version of the bilevel parameter learning model (3.13) formulated as follows

$$\min_{(\lambda,\alpha)} \quad J(u(\lambda,\alpha);u^{\text{true}}) \tag{3.15a}$$

$$\text{s.t.} \quad u(\lambda,\alpha) \in \arg\min_{u\in\mathbb{R}^n} \mathcal{F}(\lambda,u) + \sum_{j=1}^{m} \alpha_j h_\gamma((\mathbb{K}u)_j), \tag{3.15b}$$

where h_γ is a C^2-Huber regularization of the Euclidean norm

$$h_\gamma(z) = \begin{cases} -\frac{\|z\|^3}{3\gamma^2} + \frac{\|z\|^2}{\gamma} & \text{if } \|z\| \leq \gamma, \\ \|z\| - \frac{\gamma}{3} & \text{if } \|z\| > \gamma. \end{cases} \tag{3.16}$$

Assuming $\mathcal{F}$ to be strongly convex, along with the convexity of the regularized term h_γ, we have the strong convexity of the lower level problem; indeed, it implies the uniqueness of its solution. Furthermore, thanks to (3.16), we now have the differentiability of the lower level problem. It allows us to reformulate the bilevel problem as follows

$$\min_{(\lambda,\alpha)} \quad J(u(\lambda,\alpha);u^{\text{true}}) \tag{3.17a}$$

$$\text{s.t.} \quad \langle \nabla_u \mathcal{F}(\lambda,u^*), v\rangle + \sum_{j=1}^{m} \alpha_j \langle h'_\gamma((\mathbb{K}u^*)_j), (\mathbb{K}v)_j\rangle = 0, \forall\, v \in \mathbb{R}^n, \tag{3.17b}$$

where we replaced the lower level optimization problem by its necessary and due to convexity sufficient optimality condition. Now, in the reformulation presented in (3.17), we see that it corresponds to a single-level optimization problem with a variational equation constraint. It relates the bilevel optimization framework to mathematical problems with equilibrium constraints (MPEC).

Moreover, the optimality system for a local solution of the regularized problem

(3.17) can be obtained by introducing an adjoint state p, as follows

$$\begin{aligned} \nabla_u J(u^*; u^{\text{true}}) + \nabla_{uu}\mathcal{F}(\lambda^*, u^*)^\top p + \mathbb{K}^\top \beta = 0, \\ \beta_j - \alpha_j^* h_\gamma''((\mathbb{K}u^*)_j)^\top (\mathbb{K}p)_j = 0, \qquad \forall j = 1, \dots, m, \end{aligned}$$

Arguing that this adjoint equation admits a unique solution, see [25, 48], there exists a Lagrange multiplier $p \in \mathbb{R}^n$ such that an optimal solution $(\lambda^*, \alpha^*, u^*)$ satisfies the following KKT optimality system

$$\nabla_u \mathcal{F}(\lambda^*, u^*) + \mathbb{K}^\top \mu = 0, \tag{3.18a}$$

$$\mu_j - \alpha_j^* h_\gamma'((\mathbb{K}u^*)_j) = 0, \qquad \forall j = 1, \dots, n, \tag{3.18b}$$

$$\nabla_u J(u^*; u^{\text{true}}) + \nabla_{uu}\mathcal{F}(\lambda^*, u^*)^\top p + \mathbb{K}^\top \beta = 0, \tag{3.18c}$$

$$\beta_j - \alpha_j^* h_\gamma''((\mathbb{K}u^*)_j)^\top (\mathbb{K}p)_j = 0, \qquad \forall j = 1, \dots, m, \tag{3.18d}$$

$$\nabla_\alpha J(\lambda^*, \alpha^*; u^*) + \zeta = 0, \tag{3.18e}$$

$$\zeta_j - h_\gamma'((\mathbb{K}u^*)_j)^\top (\mathbb{K}p)_j = 0, \qquad \forall j = 1, \dots, m, \tag{3.18f}$$

$$\nabla_\lambda J(\lambda^*, \alpha^*; u^*) + \nabla_{u\lambda}\mathcal{F}(\lambda^*, u^*)^\top p = 0. \tag{3.18g}$$

As a final note, in [25, 48], the authors employed this technique and, thanks to a limiting procedure $\gamma \to \infty$ in (3.18), they managed to obtain a stationarity system for the original bilevel problem (3.3). Moreover, it could be seen that the stationarity system obtained using this procedure corresponds to a C-stationary point.

The rest of this work will split the analysis of the nonsmooth bilevel parameter learning problem for the two different parameters presented λ and α separately. In Chapter 4, we will focus on learning the parameter λ in (3.13), by assuming $\alpha = 1$. While Chapter 5 will study the case for learning the regularization parameter α and assuming $\lambda = 1$.

Chapter 4

Optimal Learning of the Data Fidelity Weight

In this section, we will find optimal parameters for the lower level problem described in (3.9), where only the parameter affecting the data fidelity term, $\lambda \in \mathbb{R}^n_+$, is considered. Indeed, we will make use of a bilevel parameter learning strategy by making use of a training dataset of P pairs (u_k^{true}, f_k), for $k = 1, \dots, P$, where each u_k^{true} corresponds to ground-truth data and f_k to the corresponding corrupted one. The optimization problem now reads

$$\min_{\lambda \in \mathbb{R}^n_+} \quad \sum_{k=1}^{P} J(u_k(\lambda), u_k^{\text{true}}) \tag{4.1a}$$

$$\text{s.t.} \qquad u_k(\lambda) = \arg\min_{u \in \mathbb{R}^n} \left\{ \mathcal{F}(\lambda, u; f_k) + \sum_{j=1}^{m} \|(\mathbb{K}u)_j\| \right\} \tag{4.1b}$$

where $\mathcal{F} : \mathbb{R}^n_+ \times \mathbb{R}^n \to \mathbb{R}$ is a **strongly convex** function with respect to u and **linear** with respect to λ. As a particular case of this formulation, we have the *spatially dependent* l_2 data fidelity term

$$\mathcal{F}(\lambda, u; f_k) := \frac{1}{2} \sum_{j=1}^{n} \lambda_j (u_j - (f_k)_j)^2.$$

Without loss of generality, we will analyze the case for a single training pair, and for readability purposes, we will omit the dependence of f_k, the data fidelity term. Additionally, we will assume the existence of an optimal parameter $\lambda^* \neq 0$[1]. This assumption will ensure the contribution of the data fidelity term to the final solution, which otherwise would only be governed by the total variation regularizer, leading to

[1] Alternatively, we may set the feasible set for λ as $\mathbb{R}^n_\epsilon := [\epsilon, \infty)^n$ with $\epsilon > 0$ sufficiently small.

a constant image regardless of the noisy input information.

Replacing the lower level optimization problem in (4.1) by its necessary and sufficient condition derived in Theorem 3.1. It leads to the following optimization problem with variational inequality constraints

$$\min_{\lambda \in \mathbb{R}^n_+} \quad J(u(\lambda), u^{\text{true}}) \tag{4.2a}$$

$$\text{s.t.} \quad \langle \nabla_u \mathcal{F}(\lambda, u^*), v - u^* \rangle + \sum_{j=1}^{m} \|(\mathbb{K}v)_j\| - \sum_{j=1}^{m} \|(\mathbb{K}u^*)_j\| \geq 0, \ \forall v \in \mathbb{R}^n \tag{4.2b}$$

Likewise, we can rewrite the primal-dual formulation for the lower level problem as a particular case of (3.12) with $\alpha_j = 1$, which reads

$$\nabla_u \mathcal{F}(\lambda, u) + \mathbb{K}^\top q = 0 \tag{4.3a}$$

$$\langle q_j, (\mathbb{K}u)_j \rangle - \|(\mathbb{K}u)_j\| = 0, \qquad \forall j = 1, \ldots, m \tag{4.3b}$$

$$\|q_j\| - 1 \leq 0, \qquad \forall j = 1, \ldots, m. \tag{4.3c}$$

4.1 Mordukhovich Stationarity

This section will address the primal-dual stationarity conditions for the bilevel problem (4.1). Motivated by the constraint qualification condition presented in Section 2.5, we can reformulate the lower-level optimization problem in (4.1b) as a generalized equation. Indeed, by introducing a dual variable $q \in \mathbb{R}^{m \times 2}$ where $q_j \in \partial(\|(\mathbb{K}u)_j\|)$ we may write the lower level problem equivalently as follows

$$0 \in \nabla_u \mathcal{F}(\lambda, u) + Q(u), \tag{4.4}$$

where $Q : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is the set-valued operator associated with the Euclidean norm

$$Q(u) := \left\{ \mathbb{K}^\top q : q \in \mathbb{R}^{m \times 2}, \begin{cases} q_j = \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, & \text{if } \|(\mathbb{K}u)_j\| \neq 0, \\ \|q_j\| \leq 1, & \text{if } \|(\mathbb{K}u)_j\| = 0. \end{cases} \right\} \tag{4.5}$$

The characterization (4.5) is obtained by first considering the case $\|(\mathbb{K}u)_j\| \neq 0$, where, in order to fulfill (4.3b), the relation $q_j = (\mathbb{K}u)_j / \|(\mathbb{K}u)_j\|$ must hold. Otherwise, if $\|(\mathbb{K}u)_j\| = 0$, the inequality (4.3c) holds. Equivalently, by making use of the definition

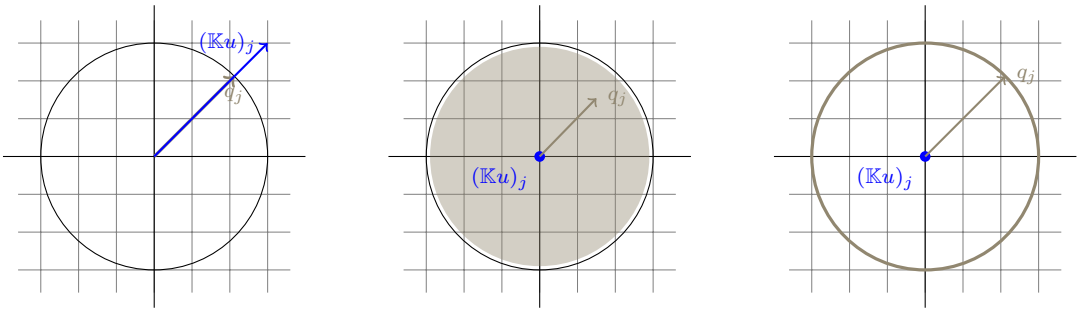
of the graph of the multifunction Q, we may rewrite (4.5) as

$$\nabla_u \mathcal{F}(\lambda, u) + \mathbb{K}^\top q = 0, \tag{4.6a}$$
$$(u, \mathbb{K}^\top q) \in \operatorname{gph} Q, \tag{4.6b}$$
$$(\lambda, u) \in \mathbb{R}^n_+ \times \mathbb{R}^n, \tag{4.6c}$$

where $\operatorname{gph} Q := \{(u, \mathbb{K}^\top q) \in \mathbb{R}^n \times \mathbb{R}^n : \mathbb{K}^\top q \in Q(u)\}$. Using this reformulation of the constraints of the bilevel problem as a generalized problem with equilibrium constraints (GMPEC), we will address the existence of Lagrange multipliers and a corresponding stationarity system. In Theorem 2.9, a constraint qualification condition for GMPECs that guarantees the existence of the Lagrange multipliers is presented; particularly, this condition requires the graph of the set-valued map to be closed. Indeed, in our case, the multifunction Q corresponds to the convex subdifferential of the total variation seminorm; therefore, the mapping $u \mapsto Q(u)$ is closed, as well as its graph [69, Theorem 24.4].

Now, the constraint qualification condition relies on fundamental definitions from Mordukhovich's generalized calculus; in particular, the Mordukhovich normal cone to the graph of the multifunction Q. We briefly review these concepts in Section 2.4. Moreover, using the structure of the set-valued operator Q presented in (4.5), let us introduce the following notation for the inactive, strongly active, and biactive sets, respectively:

$$\begin{aligned}
\mathcal{I}(u) &:= \{j \in \{1, \dots, n\} : (\mathbb{K}u)_j \neq 0\}, \\
\mathcal{A}_s(u) &:= \{j \in \{1, \dots, n\} : \|q_j\| < 1\}, \\
\mathcal{B}(u) &:= \{j \in \{1, \dots, n\} : \|q_j\| = 1,\ (\mathbb{K}u)_j = 0\}.
\end{aligned}$$



$j \in \mathcal{I}(u)$ $\quad$ $j \in \mathcal{A}_s(u)$ $\quad$ $j \in \mathcal{B}(u)$

Figure 4.1: Geometric interpretation of the primal-dual system for the different index sets.

For the sake of readability, we will omit the arguments in the set notation whenever

they can be inferred from the context.

We will start our analysis of the constraint qualification condition by providing a precise characterization of the Bouligand tangent cone, the Fréchet normal cone, and the Mordukhovich normal cone to the graph of the multifunction Q in the following lemmata.

Lemma 4.1. *The Bouligand tangent cone to the graph of Q, described in (4.5), is given by*

$$T_{\operatorname{gph} Q}(u, \mathbb{K}^\top q) = \left\{ (\delta_u, \mathbb{K}^\top \delta_q) : \begin{cases} (\delta_q)_j - T_j(\mathbb{K}\delta_u)_j = 0, & \text{if } j \in \mathcal{I}, \\ (\mathbb{K}\delta_u)_j = 0, & \text{if } j \in \mathcal{A}_s, \\ \left.\begin{array}{l} (\mathbb{K}\delta_u)_j = 0, \langle (\delta_q)_j, q_j \rangle \leq 0 \vee \\ (\mathbb{K}\delta_u)_j = \tilde{c} q_j (\tilde{c} \geq 0), \langle (\delta_q)_j, q_j \rangle = 0 \end{array}\right\} & \text{if } j \in \mathcal{B}, \end{cases} \right\} \tag{4.7}$$

where

$$T_j(\mathbb{K}v)_j = \frac{(\mathbb{K}v)_j}{\|(\mathbb{K}u)_j\|} - \frac{(\mathbb{K}u)_j (\mathbb{K}u)_j^\top (\mathbb{K}v)_j}{\|(\mathbb{K}u)_j\|^3}, \text{ for } v \in \mathbb{R}^n.$$

Proof. Using the definition of the tangent cone,

$$T_{\operatorname{gph} Q}(u, \mathbb{K}^\top q) = \{(\delta_u, \mathbb{K}^\top \delta_q) \in \mathbb{R}^n \times \mathbb{R}^n : \exists t_k \to 0, \exists \{(u_k, \mathbb{K}^\top q_k)\} \subset \operatorname{gph} Q,$$
$$\text{s.t.} \begin{cases} (\delta_u, \mathbb{K}^\top \delta_q) = \lim_{k\to\infty} t_k^{-1}((u_k, \mathbb{K}^\top q_k) - (u, \mathbb{K}^\top q)) & \text{if } (u, \mathbb{K}^\top q) \in \operatorname{gph} Q \\ \emptyset & \text{if } (u, \mathbb{K}^\top q) \notin \operatorname{gph} Q. \end{cases} \Bigg\} \tag{4.8}$$

Let us note that in this definition, we take sequences of elements in $\operatorname{gph} Q$ which, due to its closedness, have a limit that also belongs to the graph. Owing in addition to the surjectivity of the discrete partial derivative matrices, the limiting elements have the form $(\delta_u, \mathbb{K}^\top \delta_q)$. Taking a $((\delta_u, \mathbb{K}^\top \delta_q)) \in T_{\operatorname{gph} Q}((u, \mathbb{K}^\top q))$, then by the definition (4.8), we know there exists a sequence $\{(u_k, \mathbb{K}^\top q_k)\} \subset \operatorname{gph} Q$ converging to $((u, \mathbb{K}^\top q)) \in \operatorname{gph} Q$ and a sequence $t_k \to 0$. Moreover, for a particular k we know that $(u_k, \mathbb{K}^\top q_k) \in \operatorname{gph} Q$ if and only if $(q_k)_j \in \partial(\|(\mathbb{K}u_k)_j\|)$ for all $j = 1, \ldots, m$. This remark allows us to split the analysis into different cases according to the definition of the multifunction Q, where we can characterize each component as

$$(\mathbb{K}\delta u)_j = \lim_{k\to\infty} \frac{(\mathbb{K}u_k)_j - (\mathbb{K}u)_j}{t_k}, \quad (\delta_q)_j = \lim_{k\to\infty} \frac{(q_k)_j - q_j}{t_k}.$$

Case 1: $j \in \mathcal{I}(u)$. When approximating an inactive component $((\mathbb{K}u)_j, q_j)$, we can only consider approximating sequences in the inactive set. Indeed, for sequences

in this index set we know $(q_k)_j = (\mathbb{K}u_k)_j / \|(\mathbb{K}u_k)_j\|$, therefore

$$\frac{(q_k)_j - q_j}{t_k} = \frac{1}{t_k}\left(\frac{(\mathbb{K}u_k)_j}{\|(\mathbb{K}u_k)_j\|} - \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}\right).$$

Then, taking the limit as $t_k \to 0$ and considering that $v/\|v\|$ is differentiable in this index set, we get

$$(\delta_q)_j = \frac{(\mathbb{K}\delta_u)_j}{\|(\mathbb{K}u)_j\|} - \frac{(\mathbb{K}u)_j (\mathbb{K}u)_j^\top (\mathbb{K}\delta_u)_j}{\|(\mathbb{K}u)_j\|^3}$$

Case 2: $j \in \mathcal{A}_s(u)$. An entry in this index set can only be approximated by sequences in the same active set. Therefore $(\mathbb{K}u_k)_j = 0$ and $\|(q_k)_j\| < 1$. Consequently, the following limit holds true

$$(\mathbb{K}\delta u)_j = \lim_{t_k \to 0} \frac{(\mathbb{K}u_k)_j - (\mathbb{K}u)_j}{t_k} = 0.$$

For the dual variable we can approximate q_j through any sequence in int $\mathbb{B}(0,1)$; therefore, in the limit $k \to \infty$ we can reconstruct $(\delta_q)_j \in \mathbb{R}^2$.

Case 3: $j \in \mathcal{B}(u)$. This case considers the approximation of a biactive component, which can be approximated using three possible sequences: inactive, active, and biactive.

When taking an approximation sequence belonging to the *inactive* set, we know $(\mathbb{K}u_k)_j \neq 0$, $\|(q_k)_j\| = 1$. Then the sequence of dual variables has the following form

$$(q_k)_j = \frac{(\mathbb{K}u_k)_j}{\|(\mathbb{K}u_k)_j\|}.$$

Now, to find a characterization for the tangent direction in this component, let us consider the following product

$$\langle (q_k)_j, (\mathbb{K}u_k)_j \rangle = \left\langle \frac{(\mathbb{K}u_k)_j}{\|(\mathbb{K}u_k)_j\|}, (\mathbb{K}u_k)_j \right\rangle = \|(\mathbb{K}u_k)_j\|,$$

dividing by t_k in both sides and taking the limit as $k \to \infty$ it yields $\langle q_j, (\mathbb{K}\delta_u)_j \rangle = \|(\mathbb{K}\delta_u)_j\|$, which implies that both vectors q_j and $(\mathbb{K}\delta_u)_j$ are collinear, i.e., c

$$(\mathbb{K}\delta u)_j = \tilde{c} q_j \text{ for some } \tilde{c} \geq 0,$$

Using that $\|(q_k)_j\| = \|q_j\| = 1$, the following relation holds

$$\begin{aligned}\left\langle \frac{(q_k)_j - q_j}{t_k}, q_j \right\rangle &= \frac{1}{t_k}(\langle (q_k)_j, q_j\rangle - \langle q_j, q_j\rangle),\\ &= \frac{1}{t_k}(\langle (q_k)_j, q_j\rangle - \langle (q_k)_j, (q_k)_j\rangle + \langle (q_k)_j, (q_k)_j\rangle - \langle q_j, q_j\rangle),\\ &= -\left\langle \frac{(q_k)_j - q_j}{t_k}, (q_k)_j \right\rangle.\end{aligned}$$

Rearranging the terms in the last equation we get

$$\left\langle \frac{(q_k)_j - q_j}{t_k}, q_j \right\rangle + \left\langle \frac{(q_k)_j - q_j}{t_k}, (q_k)_j \right\rangle = 0.$$

Taking the limit as $k \to \infty$, we get that $\langle (\delta_q)_j, q_j\rangle = 0$, finishing this part of the proof.

Now, when taking an approximation through a sequence of *biactive* points, it holds $(\mathbb{K}u_k)_j = 0$ and $\|(q_k)_j\| = 1$. It implies that it must hold

$$(\mathbb{K}\delta u)_j = \lim_{t_k \to 0} \frac{(\mathbb{K}u_k)_j - (\mathbb{K}u)_j}{t_k} = 0.$$

Furthermore, by using the Cauchy-Schwarz inequality, we can upper bound the following product

$$\begin{aligned}\left\langle \frac{(q_k)_j - q_j}{t_k}, q_j \right\rangle &= \frac{1}{t_k}(\langle (q_k)_j, q_j\rangle - 1),\\ &\leq \frac{1}{t_k}(\|(q_k)_j\|\|q_j\| - 1),\\ &= 0,\end{aligned}$$

where we used the property that sequences in the biactive set must have $\|(q_k)_j\| = 1$. Consequently, taking the limit as $k \to \infty$, it holds $\langle (\delta_q)_j, q_j\rangle \leq 0$.

The last possible approximation of a biactive component can be made using a sequence that belongs to the *active* set. In this index set it holds $(\mathbb{K}u_k)_j = 0$ and $\|(q_k)_j\| < 1$. Then we have $(\mathbb{K}\delta u)_j = 0$ and the product for the dual variable reads

$$\begin{aligned}\left\langle \frac{(q_k)_j - q_j}{t_k}, q_j \right\rangle &= \frac{1}{t_k}(\langle (q_k)_j, q_j\rangle - 1),\\ &\leq \frac{1}{t_k}(\underbrace{\|(q_k)_j\|}_{<1}\|q_j\| - 1) < 0,\end{aligned}$$

taking the limit as $k \to \infty$ we have that $\langle (\delta_q)_j, q_j \rangle \leq 0$.

Let us name $\mathcal{M}(u, \mathbb{K}^\top q)$ the right-hand side of (4.7). Using this notation, so far, we have proven that $T_{\operatorname{gph} Q}(u, \mathbb{K}^\top q) \subseteq \mathcal{M}(u, \mathbb{K}^\top q)$. To prove the reverse inclusion let us take a $(\delta_u, \mathbb{K}^\top \delta_q) \in \mathcal{M}(u, \mathbb{K}^\top q)$, thanks to the result in (2.5), we know that a pair $(\delta_u, \mathbb{K}^\top \delta_q)$ is tangent to $\operatorname{gph} Q$ at $(u, \mathbb{K}^\top q)$ if

$$\lim_{t \to 0} \frac{\operatorname{dist}((u + t\delta_u, \mathbb{K}^\top q + t\mathbb{K}^\top \delta_q), \operatorname{gph} Q)}{t} = 0,$$

where $\operatorname{dist}(v, S)$ stands for the distance function of a vector v to the set S, presented previously in definition 2.7. We will prove in this section that $(\delta_u, \mathbb{K}^\top \delta_q) \in T_{\operatorname{gph} Q}(u, \mathbb{K}^\top q)$.

Since the elements in $\mathcal{M}(u, \mathbb{K}^\top q)$ are characterized by index set, let us consider each case separately. The $\operatorname{gph} Q$ is a smooth manifold for the inactive components, and the tangent elements are fully characterized by its derivative [68, Example 6.8]. Consequently, the elements defined in this index set are also contained in $T_{\operatorname{gph} Q}(u, \mathbb{K}^\top q)$. Likewise, the strongly active components lie in the interior of $\operatorname{gph} Q$, which by definition of tangency, coincides with the definition provided in $\mathcal{M}$.

Now, we will verify the biactive components. Let us recall that in this index set, the $\operatorname{gph} Q$ has the form $\operatorname{gph} Q = \{((\mathbb{K}u)_j, q_j) : (\mathbb{K}u)_j = 0,\ \|q_j\| = 1\}$. Since components in this index set can have two possible characterizations, we will analyze each individually. The first case corresponds to pairs $((\mathbb{K}\delta_u)_j, (\delta_q)_j)$ such that $(\mathbb{K}\delta_u)_j = 0$ and $\langle (\delta_q)_j, q_j \rangle \leq 0$. Taking $t > 0$, and the pair $((\mathbb{K}u)_j + t(\mathbb{K}\delta_u)_j, q_j + t(\delta_q)_j)$, its distance to the $\operatorname{gph} Q$ is given by

$$\inf_{((\mathbb{K}x)_j, y_j) \in \operatorname{gph} Q} \|((\mathbb{K}u)_j + t(\mathbb{K}\delta_u)_j, q_j + t(\delta_q)_j) - ((\mathbb{K}x)_j, y_j)\|.$$

Furthermore, considering that in this index set both $(\mathbb{K}u)_j = (\mathbb{K}x)_j = 0$, this problem now reads

$$\inf_{\|y_j\| = 1} \|q_j + t(\delta_q)_j - y_j\|,$$

and recalling that the solution y_j to this problem is the projection onto the l_2-ball, we have the following assertion

$$\inf_{\|y_j\| = 1} \|q_j + t(\delta_q)_j - y_j\| = \left\| q_j + t(\delta_q)_j - \frac{q_j + t(\delta_q)_j}{\|q_j + t(\delta_q)_j\|} \right\|. \tag{4.9}$$

Now, let us consider the following bound

$$\begin{aligned}\|q_j + t(\delta_q)_j\|^2 &= \|q_j\| + 2t\langle(\delta_q)_j, q_j\rangle + t^2\|(\delta_q)_j\|^2,\\ &\leq \|q_j\|^2 + t^2\|(\delta_q)_j\|^2,\\ &= 1 + t^2\|(\delta_q)_j\|^2,\end{aligned} \tag{4.10}$$

where we used the property $\langle(\delta_q)_j, q_j\rangle \leq 0$ in this index set. Furthermore, by squaring the norm in the right-hand side of (4.9), we get

$$\begin{aligned}\left\|q_j + t(\delta_q)_j - \frac{q_j + t(\delta_q)_j}{\|q_j + t(\delta_q)_j\|}\right\|^2 &= \|q_j + t(\delta_q)_j\|^2 - 2\|q_j + t(\delta_q)_j\| + 1,\\ &= (\|q_j + t(\delta_q)_j\| - 1)^2.\end{aligned} \tag{4.11}$$

Now, applying (4.10) and (4.11) in (4.9), dividing it by t, it yields

$$\inf_{\|y_j\|=1} \frac{\|q_j + t(\delta_q)_j - y_j\|}{t} = \frac{|\|q_j + t(\delta_q)_j\| - 1|}{t} \leq \frac{\sqrt{1 + t^2\|(\delta_q)_j\|^2} - 1}{t}.$$

Finally, taking the limit as $t \to 0$ we get $\lim_{t\to 0} \operatorname{dist}(q_j + t(\delta_q)_j, \operatorname{gph} Q)/t = 0$. Implying that $(\delta_q)_j$ is also a tangent vector.

The second characterization of vectors in the biactive set corresponds to the pairs $((\mathbb{K}\delta_u)_j, (\delta_q)_j)$ such that $(\mathbb{K}\delta_u)_j = \tilde{c}q_j$ with $\tilde{c} \geq 0$ or equivalently, due to colinearity, $\langle(\mathbb{K}\delta_u)_j, q_j\rangle = \|(\mathbb{K}\delta_u)_j\|$, and $\langle(\delta_q)_j, q_j\rangle = 0$. In this section we will show that for any $t > 0$, the pair $((\mathbb{K}u)_j + t(\mathbb{K}\delta_u)_j, q_j + t(\delta_q)_j) \in \operatorname{gph} Q$. With this goal in mind, let us consider the following product

$$\begin{aligned}\langle q_j + t(\delta_q)_j, (\mathbb{K}u)_j + t(\mathbb{K}\delta_u)_j\rangle &= \underbrace{\langle q_j + t(\delta_q)_j, (\mathbb{K}u)_j\rangle}_{=0} + t\langle q_j, (\mathbb{K}\delta_u)_j\rangle + t^2\langle(\delta_q)_j, (\mathbb{K}\delta_u)_j\rangle,\\ &= t\langle q_j, (\mathbb{K}\delta_u)_j\rangle + t^2\langle(\delta_q)_j, (\mathbb{K}\delta_u)_j\rangle,\end{aligned} \tag{4.12}$$

where we used the property of $(\mathbb{K}u)_j = 0$ for components in the biactive set. Now, recalling that $\langle(\mathbb{K}\delta_u)_j, q_j\rangle = \|(\mathbb{K}\delta_u)_j\|$, and using this property in (4.12), we get

$$\begin{aligned}\langle q_j + t(\delta_q)_j + t(\delta_q)_j, (\mathbb{K}u)_j + t(\mathbb{K}\delta_u)_j\rangle &= t\|(\mathbb{K}\delta_u)_j\| + t^2\langle(\delta_q)_j, \tilde{c}q_j\rangle,\\ &= \|(\mathbb{K}u)_j + t(\mathbb{K}\delta_u)_j\| + t^2\tilde{c}\underbrace{\langle(\delta_q)_j, q_j\rangle}_{=0},\end{aligned}$$

where we used the property of $\langle(\delta_q)_j, q_j\rangle = 0$. From the last equation, we may conclude that $q_j + t(\delta_q)_j = (\mathbb{K}u)_j + t(\mathbb{K}\delta_u)_j/\|(\mathbb{K}u)_j + t(\mathbb{K}\delta_u)_j\|$ with norm $\|q_j + t(\delta_q)_j\| = 1$, which implies that the pair $((\mathbb{K}u)_j + t(\mathbb{K}\delta_u)_j, q_j + t(\delta_q)_j)$ is contained in $\operatorname{gph} Q$. Furthermore, taking the distance of this pair to the graph, we get $\operatorname{dist}(((\mathbb{K}u)_j + t(\mathbb{K}\delta_u)_j, q_j +$

$t(\delta_q)_j), \operatorname{gph} Q) = 0$ for all t. Consequently, this pair is a tangent vector, finishing the proof.

□

Lemma 4.2. *The Fréchet normal cone to the graph of Q is given by*

$$N^F_{\operatorname{gph} Q}(u, \mathbb{K}^\top q) = \left\{ (\mathbb{K}^\top \mu, p) : \begin{cases} \mu_j + T_j(\mathbb{K}p)_j = 0, & \text{if } j \in \mathcal{I}, \\ (\mathbb{K}p)_j = 0 & \text{if } j \in \mathcal{A}_s, \\ (\mathbb{K}p)_j = cq_j (c \geq 0), \langle \mu_j, q_j \rangle \leq 0, & \text{if } j \in \mathcal{B}. \end{cases} \right\} \tag{4.13}$$

Proof. Recalling the definition of the Fréchet normal cone definition 2.8, we know that we can build it as the polar of the tangent cone. Considering, in particular, directions of the form $\delta_u \in \ker(\mathbb{K})$ and $\delta_q = 0$, it follows that, for a general normal vector (φ, p), $\langle \varphi, \delta_u \rangle \leq 0$ for all $\delta_u \in \ker(\mathbb{K})$ must hold. This implies $\varphi \in \ker(\mathbb{K})^\perp = \operatorname{range}(\mathbb{K}^\top)$. Consequently, for $(\delta_u, \mathbb{K}^\top \delta_q) \in T_{gphQ}(u, \mathbb{K}^\top q)$ we have that the Fréchet normal cone can be calculated as

$$N^F_{\operatorname{gph} Q}(u, \mathbb{K}^\top q) = \{ (\mathbb{K}^\top \mu, p) \in \mathbb{R}^n \times \mathbb{R}^n : \langle (\mathbb{K}^\top \mu, p), (\delta_u, \mathbb{K}^\top \delta_q) \rangle \leq 0 \}.$$

We can rewrite the inequality as

$$\sum_{j=1}^{n} \langle (\mathbb{K}\delta_u)_j, \mu_j \rangle + \langle (\delta_q)_j, (\mathbb{K}p)_j \rangle \leq 0.$$

Using this characterization, along with the tangent cone presented in Lemma 4.1, we analyze the different cases according to their index set

Case 1: $j \in \mathcal{I}(u)$. Using the characterization of the tangent cone, we have

$$0 \geq \langle (\mathbb{K}\delta_u)_j, \mu_j \rangle + \langle T_j(\mathbb{K}\delta_u), (\mathbb{K}p)_j \rangle = \langle (\mathbb{K}\delta_u)_j, \mu_j + T_j(\mathbb{K}p)_j \rangle,$$

where we used the symmetry of T_j. Since there are no constraints over $(\mathbb{K}\delta_u)_j$, it must necessarily hold $\mu_j + T_j(\mathbb{K}p)_j = 0$.

Case 2: $j \in \mathcal{A}_s(u)$. In this index set we know that $(\mathbb{K}\delta_u)_j = 0$ and $(\delta_q)_j \in \mathbb{R}^2$. Consequently, when we consider the following product

$$\langle (\delta_q)_j, (\mathbb{K}p)_j \rangle \leq 0, \ \forall (\delta_q)_j$$

we obtain that $(\mathbb{K}p)_j = 0$.

Case 3: $j \in \mathcal{B}(u)$. In this index set, there are two conditions in the normal directions.

For the first one, we take $(\mathbb{K}\delta_u)_j = 0$, and the cone inequality reads

$$\langle (\delta_q)_j, (\mathbb{K}p)_j \rangle \leq 0, \forall (\delta_q)_j \text{ s.t. } \langle (\delta_q)_j, q_j \rangle \leq 0. \tag{4.14}$$

By considering the feasible set for $(\delta_q)_j$ as $\Pi(q_j) := \{(\delta_q)_j : \langle (\delta_q)_j, q_j \rangle \leq 0\}$, we can rewrite (4.14) as follows

$$\langle (\delta_q)_j, (\mathbb{K}p)_j \rangle \leq 0, \ \forall (\delta_q)_j \in \Pi(q_j).$$

It implies that $(\mathbb{K}p)_j$ belongs to the polar cone of $\Pi(q_j)$, denoted by $\Pi^\circ(q_j)$; therefore, it must hold $(\mathbb{K}p)_j = cq_j$ with $c \geq 0$ as shown in Figure 4.2a.

For the second case, we take $(\mathbb{K}\delta_u)_j = \tilde{c}_j q_j (\tilde{c}_j \geq 0)$ and the cone inequality reads

$$\langle \tilde{c}_j q_j, \mu_j \rangle + \langle (\mathbb{K}p)_j, (\delta_q)_j \rangle \leq 0, \forall (\delta_q)_j \text{ s.t. } \langle (\delta_q)_j, q_j \rangle = 0. \tag{4.15}$$

Again, we define the feasible set for $(\delta_q)_j$ as $\Pi(q_j) := \{(\delta_q)_j : \langle (\delta_q)_j, q_j \rangle = 0\}$, see Figure 4.2b. Then, (4.15) may be rewritten as

$$\langle \tilde{c}_j q_j, \mu_j \rangle + \langle (\mathbb{K}p)_j, (\delta_q)_j \rangle \leq 0, \forall (\delta_q)_j \in \Pi(q_j). \tag{4.16}$$

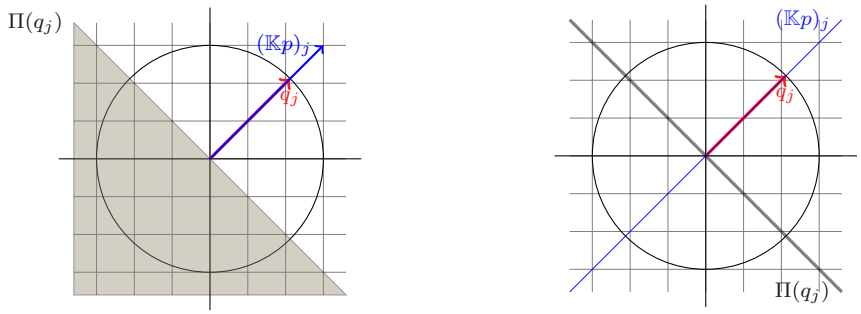
Now, considering in particular $\tilde{c}_j = 0$ we may rewrite (4.15) as follows

$$\langle (\mathbb{K}p)_j, (\delta_q)_j \rangle \leq 0, \forall (\delta_q)_j \in \Pi(q_j).$$

Similarly to the previous case, we get $(\mathbb{K}p)_j \in \Pi^\circ(q_j)$, yielding that $(\mathbb{K}p)_j = cq_j$ with $c \in \mathbb{R}$. Furthermore, taking in particular $(\delta_q)_j = 0$ in (4.16), we obtain that $\langle q_j, \mu_j \rangle \leq 0$.

Finally, considering both cases, it yields the result.

□



(a) Case 1 **(b)** Case 2

Figure 4.2: Frechet normal cone in the biactive (set geometric interpretation).

Lemma 4.3. *Let $(u, \mathbb{K}^\top q) \in \operatorname{gph} Q$, described in (4.5) and a pair $(\mathbb{K}^\top \mu, p) \in \mathbb{R}^n \times \mathbb{R}^n$. If $(\mathbb{K}^\top \mu, p)$ satisfies the following conditions*

$$\left.\begin{array}{ll} \mu_j + T_j(\mathbb{K}p)_j = 0, & \text{if } j \in \mathcal{I}, \\ (\mathbb{K}p)_j = 0, & \text{if } j \in \mathcal{A}_s, \\ (\mathbb{K}p)_j = 0, \ \vee & \\ (\mathbb{K}p)_j = cq_j (c \in \mathbb{R}), \ \langle \mu_j, q_j \rangle = 0, \ \vee & \text{if } j \in \mathcal{B}. \\ (\mathbb{K}p)_j = cq_j (c \geq 0), \langle \mu_j, q_j \rangle \leq 0. & \end{array}\right\} \tag{4.17}$$

Then, $(\mathbb{K}^\top \mu, p) \in N^M_{\operatorname{gph} Q}(u, \mathbb{K}^\top q)$, where $N^M_{\operatorname{gph} Q}(u, \mathbb{K}^\top q)$ is the Mordukhovich normal cone to the graph of Q at $(u, \mathbb{K}^\top q)$.

Proof. Let us recall the definition of the Mordukhovich normal cone for our problem

$$N^M_{\operatorname{gph} Q}(u, \mathbb{K}^\top q) = \{(\mathbb{K}^\top \mu, p) \in \mathbb{R}^n \times \mathbb{R}^n : (\mathbb{K}^\top \mu_k, p_k) \in N^F_{\operatorname{gph} Q}(u_k, \mathbb{K}^\top q_k) : (\mathbb{K}^\top \mu_k, p_k) \to (\mathbb{K}^\top \mu, p), (u_k, \mathbb{K}^\top q_k) \to (u, \mathbb{K}^\top q)\}.$$

Considering limiting sequences in the inactive and active sets, we obtain the same directions as those for the Fréchet normal cone. The differences lie in the biactive set, where we can consider several approximations.

We may take approximation sequences with *inactive* components and from Lemma 4.2 we know

$$0 = (\mu_k)_j + \frac{(\mathbb{K}p_k)_j}{\|(\mathbb{K}u_k)_j\|} - \frac{(\mathbb{K}u_k)_j \langle (\mathbb{K}u_k)_j, (\mathbb{K}p_k)_j \rangle}{\|(\mathbb{K}u_k)_j\|^3}. \tag{4.18}$$

Testing (4.18) with $(\mathbb{K}p_k)_j$, yields

$$0 = \langle (\mu_k)_j, (\mathbb{K}p_k)_j \rangle + \frac{\|(\mathbb{K}p_k)_j\|^2}{\|(\mathbb{K}u_k)_j\|} - \frac{1}{\|(\mathbb{K}u_k)_j\|} \left\langle \frac{(\mathbb{K}u_k)_j}{\|(\mathbb{K}u_k)_j\|}, (\mathbb{K}p_k)_j \right\rangle^2.$$

Recalling $(q_k)_j = (\mathbb{K}u_k)_j / \|(\mathbb{K}u_k)_j\|$ if $(\mathbb{K}u_k)_j = 0$

$$\begin{aligned} 0 &= \langle (\mu_k)_j, (\mathbb{K}p_k)_j \rangle + \frac{\|(\mathbb{K}p_k)_j\|^2}{\|(\mathbb{K}u_k)_j\|} - \frac{1}{\|(\mathbb{K}u_k)_j\|} \langle (q_k)_j, (\mathbb{K}p_k)_j \rangle^2, \\ &\geq \langle (\mu_k)_j, (\mathbb{K}p_k)_j \rangle + \frac{\|(\mathbb{K}p_k)_j\|^2}{\|(\mathbb{K}u_k)_j\|} - \frac{\|q_k\|^2 \|(\mathbb{K}p_k)_j\|^2}{\|(\mathbb{K}u_k)_j\|}, \\ &= \langle (\mu_k)_j, (\mathbb{K}p_k)_j \rangle, \end{aligned}$$

where we used the property of $\|(q_k)_j\| = 1$. Furthermore, taking the limit as $k \to \infty$, we obtain that

$$\langle \mu_j, (\mathbb{K}p)_j \rangle \leq 0. \tag{4.19}$$

Now, testing (4.18) with $(q_k)_j$ we get the following product

$$\begin{aligned} 0 &= \langle (\mu_k)_j, (q_k)_j \rangle + \frac{\langle (q_k)_j, (\mathbb{K}p_k)_j \rangle}{\|(\mathbb{K}u_k)_j\|} - \frac{\langle (q_k)_j, (\mathbb{K}u_k)_j \rangle \langle (\mathbb{K}u_k)_j, (\mathbb{K}p_k)_j \rangle}{\|(\mathbb{K}u_k)_j\|^3}, \\ &= \langle (\mu_k)_j, (q_k)_j \rangle + \frac{\langle (q_k)_j, (\mathbb{K}p_k)_j \rangle}{\|(\mathbb{K}u_k)_j\|} - \frac{\|(q_k)_j\|^2}{\|(\mathbb{K}u_k)\|} \langle (q_k)_j, (\mathbb{K}p_k)_j \rangle, \end{aligned}$$

where using the fact that $\|(q_k)_j\| = 1$ and taking the limit as $k \to \infty$ we get

$$\langle \mu_j, q_j \rangle = 0. \tag{4.20}$$

Furthermore, defining the set of feasible μ_j as $\Pi(q_j) := \{\mu_j : \langle \mu_j, q_j \rangle = 0\}$, we may constrain (4.19) to this set as follows

$$\langle \mu_j, (\mathbb{K}p)_j \rangle \leq 0, \ \forall \mu_j \in \Pi(q_j).$$

Consequently, it must hold $(\mathbb{K}p)_j \in \Pi^\circ(q_j)$ and that $(\mathbb{K}p)_j = cq_j$ with $c \in \mathbb{R}$.

For the case we take approximations through sequences in the *active* set, we know $(\mathbb{K}p_k) = 0$, which, when taking the limit as $k \to \infty$, yields $(\mathbb{K}p)_j = 0$.

When the approximation is taken using sequences in the *biactive* set, we have $(\mathbb{K}p_k)_j = c(q_k)_j$ with $c \geq 0$; which in the limit as $k \to \infty$ reads $(\mathbb{K}p)_j = cq_j$. Likewise, for sequences of components in the biactive set, we know the following bound holds

$$\langle (\mu_k)_j, (q_k)_j \rangle \leq 0.$$

Taking the limit as $k \to \infty$ it yields $\langle \mu_j, q_j \rangle \leq 0$. In both cases, the cone directions coincide with the Fréchet normal ones.

Finally, since we took sequences $(\mathbb{K}^\top \mu_k, p_k) \in N^F_{\text{gph}\, Q}(u_k, \mathbb{K}^\top q_k) \subset N^M_{\text{gph}\, Q}(u_k, \mathbb{K}^\top q_k)$ and $N^M_{\text{gph}\, Q}(u, \mathbb{K}^\top q) = clN^F_{\text{gph}\, Q}(u, \mathbb{K}^\top q)$, the proof is complete. □

Theorem 4.1 (M-Stationarity)**.** *Let $J : \mathbb{R}^m \to \mathbb{R}$ be continuously differentiable, $\mathcal{F} : \mathbb{R}^n_+ \times \mathbb{R}^n \to \mathbb{R}$ twice continuously differentiable and strongly convex with respect to u, and (λ^*, u^*, q^*) be a local solution to (4.1). Then, there exist KKT multipliers $(\mathbb{K}^\top \mu, p)$*

and a $\zeta \in \mathbb{R}^n$ *such that*

$$\nabla_u \mathcal{F}(\lambda^*, u^*) + \mathbb{K}^\top q^* = 0, \tag{4.21a}$$

$$\langle q_j^*, (\mathbb{K}u^*)_j \rangle - \|(\mathbb{K}u^*)_j\| = 0, \qquad \forall j = 1, \dots, m, \tag{4.21b}$$

$$\|q_j^*\| \leq 1, \qquad \forall j = 1, \dots, m, \tag{4.21c}$$

$$\nabla_{uu} \mathcal{F}(\lambda^*, u^*)^\top p - \mathbb{K}^\top \mu - \nabla J(u^*) = 0, \tag{4.21d}$$

$$\nabla_{u\lambda} \mathcal{F}(\lambda^*, u^*)^\top p - \zeta = 0, \tag{4.21e}$$

$$\langle \zeta, \lambda^* \rangle = 0, \tag{4.21f}$$

$$\zeta \leq 0, \tag{4.21g}$$

$$(\mathbb{K}^\top \mu, p) \in N^M_{\operatorname{gph} Q}(u^*, \mathbb{K}^\top q^*) \tag{4.21h}$$

Proof. To justify the existence of KKT multipliers, we will use the constraint qualification condition presented in Theorem 2.9 with $F_1(x, y) = \nabla \mathcal{F}(\lambda, u)$, $F_2(x, y) = u$. This theorem guarantees the existence of said multipliers if the following inclusion

$$\begin{bmatrix} \mathbf{0} & -\nabla_{u\lambda} \mathcal{F}(\lambda^*, u^*)^\top \\ \boldsymbol{I} & -\nabla_{uu} \mathcal{F}(\lambda^*, u^*)^\top \end{bmatrix} \begin{bmatrix} \mathbb{K}^\top \mu \\ p \end{bmatrix} \in -N^M_{\mathbb{R}^n_+}(\lambda^*) \times N^M_{\mathbb{R}^n}(u^*) = -N_{\mathbb{R}^n_+}(\lambda^*) \times \{0\} \tag{4.22}$$

implies $\mathbb{K}^\top \mu = 0$ and $p = 0$. Here, we used Remark 2.2 to characterize the Mordukhovich normal cone for the feasible set of λ^*, which coincides with the convex normal cone. Therefore, $N^M_{\mathbb{R}^n_+}(\lambda^*) = N_{\mathbb{R}^n_+}(\lambda^*) = \{v \in \mathbb{R}^n : \ \langle v, \lambda^* \rangle = 0, v \leq 0\}$.

Consequently, (4.22) can be written as

$$\mathbb{K}^\top \mu - \nabla_{uu} \mathcal{F}(\lambda^*, u^*)^\top p = 0, \tag{4.23}$$

$$\langle \nabla_{u\lambda} \mathcal{F}(\lambda^*, u^*)^\top p, \lambda^* \rangle = 0, \tag{4.24}$$

$$\nabla_{u\lambda} \mathcal{F}(\lambda^*, u^*)^\top p \leq 0. \tag{4.25}$$

Now, let us take $(\mathbb{K}^\top \mu, p) \in N^M_{\operatorname{gph} Q}(u^*, \mathbb{K}^\top q^*)$ and let us multiply (4.23) by p on the left. The product now reads

$$\langle p, \mathbb{K}^\top \mu \rangle - \langle p, \nabla_{uu} \mathcal{F}(\lambda^*, u^*)^\top p \rangle = 0.$$

By splitting the product according to the different index sets, we have

$$\langle p, \nabla_{uu} \mathcal{F}(\lambda^*, u^*)^\top p \rangle = \sum_{j \in \mathcal{I}} \langle \mu_j, (\mathbb{K}p)_j \rangle + \sum_{j \in \mathcal{A}_s} \langle \mu_j, (\mathbb{K}p)_j \rangle + \sum_{j \in \mathcal{B}} \langle \mu_j, (\mathbb{K}p)_j \rangle$$

Considering the characterization of the Mordukhovich normal cone, particularly for

any $j \in \mathcal{A}_s$, we know $(\mathbb{K}p)_j = 0$. Then, the previous product now reads

$$\langle p, \nabla_{uu}\mathcal{F}(\lambda^*, u^*)^\top p\rangle = \sum_{j\in\mathcal{I}}\langle \mu_j, (\mathbb{K}p)_j\rangle + \sum_{j\in\mathcal{B}}\langle \mu_j, (\mathbb{K}p)_j\rangle.$$

Furthermore, for $j \in \mathcal{B}$ we have that $(\mathbb{K}p)_j$ is either equal to zero, $(\mathbb{K}p)_j = cq_j$ with $c \in \mathbb{R}$ and $\langle \mu_j, q_j\rangle = 0$, or $(\mathbb{K}p)_j = cq_j$ with $c \geq 0$ and $\langle \mu_j, q_j\rangle \leq 0$; consequently, the product yields

$$\langle p, \nabla_{uu}\mathcal{F}(\lambda^*, u^*)^\top p\rangle = \sum_{j\in\mathcal{I}} \underbrace{-\langle T_j(\mathbb{K}p)_j, (\mathbb{K}p)_j\rangle}_{\leq 0} + \sum_{j\in\mathcal{B}} c\,\underbrace{\langle \mu_j, q_j\rangle}_{\leq 0} \leq 0,$$

where we used the positive semi-definiteness of the matrix T_j and the characterization of the Mordukhovich normal cone for $j \in \mathcal{B}$. Furthermore, using the strong convexity of $\mathcal{F}$ we know $\langle p, \nabla_{uu}\mathcal{F}(\lambda^*, u^*)^\top p\rangle \geq 0$. Both inequalities imply $p = 0$ and consequently, replacing this result in (4.23), it yields $\mathbb{K}^\top\mu = 0$.

This previous result, allow us to guarantee the existence of KKT multipliers $(\mathbb{K}^\top\mu, p) \in N^M_{\operatorname{gph} Q}(u^*, \mathbb{K}^\top q^*)$ and a $\zeta \in N^M_{\mathbb{R}^n_+}(\lambda^*)$, such that

$$0 = \nabla J(u^*) + \mathbb{K}^\top\mu - \nabla_{uu}\mathcal{F}(\lambda^*, u^*)^\top p, \tag{4.26}$$

$$0 = -\nabla_{u\lambda}\mathcal{F}(\lambda^*, u^*)^\top p + \zeta. \tag{4.27}$$

To recover the optimality system in (4.21), let us take (λ^*, u^*, q^*), a local optimal solution of (4.1). Then, note that equations in (4.26) and (4.27) correspond to equations (d) and (e) respectively. Taking a $\zeta \in \mathbb{R}^n$ we must add the conditions $\langle \zeta, \lambda^*\rangle = 0$ and $\zeta \leq 0$ to guarantee it is contained in $N^M_{\mathbb{R}^n_+}(\lambda^*)$, yielding equations (f) and (g). Finally, equations (a-c) correspond to the state constraints of the original problem. □

4.2 Bouligand Stationarity

Let us now introduce the solution operator for the lower-level problem $S : \mathbb{R}^n_+ \ni \lambda \to u \in \mathbb{R}^n$ that maps each parameter $\lambda \in \mathbb{R}^n_+$ to the corresponding reconstruction $u \in \mathbb{R}^n$. If this mapping is single-valued, we can make use of it to formulate (3.3) as a reduced optimization problem

$$\min_{\lambda\,\in\,\mathbb{R}^n_+} \quad j(\lambda) := J(S(\lambda)). \tag{4.28}$$

Furthermore, if the solution operator is Bouligand (B)-differentiable, i.e., it is locally Lipschitz continuous and directionally differentiable, we can make use of the chain rule for B-differentiable functions, see theorem 2.8, to conclude that the composite mapping J, as a function of λ, is B-differentiable as well. In such a case, its directional derivative

in a direction h is given by

$$j'(\lambda; h) = \langle \nabla J(u), S'(\lambda; h) \rangle, \tag{4.29}$$

where $S'(\lambda; h)$ is the directional derivative of the solution operator in direction h. Moreover, if λ^* is a local optimal solution and $u^* = S(\lambda^*)$ its corresponding reconstruction, then it satisfies the following necessary optimality condition

$$j'(\lambda^*; \lambda - \lambda^*) = \langle \nabla J(u^*), S'(\lambda^*; \lambda - \lambda^*) \rangle \geq 0, \ \forall \lambda \in \mathbb{R}^n_+. \tag{4.30}$$

A point λ^* satisfying the necessary condition (4.30) is called *Bouligand (B)-stationary*. This type of stationarity condition is based on the tangent cone to our feasible parameter set and can be interpreted as the counterpart of the implicit programming approach in the discussion of finite-dimensional MPECs, see [51, Lemma 4.2.5].

However, to fully characterize the Bouligand-stationarity condition (4.30), we need a characterization for the directional derivative of the solution operator. Now, regarding the solution operator for the lower-level problem (4.1b), we obtain the following result.

Theorem 4.2. *The solution operator for the lower-level problem (4.1b) $S : \mathbb{R}^n_+ \ni \lambda \to u \in \mathbb{R}^n$ is locally Lipschitz continuous.*

Proof. Thanks to Theorem 3.1, we know the lower-level problem has a unique solution. Moreover, $\lambda_1, \lambda_2 \in \mathbb{R}^n_+$ and its corresponding solutions u_1, u_2 satisfy

$$\begin{aligned}
\langle \nabla_u \mathcal{F}(\lambda_1, u_1), v - u_1 \rangle + \sum_{j=1}^{n} \|(\mathbb{K}v)_j\| - \sum_{j=1}^{n} \|(\mathbb{K}u_1)_j\| \geq 0, \ \forall v \in \mathbb{R}^n \\
\langle \nabla_u \mathcal{F}(\lambda_2, u_2), w - u_2 \rangle + \sum_{j=1}^{n} \|(\mathbb{K}w)_j\| - \sum_{j=1}^{n} \|(\mathbb{K}u_2)_j\| \geq 0, \ \forall w \in \mathbb{R}^n.
\end{aligned}$$

Taking in particular $v = u_2$ and $w = u_1$ and adding the inequalities, it yields

$$\langle \nabla_u \mathcal{F}(\lambda_2, u_2) - \nabla_u \mathcal{F}(\lambda_1, u_1), u_2 - u_1 \rangle \leq 0. \tag{4.31}$$

Since $\mathcal{F}$ is linear with respect to λ, it holds that $\nabla_u \mathcal{F}$ maintains the linearity property with respect to λ as well. Using this property and adding a zero to (4.31), it reads

$$\langle \nabla_u \mathcal{F}(\lambda_2 - \lambda_1, u_2), u_2 - u_1 \rangle + \langle \nabla_u \mathcal{F}(\lambda_1, u_2) - \nabla_u \mathcal{F}(\lambda_1, u_1), u_2 - u_1 \rangle \leq 0.$$

Using the strong convexity with respect to u, we know there exists a constant $\mu > 0$ such that the following bound holds true

$$\langle \nabla_u \mathcal{F}(\lambda_2 - \lambda_1, u_2), u_2 - u_1 \rangle + \mu \|u_2 - u_1\|^2 \leq 0. \tag{4.32}$$

Furthermore, the linearity of $\mathcal{F}$ with respect to λ also implies it is Lipschitz continuous with respect to λ as well. We will name L_λ to its Lipschitz constant, and we may bound (4.32) using the Cauchy-Schwarz inequality as follows

$$\begin{aligned}\mu\|u_2 - u_1\|^2 &\leq \langle \nabla_u \mathcal{F}(\lambda_2 - \lambda_1, u_2), u_1 - u_2\rangle, \\ &\leq \|\nabla_u \mathcal{F}(\lambda_2, u_2) - \nabla_u \mathcal{F}(\lambda_1, u_2)\|\|u_2 - u_1\|, \\ &\leq L_\lambda(u_2)\|\lambda_2 - \lambda_1\|\|u_2 - u_1\|.\end{aligned}$$

Finally, by rearranging the terms, we obtain the result

$$\|u_2 - u_1\| \leq \frac{L_\lambda(u_2)}{\mu}\|\lambda_2 - \lambda_1\|.$$

□

4.2.1 Directional Differentiability

In this section, we will study the differentiability properties of the solution operator for the lower-level problem (4.1b). This study will require a sensitivity analysis of the solution operator concerning the regularization parameter λ. By taking a perturbed regularization parameter λ^t in the primal-dual formulation for the lower-level problem (4.3) such that $\lambda_j^t = \lambda_j + th_j \geq 0$ we get the following perturbed lower-level problem

$$\nabla_u \mathcal{F}(\lambda^t, u^t) + \mathbb{K}^\top q^t = 0, \tag{4.33a}$$

$$\langle q_j^t, (\mathbb{K}u^t)_j\rangle - \|(\mathbb{K}u^t)_j\| = 0, \qquad \forall j = 1, \dots, m, \tag{4.33b}$$

$$\|q_j^t\| \leq 1, \qquad \forall j = 1, \dots, m. \tag{4.33c}$$

Thanks to the boundedness of q^t, there exist a subsequence, denoted the same, so that $q^t \to \tilde{q} \in \mathbb{R}^{m\times 2}$, to some $\tilde{q}$, as $t \to 0$. Additionally, thanks to the Lipschitz continuity of the solution operator, we know that the following sequence is bounded

$$\left\|\frac{u^t - u}{t}\right\| \leq \frac{L_\lambda(u)}{\mu}\left\|\frac{\lambda^t - \lambda}{t}\right\| = \frac{L_\lambda(u)}{\mu}\|h\| < \infty.$$

Therefore, we can guarantee the existence of a subsequence of $\{(u^t - u)/t\}$, denoted with the same symbol, satisfying

$$\lim_{t\to 0}\frac{u^t - u}{t} \to \eta \in \mathbb{R}^n. \tag{4.34}$$

Theorem 4.3. *The limit described in (4.34) satisfies $\eta \in \mathcal{C}(u)$ where*

$$\mathcal{C}(u) := \left\{ v \in \mathbb{R}^n : \begin{cases} (\mathbb{K}v)_j = 0, & \forall j \in \mathcal{A}_s, \\ \langle q_j, (\mathbb{K}v)_j \rangle = \|(\mathbb{K}v)_j\|, & \forall j \in \mathcal{B}. \end{cases} \right\} \tag{4.35}$$

Proof. By adding the complementarity relations (4.33) and (4.3) and dividing by t we get

$$0 = \left\langle \frac{q_j^t - q_j}{t}, (\mathbb{K}u^t)_j \right\rangle + \left\langle q_j, \frac{(\mathbb{K}u^t)_j - (\mathbb{K}u)_j}{t} \right\rangle - \left(\frac{\|(\mathbb{K}u^t)_j\| - \|(\mathbb{K}u)_j\|}{t} \right). \tag{4.36}$$

When considering a point in $j \in \mathcal{A}_s \cup \mathcal{B}$, we know the sequence $(\mathbb{K}u^t)_j \to (\mathbb{K}u)_j = 0$. Therefore, taking the limit as $t \to 0$ in (4.36), we get

$$0 = \langle q_j, (\mathbb{K}\eta)_j \rangle - \|(\mathbb{K}\eta)_j\|, \forall j \in \mathcal{A}_s \cup \mathcal{B}. \tag{4.37}$$

Now, considering $j \in \mathcal{A}_s$, we know in this index set $\|q_j\| < 1$. Using Cauchy-Schwarz in (4.37) we get

$$0 = \langle q_j, (\mathbb{K}\eta)_j \rangle - \|(\mathbb{K}\eta)_j\| \leq \|(\mathbb{K}\eta)_j\|(\|q_j\| - 1),$$

which implies $(\mathbb{K}\eta)_j = 0$ for all $j \in \mathcal{A}_s$. □

Remark 4.1. *If q^1 and q^2 are two different slack variables associated with the solution u in (4.3), then the two sets $\mathcal{C}_i$ for $i = 1, 2$ defined as follows*

$$\mathcal{C}_i := \left\{ v \in \mathbb{R}^n : \begin{cases} (\mathbb{K}v)_j = 0, & \text{if } \|q_j^i\| < 1, \\ \langle q_j^i, (\mathbb{K}v)_j \rangle = \|(\mathbb{K}v)_j\|, & \text{if } (\mathbb{K}u)_j = 0,\ \|q_j^i\| = 1. \end{cases} \right\},$$

coincide, since $\mathbb{K}^\top q^1 = -\nabla_u \mathcal{F}(\lambda, u) = \mathbb{K}^\top q^2$. Consequently, the set $\mathcal{C}(u)$ does not depend on the slack variable, only on the solution u.

Lemma 4.4. *The cone $\mathcal{C}(u)$ can alternatively be written as*

$$\mathcal{C}(u) = \left\{ v \in \mathbb{R}^n : \langle \mathbb{K}^\top q, v \rangle \geq \sum_{j \in \mathcal{I}} \left\langle \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, (\mathbb{K}v)_j \right\rangle + \sum_{j \in \mathcal{A}_s \cup \mathcal{B}} \|(\mathbb{K}v)_j\| \right\} \tag{4.38}$$

Proof. Let us denote the set in (4.38) as $\mathcal{M}$. Taking $v \in \mathcal{C}$, as in (4.35), and using its

definition, we obtain

$$\begin{aligned}\langle \mathbb{K}^\top q, v\rangle &= \sum_{j\in\mathcal{I}} \langle q_j, (\mathbb{K}v)_j\rangle + \sum_{j\in\mathcal{A}_s} \langle q_j, (\mathbb{K}v)_j\rangle + \sum_{j\in\mathcal{B}} \langle q_j, (\mathbb{K}v)_j\rangle,\\ &= \sum_{j\in\mathcal{I}} \left\langle \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, (\mathbb{K}v)_j \right\rangle + \sum_{j\in\mathcal{A}_s} \langle q_j, \underbrace{(\mathbb{K}v)_j}_{=0}\rangle + \sum_{j\in\mathcal{B}} \|(\mathbb{K}v)_j\|,\end{aligned}$$

and, consequently, $\mathcal{C} \subset \mathcal{M}$.

To prove the reverse inclusion, let us take $v \in \mathcal{M}$. Then, we may rewrite the inequality in (4.38) as follows

$$\sum_{j\in\mathcal{I}} \left\langle \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, (\mathbb{K}v)_j \right\rangle + \sum_{j\in\mathcal{B}\cup\mathcal{A}_s} \langle q_j, (\mathbb{K}v)_j\rangle \geq \sum_{j\in\mathcal{I}} \left\langle \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, (\mathbb{K}v)_j \right\rangle + \sum_{j\in\mathcal{A}_s\cup\mathcal{B}} \|(\mathbb{K}v)_j\|, \quad (4.39)$$

which can be rewritten as

$$\sum_{j\in\mathcal{B}\cup\mathcal{A}_s} \langle q_j, (\mathbb{K}v)_j\rangle \geq \sum_{j\in\mathcal{A}_s\cup\mathcal{B}} \|(\mathbb{K}v)_j\|.$$

Now, using the Cauchy-Schwarz inequality and $\|q_j\| \leq 1$, for all $j \in \mathcal{A}_s \cup \mathcal{B}$, we can upper bound this term as follows

$$\sum_{j\in\mathcal{A}_s\cup\mathcal{B}} \|(\mathbb{K}v)_j\| \leq \sum_{j\in\mathcal{A}_s\cup\mathcal{B}} \langle q_j, (\mathbb{K}v)_j\rangle \leq \sum_{j\in\mathcal{A}_s\cup\mathcal{B}} \underbrace{\|q_j\|}_{\leq 1} \|(\mathbb{K}v)_j\| \leq \sum_{j\in\mathcal{A}_s\cup\mathcal{B}} \|(\mathbb{K}v)_j\|. \quad (4.40)$$

Therefore, since the lower and upper bounds are the same, it holds

$$\sum_{j\in\mathcal{A}_s\cup\mathcal{B}} \langle q_j, (\mathbb{K}v)_j\rangle - \sum_{j\in\mathcal{A}_s\cup\mathcal{B}} \|(\mathbb{K}v)_j\| = 0. \quad (4.41)$$

Consequently, for each index in $\mathcal{A}_s \cup \mathcal{B}$ we have

$$\langle q_j, (\mathbb{K}v)_j\rangle = \|(\mathbb{K}v)_j\|, \ \forall j \in \mathcal{A}_s \cup \mathcal{B}.$$

Taking, in particular, $j \in \mathcal{A}_s$ and using the Cauchy-Schwarz inequality, along with the property $\|q_j\| < 1$ in the active set, it yields

$$\|(\mathbb{K}v)_j\| = \langle q_j, (\mathbb{K}v)_j\rangle \leq \underbrace{\|q_j\|}_{<1} \|(\mathbb{K}v)_j\| < \|(\mathbb{K}v)_j\|,$$

which implies that $(\mathbb{K}v)_j = 0$ for all $j \in \mathcal{A}_s$ and it follows that $\mathcal{M} \subset \mathcal{C}$, concluding the proof. □

Now, to prove the directional differentiability of the solution operator for the lower-level problem (4.1b), we will first demonstrate the following lemmata.

Lemma 4.5. *Let $\mathbb{R}^n_+ \ni \lambda$ and $\mathbb{R}^n_+ \ni \lambda + th$. Then for every $v \in \mathcal{C}$, it holds*

$$\left\langle \mathbb{K}^\top \left(\frac{q^t - q}{t} \right), v \right\rangle \leq \sum_{j \in \mathcal{I}} \frac{1}{t} \left\langle \frac{(\mathbb{K}u^t)_j}{\|(\mathbb{K}u^t)_j\|} - \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, (\mathbb{K}v)_j \right\rangle. \tag{4.42}$$

Proof. Let us first start by taking $v \in \mathcal{C}(u)$ and bound the following product

$$\begin{aligned} \langle \mathbb{K}^\top q^t, v \rangle &= \sum_{j \in \mathcal{I}} \langle q_j^t, (\mathbb{K}v)_j \rangle + \sum_{j \in \mathcal{A}_s} \langle q_j^t, \underbrace{(\mathbb{K}v)_j}_{=0} \rangle + \sum_{j \in \mathcal{B}} \langle q_j^t, (\mathbb{K}v)_j \rangle, \\ &\leq \sum_{j \in \mathcal{I}} \left\langle \frac{(\mathbb{K}u^t)_j}{\|(\mathbb{K}u^t)_j\|}, (\mathbb{K}v)_j \right\rangle + \sum_{j \in \mathcal{B}} \|(\mathbb{K}v)_j\|, \end{aligned}$$

Now, given that we took $v \in \mathcal{C}(u)$, we know that the bound in Lemma 4.4 holds, i.e.,

$$\langle \mathbb{K}^\top q, v \rangle \geq \sum_{j \in \mathcal{I}} \left\langle \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, (\mathbb{K}v)_j \right\rangle + \sum_{j \in \mathcal{B}} \|(\mathbb{K}v)_j\|.$$

Adding both inequalities and dividing by t yields the result. □

Lemma 4.6. *Let $\mathbb{R}^n_+ \ni \lambda$ and $\mathbb{R}^n_+ \ni \lambda + th$. It holds*

$$\left\langle \mathbb{K}^\top \left(\frac{q^t - q}{t} \right), \frac{u^t - u}{t} \right\rangle \geq \sum_{j \in \mathcal{I}} \frac{1}{t} \left\langle \frac{(\mathbb{K}u^t)_j}{\|(\mathbb{K}u^t)_j\|} - \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, \frac{(\mathbb{K}u^t)_j - (\mathbb{K}u)_j}{t} \right\rangle.$$

Proof. For t sufficiently small, we can split the product by their index set

$$\begin{aligned} \left\langle \mathbb{K}^\top \left(\frac{q^t - q}{t} \right), \frac{u^t - u}{t} \right\rangle = \sum_{j \in \mathcal{I}} \frac{1}{t} \left\langle \frac{(\mathbb{K}u^t)_j}{\|(\mathbb{K}u^t)_j\|} - \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, \frac{(\mathbb{K}u^t)_j - (\mathbb{K}u)_j}{t} \right\rangle \\ + \sum_{j \in \mathcal{A}_s \cup \mathcal{B}} \left\langle \frac{q_j^t - q_j}{t}, \frac{(\mathbb{K}u^t)_j - (\mathbb{K}u)_j}{t} \right\rangle. \end{aligned}$$

Considering the index set $\mathcal{A}_s \cup \mathcal{B}$, the complementarity relations in (4.3) and (4.33)

can be used to bound the following product

$$\begin{aligned}
&\sum_{j\in\mathcal{A}_s\cup\mathcal{B}} \frac{1}{t^2}\langle q_j^t - q_j, (\mathbb{K}u^t)_j - (\mathbb{K}u)_j\rangle \\
&= \frac{1}{t^2}\sum_{j\in\mathcal{A}_s\cup\mathcal{B}} \langle q_j^t, (\mathbb{K}u^t)_j\rangle - \langle q_j^t, (\mathbb{K}u)_j\rangle - \langle q_j, (\mathbb{K}u^t)_j\rangle + \langle q_j, (\mathbb{K}u)_j\rangle, \\
&\geq \frac{1}{t^2}\sum_{\mathcal{A}_s\cup\mathcal{B}} \|(\mathbb{K}u^t)_j\| - \|q_j\|\|(\mathbb{K}u^t)_j\| - \|q_j^t\|\|(\mathbb{K}u)_j\| + \|(\mathbb{K}u)_j\|, \\
&\geq 0.
\end{aligned}$$

which implies the result. □

Theorem 4.4. *Let $\lambda \in \mathbb{R}^n_+$ and $h \in \mathbb{R}^n$ be a direction such that $\lambda + th \geq 0$, for t small enough. The solution operator $S : \lambda \to S(\lambda) = u \in \mathbb{R}^n$ is directionally differentiable and its directional derivative $\eta \in \mathcal{C}(u)$ at u, in direction h, is given by the solution of the following variational inequality:*

$$\langle \nabla_{uu}\mathcal{F}(\lambda, u)\eta + \nabla_u\mathcal{F}(h, u), v - \eta\rangle + \sum_{j\in\mathcal{I}} \langle T_j(\mathbb{K}\eta)_j, (\mathbb{K}v)_j - (\mathbb{K}\eta)_j\rangle \geq 0,\ \forall v \in \mathcal{C}, \quad (4.43)$$

where $T_j(\mathbb{K}v)_j = \frac{(\mathbb{K}v)_j}{\|(\mathbb{K}u)_j\|} - \frac{(\mathbb{K}u)_j(\mathbb{K}u)_j^\top(\mathbb{K}v)_j}{\|(\mathbb{K}u)_j\|^3}$ for $v \in \mathbb{R}^n$.

Proof. Taking (4.3) and (4.33) and testing it with $v - \frac{u^t - u}{t}$ we get

$$\begin{aligned}
0 &= \left\langle \nabla_u\mathcal{F}(\lambda + th, u^t) + \mathbb{K}^\top q^t - \nabla_u\mathcal{F}(\lambda, u) - \mathbb{K}^\top q, v - \frac{u^t - u}{t}\right\rangle, \\
&= \left\langle \nabla_u\mathcal{F}(\lambda, u^t) - \nabla_u\mathcal{F}(\lambda, u), v - \frac{u^t - u}{t}\right\rangle + t\left\langle \nabla_u\mathcal{F}(h, u^t), v - \frac{u^t - u}{t}\right\rangle \\
&\quad + \left\langle \mathbb{K}^\top(q^t - q), v - \frac{u^t - u}{t}\right\rangle,
\end{aligned}$$

where we used the linearity of $\mathcal{F}$ with respect to λ and the fact that $\mathbb{K}$ is a linear operator. Furthermore, dividing by t, it now reads

$$\begin{aligned}
0 = \left\langle \frac{\nabla_u\mathcal{F}(\lambda, u^t) - \nabla_u\mathcal{F}(\lambda, u)}{t}, v - \frac{u^t - u}{t}\right\rangle + \left\langle \nabla_u\mathcal{F}(h, u^t), v - \frac{u^t - u}{t}\right\rangle \\
+ \left\langle \mathbb{K}\left(\frac{q^t - q}{t}\right), v - \frac{u^t - u}{t}\right\rangle,
\end{aligned}$$

and using the bounds presented in Lemmas 4.5 and 4.6 it yields

$$0 \leq \left\langle \frac{\nabla_u \mathcal{F}(\lambda, u^t) - \nabla_u \mathcal{F}(\lambda, u)}{t}, v - \frac{u^t - u}{t} \right\rangle + \left\langle \nabla_u \mathcal{F}(h, u^t), v - \frac{u^t - u}{t} \right\rangle + \sum_{j \in \mathcal{I}} \frac{1}{t} \left\langle \frac{(\mathbb{K}u^t)_j}{\|(\mathbb{K}u^t)_j\|} - \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, (\mathbb{K}v)_j - \frac{(\mathbb{K}u^t)_j - (\mathbb{K}u)_j}{t} \right\rangle.$$

Taking the limit as $t \to 0$, as well as the differentiability of the term $x/\|x\|$ in the inactive set, it yields

$$0 \leq \langle \nabla_{uu} \mathcal{F}(\lambda, u)\eta, v - \eta \rangle + \langle \nabla_u \mathcal{F}(h, u), v - \eta \rangle + \sum_{j \in \mathcal{I}} \left\langle \frac{(\mathbb{K}\eta)_j}{\|(\mathbb{K}u)_j\|} - \frac{(\mathbb{K}u)_j (\mathbb{K}u)_j^\top (\mathbb{K}\eta)_j}{\|(\mathbb{K}u)_j\|^3}, (\mathbb{K}v)_j - (\mathbb{K}\eta)_j \right\rangle.$$

Using the definition of T_j and recalling $v, \eta \in \mathcal{C}$, the inequality takes the form in (4.43). Now it is required to verify the uniqueness of the limit. For this purpose, let us note that (4.43) is a variational inequality of the first kind

$$\langle \nabla_{uu} \mathcal{F}(\lambda, u)\eta, v - \eta \rangle + \sum_{j \in \mathcal{I}} \langle T_j (\mathbb{K}\eta)_j, (\mathbb{K}v)_j - (\mathbb{K}\eta)_j \rangle \geq \langle -\nabla_u \mathcal{F}(h, u), v - \eta \rangle, \ \forall v, \eta \in \mathcal{C}.$$

Using the strong convexity of $\mathcal{F}$ and the positive semi-definiteness of T_j, the bilinear form on the left-hand side is V-elliptic, i.e.,

$$\langle \nabla_{uu} \mathcal{F}(\lambda, u)v, v \rangle + \sum_{j \in \mathcal{I}} \langle T_j (\mathbb{K}v)_j, (\mathbb{K}v)_j \rangle \geq c\|v\|^2, \text{ for some } c > 0.$$

Moreover, given that the right-hand side is linear and continuous with respect to $v - \eta$, we know by [34, Chapter I, Theorem 3.1] that there exists a unique solution for this variational inequality. □

Once we have demonstrated the Bouligand differentiability of the solution operator and the corresponding characterization of its directional derivative, described in this section, we arrive at the following result.

Theorem 4.5. *Let $\lambda^* \in \mathbb{R}^n_+$ be a local optimal solution of (4.28) and $u^* = S(\lambda^*)$. Then λ^* is a B-stationary point, i.e., it satisfies the following inequality*

$$\langle \nabla J(u^*), S'(\lambda^*; \lambda - \lambda^*) \rangle \geq 0, \quad \forall \lambda \in \mathbb{R}^n_+, \tag{4.44}$$

where $S'(\lambda^*; \lambda - \lambda^*) =: \eta$ is the unique solution to (4.43) with $h = \lambda - \lambda^*$.

Proof. Since we know that the solution operator is directionally differentiable, as shown

in theorem 4.4, along with its local Lipschitz continuity as demonstrated in theorem 4.2, we have that the solution operator is Bouligand differentiable. Consequently, a local optimal solution λ^* for problem (4.28) and $u^* = S(\lambda^*)$ its optimal reconstruction, satisfy the necessary optimality condition (4.30). □

4.2.2 Strict Complementarity

The characterization of the directional differentiability can take a different formulation if the biactive set is empty, i.e., $\mathcal{B} = \emptyset$. Then, the solution operator has stronger differentiability properties.

Theorem 4.6. *Assuming the index set $\mathcal{B}$ is empty. Then, the solution operator for the lower level problem* (4.1b) *is Fréchet differentiable, and the derivative can be computed as the solution of the following system of equations*

$$\nabla_{uu}\mathcal{F}(\lambda,u)\eta + \nabla_u\mathcal{F}(h,u) + \mathbb{K}^\top\xi = 0, \tag{4.45a}$$

$$\xi_j - T_j(\mathbb{K}\eta)_j = 0, \qquad \forall j \in \mathcal{I}, \tag{4.45b}$$

$$(\mathbb{K}\eta)_j = 0, \qquad \forall j \in \mathcal{A}_s. \tag{4.45c}$$

Proof. Using the empty biactive set assumption, we get that the cone $\mathcal{C}$ becomes the following linear subspace $\mathcal{C} = \{v \in \mathbb{R}^n : (\mathbb{K}v)_j = 0 \text{ if } (\mathbb{K}u)_j = 0\}$. Thus, the variational inequality (4.43) becomes the following variational equation

$$\langle \nabla_{uu}\mathcal{F}(\lambda,u)\eta + \nabla_u\mathcal{F}(h,u), v-\eta\rangle + \sum_{j\in\mathcal{I}}\langle T_j(\mathbb{K}\eta)_j, (\mathbb{K}v)_j - (\mathbb{K}\eta)_j\rangle = 0,\ \forall v \in \mathcal{C}. \tag{4.46}$$

This guarantees that the solution operator's directional derivative is a linear mapping w.r.t the direction h. Since S is Bouligand differentiable, it implies its Fréchet differentiability [74, Proposition 3.1.2]. Furthermore, (4.46) is the necessary and (due to convexity) sufficient optimality condition of the following optimization problem

$$\min_{\eta\,\in\,\mathcal{C}} \quad \frac{1}{2}\langle \nabla_{uu}\mathcal{F}(\lambda,u)\eta,\eta\rangle + \langle \nabla_u\mathcal{F}(h,u),\eta\rangle + \sum_{j\in\mathcal{I}}\left(\frac{\|(\mathbb{K}\eta)_j\|^2}{\|(\mathbb{K}u)_j\|} - \frac{\langle(\mathbb{K}u)_j,(\mathbb{K}\eta)_j\rangle^2}{\|(\mathbb{K}u)_j\|^3}\right). \tag{4.47}$$

Since all constraints are linear, the Abadie constraint qualification condition [33, Definition 2.33] is satisfied. Then, there exist Lagrange multipliers $\nu_j \in \mathbb{R}^2$, such that the KKT-optimality conditions for (4.47) look as follows

$$\langle \nabla_{uu}\mathcal{F}(\lambda,u)\eta, v\rangle + \langle \nabla_u\mathcal{F}(h,u), v\rangle + \sum_{j\in\mathcal{I}}\langle T_j(\mathbb{K}\eta)_j, (\mathbb{K}v)_j\rangle + \sum_{j\in\mathcal{A}_s}\langle \nu_j, (\mathbb{K}v)_j\rangle = 0,\ \forall v \in \mathbb{R}^n$$

$$(\mathbb{K}\eta)_j = 0, \forall j \in \mathcal{A}_s.$$

Finally, by introducing $\xi \in \mathbb{R}^{m\times 2}$ as

$$\xi_j = \begin{cases} \nu_j, & \forall j \in \mathcal{A}_s, \\ T_j(\mathbb{K}\eta)_j, & \forall j \in \mathcal{I} \end{cases}$$

the result is obtained. □

4.2.3 Bouligand Subdifferential of the Solution Operator

Even though the Bouligand stationarity condition presented in Section 4.2 holds for any local optimal solution, without requiring any constraint qualification, its purely primal form is in general not amenable for algorithmic purposes. Indeed, this limitation is related to the non-linearity of the directional derivative. As a remedy, in this section, we will focus on studying the Bouligand subdifferential of the solution operator S. Characterizing the linear elements of this subdifferential is helpful when devising a numerical algorithm to solve the bilevel problem (4.1).

Thanks to the local Lipschitz continuity of S, shown in Theorem 4.2, and Rademacher's theorem, we know that the solution operator is differentiable almost everywhere. Moreover, we will denote the set of points where this function is differentiable as D_S.

The following result characterizes the elements of the Bouligand subdifferential of the solution operator.

Theorem 4.7. *Let $G \in \partial_B S(\lambda)$ with $\lambda > 0$ and let us introduce the following subspace*

$$V := \{v \in \mathbb{R}^n : \ (\mathbb{K}v)_j = 0, \ \forall j \in \mathcal{A}_s \cup \mathcal{B}_1; (\mathbb{K}v)_j \in \operatorname{span}(q_j), \forall j \in \mathcal{B}_2\}. \tag{4.48}$$

Then, there exists a partition of the biactive set $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$ such that, for any $h \in \mathbb{R}^n$ such that $\lambda + th \geq 0$, $Gh =: \tilde{\eta} \in V$ is the unique solution of the system

$$\langle \nabla_{uu}\mathcal{F}(\lambda, u)\tilde{\eta}, v\rangle + \langle \nabla_u \mathcal{F}(h, u), v\rangle + \sum_{j\in\mathcal{I}} \langle \tilde{\xi}_j, (\mathbb{K}v)_j\rangle = 0, \qquad \forall v \in V \tag{4.49a}$$

$$\tilde{\xi}_j - T_j(\mathbb{K}\tilde{\eta})_j = 0, \qquad \forall j \in \mathcal{I}. \tag{4.49b}$$

Proof. To derive the characterization (4.49), we will make use of the local Lipschitz continuity of the solution operator, as described in theorem 4.2. This property implies that this map is almost everywhere differentiable. Therefore, let us consider a sequence $\{\lambda_k\} \subset D_S$ such that $\lambda_k \to \lambda$ and $S'(\lambda_k) \to G$. Since we have that $\lambda > 0$, for a sufficiently large k, it holds $\lambda_k > 0$. Using this result, along with the continuity of

$\nabla_u \mathcal{F}$, the following limits hold true

$$u_k = S(\lambda_k) \to S(\lambda) = u, \tag{4.50a}$$

$$\mathbb{K}^\top q_k = -\nabla_u \mathcal{F}(\lambda_k, u_k) \to -\nabla_u \mathcal{F}(\lambda, u) = \mathbb{K}^\top q. \tag{4.50b}$$

Now, each of this subsequence elements (u_k, q_k) define their respective inactive $\mathcal{I}^k := \mathcal{I}(u_k)$ and strongly active $\mathcal{A}_s^k := \mathcal{A}_s(u_k)$ sets. Moreover, from (4.50), we deduce the existence of an $N \in \mathbb{N}$ such that

$$\mathcal{I} \subset \mathcal{I}^k \quad \text{and } \mathcal{A}_s \subset \mathcal{A}_s^k \quad \forall n \geq N.$$

Then, by introducing the subspace $V^k := \{v \in \mathbb{R}^n : (\mathbb{K}v)_j = 0, \forall j \in \mathcal{A}_s^k\}$ and since $\{\lambda_k\} \subset D_S$. It then follows that, for $h \in \mathbb{R}^n$, we have the directional derivative of the solution operator in direction h, i.e., $S'(\lambda_k)h =: \eta_k \in V^k$ satisfies the system (4.45), as detailed below

$$\nabla_{uu} \mathcal{F}(\lambda_k, u_k)\eta_k + \nabla_u \mathcal{F}(h, u_k) + \mathbb{K}^\top \xi_k = 0, \tag{4.51a}$$

$$(\xi_k)_j - (T_k)_j (\mathbb{K}\eta_k)_j = 0, \ \forall j \in \mathcal{I}^k, \tag{4.51b}$$

$$(\mathbb{K}\eta_k)_j = 0, \ \forall j \in \mathcal{A}_s^k, \tag{4.51c}$$

or equivalently

$$\langle \nabla_{uu} \mathcal{F}(\lambda_k, u_k)\eta_k, v\rangle + \langle \nabla_u \mathcal{F}(h, u_k), v\rangle + \sum_{j \in \mathcal{I}^k} \langle (T_k)_j (\mathbb{K}\eta_k)_j, (\mathbb{K}v)_j \rangle = 0, \ \forall v \in V^k. \tag{4.52}$$

Now, to obtain (4.49) we have to apply the limit as $k \to \infty$ in (4.52). Even though from the definition of the Bouligand subdifferential, it follows that $\tilde{\eta} = \lim_{k\to\infty} \eta_k$, we need to guarantee the boundedness of the sequence $\{\xi_k\}$ for the limit to be well defined. In this spirit, for $j \in \mathcal{I}^k$, the sequence $\{(\xi_k)_j\}$ satisfies $\|(\xi_k)_j\| \leq \|(T_k)_j\|_2 \|(\mathbb{K}\eta_k)_j\|$, where $\|\cdot\|_2$ is the matrix norm consistent with the Euclidean norm. Moreover, using the definition for $(T_k)_j$, we have

$$\begin{aligned}
\|(\xi_k)_j\| &\leq \left\| \frac{I}{\|(\mathbb{K}u_k)_j\|} - \frac{(\mathbb{K}u_k)_j (\mathbb{K}u_k)_j^\top}{\|(\mathbb{K}u_k)_j\|^3} \right\|_2 \|(\mathbb{K}\eta_k)_j\|, \\
&= \frac{\|(\mathbb{K}\eta_k)_j\|}{\|(\mathbb{K}u_k)_j\|} \left\| I - \frac{(\mathbb{K}u_k)_j (\mathbb{K}u_k)_j^\top}{(\mathbb{K}u_k)_j^\top (\mathbb{K}u_k)_j} \right\|_2, \\
&\leq \frac{\|(\mathbb{K}\eta_k)_j\|}{\|(\mathbb{K}u_k)_j\|} \left\| I - \frac{(\mathbb{K}u_k)_j (\mathbb{K}u_k)_j^\top}{(\mathbb{K}u_k)_j^\top (\mathbb{K}u_k)_j} \right\|_F
\end{aligned}$$

Furthermore, for the given form of the matrix inside the Frobenius norm, we may

bound this norm using [32, Lemma 11.15] and obtain the following bound

$$\|(\xi_k)_j\| \leq \frac{3\sqrt{2}}{4} \frac{\|(\mathbb{K}\eta_k)_j\|}{\|(\mathbb{K}u_k)_j\|}, \ \forall k.$$

This result implies the boundedness of the sequence $\{(\xi_k)_j\}$, then there exists a subsequence that converges to a limit point $\tilde{\xi}_j$. Likewise, for $j \in \mathcal{A}_s^k$ it holds $(\mathbb{K}\eta_k)_j = (\mathbb{K}u_k)_j = 0$. Therefore, up to a subsequence, by passing to the limit, we get

$$\begin{aligned} \tilde{\xi}_j - T_j(\mathbb{K}\tilde{\eta})_j = 0, \ \forall j \in \mathcal{I}, \\ (\mathbb{K}\tilde{\eta})_j = 0, \ \forall j \in \mathcal{A}_s. \end{aligned}$$

Let us now consider a partition of the biactive set $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$, with

$$\mathcal{B}_1 := \{j \in \mathcal{B} : \exists \{u_{k_l}\} : (\mathbb{K}u_{k_l})_j = 0, \forall l\} \quad \text{and} \quad \mathcal{B}_2 := \mathcal{B} \backslash \mathcal{B}_1.$$

In the index set $\mathcal{B}_1$ we know that $(\mathbb{K}u_{k_l})_j = 0, \ \forall l$, i.e., the components are strongly active. Consequently, from (4.51c), it follows that the subsequence $(\mathbb{K}\eta_{k_l})_j = 0$, for all l. Since $\eta_k \to \tilde{\eta}$, we get that

$$(\mathbb{K}\tilde{\eta})_j = 0, \ \forall j \in \mathcal{A}_s \cup \mathcal{B}_1.$$

Considering the partition $\mathcal{B}_2$, we approach a biactive point using a sequence of points such that $(\mathbb{K}u_{k_l})_j \neq 0$, i.e., $j \in \mathcal{I}^k$. Now, taking $(\xi_k)_j = (T_k)_j(\mathbb{K}\eta_k)_j$ for $j \in \mathcal{I}^k$ we get

$$\langle (\xi_k)_j, (\mathbb{K}\eta_k)_j \rangle = \frac{1}{\|(\mathbb{K}u_k)_j\|} \left(\|(\mathbb{K}\eta_k)_j\|^2 - \frac{\langle (\mathbb{K}\eta_k)_j, (\mathbb{K}u_k)_j \rangle^2}{\|(\mathbb{K}u_k)_j\|^2} \right) \geq 0, \ \forall j \in \mathcal{I}^k. \quad (4.53)$$

Using the positivity of the term $\langle (\xi_k)_j, (\mathbb{K}\eta_k)_j \rangle$ we have

$$0 \leq \langle (\xi_k)_j, (\mathbb{K}\eta_k)_j \rangle \leq \sum_{j \in \mathcal{I}^k} \langle (\xi_k)_j, (\mathbb{K}\eta_k)_j \rangle. \quad (4.54)$$

Furthermore, using the semi-positive definiteness of $\nabla_{uu}\mathcal{F}(\lambda_k, u_k)$, we may upper bound (4.52) for $v = \eta_k$, as follows

$$\begin{aligned} \sum_{j \in \mathcal{I}^k} \langle (\xi_k)_j, (\mathbb{K}\eta_k)_j \rangle &\leq \langle \nabla_{uu}\mathcal{F}(\lambda_k, u_k)\eta_k, \eta_k \rangle + \sum_{j \in \mathcal{I}^k} \langle (T_k)_j (\mathbb{K}\eta_k)_j, (\mathbb{K}\eta_k)_j \rangle, \\ &= -\langle \nabla_u \mathcal{F}(h, u_k), \eta_k \rangle, \\ &\leq \|\nabla_u \mathcal{F}(h, u_k)\| \|\eta_k\|. \end{aligned} \quad (4.55)$$

Consequently, joining bounds (4.53)–(4.55), it reads

$$0 \leq \langle (\xi_k)_j, (\mathbb{K}\eta_k)_j \rangle \leq \langle \nabla_{uu}\mathcal{F}(\lambda_k, u_k)\eta_k, \eta_k \rangle + \sum_{j \in \mathcal{I}^k} \langle (\xi_k)_j, (\mathbb{K}v)_j \rangle \leq \|\nabla_u \mathcal{F}(h, u_k)\| \|\eta_k\|, \tag{4.56}$$

which, since $\eta_k \to \tilde{\eta}$, as $k \to \infty$, implies that $\langle (\xi_k)_j, (\mathbb{K}\eta_k)_j \rangle$ is uniformly bounded. Moreover, multiplying (4.56) with $\|(\mathbb{K}u_k)_j\|$, it reads

$$0 \leq \|(\mathbb{K}u_k)_j\| \langle (T_k)_j (\mathbb{K}\eta_k)_j, (\mathbb{K}\eta_k)_j \rangle \leq \|\nabla_u \mathcal{F}(h, u_k)\| \|\eta_k\| \|(\mathbb{K}u_k)_j\|.$$

Replacing the form for the term $(T_k)_j$ and the property for $(q_k)_j = (\mathbb{K}u_k)_j / \|(\mathbb{K}u_k)_j\|$ for $j \in \mathcal{I}^k$, we get

$$0 \leq \|(\mathbb{K}\eta_k)\|^2 - \langle (q_k)_j, (\mathbb{K}\eta_k) \rangle^2 \leq \|\nabla_u \mathcal{F}(h, u_k)\| \|\eta_k\| \|(\mathbb{K}u_k)_j\|.$$

Since for $j \in \mathcal{B}_2$ we know $(\mathbb{K}u_k)_j \to 0$, we get that the limit as $k \to \infty$ reads

$$\|(\mathbb{K}\tilde{\eta})_j\|^2 - \langle q_j, (\mathbb{K}\tilde{\eta})_j \rangle^2 = \lim_{k \to \infty} \|(\mathbb{K}\eta_k)_j\|^2 - \langle (q_k)_j, (\mathbb{K}\eta_k)_j \rangle^2 = 0.$$

which implies that $(\mathbb{K}\tilde{\eta})_j \in \operatorname{span}(q_j)$ for all $j \in \mathcal{B}_2$. Consequently, we have shown that $\tilde{\eta} \in V$.

Now, when taking the limit as $k \to \infty$ in (4.52), there may exist sequences in $\mathcal{I}^k$ that converge to a component in $\mathcal{B}_2$. Therefore, let us take a $v \in V$ and find the limit for the following term

$$\begin{aligned}
\lim_{k \to \infty} \langle (T_k)_j (\mathbb{K}\eta_k)_j, (\mathbb{K}v)_j \rangle &= \lim_{k \to \infty} \langle (T_k)_j (\mathbb{K}\eta_k)_j, c(q_k)_j \rangle, \\
&= \lim_{k \to \infty} \left\langle (T_k)_j (\mathbb{K}\eta_k)_j, c \frac{(\mathbb{K}u_k)_j}{\|(\mathbb{K}u_k)_j\|} \right\rangle, \\
&= c \lim_{k \to \infty} \left\langle \frac{(\mathbb{K}\eta_k)_j}{\|(\mathbb{K}u_k)_j\|}, \frac{(\mathbb{K}u_k)_j}{\|(\mathbb{K}u_k)_j\|} \right\rangle \\
&\quad - \left\langle \frac{\langle (\mathbb{K}\eta_k)_j, (\mathbb{K}u_k)_j \rangle (\mathbb{K}u_k)_j}{\|(\mathbb{K}u_k)_j\|^3}, \frac{(\mathbb{K}u_k)_j}{\|(\mathbb{K}u_k)_j\|} \right\rangle, \\
&= c \lim_{k \to \infty} \frac{\langle (\mathbb{K}\eta_k)_j, (\mathbb{K}u_k)_j \rangle}{\|(\mathbb{K}u_k)_j\|^2} - \frac{\langle (\mathbb{K}\eta_k)_j, (\mathbb{K}u_k)_j \rangle \|(\mathbb{K}u_k)_j\|^2}{\|(\mathbb{K}u_k)_j\|^4}, \\
&= 0.
\end{aligned}$$

Consequently, we may see that this product's limit vanishes for sequences coming from components either from $\mathcal{A}_s \cup \mathcal{B}_1$, where $(\mathbb{K}v)_j = 0$, and from $\mathcal{B}_2$ as $k \to \infty$. Therefore, taking the limit as $k \to \infty$ in (4.52), yields the result. □

Corollary 4.1. *Let $G \in \partial_B S(\lambda)$. There exists a partition of the biactive set $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$ and a multiplier $\theta \in \mathbb{R}^m$ such that, for any h such that $\lambda + th \geq 0$, $\tilde{\eta} := Gh$ is the unique solution of the system*

$$\nabla_{uu}\mathcal{F}(\lambda, u)\tilde{\eta} + \nabla_u \mathcal{F}(h, u) + \mathbb{K}^\top \theta = 0 \tag{4.57a}$$

$$\theta_j - T_j(\mathbb{K}\tilde{\eta})_j = 0, \qquad \forall j \in \mathcal{I}, \tag{4.57b}$$

$$\langle \theta_j, q_j \rangle = 0, \qquad \forall j \in \mathcal{B}_2. \tag{4.57c}$$

Proof. Let us consider the functional $M \in \mathbb{R}^n$ defined by

$$(M, v) := (\nabla_{uu}\mathcal{F}(\lambda, u)\tilde{\eta}, v) + (\nabla_u \mathcal{F}(h, u), v) + \sum_{j \in \mathcal{I}} \langle T_j(\mathbb{K}\tilde{\eta})_j, (\mathbb{K}v)_j \rangle, \ \forall v \in V.$$

(4.49a) can then be written as $M \in V^\perp$, where $V^\perp$ is the orthogonal complement of the set V defined in (4.48). Thanks to the structure of the linear subspace V, it can be represented in a separate way as $V = \left(\bigcap_{j \in \mathcal{A}_s \cup \mathcal{B}_1} V_j^1\right) \cap \left(\bigcap_{j \in \mathcal{B}_2} V_j^2\right)$, where

$$V_j^1 := \{v \in \mathbb{R}^n : (\mathbb{K}v)_j = 0\}, \qquad j \in \mathcal{A}_s \cup \mathcal{B}_1,$$
$$V_j^2 := \{v \in \mathbb{R}^n : (\mathbb{K}v)_j \in \operatorname{span}(q_j)\}, \qquad j \in \mathcal{B}_2.$$

Consequently, using the properties of the orthogonal complement described in Theorem 2.3, we know $V^\perp = \sum_{j \in \mathcal{A}_s \cup \mathcal{B}_1} (V_j^1)^\perp + \sum_{j \in \mathcal{B}_2} (V_j^2)^\perp$.

For $j \in \mathcal{A}_s \cup \mathcal{B}_1$, we get that $(V_j^1)^\perp = \ker(\mathbb{K}_j)^\perp$. Thanks to the orthogonality relations, see Theorem 2.4, it follows that $\ker(\mathbb{K}_j)^\perp = \operatorname{range}(\mathbb{K}_j^\top)$. Hence, for any $\xi_j \in (V_j^1)^\perp$, there exist π_j such that $\xi_j = \mathbb{K}_j^\top \pi_j$. Consequently,

$$\sum_{j \in \mathcal{A}_s \cup \mathcal{B}_1} (V_j^1)^\perp = \sum_{j \in \mathcal{A}_s \cup \mathcal{B}_1} \mathbb{K}_j^\top \pi_j, \quad \pi_j \in \mathbb{R}^2.$$

For $j \in \mathcal{B}_2$, any $v \in V_j^2$ can be represented as a sum between an element from the nullspace and an element from the row space of $\mathbb{K}_j$ (Theorem 2.2) as follows

$$v = \phi + \varphi, \quad \text{with } (\mathbb{K}_j\varphi) = 0 \text{ and } \phi \in \operatorname{range}(\mathbb{K}_j^\top).$$

Since $(\mathbb{K}v)_j \in \operatorname{span}(q_j)$ and $(\mathbb{K}_j\varphi) = 0$, it follows that $(\mathbb{K}v)_j \in \operatorname{span}(q_j)$ as well. Let us now consider $w_j \in (V_j^2)^\perp$, which can be represented as $w_j = \tilde{w}_j + \hat{w}_j$, where $\tilde{w}_j \in \operatorname{range}(\mathbb{K}_j^\top)$ and $\hat{w}_j \in \operatorname{range}(\mathbb{K}_j^\top)^\perp = \ker(\mathbb{K}_j)$. Consequently, there exists ψ_j such that

$$w_j = \mathbb{K}_j^\top \psi_j + \hat{w}_j, \quad \text{with } \mathbb{K}_j \hat{w}_j = 0.$$

Taking the scalar product with $v_j \in V_j^2$, we get

$$(w_j, v_j) = (\mathbb{K}_j^\top \psi_j + \hat{w}_j, \phi + \varphi) = \langle \psi_j, \mathbb{K}_j \phi \rangle + (\hat{w}_j, \mathbb{K}_j^\top \psi) + (\hat{w}_j, \varphi) = c\langle \psi_j, q_j \rangle + (\hat{w}_j, \varphi),$$

since $\mathbb{K}_j \varphi = \mathbb{K}_j \hat{w}_j = 0$. For the product to be zero, it is then required that $(\hat{w}_j, \varphi) = 0, \forall \varphi \in \ker(\mathbb{K}_j)$ and $\langle \psi_j, q_j \rangle = 0$. Since $\hat{w}_j$ belongs to $\ker(\mathbb{K}_j)$ as well, it follows that $\hat{w}_j = 0$. Consequently,

$$\sum_{j \in \mathcal{B}_2} (V_j^2)^\perp = \sum_{j \in \mathcal{B}_2} \mathbb{K}_j^\top \psi_j, \quad \psi_j \in \mathbb{R}^2 : \langle \psi_j, q_j \rangle = 0.$$

Altogether, we then obtain that there exist multipliers π_j and ψ_j such that

$$M + \sum_{j \in \mathcal{A}_s \cup \mathcal{B}_1} \mathbb{K}_j^\top \pi_j + \sum_{j \in \mathcal{B}_2} \mathbb{K}_j^\top \psi_j = 0,$$

with $\langle \psi_j, q_j \rangle = 0$. Defining

$$\theta_j := \begin{cases} T_j (\mathbb{K}\tilde{\eta})_j, & j \in \mathcal{I}, \\ \pi_j, & j \in \mathcal{A}_s \cup \mathcal{B}_1, \\ \psi_j, & j \in \mathcal{B}_2, \end{cases}$$

the result is obtained. □

Next, we verify that along a given direction, there exists a solution of system (4.49) which coincides with the directional derivative. When properly characterized, we can use a linear representative of the (otherwise nonlinear) directional derivative within a solution algorithm.

Theorem 4.8. *For any $\lambda \in \mathbb{R}^n_+$ and $h \in \mathbb{R}^n$ such that $\lambda + th \geq 0$, there exists a linearized element $\tilde{\eta} = Gh$ such that $S'(\lambda; h) = Gh$.*

Proof. Let us recall that the directional derivative of the solution mapping, in direction h, is given by the unique solution of the following variational inequality of the first kind

$$\langle \nabla_{uu} \mathcal{F}(\lambda, u)\eta + \nabla_u \mathcal{F}(h, u), v - \eta \rangle + \sum_{j \in \mathcal{I}} \langle T_j (\mathbb{K}\eta)_j, (\mathbb{K}v)_j - (\mathbb{K}\eta)_j \rangle \geq 0, \ \forall v \in \mathcal{C}. \quad (4.58)$$

Considering the sets $\mathcal{B}_1 := \{j \in \mathcal{B} : (\mathbb{K}\eta)_j = 0\}$ and $\mathcal{B}_2 := \mathcal{B} \backslash \mathcal{B}_1$, and since $\eta \in \mathcal{C}$, it also follows $(\mathbb{K}\eta)_j = cq_j$ for all $j \in \mathcal{B}_2$, for some $c > 0$. Consequently, η belongs to the

subspace

$$V := \{v \in \mathbb{R}^n : (\mathbb{K}v)_j = 0,\ \forall j \in \mathcal{A}_s \cup \mathcal{B}_1;\ (\mathbb{K}v)_j \in \operatorname{span}(q_j),\ \forall j \in \mathcal{B}_2\}.$$

Moreover, for any $w \in V$ and t sufficiently small, we have $\eta \pm tw \in \mathcal{C}$. Testing (4.58) with these vectors and using the linearity of $\mathbb{K}$, we get

$$\langle \nabla_{uu}\mathcal{F}(\lambda,u)\eta + \nabla_u \mathcal{F}(h,u), \eta + tw\rangle + \sum_{j\in\mathcal{I}} \langle T_j(\mathbb{K}\eta)_j, (\mathbb{K}\eta)_j + t(\mathbb{K}w)_j\rangle = 0,$$
$$\langle \nabla_{uu}\mathcal{F}(\lambda,u)\eta + \nabla_u \mathcal{F}(h,u), tw - \eta\rangle + \sum_{j\in\mathcal{I}} \langle T_j(\mathbb{K}\eta)_j, t(\mathbb{K}w)_j - (\mathbb{K}\eta)_j\rangle = 0.$$

Adding both equations, yields

$$\langle \nabla_{uu}\mathcal{F}(\lambda,u)\eta + \nabla_u \mathcal{F}(h,u), w\rangle + \sum_{j\in\mathcal{I}} \langle T_j(\mathbb{K}\eta)_j, (\mathbb{K}w)_j\rangle = 0,\ \forall w \in V.$$

Indeed, the directional derivative takes the form $\eta = Gh$ solution of (4.49), with $\mathcal{B}_2$ as defined above. □

4.3 Nonsmooth Trust Region Algorithm

In this section, we will use the characterization of the linear elements of the Bouligand subdifferential to find optimal denoising parameters for the data learning model (4.1). The directional derivative of the reduced cost problem can be written as

$$\langle j'(\lambda), h\rangle = \langle \nabla J(u), S'(\lambda;h)\rangle = \langle \nabla J(u), \tilde{\eta}\rangle, \tag{4.59}$$

where $\tilde{\eta}$ is the solution of (4.57) for a particular partition of the biactive set $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$ and $S'(\lambda;h)$ is the directional derivative of the solution operator at λ in direction h.

By defining a *generalized adjoint* $p \in \mathbb{R}^n$ as the solution of the following system

$$\begin{aligned} \langle \nabla_{uu}\mathcal{F}(\lambda,u)^\top p, v\rangle + \sum_{j\in\mathcal{I}} \langle \mu_j, (\mathbb{K}v)_j\rangle - \langle \nabla J(u), v\rangle = 0, \qquad & \forall v \in V \\ \mu_j - T_j(\mathbb{K}p)_j = 0, \qquad & \forall j \in \mathcal{I}, \end{aligned}$$

where V is defined as in Theorem 4.7. Thanks to Theorem 4.8 we know that $\tilde{\eta} \in V$ is a linear representative of the directional derivative of the solution operator. Consequently (4.59) reads

$$\langle j'(\lambda), h\rangle = \langle \nabla J(u), \tilde{\eta}\rangle = \langle \nabla_{uu}\mathcal{F}(\lambda,u)p, \tilde{\eta}\rangle + \sum_{j\in\mathcal{I}} \langle T_j(\mathbb{K}p)_j, (\mathbb{K}\tilde{\eta})_j\rangle.$$

Using the symmetry of T_j and (4.49a) it yields

$$\begin{aligned}\langle j'(\lambda), h\rangle &= \langle \nabla_{uu}\mathcal{F}(\lambda,u)^\top p, \tilde{\eta}\rangle + \sum_{j\in\mathcal{I}} \langle (\mathbb{K}p)_j, T_j(\mathbb{K}\tilde{\eta})_j\rangle, \\ &= \langle \nabla_{uu}\mathcal{F}(\lambda,u)^\top p, \tilde{\eta}\rangle + \sum_{j\in\mathcal{I}} \langle (\mathbb{K}p)_j, \tilde{\xi}_j\rangle,\end{aligned}$$

which gives us the following characterization for the directional derivative

$$\langle j'(\lambda), h\rangle = \langle g, h\rangle = -\langle \nabla_u \mathcal{F}(h,u), p\rangle. \tag{4.60}$$

With (4.60) it is now possible to define a trust region algorithm, see Section 2.7, for solving the bilevel problem.

We will use the nonsmooth trust-region method described in Section 2.7. Recalling this method is built using two switching models, see Algorithm 4.1, the implementation will consider two models as well. A first model will consider the linear representative of the Bouligand subdifferential described in Theorem 4.7 for a particular partition of the biactive set $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$ and its associated generalized adjoint state as well as its reduced cost directional derivative representative (4.60), yielding the following model

$$m_k(\lambda_k + d_k) = j(\lambda_k) - \langle \nabla_u \mathcal{F}(d_k, u_k), p_k\rangle + \frac{1}{2}\langle d_k, B_k d_k\rangle,$$

where B_k is a BFGS second-order approximation of the Hessian matrix.

Furthermore, when the trust region radius falls below a threshold radius Δ_t, the algorithm switches to a model built using a regularized version of the problem as described in Section 3.5. In particular, considering the KKT optimality system for the smooth bilevel problem (3.18) with $\alpha_j = 1$. Then, we may obtain a *regularized adjoint* $p_\gamma \in \mathbb{R}^n$ as the solution of the following system

$$\begin{aligned}\nabla_u J(u) + \nabla_{uu}\mathcal{F}(\lambda,u)^\top p_\gamma + \mathbb{K}^\top \beta &= 0, \\ \beta_j - h_\gamma''((\mathbb{K}u)_j)^\top (\mathbb{K}p_\gamma)_j &= 0, \qquad \forall j = 1,\dots,m.\end{aligned}$$

Using the obtained solution p_γ, we know the gradient of the reduced cost function corresponds to

$$\langle j'(\lambda), h\rangle = \langle g_\gamma, h\rangle = -\langle \nabla_{u\lambda}\mathcal{F}(\lambda,u)^\top p_\gamma, h\rangle, \tag{4.61}$$

With this result, we can now build a regularized model for the trust-region algorithm

$$m_k(\lambda_k + d_k) = j(\lambda_k) - \langle \nabla_{u\lambda}\mathcal{F}(\lambda_k, u_k)^\top p_{\gamma_k}, d_k\rangle + \frac{1}{2}\langle d_k, B_k d_k\rangle.$$

A complete description of the trust-region algorithm used for data parameter learning

is depicted in Algorithm 4.1.

Algorithm 4.1 Non-smooth Trust-Region for Learning the Data Fidelity Weight

Choose initial parameter λ_0, radius Δ_0, $0 < \eta_1 \leq \eta_2 < 1$, $0 < \gamma_1 \leq 1 \leq \gamma_2$ and $tol > 0$
Choose initial second order matrix B_0 and a threshold radius Δ_t
Compute $j(\lambda_0)$ and set $k = 0$.
while $\Delta_k > tol$ **do**
  if $\Delta_k >= \Delta_t$ **then**
    Compute a linear element of the Bouligand subdifferential g_k at λ_k as the solution of (4.60) for a particular partition of the biactive set $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$.
    Build the model function as: $m_k(\lambda_k + d_k) = j(\lambda_k) + g_k^\top d_k + \frac{1}{2} d_k^\top B_k d_k$.
  else
    Compute a regularized gradient $g_{\gamma,k}$ at λ_k using (4.61).
    Build the model function as: $m_k(\lambda_k + d_k) = j(\lambda_k) + g_{\gamma,k}^\top d_k + \frac{1}{2} d_k^\top B_k d_k$.
  end if
  Compute a step s_k that "sufficiently" reduces the model m_k such that $\lambda_k + s_k \in B_{\Delta_k}$
  Update second order matrix B_k using limited memory BFGS.
  Calculate the predicted and actual reduction

$$pred_k = m_k(\lambda_k) - m_k(\lambda_k + s_k),$$
$$ared_k = j(\lambda_k) - j(\lambda_k + s_k).$$

  Compute the quality measure $\rho_k = ared_k / pred_k$.
  $\lambda_{k+1} = \begin{cases} \lambda_k & \text{if } \rho_k \leq \eta_1, \\ \lambda_k + s_k & \text{otherwise.} \end{cases}$
  $\Delta_{k+1} = \begin{cases} \gamma_2 \Delta_k & \text{if } \rho_k \geq \eta_2, \\ \Delta_k & \text{if } \rho_k \in (\eta_1, \eta_2), \\ \gamma_1 \Delta_k & \text{if } \rho_k < \eta_1. \end{cases}$
  $k \leftarrow k + 1$
end while
return λ_k

4.4 Numerical Experiments

In this section, we will describe the algorithm's performance described in Section 4.3 to obtain optimal patch parameters for the variational image denoising problem. This problem is built with an upper-level loss corresponding to the following quadratic function

$$J(u, \bar{u}) := \frac{1}{N} \sum_{k=1}^{N} \|u_k - u_k^{\text{true}}\|^2,$$

and for the lower level problem, we will consider the patch parameter total variation denoising

$$u_k = \arg\min_u \sum_{j=1}^{n} \mathcal{Q}(\lambda)_j (u_j - (f_k)_j)^2 + \sum_{j=1}^{m} \|(\mathbb{K}u)_j\|,$$

where $(f_k, \bar{u}_k)$ is a given training dataset, $\lambda \in \mathbb{R}^p_+$ with $p << n$ and $\mathcal{Q} : \mathbb{R}^p \to \mathbb{R}^n_+$ is a linear patch operator defined as

$$\mathcal{Q}(\lambda) := \lambda_{\sqrt{p}\times\sqrt{p}} \otimes \mathbb{I}_{\frac{\sqrt{n}}{\sqrt{p}}\times\frac{\sqrt{n}}{\sqrt{p}}} \in \mathbb{R}^{\sqrt{n}\times\sqrt{n}},$$

here, $\otimes$ is the Kronecker product, $\lambda_{\sqrt{p}\times\sqrt{p}}$ is a matrix built by reordering the elements of λ into a matrix of size $\sqrt{p} \times \sqrt{p}$, and $\mathbb{I}_{\frac{\sqrt{n}}{\sqrt{p}}\times\frac{\sqrt{n}}{\sqrt{p}}}$ is a matrix of ones of size $\frac{\sqrt{n}}{\sqrt{p}} \times \frac{\sqrt{n}}{\sqrt{p}}$. This product outputs a matrix of size $\sqrt{n} \times \sqrt{n}$ that is reshaped into a vector of n components.

The trust-region algorithm used for solving this problem along with the lower level solvers used for the denoising problem were coded in `Python` making use of the libraries: `numpy`, `scipy`, `pillow`, `pylops` and `pyprox`. The source code and instructions for repeating the computations are provided in [83].

We divided the experiments of this bilevel learning problem into several parts. First, we will explore the optimal denoising results on the Cameraman training pair to verify the results in a generic natural image for both scalar and patch-dependent denoise models. Furthermore, we continue with a patch behavior exploration using a test image with a noise contaminating specific patches of the picture. After, we will finish this exploration with an application with a larger dataset; indeed, in Section 4.4.3 we will use our bilevel learning strategy to obtain optimal denoising patch-based parameters for a subset of the CelebA Faces dataset [50].

As a final comment on the implementation of the algorithm parameters, we consider a threshold radius $\Delta_t = 1 \times 10^{-4}$, the smoothing parameter on the regularized gradient $\gamma = 1 \times 10^{-10}$. Finally, regarding the initialization procedure, we rely on a warm-start method. This strategy involves solving the bilevel problem for a parameter with increasing patches and using the previous solution with the smaller patch size to set the initial parameter.

4.4.1 Cameraman Training Pair

Here, we analyze the algorithm's behavior with a single training pair based on the cameraman test image of size 128×128 pixels. This training pair was built by adding Gaussian noise with zero mean and standard deviation $\sigma = 0.05$ to the ground-truth image. When considering a scalar and a two-dimensional parameter space, Figure 4.3

Original	SSIM=0.7494 PSNR=23.1698	$\lambda^* = 64.36$ SSIM=0.8426 PSNR=23.8045	$\lambda^* = [46.3617; 90.8444]$ SSIM=0.8537 PSNR=23.8322

Figure 4.3: Optimal reconstructions of the cameraman training pair using a scalar regularization parameter and a 2 dimensional regularization parameter.

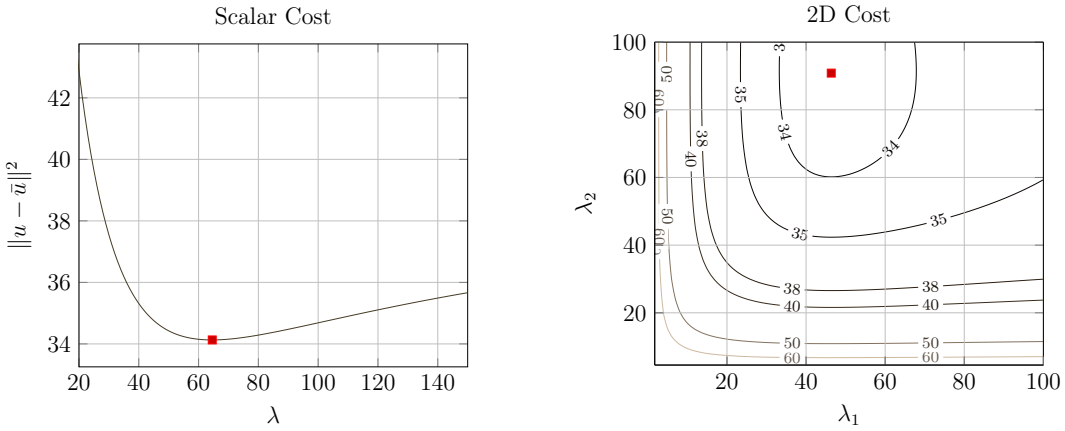


Figure 4.4: Values for the l_2 squared cost function using a scalar regularization parameter and a two dimensional regularization parameter using the Cameraman training pair.

shows this training image pair along with the optimal denoised images computed using the trust-region algorithm. An improvement on the SSIM quality metric can be verified when using a two-dimensional parameter compared to a scalar parameter optimal denoising. Furthermore, we plot the reduced cost function in Figure 4.4 for the scalar and two-dimensional parameter cases. As presented in the figure, the non-convexity of said function can be inferred.

The non-convexity of the reduced cost function leads to several limitations in our choice for the initial parameter considered by the algorithm. Indeed, Table 4.1 shows the obtained solution for the scalar case for different initialization values λ_0. The figure shows that the algorithm can find the optimal value within a region with good enough curvature information. Nevertheless, it is not the case for high values on the initial parameter value ($\lambda_0 = 80$). At this parameter value, the gradient of the reduced cost function satisfies the stopping criteria, and consequently, the algorithm stops its

					Reconstruction		
λ_0	nit	nfev	ngev	nreggev	COST	PSNR	SSIM
10	15	17	17	0	34.114227	23.804544	0.842666
20	14	16	15	1	34.114202	23.804547	0.842691
30	15	17	16	1	34.114088	23.804562	0.842909
40	11	13	12	1	34.114125	23.804557	0.842792
50	16	18	18	0	34.114088	23.804562	0.842923

Table 4.1: Changes on the initial parameter λ_0 regularization parameter for the cameraman training pair.

					Reconstruction		
patch	nit	nfev	ngev	nreggev	COST	PSNR	SSIM
1×1	13	15	14	1	34.114243	23.804542	0.842652
2×2	40	42	41	1	33.673749	23.860985	0.854657
4×4	43	45	45	0	33.017201	23.946497	0.859556
8×8	37	39	39	0	32.236732	24.050389	0.868040
16×16	30	32	32	0	31.175826	24.195720	0.869611
32×32	32	34	34	0	29.294562	24.466029	0.873243

Table 4.2: Trust Region Algorithm behavior on the cameraman training pair.

execution even though it is not a stationary point.

Regarding the performance of the algorithm in this dataset, Table 4.2 shows a summary of the number of iterations, number of function evaluations (nfev), number of gradient evaluations (ngev), and number of regularized gradient evaluations (nreggev). Indeed, we can see an improvement in the quality of the reconstruction obtained using the l_2, SSIM, and PSNR metrics; it implies that an increasing number of patches used in the parameter leads to a better reconstruction.

4.4.2 Circles Training Pair

This next experiment will explore the patch mechanism and its spatial adaptation. For this purpose, we built a synthetic dataset where four patches of noise with zero mean and variance of 0.1, see Figure 4.6. In this training pair, when using the trust-region algorithm for an increasing number of patches in the parameter λ, we see an increase in the reconstruction quality according to the l_2, PSNR, and SSIM metrics as detailed in Table 4.3. Furthermore, Figure 4.7 shows the spatial adaptation of the patches. Recalling that dark values in the figure corresponds to zones where the data term has less influence on the final solution, we see that it correlates with the original distribution of the noise in the test image.

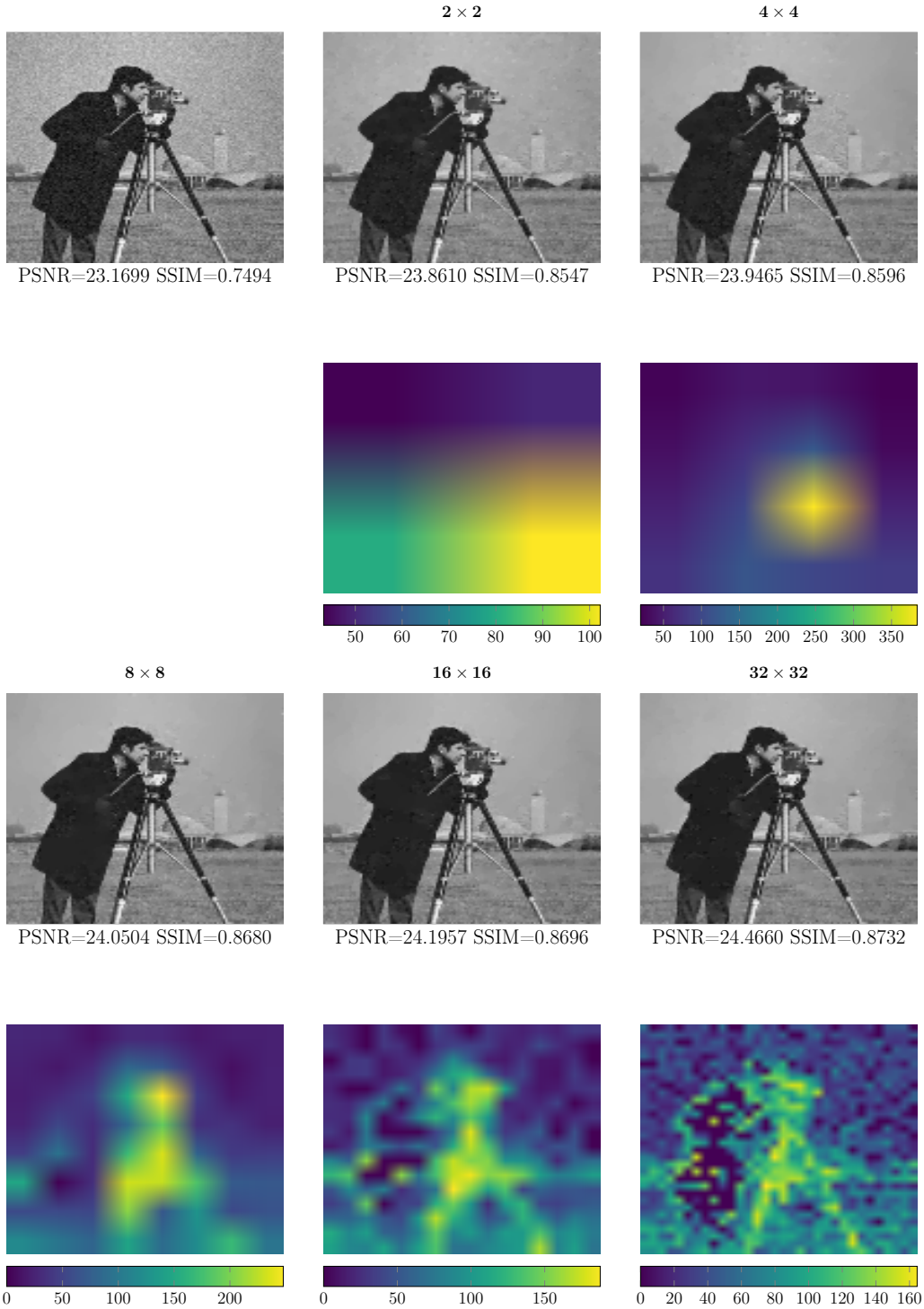


Figure 4.5: Learned optimal patch parameter for an increasing number of patches for the cameraman training pair.

Figure 4.6: Optimal Scalar Parameter Circles Training Pair

					Reconstruction		
patch	nit	nfev	ngev	nreggev	COST	PSNR	SSIM
1×1	12	14	13	1	76.902703	20.274483	0.763919
2×2	202	204	204	0	69.025173	20.743824	0.860534
4×4	14	16	15	1	47.882484	22.332133	0.856216
8×8	14	16	15	1	42.264715	22.874120	0.853280
16×16	19	21	21	0	34.827006	23.714738	0.844906
32×32	42	44	44	0	20.107619	26.100293	0.867166

Table 4.3: Trust Region Algorithm behavior on the circles training pair.

4.4.3 Multiple Training Pairs

For the second experiment, we used ten image pairs containing images of faces to generate a training dataset and ten different image pairs to generate a validation dataset; both datasets were based on the CelebA dataset [50]. Said images are of size 128 by 128 pixels, and in both datasets, we obtained the degenerated pairs by adding Gaussian noise with zero-mean and standard deviation $\sigma = 0.1$. A subset of the training dataset is depicted in Figure 4.8. In Figure 4.9, we plot the reduced cost functions corresponding to a scalar parameter and two-dimensional patch parameter, along with the optimal value calculated by the algorithm. Again, we can confirm experimentally that the optimal value was calculated.

According to the results presented in Figure 4.10, the learned parameter does not adjust to a particular image. Still, it changes according to the training set data. Furthermore, for the training set, Table 4.4 shows an improvement in the averaged quality of the obtained images as more patches are considered in the parameter. Moreover, the number of iterations used to get said solution is detailed, along with the number of function, gradient, and regularized gradient evaluations.

Finally, regarding the generalization capabilities of the learned parameter, we may see a degradation in the averaged quality of the reconstruction images in Table 4.5. This effect may indicate overfitting of the parameter when it uses a high number of

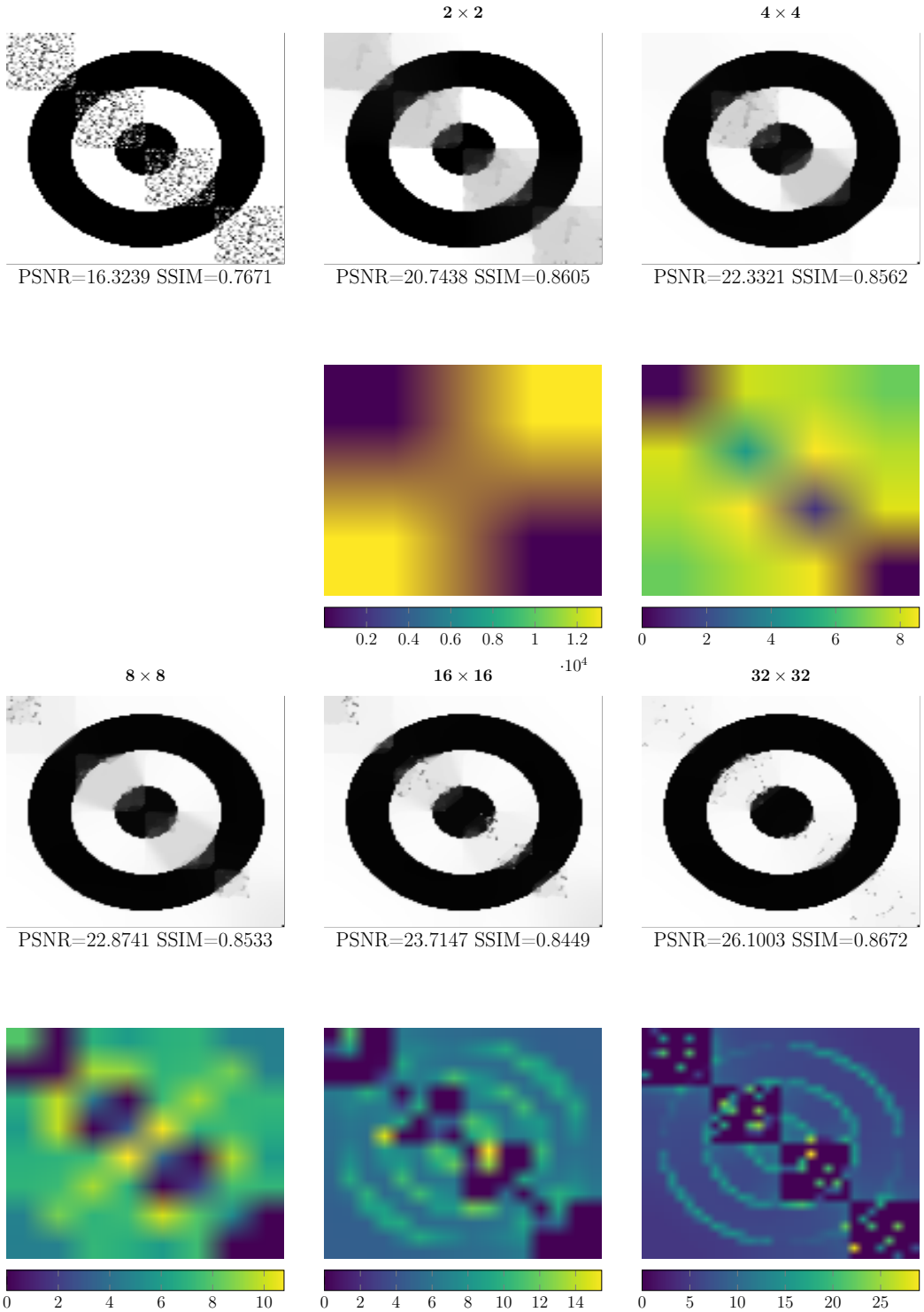


Figure 4.7: Learned optimal parameters for an increasing patch number on the circles training pair.

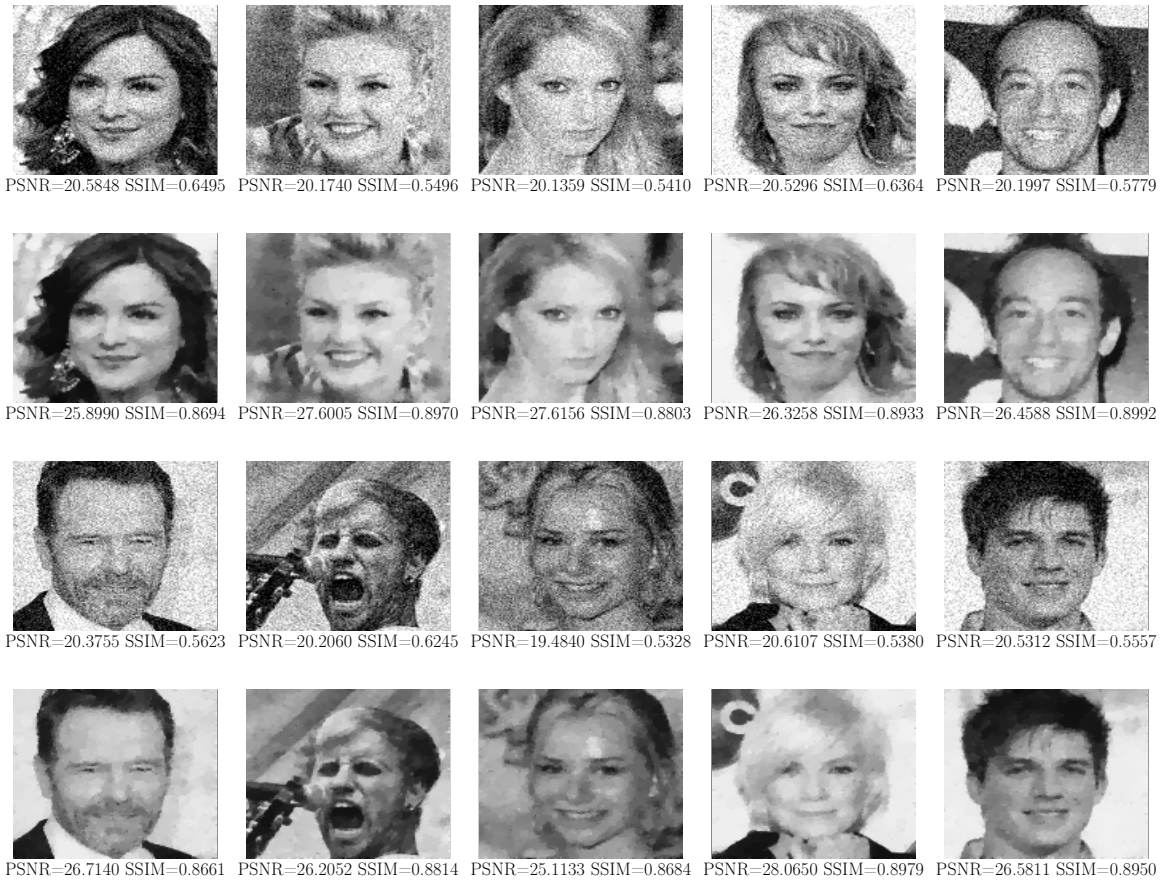


Figure 4.8: A subset of the CelebA dataset corrupted with Gaussian noise.

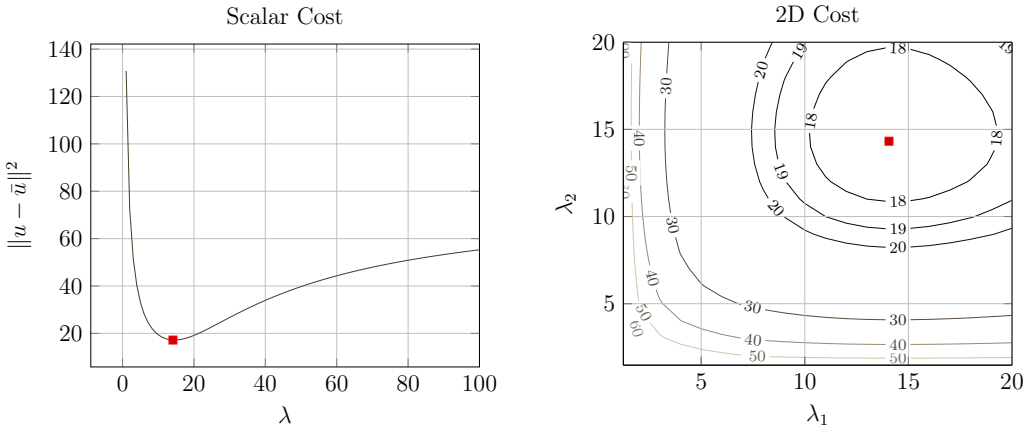


Figure 4.9: Values for the l_2 squared cost function using a scalar regularization parameter and a two dimensional regularization parameter using the faces dataset.

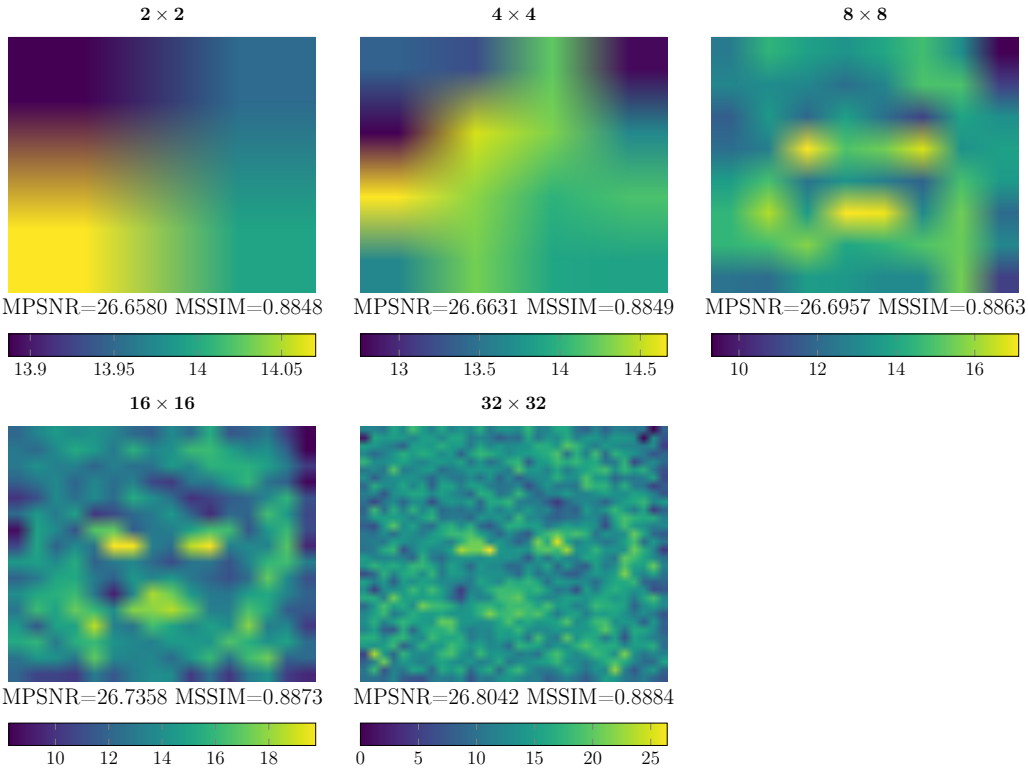


Figure 4.10: Values for the optimal parameters calculated for different parameter patch sizes on the faces dataset.

					Reconstruction		
patch	nit	nfev	ngev	nreggev	COST	MPSNR	MSSIM
1×1	7	9	8	1	18.020032	26.657829	0.884794
2×2	9	11	11	0	18.019781	26.657994	0.884803
4×4	9	11	11	0	18.001191	26.663121	0.884901
8×8	9	11	11	0	17.881697	26.695693	0.886269
16×16	14	16	15	1	17.716941	26.735823	0.887308
32×32	69	71	70	1	17.435898	26.804171	0.888396

Table 4.4: Trust Region Algorithm behavior on the Faces dataset.

num	noisy	scalar	2×2	4×4	8×8	16×16	32×32
1	0.6524	0.8748	0.8754	0.8752	0.8749	0.8745	0.8718
2	0.5840	0.7656	0.7681	0.7675	0.7654	0.7640	0.7579
3	0.5623	0.8668	0.8669	0.8668	0.8670	0.8667	0.8623
4	0.5350	0.8204	0.8207	0.8206	0.8205	0.8203	0.8160
5	0.5979	0.8737	0.8735	0.8735	0.8732	0.8728	0.8694
6	0.5807	0.8439	0.8444	0.8443	0.8443	0.8445	0.8427
7	0.5640	0.7460	0.7484	0.7478	0.7459	0.7444	0.7385
8	0.5631	0.8467	0.8470	0.8471	0.8473	0.8471	0.8441
9	0.5910	0.8354	0.8371	0.8365	0.8342	0.8316	0.8205
10	0.6622	0.8753	0.8765	0.8761	0.8749	0.8737	0.8696
MSSIM	**0.5892**	**0.8348**	**0.8358**	**0.8355**	**0.8348**	**0.8340**	**0.8293**

Table 4.5: Faces Dataset SSIM Quality Measures in the validation dataset.

patches. Consequently, according to this validation dataset, a 2×2 patch has the best generalization properties.

Chapter 5

Optimal Learning of the Regularization Weight

This section will focus on finding the optimal parameters for the lower level problem shown in (3.9) where we are only considering the parameter affecting the regularization term $\alpha \in \mathbb{R}^m_+$. With this goal in mind, we will rely on a bilevel parameter learning strategy where we make use of a training dataset of P pairs (u_k^{true}, f_k), for $k = 1, \ldots, P$, where each u_k^{true} corresponds to ground-truth data and f_k to the corresponding corrupted one, the optimization problem now reads

$$\min_{\alpha \in \mathbb{R}^n_+} \quad \sum_{k=1}^{P} J(u_k(\alpha), u_k^{\text{true}}) \tag{5.1a}$$

$$\text{s.t.} \qquad u_k(\alpha) = \arg\min_{u \in \mathbb{R}^n} \left\{ \mathcal{F}(u; f_k) + \sum_{j=1}^{m} \alpha_j \|(\mathbb{K}u)_j\|, \right\} \tag{5.1b}$$

where $\mathbb{K} : \mathbb{R}^n \to \mathbb{R}^{m \times 2}$ is a finite differences discretization of the gradient operator. Furthermore, we will analyze the single training pair case since the multiple training pairs case is a natural extension. Moreover, for readability purposes, we will omit the dependence of corrupted data f_k in the data fidelity term.

Existence of optimal solutions for the bilevel problem (5.1) has been previously investigated in [48] for the scalar parameter case. If the data (noisy image, ground truth), the fidelity term, and the loss function fulfill certain conditions, no box constraints are required to prove the existence of optimal scalar and scale-dependent parameters (see [48, Proposition 4.1]). The proof can be easily extended to the case of spatially-dependent parameters using similar arguments.

Now we can replace the lower level optimization problem in (5.1) by its necessary and sufficient condition, leading to the following optimization problem with variational

inequality constraints

$$\min_{\alpha \in \mathbb{R}^n_+} \quad J(u(\alpha), u^{\text{true}}) \tag{5.2a}$$

$$\text{s.t.} \quad \langle \nabla \mathcal{F}(u), v - u \rangle + \sum_{j=1}^{m} \alpha_j \|(\mathbb{K}v)_j\| - \sum_{j=1}^{m} \alpha_j \|(\mathbb{K}u)_j\| \geq 0, \ \forall v \in \mathbb{R}^n \tag{5.2b}$$

Using duality techniques [28], we know that there exists a dual variable $q \in \mathbb{R}^{m \times 2}$ with $q_j \in \partial(\alpha_j \|(\mathbb{K}u)_j\|)$, such that the variational inequality of the second kind (5.2b) can be equivalently written in primal-dual form, yielding the following reformulation

$$\nabla \mathcal{F}(u) + \mathbb{K}^\top q = 0 \tag{5.3a}$$

$$\langle q_j, (\mathbb{K}u)_j \rangle - \alpha_j \|(\mathbb{K}u)_j\| = 0, \qquad \forall j = 1, \dots, m \tag{5.3b}$$

$$\|q_j\| - \alpha_j \leq 0, \qquad \forall j = 1, \dots, m. \tag{5.3c}$$

Consequently, the bilevel parameter learning problem for a single training pair can be written as:

$$\min_{\alpha \in \mathbb{R}^m} \quad J(u, \bar{u}) \tag{5.4a}$$

$$\text{s.t.} \quad \nabla \mathcal{F}(u) + \mathbb{K}^\top q = 0, \tag{5.4b}$$

$$\langle q_j, (\mathbb{K}u)_j \rangle - \alpha_j \|(\mathbb{K}u)_j\| = 0, \qquad \forall j = 1, \dots, m, \tag{5.4c}$$

$$\|q_j\| - \alpha_j \leq 0, \qquad \forall j = 1, \dots, m, \tag{5.4d}$$

$$\alpha_j \geq 0, \qquad \forall j = 1, \dots, m. \tag{5.4e}$$

For clarity in the exposition, we restrict the analysis to the case of a single training pair. The results are, however, easily extendable to larger training sets.

5.1 Mordukhovich Stationarity

This section will address the primal-dual stationarity conditions for the bilevel problem (5.1). Then, motivated by the constraint qualification condition presented in Section 2.5, we can reformulate the lower level optimization problem in (5.1b) as a generalized equation. Then, by verifying the constraint qualification condition for generalized mathematical problems with equilibrium constraints (GMPEC) presented in Theorem 2.9, we can guarantee the existence of Lagrange multipliers and a corresponding stationarity system.

Indeed, by introducing a dual variable $q \in \mathbb{R}^{m \times 2}$, where $q_j \in \partial(\alpha_j \|(\mathbb{K}u)_j\|)$, we may

write the lower level problem equivalently as follows

$$0 \in \nabla \mathcal{F}(u) + Q(\alpha, u), \tag{5.5}$$

where $Q : \mathbb{R}^m_+ \times \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is the set-valued operator associated with the subdifferential of the Euclidean norm, i.e.,

$$Q(\alpha, u) := \left\{ \begin{array}{ll} \mathbb{K}^\top q, \text{ with } q \in \mathbb{R}^{m\times 2} : \begin{cases} q_j = \alpha_j \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, & \text{if } \|(\mathbb{K}u)_j\| \neq 0, \\ \|q_j\| \leq \alpha_j, & \text{if } \|(\mathbb{K}u)_j\| = 0, \end{cases} & \text{if } \alpha_j \geq 0 \\ \emptyset, & \text{if } \alpha_j < 0 \end{array} \right\}. \tag{5.6}$$

Let us note that the reformulation (5.5) is a particular case of the first constraint in (2.10), if $F_1(\alpha, u) = \nabla \mathcal{F}(u)$, $F_2(\alpha, u) = (\alpha, u)$, and $\omega \in \mathbb{R}^m_+ \times \mathbb{R}^n$ are chosen. Additionally, the characterization (5.6) is obtained by first considering the case $\|(\mathbb{K}u)_j\| \neq 0$, where, in order to fulfill (5.3b), the relation $q_j = \alpha_j (\mathbb{K}u)_j / \|(\mathbb{K}u)_j\|$ must hold. Otherwise, if $\|(\mathbb{K}u)_j\| = 0$, the inequality (5.3c) holds. Equivalently, by using the definition of the graph of the multifunction Q, we may rewrite (5.5) as

$$\nabla \mathcal{F}(u) + \mathbb{K}^\top q = 0, \tag{5.7a}$$

$$(\alpha, u, \mathbb{K}^\top q) \in gph\ Q, \tag{5.7b}$$

$$(\alpha, u) \in \mathbb{R}^m_+ \times \mathbb{R}^n, \tag{5.7c}$$

where $\operatorname{gph} Q := \{(\alpha, u, \mathbb{K}^\top q) \in \mathbb{R}^m_+ \times \mathbb{R}^n \times \mathbb{R}^n : \mathbb{K}^\top q \in Q(\alpha, u)\}$. Since, for each $\alpha \geq 0$, Q corresponds to the convex subdifferential of the multi-parameter extension of the total variation seminorm, the mapping $(\alpha, u) \mapsto Q(\alpha, u)$ is closed, as well as its graph [69, Theorem 24.4].

The constraint qualification condition presented in [59] guarantees the existence of multipliers that allow the derivation of a stationarity system. Unlike the case presented when learning the data fidelity term, the multivalued function Q now depends on α and u. Consequently, we cannot use the Robinson regularity property as a guarantee for the existence of KKT multipliers as previously done in [42]. An alternative for accomplishing this goal is to prove the constraint qualification condition [59] presented in Section 2.6.

Using the structure of the set-valued operator Q presented in (5.6), let us introduce the following notation for the inactive, strongly active, biactive, zero-inactive and

triactive sets respectively:

$$\begin{aligned}
\mathcal{I}(\alpha,u) &:= \{j\in\{1,\dots,m\} : (\mathbb{K}u)_j \neq 0, \alpha_j > 0\},\\
\mathcal{A}_s(\alpha,u) &:= \{j\in\{1,\dots,m\} : \|q_j\| < \alpha_j, \alpha_j > 0\},\\
\mathcal{B}(\alpha,u) &:= \{j\in\{1,\dots,m\} : \|q_j\| = \alpha_j,\ (\mathbb{K}u)_j = 0, \alpha_j > 0\},\\
\mathcal{I}_0(\alpha,u) &:= \{j\in\{1,\dots,m\} : (\mathbb{K}u)_j \neq 0, \alpha_j = 0\},\\
\mathcal{T}(\alpha,u) &:= \{j\in\{1,\dots,m\} : \|q_j\| = \alpha_j,\ (\mathbb{K}u)_j = 0, \alpha_j = 0\},
\end{aligned}$$

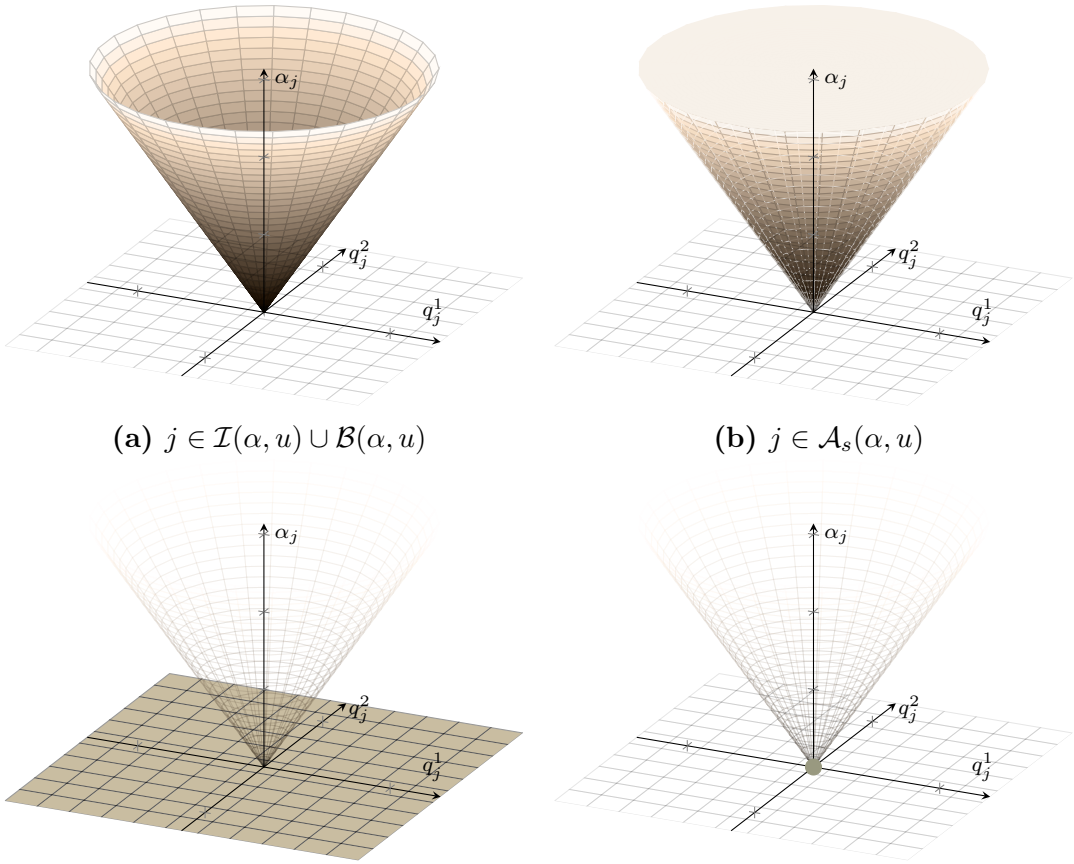
We will hereafter omit the arguments in the set notation whenever they can be inferred from the context. Let us note that the condition in the strongly active set $\mathcal{A}_s$ implies a strict positive parameter $\alpha_j > 0$ for this index set.

In Figure 5.1, we present a geometric representation of these index sets. Particularly, for $j \in \mathcal{I}\cup\mathcal{B}$, we see that the norm of the dual variable must hold $\|q_j\| = \alpha_j$, which corresponds to the cone described in Figure 5.1a. When $j\in\mathcal{A}_s$ the dual variable q will exists only in the interior of the cone, see Figure 5.1b. For the case $j\in\mathcal{I}_0$, the dual variable must be in the plane $\alpha_j = 0$. Finally, the triactive set depicted in Figure 5.1d contains a single element corresponding to the case where both $(\mathbb{K}u)_j = 0$ and $\alpha_j = 0$.

In the following lemmata, we obtain the Bouligand tangent cone, the Fréchet normal cone, and the Mordukhovich normal cone to the graph of the multifunction Q.

Lemma 5.1. *The Bouligand tangent cone to the graph of Q, described in (5.7), is given by*

$$T_{\operatorname{gph} Q}(\alpha,u,\mathbb{K}^\top q) =$$
$$\left\{(\delta_\alpha,\delta_u,\mathbb{K}^\top\delta_q) : \begin{cases}
(\delta_q)_j - (\delta_\alpha)_j \dfrac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|} - \alpha_j T_j(\mathbb{K}\delta_u)_j = 0, & \text{if } j\in\mathcal{I},\\
(\mathbb{K}\delta_u)_j = 0, & \text{if } j\in\mathcal{A}_s,\\
\left.\begin{array}{l}(\mathbb{K}\delta_u)_j = 0, \langle(\delta_q)_j, q_j\rangle \le \alpha_j(\delta_\alpha)_j \vee\\ (\mathbb{K}\delta_u)_j = \tilde{c}q_j(\tilde{c}\ge 0), \langle(\delta_q)_j, q_j\rangle = \alpha_j(\delta_\alpha)_j\end{array}\right\} & \text{if } j\in\mathcal{B},\\
(\delta_q)_j - (\delta_\alpha)_j \dfrac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|} = 0,\ (\delta_\alpha)_j \ge 0 & \text{if } j\in\mathcal{I}_0,\\
\left.\begin{array}{l}(\delta_\alpha)_j \ge 0, (\mathbb{K}\delta_u)_j \in \mathbb{R}^2\backslash\{0\}, (\delta_q)_j - (\delta_\alpha)_j \dfrac{(\mathbb{K}\delta_u)_j}{\|(\mathbb{K}\delta_u)_j\|} = 0 \vee\\ (\delta_\alpha)_j \ge 0, (\mathbb{K}\delta_u)_j = 0, \|(\delta_q)_j\| - (\delta_\alpha)_j \le 0\end{array}\right\} & \text{if } j\in\mathcal{T},
\end{cases}\right\} \tag{5.8}$$



(a) $j \in \mathcal{I}(\alpha, u) \cup \mathcal{B}(\alpha, u)$ **(b)** $j \in \mathcal{A}_s(\alpha, u)$

(c) $j \in \mathcal{I}_0(\alpha, u)$ **(d)** $j \in \mathcal{T}(\alpha, u)$

Figure 5.1: Geometric interpretation of the primal-dual system for different index sets.

where

$$T_j(\mathbb{K}v)_j = \frac{(\mathbb{K}v)_j}{\|(\mathbb{K}u)_j\|} - \frac{(\mathbb{K}u)_j(\mathbb{K}u)_j^\top(\mathbb{K}v)_j}{\|(\mathbb{K}u)_j\|^3}, \text{ for } v \in \mathbb{R}^n.$$

Proof. The tangent cone to the graph of the multifunction Q, see Definition 2.6, is defined as

$$T_{\operatorname{gph} Q}(\alpha, u, \mathbb{K}^\top q) = \{(\delta_\alpha, \delta_u, \mathbb{K}^\top \delta_q) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n : \exists t_k \to 0, (\alpha_k, u_k, \mathbb{K}^\top q_k) \in \operatorname{gph} Q : \frac{1}{t_k}((\alpha_k, u_k, \mathbb{K}^\top q_k) - (\alpha, u, \mathbb{K}^\top q)) \to (\delta_\alpha, \delta_u, \mathbb{K}^\top \delta_q)\}. \quad (5.9)$$

Let us note that in this definition, we take sequences of elements of $\operatorname{gph} Q$, and because of the closeness of $\operatorname{gph} Q$, this limit belongs to the graph as well. Consequently, the limiting elements are of the form $(\delta_\alpha, \delta_u, \mathbb{K}^\top \delta_q)$. Taking $(\delta_\alpha, \delta_u, \mathbb{K}^\top \delta_q) \in T_{\operatorname{gph} Q}(\alpha, u, \mathbb{K}^\top q)$, then by definition of the tangent cone, see (5.9), there exist a sequence $\{(\alpha_k, u_k, \mathbb{K}^\top q_k)\} \subset \operatorname{gph} Q$ and a sequence $t_k \to 0$. Moreover, for a particular k we know that $(\alpha_k, u_k, \mathbb{K}^\top q_k) \in \operatorname{gph} Q$ if and only if $(q_k)_j \in \partial((\alpha_k)_j \|(\mathbb{K}u_k)_j\|)$ for all $j = 1, \ldots, m$. This remark allows us to split the analysis into different cases according to the definition of the multifunction Q.

Case 1: $j \in \mathcal{I}(\alpha, u)$. In this index set, the dual variable can be uniquely characterized. According to (5.6), the graph of the set-valued map corresponds to the following differentiable manifold:

$$h_j(\alpha, u, \mathbb{K}^\top q) := q_j - \alpha_j \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|} = 0. \quad (5.10)$$

Using Lyusternik's theorem [43, Theorem 4.21], the $j-$th component of the tangent direction, $((\delta_\alpha)_j, (\mathbb{K}\delta_u)_j, (\delta_q)_j)$, then satisfies

$$(\delta_q)_j - (\delta_\alpha)_j \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|} - \alpha_j T_j(\mathbb{K}\delta_u)_j = 0.$$

Case 2: $j \in \mathcal{A}_s(\alpha, u)$. In this index set we know $\|q_j\| < \alpha_j$. Therefore, a component in this index set can only be approximated by taking sequences of strongly active components. For n sufficiently large we then take sequences such that $(\mathbb{K}u_k)_j = 0$, $\|(q_k)_j\| < (\alpha_k)_j$. Taking the limit in $(\mathbb{K}u_k)_j = 0$, as $k \to \infty$, yields $(\mathbb{K}\delta_u)_j = 0$. For the dual variable, let us take the sequence $(q_k)_j = q_j + t_k d$ with arbitrary $d \in \mathbb{R}^2$. It then follows that

$$(\delta_q)_j = \lim_{k\to\infty} \frac{(q_k)_j - q_j}{t_k} = d.$$

Since we took an arbitrary direction d it yields $(\delta_q)_j \in \mathbb{R}^2$. Similarly, we get that $(\delta_\alpha)_j \in \mathbb{R}$, as $\alpha_j > 0$.

Case 3: $j \in \mathcal{B}(\alpha, u)$. There are three possible approximations to a component in this index set either via inactive, strongly active, or biactive sequences. Since $\alpha_j > 0$ in all these three cases, we can approximate it by sequences $(\alpha_k)_j < \alpha_j$ or $(\alpha_k)_j > \alpha_j$. Consequently, we get $(\delta_\alpha)_j \in \mathbb{R}$.

Now, if we approach a biactive point using a sequence in the *inactive* set, this sequence satisfies $(\mathbb{K}u_k)_j \neq 0$. Then the sequence of dual variables has the following form

$$(q_k)_j = (\alpha_k)_j \frac{(\mathbb{K}u_k)_j}{\|(\mathbb{K}u_k)_j\|}. \tag{5.11}$$

Furthermore, considering the following product

$$\langle (q_k)_j, (\mathbb{K}u_k)_j \rangle = \left\langle (\alpha_k)_j \frac{(\mathbb{K}u_k)_j}{\|(\mathbb{K}u_k)_j\|}, (\mathbb{K}u_k)_j \right\rangle = (\alpha_k)_j \|(\mathbb{K}u_k)_j\|, \tag{5.12}$$

dividing by t_k in both sides and taking the limit as $k \to \infty$, we get $\langle q_j, (\mathbb{K}\delta_u)_j \rangle = \alpha_j \|(\mathbb{K}\delta_u)_j\|$. Recalling that in this index set $\|q_j\| = \alpha_j > 0$, we know both vectors are collinear, i.e.,

$$(\mathbb{K}\delta_u)_j = \tilde{c} q_j, \quad \text{for some } \tilde{c} \geq 0. \tag{5.13}$$

Using that $\|q_j\| = \alpha_j$, the following holds

$$\begin{aligned}
\left\langle \frac{(q_k)_j - q_j}{t_k}, q_j \right\rangle &= \frac{1}{t_k}(\langle (q_k)_j, q_j \rangle - \langle q_j, q_j \rangle) \\
&= \frac{1}{t_k}(\langle (q_k)_j, q_j \rangle - \langle (q_k)_j, (q_k)_j \rangle + \langle (q_k)_j, (q_k)_j \rangle - \alpha_j^2), \\
&= -\left\langle \frac{(q_k)_j - q_j}{t_k}, (q_k)_j \right\rangle + \frac{(\alpha_k)_j^2 - \alpha_j^2}{t_k},
\end{aligned}$$

where we used the property $\|(q_k)_j\| = (\alpha_k)_j$. Now, rearranging the terms in the last equation, we get

$$\left\langle \frac{(q_k)_j - q_j}{t_k}, q_j \right\rangle + \left\langle \frac{(q_k)_j - q_j}{t_k}, (q_k)_j \right\rangle = \frac{(\alpha_k)_j^2 - \alpha_j^2}{t_k}.$$

Consequently, taking the limit as $k \to \infty$, the following equation holds true

$$2\langle (\delta q)_j, q_j \rangle = 2\alpha_j (\delta_\alpha)_j,$$

and consequently we get $\langle (\delta q)_j, q_j \rangle = \alpha_j (\delta_\alpha)_j$.

Now, if the approximation is done through a sequence of *strongly active* compo-

nents, we know the sequence satisfies $(\mathbb{K}u_k)_j = 0$ and $\|(q_k)_j\| < (\alpha_k)_j$. In this case, we know $(\mathbb{K}\delta_u)_j = 0$ and, using the Cauchy-Schwarz inequality, we get

$$\left\langle \frac{(q_k)_j - q_j}{t_k}, q_j \right\rangle \leq \frac{\alpha_j}{t_k}(\|(q_k)_j\| - \|q_j\|) < \alpha_j \left(\frac{(\alpha_k)_j - \alpha_j}{t_k} \right),$$

which implies that $\langle (\delta_q)_j, q_j \rangle \leq \alpha_j (\delta_\alpha)_j$.

Finally, approximating through a *biactive* set sequence, we know again $(\mathbb{K}\delta_u)_j = 0$ and, estimating the product

$$\left\langle \frac{(q_k)_j - q_j}{t_k}, q_j \right\rangle \leq \frac{\alpha_j}{t_k}(\|(q_k)_j\| - \|q_j\|) = \alpha_j \left(\frac{(\alpha_k)_j - \alpha_j}{t_k} \right),$$

taking the limit as $k \to \infty$, we get

$$\langle (\delta_q)_j, q_j \rangle \leq \alpha_j (\delta_\alpha)_j. \tag{5.14}$$

Case 4: $j \in \mathcal{I}_0(\alpha, u)$. We can approximate a component in the zero-inactive set by sequences in the *zero-inactive* set and the *inactive* set. Therefore, considering an sequence in the *inactive* components, let us take $(\mathbb{K}u_k)_j = (\mathbb{K}u)_j + t_k v$ with $v \in \mathbb{R}^2$ arbitrary. Then, the following limit holds true

$$(\mathbb{K}\delta_u)_j = \lim_{k\to\infty} \frac{(\mathbb{K}u_k)_j - (\mathbb{K}u)_j}{t_k} = v. \tag{5.15}$$

Since we took $v \in \mathbb{R}^2$ arbitrary, it follows that $(\mathbb{K}\delta_u) \in \mathbb{R}^2$. Furthermore, since $q_j = \alpha_j((\mathbb{K}u)_j / \|(\mathbb{K}u)_j\|)$ and $\alpha_j = 0$, then $q_j = 0$ in this index set. Then, the limiting process for the dual variable reads

$$(\delta_q)_j = \lim_{k\to\infty} \frac{(q_k)_j - q_j}{t_k} = \lim_{k\to\infty} \frac{q_k}{t_k} = \lim_{k\to\infty} \frac{(\alpha_k)_j}{t_k} \frac{(\mathbb{K}u_k)_j}{\|(\mathbb{K}u_k)_j\|},$$

where we used the characterization $(q_k)_j = (\mathbb{K}u_k)_j / \|(\mathbb{K}u_k)_j\|$ for the dual variable in the inactive set. Now, considering the sequence $(\alpha_k)_j = \alpha_j + t_k(\delta_\alpha)_j$ we obtain

$$\begin{aligned}(\delta_q)_j &= \lim_{k\to\infty} \frac{\alpha_j + t_k(\delta_\alpha)_j}{t_k} \frac{(\mathbb{K}u_k)_j}{\|(\mathbb{K}u_k)_j\|}, \\ &= \lim_{k\to\infty} (\delta_\alpha)_j \frac{(\mathbb{K}u_k)_j}{\|(\mathbb{K}u_k)_j\|} = (\delta_\alpha)_j \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}.\end{aligned} \tag{5.16}$$

Since $\alpha_j = 0$, the only valid approximations are the ones coming from positive elements, thus $(\delta_\alpha)_j \geq 0$. Another possible approximation can be done through *zero-inactive* components, meaning $(q_k)_j = 0$ and $(\alpha_k)_j = 0$. This case implies

$(\delta_q)_j = 0$, $(\delta_\alpha)_j = 0$ and $(\mathbb{K}\delta_u)_j \in \mathbb{R}^2$, which is a particular case of (5.16).

Case 5: $j \in \mathcal{T}(\alpha, u)$. There are four ways to approach a triactive component. As in the zero-inactive case, all valid approximations come from $(\alpha_k)_j \geq 0$, which again implies $(\delta_\alpha)_j \geq 0$. Similarly to the zero-inactive case (5.15), we also get $(\mathbb{K}\delta_u)_j \in \mathbb{R}^2$.

Approximating through a sequence in the *inactive* set, we obtain

$$(\delta_q)_j = \lim_{t_k \to 0} \frac{1}{t_k} \left((\alpha_k)_j \frac{(\mathbb{K}u_k)_j}{\|(\mathbb{K}u_k)_j\|} \right) = (\delta_\alpha)_j \frac{(\mathbb{K}\delta_u)_j}{\|(\mathbb{K}\delta_u)_j\|}. \tag{5.17}$$

Likewise, the approximation can be made using *zero-inactive* components. In this case $(\mathbb{K}\delta_u)_j \neq 0$, $(q_k)_j = 0$ and $(\alpha_k)_j = 0$. From this sequence we can derive $(\delta_q)_j = 0$, $(\delta_\alpha)_j = 0$ and $(\mathbb{K}\delta_u)_j \in \mathbb{R}^2 \backslash \{0\}$, which is included in (5.17).

Moving forward, we can also approximate through *strongly active* entries, i.e., $(\mathbb{K}u_k)_j = 0$ and $\|(q_k)_j\| < (\alpha_k)_j$. From this sequence we know $(\mathbb{K}\delta_u)_j = 0$, $(\delta_\alpha)_j \geq 0$, and the dual variable direction will satisfy

$$\|(\delta_q)_j\| = \left\| \lim_{t_k \to 0} \frac{(q_k)_j}{t_k} \right\| = \lim_{t_k \to 0} \frac{1}{t_k} \|(q_k)_j\| \leq \lim_{t_k \to 0} \frac{1}{t_k} (\alpha_k)_j = (\delta_\alpha)_j, \tag{5.18}$$

yielding $\|(\delta_q)_j\| \leq (\delta_\alpha)_j$.

Finally, we consider an approximation through *biactive* components, meaning $(\mathbb{K}u_k)_j = 0$ and $\|(q_k)_j\| = (\alpha_k)_j$. We then obtain $(\mathbb{K}\delta_u)_j = 0$, $(\delta_\alpha)_j \geq 0$ and

$$\|(\delta_q)_j\| = \left\| \lim_{t_k \to 0} \frac{(q_k)_j}{t_k} \right\| = \lim_{t_k \to 0} \frac{1}{t_k} \|(q_k)_j\| = \lim_{t_k \to 0} \frac{1}{t_k} (\alpha_k)_j = (\delta_\alpha)_j,$$

which is a particular case of (5.18).

Now, let us name $\mathcal{M}(\alpha, u, \mathbb{K}^\top q)$ the right-hand side of (5.8). Using this notation, so far, we have proven that $T_{\operatorname{gph} Q}(\alpha, u, \mathbb{K}^\top q) \subseteq \mathcal{M}(\alpha, u, \mathbb{K}^\top q)$. To prove the reverse inclusion, let us take $(\delta_\alpha, \delta_u, \mathbb{K}^\top \delta_q) \in \mathcal{M}(\alpha, u, \mathbb{K}^\top q)$ and in the rest of this section we will prove that $(\delta_\alpha, \delta_u, \mathbb{K}^\top \delta_q) \in T_{\operatorname{gph} Q}(\alpha, u, \mathbb{K}^\top q)$. Thanks to the result in (2.5), we know that a triplet $(\delta_\alpha, \delta_u, \mathbb{K}^\top \delta_q)$ is tangent to $\operatorname{gph} Q$ at $(\alpha, u, \mathbb{K}^\top q)$ if

$$\lim_{t \to 0} \frac{\operatorname{dist}((\alpha + t\delta_\alpha, u + t\delta_u, \mathbb{K}^\top q + t\mathbb{K}^\top \delta_q), \operatorname{gph} Q)}{t} = 0,$$

where $\operatorname{dist}(v, S)$ stands for the distance function of a vector v to the set S, presented previously in definition 2.7.

Since the elements in $\mathcal{M}(\alpha, u, \mathbb{K}^\top q)$ are characterized by index set, let us consider

each case separately. The $\operatorname{gph} Q$ is a smooth manifold for the *inactive* and *zero-inactive* components, and the tangent elements are fully characterized by its derivative [68, Example 6.8]. Consequently, the elements defined in this index set are also contained in $T_{\operatorname{gph} Q}(\alpha, u, \mathbb{K}^\top q)$. Likewise, the *strongly active* components lie in the interior of $\operatorname{gph} Q$, which by definition of tangency, coincides with the definition provided in $\mathcal{M}$.

Now, concerning the *biactive* components, we know the set $\mathcal{M}$ provides two representations for the components in this index set. The first being $(\mathbb{K}\delta_u)_j = 0$ and $\langle (\delta_q)_j, q_j \rangle \leq \alpha_j (\delta_\alpha)_j$. Furthermore, taking $t > 0$ and the triplet $(\alpha_j + t(\delta_\alpha)_j, (\mathbb{K}u)_j + t(\mathbb{K}\delta_u)_j, q_j + t(\delta_q)_j)$, its distance to the $\operatorname{gph} Q$ is defined as

$$\inf_{(x_j, (\mathbb{K}y)_j, z_j) \in \operatorname{gph} Q} \|(\alpha_j + t(\delta_\alpha)_j, (\mathbb{K}u)_j + t(\mathbb{K}\delta_u)_j, q_j + t(\delta_q)_j) - (x_j, (\mathbb{K}y)_j, z_j)\|.$$

Using the properties of the $\operatorname{gph} Q$ and $\mathcal{M}$ in this index set, namely, $(\mathbb{K}u)_j = 0$ and $(\mathbb{K}\delta_u)_j = 0$, we may rewrite the previous equation as

$$\inf_{\|z_j\| = x_j} \|(\alpha_j + t(\delta_\alpha)_j, q_j + t(\delta_q)_j) - (x_j, z_j)\|. \tag{5.19}$$

Considering the optimal values for (x_j, z_j), it corresponds to the projection of the tuple $(\alpha_j + t(\delta_\alpha)_j, q_j + t(\delta_q)_j)$ onto the cone $\|q_j\| = \alpha_j$ and can be derived to be the following expressions

$$\begin{aligned} x_j &= \left(\frac{\alpha_j + t(\delta_\alpha)_j + \|q_j + t(\delta_q)_j\|}{2} \right), \\ z_j &= \left(\frac{\alpha_j + t(\delta_\alpha)_j + \|q_j + t(\delta_q)_j\|}{2} \right) \frac{q_j + t(\delta_q)_j}{\|q_j + t(\delta_q)_j\|}. \end{aligned}$$

Replacing these results into (5.19) yields

$$\begin{aligned} &\inf_{\|z_j\| = x_j} \|(\alpha_j + t(\delta_\alpha)_j - x_j, q_j + t(\delta_q)_j - z_j)\| = \\ &\left\| \left(\frac{\alpha_j + t(\delta_\alpha)_j - \|q_j + t(\delta_q)_j\|}{2}, q_j + t(\delta_q)_j - \left(\frac{\alpha_j + t(\delta_\alpha)_j + \|q_j + t(\delta_q)_j\|}{2} \right) \frac{q_j + t(\delta_q)_j}{\|q_j + t(\delta_q)_j\|} \right) \right\|. \end{aligned}$$

Now, using the properties of the norm in product spaces, we may upper bound this norm as follows

$$\begin{aligned} \inf_{\|z_j\| = x_j} \|(\alpha_j + t(\delta_\alpha)_j - x_j, q_j + t(\delta_q)_j - z_j)\| \leq &\left(\frac{\alpha_j + t(\delta_\alpha)_j - \|q_j + t(\delta_q)_j\|}{2} \right) \\ &+ \left\| q_j + t(\delta_q)_j - \left(\frac{\alpha_j + t(\delta_\alpha)_j + \|q_j + t(\delta_q)_j\|}{2} \right) \frac{q_j + t(\delta_q)_j}{\|q_j + t(\delta_q)_j\|} \right\|. \end{aligned} \tag{5.20}$$

Using the same procedure described in (4.11), we can write the second term in the

right-hand side equivalently as

$$\left\| q_j + t(\delta_q)_j - \left(\frac{\alpha_j + t(\delta_\alpha)_j + \|q_j + t(\delta_q)_j\|}{2} \right) \frac{q_j + t(\delta_q)_j}{\|q_j + t(\delta_q)_j\|} \right\| = \frac{\|q_j + t(\delta_q)_j\|}{2} - \left(\frac{\alpha_j + t(\delta_\alpha)_j}{2} \right)$$

Furthermore, finding the upper bound of the norm

$$\begin{aligned} \|q_j + t(\delta_q)_j\|^2 &= \|q_j\|^2 + 2t\langle q_j, (\delta_q)_j \rangle + t^2 \|(\delta_q)_j\|^2, \\ &\leq \alpha_j^2 + 2t\alpha_j(\delta_\alpha)_j + t^2\|(\delta_q)_j\|^2. \end{aligned} \tag{5.21}$$

From this result, we will divide it by t and take the limit as $t \to 0$ for each term in the right-hand side of (5.20). For the first one, we have

$$\begin{aligned} \lim_{t\to 0} \left(\frac{\alpha_j + t(\delta_\alpha)_j - \|q_j + t(\delta_q)_j\|}{2t} \right) &= \lim_{t\to 0} \left(\frac{\alpha + t(\delta_\alpha)_j - \sqrt{\alpha_j^2 + t\alpha_j(\delta_\alpha)_j + t^2\|(\delta_q)_j\|^2}}{2t} \right) \\ &= -\frac{(\delta_\alpha)_j}{2} \end{aligned} \tag{5.22}$$

Now, for the second one, its limit reads

$$\begin{aligned} \lim_{t\to 0} \frac{\|q_j + t(\delta_q)_j\|}{2t} - \left(\frac{\alpha_j + t(\delta_\alpha)_j}{2t} \right) &\leq \lim_{t\to 0} \frac{\sqrt{\alpha_j^2 + 2t\alpha_j(\delta_\alpha)_j + t^2\|(\delta_q)_j\|^2} - \alpha_j - t(\delta_\alpha)_j}{2t}, \\ &= \frac{(\delta_\alpha)_j}{2} \end{aligned}$$

Using both results, we get that the following limit holds true

$$\lim_{t\to 0} \frac{\inf_{\|z_j\|=x_j} \|(\alpha_j + t(\delta_\alpha)_j - x_j, q_j + t(\delta_q)_j - z_j)\|}{t} \leq \frac{1}{2}\left((\delta_\alpha)_j - (\delta_\alpha)_j\right) = 0.$$

Consequently, a triplet $((\delta_\alpha)_j, (\mathbb{K}\delta_u)_j, (\delta_q)_j)$ on $\mathcal{M}$ for the biactive components of the form $(\mathbb{K}\delta_u)_j = 0$ and $\langle q_j, (\delta_q)_j \rangle \leq \alpha_j(\delta_\alpha)_j$ is indeed part of the tangent cone.

Moving forward, we have a second representation for vectors in the *biactive* set. We can also consider the triplets $((\delta_\alpha)_j, (\mathbb{K}\delta_u)_j, (\delta_q)_j)$ such that $(\mathbb{K}\delta_u)_j = \tilde{c}q_j$ with $\tilde{c} \geq 0$, or equivalently, see (5.12), $\langle (\mathbb{K}\delta_u)_j, q_j \rangle = \alpha_j \|(\mathbb{K}\delta_u)_j\|$ and $\langle (\delta_q)_j, q_j \rangle = \alpha_j(\delta_\alpha)_j$. Taking a $t > 0$ we will show in the following that a triplet $(\alpha_j + t(\delta_\alpha)_j, (\mathbb{K}u)_j + t(\mathbb{K}\delta_u)_j, q_j + t(\delta_q)_j)$ is contained in the $\operatorname{gph} Q$. With that goal in mind, let us consider the following product

$$\begin{aligned} \langle (\mathbb{K}u)_j + t(\mathbb{K}\delta_u)_j, q_j + t(\delta_q)_j \rangle &= \underbrace{\langle q_j + t(\delta_q)_j, (\mathbb{K}u)_j \rangle}_{=0} + t\langle q_j, (\mathbb{K}\delta_u)_j \rangle + t^2 \langle (\delta_q)_j, (\mathbb{K}\delta_u)_j \rangle, \\ &= t\alpha_j \|(\mathbb{K}\delta_u)_j\| + t^2 \tilde{c} \langle (\delta_q)_j, q_j \rangle \end{aligned}$$

where we used (5.12) and (5.13). Furthermore, in this representation holds $\langle(\delta_q)_j, q_j\rangle = \alpha_j(\delta_\alpha)_j$, replacing this on the previous equation it yields

$$\langle(\mathbb{K}u)_j + t(\mathbb{K}\delta_u)_j, q_j + t(\delta_q)_j\rangle = t\alpha_j\|(\mathbb{K}\delta_u)_j\| + t^2\tilde{c}\alpha_j(\delta_\alpha)_j.$$

Now, representing the constant $\tilde{c}$ explicitly, we get $\tilde{c} = \|(\mathbb{K}\delta_u)_j\|/\alpha_j$. The product then reads

$$\begin{aligned}\langle(\mathbb{K}u)_j + t(\mathbb{K}\delta_u)_j, q_j + t(\delta_q)_j\rangle &= t\alpha_j\|(\mathbb{K}\delta_u)_j\| + t^2\|(\mathbb{K}\delta_u)_j\|(\delta_\alpha)_j,\\ &= t\|(\mathbb{K}\delta_u)_j\|(\alpha_j + t(\delta_\alpha)_j),\\ &= (\alpha_j + t(\delta_\alpha)_j)\|(\mathbb{K}u)_j + t(\mathbb{K}\delta_u)_j\|.\end{aligned}$$

Consequently, thanks to the previous result and the fact that $\|q_j + t(\delta_q)_j\| = (\alpha_j + t(\delta_\alpha)_j)$, we have shown that the triplet $(\alpha_j + t(\delta_\alpha)_j, (\mathbb{K}u)_j + t(\mathbb{K}\delta_u)_j, q_j + t(\delta_q)_j) \in \operatorname{gph} Q$. This result has further implications relating its distance to $\operatorname{gph} Q$. In particular, we know that

$$\operatorname{dist}((\alpha_j + t(\delta_\alpha)_j, (\mathbb{K}u)_j + t(\mathbb{K}\delta_u)_j, q_j + t(\delta_q)_j), \operatorname{gph} Q) = 0,$$

from where we can conclude that this triplet corresponds to a tangent to $\operatorname{gph} Q$.

Finally, regarding the triactive set, both characterizations trivially imply that for any $t > 0$ the triplet $(\alpha_j + t(\delta_\alpha)_j, (\mathbb{K}u)_j + t(\mathbb{K}\delta_u)_j, q_j + t(\delta_q)_j)$ are included in the $\operatorname{gph} Q$. Consequently, using the same argument as in the previous case, we conclude these are tangent as well, finishing the proof.

□

Lemma 5.2. *The Fréchet normal cone to the graph of Q is given by*

$$N^F_{\operatorname{gph} Q}(\alpha, u, \mathbb{K}^\top q) =$$
$$\left\{(\vartheta, \mathbb{K}^\top\mu, p) : \begin{cases} \mu_j + \alpha_j T_j(\mathbb{K}p)_j = 0, & \text{if } j \in \mathcal{I},\\ \vartheta_j + \dfrac{\langle(\mathbb{K}u)_j, (\mathbb{K}p)_j\rangle}{\|(\mathbb{K}u)_j\|} = 0, & \text{if } j \in \mathcal{I},\\ \vartheta_j = 0,\ (\mathbb{K}p)_j = 0 & \text{if } j \in \mathcal{A}_s,\\ \vartheta_j + c\alpha_j = 0, (\mathbb{K}p)_j = cq_j (c \geq 0), \langle\mu_j, q_j\rangle \leq 0, & \text{if } j \in \mathcal{B},\\ \vartheta_j + \dfrac{\langle(\mathbb{K}u)_j, (\mathbb{K}p)_j\rangle}{\|(\mathbb{K}u)_j\|} \leq 0, \quad \mu_j = 0, & \text{if } j \in \mathcal{I}_0,\\ \vartheta_j + \|(\mathbb{K}p)_j\| \leq 0,\ \mu_j = 0, & \text{if } j \in \mathcal{T}. \end{cases}\right\} \tag{5.23}$$

Proof. Using the definition of the Fréchet normal cone, see Definition 2.8, we know that we can build it as the polar of the tangent cone. Considering, in particular,

directions of the form $\delta_\alpha = 0$, $\delta_u \in \ker(\mathbb{K})$ and $\delta_q = 0$, it follows that, for a general normal vector (ϑ, φ, p), $\langle \varphi, \delta_u \rangle \leq 0, \forall \delta_u \in \ker(\mathbb{K})$ must hold. This implies that $\varphi \in \ker(\mathbb{K})^\perp = \operatorname{range}(\mathbb{K}^\top)$. Consequently, for $(\delta_\alpha, \delta_u, \mathbb{K}^\top \delta_q) \in T_{\operatorname{gph} Q}(\alpha, u, \mathbb{K}^\top q)$ we have that the Fréchet normal cone can be calculated as

$$N^F_{\operatorname{gph} Q}(\alpha, u, \mathbb{K}^\top q) = \{(\vartheta, \mathbb{K}^\top \mu, p) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n : \langle (\vartheta, \mathbb{K}^\top \mu, p), (\delta_\alpha, \delta_u, \mathbb{K}^\top \delta_q) \rangle \leq 0\}.$$

Indeed, we can rewrite this inequality as

$$\sum_{j=1}^{n} ((\delta_\alpha)_j \vartheta_j + \langle (\mathbb{K}\delta_u)_j, \mu_j \rangle + \langle (\delta_q)_j, (\mathbb{K}p)_j \rangle) \leq 0.$$

We analyze the different cases according to their index set using this representation, along with the tangent cone characterization from Lemma 5.1.

Case 1: $j \in \mathcal{I}$. Using the characterization of the elements in the tangent cone, we get

$$(\delta_\alpha)_j \vartheta_j + \langle (\mathbb{K}\delta_u)_j, \mu_j \rangle + \left\langle (\delta_\alpha)_j \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, (\mathbb{K}p)_j \right\rangle + \langle \alpha_j T_j (\mathbb{K}\delta_u)_j, (\mathbb{K}p)_j \rangle \leq 0.$$

Rearranging the terms and using the symmetry of T_j,

$$(\delta_\alpha)_j \left(\vartheta_j + \frac{\langle (\mathbb{K}u)_j, (\mathbb{K}p)_j \rangle}{\|(\mathbb{K}u)_j\|} \right) + \langle (\mathbb{K}\delta_u)_j, \mu_j + \alpha_j T_j (\mathbb{K}p)_j \rangle \leq 0.$$

Since $(\mathbb{K}\delta_u)_j \in \mathbb{R}^2$ and $(\delta_\alpha)_j \in \mathbb{R}$, it necessarily must hold

$$\vartheta_j + \frac{\langle (\mathbb{K}u)_j, (\mathbb{K}p)_j \rangle}{\|(\mathbb{K}u)_j\|} = 0, \qquad \mu_j + \alpha_j T_j (\mathbb{K}p)_j = 0.$$

Case 2: $j \in \mathcal{A}_s$. In this index set we know that $(\mathbb{K}\delta_u)_j = 0$, $(\delta_\alpha)_j \in \mathbb{R}$ and $(\delta_q)_j \in \mathbb{R}^2$. Consequently, the product reads

$$(\delta_\alpha)_j \vartheta_j + \langle (\delta_q)_j, (\mathbb{K}p)_j \rangle \leq 0,$$

and we obtain that $(\mathbb{K}p)_j = 0$ and $\vartheta_j = 0$.

Case 3: $j \in \mathcal{B}$. In this index set, there are two conditions in the normal directions. For the first one, we take $(\mathbb{K}\delta_u)_j = 0$, and the cone inequality reads

$$(\delta_\alpha)_j \vartheta_j + \langle (\delta_q)_j, (\mathbb{K}p)_j \rangle \leq 0, \quad \forall (\delta_\alpha)_j, (\delta_q)_j \text{ s.t. } \langle (\delta_q)_j, q_j \rangle \leq \alpha_j (\delta_\alpha)_j. \tag{5.24}$$

Taking in particular, $(\delta_\alpha)_j = 0$ we get

$$\langle (\delta_q)_j, (\mathbb{K}p)_j \rangle \leq 0, \forall (\delta_q)_j \text{ s.t. } \langle (\delta_q)_j, q_j \rangle \leq 0.$$

Therefore, $(\mathbb{K}p)_j = cq_j$ for some $c \geq 0$. Using this result in (5.24) for the particular case $(\delta_\alpha)_j = \frac{1}{\alpha_j}\langle (\delta_q)_j, q_j \rangle$, we obtain

$$0 \geq (\delta_\alpha)_j \vartheta_j + \langle (\delta_q)_j, cq_j \rangle = (\delta_\alpha)_j (\vartheta_j + c\alpha_j),$$

from where $\vartheta_j + c\alpha_j = 0$ holds. The resulting cone reads

$$\vartheta_j + c\alpha_j = 0, \quad (\mathbb{K}p)_j = cq_j, \quad c \geq 0, \quad \mu_j \in \mathbb{R}^2. \tag{5.25}$$

For the second case we take $(\mathbb{K}\delta_u) = \tilde{c}q_j$ ($\tilde{c} \geq 0$) in the normal cone inequality,

$$(\delta_\alpha)_j \vartheta_j + \langle \tilde{c}q_j, \mu_j \rangle + \langle (\mathbb{K}p)_j, (\delta_q)_j \rangle \leq 0, \forall (\delta_\alpha)_j, (\delta_q)_j \text{ s.t. } \langle (\delta_q)_j, q_j \rangle = \alpha_j (\delta_\alpha)_j. \tag{5.26}$$

Again, considering $(\delta_\alpha)_j = 0$ and $(\mathbb{K}\delta_u)_j = 0$, we get

$$\langle (\delta_q)_j, (\mathbb{K}p)_j \rangle \leq 0, \forall (\delta_q)_j \text{ s.t. } \langle (\delta_q)_j, q_j \rangle = 0.$$

Consequently, $(\mathbb{K}p)_j = cq_j$ with $c \in \mathbb{R}$. Using this result in (5.26), while keeping $(\delta_\alpha)_j = 0$, yields $\tilde{c}\langle q_j, \mu_j \rangle \leq 0$. Thanks to the positiveness of $\tilde{c}$ we then get $\langle q_j, \mu_j \rangle \leq 0$. Now, applying all previous results to the normal cone inequality, we get

$$0 \geq (\delta_\alpha)_j \vartheta_j + \langle cq_j, (\delta_q)_j \rangle = (\delta_\alpha)_j \vartheta_j + c\alpha_j (\delta_\alpha)_j, \forall (\delta_\alpha)_j \in \mathbb{R}, c \in \mathbb{R},$$

yielding $\vartheta_j + c\alpha_j = 0$. Therefore, the resulting cone for the second case reads

$$\vartheta_j + c\alpha_j = 0, \quad \langle q_j, \mu_j \rangle \leq 0, \quad (\mathbb{K}p)_j = cq_j, \quad c \in \mathbb{R}. \tag{5.27}$$

Finally, considering both cases, we obtain

$$\vartheta_j + c\alpha_j = 0 \wedge \langle \mu_j, q_j \rangle \leq 0 \wedge (\mathbb{K}p)_j = cq_j \wedge c \geq 0.$$

Case 4: $j \in \mathcal{I}_0$. By using the characterization of the tangent cone in this index set, we get

$$(\delta_\alpha)_j \left(\vartheta_j + \frac{\langle (\mathbb{K}u)_j, (\mathbb{K}p)_j \rangle}{\|(\mathbb{K}u)_j\|} \right) + \langle (\mathbb{K}\delta_u)_j, \mu_j \rangle \leq 0.$$

This relationship must hold for all $(\mathbb{K}\delta_u)_j \in \mathbb{R}^2$ and $(\delta_\alpha)_j \geq 0$, which implies

$$\vartheta_j + \frac{\langle (\mathbb{K}u)_j, (\mathbb{K}p)_j \rangle}{\|(\mathbb{K}u)_j\|} \leq 0, \quad \langle (\mathbb{K}\delta_u)_j, \mu_j \rangle \leq 0.$$

Since $(\mathbb{K}\delta_u)_j \in \mathbb{R}^2$ we get $\mu_j = 0$.

Case 5: $j \in \mathcal{T}$. For the first case in this index set, we know the elements of the tangent cone satisfy

$$(\delta_q)_j = (\delta_\alpha)_j \frac{(\mathbb{K}\delta_u)_j}{\|(\mathbb{K}\delta_u)_j\|}, \ (\delta_\alpha)_j \geq 0, \ (\mathbb{K}\delta_u)_j \in \mathbb{R}^2 \backslash \{0\}.$$

Replacing these terms with the normal cone inequality,

$$(\delta_\alpha)_j \left(\vartheta_j + \frac{\langle (\mathbb{K}p)_j, (\mathbb{K}\delta_u)_j \rangle}{\|(\mathbb{K}\delta_u)_j\|} \right) + \langle \mu_j, (\mathbb{K}\delta_u)_j \rangle \leq 0, \ \forall (\delta_\alpha)_j \geq 0, (\mathbb{K}\delta_u)_j \in \mathbb{R}^2 \backslash \{0\}.$$

In particular, for $(\delta_\alpha)_j = 0$, we get that $\langle \mu_j, (\mathbb{K}\delta_u)_j \rangle \leq 0$, for all $(\mathbb{K}\delta_u)_j \in \mathbb{R}^2 \backslash \{0\}$, which implies that $\mu_j = 0$. Moreover, thanks to the positiveness of $(\delta_\alpha)_j \geq 0$, we get

$$\vartheta_j + \frac{\langle (\mathbb{K}p)_j, (\mathbb{K}\delta_u)_j \rangle}{\|(\mathbb{K}\delta_u)_j\|} \leq 0, \forall (\mathbb{K}\delta_u)_j \in \mathbb{R}^2 \backslash \{0\}. \tag{5.28}$$

Since (5.28) must hold for all $(\mathbb{K}\delta_u)_j \in \mathbb{R}^2 \backslash \{0\}$, we can test it with $(\mathbb{K}\delta_u)_j = (\mathbb{K}p)_j$, from where we get

$$\vartheta_j + \|(\mathbb{K}p)_j\| \leq 0.$$

Now, regarding the second condition, we know $\|(\delta_q)_j\| - (\delta_\alpha)_j \leq 0, \ (\mathbb{K}\delta_u)_j = 0$ and $(\delta_\alpha)_j \geq 0$. Since $(\mathbb{K}\delta_u)_j = 0$, it follows $\mu_j \in \mathbb{R}^2$. Using the normal cone inequality,

$$0 \geq (\delta_\alpha)_j \vartheta_j + \langle (\delta_q)_j, (\mathbb{K}p)_j \rangle \geq (\delta_\alpha)_j \vartheta_j - \|(\delta_q)_j\| \|(\mathbb{K}p)_j\| \geq (\delta_\alpha)_j \vartheta_j - (\delta_\alpha)_j \|(\mathbb{K}p)_j\|,$$

from where we derive $(\delta_\alpha)_j (\vartheta_j - \|(\mathbb{K}p)_j\|) \leq 0$. Along with the positivity of $(\delta_\alpha)_j$, this implies $\vartheta_j - \|(\mathbb{K}p)_j\| \leq 0$. Finally, considering both conditions, the result is obtained.

□

Lemma 5.3. *Let $(\alpha, u, \mathbb{K}^\top q) \in \operatorname{gph} Q$, described in (5.7), and a triplet $(\vartheta, \mathbb{K}^\top \mu, p) \in$*

$\mathbb{R}^m_+ \times \mathbb{R}^n \times \mathbb{R}^n$. *If* $(\vartheta, \mathbb{K}^\top \mu, p)$ *satisfies the following conditions*

$$\mu_j + \alpha_j T_j(\mathbb{K}p)_j = 0, \qquad \text{if } j \in \mathcal{I}, \tag{5.29a}$$

$$\vartheta_j + \frac{\langle (\mathbb{K}u)_j, (\mathbb{K}p)_j \rangle}{\|(\mathbb{K}u)_j\|} = 0, \qquad \text{if } j \in \mathcal{I}, \tag{5.29b}$$

$$\vartheta_j = 0, (\mathbb{K}p)_j = 0, \qquad \text{if } j \in \mathcal{A}_s, \tag{5.29c}$$

$$\left.\begin{array}{l} \vartheta_j = 0, (\mathbb{K}p)_j = 0, \ \vee \\ (\mathbb{K}p)_j = cq_j (c \in \mathbb{R}), \ \langle \mu_j, q_j \rangle = 0, \ \vee \\ \vartheta_j + c\alpha_j = 0, (\mathbb{K}p)_j = cq_j (c \geq 0), \langle \mu_j, q_j \rangle \leq 0. \end{array}\right\} \qquad \text{if } j \in \mathcal{B}, \tag{5.29d}$$

$$\vartheta_j + \frac{\langle (\mathbb{K}u)_j, (\mathbb{K}p)_j \rangle}{\|(\mathbb{K}u)_j\|} \leq 0, \quad \mu_j = 0, \qquad \text{if } j \in \mathcal{I}_0, \tag{5.29e}$$

$$\left.\begin{array}{l} \vartheta_j = 0, (\mathbb{K}p)_j = 0, \ \vee \\ \vartheta_j + \|(\mathbb{K}p)_j\| \leq 0, \mu_j = 0 \end{array}\right\} \qquad \text{if } j \in \mathcal{T}. \tag{5.29f}$$

Then, $(\vartheta, \mathbb{K}^\top \mu, p) \in N^M_{\operatorname{gph} Q}(\alpha, u, \mathbb{K}^\top q)$*, where* $N^M_{\operatorname{gph} Q}(\alpha, u, \mathbb{K}^\top q)$ *stands for the Mordukhovich normal cone to the graph of* Q *at* $(\alpha, u, \mathbb{K}^\top q)$*.*

Proof. Let us recall the definition of the Mordukhovich normal cone for our problem (see Definition 2.9)

$$N^M_{\operatorname{gph} Q}(\alpha, u, \mathbb{K}^\top q) = \{(\vartheta, \mathbb{K}^\top \mu, p) : (\vartheta_k, \mathbb{K}^\top \mu_k, p_k) \in N^F_{\operatorname{gph} Q}(\alpha_k, u_k, \mathbb{K}^\top q_k) : \\ (\vartheta_k, \mathbb{K}^\top \mu_k, p_k) \to (\vartheta, \mathbb{K}^\top \mu, p), (\alpha_k, u_k, \mathbb{K}^\top q_k) \to (\alpha, u, \mathbb{K}^\top q)\}.$$

Considering limiting sequences to the inactive, strongly active, and zero-inactive sets, the same directions as for the Fréchet normal cone are obtained. The differences lie in the biactive and triactive sets, where several approximations may be considered.

Case 1: $\boldsymbol{j \in \mathcal{B}}$**.** By taking approximation sequences in the *inactive* set, from Lemma 5.2 we know

$$0 = (\mu_k)_j + (\alpha_k)_j \frac{(\mathbb{K}p_k)_j}{\|(\mathbb{K}u_k)_j\|} - (\alpha_k)_j \frac{(\mathbb{K}u_k)\langle ((\mathbb{K}u_k)_j), (\mathbb{K}p_k)_j \rangle}{\|(\mathbb{K}u_k)_j\|^3}. \tag{5.30}$$

Testing (5.30) with $(\mathbb{K}p_k)_j$ yields

$$\langle (\mu_k)_j, (\mathbb{K}p_k)_j \rangle = (\alpha_k)_j \frac{\langle (\mathbb{K}u_k)_j, (\mathbb{K}p_k)_j \rangle^2}{\|(\mathbb{K}u_k)_j\|^3} - (\alpha_k)_j \frac{\|(\mathbb{K}p_k)_j\|^2}{\|(\mathbb{K}u_k)_j\|} \tag{5.31}$$

Multiplying with $(\alpha_k)_j \|(\mathbb{K}u_k)_j\|$ on both sides and recalling that the dual variable

in the inactive set can be uniquely determined by

$$(q_k)_j = (\alpha_k)_j \frac{(\mathbb{K}u_k)_j}{\|(\mathbb{K}u_k)_j\|},$$

we may rewrite (5.31) as follows

$$(\alpha_k)_j \|(\mathbb{K}u_k)_j\| \langle (\mu_k)_j, (\mathbb{K}p_k)_j \rangle = \langle (q_k)_j, (\mathbb{K}p_k)_j \rangle^2 - (\alpha_k)_j^2 \|(\mathbb{K}p_k)_j\|^2.$$

Taking the limit as $k \to \infty$ and recalling $(\alpha_k)_j = \|(q_k)_j\|$ in this index set, we obtain

$$\langle q_j, (\mathbb{K}p)_j \rangle^2 = \|q_j\|^2 \|(\mathbb{K}p)_j\|^2,$$

which implies that $(\mathbb{K}p)_j = cq_j (c \in \mathbb{R})$. Now, testing (5.30) with $(q_k)_j$ we get the following product

$$\begin{aligned} \langle (\mu_k)_j, (q_k)_j \rangle &= (\alpha_k)_j \frac{\langle (q_k)_j, (\mathbb{K}u_k)_j \rangle \langle (\mathbb{K}u_k)_j, (\mathbb{K}p_k)_j \rangle}{\|(\mathbb{K}u_k)_j\|^3} - (\alpha_k)_j \frac{\langle (q_k)_j, (\mathbb{K}p_k)_j \rangle}{\|(\mathbb{K}u_k)_j\|}, \\ &= (\alpha_k)_j^2 \frac{\langle (\mathbb{K}u_k)_j, (\mathbb{K}p_k)_j \rangle}{\|(\mathbb{K}u_k)_j\|^2} - (\alpha_k)_j^2 \frac{\langle (\mathbb{K}u_k)_j, (\mathbb{K}p_k)_j \rangle}{\|(\mathbb{K}u_k)_j\|^2} = 0. \end{aligned} \tag{5.32}$$

Taking the limit, we get that $\langle \mu_j, q_j \rangle = 0$. Regarding ϑ_j we have $(\alpha_k)_j (\vartheta_k)_j + \langle (q_k)_j, (\mathbb{K}p_k)_j \rangle = 0$. Taking the limit as $k \to \infty$, we obtain

$$0 = \alpha_j \vartheta_j + \langle q_j, (\mathbb{K}p)_j \rangle = \alpha_j \vartheta_j + c\|q_j\|^2, (c \in \mathbb{R}),$$

which implies that $\vartheta_j = -c\alpha_j$ with $c \in \mathbb{R}$ and consequently $\vartheta_j \in \mathbb{R}$.

When taking the approximation through the *strongly active* set, from Lemma 5.2 we know

$$(\vartheta_k)_j = 0, \quad (\mathbb{K}p_k)_j = 0.$$

Taking the limit as $k \to \infty$ it reads $\vartheta_j = 0$ and $(\mathbb{K}p)_j = 0$. Finally, considering sequences in the *biactive* set, such approximations take the form

$$(\vartheta_k)_j + c(\alpha_k)_j = 0, \quad (\mathbb{K}p_k)_j = c(q_k)_j, \quad \langle (\mu_k)_j, (q_k)_j \rangle \leq 0.$$

with $c \geq 0$. When considering the limit as $k \to \infty$, the cone directions coincide with the Fréchet normal one.

Case 2: $j \in \mathcal{T}$. This index set can be approximated by sequences belonging either to the inactive, biactive, strongly active, or zero-inactive sets. Considering *strongly active* sequences, $(\mathbb{K}p_k)_j = 0$ and $(\vartheta_k)_j = 0$. Taking the limit as $k \to \infty$ we get $(\mathbb{K}p)_j = 0$ and $\vartheta_j = 0$ as well.

Likewise, when taking *biactive* sequences we get $(\vartheta_k)_j + c(\alpha_k)_j = 0$, $(\mathbb{K}p_k)_j = c(q_k)_j (c \geq 0)$ and $\langle (\mu_k)_j, (q_k)_j \rangle \leq 0$. Again, taking the limit as $k \to \infty$ we get, since $q_j = 0$ and $\alpha_j = 0$, that $\mu_j \in \mathbb{R}^2$, $(\mathbb{K}p)_j = 0$ and $\vartheta_j = 0$.

Furthermore, taking sequences in the *zero-inactive* set we have $(\mu_k)_j = 0$, which implies that $\mu_j = 0$. Recalling that for a component in the triactive set, we have $(\mathbb{K}u)_j = 0$, then the following bound holds true

$$0 \geq (\vartheta_k)_j + \frac{\langle (\mathbb{K}u_k)_j, (\mathbb{K}p_k)_j \rangle}{\|(\mathbb{K}u_k)_j\|} = (\vartheta_k)_j + \frac{\langle (\mathbb{K}\delta u)_j, (\mathbb{K}p_k)_j \rangle}{\|(\mathbb{K}\delta u)_j\|}.$$

Since $(\mathbb{K}\delta u)_j$ is free in this index set, we may take in particular $(\mathbb{K}\delta u)_j = (\mathbb{K}p_k)_j$. Then, taking the limit as $k \to \infty$ yields $\vartheta_j + \|(\mathbb{K}p)_j\| \leq 0$.

When taking *inactive sequences*, we know that

$$(\mu_k)_j + (\alpha_k)_j \left(\frac{I}{\|(\mathbb{K}u_k)_j\|} - \frac{(\mathbb{K}u_k)_j (\mathbb{K}u_k)_j^\top}{\|(\mathbb{K}u_k)_j\|^3} \right) (\mathbb{K}p_k)_j = 0, \tag{5.33}$$

$$(\vartheta_k)_j + \frac{\langle (\mathbb{K}u_k)_j, (\mathbb{K}p_k)_j \rangle}{\|(\mathbb{K}u_k)_j\|} = 0. \tag{5.34}$$

Multiplying (5.33) with $(q_k)_j$ it yields

$$0 = \langle (\mu_k)_j, (q_k)_j \rangle + (\alpha_k)_j \left(\frac{\langle (\mathbb{K}p_k)_j, (q_k)_j \rangle}{\|(\mathbb{K}u_k)_j\|} - \frac{1}{(\alpha_k)_j^2} \frac{\langle (q_k)_j, (\mathbb{K}p_k)_j \rangle \|(q_k)_j\|^2}{\|(\mathbb{K}u_k)_j\|} \right),$$

which implies that

$$\langle (\mu_k)_j, (q_k)_j \rangle = 0. \tag{5.35}$$

Taking the limit as $k \to \infty$ it reads $\langle \mu_j, q_j \rangle = 0$. Furthermore, since the limit corresponds to a triactive component, we have $\|q_j\| = 0$; consequently, it implies $\mu_j \in \mathbb{R}^2$. Moreover, thanks to the property $(\mathbb{K}u)_j = 0$ in a triactive component, we may rewrite (5.34) as follows

$$(\vartheta_k)_j + \frac{\langle (\mathbb{K}\delta u)_j, (\mathbb{K}p_k)_j \rangle}{\|(\mathbb{K}\delta u)_j\|} = 0.$$

Taking in particular $(\mathbb{K}\delta u)_j = (\mathbb{K}p_k)_j$ we have $(\vartheta_k)_j + \|(\mathbb{K}p_k)_j\| = 0$. Furthermore, if taking $(\mathbb{K}\delta u)_j = -(\mathbb{K}p_k)_j$ it also yields $(\vartheta_k)_j - \|(\mathbb{K}p_k)_j\| = 0$. Consequently, it must hold $(\vartheta_k)_j = 0$ and taking the limit as $k \to \infty$ we get $\vartheta_j = 0$ and $(\mathbb{K}p)_j = 0$ in this index set.

Finally, since we took sequences

$$(\vartheta_k, \mathbb{K}^\top \mu_k, p_k) \in N^F_{\operatorname{gph} Q}(\alpha_k, u_k, \mathbb{K}^\top q_k) \subset N^M_{\operatorname{gph} Q}(\alpha_k, u_k, \mathbb{K}^\top q_k)$$

and $N^M_{\mathrm{gph}\, Q}(\alpha, u, \mathbb{K}^\top q) = clN^F_{\mathrm{gph}\, Q}(\alpha, u, \mathbb{K}^\top q)$, the proof is complete.

□

Theorem 5.1 (M-Stationarity). *Let $J : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, $\mathcal{F} : \mathbb{R}^n \to \mathbb{R}$ twice continuously differentiable and strongly convex, and (α^*, u^*, q^*) be a local solution to (5.1). Then there exist KKT multipliers $(\vartheta, \mathbb{K}^\top \mu, p)$ such that*

$$\nabla \mathcal{F}(u^*) + \mathbb{K}^\top q^* = 0, \tag{5.36a}$$

$$\langle q_j^*, (\mathbb{K}u^*)_j \rangle - \alpha_j^* \|(\mathbb{K}u^*)_j\| = 0, \qquad \forall j = 1, \dots, m, \tag{5.36b}$$

$$\|q_j^*\| \leq \alpha_j^*, \qquad \forall j = 1, \dots, m, \tag{5.36c}$$

$$\nabla_{uu} \mathcal{F}(u^*)^\top p - \mathbb{K}^\top \mu - \nabla J(u^*) = 0, \tag{5.36d}$$

$$\vartheta + \rho = 0, \tag{5.36e}$$

$$\langle \alpha^*, \rho \rangle = 0, \tag{5.36f}$$

$$\rho \leq 0, \tag{5.36g}$$

$$\alpha^* \geq 0, \tag{5.36h}$$

$$(\vartheta, \mathbb{K}^\top \mu, p) \in N^M_{\mathrm{gph}\, Q}(\alpha^*, u^*, \mathbb{K}^\top q^*) \tag{5.36i}$$

Proof. Referring to Theorem 2.9 let us take $F_1(\alpha, u) = \nabla \mathcal{F}(u) \in \mathbb{R}^n$ and $F_2(\alpha, u) = (\alpha, u) \in \mathbb{R}^m_+ \times \mathbb{R}^n$. Existence of KKT multipliers is guaranteed if the following constraint qualification condition holds for $(\vartheta, \mathbb{K}^\top \mu, p) \in N^M_{\mathrm{gph}\, Q}(\alpha^*, u^*, \mathbb{K}^\top q^*)$

$$\begin{bmatrix} I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I & -\nabla_{uu} \mathcal{F}(u^*)^\top \end{bmatrix} \begin{bmatrix} \vartheta \\ \mathbb{K}^\top \mu \\ p \end{bmatrix} \in -N^M_{\mathbb{R}^m_+}(\alpha^*) \times \{0\} \text{ implies } \vartheta = 0,\ \mathbb{K}^\top \mu = 0,\ p = 0. \tag{5.37}$$

Recalling Remark 2.2 and using the expression of the Mordukhovich normal cone $N^M_{R^m_+}(\alpha^*) = N_{R^m_+}(\alpha^*) = \{v \in \mathbb{R}^n : \langle v, \alpha^* \rangle = 0,\ v \leq 0\}$, condition in (5.37) can also be written as

$$\mathbb{K}^\top \mu - \nabla_{uu} \mathcal{F}(u^*)^\top p = 0, \tag{5.38}$$

$$\langle \alpha^*, \vartheta \rangle = 0, \tag{5.39}$$

$$\vartheta \geq 0. \tag{5.40}$$

Let us take $(\vartheta, \mathbb{K}^\top \mu, p) \in N^M_{\mathrm{gph}\, Q}(\alpha^*, u^*, \mathbb{K}^\top q^*)$ and let us multiply (5.38) by p on the

left. Recalling $(\mathbb{K}p)_j = 0$ in $\mathcal{A}_s$ and $\mu_j = 0$ in $\mathcal{I}_0$, we have for each remaining index set

$$\begin{aligned}\langle p, \nabla_{uu}\mathcal{F}(u^*)^\top p\rangle = \langle p, \mathbb{K}^\top\mu\rangle &= \sum_{j\in\mathcal{I}}\langle \mu_j, (\mathbb{K}p)_j\rangle + \sum_{j\in\mathcal{B}}\langle \mu_j, (\mathbb{K}p)_j\rangle + \sum_{j\in\mathcal{B}_0}\langle \mu_j, (\mathbb{K}p)_j\rangle,\\ &= \sum_{j\in\mathcal{I}} -\alpha_j\langle (\mathbb{K}p)_j, T_j(\mathbb{K}p)_j\rangle + \sum_{j\in\mathcal{B}} c\underbrace{\langle \mu_j, q_j\rangle}_{\leq 0} + \sum_{j\in\mathcal{T}}\underbrace{\langle \mu_j, (\mathbb{K}p)_j\rangle}_{\leq 0} \leq 0,\end{aligned}$$

where we used the positive semi-definiteness of the matrix T_j and the characterization of the Mordukhovich normal cone. Furthermore, using the strong convexity of the function $\mathcal{F}$ we have $\langle p, \nabla_{uu}\mathcal{F}(u^*)^\top p\rangle > 0,\ \forall p\backslash\{0\}$. Both inequalities imply $p = 0$ and, according to (5.38), it also yields $\mathbb{K}^\top\mu = 0$. Moreover, if we consider the index set $\mathcal{I}\cup\mathcal{A}_s\cup\mathcal{B}$, we know in all these sets $\alpha_j^* > 0$, and therefore, to satisfy equation (5.39) it must hold $\vartheta_j = 0$. Since $p = 0$ in $\mathcal{T}$ we know $\vartheta_j = 0$ or $\vartheta_j \leq \|(\mathbb{K}p)_j\|$ for this index set. In both cases, it leads to $\vartheta_j = 0$. In $\mathcal{I}_0$, we have $\vartheta_j \leq -\frac{\langle(\mathbb{K}u)_j, (\mathbb{K}p)_j\rangle}{\|(\mathbb{K}u)_j\|} = 0$ and (5.40) yields $\vartheta_j = 0$. Therefore, $\vartheta_j = 0$ for all j.

Consequently, the existence of multipliers is guaranteed and there exists a vector $\rho \in N^M_{R^m_+}(\alpha^*) = \{v \in \mathbb{R}^n : \langle v, \alpha^*\rangle = 0,\ v \leq 0\}$ and KKT multipliers $(\vartheta, \mathbb{K}^\top\mu, p) \in N^M_{\operatorname{gph} Q}(\alpha^*, u^*, \mathbb{K}^\top q^*)$ such that

$$0 = \nabla J(u^*) + \nabla_u F_2(\alpha^*, u^*)^\top \begin{bmatrix}\vartheta\\ \mathbb{K}^\top\mu\end{bmatrix} - \nabla_u F_1(\alpha^*, u^*)^\top p, \tag{5.41}$$

$$0 = \nabla_\alpha F_2(\alpha^*, u^*)^\top \begin{bmatrix}\vartheta\\ \mathbb{K}^\top\mu\end{bmatrix} - \nabla_\alpha F_1(\alpha^*, u^*)^\top p + \rho. \tag{5.42}$$

To recover the optimality system in (5.36), let us take (α^*, u^*, q^*), a local optimal solution of (4.1). Then, note that equations in (5.41) and (5.42) correspond to equations (d) and (e) respectively. Taking a $\rho \in \mathbb{R}^n$ we must add the conditions $\langle\alpha^*, \rho\rangle = 0$ and $\rho \leq 0$ to guarantee it is contained in $N^M_{\mathbb{R}^m_+}(\alpha^*)$, yielding equations (f) and (g). Finally, equations (a-c) correspond to the state constraints of the original problem. □

5.2 Bouligand Stationarity

In this section we study the Bouligand stationarity condition for (3.3). With this goal in mind, let us introduce the solution operator for the lower-level problem $S : \mathbb{R}^m_+ \ni \alpha \to u \in \mathbb{R}^n$ that maps each parameter $\alpha \in \mathbb{R}^m_+$ to the corresponding reconstruction $u \in \mathbb{R}^n$. Since this mapping is single-valued, we can make use of it to formulate (3.3)

as a reduced optimization problem

$$\min_{\alpha \in \mathbb{R}^m_+} \quad j(\alpha) := J(S(\alpha), u^{\text{true}}). \tag{5.43}$$

Furthermore, assuming the solution operator is Bouligand (B)-differentiable, i.e., locally Lipschitz continuous and directionally differentiable. Then, we can use the chain rule for B-differentiable functions to conclude that the composite mapping j, as a function of α, is also B-differentiable. In this case, its directional derivative in a direction h is given by

$$j'(\alpha; h) = \langle \nabla J(u), S'(\alpha; h) \rangle, \tag{5.44}$$

where $S'(\alpha; h)$ is the directional derivative of the solution operator in direction h. Moreover, if α^* is a local optimal solution and $u^* = S(\alpha^*)$ its corresponding reconstruction, then it satisfies the following necessary condition:

$$j'(\alpha^*; \alpha - \alpha^*) = \langle \nabla J(u^*), S'(\alpha^*; \alpha - \alpha^*) \rangle \geq 0, \ \forall \alpha \in \mathbb{R}^m_+. \tag{5.45}$$

A point α^* satisfying the necessary condition (5.45) is called *Bouligand (B)-stationary*. This type of stationarity condition is based on the tangent cone to our feasible parameter set and can be interpreted as the counterpart of the implicit programming approach in the discussion of finite-dimensional MPECs, see [51, Lemma 4.2.5].

Therefore, to fully characterize the B-stationarity condition (5.45), we need to prove that the solution map is Lipschitz continuous and obtain a proper expression for the directional derivative of the solution map $S'(\alpha)$.

Indeed, using the analysis presented in Section 3.1, we already argued there exists a unique solution for the lower-level problem, with (5.1b) being a particular case. Consequently, the solution map $S : \mathbb{R}^m_+ \to \mathbb{R}^n$ is singled valued. Moreover, this property will allow us to formulate the following result.

Theorem 5.2. *Let $\mathcal{F}$ in* (5.1b) *be a strongly convex function. Then, the solution operator for the lower-level problem (5.1b) $S : \mathbb{R}^m_+ \ni \alpha \to u \in \mathbb{R}^n$ is Lipschitz continuous.*

Proof. Thanks to Theorem 3.1, we know the lower-level problem has a unique solution. Moreover, $\alpha_1, \alpha_2 \in \mathbb{R}^m_+$ and its corresponding solutions u_1, u_2 satisfy

$$\langle \nabla \mathcal{F}(u_1), v - u_1 \rangle + \sum_{j=1}^{m} (\alpha_1)_j \|(\mathbb{K}v)_j\| - \sum_{j=1}^{m} (\alpha_1)_j \|(\mathbb{K}u_1)_j\| \geq 0, \ \forall v \in \mathbb{R}^n$$

$$\langle \nabla \mathcal{F}(u_2), w - u_2 \rangle + \sum_{j=1}^{m} (\alpha_2)_j \|(\mathbb{K}w)_j\| - \sum_{j=1}^{m} (\alpha_2)_j \|(\mathbb{K}u_2)_j\| \geq 0, \ \forall w \in \mathbb{R}^n.$$

Taking in particular $v = u_2$ and $w = u_1$ and adding the inequalities, it yields

$$\langle \nabla \mathcal{F}(u_2) - \nabla \mathcal{F}(u_1), u_2 - u_1 \rangle \leq \sum_{j=1}^{m} ((\alpha_1)_j - (\alpha_2)_j)(\|(\mathbb{K}u_2)_j\| - \|(\mathbb{K}u_1)_j\|).$$

Moreover, since $\mathcal{F}$ is strongly convex, with constant $c > 0$, and using the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} c\|u_2 - u_1\|^2 &\leq \sum_{j=1}^{m} (\alpha_{2,j} - \alpha_{1,j}) \|(\mathbb{K}(u_2 - u_1))_j\| \leq \|\alpha_2 - \alpha_1\| \sum_{j=1}^{m} \|(\mathbb{K}(u_2 - u_1))_j\|, \\ &\leq m\|\alpha_2 - \alpha_1\| \|\mathbb{K}\| \|u_2 - u_1\|, \end{aligned}$$

where $\|\mathbb{K}\|$ is the operator norm of the linear operator $\mathbb{K}$. Finally, rearranging the terms of the inequality, it yields the result

$$\|u_2 - u_1\| \leq \frac{m}{c} \|K\| \|\alpha_2 - \alpha_1\|.$$

□

5.2.1 Directional Differentiability

Now, we are interested in the differentiability properties of the solution operator for the lower-level problem (5.1b). It will require a sensitivity analysis of the solution operator with respect to the regularization parameter α. Indeed, by taking a perturbed regularization parameter α^t in the primal-dual formulation for the lower-level problem (3.12b), such that $\alpha_j^t = \alpha_j + th_j \geq 0$, we get the following perturbed lower-level problem:

$$\nabla \mathcal{F}(u^t) + \mathbb{K}^\top q^t = 0, \tag{5.46a}$$

$$\langle q_j^t, (\mathbb{K}u^t)_j \rangle - (\alpha_j + th_j)\|(\mathbb{K}u^t)_j\| = 0, \qquad \forall j = 1, \dots, m, \tag{5.46b}$$

$$\|q_j^t\| - (\alpha_j + th_j) \leq 0, \qquad \forall j = 1, \dots, m. \tag{5.46c}$$

Thanks to the uniform boundedness of q^t, there exists a subsequence, denoted the same, that converges to an element $\tilde{q} \in \mathbb{R}^{m \times 2}$, as $t \to 0$. Additionally, using the Lipschitz continuity of the solution operator, we know the following sequence is bounded

$$\left\| \frac{u^t - u}{t} \right\| \leq \frac{m}{c} \left\| \frac{\alpha^t - \alpha}{t} \right\| = \frac{m}{c} \|h\| < \infty.$$

Consequently, we can guarantee the existence of a subsequence of $\{(u^t - u)/t\}$, denoted with the same symbol, satisfying the following limit

$$\lim_{t \to 0} \frac{u^t - u}{t} \to \eta \in \mathbb{R}^n. \tag{5.47}$$

Theorem 5.3. *The limit described in (5.47) satisfies $\eta \in \mathcal{C}(\alpha, u)$ where*

$$\mathcal{C}(\alpha, u) := \left\{ v \in \mathbb{R}^n : \begin{cases} (\mathbb{K}v)_j = 0, & \forall j \in \mathcal{A}_s, \\ \langle q_j, (\mathbb{K}v)_j \rangle = \alpha_j \|(\mathbb{K}v)_j\|, & \forall j \in \mathcal{B}. \end{cases} \right\} \tag{5.48}$$

Proof. By adding the complementarity relationships in (5.3b) and (5.46b), and dividing by t, we get

$$\left\langle \frac{q_j^t - q_j}{t}, (\mathbb{K}u)_j \right\rangle + \left\langle q_j^t, \frac{(\mathbb{K}u^t)_j - (\mathbb{K}u)_j}{t} \right\rangle - \alpha_j \left(\frac{\|(\mathbb{K}u^t)_j\| - \|(\mathbb{K}u)_j\|}{t} \right) - h_j \|(\mathbb{K}u^t)_j\| = 0. \tag{5.49}$$

As previously stated, the uniform boundedness of q^t guarantees the existence of a subsequence that converges to an element $\tilde{q}$. Therefore, taking the limit as $t \to 0$ in (5.49), for $j \in \mathcal{A}_s \cup \mathcal{B}$ it yields

$$\langle \tilde{q}_j, (\mathbb{K}\eta)_j \rangle - \alpha_j \|(\mathbb{K}\eta)_j\| - h_j \|(\mathbb{K}u)_j\| = 0,$$

where we used the Bouligand differentiability of the Euclidean norm. Furthermore, since $(\mathbb{K}u)_j = 0$ for $j \in \mathcal{A}_s \cup \mathcal{B}$, we get that

$$\langle \tilde{q}_j, (\mathbb{K}\eta)_j \rangle - \alpha_j \|(\mathbb{K}\eta)_j\| = 0.$$

Moreover, for $j \in \mathcal{A}_s$, recalling $\alpha_j > 0$ in this index set, we get

$$\alpha_j \|(\mathbb{K}\eta)_j\| = \langle \tilde{q}_j, (\mathbb{K}\eta)_j \rangle \leq \|\tilde{q}_j\| \|(\mathbb{K}\eta)_j\| < \alpha_j \|(\mathbb{K}\eta)_j\|,$$

which only holds if $(\mathbb{K}\eta)_j = 0$ in this index set, finishing the proof. □

Remark 5.1. *If q^1 and q^2 are two different slack variables associated with the parameter α and its corresponding solution u in (3.12), then the two sets $\mathcal{C}_i$ defined as*

follows

$$\mathcal{C}_i(\alpha,u) := \left\{ v\in\mathbb{R}^n : \begin{cases} (\mathbb{K}v)_j = 0, & \text{if } \|q_j^i\| < \alpha_j, \\ \langle q_j^i, (\mathbb{K}v)_j\rangle = \alpha_j\|(\mathbb{K}v)_j\|, & \text{if } (\mathbb{K}u)_j = 0,\ \alpha_j>0,\ \|q_j^i\|=\alpha_j. \end{cases} \right\},\ i=1,2$$

coincide, since $\mathbb{K}^\top q^1 = -\nabla\mathcal{F}(u) = \mathbb{K}^\top q^2$. *Consequently, the set* $\mathcal{C}(\alpha,u)$ *does not depend on the slack variable, only on the solution* u *and the parameter* α. *Hereafter, to simplify the notation, we will omit the arguments in the set notation as follows* $\mathcal{C} := \mathcal{C}(\alpha,u)$.

Lemma 5.4. *The cone* $\mathcal{C}$ *can alternatively be written as*

$$\mathcal{C} = \left\{ v\in\mathbb{R}^n : \langle \mathbb{K}^\top q, v\rangle \geq \sum_{j\in\mathcal{I}} \left\langle \alpha_j \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, (\mathbb{K}v)_j \right\rangle + \sum_{j\in\mathcal{A}_s\cup\mathcal{B}} \alpha_j \|(\mathbb{K}v)_j\| \right\} \tag{5.50}$$

Proof. Let us denote the set on the right-hand side in (5.50) as $\mathcal{M}$. Taking $v\in\mathcal{C}$, as in (5.48), and using its definition, we obtain

$$\begin{aligned} \langle \mathbb{K}^\top q, v\rangle &= \sum_{j\in\mathcal{I}} \langle q_j, (\mathbb{K}v)_j\rangle + \sum_{j\in\mathcal{A}_s} \langle q_j, (\mathbb{K}v)_j\rangle + \sum_{j\in\mathcal{B}} \langle q_j, (\mathbb{K}v)_j\rangle, \\ &= \sum_{j\in\mathcal{I}} \left\langle \alpha_j \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, (\mathbb{K}v)_j \right\rangle + \sum_{j\in\mathcal{A}_s} \langle q_j, \underbrace{(\mathbb{K}v)_j}_{=0}\rangle + \sum_{j\in\mathcal{B}} \alpha_j \|(\mathbb{K}v)_j\|, \end{aligned}$$

and, consequently, $\mathcal{C}\subset\mathcal{M}$.

To prove the reverse inclusion, let us take $v\in\mathcal{M}$. Then, we may rewrite (5.50) as follows

$$\sum_{j\in\mathcal{I}} \left\langle \alpha_j \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, + \right\rangle \sum_{j\in\mathcal{A}_s\cup\mathcal{B}} \langle q_j, (\mathbb{K}v)_j\rangle \geq \sum_{j\in\mathcal{I}} \left\langle \alpha_j \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, + \right\rangle \sum_{j\in\mathcal{A}_s\cup\mathcal{B}} \alpha_j \|(\mathbb{K}v)_j\|.$$

using the Cauchy-Schwarz inequality and $\|q_j\|\leq\alpha_j$ for all $j\in\mathcal{A}_s\cup\mathcal{B}$ we get

$$\sum_{j\in\mathcal{A}_s\cup\mathcal{B}} \alpha_j\|(\mathbb{K}v)_j\| \leq \sum_{j\in\mathcal{A}_s\cup\mathcal{B}} \langle q_j, (\mathbb{K}v)_j\rangle \leq \sum_{j\in\mathcal{A}_s\cup\mathcal{B}} \underbrace{\|q_j\|}_{\leq\alpha_j} \|(\mathbb{K}v)_j\| \leq \sum_{j\in\mathcal{A}_s\cup\mathcal{B}} \alpha_j\|(\mathbb{K}v)_j\|.$$

Consequently, it holds

$$\sum_{j\in\mathcal{A}_s\cup\mathcal{B}} \langle q_j, (\mathbb{K}v)_j\rangle - \sum_{j\in\mathcal{A}_s\cup\mathcal{B}} \alpha_j\|(\mathbb{K}v)_j\| = 0, \tag{5.51}$$

Therefore, for each index in $\mathcal{A}_s \cup \mathcal{B}$ we have

$$\langle q_j, (\mathbb{K}v)_j \rangle - \alpha_j \|(\mathbb{K}v)_j\| = 0.$$

Taking, in particular, $j \in \mathcal{A}_s$, and again using the Cauchy-Schwarz inequality, along with the property $\|q_j\| < \alpha_j$, it yields

$$\alpha_j \|(\mathbb{K}v)_j\| = \langle q_j, (\mathbb{K}v)_j \rangle \leq \underbrace{\|q_j\|}_{<\alpha_j} \|(\mathbb{K}v)_j\| < \alpha_j \|(\mathbb{K}v)_j\|,$$

which implies that $(\mathbb{K}v)_j = 0$ for all $j \in \mathcal{A}_s$, and it follows that $\mathcal{M} \subset \mathcal{C}$, concluding the proof. □

Now, to prove the directional differentiability of the solution operator for the lower-level problem (5.1b), we will first demonstrate the following lemmata.

Lemma 5.5. *Let $\mathbb{R}^m_+ \ni \alpha$ and $\mathbb{R}^m_+ \ni \alpha + th$. Then for every $v \in \mathcal{C}$, it holds*

$$\begin{aligned}\left\langle \mathbb{K}^\top \left(\frac{q^t - q}{t} \right), v \right\rangle \leq \sum_{j \in \mathcal{I}} \frac{\alpha_j}{t} \left\langle \frac{(\mathbb{K}u^t)_j}{\|(\mathbb{K}u^t)_j\|} - \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, (\mathbb{K}v)_j \right\rangle \\ + \sum_{j \in \mathcal{I}} h_j \left\langle \frac{(\mathbb{K}u^t)}{\|(\mathbb{K}u^t)_j\|}, (\mathbb{K}v)_j \right\rangle + \sum_{j \in \mathcal{B}} h_j \|(\mathbb{K}v)_j\| + \sum_{j \in \mathcal{T} \cup \mathcal{I}_0} h_j \|(\mathbb{K}v)_j\|. \quad (5.52)\end{aligned}$$

Proof. Given that $v \in \mathcal{C}$, let us first bound the following product

$$\begin{aligned}\langle \mathbb{K}^\top q^t, v \rangle &= \sum_{j \in \mathcal{I}} \langle q^t_j, (\mathbb{K}v)_j \rangle + \sum_{j \in \mathcal{A}_s} \langle q^t_j, \underbrace{(\mathbb{K}v)_j}_{=0} \rangle + \sum_{j \in \mathcal{B}} \langle q^t_j, (\mathbb{K}v)_j \rangle + \sum_{j \in \mathcal{T} \cup \mathcal{I}_0} \langle q^t_j, (\mathbb{K}v)_j \rangle, \\ &\leq \sum_{j \in \mathcal{I}} \left\langle (\alpha_j + th_j) \frac{(\mathbb{K}u^t)_j}{\|(\mathbb{K}u^t)_j\|}, (\mathbb{K}v)_j \right\rangle + \sum_{j \in \mathcal{B}} (\alpha_j + th_j) \|(\mathbb{K}v)_j\| + \sum_{j \in \mathcal{T} \cup \mathcal{I}_0} th_j \|(\mathbb{K}v)_j\|,\end{aligned}$$

for t sufficiently small, since $u^t \to u$ implies $\mathcal{I}(\alpha, u) \subset \mathcal{I}(\alpha + th, u^t)$, where we used the property $(\mathbb{K}v)_j = 0$ for $j \in \mathcal{A}_s$ and $\alpha_j = 0$ for $j \in \mathcal{T} \cup \mathcal{I}_0$, along with Cauchy-Schwarz inequality and $\|q^t_j\| \leq \alpha_j + th_j$. Now, as $v \in \mathcal{C}$ we know the bound in (5.50) holds, i.e.,

$$\langle \mathbb{K}^\top q, v \rangle \geq \sum_{j \in \mathcal{I}} \left\langle \alpha_j \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, (\mathbb{K}v)_j \right\rangle + \sum_{j \in \mathcal{B}} \alpha_j \|(\mathbb{K}v)_j\|.$$

Therefore,

$$\langle \mathbb{K}^\top (q^t - q), v\rangle \leq \sum_{j\in\mathcal{I}} \alpha_j \left\langle \frac{(\mathbb{K}u^t)_j}{\|(\mathbb{K}u^t)_j\|} - \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, (\mathbb{K}v)_j \right\rangle + \sum_{j\in\mathcal{I}} t h_j \left\langle \frac{(\mathbb{K}u^t)_j}{\|(\mathbb{K}u^t)_j\|}, (\mathbb{K}v)_j \right\rangle + \sum_{j\in\mathcal{B}} t h_j \|(\mathbb{K}v)_j\| + \sum_{j\in\mathcal{T}\cup\mathcal{I}_0} t h_j \|(\mathbb{K}v)_j\|.$$

Finally, dividing both sides by t yields the result. □

Lemma 5.6. *Let $\mathbb{R}^m_+ \ni \alpha$ and $\mathbb{R}^m_+ \ni \alpha + th$. Then, it holds*

$$\left\langle \mathbb{K}^\top \left(\frac{q^t - q}{t}\right), \frac{u^t - u}{t} \right\rangle \geq \sum_{j\in\mathcal{I}} \frac{\alpha_j}{t} \left\langle \frac{(\mathbb{K}u^t)_j}{\|(\mathbb{K}u^t)_j\|} - \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, \frac{(\mathbb{K}u^t)_j - (\mathbb{K}u)_j}{t} \right\rangle + \sum_{j\in\mathcal{I}} h_j \left\langle \frac{(\mathbb{K}u^t)}{\|(\mathbb{K}u^t)_j\|}, \frac{(\mathbb{K}u^t)_j - (\mathbb{K}u)_j}{t} \right\rangle + \frac{1}{t} \sum_{j\in\mathcal{A}\cup\mathcal{I}_0} h_j (\|(\mathbb{K}u^t)_j\| - \|(\mathbb{K}u)_j\|),$$

where $\mathcal{A} = \mathcal{A}_s \cup \mathcal{B} \cup \mathcal{T}$.

Proof. For t small enough, we can split the product by their index set

$$\left\langle \mathbb{K}^\top \left(\frac{q^t - q}{t}\right), \frac{u^t - u}{t} \right\rangle = \sum_{j\in\mathcal{I}} \frac{\alpha_j}{t} \left\langle \frac{(\mathbb{K}u^t)_j}{\|(\mathbb{K}u^t)_j\|} - \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, \frac{(\mathbb{K}u^t)_j - (\mathbb{K}u)_j}{t} \right\rangle + \sum_{j\in\mathcal{I}} h_j \left\langle \frac{(\mathbb{K}u^t)_j}{\|(\mathbb{K}u^t)_j\|}, \frac{(\mathbb{K}u^t)_j - (\mathbb{K}u)_j}{t} \right\rangle + \frac{1}{t} \sum_{j\in\mathcal{A}_s\cup\mathcal{B}} \left\langle q_j^t - q_j, \frac{(\mathbb{K}u^t)_j - (\mathbb{K}u)_j}{t} \right\rangle + \frac{1}{t} \sum_{j\in\mathcal{I}_0\cup\mathcal{T}} \left\langle q_j^t - q_j, \frac{(\mathbb{K}u^t)_j - (\mathbb{K}u)_j}{t} \right\rangle.$$

Focusing, on the index set $\mathcal{A}_s \cup \mathcal{B}$, the complementarity relations in (3.12) and (5.46) yield

$$\begin{aligned}
&\frac{1}{t^2} \sum_{j\in\mathcal{A}_s\cup\mathcal{B}} \langle q_j^t - q_j, (\mathbb{K}u^t)_j - (\mathbb{K}u)_j\rangle \\
&= \frac{1}{t^2} \sum_{j\in\mathcal{A}_s\cup\mathcal{B}} \langle q_j^t, (\mathbb{K}u^t)_j\rangle - \langle q_j^t, (\mathbb{K}u)_j\rangle - \langle q_j, (\mathbb{K}u^t)_j\rangle + \langle q_j, (\mathbb{K}u)_j\rangle, \\
&\geq \frac{1}{t^2} \sum_{j\in\mathcal{A}_s\cup\mathcal{B}} (\alpha_j + t h_j)\|(\mathbb{K}u^t)_j\| - \underbrace{\|q_j^t\|}_{\leq \alpha_j + t h_j} \|(\mathbb{K}u)_j\| - \underbrace{\|q_j\|}_{\leq \alpha_j} \|(\mathbb{K}u^t)_j\| + \alpha_j \|(\mathbb{K}u)_j\|, \\
&\geq \frac{1}{t} \sum_{j\in\mathcal{A}_s\cup\mathcal{B}} h_j (\|(\mathbb{K}u^t)_j\| - \|(\mathbb{K}u)_j\|).
\end{aligned}$$

Using the same analysis over the set $\mathcal{I}_0 \cup \mathcal{T}$ we get

$$\frac{1}{t^2}\sum_{j\in\mathcal{I}_0\cup\mathcal{T}}\langle q_j^t - q_j, (\mathbb{K}u^t)_j - (\mathbb{K}u)_j\rangle = \frac{1}{t^2}\sum_{j\in\mathcal{I}_0\cup\mathcal{T}}\langle q_j^t, (\mathbb{K}u^t)_j\rangle - \langle q_j^t, (\mathbb{K}u)_j\rangle,$$
$$\geq \frac{1}{t^2}\sum_{j\in\mathcal{I}_0\cup\mathcal{T}} th_j\|(\mathbb{K}u^t)_j\| - \underbrace{\|q_j^t\|}_{\leq th_j}\|(\mathbb{K}u)_j\| \geq \frac{1}{t}\sum_{j\in\mathcal{I}_0\cup\mathcal{T}} h_j(\|(\mathbb{K}u^t)_j\| - \|(\mathbb{K}u)_j\|).$$

□

Theorem 5.4. *Let $\alpha \in \mathbb{R}^m_+$ and $h \in \mathbb{R}^n$ be a direction such that $\alpha + th \geq 0$ for t small enough. The solution operator $S : \alpha \to S(\alpha) = u \in \mathbb{R}^n$ is directionally differentiable and its directional derivative $\eta \in \mathcal{C}$ at u, in direction h, is given by the solution of the following variational inequality*

$$\langle \nabla_{uu}\mathcal{F}(u)\eta, v-\eta\rangle + \sum_{j\in\mathcal{I}}\alpha_j\langle T_j(\mathbb{K}\eta)_j, (\mathbb{K}v)_j - (\mathbb{K}\eta)_j\rangle + h_j\left\langle \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, (\mathbb{K}v)_j - (\mathbb{K}\eta)_j\right\rangle$$
$$+\sum_{j\in\mathcal{B}}\frac{h_j}{\alpha_j}\langle q_j, (\mathbb{K}v)_j - (\mathbb{K}\eta)_j\rangle + \sum_{j\in\mathcal{I}_0\cup\mathcal{T}} h_j(\|(\mathbb{K}v)_j\| - \|(\mathbb{K}\eta)_j\|) \geq 0,\ \forall v\in\mathcal{C}, \quad (5.53)$$

where $T_j(\mathbb{K}v)_j = \frac{(\mathbb{K}v)_j}{\|(\mathbb{K}u)_j\|} - \frac{(\mathbb{K}u)_j(\mathbb{K}u)_j^\top(\mathbb{K}v)_j}{\|(\mathbb{K}u)_j\|^3}$ for $v \in \mathbb{R}^n$.

Proof. To verify the variational inequality, let us take (5.46), (3.12) and test them with $v - \frac{u^t-u}{t}$, with $v \in \mathcal{C}$

$$0 = \left\langle \frac{\nabla\mathcal{F}(u^t) - \nabla\mathcal{F}(u)}{t}, v - \frac{u^t-u}{t}\right\rangle + \left\langle \mathbb{K}^\top\left(\frac{q^t-q}{t}\right), v - \frac{u^t-u}{t}\right\rangle,$$
$$= \left\langle \frac{\nabla\mathcal{F}(u^t) - \nabla\mathcal{F}(u)}{t}, v - \frac{u^t-u}{t}\right\rangle + \left\langle \mathbb{K}^\top\left(\frac{q^t-q}{t}\right), v\right\rangle - \left\langle \mathbb{K}^\top\left(\frac{q^t-q}{t}\right), \frac{u^t-u}{t}\right\rangle$$

Now, applying the bounds in Lemmas 5.5 and 5.6 we have

$$0 \leq \left\langle \frac{\nabla\mathcal{F}(u^t) - \nabla\mathcal{F}(u)}{t}, v - \frac{u^t-u}{t}\right\rangle$$
$$+\sum_{j\in\mathcal{I}}\frac{\alpha_j}{t}\left\langle \frac{(\mathbb{K}u^t)_j}{\|(\mathbb{K}u^t)_j\|} - \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, (\mathbb{K}v)_j - \frac{(\mathbb{K}u^t)_j - (\mathbb{K}u)_j}{t}\right\rangle$$
$$+ h_j\left\langle \frac{(\mathbb{K}u^t)}{\|(\mathbb{K}u^t)_j\|}, (\mathbb{K}v)_j - \frac{(\mathbb{K}u^t)_j - (\mathbb{K}u)_j}{t}\right\rangle + \sum_{j\in\mathcal{B}} h_j\|(\mathbb{K}v)_j\|$$
$$+\sum_{j\in\mathcal{I}_0\cup\mathcal{T}} h_j\|(\mathbb{K}v)_j\| - \frac{1}{t}\sum_{j\in\mathcal{A}_s\cup\mathcal{B}\cup\mathcal{I}_0\cup\mathcal{T}} h_j(\|(\mathbb{K}u^t)_j\| - \|(\mathbb{K}u)_j\|).$$

Taking the limit $t \to 0$, as well as the differentiability of the term $x/\|x\|$ in the inactive

set, and given that $(\mathbb{K}\eta)_j = 0$ in the strongly active set $\mathcal{A}_s$, it yields

$$0 \leq \langle \nabla_{uu}\mathcal{F}(u)\eta, v-\eta\rangle + \sum_{j\in\mathcal{I}} \alpha_j \left\langle \left(\frac{I}{\|(\mathbb{K}u)_j\|} - \frac{(\mathbb{K}u)_j (\mathbb{K}u)_j^\top}{\|(\mathbb{K}u)_j\|^3} \right) (\mathbb{K}\eta)_j, (\mathbb{K}v)_j - (\mathbb{K}\eta)_j \right\rangle$$
$$+h_j \left\langle \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, (\mathbb{K}v)_j - (\mathbb{K}\eta)_j \right\rangle + \sum_{j\in\mathcal{B}} h_j(\|(\mathbb{K}v)_j\| - \|(\mathbb{K}\eta)_j\|) + \sum_{j\in\mathcal{I}_0\cup\mathcal{T}} h_j(\|(\mathbb{K}v)_j\| - \|(\mathbb{K}\eta)_j\|).$$

Using the definition for T_j and recalling $v, \eta \in \mathcal{C}$, the inequality takes the form in (5.53).

Now it is required to verify the uniqueness of the limit. For this purpose, let us note that (5.53) is a variational inequality

$$\langle \nabla_{uu}\mathcal{F}(u)\eta, v-\eta\rangle + \sum_{j\in\mathcal{I}} \alpha_j \langle T_j(\mathbb{K}\eta)_j, (\mathbb{K}v)_j - (\mathbb{K}\eta)_j\rangle + \sum_{j\in\mathcal{I}_0\cup\mathcal{T}} h_j(\|(\mathbb{K}v)_j\| - \|(\mathbb{K}\eta)_j\|)$$
$$\geq -\sum_{j\in\mathcal{I}\cup\mathcal{B}} \frac{h_j}{\alpha_j} \left\langle q_j, (\mathbb{K}v)_j - (\mathbb{K}\eta)_j \right\rangle, \forall v \in \mathcal{C}$$

Now, recalling that the function $f(z) := \sum_{j=1}^m \|(\mathbb{K}z)_j\|$ is indeed convex, lower semi-continuous, and proper, the right-hand side is continuous and linear, and finally, using the strong convexity of $\mathcal{F}$ and the positive semi-definiteness of T_j, the bilinear form in the smooth part of the left-hand side is V-elliptic, i.e.,

$$\langle \nabla_{uu}\mathcal{F}(u)v, v\rangle + \sum_{j\in\mathcal{I}} \alpha_j \langle T_j(\mathbb{K}v)_j, (\mathbb{K}v)_j\rangle \geq c\|v\|^2.$$

We know by [34, Chapter I, Theorem 4.1], that there exists a unique solution for this variational inequality. □

Finally, we have proven the following result by using the demonstrated Bouligand differentiability of the solution operator and the corresponding characterization of the directional derivative.

Theorem 5.5. *Let $\alpha^* \in \mathbb{R}^m_+$ be a local optimal solution of (5.43) and $u^* = S(\alpha^*)$. Then α^* is a B-stationary point, i.e., it satisfies the following inequality*

$$\langle \nabla J(u^*), S'(\alpha^*; \alpha - \alpha^*)\rangle \geq 0, \quad \forall \alpha \in \mathbb{R}^m_+, \tag{5.54}$$

where $S'(\alpha^*; \alpha - \alpha^*) =: \eta$ is the unique solution to (5.53).

Proof. Since we know that the solution operator is directionally differentiable, as shown in theorem 5.4, along with its local Lipschitz continuity as demonstrated in theorem 5.2,

we have that the solution operator is Bouligand differentiable. Consequently, a local optimal solution α^* for problem (5.43) and $u^* = S(\alpha^*)$ its optimal reconstruction, satisfy the necessary optimality condition (5.45). □

5.2.2 Strict Complementarity

The characterization of the directional differentiability can take different formulations if any of the active sets becomes empty. For instance, assuming the zero-inactive and triactive sets are empty, i.e., $\mathcal{I}_0 \cup \mathcal{T} = \emptyset$, then the directional derivative of the solution operator can be written as the following variational inequality of the first kind

$$\langle \nabla_{uu}\mathcal{F}(u)\eta, v-\eta\rangle + \sum_{j\in\mathcal{I}} \alpha_j \langle T_j(\mathbb{K}\eta)_j, (\mathbb{K}v)_j - (\mathbb{K}\eta)_j\rangle + h_j \left\langle \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, (\mathbb{K}v)_j - (\mathbb{K}\eta)_j \right\rangle + \sum_{j\in\mathcal{B}} \frac{h_j}{\alpha_j}(\langle q_j, (\mathbb{K}v)_j - (\mathbb{K}\eta)_j\rangle) \geq 0,\ \forall v \in \mathcal{C}. \quad (5.55)$$

Furthermore, assuming an empty biactive set and $\alpha_j > 0$, for all j, we obtain that the solution operator is Fréchet differentiable, as stated in the following theorem.

Theorem 5.6. *Let us assume the index set $\mathcal{B} \cup \mathcal{I}_0 \cup \mathcal{T}$ is empty. Then, the solution operator is Fréchet differentiable, and the derivative can be computed as the solution of the following system of equations for some $\xi \in \mathbb{R}^{m\times 2}$*

$$\nabla_{uu}\mathcal{F}(u)\eta + \mathbb{K}^\top \xi = 0, \quad (5.56a)$$

$$\xi_j - \alpha_j T_j(\mathbb{K}\eta)_j - \frac{h_j}{\alpha_j} q_j = 0, \qquad \forall j \in \mathcal{I}, \quad (5.56b)$$

$$(\mathbb{K}\eta)_j = 0, \qquad \forall j \in \mathcal{A}_s. \quad (5.56c)$$

Proof. Using the empty biactive set assumption, we get that the cone $\mathcal{C}$ becomes the following linear subspace $\mathcal{C} = \{v \in \mathbb{R}^n : (\mathbb{K}v)_j = 0 \text{ if } (\mathbb{K}u)_j = 0\}$. Thus, the variational inequality in (5.55) becomes the following variational equation

$$\langle \nabla_{uu}\mathcal{F}(u)\eta, v-\eta\rangle + \sum_{j\in\mathcal{I}} \alpha_j \langle T_j(\mathbb{K}\eta)_j, (\mathbb{K}v)_j - (\mathbb{K}\eta)_j\rangle + h_j \left\langle \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, (\mathbb{K}v)_j - (\mathbb{K}\eta)_j \right\rangle = 0,\ \forall v \in \mathcal{C}. \quad (5.57)$$

(5.57) guarantees that the directional derivative of the solution operator is a linear mapping w.r.t. the direction h. Since S is Bouligand differentiable, it implies the Fréchet differentiability [74, Proposition 3.1.2]. Furthermore, (5.57) is equivalent to the following optimization problem

$$\min_{\eta \in \mathcal{C}} \quad \frac{1}{2}\langle \eta, \nabla_{uu}\mathcal{F}(u)\eta\rangle + \sum_{j\in\mathcal{I}} \alpha_j \left(\frac{\|(\mathbb{K}\eta)_j\|^2}{\|(\mathbb{K}u)_j\|} - \frac{\langle (\mathbb{K}u)_j, (\mathbb{K}\eta)_j\rangle^2}{\|(\mathbb{K}u)_j\|^3} \right) + h_j \left\langle (\mathbb{K}\eta)_j, \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|} \right\rangle \tag{5.58}$$

Then the KKT-optimality conditions for this problem look as follows

$$\langle \nabla_{uu}\mathcal{F}(u)\eta, v\rangle + \sum_{j\in\mathcal{I}} \alpha_j \langle T_j(\mathbb{K}\eta)_j, (\mathbb{K}v)_j\rangle + h_j \left\langle \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, (\mathbb{K}v)_j \right\rangle + \sum_{j\in\mathcal{A}_s} \langle \nu_j, (\mathbb{K}v)_j\rangle = 0, \ \forall v \in \mathbb{R}^n$$
$$(\mathbb{K}\eta)_j = 0, \forall j \in \mathcal{A}_s,$$

with Lagrange multipliers $\nu_j \in \mathbb{R}^2$. Since all the constraints are linear, the Abadie constraint qualification condition [33, Definition 2.33] is satisfied. By introducing $\xi \in \mathbb{R}^{m\times 2}$ as

$$\xi_j := \begin{cases} \nu_j, & \forall j \in \mathcal{A}_s \\ \alpha_j T_j(\mathbb{K}\eta)_j + \frac{h_j}{\alpha_j} q_j, & \forall j \in \mathcal{I} \end{cases}$$

the result is obtained. □

5.2.3 Bouligand Subdifferential

Even though the Bouligand stationarity condition presented in Section 5.2 holds for any local optimal solution without requiring any constraint qualification, its purely primal form is generally not amenable for algorithmic purposes; this limitation is related to the non-linearity of the directional derivative. As a remedy, in this section, we will focus on studying the Bouligand subdifferential of the solution operator S. Characterizing the linear elements of this subdifferential turns out to be helpful when devising a numerical algorithm to solve the bilevel problem.

Thanks to the local Lipschitz continuity of S, shown in Section 5.2, and Rademacher's theorem, we know the solution operator is differentiable almost everywhere. Furthermore, denoting the set of points where this function is differentiable as D_S, the Bouligand subdifferential of the solution map $\partial_B S(\alpha)$ is defined as in Definition 2.3.

In the next result, a characterization of the elements of the Bouligand subdifferential is provided. We assume only along this section that $\alpha_j > 0$, i.e., $\mathcal{T} \cup \mathcal{I}_0 = \emptyset$.

Theorem 5.7. *Let $G \in \partial_B S(\alpha)$ with $\alpha > 0$ and let us introduce the following subspace*

$$V := \{v \in \mathbb{R}^n : \ (\mathbb{K}v)_j = 0, \ \forall j \in \mathcal{A}_s \cup \mathcal{B}_1; (\mathbb{K}v)_j \in \operatorname{span}(q_j), \forall j \in \mathcal{B}_2\} \tag{5.59}$$

Then, there exists a partition of the biactive set $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$ such that, for any $h \in \mathbb{R}^m$

such that $\alpha + th \geq 0$, $Gh =: \tilde{\eta} \in V$ *is the solution of the system*

$$\langle \nabla_{uu}\mathcal{F}(u)\tilde{\eta}, v\rangle + \sum_{j\in\mathcal{I}\cup\mathcal{B}_2} \langle \tilde{\xi}_j, (\mathbb{K}v)_j\rangle = 0, \qquad \forall v \in V \tag{5.60a}$$

$$\tilde{\xi}_j - \alpha_j T_j (\mathbb{K}\tilde{\eta})_j - \frac{h_j}{\alpha_j} q_j = 0, \qquad \forall j \in \mathcal{I}, \tag{5.60b}$$

$$\tilde{\xi}_j - \frac{h_j}{\alpha_j} q_j = 0, \qquad \forall j \in \mathcal{B}_2 \tag{5.60c}$$

Proof. We know the solution operator is locally Lipschitz continuous (see Theorem 5.2), which implies it is differentiable almost everywhere. Let us consider a sequence $\{\alpha_k\} \subset D_S$ such that $\alpha_k \to \alpha$ and $S'(\alpha_k) \to G$. Since we assumed that $\alpha_j > 0$, we know that for k sufficiently large, $(\alpha_k)_j > 0$, for all $j = 1, \ldots, n$. Moreover, thanks to the Lipschitz continuity of S, we know that

$$u_k = S(\alpha_k) \to S(\alpha) = u,$$
$$\mathbb{K}^\top q_k = -\nabla\mathcal{F}(u_k) \to -\nabla\mathcal{F}(u) = \mathbb{K}^\top q.$$

This last statement follows from the fact that $u_k \to u$ and the continuity of $\nabla\mathcal{F}$. Now, each of this subsequence elements (u_k, q_k) define their respective inactive $\mathcal{I}^k := \mathcal{I}(\alpha_k, u_k)$ and strongly active $\mathcal{A}_s^k := \mathcal{A}_s(\alpha_k, u_k)$ sets.
By continuity, we know that $\mathcal{I} \subset \mathcal{I}^k$ and $\mathcal{A}_s \subset \mathcal{A}_s^k$, for k sufficiently large. Introducing the subspace $V^k := \{v \in \mathbb{R}^n : (\mathbb{K}v)_j = 0, \forall j \in \mathcal{A}_s^k\}$ and since $\{\alpha_k\} \subset D_S$, it follows that, for $h \in \mathbb{R}^m$, we have that the directional derivative of the solution operator in direction h, i.e., $S'(\alpha_k)h =: \eta_k \in V^k$ satisfies the system

$$\nabla_{uu}\mathcal{F}(u_k)\eta_k + \mathbb{K}^\top \xi_k = 0, \tag{5.61a}$$

$$(\xi_k)_j - (\alpha_k)_j (T_k)_j (\mathbb{K}\eta_k)_j = \frac{h_j}{(\alpha_k)_j} \mathbb{K}_j^\top (q_k)_j, \qquad \forall j \in \mathcal{I}^k, \tag{5.61b}$$

$$(\mathbb{K}\eta_k)_j = 0, \qquad \forall j \in \mathcal{A}_s^k, \tag{5.61c}$$

or equivalently,

$$\langle \nabla_{uu}\mathcal{F}(u_k)\eta_k, v\rangle + \sum_{j\in\mathcal{I}^k} \left\langle (\alpha_k)_j (T_k)_j (\mathbb{K}\eta_k)_j, (\mathbb{K}v)_j \right\rangle + \frac{h_j}{(\alpha_k)_j} \langle (q_k)_j, (\mathbb{K}v)_j\rangle = 0, \forall v \in V^k, \tag{5.62}$$

From the definition of the Bouligand subdifferential it follows that $\tilde{\eta} = \lim_{k\to\infty} \eta_k$. Moreover, since for $j \in \mathcal{I}$ the sequence $\{(\xi_k)_j\}$ is bounded, then there exists a subsequence that converges to a limit point $\tilde{\xi}_j$. Therefore, up to a subsequence, by passing

to the limit, we get

$$\tilde{\xi}_j - \alpha_j T_j (\mathbb{K}\tilde{\eta})_j - \frac{h_j}{\alpha_j} q_j = 0, \ \forall j \in \mathcal{I},$$
$$(\mathbb{K}\tilde{\eta})_j = 0, \ \forall j \in \mathcal{A}_s.$$

Let us now consider a partition of the biactive set $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$, with

$$\mathcal{B}_1 := \{j \in \mathcal{B} : \exists \{u_{k_l}\} : (\mathbb{K}u_{k_l})_j = 0, \forall l\} \quad \text{and} \quad \mathcal{B}_2 := \mathcal{B} \backslash \mathcal{B}_1.$$

In the index set $\mathcal{B}_1$ we know that $(\mathbb{K}u_{k_l})_j = 0, \forall l$, i.e., the components are strongly active. Consequently, from (5.61), it follows that the subsequence $(\mathbb{K}\eta_{k_l})_j = 0$, for all l. Since $\eta_k \to \tilde{\eta}$, we get that

$$(\mathbb{K}\tilde{\eta})_j = 0, \ \forall j \in \mathcal{A}_s \cup \mathcal{B}_1.$$

Considering the partition $\mathcal{B}_2$, we approach a biactive point by a sequence of points such that $(\mathbb{K}u_k)_j \neq 0$, i.e., $j \in \mathcal{I}^k$. Let us first notice that the term on the right-hand side of (5.61b) is uniformly bounded and, therefore, as $k \to \infty$,

$$\sum_{j \in \mathcal{I}^k} \frac{h_j}{(\alpha_k)_j} \langle (q_k)_j, (\mathbb{K}v)_j \rangle \to \sum_{j \in \mathcal{I} \cup \mathcal{B}_2} \frac{h_j}{\alpha_j} \langle q_j, (\mathbb{K}v)_j \rangle, \quad \forall v \in V.$$

In addition, defining $(\zeta_k)_j = (\alpha_k)_j (T_k)_j (\mathbb{K}\eta_k)_j$, for $j \in \mathcal{I}^k$, we get that

$$\langle (\zeta_k)_j, (\mathbb{K}\eta_k)_j \rangle = \frac{(\alpha_k)_j}{\|(\mathbb{K}u_k)_j\|} \left(\|(\mathbb{K}\eta_k)_j\|^2 - \frac{1}{\|(\mathbb{K}u_k)_j\|^2} \langle (\mathbb{K}\eta_k)_j, (\mathbb{K}u_k)_j \rangle^2 \right) \geq 0, \ \forall j \in \mathcal{I}^k. \tag{5.64}$$

Using the positivity of the term $\langle (\zeta_k)_j, (\mathbb{K}\eta_k)_j \rangle$ we have

$$0 \leq \langle (\zeta_k)_j, (\mathbb{K}\eta_k)_j \rangle \leq \sum_{j \in \mathcal{I}^k} \langle (\zeta_k)_j, (\mathbb{K}\eta_k)_j \rangle. \tag{5.65}$$

Furthermore, using the semi-positive definiteness of $\nabla_{uu}\mathcal{F}(u_k)$, we may upper bound (5.62) for $v = \eta_k$, as follows

$$\begin{aligned} \sum_{j \in \mathcal{I}^k} \langle (\zeta_k)_j, (\mathbb{K}\eta_k)_j \rangle &\leq \langle \nabla_{uu}\mathcal{F}(u_k)\eta_k, \eta_k \rangle + \sum_{j \in \mathcal{I}^k} \langle (\zeta_k)_j, (\mathbb{K}\eta_k)_j \rangle, \\ &= -\sum_{j \in \mathcal{I}^k} \frac{h_j}{(\alpha_k)_j} \langle (q_k)_j, (\mathbb{K}\eta_k)_j \rangle, \\ &\leq \sum_{j \in \mathcal{I}^k} |h_j| \|(\mathbb{K}\eta_k)_j\|. \end{aligned} \tag{5.66}$$

Consequently, joining bounds (5.64)–(5.66), it reads

$$0 \leq \langle (\zeta_k)_j, (\mathbb{K}\eta_k)_j \rangle \leq \langle \nabla_{uu}\mathcal{F}(u_k)\eta_k, \eta_k \rangle + \sum_{j\in\mathcal{I}^k} \langle (\zeta_k)_j, (\mathbb{K}\eta_k)_j \rangle \leq \sum_{j\in\mathcal{I}^k} |h_j| \|(\mathbb{K}\eta_k)_j\|, \tag{5.67}$$

which, since $\eta_k \to \tilde{\eta}$, as $k \to \infty$, implies that $\langle (\zeta_k)_j, (\mathbb{K}\eta_k)_j \rangle$ is uniformly bounded. Since for $j \in \mathcal{B}_2$ we know that $(\mathbb{K}u_k)_j \to 0$, it follows from the previous relations that

$$\alpha_j^2 \|(\mathbb{K}\tilde{\eta})_j\|^2 - \langle q_j, (\mathbb{K}\tilde{\eta})_j \rangle^2 = \lim_{k\to\infty} (\alpha_k)_j^2 \|(\mathbb{K}\eta_k)_j\|^2 - \langle (q_k)_j, (\mathbb{K}\eta_k)_j \rangle^2 = 0,$$

which implies that $(\mathbb{K}\tilde{\eta})_j \in \operatorname{span}(q_j), \forall j \in \mathcal{B}_2$. Consequently, we have shown that $\tilde{\eta} \in V$.

Now, when taking the limit as $k \to \infty$ in (5.62), there may exist sequences in $\mathcal{I}^k$ that converge to a component in $\mathcal{B}_2$. To verify that this does not occur, let us take a $v \in V$ and find the limit for the following term

$$\begin{aligned}
\lim_{k\to\infty} \langle (\zeta_k)_j, (\mathbb{K}v)_j \rangle &= \lim_{k\to\infty} \langle (\zeta_k)_j, c(q_k)_j \rangle = \lim_{k\to\infty} \left\langle (\zeta_k)_j, c(\alpha_k)_j \frac{(\mathbb{K}u_k)_j}{\|(\mathbb{K}u_k)_j\|} \right\rangle, \\
&= c \lim_{k\to\infty} \left\langle (\alpha_k)_j \frac{(\mathbb{K}\eta_k)_j}{\|(\mathbb{K}u_k)_j\|}, (\alpha_k)_j \frac{(\mathbb{K}u_k)_j}{\|(\mathbb{K}u_k)_j\|} \right\rangle \\
&\qquad - \left\langle (\alpha_k)_j \frac{\langle (\mathbb{K}\eta_k)_j, (\mathbb{K}u_k)_j \rangle (\mathbb{K}u_k)_j}{\|(\mathbb{K}u_k)_j\|^3}, (\alpha_k)_j \frac{(\mathbb{K}u_k)_j}{\|(\mathbb{K}u_k)_j\|} \right\rangle, \\
&= c \lim_{k\to\infty} \frac{(\alpha_k)_j^2}{\|(\mathbb{K}u_k)_j\|^2} \langle (\mathbb{K}\eta_k)_j, (\mathbb{K}u_k)_j \rangle \\
&\qquad - \frac{(\alpha_k)_j^2}{\|(\mathbb{K}u_k)_j\|^4} \langle (\mathbb{K}\eta_k)_j, (\mathbb{K}u_k)_j \rangle \langle (\mathbb{K}u_k)_j, (\mathbb{K}u_k)_j \rangle = 0.
\end{aligned}$$

Consequently, we can see that this product's limit vanishes for sequences coming from components either from $\mathcal{A}_s \cup \mathcal{B}_1$, where $(\mathbb{K}v)_j = 0$, and from $\mathcal{B}_2$ as $k \to \infty$. Therefore, taking the limit as $k \to \infty$ in (5.62), yields the result. □

Corollary 5.1. *Let $G \in \partial_B S(\alpha)$. There exists a partition of the biactive set $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$ and a multiplier $\theta \in \mathbb{R}^n$ such that, for any h such that $\alpha + th \geq 0$, $\tilde{\eta} := Gh$ is the unique solution of the system*

$$\nabla_{uu}\mathcal{F}(u)\tilde{\eta} + \mathbb{K}^T\theta = 0 \tag{5.68a}$$

$$\theta_j - \alpha_j T_j (\mathbb{K}\tilde{\eta})_j - \frac{h_j}{\alpha_j} q_j = 0, \qquad \forall j \in \mathcal{I}, \tag{5.68b}$$

$$\langle \theta_j, q_j \rangle - \alpha_j h_j = 0, \qquad \forall j \in \mathcal{B}_2. \tag{5.68c}$$

Proof. Let us consider the functional $M \in \mathbb{R}^n$ defined by

$$(M, v) := (\nabla_{uu}\mathcal{F}(u)\tilde{\eta}, v) + \sum_{j\in\mathcal{I}} \langle \alpha_j T_j (\mathbb{K}\tilde{\eta})_j, (\mathbb{K}v)_j \rangle + \sum_{j\in\mathcal{I}\cup\mathcal{B}_2} \frac{h_j}{\alpha_j} \langle q_j, (\mathbb{K}v)_j \rangle, \ \forall v \in V.$$

(5.60a) can then be written as $M \in V^\perp$. Thanks to the structure of the linear subspace V, it can be represented in a separate way as $V = \left(\bigcap_{j\in\mathcal{A}_S\cup\mathcal{B}_1} V_j^1\right) \cap \left(\bigcap_{j\in\mathcal{B}_2} V_j^2\right)$, where

$$\begin{aligned} V_j^1 &:= \{v \in \mathbb{R}^n : (\mathbb{K}v)_j = 0\}, && j \in \mathcal{A}_S \cup \mathcal{B}_1, \\ V_j^2 &:= \{v \in \mathbb{R}^n : (\mathbb{K}v)_j \in \text{span}(q_j)\}, && j \in \mathcal{B}_2. \end{aligned}$$

Consequently, $V^\perp = \sum_{j\in\mathcal{A}_S\cup\mathcal{B}_1} (V_j^1)^\perp + \sum_{j\in\mathcal{B}_2} (V_j^2)^\perp$.

For $j \in \mathcal{A}_S \cup \mathcal{B}_1$, we get that $(V_j^1)^\perp = \ker(\mathbb{K}_j)^\perp$. Thanks to the orthogonality relations, it follows that $\ker(\mathbb{K}_j)^\perp = \text{range}(\mathbb{K}_j^\top)$. Hence, for any $\xi_j \in (V_j^1)^\perp$, there exist π_j such that $\xi_j = \mathbb{K}_j^\top \pi_j$. Consequently,

$$\sum_{j\in\mathcal{A}_S\cup\mathcal{B}_1} (V_j^1)^\perp = \sum_{j\in\mathcal{A}_S\cup\mathcal{B}_1} \mathbb{K}_j^\top \pi_j, \quad \pi_j \in \mathbb{R}^2.$$

For $j \in \mathcal{B}_2$, any $v \in V_j^2$ can be represented as a sum of an element from the nullspace and the row space of $\mathbb{K}_j$, see Theorem 2.2, as follows

$$v = \phi + \varphi, \quad \text{with } (\mathbb{K}_j\varphi) = 0 \text{ and } \phi \in \text{range}(\mathbb{K}_j^\top).$$

Since $(\mathbb{K}v)_j \in \text{span}(q_j)$ and $(\mathbb{K}_j\varphi) = 0$, it follows that $(\mathbb{K}v)_j \in \text{span}(q_j)$ as well. Let us now consider $w_j \in (V_j^2)^\perp$, which can be represented as $w_j = \tilde{w}_j + \hat{w}_j$, where $\tilde{w}_j \in \text{range}(\mathbb{K}_j^\top)$ and $\hat{w}_j \in \text{range}(\mathbb{K}_j^\top)^\perp = \ker(\mathbb{K}_j)$. Consequently, there exists ψ_j such that

$$w_j = \mathbb{K}_j^\top \psi_j + \hat{w}_j, \quad \text{with } \mathbb{K}_j\hat{w}_j = 0.$$

Taking the scalar product with $v_j \in V_j^2$, we get

$$(w_j, v_j) = (\mathbb{K}_j^\top\psi_j + \hat{w}_j, \phi + \varphi) = \langle \psi_j, \mathbb{K}_j\phi \rangle + (\hat{w}_j, \mathbb{K}_j^\top\psi) + (\hat{w}_j, \varphi) = c\langle \psi_j, q_j \rangle + (\hat{w}_j, \varphi),$$

since $\mathbb{K}_j\varphi = \mathbb{K}_j\hat{w}_j = 0$. For the product to be zero, it is then required that $(\hat{w}_j, \varphi) = 0, \forall\varphi \in \ker(\mathbb{K}_j)$ and $\langle \psi_j, q_j \rangle = 0$. Since $\hat{w}_j$ belongs to $\ker(\mathbb{K}_j)$ as well, it follows that $\hat{w}_j = 0$. Consequently,

$$\sum_{j\in\mathcal{B}_2} (V_j^2)^\perp = \sum_{j\in\mathcal{B}_2} \mathbb{K}_j^\top \psi_j, \quad \psi_j \in \mathbb{R}^2 : \langle \psi_j, q_j \rangle = 0.$$

Altogether, we then obtain that there exist multipliers π_j and ψ_j such that

$$M + \sum_{j\in\mathcal{A}_S\cup\mathcal{B}_1} \mathbb{K}_j^\top \pi_j + \sum_{j\in\mathcal{B}_2} \mathbb{K}_j^\top \psi_j = 0,$$

with $\langle \psi_j, q_j\rangle = 0$. Defining

$$\theta_j := \begin{cases} \alpha_j T_j(\mathbb{K}\tilde{\eta})_j + \frac{h_j}{\alpha_j} q_j, & j \in \mathcal{I}, \\ \pi_j, & j \in \mathcal{A}_S \cup \mathcal{B}_1, \\ \psi_j + \frac{h_j}{\alpha_j} q_j, & j \in \mathcal{B}_2, \end{cases}$$

the result is obtained. □

Next, we verify that along a given direction, there exists a solution of system (5.60) which coincides with the directional derivative. When properly characterized, this enables us to use a linear representative of the (otherwise nonlinear) directional derivative within a solution algorithm.

Theorem 5.8. *For any $\alpha \in \mathbb{R}^m_+$ and $h \in \mathbb{R}^m$ such that $\alpha + th \geq 0$, there exists a linear element $\tilde{\eta} = Gh$ such that $S'(\alpha)h = Gh$.*

Proof. Let us recall that, since by assumption $\mathcal{T} \cup \mathcal{I}_0 = \emptyset$, the directional derivative of the solution mapping, in direction h, is given by the unique $\eta \in \mathcal{C}$ solution of

$$\begin{aligned} \langle \nabla_{uu}\mathcal{F}(u)\eta, v-\eta\rangle + \sum_{j\in\mathcal{I}} \langle \alpha_j T_j(\mathbb{K}\eta)_j, (\mathbb{K}v)_j - (\mathbb{K}\eta)_j\rangle \geq \\ - \sum_{j\in\mathcal{I}} h_j \left\langle \frac{(\mathbb{K}u)_j}{\|(\mathbb{K}u)_j\|}, (\mathbb{K}v)_j - (\mathbb{K}\eta)_j \right\rangle - \sum_{j\in\mathcal{B}} \frac{h_j}{\alpha_j} \langle q_j, (\mathbb{K}v)_j - (\mathbb{K}\eta)_j\rangle, \end{aligned} \quad (5.69)$$

for all $v \in \mathcal{C}$. Considering the sets $\mathcal{B}_1 := \{j \in \mathcal{B} : \ (\mathbb{K}\eta)_j = 0\}$ and $\mathcal{B}_2 := \mathcal{B}\backslash\mathcal{B}_1$, and since $\eta \in \mathcal{C}$, it also follows that $(\mathbb{K}\eta) = c_j q_j$, for all $j \in \mathcal{B}_2$, for some $c_j > 0$. Consequently, η belongs to the subspace

$$V := \{v \in \mathbb{R}^n : \ (\mathbb{K}v)_j = 0, \ \forall j \in \mathcal{A}_s \cup \mathcal{B}_1; \ (\mathbb{K}v)_j \in \operatorname{span}(q_j), \ \forall j \in \mathcal{B}_2\}.$$

Moreover, for any $w \in V$ it follows that, for t sufficiently small, $\eta \pm tw \in \mathcal{C}(u)$. Testing (5.69) with these vectors, we then get that

$$\langle \nabla_{uu}\mathcal{F}(u)\eta, w\rangle + \sum_{j\in\mathcal{I}} \langle \alpha_j T_j(\mathbb{K}\eta)_j, (\mathbb{K}w)_j\rangle = - \sum_{j\in\mathcal{I}\cup\mathcal{B}_2} \frac{h_j}{\alpha_j} \langle q_j, (\mathbb{K}w)_j\rangle, \quad \forall w \in V,$$

and, consequently, the directional derivative takes the form $\eta = Gh$, solution of (5.60),

with $\mathcal{B}_2$ as defined above. □

5.3 Nonsmooth Trust Region Algorithm

In this section, we describe the numerical algorithm used for finding optimal parameters of (5.1). Thanks to the Bouligand subdifferential characterization given in Section 5.2.3, it is possible to compute a linearized representative of the directional derivative via the linear system (5.60). Using this information, we can make use of a descent-like algorithm for its numerical solution. Indeed, by using the uniqueness properties of the solution operator, we can write a *reduced optimization problem*:

$$\min_{\alpha \in \mathbb{R}^m_+} \quad j(u(\alpha)), \tag{5.70}$$

where $u(\alpha)$ is the image reconstruction corresponding to a particular value of α (see (5.43) as well). With this reduced problem, we can use the stationarity condition for the bilevel problem described in (5.44) and the directional derivative characterization (5.56). By using the definition of the directional derivative for the reduced optimization problem, we get

$$\langle j'(\alpha), h \rangle = \langle \nabla J(u), S'(\alpha; h) \rangle = \langle \nabla J(u), \tilde{\eta} \rangle. \tag{5.71}$$

where $\tilde{\eta}$ is a solution of system (5.60) for a particular partition of the biactive set $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$ and $S'(\alpha; h)$ is the directional derivative of the solution operator at α in direction h. Let us now define a *generalized adjoint* $p \in \mathbb{R}^n$ as the solution of the following system:

$$\begin{aligned} \langle \nabla_{uu} \mathcal{F}(u)^\top p, v \rangle + \sum_{j \in \mathcal{I}} \langle \mu_j, (\mathbb{K}v)_j \rangle - \langle \nabla J(u), v \rangle = 0, \qquad & \forall v \in V, \\ \mu_j - \alpha_j T_j (\mathbb{K}p)_j = 0, \qquad & \forall j \in \mathcal{I}, \end{aligned}$$

where V is defined as in Theorem 5.7. Using the results in Theorem 5.8, we know that $\tilde{\eta} \in V$ is a linear representative of the directional derivative. Consequently, (5.71) reads

$$\langle j'(\alpha), h \rangle = \langle \nabla J(u), \tilde{\eta} \rangle = \langle \nabla_{uu} \mathcal{F}(u)^\top p, \tilde{\eta} \rangle + \sum_{j \in \mathcal{I}} \langle \alpha_j T_j (\mathbb{K}p)_j, (\mathbb{K}\tilde{\eta})_j \rangle.$$

Rearranging the terms, we get

$$\langle j'(\alpha), h \rangle = \langle p, \nabla_{uu} \mathcal{F}(u) \tilde{\eta} \rangle + \sum_{j \in \mathcal{I}} \langle (\mathbb{K}p)_j, \alpha_j T_j (\mathbb{K}\tilde{\eta})_j \rangle = \langle p, \nabla_{uu} \mathcal{F}(u) \tilde{\eta} \rangle + \sum_{j \in \mathcal{I}} \langle (\mathbb{K}p)_j, \tilde{\lambda}_j \rangle.$$

Finally, using (5.56) we get

$$\langle j'(\alpha), h\rangle = \langle g, h\rangle = -\sum_{j\in\mathcal{I}\cup\mathcal{B}_2} \frac{h_j}{\alpha_j}\langle q_j, (\mathbb{K}p)_j\rangle. \tag{5.72}$$

In our algorithm, we will use the nonsmooth trust-region method detailed in Section 5.3. This method uses two different model functions and switches between them according to the size of the trust-region radius. Along the same line, we will implement two models in this algorithm. A non-regularized model that use (5.72) for a particular partition of the biactive set $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$ to build a quadratic model. Indeed, if the value for the radius runs above a predetermined threshold value Δ_t, the model reads

$$\begin{aligned} m_k(\alpha_k + d_k) &= j(\alpha_k) - \langle g_k, d_k\rangle + \frac{1}{2}\langle d_k, B_k d_k\rangle, \\ &= j(\alpha_k) - \sum_{j\in\mathcal{I}\cup\mathcal{B}_2} \frac{(d_k)_j}{(\alpha_k)_j}\langle (q_k)_j, (\mathbb{K}p_k)_j\rangle + \frac{1}{2}\langle d_k, B_k d_k\rangle \end{aligned}$$

where B_k is a BFGS approximation of the Hessian matrix.

Furthermore, when the trust-region radius falls below a threshold value, the algorithm switches the model based on a gradient built using a regularization, as presented in Section 3.5. In particular, considering the KKT optimality system for the smooth bilevel problem (3.18) with $\lambda_j = 1$, we can find the *regularized adjoint* $p_\gamma \in \mathbb{R}^n$ as the solution of the following system

$$\begin{aligned} &\nabla_u J(u) + \nabla_{uu}\mathcal{F}(u)^\top p_\gamma + \mathbb{K}^\top \beta = 0, \\ &\beta_j - \alpha_j h''_\gamma((\mathbb{K}u)_j)^\top (\mathbb{K}p_\gamma)_j = 0, \qquad\qquad \forall j = 1, \dots, m. \end{aligned}$$

Indeed, we can make use of the regularized adjoint p_γ, we know the derivative of the regularized reduced cost function $j(\alpha) = J(u_\gamma(\alpha))$ reads

$$(j'(\alpha))_j = (g_\gamma)_j = \langle h'_\gamma((\mathbb{K}u)_j), (\mathbb{K}p_\gamma)_j\rangle. \tag{5.73}$$

Consequently, the model to be used when the trust-region radius falls below the threshold value reads

$$m_k(\alpha_k + d_k) = j(\alpha_k) - \langle g_{\gamma,k}, d_k\rangle + \frac{1}{2}\langle d_k, B_k d_k\rangle.$$

Finally, using both the element of the Bouligand subdifferential, the step selection procedure, and the regularized gradient described in this section, we can propose a trust-region algorithm for solving this bilevel problem following the steps that are provided in Algorithm 5.1.

Algorithm 5.1 Non-smooth Trust-Region for Learning the Regularization Weight

1: Choose initial parameter α_0, radius Δ_0, $0 < \eta_1 \leq \eta_2 < 1$, $0 < \gamma_1 \leq 1 \leq \gamma_2$ and $tol > 0$
2: Choose initial second order matrix B_0 and a threshold radius Δ_t
3: Compute $j(\alpha_0)$ and set $k = 0$.
4: **while** $\Delta_k > tol$ **do**
5: **if** $\Delta_k >= \Delta_t$ **then**
6: Compute a linear element of the Bouligand subdifferential g_k at α_k as the solution of (5.72) for a particular partition of the biactive set $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$.
7: Build the model function as: $m_k(\alpha_k + d_k) = j(\alpha_k) + g_k^\top d_k + \frac{1}{2} d_k^\top B_k d_k$.
8: **else**
9: Compute a regularized gradient $g_{\gamma,k}$ at α_k using (5.73).
10: Build the model function as: $m_k(\alpha_k + d_k) = j(\alpha_k) + g_{\gamma,k}^\top d_k + \frac{1}{2} d_k^\top B_k d_k$.
11: **end if**
12: Compute a step s_k that "sufficiently" reduces the model m_k such that $\alpha_k + s_k \in B_{\Delta_k}$
13: Update second order matrix B_k using limited memory BFGS.
14: Calculate the predicted and actual reduction

$$pred_k = m_k(\alpha_k) - m_k(\alpha_k + s_k),$$
$$ared_k = j(\alpha_k) - j(\alpha_k + s_k).$$

15: Compute the quality measure $\rho_k = ared_k / pred_k$.
16: $\alpha_{k+1} = \begin{cases} \alpha_k & \text{if } \rho_k \leq \eta_1, \\ \alpha_k + s_k & \text{otherwise.} \end{cases}$
17: $\Delta_{k+1} = \begin{cases} \gamma_2 \Delta_k & \text{if } \rho_k \geq \eta_2, \\ \Delta_k & \text{if } \rho_k \in (\eta_1, \eta_2), \\ \gamma_1 \Delta_k & \text{if } \rho_k < \eta_1. \end{cases}$
18: $k \leftarrow k + 1$
19: **end while**
20: **return** α_k

5.4 Numerical Experiments

In this section we report on the performance of Algorithm 5.1 presented in Section 5.3 to find optimal parameters. For the upper-level cost function, we chose the following quadratic function

$$J(u, \bar{u}) := \frac{1}{N} \sum_{k=1}^{N} \|u_k - \bar{u}_k\|^2$$

and for the lower level problem, we will consider the patch parameter total variation denoising

$$u_k = \arg\min_u \frac{1}{2}\|u - f\|^2 + \sum_{j=1}^{m} \mathcal{P}(\alpha)_j \|(\mathbb{K}u)_j\|$$

where $\alpha \in \mathbb{R}^p_+$ with $p << m$ and $\mathcal{P} : \mathbb{R}^p \to \mathbb{R}^m$ is a linear patch operator defined as follows

$$\mathcal{P}(\alpha) := \alpha_{\sqrt{p}\times\sqrt{p}} \otimes \mathbb{I}_{\frac{\sqrt{m}}{\sqrt{p}}\times\frac{\sqrt{m}}{\sqrt{p}}} \in \mathbb{R}^{\sqrt{m}\times\sqrt{m}},$$

here, $\otimes$ is the Kronecker product, $\lambda_{\sqrt{p}\times\sqrt{p}}$ is a matrix built by reordering the elements of λ into a matrix of size $\sqrt{p} \times \sqrt{p}$, and $\mathbb{I}_{\frac{\sqrt{m}}{\sqrt{p}}\times\frac{\sqrt{m}}{\sqrt{p}}}$ is a matrix of ones of size $\frac{\sqrt{m}}{\sqrt{p}} \times \frac{\sqrt{m}}{\sqrt{p}}$. This product outputs a matrix of size $\sqrt{m} \times \sqrt{m}$ that is reshaped into a vector of m components.

For all experiments we chose $\Delta_0 = 0.1$, $\Delta_t = 1 \cdot 10^{-4}$ and the parameter $\gamma = 1 \cdot 10^{-5}$ for the local regularization in the second phase of the algorithm. With this goal in mind, we prepared two training image datasets.

5.4.1 Single Training Pair

The first experiment we will explore is a single 128 by 128-pixel image training pair based on the cameraman image and a corrupted version, obtained by adding Gaussian noise with zero mean and standard deviation $\sigma = 0.05$. Figure 5.2 shows this training image pair along with the optimal parameter obtained using the trust-region algorithm for both a scalar and a two-dimensional patch parameter. An improvement in the reconstruction quality when using a two-dimensional patch parameter is verified according to the SSIM value of the image reconstructions. Moreover, when using a scalar and a two-dimensional parameter, the reduced cost function is shown in Figure 5.3, where the non-convexity of the reduced cost functions is inferred.

Table 5.3 shows the reconstruction quality measures on the cameraman training dataset when using different patches sizes, along with the number of trust-region iterations, cost evolution, the number of function evaluations *nfev* and both the number of times the gradient *ngev* and smooth regularized gradient *nreggev* were used. As expected, an improvement in the quality of the reconstructed images can be confirmed

Original	SSIM=0.7494 PSNR=23.1698	$\alpha^* = 0.0155$ SSIM=0.8430 PSNR=23.8045	$\alpha^* = [0.0233; 0.0126]$ SSIM=0.8537 PSNR=23.8345

Figure 5.2: Optimal reconstructions using a scalar regularization parameter and a 2 dimensional regularization parameter.

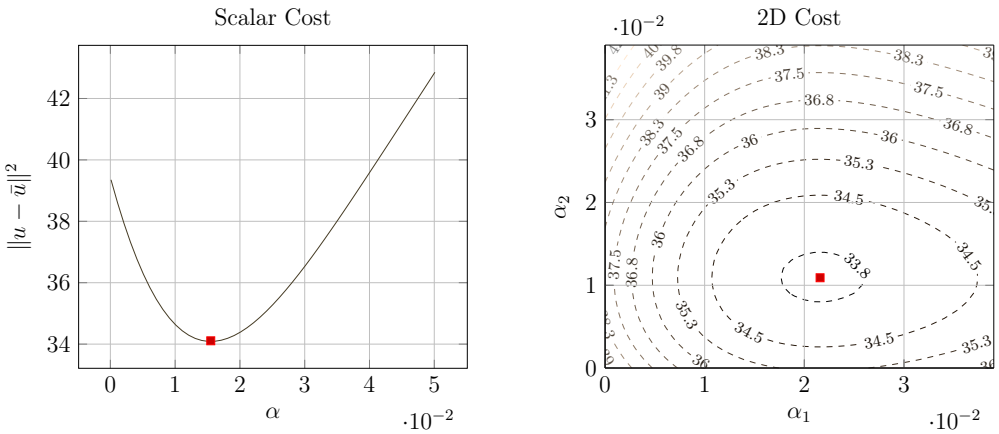


Figure 5.3: Values for the l_2 squared cost function using a scalar regularization parameter and a two-dimensional regularization parameter using the Cameraman training pair.

								Reconstruction		
patch	γ	iterations	nfev	ngev	nreggev	cond mean	time (s)	COST	SSIM	PSNR
	10^2	13	15	-	13	19.65	47.70	33.681	0.853	23.860
	10^3	10	12	-	10	18.58	41.39	33.676	0.854	23.860
2×2	10^4	6	8	-	6	17.56	31.02	33.675	0.855	23.861
	10^5	6	8	-	6	17.59	38.93	33.675	0.855	23.861
	Alg. 6.2	7	9	7	-	9.14	27.55	33.676	0.855	23.859
	10^2	5	7	-	5	31.43	25.72	33.082	0.859	23.938
	10^3	9	12	-	9	29.20	45.47	33.115	0.859	23.934
4×4	10^4	6	8	-	6	15.82	40.22	33.134	0.859	23.931
	10^5	6	8	-	6	15.76	57.74	33.136	0.859	23.931
	Alg. 6.2	6	8	6	-	9.44	24.97	33.027	0.856	23.926
	10^2	13	15	-	13	668.87	52.93	32.437	0.865	24.023
	10^3	14	16	-	14	2027.77	75.22	32.450	0.867	24.022
8×8	10^4	13	15	-	13	2121.60	94.96	32.450	0.866	24.022
	10^5	14	16	-	14	3297.69	157.06	32.424	0.866	24.025
	Alg. 6.2	18	20	19	-	281.50	54.30	32.225	0.864	24.012

Table 5.1: Comparison between smooth and nonsmooth trust-region algorithms for the cameraman training pair.

using both the PSNR and SSIM metrics. Figure 5.4 shows the optimal reconstructed images and the corresponding optimal parameters. It can be observed how the algorithm adjusts the regularization parameter to specific zones in the image to obtain a better reconstruction.

When dealing with different initialization values for α, algorithm 5.1 appears to converge to the same local minima, although with different computational efforts. Table 5.2 shows the number of iterations along with the quality measures for the scalar parameter problem; more iterations are required when dealing with high values for α_0.

In Table 5.1 the performance of algorithm 5.1 with respect to the one-phase trust-region algorithm, using solely the smoothed gradient (for different values of γ), is tested. As reported in this table, even though the solution obtained solely with the regularized gradient is not different from the non-regularized one, algorithm 5.1 requires less running time. Furthermore, in column *cond mean*, the average condition number of the BFGS approximation matrices B_k along the iterations is registered. These matrices are used in step one of algorithm 2.3, and are an indicator of how fast the linear systems may be solved. The average condition number computed with the non-smooth model is consistently smaller than the one obtained using a regularized procedure. As, for this example, the algorithm does not enter the second phase, an advantage in choosing the non-smooth model (with the Bouligand element) may be clearly inferred.

					Reconstruction		
α_0	nit	nfev	ngev	nreggev	COST	PSNR	SSIM
1e-05	6	8	8	0	34.113114	23.804686	0.842906
0.0001	6	8	8	0	34.113114	23.804686	0.842906
0.001	6	8	8	0	34.113114	23.804686	0.842906
0.01	5	7	7	0	34.113114	23.804686	0.842906
0.1	8	10	10	0	34.113114	23.804686	0.842906
0.2	7	9	9	0	34.113114	23.804686	0.842906
0.3	8	10	10	0	34.113115	23.804686	0.842904
0.4	11	13	13	0	34.113114	23.804686	0.842906
0.5	6	8	8	0	34.113115	23.804686	0.842904

Table 5.2: Dependence of the algorithm on the initial value of the scalar parameter for the Cameraman training pair.

					Reconstruction		
patch	nit	nfev	ngev	nreggev	COST	PSNR	SSIM
1×1	6	8	8	0	34.113114	23.804686	0.842906
2×2	9	11	11	0	33.675627	23.860742	0.855044
4×4	12	14	14	0	33.027442	23.945150	0.859578
8×8	13	15	14	1	32.224837	24.051992	0.868038
16×16	26	28	27	1	31.728896	24.119350	0.871783
32×32	31	33	33	0	31.000720	24.220182	0.877032

Table 5.3: Nonsmooth trust-region algorithm behavior for the Cameraman training pair.

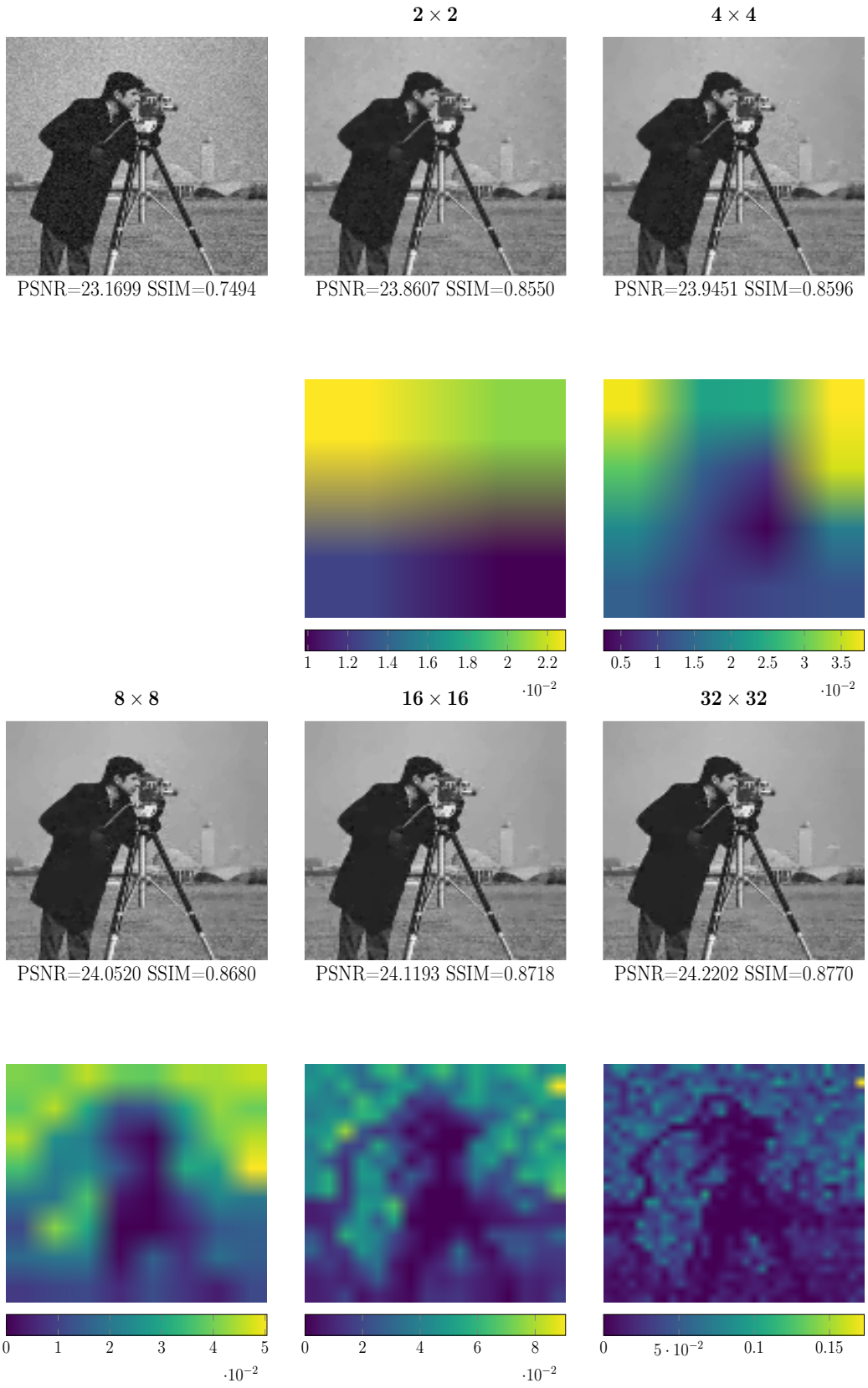


Figure 5.4: Learned optimal patch parameter for an increasing number of patches for the Cameraman training pair.

Figure 5.5: Optimal scalar reconstruction for the circles training pair. In this experiment, the optimal parameter found is $\alpha^* = 0.21629352$.

					Reconstruction		
patch	nit	nfev	ngev	nreggev	COST	PSNR	SSIM
1×1	6	8	8	0	76.752394	20.282980	0.760610
2×2	7	9	9	0	72.654528	20.521273	0.821846
4×4	19	21	20	1	71.266465	20.605047	0.849556
8×8	52	54	53	1	67.082882	20.867782	0.865338
16×16	55	57	57	0	49.507960	22.187149	0.881588
32×32	60	62	61	1	39.249873	23.195517	0.886326

Table 5.4: Trust Region Algorithm behavior on the circles training pair.

5.4.2 Circles Training Pair

For this experiment, we explore the patch parameter adaptation for a test image with noise distributed differently along with the domain. This training pair is shown in Figure 5.5 along with its optimal scalar parameter reconstruction. When moving to a patch-dependent parameter learning, we can see in Figure 5.6 that the optimal patch parameter adjusts to the original noise distribution. Furthermore, we can see that this adaptation leads to a better reconstruction quality according to the SSIM metric. Finally, regarding the behavior of the trust-region algorithm, Table 5.4 shows the number of iterations, number of function evaluations, gradient and regularized gradient evaluations (*nfev*,*ngev*,*nregev*) respectively. Again, we can see an improvement in the reconstruction quality as more patches are considered for the parameter.

5.4.3 Multiple Training Pairs

In this experiment, we used ten image pairs containing images of faces to generate a training dataset and ten different image pairs to generate a validation dataset; both datasets were based on the CelebA dataset [50]. These images are of size 128 by 128 pixels, and in both datasets, we created the degenerated pairs by adding Gaussian noise with zero-mean and standard deviation $\sigma = 0.1$. A subset of the training dataset

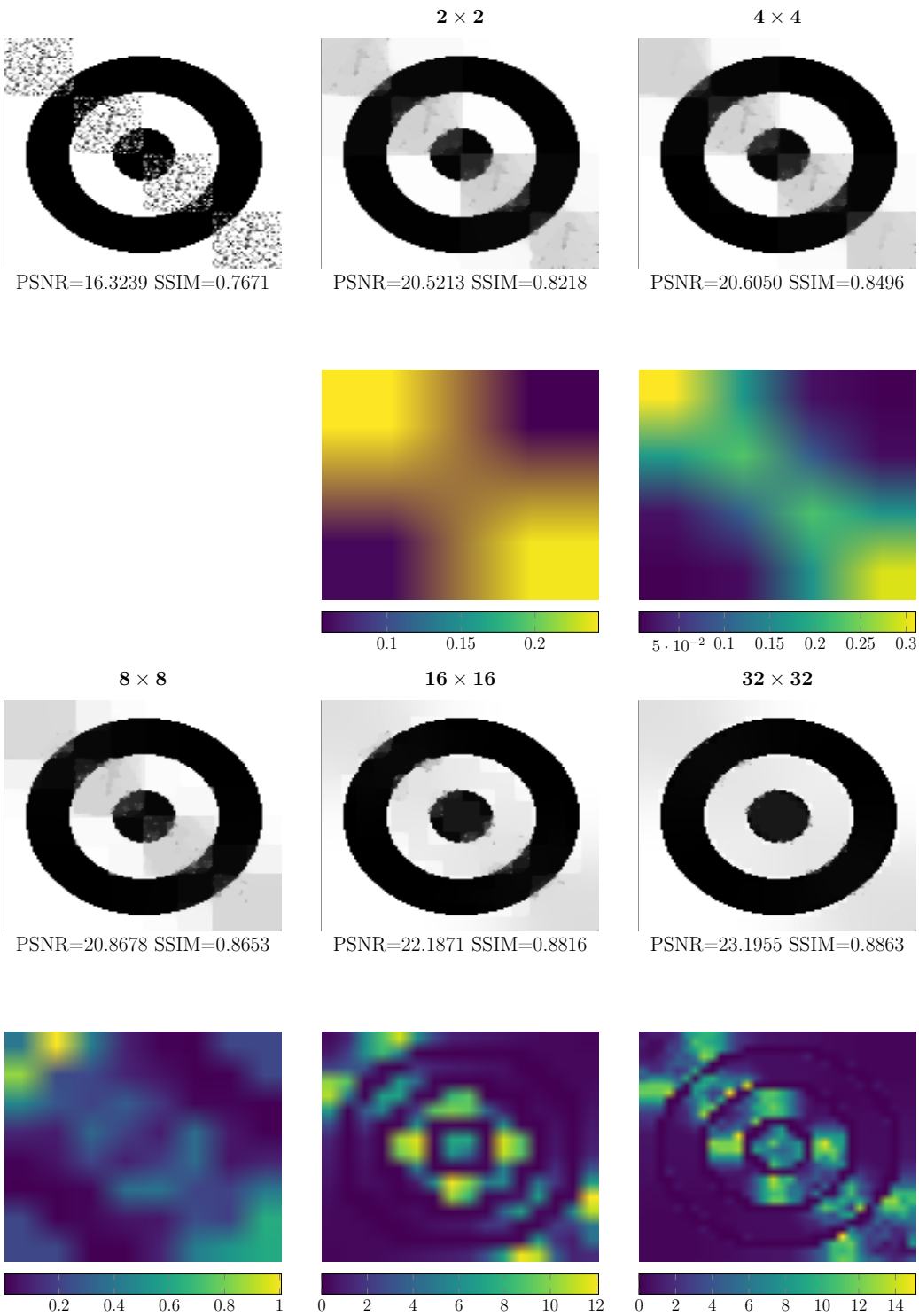


Figure 5.6: Learned optimal patch parameter for an increasing number of patches for the circles training pair.

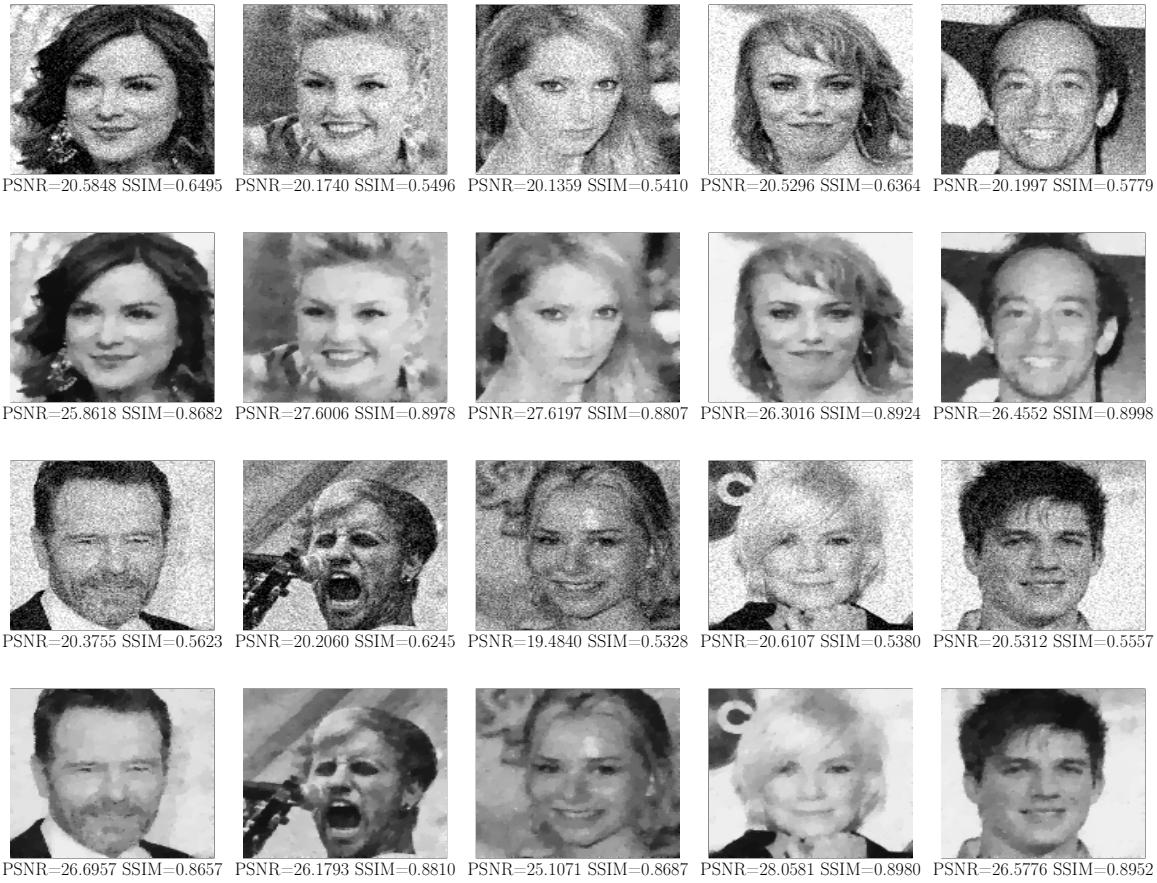


Figure 5.7: Noisy images used for the faces training dataset corrupted with gaussian noise and their corresponding optimal scalar reconstructions for $\alpha^* = 0.07311238$.

is depicted in Figure 5.7.

In Figure 5.8, we plot the cost function corresponding to a scalar parameter and two-dimensional patch parameter along with the cost function corresponding to the optimal value calculated by the algorithm. Again, we can confirm experimentally that the optimal value was calculated. When incrementing the number of patches used in the reconstruction, Figure 5.9 shows the optimal patch parameters obtained. We can now see that the parameter value doesn't adjust to a specific image but to the whole training set.

For the training dataset proposed, Table 5.5 shows the mean SSIM (MSSIM) and mean PSNR (MPSNR) reconstruction quality metrics for different patch sizes, along with the number of iterations, cost evolution, and the number of function and gradient evaluations used by the algorithm. As was also the case for the cameraman training pair, an improvement in the reconstruction quality can be seen as the degrees of freedom for the regularization parameter increase.

Finally, we can estimate the denoiser performance in images from the validation dataset. This experiment shows a slight *overfitting* phenomenon that may occur when

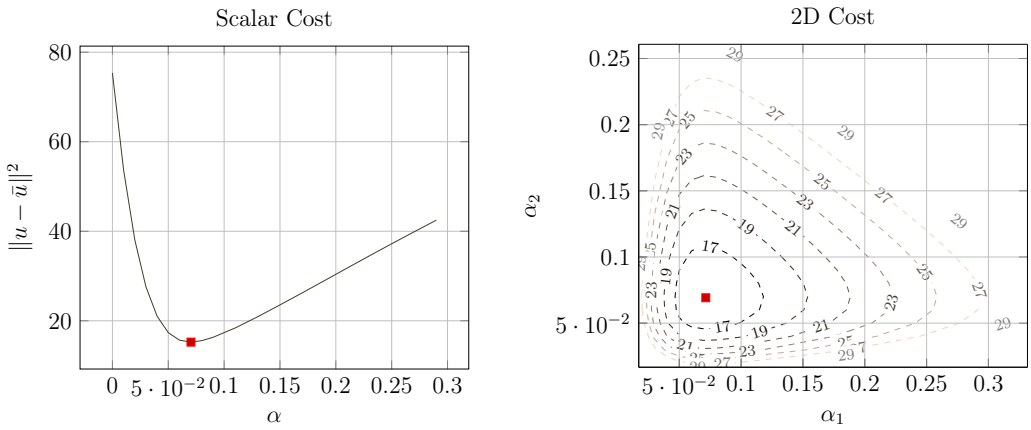


Figure 5.8: Values for the l_2 squared cost function using a scalar regularization parameter and a two-dimensional regularization parameter using the faces dataset.

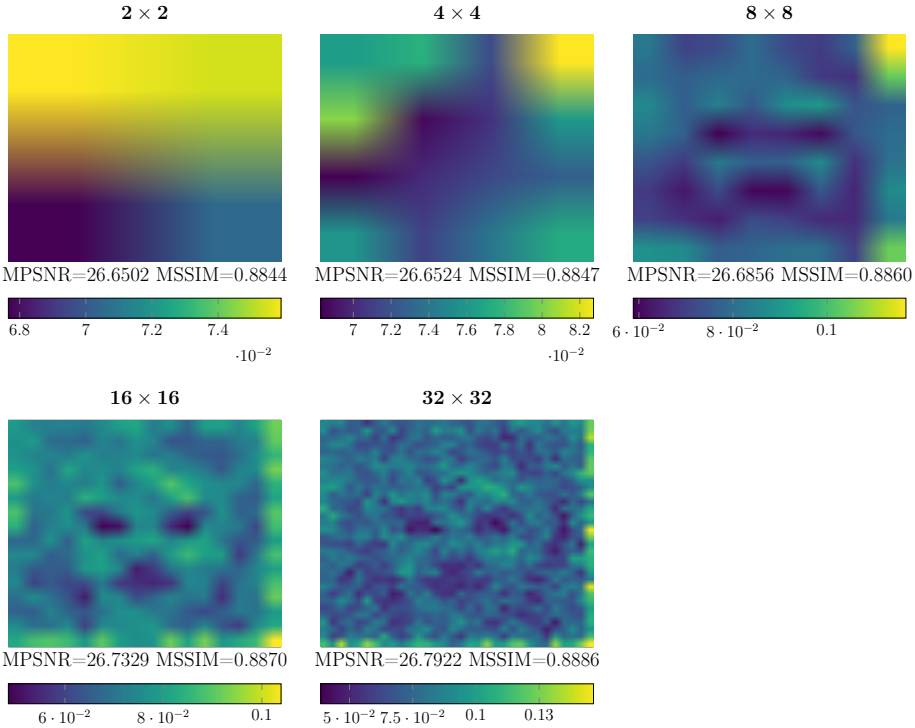


Figure 5.9: Values for the optimal parameters calculated for different parameter patch sizes on the faces dataset.

					Reconstruction		
patch	nit	nfev	ngev	nreggev	COST	MPSNR	MSSIM
1×1	9	11	11	0	18.075146	26.645679	0.884746
2×2	10	12	12	0	18.054696	26.650206	0.884403
4×4	27	29	28	1	18.049779	26.652420	0.884702
8×8	17	19	19	0	17.927620	26.685560	0.886005
16×16	13	15	14	1	17.723099	26.732897	0.886991
32×32	19	21	20	1	17.487146	26.792166	0.888616

Table 5.5: Trust Region Algorithm behavior on the Faces dataset.

num	**noisy**	**scalar**	**2×2**	**4×4**	**8×8**	**16×16**	**32×32**
1	0.6524	0.8748	0.8753	0.8751	0.8752	0.8751	0.8751
2	0.5840	0.7656	0.7688	0.7690	0.7695	0.7698	0.7693
3	0.5623	0.8668	0.8668	0.8663	0.8654	0.8638	0.8620
4	0.5350	0.8204	0.8206	0.8207	0.8208	0.8209	0.8204
5	0.5979	0.8737	0.8729	0.8726	0.8719	0.8708	0.8695
6	0.5807	0.8439	0.8446	0.8445	0.8447	0.8448	0.8449
7	0.5640	0.7460	0.7490	0.7495	0.7501	0.7506	0.7504
8	0.5631	0.8467	0.8471	0.8471	0.8471	0.8467	0.8461
9	0.5910	0.8354	0.8368	0.8366	0.8359	0.8348	0.8335
10	0.6622	0.8753	0.8768	0.8765	0.8761	0.8756	0.8752
MSSIM	**0.5892**	**0.8348**	**0.8358**	**0.8358**	**0.8356**	**0.8352**	**0.8346**

Table 5.6: Faces Dataset SSIM Quality Measures in the validation dataset.

dealing with a large number of patches, as described in Table 5.6. Indeed, it can be seen in the validation dataset an increment on the mean SSIM (MSSIM) for the reconstructed images from the validation dataset up to a 4×4 patch size. Any higher number of patches results in quality degradation. It is indeed the expected behavior when dealing with overfitting problems.

5.4.4 Learning Optimal Total Variation Discretization

The selection of an adequate discretization for the total variation seminorm in the context of image reconstruction problems is still an open problem [10, 16]. In [17], the authors propose a methodology for finding optimal discretizations, where a bilevel learning strategy is proposed instead of using hand-crafted schemes.

The bilevel framework presented in this work can also be used to learn optimal gradient discretization by considering different schemes and their corresponding regularization parameters. We will use a training dataset to estimate the optimal regular-

ization parameters related to the "contributions" of each discretization scheme to the final solution.

Let us consider the following variational denoising model

$$\min_{u\in\mathbb{R}^n} \mathcal{E}(u) := \frac{1}{2}\|u-f\|^2 + \sum_{i=1}^{3}\sum_{j=1}^{m} \left(\mathcal{P}(\alpha_i)\right)_j \|(\mathbb{K}_i u)_j\|, \tag{5.74}$$

where $\mathbb{K}_1, \mathbb{K}_2$ and $\mathbb{K}_3$ are the forward, backward and centered finite difference schemes of the gradient operator. The goal is to determine optimal parameters $(\alpha_1, \alpha_2, \alpha_3)^\top \in \mathbb{R}_+^{3\times m}$ that lead to an optimal patchwise discretization of the total variation operator. We can make use of a similar analysis as the one leading to the Bouligand candidate in Theorem 5.7, by defining the following adjoint state:

$$\begin{aligned} \langle p, v\rangle + \sum_{i=1}^{3}\sum_{j\in\mathcal{I}^i} \langle \mu_{ij}, (\mathbb{K}_i v)_j\rangle - \langle \nabla J(u), v\rangle = 0, \forall v \in \mathcal{V} \\ \mu_{ij} - \mathcal{P}(\alpha_i)_j T_{ij} (\mathbb{K}_i p)_j = 0,\ \forall j \in \mathcal{I}^i,\ i = 1,2,3, \end{aligned}$$

where $\mathcal{V} := \{v \in \mathbb{R}^n : (\mathbb{K}_i v)_j = 0,\ \forall \mathcal{A}_s^i \cup \mathcal{B}_1^i,\ (\mathbb{K}^i v)_j \in span(q_{ij}),\ \forall j \in \mathcal{B}_2^i,\ i = 1,2,3\}$, and the following gradient form

$$\langle j'(\alpha_i), h_i\rangle = -\sum_{j\in\mathcal{I}^i\cup\mathcal{B}_2^i} \frac{h_{ij}}{\mathcal{P}(\alpha_i)_j} \langle q_{ij}, (\mathbb{K}_i p)_j\rangle, \quad \text{for } i = 1,2,3.$$

Figure 5.10 shows the 4×4 and 8×8 optimal patchwise parameters α_1, α_2 and α_3 obtained for the cameraman training pair and Figure 5.11 for the faces dataset. We can observe an improvement in the reconstruction quality by using more patches in the training dataset for both experiments. For the faces dataset, we also run a validation experiment comparing different learned parameters of increasing size (see Table 5.7). A higher mean SSIM value is registered when using an 8×8 patch size, with the reconstruction quality slightly degrading for larger patch sizes.

5.4.5 Comparison with Derivative-Free Bilevel Parameter Learning

This section compares the method proposed in this work and the parameter bilevel learning using inexact derivative-free optimization (DFO) as described in [27]. This methodology removes the requirement for exactly solving the lower-level problem, reducing its computational cost significantly. In particular, the authors consider a

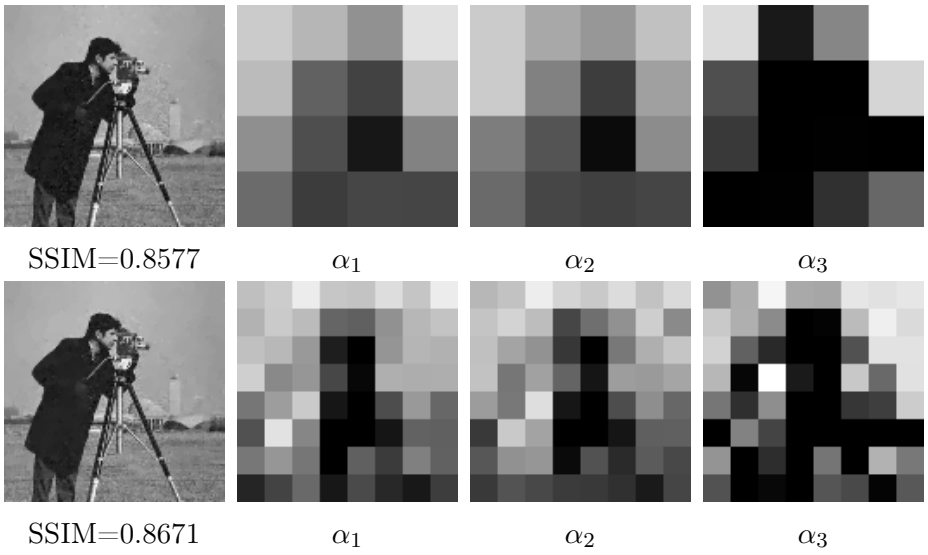


Figure 5.10: Values for the optimal parameter calculated for the Cameraman training pair for different patch sizes (white - higher, black - lower). For each parameter, the higher the parameter, the more the final solution used its corresponding type of discretization.

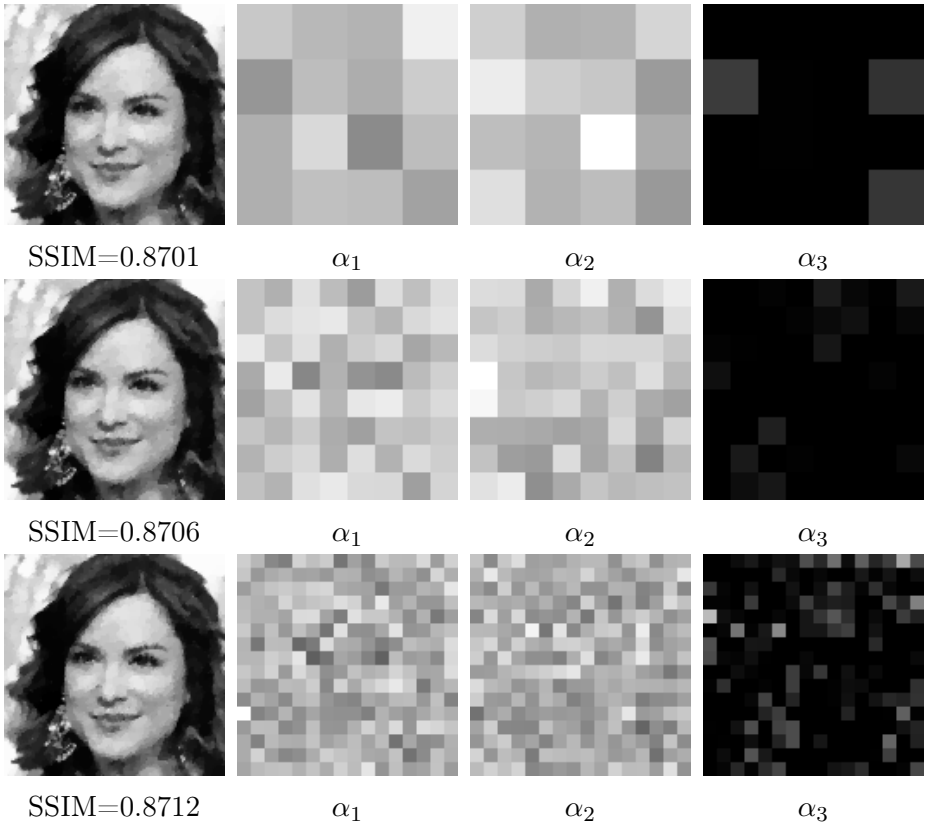


Figure 5.11: Values for the optimal parameters calculated for different parameter patch sizes (white - higher, black - lower).

img num	noisy	scalar	2×2	4×4	8×8	16×16
1	0.5247	0.7922	0.7952	0.7951	0.7977	0.7965
2	0.4588	0.6568	0.6676	0.6663	0.6649	0.6653
3	0.4267	0.7745	0.7742	0.7729	0.7748	0.7741
4	0.3836	0.7237	0.7239	0.7226	0.7244	0.7206
5	0.4580	0.7814	0.7796	0.7779	0.7806	0.7809
6	0.4263	0.7394	0.7406	0.7414	0.7432	0.7441
7	0.4547	0.6262	0.6372	0.6357	0.6338	0.6333
8	0.4117	0.7377	0.7380	0.7413	0.7429	0.7421
9	0.4655	0.7438	0.7481	0.7445	0.7432	0.7420
10	0.5081	0.8185	0.8225	0.8217	0.8219	0.8211
MSSIM		**0.7395**	**0.7420**	**0.7419**	**0.7428**	**0.7424**

Table 5.7: Faces Dataset SSIM Quality Measures - Optimal gradient discretization for the validation dataset

smoothed version of the scalar ROF model for the lower-level optimization problem for a vector of parameters θ

$$\Phi_{i,\theta} = \frac{1}{2}\|u - f_i\|^2 + \alpha_\theta \sum_{j=1}^{m} \|(\mathbb{K}u)_j\|_{\nu_\theta} + \frac{\xi_\theta}{2}\|u\|^2,$$

where α_θ is the scalar regularization parameter, $(\mathbb{K}u)_j$ is the finite forward difference discretization of the spatial gradient of u at pixel j, $\|v\|_{\nu_\theta} = \sqrt{\|v\|^2 + \nu_\theta^2}$ is a smoothed version of the total variation, and ξ_θ is the weight assigned for the strongly convex term. Furthermore, the lower-level problem is solved numerically using the FISTA [5] algorithm. The upper-level solver is based on an inexact trust-region algorithm [27, Algorithm 1] that can be configured to work in two different regimes. A first regime, called *dynamic*, solves the lower-level problem with "sufficient" accuracy, and the *fixed* regime, where the lower-level problem is solved in a fixed number of iterations.

In this comparison experiment, we will compare algorithm 5.1 for calculating the optimal scalar parameter for the non-smoothed version of the ROF model with a lower-level solver PDHGM and both the dynamic version and the fixed version of the DFO algorithm with 1000 iterations of the FISTA algorithm for the latter. In the spirit of comparing this algorithm with the non-smooth model proposed in this work, we fixed the smoothing parameters ν_θ and ξ_θ to the lowest possible value that allowed computation of the lower level problem in a reasonable time. Particularly, we chose $\nu_\theta = \xi_\theta = 1 \cdot 10^{-3}$. Both experiments will make use of the Kodak dataset. The ground truth contains 24 images resized to 256×256-pixels and converted to black and white, and the corresponding noisy pairs are synthetically generated by adding gaussian noise with zero mead and 0.1 variance.

	l_2			PSNR			SSIM		
img	algorithm 5.1	dynamic	fixed	algorithm 5.1	dynamic	fixed	algorithm 5.1	dynamic	fixed
1	108.399640	108.028863	108.025305	24.804221	24.819101	24.819244	0.799179	0.799593	0.799601
2	36.692144	36.071444	36.070946	29.508769	29.582864	29.582924	0.860296	0.860352	0.860355
3	34.094641	33.517839	33.519153	29.827638	29.901739	29.901569	0.884152	0.884052	0.884043
4	24.244099	23.698494	23.698925	31.308439	31.407292	31.407213	0.899989	0.899493	0.899489
5	113.720294	112.958227	112.954822	24.596120	24.625321	24.625452	0.847932	0.848568	0.848575
6	111.569222	110.649647	110.642370	24.679055	24.714999	24.715285	0.754031	0.755755	0.755775
7	62.766173	62.186950	62.183051	27.177243	27.217507	27.217779	0.874712	0.875072	0.875080
8	96.430437	95.444597	95.443889	25.312358	25.356986	25.357018	0.862241	0.862743	0.862744
9	35.106264	34.611217	34.613604	29.700653	29.762331	29.762031	0.901842	0.901535	0.901522
10	23.723895	23.172592	23.175260	31.402639	31.504753	31.504253	0.901620	0.901163	0.901150
11	96.549526	95.546500	95.542751	25.306998	25.352352	25.352522	0.827413	0.828732	0.828743
12	44.489950	43.593604	43.590994	28.671880	28.760272	28.760532	0.841915	0.843136	0.843153
13	125.810563	125.127508	125.120610	24.157328	24.180971	24.181211	0.752221	0.753796	0.753819
14	76.524528	75.979117	75.974868	26.316493	26.347557	26.347800	0.848706	0.849095	0.849107
15	44.312576	43.606816	43.604357	28.689229	28.758956	28.759200	0.859184	0.859698	0.859707
16	49.316021	48.645422	48.643158	28.224619	28.284080	28.284282	0.834352	0.834577	0.834589
17	42.687262	42.182209	42.178122	28.851516	28.903206	28.903627	0.867394	0.867776	0.867791
18	67.882205	67.099766	67.097877	26.836940	26.887289	26.887412	0.859794	0.860026	0.860032
19	85.897165	85.514149	85.512234	25.814711	25.834120	25.834217	0.834928	0.835329	0.835333
20	39.751725	39.050264	39.054558	29.160940	29.238260	29.237782	0.917254	0.916472	0.916447
21	73.936611	73.350698	73.350563	26.465904	26.500457	26.500465	0.869522	0.869387	0.869380
22	61.397514	60.418329	60.414806	27.272991	27.342812	27.343066	0.837474	0.838404	0.838419
23	33.024632	32.532794	32.532778	29.966120	30.031286	30.031288	0.906590	0.906264	0.906260
24	48.111877	47.678912	47.677969	28.331976	28.371236	28.371322	0.868742	0.868953	0.868957
25	75.300547	74.505264	74.497676	26.386518	26.432630	26.433072	0.817761	0.819139	0.819162
mean	**64.469580**	**63.806849**	**63.804826**	**27.550852**	**27.604735**	**27.604823**	**0.853170**	**0.853564**	**0.853569**

Table 5.8: Comparison between the l_2, PSNR, and SSIM metric for the nonsmooth trust region method, the derivative-free with dynamic accuracy and the derivative-free method with a fixed number of iterations when solving the bilevel parameter learning problem for the Kodak dataset.

	Algorithm 5.1	**dynamic**	**fixed**
α^*	0.06985138	0.06946807	0.06947291
ν^*	-	0.001	0.001
ξ^*	-	0.001	0.001

Table 5.9: Optimal parameters found using the nonsmooth trust-region algorithm and the optimal parameters calculated using dynamic and fixed versions of the DFO algorithm.

Table 5.8 shows the quality reconstruction metrics for each image in the training set and their respective mean. As shown in the table, all three metrics considered are very similar, corresponding to the value of the optimal parameter obtained as detailed in table 5.9, which are very similar as well.

Regarding the computational time, we saw a dramatic computation speedup when solving the dynamic DFO, solving the problem in about 8 hours of computation. In comparison, the fixed DFO algorithm and the nonsmooth trust-region algorithm presented here took a similar computation time of about 24 hours. An important observation regarding this experiment concerns the size of the parameters calculated. DFO techniques generally cannot be extended to a large parameter size, as in a scale or patch-dependent parameter problem. This case prevents us from comparing these algorithms

Dataset	Data Fidelity Term				Regularization Term			
	λ^*	L2	PSNR	SSIM	α^*	L2	PSNR	SSIM
Cameraman	64.357504	34.114243	23.804542	0.842652	0.015622	34.113114	23.804686	0.842906
Circles	4.663291	76.902703	20.274483	0.763919	0.216294	76.752394	20.282980	0.760610
Faces	14.001192	18.020032	26.657829	0.884794	0.073112	18.075146	26.645679	0.884746

Table 5.10: Optimal scalar parameters and their corresponding reconstruction quality metrics for both the data parameter learning problem and the regularization parameter learning problem.

in a patch-dependent parameter scenario. Furthermore, the DFO algorithm seems to rely heavily on the smoothing parameters ν_θ and ξ_θ since, as its value decreases, the lower-level problem becomes increasingly harder to solve. This phenomenon prevents us from approximating the solution of a non-smooth bilevel learning problem.

5.5 Learning Data Weight vs Regularization Weight

In this section, we will compare the optimal data fidelity weight learned using the approach described in chapter 4 and the optimal regularization weight learned using the methodology described in chapter 5.

To start our comparison, we will first explore the scalar parameter model for learning the data and the regularization term. Then, in table 5.10, a comparison between the quality metrics of the reconstructed images using the optimal learned parameter for the Cameraman, Circles, and the Faces dataset. In this table, we can see that the obtained reconstructions have very similar quality metrics for all the datasets explored. Moreover, when comparing the obtained values for the data fidelity parameter, their inverses are very close to the calculated optimal regularization parameter.

Now, when moving into a patch-based parameter regime, Figures 5.12 and 5.13 show the calculated patch parameter for both the data fidelity and the regularization learning models for the Cameraman and the Circles datasets. In addition, below each patch, the reconstructed image's PSNR and SSIM quality metrics are shown. Here, even though the reconstruction quality grows higher as a large number of patches is used, the solutions diverge. Remarkably, this phenomenon is more evident when using a patch of size larger than 16×16 in the Cameraman dataset and 4×4 in the Circles dataset.

A similar comparison is shown in Figure 5.14 for the Faces dataset. In this figure, the quality of the optimal reconstructed dataset obtained using the learned data fidelity term remains very close to the one obtained when learning the optimal regularization term according to the MPSNR and MSSIM metrics.

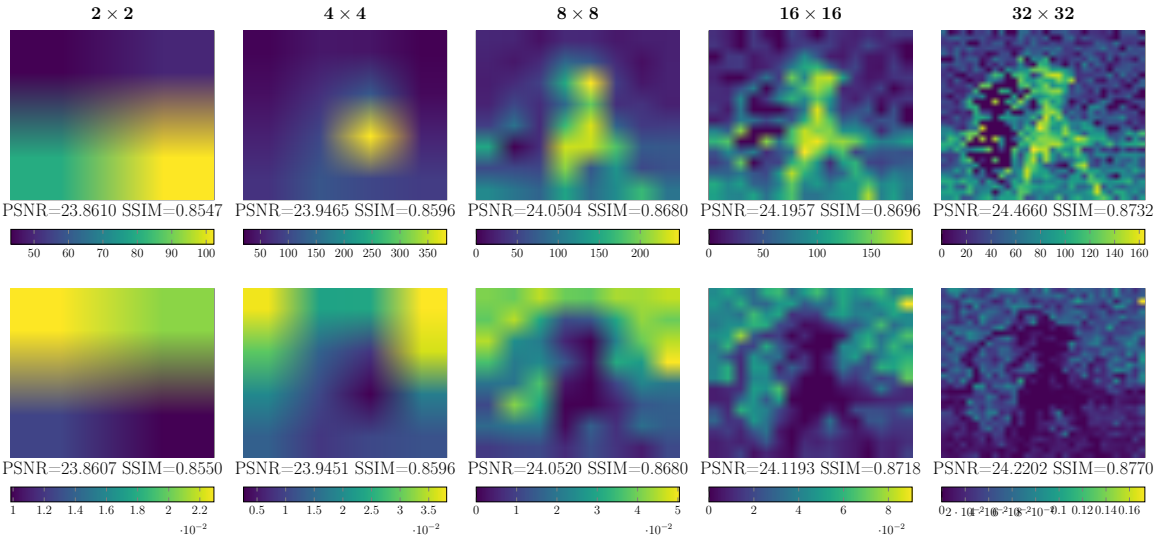


Figure 5.12: Reconstruction and patches comparison for data and reg.

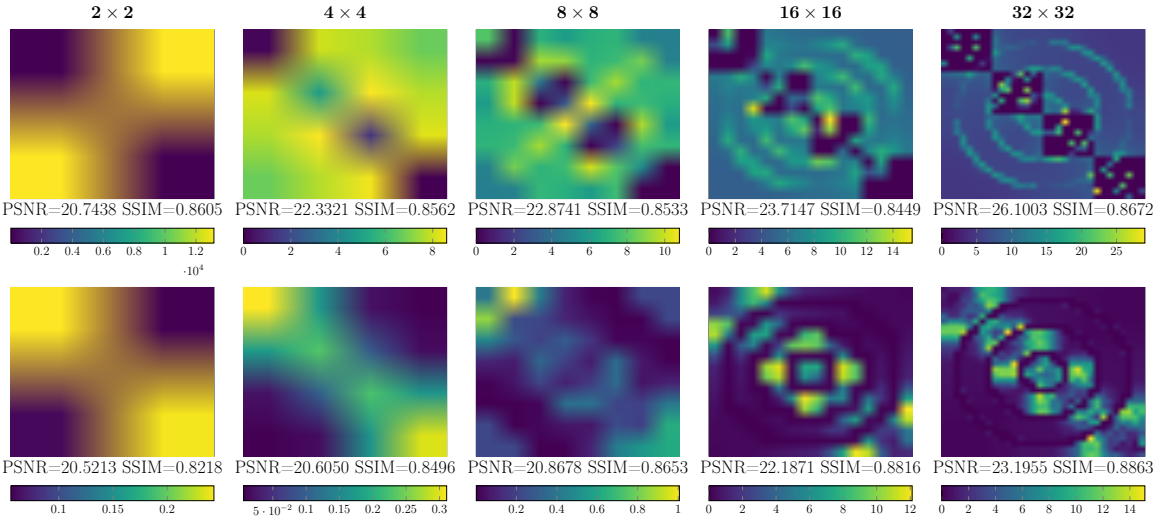


Figure 5.13: Reconstruction and patches comparison for data and reg.

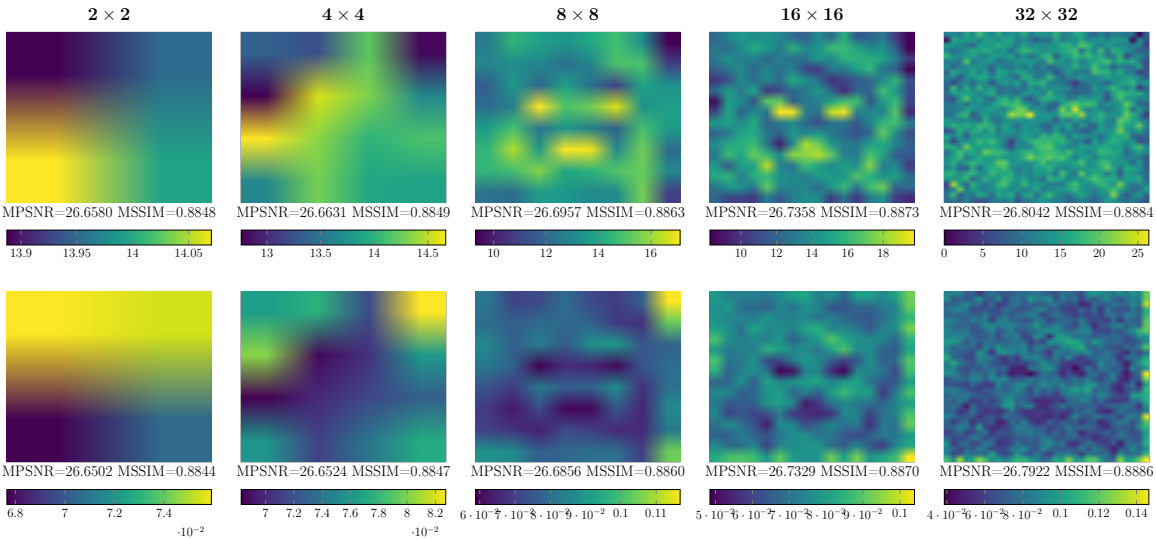


Figure 5.14: Reconstruction and patches comparison for data and reg.

As a summary of all previous results, even though we can see that in the scalar case, regardless of learning either the data fidelity term or the regularization term, both obtain the same solution, this behavior doesn't repeat when dealing with higher dimensional parameters. In this regime, it can be seen that as the parameter grows larger, the solutions obtained differ. A similar result has been seen previously for the one-dimensional total variation denoising in [37, Figure 12], and its extension to two dimensions is a matter for further research. Furthermore, the observed phenomenon is particularly interesting when dealing with training sets, where the averaged reconstruction quality metric doesn't present such a considerable variation as in the case of a single training pair.

Chapter 6

Conclusions and Outlook

In this research, we investigated optimality conditions for bilevel parameter learning problems when the lower level problem corresponds to a variational image denoising problem. This problem is particularly challenging due to the non-differentiability induced by the use of the total variation seminorm as a regularizer. Traditionally, this problem is often approached using a tailored regularization of the non-smooth term, from where Clarke (C-) optimality conditions can be derived.

This research aimed to characterize further stationary points for two bilevel parameter learning problems. The first one, presented in Chapter 4 deals with the learning of optimal parameters affecting the data fidelity term in the variational denoising model; followed by an analysis and experiments dealing with learning optimal parameters affecting the regularization term in the variational denoising model in Chapter 5. Indeed, by reformulating each of the problems as a Generalized Mathematical Program with Equilibrium Constraints (GMPEC) and verifying the fulfillment of suitable constraint qualification conditions, we managed to find Mordukhovich (M)- stationary points. Furthermore, by investigating the differentiability properties of the solution map, we were able to characterize Bouligand (B-) stationary points.

In this work, we provide a precise characterization of the Tangent, Fréchet, and Mordukhovich normal cones for the subdifferential of the Euclidean norm. Moreover, we proved and characterized the directional derivative of the solution map and its Fréchet differentiability in the case that strict complementarity holds.

Regarding the numerical solution of the problem, we used a two-phase trust-region algorithm tailored to deal with Lipschitz continuous functions and positiveness constraints. A crucial component of this algorithm is the definition of the model to be used within the trust region; in our case, we used a characterization of the linear elements in the Bouligand subdifferential of the solution map to define it. For the lower level problem, we implemented a matrix-free version of the classical primal-dual

techniques for solving the image denoising model. This implementation presented improvements for both the computation time and the memory usage when compared with sparse-matrix and full-matrix implementations.

Moreover, the numeric part explored the use of different parameter structures; in particular, scalar and patch-dependent parameters were compared. Experimentally, we could see that using a patch-dependent structure allows us to validate the generalization capabilities of the learned parameter when dealing with images not used for training. Indeed, the optimal patch size can be found using traditional validation techniques.

As presented in the experimental part, the structure of the parameters is an essential part of the image reconstruction process. Therefore, learning the parameter structure along with the parameter is a promising research direction in the future. In particular, exploring the enforcement of sparsity properties on the learned parameter in the upper-level problem is a recommended continuation of this work. Furthermore, the analysis presented in this work focuses on the isotropic version of the total variation seminorm. It opens the door to exploring the geometric structure of other norms within the regularizer, with the anisotropic total variation being a natural extension.

Bibliography

[1] Apkarian, Noll & Ravanbod, Nonsmooth bundle trust-region algorithm with applications to robust stability, *Set-Valued and variational analysis* 24 (2016), 115–148.

[2] Bao et al., Stability and Generalization of Bilevel Programming in Hyperparameter Optimization, in: *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2021, 4529–4541, URL: `proceedings.neurips.cc/paper/2021/file/2406a0a94c80406914ff2f6c9fdd67d5-Paper.pdf`.

[3] Barbu, Optimal control of variational inequalities, *Research Notes in Math.* 100 (1984).

[4] Beck, *First-order methods in optimization*, SIAM, 2017.

[5] Beck & Teboulle, A fast iterative shrinkage-thresholding algorithm with application to wavelet-based image deblurring, in: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2009, 693–696.

[6] Bergounioux, Optimal control of problems governed by abstract elliptic variational inequalities with state constraints, *SIAM Journal on Control and Optimization* 36 (1998), 273–289.

[7] Bergstra & Bengio, Random search for hyper-parameter optimization. *Journal of machine learning research* 13 (2012).

[8] Bertsekas, Nedic & Ozdaglar, *Convex analysis and optimization*, Athena Scientific, 2003.

[9] Bredies, Dong & Hintermüller, Spatially dependent regularization parameter selection in total generalized variation models for image restoration, *International Journal of Computer Mathematics* 90 (2013), 109–123.

[10] Caillaud & Chambolle, Error estimates for finite differences approximations of the total variation (2020).

[11] Calatroni, De Los Reyes & Schönlieb, Dynamic sampling schemes for optimal noise learning under multiple nonsmooth constraints, in: *IFIP Conference on System Modeling and Optimization*, Springer, 2013, 85–95.

[12] Calatroni & Papafitsoros, Analysis and automatic parameter selection of a variational model for mixed Gaussian and salt-and-pepper noise removal, *Inverse Problems* 35 (2019), 114001.

[13] Calatroni et al., Bilevel approaches for learning of variational imaging models, in: *Variational methods*, De Gruyter, 2017, 252–290.

[14] Chambolle & Lions, Image recovery via total variation minimization and related problems, *Numerische Mathematik* 76 (1997), 167–188.

[15] Chambolle & Pock, A first-order primal-dual algorithm for convex problems with applications to imaging, *Journal of mathematical imaging and vision* 40 (2011), 120–145.

[16] Chambolle & Pock, Crouzeix–Raviart approximation of the total variation on simplicial meshes, *Journal of Mathematical Imaging and Vision* 62 (2020), 872–899.

[17] Chambolle & Pock, Learning consistent discretizations of the total variation, *SIAM Journal on Imaging Sciences* 14 (2021), 778–813.

[18] Christof, De los Reyes & Meyer, A Nonsmooth Trust-Region Method for Locally Lipschitz Functions with Application to Optimization Problems Constrained by Variational Inequalities, *SIAM Journal on Optimization* 30 (2020), 2163–2196.

[19] D'Elia, De Los Reyes & Miniguano-Trujillo, Bilevel Parameter Learning for Nonlocal Image Denoising Models, *Journal of Mathematical Imaging and Vision* (2021), 1–23.

[20] De los Reyes & Herrera, Parameter space study of optimal scale-dependent weights in TV image denoising, *Applicable Analysis* (2022), 1–25.

[21] De los Reyes, Optimal control of a class of variational inequalities of the second kind, *SIAM Journal on Control and Optimization* 49 (2011), 1629–1658.

[22] De los Reyes & Meyer, Strong stationarity conditions for a class of optimization problems governed by variational inequalities of the second kind, *Journal of Optimization Theory and Applications* 168 (2016), 375–409.

[23] De Los Reyes, Schönlieb & Valkonen, The structure of optimal parameters for image restoration problems, *Journal of Mathematical Analysis and Applications* 434 (2016), 464–500.

[24] De los Reyes, Schönlieb & Valkonen, Bilevel parameter learning for higher-order total variation regularisation models, *Journal of Mathematical Imaging and Vision* 57 (2017), 1–25.

[25] De los Reyes & Schönlieb, Image denoising: learning the noise model via nonsmooth PDE-constrained optimization, *Inverse Problems & Imaging* 7 (2013), 1183–1214.

[26] De los Reyes & Villacís, Bilevel Optimization Methods in Imaging, *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging* 33 (2021), 074005.

[27] Ehrhardt & Roberts, Inexact derivative-free optimization for bilevel learning, *Journal of Mathematical Imaging and Vision* 63 (2021), 580–600.

[28] Ekeland & Temam, *Convex analysis and variational problems*, SIAM, 1999.

[29] Engl, Hanke & Neubauer, *Regularization of inverse problems*, Springer Science & Business Media, 1996.

[30] Franceschi et al., Bilevel programming for hyperparameter optimization and meta-learning, in: *International Conference on Machine Learning*, PMLR, 2018, 1568–1577.

[31] Friedberg, Insel & Spence, *Linear Algebra: Pearson New International Edition PDF eBook*, Pearson Higher Ed, 2013.

[32] Geiger & Kanzow, *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*, Springer-Verlag, 1999.

[33] Geiger & Kanzow, *Theorie und numerik restringierter optimierungsaufgaben*, Springer-Verlag, 2013.

[34] Glowinski & Oden, Numerical methods for nonlinear variational problems, *Journal of Applied Mechanics* 52 (1985), 739.

[35] Golub, Heath & Wahba, Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics* 21 (1979), 215–223.

[36] Hammernik & Knoll, Machine learning for image reconstruction, in: *Handbook of medical image computing and computer assisted intervention*, Elsevier, 2020, 25–64.

[37] Hintermüller, Papafitsoros & Rautenberg, Analytical aspects of spatially adapted total variation regularisation, *Journal of Mathematical Analysis and Applications* 454 (2017), 891–935.

[38] Hintermüller, Papafitsoros, Rautenberg & Sun, Dualization and automatic distributed parameter selection of total generalized variation via bilevel optimization, *Numerical Functional Analysis and Optimization* 43 (2022), 887–932.

[39] Hintermüller & Rautenberg, Optimal selection of the regularization function in a weighted total variation model. Part I: Modelling and theory, *Journal of Mathematical Imaging and Vision* 59 (2017), 498–514.

[40] Hintermüller, Rautenberg, Wu & Langer, Optimal selection of the regularization function in a weighted total variation model. Part II: Algorithm, its analysis and numerical tests, *Journal of Mathematical Imaging and Vision* 59 (2017), 515–533.

[41] Hintermüller & Stadler, An infeasible primal-dual algorithm for total bounded variation–based inf-convolution-type image restoration, *SIAM Journal on Scientific Computing* 28 (2006), 1–23.

[42] Hintermüller & Wu, Bilevel optimization for calibrating point spread functions in blind deconvolution, *Inverse Problems & Imaging* 9 (2015).

[43] Jahn, *Introduction to the theory of nonlinear optimization*, Springer Nature, 2020.

[44] Karush, Minima of functions of several variables with inequalities as side constraints, *M. Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago* (1939).

[45] Khan, Tammer & Zalinescu, *Set-valued optimization*, Springer, 2016.

[46] Klatzer & Pock, Continuous hyper-parameter learning for support vector machines, in: *Computer Vision Winter Workshop (CVWW)*, Citeseer, 2015, 39–47.

[47] Kuhn & Tucker, Non-linear Programming, in: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, Hrsg.: J. Neyman. S*, 1950, 481–492.

[48] Kunisch & Pock, A bilevel optimization approach for parameter learning in variational models, *SIAM Journal on Imaging Sciences* 6 (2013), 938–983.

[49] Le, Chartrand & Asaki, A variational approach to reconstructing images corrupted by Poisson noise, *Journal of mathematical imaging and vision* 27 (2007), 257–263.

[50] Liu, Luo, Wang & Tang, Deep Learning Face Attributes in the Wild, in: *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[51] Luo, Pang & Ralph, *Mathematical programs with equilibrium constraints*, Cambridge University Press, 1996.

[52] Mairal, Bach, Ponce & Sapiro, Online dictionary learning for sparse coding, in: *Proceedings of the 26th annual international conference on machine learning*, 2009, 689–696.

[53] Nikolova, A variational approach to remove outliers and impulse noise, *Journal of Mathematical Imaging and Vision* 20 (2004), 99–120.

[54] Nocedal & Wright, *Numerical Optimization*, Springer, 2006.

[55] Nocedal & Wright, *Numerical optimization*, Springer, 1999.

[56] Ochs, Ranftl, Brox & Pock, Bilevel optimization with nonsmooth lower level problems, in: *International Conference on Scale Space and Variational Methods in Computer Vision*, Springer, 2015, 654–665.

[57] Ochs, Ranftl, Brox & Pock, Techniques for gradient-based bilevel optimization with non-smooth lower level problems, *Journal of Mathematical Imaging and Vision* 56 (2016), 175–194.

[58] Olshausen & Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, *Nature* 381 (1996), 607–609.

[59] Outrata, A generalized mathematical program with equilibrium constraints, *SIAM Journal on Control and Optimization* 38 (2000), 1623–1638.

[60] Peyré & Fadili, Learning analysis sparsity priors, in: *Sampta'11*, 2011, 4–pp.

[61] Phillips, A technique for the numerical solution of certain integral equations of the first kind, *Journal of the ACM (JACM)* 9 (1962), 84–97.

[62] Poon & Peyré, Smooth Bilevel Programming for Sparse Regularization, in: *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2021, 1543–1555, URL: `proceedings.neurips.cc/paper/2021/file/0bed45bd5774ffddc95ffe5000Paper.pdf`.

[63] Qi & Sun, A trust region algorithm for minimization of locally Lipschitzian functions, *Mathematical Programming* 66 (1994), 25–43.

[64] Ranftl & Pock, A deep variational model for image segmentation, in: *German Conference on Pattern Recognition*, Springer, 2014, 107–118.

[65] Riis, Ehrhardt, Quispel & Schönlieb, A geometric integration approach to nonsmooth, nonconvex optimisation, *Foundations of Computational Mathematics* (2021), 1–44.

[66] Ring, Structural properties of solutions to total variation regularization problems, *ESAIM: Mathematical Modelling and Numerical Analysis* 34 (2000), 799–810.

[67] Rockafellar, *Convex analysis*, Princeton university press, 1970.

[68] Rockafellar & Wets, *Variational analysis*, Springer Science & Business Media, 2009.

[69] Rockafellar, *Convex analysis*, Princeton university press, 2015.

[70] Roth & Black, Fields of experts: A framework for learning image priors, in: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, IEEE, 2005, 860–867.

[71] Rudin, Osher & Fatemi, Nonlinear total variation based noise removal algorithms, *Physica D: nonlinear phenomena* 60 (1992), 259–268.

[72] Saquib, Bouman & Sauer, ML parameter estimation for Markov random fields with applications to Bayesian tomography, *IEEE transactions on Image Processing* 7 (1998), 1029–1044.

[73] Sawatzky, Brune, Müller & Burger, Total variation processing of images with Poisson statistics, in: *International Conference on Computer Analysis of Images and Patterns*, Springer, 2009, 533–540.

[74] Scholtes, *Introduction to piecewise differentiable equations*, Springer Science & Business Media, 2012.

[75] Sherry et al., Learning the sampling pattern for MRI, *IEEE Transactions on Medical Imaging* 39 (2020), 4310–4321.

[76] Stone, Cross-validation: A review, *Statistics: A Journal of Theoretical and Applied Statistics* 9 (1978), 127–139.

[77] Strang, *Linear algebra and its applications.* Belmont, CA: Thomson, Brooks/Cole, 2006.

[78] Tappen, Utilizing variational optimization to learn markov random fields, in: *2007 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2007, 1–8.

[79] Tikhonov & Arsenin, *Solutions of ill-posed problems*, Winston, Washington, DC, 1977.

[80] Valkonen, A primal–dual hybrid gradient method for nonlinear operators with applications to MRI, *Inverse Problems* 30 (2014), 055012.

[81] Van Chung, De los Reyes & Schönlieb, Learning optimal spatially-dependent regularization parameters in total variation image denoising, *Inverse Problems* 33 (2017), 074005.

[82] Villacís, First Order Methods for High Resolution Image Denoising, *Latin American Journal of Computing Faculty of Systems Engineering Escuela Politécnica Nacional Quito-Ecuador* 4 (2017), 37–42.

[83] Villacís, Bilevel Parameter Learning, `https://github.com/dvillacis/nonsmooth_bilevel_learning`, 2022.

[84] Voglis & Lagaris, A rectangular trust region dogleg approach for unconstrained and bound constrained nonlinear optimization, in: *WSEAS International Conference on Applied Mathematics*, Citeseer, 2004.

[85] Wang, Bovik, Sheikh, Simoncelli, et al., Image quality assessment: from error visibility to structural similarity, *IEEE transactions on image processing* 13 (2004), 600–612.

[86] Xu & Burke, ASTRAL: An active set l∞-trust-region algorithm for box constrained optimization, *Optimization online, July* (2007).

[87] Xue et al., Rethinking Bi-Level Optimization in Neural Architecture Search: A Gibbs Sampling Perspective, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 12, 2021, 10551–10559.

[88] Yu, Sapiro & Mallat, Image modeling and enhancement via structured sparse model selection, in: *2010 IEEE International Conference on Image Processing*, IEEE, 2010, 1641–1644.

[89] Zhang et al., SBO-RNN: Reformulating Recurrent Neural Networks via Stochastic Bilevel Optimization, in: *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2021, 25839–25851, URL: proceedings.neurips.cc/paper/2021/file/d87ca511e2a8593c8039ef732f5bffed-Paper.pdf.

Index