Higher Order Regularization for Model Based Data Decompression

Dissertation

zur Erlangung des akademischen Grades eines Doktor der Naturwissenschaften an der Karl-Franzens-Universität Graz

vorgelegt von

Mag. Martin Holler

am Institut für Mathematik und wissenschaftliches Rechnen

Erstbegutachter:
O.Univ.-Prof. Karl Kunisch
Zweitbegutachter:
Univ.-Prof. Antonin Chambolle, A.Univ.-Prof. Andreas Neubauer

2013

Gewidmet August Holler

Abstract

This thesis deals with the development, analysis and application of a mathematical model for image reconstruction. The model is realized by minimizing the sum of two convex functionals, one ensuring data fidelity and the other being a regularization term. The novelty of the considered approach lies both in the general definition of the data term as well as the application of the non-standard Total Generalized Variation (TGV) functional for regularization in this context. The main part of the work is the extension of theory for the TGV functional, the definition and analysis of the general reconstruction model and its application to imaging tasks. In particular, existence of a solution and optimality conditions for the resulting minimization problem are obtained for a setting that covers diverse applications: Suitable problem formulations for JPEG and JPEG 2000 decompression as well as wavelet based zooming are defined in function space setting and numerical solution schemes for the discretized problems are developed.

Zusammenfassung

Diese Doktorarbeit befasst sich mit der Entwicklung, Analyse und Anwendung eines mathematischen Modells zur Bildrekonstruktion. Das Modell wird durch die Minimierung der Summe zweier konvexer Funktionale realisiert. Eines dieser Funktionale stellt die Datentreue sicher während das andere zur Regularisierung dient. Neu an dem hier vorgestellten Zugang ist sowohl die allgemeine Formulierung des Datenterms als auch die Anwendung des *Total Generalized Variation* (TGV) Funktionals in diesem Kontext. Der Fokus der Arbeit liegt in der Erweiterung der Theorie über das TGV Funktional, der Definition und Analyse des allgemeinen Modells zur Bildrekonstruktion und dessen Anwendung für Problemstellungen der Bildverarbeitung. Im Speziellen kann für das aus dem Modell resultierende Minimierungsproblem die Existenz von Lösungen sichergestellt- und Optimalitätsbedingungen hergeleitet werden. Dies wird in einer allgemeinen Formulierung erreicht, die unterschiedliche Anwendungen erlaubt: Es werden geeignete Problemstellungen sowohl für JPEG und JPEG 2000 Dekomprimierung als auch Wavelet-basierte Bildvergrößerung im Funktionenraum formuliert und numerische Lösungsverfahren für die resultierenden, diskretisierten Probleme werden entwickelt.

Acknowledgements

First of all I want to thank Prof. Kunisch for being a fantastic supervisor and for always providing support for my work in every way possible. Without his untiring commitment, both the stimulating working environment at the institute and this thesis would not have been possible.

Next I want to express my deep gratitude to Prof. Bredies for more than three years of fruitful cooperation, in which he significantly influenced my scientific education. Without his ideas and advice, the content of this thesis would certainly not be the same. I also want to thank Prof. Chambolle and Prof. Neubauer for reviewing this thesis.

All of this would not have been possible without my parents, who I want to thank at this point for always providing unconditional support and trust and laying the basis for all I will ever achieve. I also want to thank my brother for being a good friend and for enabling mathematical discussions at my parents' home, even though this might sometimes be unfortunate for my other family members. Last but not least I want to thank Tanja for having been at my side for many years and for being a source of love and happiness in my life.

Contents

Introduction

The present work is concerned with mathematical image processing. A main part is the development and analysis of a particular type of image reconstruction model in its full generality. While great interest is put on a sound mathematical analysis and a broad applicability of this model, we point out that also successful real life applications are possible and have been fully developed within this work. In particular, a framework for the improved reconstruction of JPEG (Joint Photographic Experts Group) and JPEG 2000 compressed images is presented.

In the past 20 years, research on digital image processing has increasingly gained the interest of the mathematical community [6, 18]. Originally being an engineering topic image processing has become an independent and important field of applied mathematics. Defining an image as a function on a continuous domain allows the application of the full spectrum of applied mathematics to imaging problems. Mathematical image processing utilizes methods from optimal control, inverse problems, partial differential equations, functional analysis, numerical analysis and geometric measure theory [6, 3, 18]. Not only do these fields provide tools for mathematical image processing, but they themselves benefit from new ideas developed in imaging.

Typical image processing tasks can be divided into two areas: *Image reconstruction* and *image analysis*. While the first is mostly concerned with the reconstruction of image data for visualization purposes, the latter aims to get some structural information about an image for further processing. Of course, this separation is not strict. A good reconstruction of an image should display structural information that may be further processed by humans, for example neurologists that evaluate an MRI scan. But also for efficient image analysis, the reduction of noise, a generic image reconstruction problem, is necessary.

Following this rough classification, the present work is situated in the area of image reconstruction. In particular the applications we present later on aim to improve visual image quality.

While data modeling for image reconstruction methods depends heavily on the considered application, a typical commonality is the usage of a regularization functional. When faced with incomplete or noisy image data, some image model is used to correct or complete this data. The minimization of a regularization functional, subject to data constraints, realizes such an image model. Obviously the choice of a regularization term heavily influences visual reconstruction quality. Consequently, a lot of research has been carried out in this direction. A classical choice is certainly the *Total Variation* (TV) [60] functional, having the advantage of being a convex functional that allows jump discontinuities. While its non-differentiability was initially compensated by using a smooth modification [1], in particular duality based algorithms [24, 26] nowadays allow the direct solution of TV regularized, convex problems. In our problem setting, we use the non-standard, convex *Total Generalized Variation* (TGV) [16] functional as regularization. For an introduction about the functional and a comparison to other regularization terms we ask for the readers patience until section 3.

Since differentiability of a data term is very convenient for a numerical realization, the generic image reconstruction problem was the L^2 regularized denoising problem [44, 22, 25, 1]. Also L^1 regularization has become increasingly popular [45, 35, 29], having the advantage of allowing outliers. In the present work we use a non-differentiable L^∞ type constraint for data fidelity. This is motivated by the application to image decompression, where, besides a linear transform, also a quantization step is involved.

Even with growing storage capacities, due to the rise of high resolution digital cameras, efficient compression of digital images remains to be of importance. In the April edition of the *National Geographic* magazine, it is reported that in 2011, Americans have taken about *80 billion* digital photos. This number is expected to grow to *105 billion* photos in 2015. Being able to reduce the memory requirement of digital images significantly, the JPEG compression standard [68] for digital images is still an important tool to handle this huge amount of digital image data. More than 10 years after the introduction of its successor, the JPEG 2000 standard [64], still most digital cameras capture JPEG compressed images.

JPEG compression is lossy, which means that typically most of the compression is obtained by a loss of data and the original image cannot be obtained from the compressed JPEG file. Even though this may not be noticeable at low compression rates, it becomes clearly visible when using higher compression rates or zooming closer to image details. In particular, blocking and ringing effects, as can also be observed in figure 1, image A, are typical JPEG compression artifacts.

Given such a JPEG compressed image file, the initial motivation for this work was to apply methods of mathematical image processing to reduce or even remove these artifacts in order to get a reconstruction very close to the original image. Doing so would allow not only to get better reconstructions form already compressed JPEG files, but also to obtain a higher compression rate without loss of image quality. For previous attempt in this direction we refer to [2, 13] and the overview of subsection 5.2.

Starting from this motivation, we develop and analyze a general image reconstruction model that not only allows the improvement of JPEG and also JPEG 2000 compressed images, but also applications besides image decompression, for example image zooming.

This reconstruction model is based on an optimization problem of specific structure:

$$\min_u R(u) + \mathcal{I}_D(Au), \tag{1}$$

where

$$\mathcal{I}_D(v) = \begin{cases} 0 & \text{if } v \in D, \\ \infty & \text{else.} \end{cases}$$

With R being a regularization functional, the problem setting can be seen as Tikhonov regularized inverse problem. Two features are important for this kind of setting:

- First of all, the conditions on the data term $\mathcal{I}_D \circ A$, which reflect the type of imaging problems captured with this framework. We will define $A : H \to \ell^2$ to be a basis transformation operator with respect to a Riesz basis and $D \subset \ell^2$ to be a convex, closed set. The general concept of Riesz basis will allow to apply this setting for any continuous, linear operator $A : H \to \ell^2$ with closed range.

- Secondly, the choice of a suitable regularization term, which heavily influences the obtained reconstruction quality. We use the total generalized variation functional [16] of arbitrary order. While this functional yields a very good visual image quality, in particular avoids first order staircasing effects, some of its analytic properties still had to be developed partly within this work.

The core of the present work are certainly the results of section 4. After non-trivial preparations in section 3, there we will proof existence of a solution, as well as optimality conditions for the minimization problem related to TGV based image reconstruction in a general setting. For that purpose, besides a characterization of the subdifferential of the TGV functional, careful considerations of sufficient assumptions for existence of a solution as well as the derivation of non standard subdifferential calculus rules are necessary.

In section 5 we then show the applicability of the results of section 4 to image reconstruction problems, in particular the improved decompression of transform coded images as well as wavelet based image zooming. We refer to figure 1 for a visualization of results obtained in this context. It will be seen in these applications that the generality of the problem setting of section 4 was indeed necessary.

Let us now discuss the two terms of the minimization problem (1), the data and the regularization term, in more detail.

Data term As already mentioned, the motivation for the specific type of data fidelity term comes from the basis transform and quantization operation in transform based lossy image coding, in particular JPEG compression. This motivation, as well as the concrete problem setting for JPEG decompression, will be elaborated in more detail in the subsections 4.1 and 5.2. We also refer to subsections 5.2 and 5.3 for a literature overview of models with a similar data term structure. At this point we give just a formal introduction: It is an empirical obervation that the human visual system is less sensitive to high frequency brightness variations than to low frequency variations. Motivated by this fact, a basis transformation of a given uncompressed image u with respect to a block-cosine orthonormal basis is performed within the JPEG compression standard. We denote this transformation by BDCT. The result is an equivalent representation of the image as linear combination of different frequencies, expressed by the non integer coefficients $(\mathrm{BDCT}(u))_n$, for indices n. These coefficients are then quantized and rounded to integer, where stronger quantization is applied to coefficients reflecting high frequency variations. Finally, the integer data and the quantization values are saved lossless to disk as compressed JPEG file. Now reading such a JPEG file, we can determine the set of all possible source images of the compression process as $U_D = \{u \,|\, (\mathrm{BDCT}(u))_n \in J_n \text{ for all } n\}$, with $(J_n)_n$ closed, nonempty intervals resulting from the quantization and rounding. Clearly this set U_D can become very large. Thus when aiming to get a good reconstruction $\hat{u} \in U_D$ it is helpful to apply regularization. This yields the problem setting

$$\min_u R(u) + \mathcal{I}_{U_D}(u), \qquad U_D = \{u \,|\, (\mathrm{BDCT}(u))_n \in J_n \text{ for all } n\}.$$

Note that for JPEG decompression, the assumptions on U_D are relatively easy: The basis transformation operator is orthonormal and all intervals $(J_n)_n$ are closed and bounded with length greater than one. This will not be the case for

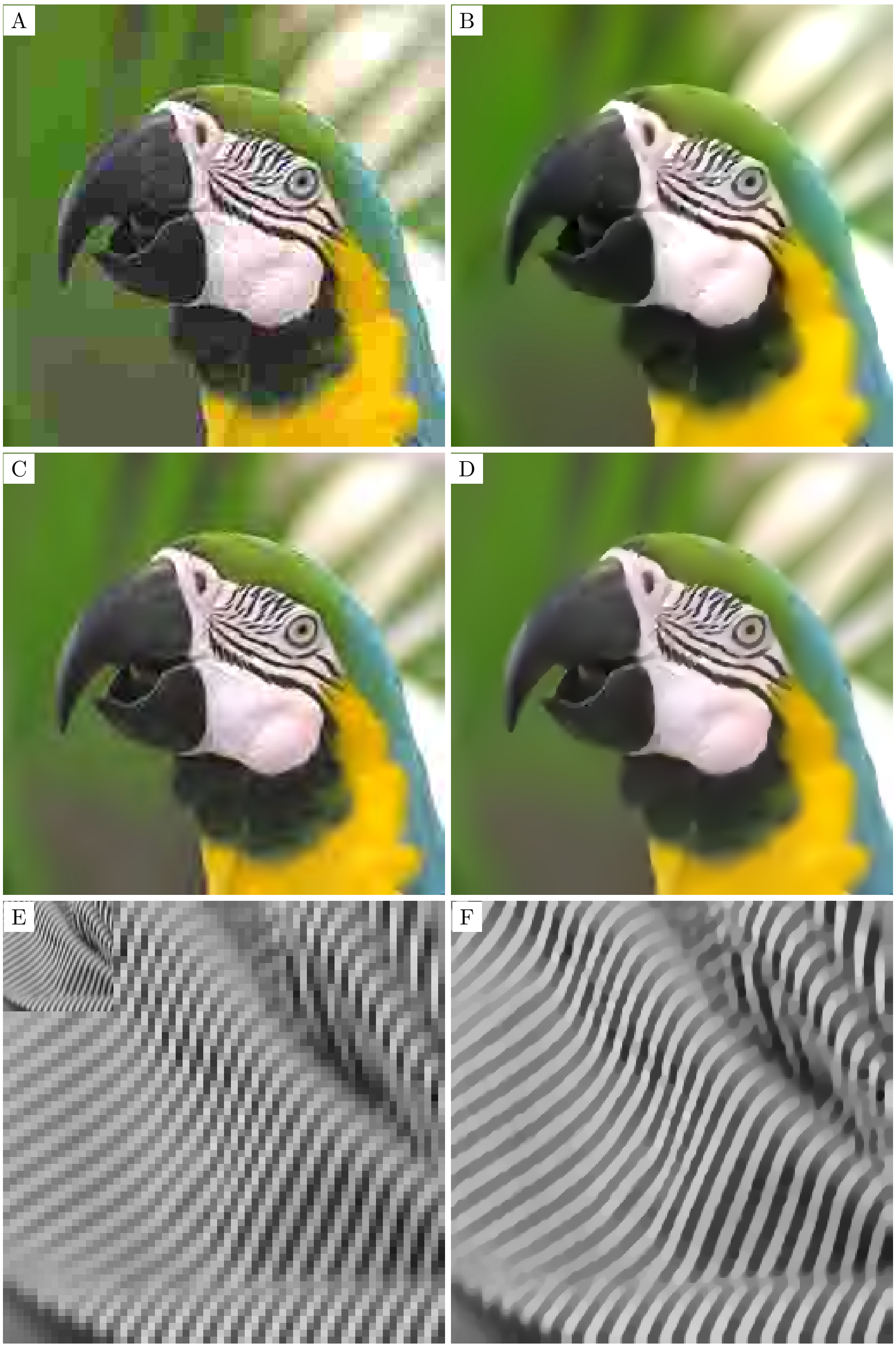


Figure 1: Applications of the TGV based image reconstruction model. A: Standard decompression of JPEG compressed image. B: TGV based reconstruction of the same JPEG compressed image. C: Standard decompression of JPEG 2000 compressed image. D: TGV based reconstruction of the same JPEG 2000 compressed image. E: Original, small image with stripe structure together with 4 times magnification by box filtering. F: TGV - wavelet based 4 times magnification of same image.

other applications, such as JPEG 2000 decompression and image zooming. To get a broad applicability of our model, we generalize the basis transformation operator BDCT to be a basis transformation operator related to a *Riezs basis*, allow finitely many J_n to be *point intervals* and all of them, except for very few, to be *unbounded.* For an exact definition of the generic problem assumption we refer to subsection 4.1.

Regularization term As regularization term we use the total generalized variation functional of arbitrary order $k \in \mathbb{N}$. It has been introduced in [16], for $u \in L^1_{\text{loc}}(\Omega)$, $\Omega \subset \mathbb{R}^d$, $\alpha \in \mathbb{R}^k_{>0}$, as

$$\mathrm{TGV}^{\mathrm{k}}_{\alpha}(u) = \sup \left\{ \int_{\Omega} u \operatorname{div}^k v \, \mathrm{d}x \,\middle|\, v \in C^k_c(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)), \right. \\ \left. \| \operatorname{div}^l v\|_\infty \leq \alpha_l, \, l = 0, \ldots, k-1 \right\}, \quad (2)$$

with $\mathrm{Sym}^k(\mathbb{R}^d)$ the space of symmetric tensors of order k (see section 2). For a detailed motivation for the usage of this functional, as well as a comparison to different regularization functionals, we ask for the readers patience until section 3. As brief motivation however, let us mention that the TGV functional can be seen as generalization of the well known total variation functional in the sense that it is a scalar multiple of the TV functional in the case $k = 1$. But, for $k \geq 2$, it incorporates information about higher order smoothness and does not suffer from first order staircasing effects [17]. It has been shown in [20] that the space of all L^1 functions having finite second order TGV^2_α functional is topologically equivalent to the well studied space $\mathrm{BV}(\Omega)$ [3]. Also, a Poincaré type inequality and an equivalent minimum representation for TGV^2_α have been obtained for $k = 2$. These results are very convenient for the usage of the TGV^2_α functional as regularization term. In the present work, in particular in section 3, we obtain that similar results also hold for the case of arbitrary order k. This allows us to obtain a characterization of the subdifferential of the $\mathrm{TGV}^{\mathrm{k}}_\alpha$ functional and, consequently, the derivation of an optimality condition for our problem setting.

We now conclude the introduction to the thesis by a short description of the contents of each section. Note that also at the beginning of each section, a short overview of its contents can be found.

- *Section 1: Preliminaries.* As the name suggests, this section repeats basic preliminaries. Only the remark on the connection between vector valued- and component wise Riesz bases, even if simple, is specific for our problem setting and should also be noted also by the advanced reader.

- *Section 2: Tensor calculus.* At first, this subsection repeats basic tensor calculus and fixes notation. Then, non standard results on functions of bounded deformation are presented, mainly taken from [10]. These functions will play an important role in the derivation of analytical properties of the TGV functional later on. Afterwards, the obtained results are generalized to product tensor spaces and, in the last subsection, the space $W^q(\operatorname{div}^k; \Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)$ is introduced.

- *Section 3: Regularization theory.* After giving an overview on regularization functionals, the total generalized variation functional is introduced. First, some properties of the TGV functional, that have already been shown in [16, 20], are stated. Afterwards, new results for the TGV functional of arbitrary order k are presented. The derivation of these results relies on [10, 11] and can also be found in [11]. At last, the generalization of the TGV functional to vector valued data as well as the notion of $\mathrm{TGV}_{\alpha}^{\mathrm{k}}$-strict convergence is introduced.

- *Section 4. The general reconstruction model.* This is one of the two main sections. Here, the generic problem setting for the TGV regularized image reconstruction model is stated and existence of a solution as well as optimality conditions are obtained. In particular, the derivation of the optimality condition includes a characterization of the subdifferential of the TGV functional of arbitrary order.

- *Section 5. Application to data reconstruction.* This is the second main section. After discussing the applicability of the reconstruction model of section 4 to a general class of problems, specific applications are considered. These applications are the decompression of JPEG and JPEG 2000 compressed images and a wavelet based zooming method. For those three particular problem settings, we define resulting minimization problems in function space and show that the results of section 4 are applicable. We also introduce a discretization and solve the discrete minimization problems by variants of a primal dual algorithm presented in [26]. Additionally, a theoretically justified stopping criterion is presented. For the application to JPEG decompression we present computation times of an optimized implementation adapted to multicore processors and graphic processing units. While the numerical solution of the JPEG decompression problem by the primal dual algorithm is relatively straightforward, for the JPEG 2000 decompression and wavelet zooming setting some work has to be done to get an efficient implementation for which convergence can be assured.

Part I
Functional analytic background

1 Preliminaries

The aim of this section is to introduce mathematical concepts that are of particular relevance for this work.

1.1 Basic function spaces and definitions

Throughout this subsection, we will fix $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ to be a domain and m a natural number, typically the dimension of an image space. At times it will be necessary to further specify Ω to be, for example, a bounded Lipschitz domain.

Definition 1.1. *Let $\mathcal{B}(\Omega)$ be the Borel σ-algebra of sets in Ω. A mapping $\mathcal{B}(\Omega) \to \mathbb{R}^m$ is called a $\mathbb{R}^m$ valued finite Radon measure on Ω if it is σ-additive and $\mu(\emptyset) = 0$. We denote the variation of $\mathbb{R}^m$ valued finite Radon measures μ by $|\mu| : \mathcal{B}(X) \to \mathbb{R}^m$, defined as*

$$|\mu|(E) = \sup \left\{ \sum_{i=0}^{\infty} |\mu(E_i)| \,\middle|\, (E_i)_{i \geq 0} \text{ in } \mathcal{B}(\Omega) \text{ pairwise disjoint, } E = \bigcup_{i=0}^{\infty} E_i \right\},$$

and by $\mathcal{M}(\Omega, \mathbb{R}^m)$ the space of all finite $\mathbb{R}^m$ valued Radon measures on Ω.

The following classical result can be found in [3, Theorem 1.54].

Proposition 1.1. *The space $\mathcal{M}(\Omega, \mathbb{R}^m)$ equipped with $\|\mu\|_{\mathcal{M}} := |\mu|$ is a Banach space. Further it can be considered as the dual of $C_0(\Omega, \mathbb{R}^m)$ with the duality pairing*

$$\langle \mu, \phi \rangle = \int_{\Omega} \phi \,\mathrm{D}\, \mu := \sum_{i=1}^{m} \int_{\Omega} \phi_i \,\mathrm{D}\, \mu_i,$$

for $\mu = (\mu_1, \ldots, \mu_m) \in \mathcal{M}(\Omega, \mathbb{R}^m)$, $\phi = (\phi_1, \ldots, \phi_m) \in C_0(\Omega, \mathbb{R}^m)$, where the norm $\| \cdot \|_{\mathcal{M}}$ is the dual norm.

Note that by $\langle \cdot, \cdot \rangle$ we always denote a duality pairing. As a consequence of the Radon-Nikodym theorem ([43, Theorem 31.B]), the following proposition holds.

Proposition 1.2. *Let μ be a finite Radon measure on Ω. Then there exists a unique function $\sigma_\mu \in L^1(\Omega, \mathbb{R}^m; |\mu|)$, the density of μ with respect to $|\mu|$, such that $|\sigma_\mu| = 1$, $|\mu|$- almost everywhere, and*

$$\mu(E) = \int_{E} \sigma_\mu \,\mathrm{D}\, |\mu|.$$

For the sake of completeness we also introduce functions of bounded variation (see [3]).

Definition 1.2. *For $m \in \mathbb{N}$ and $u = (u_1, \dots, u_m) \in L^1_{\text{loc}}(\Omega, \mathbb{R}^m)$ we define the* Total Variation *functional* $\mathrm{TV} : L^1_{\text{loc}}(\Omega, \mathbb{R}^m) \to \mathbb{R} \cup \{\infty\}$ *as*

$$\mathrm{TV}(u) = \sup \left\{ \sum_{i=1}^{m} \int_{\Omega} u_i \operatorname{div} \phi_i \,|\, \phi = (\phi_1 \dots, \phi_m)^T \in \mathcal{C}^1_c(\Omega, \mathbb{R}^{m \times d}),\ \|\phi\|_\infty \leq 1 \right\},$$

where $\|\phi\|_\infty = \sup_{x \in \Omega} \sqrt{\sum_{j=1}^m |\phi_j(x)|^2}$ *and* $|\cdot|$ *denotes the Euclidean norm on* $\mathbb{R}^d$*. We further define the space of functions of bounded variation*

$$\mathrm{BV}(\Omega, \mathbb{R}^m) = \{u \in L^1(\Omega, \mathbb{R}^m) \mid \mathrm{TV}(u) < \infty\}$$

and

$$\|u\|_{\mathrm{BV}} = \|u\|_{L^1} + \mathrm{TV}(u).$$

Remark 1.1. *Functions of bounded variation are well known, in particular in the field of mathematical image processing, and have been extensively studied in the literature. We repeat just some properties of functions of bounded variation that are the closest related to our work and refer to [3, 38, 75] for proofs and further information.*

Proposition 1.3. *A function* $u = (u_1, \dots, u_m) \in L^1(\Omega, \mathbb{R}^m)$ *belongs to* $\mathrm{BV}(\Omega, \mathbb{R}^m)$ *if and only if there exist finite Radon measures* $\mathrm{D}\, u_j = (\mathrm{D}_1\, u_j, \dots, \mathrm{D}_d\, u_j) \in \mathcal{M}(\Omega, \mathbb{R}^m)$, $1 \leq j \leq m$, *such that*

$$\int_\Omega u_j \operatorname{div} \phi \,\mathrm{d}x = \int_\Omega \phi \,\mathrm{d}\,\mathrm{D}\, u_j \quad \forall \phi \in C^\infty_c(\Omega, \mathbb{R}^d).$$

Proposition 1.4. *The functional* TV *is proper, convex and lower semi continuous (with respect to norm topology) in* $L^1(\Omega, \mathbb{R}^m)$*. Further* $\mathrm{TV}(u) = 0$ *if and only if there exist constants* $c_1, \dots c_m$ *such that* $u = (c_1, \dots, c_m)$*.*

Proposition 1.5. *Let* Ω *be a bounded Lipschitz domain. Then the embedding*

$$\begin{aligned} i : \mathrm{BV}(\Omega, \mathbb{R}^m) &\to L^p(\Omega, \mathbb{R}^m) \\ u &\mapsto i(u) = u \end{aligned}$$

is continuous for $1 \leq p \leq \frac{d}{d-1}$ *and compact for* $1 \leq p < \frac{d}{d-1}$*.*

Proposition 1.6. *Let* $u \in L^1(\Omega, \mathbb{R}^m)$*. Then* $u \in \mathrm{BV}(\Omega)$ *if and only if there exists a sequence* $(u_n)_{n \in \mathbb{N}}$ *in* $C^\infty(\Omega, \mathbb{R}^m)$ *such that*

$$\|u_n - u\|_{L_1} \to 0 \quad \text{and} \quad \mathrm{TV}(u_n) \to \mathrm{TV}(u).$$

1.2 The notion of Riesz basis

The following concept extents the classical notion of an orthonormal basis.

Definition 1.3 (Riesz Basis)**.** *Let H be a Hilbert space. We say that a sequence* $(\psi_n)_{n \in \mathbb{N}}$ *with* $\psi_n \in H$ *for all* $n \in \mathbb{N}$ *is a* Riesz Basis *of H if* $\mathrm{span}(\{\psi_n | n \in \mathbb{N}\})$ *is dense in H and there exist* $0 < A \leq B$ *such that, for any* $c = (c_i)_{i \in \mathbb{N}} \in \ell^2$*, we have*

$$A \sum_{n \in \mathbb{N}} c_n^2 \leq \Big\| \sum_{n \in \mathbb{N}} c_n \psi_n \Big\|_H^2 \leq B \sum_{n \in \mathbb{N}} c_n^2. \tag{3}$$

If $(\psi_n)_{n\in\mathbb{N}}$ is an orthonormal basis, equation (3) holds with $A = B = 1$. Thus orthonormal bases are indeed Riesz bases.

As the following proposition states, for any Riesz basis $(\psi_n)_{n\in\mathbb{N}}$ we can find a dual sequence $(\tilde{\psi}_n)_{n\in\mathbb{N}}$ that again is a Riesz basis.

Proposition 1.7. *Let $(\psi_n)_{n\in\mathbb{N}}$ be a Riesz basis of a Hilbert space H. Then there exists a sequence $(\tilde{\psi}_n)_{n\in\mathbb{N}}$ such that also $(\tilde{\psi}_n)_{n\in\mathbb{N}}$ is a Riesz basis of H and*

$$(\psi_i, \tilde{\psi}_j)_H = \delta_{i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{else.} \end{cases}$$

Proof. See [72, Theorem 1.9]. □

The following proposition summarizes the relation between Riesz bases and transformation operators

Proposition 1.8. *Let $(\psi_n)_{n\in\mathbb{N}}$ and $(\tilde{\psi}_n)_{n\in\mathbb{N}}$ be two dual Riesz basis. Then, the operators*

$$\begin{aligned} T :& H \to \ell^2 & \qquad \tilde{T} :& H \to \ell^2 \\ & u \mapsto ((u, \psi_n)_H)_{n\in\mathbb{N}} & & u \mapsto ((u, \tilde{\psi}_n)_H)_{n\in\mathbb{N}} \end{aligned}$$

are both continuous and possess continuous inverses with

$$T^{-1} = \tilde{T}^* \quad \tilde{T}^{-1} = T^*.$$

Their adjoints are given by

$$T^*((\lambda_n)_{n\in\mathbb{N}}) = \sum_{n\in\mathbb{N}} \lambda_n \psi_n \quad \tilde{T}^*((\lambda_n)_{n\in\mathbb{N}}) = \sum_{n\in\mathbb{N}} \lambda_n \tilde{\psi}_n.$$

Proof. It follows from [72, Theorem 1.9] that both $T, \tilde{T}$ are well defined. Equation (3) immediately implies that the operators T^* and $\tilde{T}^*$ are also well defined and, as ease consequence, are adjoint to T and $\tilde{T}$, respectively. Injectivity also follows and, together with density of $(\psi_n)_{n\in\mathbb{N}}, (\tilde{\psi}_n)_{n\in\mathbb{N}}$, implies bijectivity of $T^*, \tilde{T}^*$. Biorthogonality finally yieds $\tilde{T}^* = T^{-1}$, $T^* = \tilde{T}^{-1}$ and thus all claimed assertions hold. □

As can be shown, the notion of Riesz basis is the most general basis concept that ensures a sequence to be complete and the resulting basis transformation to be continuous and continuously invertible. As we will see later on, this fits perfectly to our data modeling in the general image reconstruction setting of part II.

There, we will mainly deal with component wise bases, e.g. m different Riesz bases of $L^2(\Omega)$. The following remark emphasizes the connection to Riesz bases in $L^2(\Omega, \mathbb{R}^m)$. In particular, also interval restrictions on the component wise bases naturally transfer to a related vector valued basis.

Remark 1.2. *For $m \in \mathbb{N}$, $1 \leq i \leq m$, let $(a_n^i)_{n\in\mathbb{N}}$ be Riesz bases of $L^2(\Omega)$. Then, $(\overline{a}_n)_{n\in\mathbb{N}}$ defined by*

$$\overline{a}_n^i = \begin{cases} a_k & \text{for } i = j+1 \text{ with } n-1 = km+j,\ k,j \in \mathbb{N}_0, \\ 0 & \text{else,} \end{cases}$$

is a Riesz basis of $L^2(\Omega,\mathbb{R}^m)$*. Further, given any* $u=(u_1,\dots,u_m)\in L^2(\Omega,\mathbb{R}^m)$*, and intervals* $(J_n^i)_{\substack{1\leq i\leq m,\\ n\in\mathbb{N}}}$

$$(a_n^i,u^i)_{L^2}\in J_n^i \quad \textit{for all } 1\leq i\leq m, n\in\mathbb{N}$$

is equivalent to

$$(\overline{a}_n,u)\in\overline{J}_n \quad \textit{for all } n\in\mathbb{N},$$

where each $\overline{J}_n$ *coincides with exactly one* J_n^i*.*

2 Tensor calculus

Essential for the usage of the total generalized variation functional, that will serve as regularization term in our models, is the notion of tensors and tensor fields. In the following we will therefore state the main definitions and results in this context.

In the first subsection, we will introduce functions mapping to tensor spaces, and in its succeeding subsection state a straightforward generalization to functions mapping to products of tensor spaces. This generalization is necessary to apply the regularization theory also for vector-valued functions, which correspond to color images. Throughout this section, we denote by $\Omega\subset\mathbb{R}^d$ a domain with $d\in\mathbb{N}$ its dimension. Again Ω will be further specified to be a bounded Lipschitz domain, if necessary. Further k and m will always be natural numbers, typically used to denote the order of tensors and again the dimension of an image space, respectively.

2.1 Tensor spaces and related mappings

This subsection is devoted to repeat basic tensor- and tensor field theory for the purpose of defining the total generalized variation for scalar valued functions. We will mainly follow the introduction given in [16, Section 2]. For more information we refer to [7, 61]. At first, remember the definition of a k-tensor space and a symmetric k-tensor space

$$\mathcal{T}^k(\mathbb{R}^d):=\left\{\xi:\left(\mathbb{R}^d\right)^k\to\mathbb{R}\,\middle|\,\xi:\ k-\text{linear}\right\}, \tag{4}$$

$$\mathrm{Sym}^k(\mathbb{R}^d):=\left\{\xi:\left(\mathbb{R}^d\right)^k\to\mathbb{R}\,\middle|\,\xi:\ k-\text{linear and symmetric}\right\}, \tag{5}$$

with the scalar product

$$\xi\cdot\eta=\mathrm{tr}^k(\overline{\xi}\otimes\eta)=\sum_{p\in\{1,\dots d\}^k}\xi(e_{p_1},\dots,e_{p_k})\eta(e_{p_1},\dots,e_{p_k}), \tag{6}$$

for $\xi,\eta\in\mathcal{T}^k(\mathbb{R}^d)$, and the induced norm

$$|\xi|=\sqrt{\xi\cdot\xi}. \tag{7}$$

Here, e_i denotes the ith standard basis element of $\mathbb{R}^d$, i.e.,

$$(e_i)_j=\begin{cases}1 & \text{if } i=j,\\ 0 & \text{else,}\end{cases}$$

and $\sum_{p\in\{1,\dots,d\}^k}$ means to take the sum over all possible $p = (p_1, \dots, p_k)$ with $p_i \in \{1, \dots, d\}$, $1 \le i \le k$.

Further, for $\xi \in \mathcal{T}^k(\mathbb{R}^d)$, $\eta \in \mathcal{T}^l(\mathbb{R}^d)$, we define

$$\overline{\xi}(a_1, \dots, a_k) = \xi(a_k, \dots, a_1),$$

$$\xi \otimes \eta \in \mathcal{T}^{k+l}(\mathbb{R}^d), \quad \xi \otimes \eta(a_1, \dots, a_{k+l}) = \xi(a_1, \dots, a_k)\eta(a_{k+1}, \dots, a_{k+l})$$

and

$$\operatorname{tr}(\xi) \in \mathcal{T}^{k-2}(\mathbb{R}^d), \quad \operatorname{tr}(\xi)(a_1, \dots, a_{k-2}) = \sum_{i=1}^{d} \xi(e_i, a_1, \dots, a_{k-2}, e_i).$$

A tensor $\xi \in \mathcal{T}$ is called symmetric if $\xi(a_1, \dots, a_k) = \xi(a_{\pi(1)}, \dots, a_{pi(k)})$ for all permutations $\pi : \{1, \dots, k\} \to \{1, \dots, k\}$.

With these definitions, $\operatorname{Sym}^k(\mathbb{R}^d)$ is a finite dimensional vector space (see [16, Section 2] for a representation of its basis elements) equipped with an inner product.

Next we define the space of symmetric k-tensor fields as mappings $\xi : \Omega \to \operatorname{Sym}^k(\mathbb{R}^d)$ and the associated Lebesgue spaces as

$$L^p(\Omega, \operatorname{Sym}^k(\mathbb{R}^d)) = \left\{ \xi : \Omega \to \operatorname{Sym}^k(\mathbb{R}^d) \text{ measureable, identified a.e.} \;\middle|\; \|\xi\|_p < \infty \right\}, \tag{8}$$

where

$$\|\xi\|_p = \left(\int_\Omega |\xi(x)|^p \, \mathrm{d}x \right)^{\frac{1}{p}}, \text{ for } 1 \le p < \infty, \quad \|\xi\|_\infty = \operatorname*{ess\,sup}_{x\in\Omega} |\xi(x)|.$$

We will need symmetric k-tensor fields that are differentiable. At first, for a given, sufficiently smooth, k-tensor field ξ, its lth derivative can be identified with a (non-symmetric) $(k+l)$ tensor field, defined by

$$(\nabla^l \otimes \xi)(x)(a_1, \dots, a_{k+l}) = \left(\mathrm{D}^l \, \xi(x)(a_1, \dots, a_l) \right) (a_{l+1}, \dots, a_{l+k}),$$

where $\mathrm{D}^l \, \xi : \Omega \to \mathcal{L}^l\left(\mathbb{R}^d, \operatorname{Sym}^k(\mathbb{R}^d)\right)$ denotes the lth Fréchet derivative of ξ and $\mathcal{L}^l(X, Y)$ the space of l-linear and continuous mappings from X^l to Y. Further we define a symmetrized derivative of a smooth tensor field $\xi : \Omega \to \operatorname{Sym}^k(\mathbb{R}^d)$ that can be identified with a symmetric tensor field:

$$\mathcal{E}^l(\xi) = |||(\nabla^l \otimes \xi) = (|||(\nabla\otimes))^l \xi, \tag{9}$$

where the last identity is shown in [16, Section 2]. Here $|||\eta$ denotes the symmetrization of a given tensor $\eta \in \mathcal{T}^k(\mathbb{R}^d)$ defined by

$$(|||\eta)(a_1, \dots, a_k) = \frac{1}{k!} \sum_{\pi \in S_k} \eta(a_{\pi(1)}, \dots a_{\pi(k)}),$$

where S_k is the set of all permutations of $\{1, \dots, k\}$. Note that, as can be verified by direct calculation, $|||$ is an orthogonal projection from $\mathcal{T}^k(\mathbb{R}^d)$ to $\operatorname{Sym}^k(\mathbb{R}^d)$.

We now define the space of continuously differentiable symmetric k-tensor fields as

$$\mathcal{C}^l(\overline{\Omega}, \mathrm{Sym}^k(\mathbb{R}^d)) = \Big\{ \xi : \overline{\Omega} \to \mathrm{Sym}^k(\mathbb{R}^d) \\ \Big| \nabla^i \otimes \xi \text{ continuous on } \overline{\Omega},\ i = 1, \ldots, l \Big\}. \quad (10)$$

In case Ω is bounded, this becomes a Banach space when equipped with the norm

$$\|\xi\|_{l,\infty} = \max_{m=0,\ldots,l} \|\nabla^m \otimes \xi\|_\infty$$

where $\|\xi\|_\infty = \max\limits_{x \in \overline{\Omega}} |\xi(x)|$. We will also need continuously differentiable k-tensor fields which are compactly supported in Ω:

$$\mathcal{C}^l_c(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)) = \Big\{ \xi \in \mathcal{C}^l(\overline{\Omega}, \mathrm{Sym}^k(\mathbb{R}^d)) \mid \mathrm{supp}\, \xi \subset\subset \Omega \Big\},$$

$$\mathcal{C}^\infty_c(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)) = \bigcap_{l \geq 0} \mathcal{C}^l_c(\overline{\Omega}, \mathrm{Sym}^k(\mathbb{R}^d)).$$

The space of distributions on Ω is further defined as

$$\mathcal{D}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)) = \mathcal{C}^\infty_c(\Omega, \mathrm{Sym}^k(\mathbb{R}^d))^*.$$

Moreover, we use the notion of l-divergence of a sufficiently smooth $(k+l)$ tensor field:

$$\mathrm{div}^l\, \eta = \mathrm{tr}^l(\nabla^l \otimes \eta)$$

for $\eta \in \mathcal{T}^{k+l}(\mathbb{R}^d)$. Note that by definition of the trace operator, the divergence of η is symmetric if $\eta \in \mathrm{Sym}^{k+l}(\mathbb{R}^d)$.

Based on that, we can define the distributional derivative and symmetrized derivative as follows:

Definition 2.1. *For $\xi \in \mathcal{D}(\Omega, \mathcal{T}^k(\mathbb{R}^d))$,*

- *$\eta \in \mathcal{D}(\Omega, \mathcal{T}^{k+1}(\mathbb{R}^d))$ is called the weak derivative of ξ if*

$$\langle \eta, \zeta \rangle = -\langle \xi, \mathrm{div}\, \zeta \rangle$$

for all $\zeta \in C^1_c(\Omega, \mathcal{T}^{k+1}(\mathbb{R}^d))$. In this case we denote $\mathrm{D}(\xi) = \eta$.

- *$\eta \in \mathcal{D}(\Omega, \mathrm{Sym}^{k+1}(\mathbb{R}^d))$ is called the weak symmetrized derivative of ξ if*

$$\langle \eta, \zeta \rangle = -\langle \xi, \mathrm{div}\, \zeta \rangle$$

for all $\zeta \in C^1_c(\Omega, \mathrm{Sym}^{k+1}(\mathbb{R}^d))$. In this case we denote $\mathcal{E}(\xi) = \eta$.

At last in this subsection, we introduce symmetric k tensor fields of bounded deformation. We refer to [65, Chapter II] as a classical reference on this topic for the case $k = 1$ and to [10] for the nontrivial generalization to arbitrary k.

Definition 2.2. *The total deformation of a function* $\xi \in L^1_{\mathrm{loc}}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d))$ *is defined as*

$$\mathrm{TD}(\xi) = \sup \left\{ \int\limits_\Omega \xi \cdot \operatorname{div} \eta \, \mathrm{d}x \middle| \eta \in C^1_c(\Omega, \mathrm{Sym}^{k+1}(\mathbb{R}^d)), \|\eta\|_\infty \leq 1 \right\}. \tag{11}$$

The space

$$\mathrm{BD}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)) = \{\xi \in L^1(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)) | \, \mathrm{TD}(\xi) < \infty\}$$

equipped with

$$\|\xi\|_{\mathrm{BD}} = \|\xi\|_1 + \mathrm{TD}(\xi)$$

is called the space of symmetric k tensor fields of bounded deformation.

The following proposition summarizes essential properties of $\mathrm{BD}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d))$ and the weak symmetrized derivative.

Proposition 2.1. *Let Ω be a bounded Lipschitz domain. Then the following holds:*

1. *The weak symmetrized gradient, as mapping from* $\mathrm{BD}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)) \subset \mathcal{M}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d))$ *to* $\mathcal{M}(\Omega, \mathrm{Sym}^{k+1}(\mathbb{R}^d))$*, is closed with respect to weak* topology.*

2. *The space* $\mathrm{BD}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d))$ *equipped with* $\| \cdot \|_{\mathrm{BD}}$ *is a Banach space.*

3. $\xi \in \mathrm{BD}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d))$ *if and only if* $\xi \in L^1(\Omega, \mathrm{Sym}^k(\mathbb{R}^d))$ *and possesses a weak symmetrized derivative* $\mathcal{E}(\xi) \in \mathcal{M}(\Omega, \mathrm{Sym}^{k+1}(\mathbb{R}^d))$.

4. *The embedding* $\mathrm{BD}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)) \hookrightarrow L^p(\Omega, \mathrm{Sym}^k(\mathbb{R}^d))$ *is continuous for* $1 \leq p \leq d/(d-1)$ *and compact for* $1 \leq p < d/(d-1)$.

5. *The kernel of $\mathcal{E}$ in* $\mathcal{D}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d))$ *is a subset of the finite dimensional space of* $\mathrm{Sym}^k(\mathbb{R}^d)$ *valued polynomials of maximal degree k.*

6. *For* $R : L^{d/(d-1)}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)) \to \ker(\mathcal{E})$ *a linear, continuous onto projection, there exists a $C > 0$ such that, for all* $u \in \mathrm{BD}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d))$,

$$\|u - Ru\|_{d/(d-1)} \leq C\|\mathcal{E}u\|_{\mathcal{M}}.$$

Proof. The first and second assertion can be found [10, Proposition 4.2] and [10, Proposition 4.4], respectively. The equivalence of the third assertion follows in one direction from $C_0(\Omega, \mathrm{Sym}^k(\mathbb{R}^d))^* = \mathcal{M}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d))$ by standard arguments and in the other direction from the definition of the weak symmetrized derivative and the fact that $\mathrm{TD}(\xi) = |\mathcal{E}(\xi)|(\Omega) < \infty$. The fourth assertion can be found in [10, Theorems 4.16, 4.17], the fifth in [10, Proposition 3.3] and the last one follows trivially from [10, Theorems 4.16, 4.19]. □

The third assertion of the previous proposition can be strengthened: It is known [65, Theorem 2.3] that a distribution $u \in \mathcal{D}(\Omega, \mathbb{R}^d)$ can be represented by a function of bounded deformation if $\mathcal{E}u \in \mathcal{M}(\Omega, \mathrm{Sym}^2(\mathbb{R}^d))$. In [11] this has been shown for the general setting:

Proposition 2.2. *Let Ω be a bounded Lipschitz domain. If for* $u \in \mathcal{D}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d))$ *we have* $\mathcal{E}u \in \mathcal{M}(\Omega, \mathrm{Sym}^{k+1}(\mathbb{R}^d))$*, then* $u \in \mathrm{BD}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d))$.

2.2 Products of tensor spaces and related mappings

Now we deal with functions mapping to products of tensor spaces, with the aim of defining the total generalized variation functional for vector valued data later on.

A straightforward generalization of $\mathrm{Sym}^k(\mathbb{R}^d)$ to a space containing m-tuples of symmetric tensors, is the following:

$$\mathrm{Sym}^k(\mathbb{R}^d)^m = \Big\{ \xi = (\xi_1, \ldots, \xi_m) \,\big|\, \xi_i \in \mathrm{Sym}^k(\mathbb{R}^d),\ i \in \{1, \ldots, m\} \Big\} \tag{12}$$

with the norm

$$|\xi| = \big|(|\xi_1|, \ldots, |\xi_m|)\big|_v \tag{13}$$

where for the moment $|\cdot|_v$ denotes an arbitrary vector norm on the space $\mathbb{R}^m$. With that, also the generalization of integrable and differentiable tensor fields is straightforward:

$$L^p(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m) = \Big\{ \xi : \Omega \to \mathrm{Sym}^k(\mathbb{R}^d)^m \text{ measureable, identified a.e.} \,\big|\, \|\xi\|_p < \infty \Big\}, \tag{14}$$

where

$$\|\xi\|_p = \Bigg(\int_\Omega |\xi(x)|^p \,\mathrm{d}x \Bigg)^{\frac{1}{p}}, \text{ for } 1 \le p < \infty, \quad \|\xi\|_\infty = \operatorname*{ess\,sup}_{x \in \Omega} |\xi(x)|.$$

Note that, by equivalence of norms in $\mathbb{R}^m$, the choice of a particular norm $|\cdot|_v$ in (13) does not influence definition (14). Similarly we generalize

$$\begin{aligned}\mathcal{C}^l(\overline{\Omega}, \mathrm{Sym}^k(\mathbb{R}^d)^m) = \Bigg\{ &\xi = (\xi_1, \ldots, \xi_m) : \overline{\Omega} \to \mathrm{Sym}^k(\mathbb{R}^d)^m \\ &\big| \nabla^j \otimes \xi = (\nabla^j \otimes \xi_1, \ldots, \nabla^j \otimes \xi_m) \text{ is continuous on } \overline{\Omega},\ j = 0, \ldots l, \Bigg\},\end{aligned} \tag{15}$$

with a corresponding norm

$$\|\xi\|_{l,\infty} = \max_{m=0,\ldots,l} \|\nabla^m \otimes \xi\|_\infty,$$

and compactly supported differentiable tensor fields

$$\mathcal{C}_c^l(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m) = \Big\{ \xi \in \mathcal{C}^l(\overline{\Omega}, \mathrm{Sym}^k(\mathbb{R}^d)^m) \mid \operatorname{supp} \xi \subset\subset \Omega \Big\},$$

$$\mathcal{C}_c^\infty(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m) = \bigcap_{l \ge 0} \mathcal{C}_c^l(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m).$$

The space of distributions on Ω is accordingly defined as

$$\mathcal{D}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m) = \mathcal{C}_c^\infty(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)^*.$$

We extend the notion of lth order divergence for a sufficiently smooth $(k+l)$ product tensor field ξ as:

$$\operatorname{div}^l \xi = (\operatorname{div}^l \xi_1, \ldots, \operatorname{div}^l \xi_m).$$

Again this yields a symmetric product tensor field in case ξ is symmetric.

Remark 2.1. *In principle, one can choose an arbitrary norm* $|\cdot|_v$ *on* $\mathbb{R}^m$ *in the definition of* $\operatorname{Sym}^k(\mathbb{R}^d)^m$. *However, if the norm does not result from an inner product, we can no longer, as in the previous section, identify* $(\operatorname{Sym}^k(\mathbb{R}^d)^m)^*$ *with* $\operatorname{Sym}^k(\mathbb{R}^d)^m$, *which has to be taken into account also for the dual spaces of the corresponding product tensor field spaces. Also, as mentioned in [9, Remark 2], a useful generalization would amount to allow an arbitrary norm on the whole* $\operatorname{Sym}^k(\mathbb{R}^d)^m$, *not only its vector components. This would result in a discussion about an improved color distance, rather than the Euclidean, that is beyond the scope of our current work and for which we refer to [41] and [18, Section 6.3].*

Motivated by the previous remark, we will in the following always assume $|\cdot|_v = |\cdot|_{\text{eukl}}$, i.e. the Euclidean norm, and hence identify $(\operatorname{Sym}^k(\mathbb{R}^d)^m)^*$ with $(\operatorname{Sym}^k(\mathbb{R}^d)^m)$. With that choice, the norms on $\mathcal{T}^k(\mathbb{R}^d)^m$ and $\operatorname{Sym}^k(\mathbb{R}^d)^m$ are again induced by the inner product

$$\xi \cdot \eta = \sum_{i=1}^{m} \xi_i \cdot \eta_i$$

for $\xi = (\xi_1, \ldots, \xi_m)$ and $\eta = (\eta_1, \ldots, \eta_m)$ contained either in $\mathcal{T}^k(\mathbb{R}^d)^m$ or $\operatorname{Sym}^k(\mathbb{R}^d)^m$.

Also the generalization of the notion of distributional derivative is straightforward.

Definition 2.3. *For* $\xi \in \mathcal{D}(\Omega, \mathcal{T}^k(\mathbb{R}^d)^m)$,

- $\eta \in \mathcal{D}(\Omega, \mathcal{T}^{k+1}(\mathbb{R}^d)^m)$ *is called the weak derivative of* ξ *if*

  $$\langle \eta, \zeta \rangle = -\langle \xi, \operatorname{div} \zeta \rangle$$

  for all $\zeta \in C_c^1(\Omega, \mathcal{T}^{k+1}(\mathbb{R}^d)^m)$. *In this case we denote* $\mathrm{D}(\xi) = \eta$.

- $\eta \in \mathcal{D}(\Omega, \operatorname{Sym}^{k+1}(\mathbb{R}^d)^m)$ *is called the weak symmetrized derivative of* ξ *if*

  $$\langle \eta, \zeta \rangle = -\langle \xi, \operatorname{div} \zeta \rangle$$

  for all $\zeta \in C_c^1(\Omega, \operatorname{Sym}^{k+1}(\mathbb{R}^d)^m)$. *In this case we denote* $\mathcal{E}(\xi) = \eta$.

Again we need the space of functions of bounded deformation, but for product tensor valued mappings:

Definition 2.4. *The total deformation of a function* $\xi \in L^1(\Omega, \operatorname{Sym}^k(\mathbb{R}^d)^m)$ *is defined as*

$$\mathrm{TD}(\xi) = \sup \left\{ \int_\Omega \xi \cdot \operatorname{div} \eta \, \mathrm{d}x \middle| \eta \in C_c^1(\Omega, \operatorname{Sym}^{k+1}(\mathbb{R}^d)^m), \|\eta\|_\infty \leq 1 \right\}. \quad (16)$$

The space

$$\mathrm{BD}(\Omega, \operatorname{Sym}^k(\mathbb{R}^d)^m) = \{\xi \in L^1(\Omega, \operatorname{Sym}^k(\mathbb{R}^d)^m) |\, \mathrm{TD}(\xi) < \infty\}$$

equipped with the norm

$$\|\xi\|_{BD} = \|\xi\|_1 + \mathrm{TD}(\xi)$$

is called the space of symmetric k *tensor fields of bounded deformation.*

Note that we abuse notation by using the same notation for the total deformation functional also in the product tensor case. Equivalence of norms in the space $\mathbb{R}^m$ implies that all results for proposition 2.1 hold also in the product tensor case:

Proposition 2.3. *Let Ω be a bounded Lipschitz domain. The following holds:*

1. *The weak symmetrized gradient, as mapping from $\mathrm{BD}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m) \subset \mathcal{M}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m) \to \mathcal{M}(\Omega, \mathrm{Sym}^{k+1}(\mathbb{R}^d)^m)$, is closed with respect to weak* convergence.*
2. *The space $\mathrm{BD}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)$ equipped with $\|\cdot\|_{\mathrm{BD}}$ is a Banach space.*
3. *$\xi \in \mathrm{BD}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)$ if and only if $\xi \in L^1(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)$ and possesses a weak symmetrized derivative $\mathcal{E}(\xi) \in \mathcal{M}(\Omega, \mathrm{Sym}^{k+1}(\mathbb{R}^d)^m)$.*
4. *The embedding $\mathrm{BD}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m) \hookrightarrow L^p(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)$ is continuous for $1 \leq p \leq d/(d-1)$ and compact for $1 \leq p < d/(d-1)$.*
5. *The kernel of $\mathcal{E}$ in $\mathcal{D}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)$ is a subset of the finite dimensional space of $\mathrm{Sym}^k(\mathbb{R}^d)^m$ valued polynomials of maximal degree k.*
6. *For $R : L^{d/(d-1)}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m) \to \ker(\mathcal{E})$ a linear, continuous, onto projection, there exists a $C > 0$ such that, for all $u \in \mathrm{BD}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)$,*

$$\|u - Ru\|_{d/(d-1)} \leq C\|\mathcal{E}u\|_{\mathcal{M}}.$$

Proof. The first assertion follows immediately from equivalence of norms in $\mathbb{R}^m$ and the first assertion from proposition 2.1. The description of $\ker(\mathcal{E}) \subset \mathcal{D}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)$ as in the fifth assertion can also be deduced from proposition 2.1 as follows: Given any $u \in \mathcal{D}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)$ we can define $u^i \in \mathcal{D}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d))$ by

$$\langle u^i, \phi \rangle = \langle u, (\underbrace{0, \ldots, 0}_{(i-1) \text{ times}}, \phi, \underbrace{0 \ldots, 0}_{(m-i) \text{ times}}) \rangle$$

for $\phi \in C_c^\infty(\Omega, \mathrm{Sym}^k(\mathbb{R}^d))$. Thus if $u \in \ker(\mathcal{E})$ it follows that $u^i \in \ker(\mathcal{E}) \subset \mathcal{D}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d))$, $1 \leq i \leq m$. Thus, by proposition 2.1 $(u^1, \ldots, u^m)$ is contained in a subset of the space of $\mathrm{Sym}^k(\mathbb{R}^d)^m$ valued polynomials of maximal degree k and it further coincides with u in $\mathcal{D}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)$, thus u is also contained in that set.

Again by equivalence of norms, there exist constants $\tilde{c}, \tilde{C} > 0$ such that

$$\tilde{c} \sum_{i=1}^m \|\eta_i\|_\infty \leq \|\eta\|_\infty \leq \tilde{C} \sum_{i=1}^m \|\eta_i\|_\infty$$

for any $\eta = (\eta_1, \ldots, \eta_m) \in \mathrm{Sym}^k(\mathbb{R}^d)^m$. Using this, it is easy to show that

$$c \sum_{i=1}^m \mathrm{TD}(\xi_i) \leq \mathrm{TD}(\xi) \leq C \sum_{i=1}^m \mathrm{TD}(\xi_i),$$

for constants $C, c > 0$ and all $\xi = (\xi_1, \ldots, \xi_m) \in \mathrm{Sym}^k(\mathbb{R}^d)^m$. From that, the remaining assertions follows by straightforward argumentation. □

Again we have the stronger result that a distribution is represented by an L^1 function if its symmetrized gradient can be represented by a Radon measure.

Proposition 2.4. *Let Ω be a bounded Lipschitz domain. If for $u \in \mathcal{D}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)$ we have $\mathcal{E}u \in \mathcal{M}(\Omega, \mathrm{Sym}^{k+1}(\mathbb{R}^d)^m)$, then $u \in \mathrm{BD}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)$.*

Proof. This can be deduced from the $\mathrm{Sym}^k(\mathbb{R}^d)$ valued case and equivalence of norms by applying proposition 2.2 to distributions $u_i \in \mathcal{D}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d))$ as defined in the proof of proposition 2.3.

□

2.3 The space $W^q(\mathrm{div}^k; \Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)$

The purpose of this section is to introduce the space $W^q(\mathrm{div}^k; \Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)$, that will be needed for the description of the subdifferential of $\mathrm{TGV}_\alpha^\mathrm{k}$ as well as the optimality condition of our general minimization problem for image reconstruction. Note that this space is a generalization of the space $H(\mathrm{div}; \Omega)$, as described for example in [40, Chapter 1], and also many properties can easily be generalized.

Definition 2.5 (The space $W^q(\mathrm{div}^k; \Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)$)**.** *Let $1 \leq q < \infty$, $g \in L^q(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)$. We say that $\mathrm{div}\, g \in L^q(\Omega, \mathrm{Sym}^{k-1}(\mathbb{R}^d)^m)$ if there exists $w \in L^q(\Omega, \mathrm{Sym}^{k-1}(\mathbb{R}^d)^m)$ such that for all $\phi \in C_c^\infty(\Omega, \mathrm{Sym}^{k-1}(\mathbb{R}^d)^m)$*

$$\int\limits_\Omega \nabla \phi \cdot g = - \int\limits_\Omega \phi \cdot w.$$

Furthermore we define

$$W^q(\mathrm{div}^k; \Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m) = \Big\{ g \in L^q(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m) \mid \\ \mathrm{div}^l\, g \in L^q(\Omega, \mathrm{Sym}^l(\mathbb{R}^d)^m) \text{ for all } 1 \leq l \leq k \Big\}$$

with the norm $\|g\|^q_{W(\mathrm{div}^k)} := \sum_{l=0}^k \|\,\mathrm{div}^l\, g\|^q_{L^q}$.

Remark 2.2. *Density of $C_c^\infty(\Omega, \mathrm{Sym}^{k-1}(\mathbb{R}^d)^m)$ in $L^q(\Omega, \mathrm{Sym}^{k-1}(\mathbb{R}^d)^m)$ implies that, if there exists $w \in L^q(\Omega, \mathrm{Sym}^{k-1}(\mathbb{R}^d)^m)$ as above, it is unique. Hence it makes sense to write $\mathrm{div}\, g = w$. By completeness of $L^q(\Omega, \mathrm{Sym}^l(\mathbb{R}^d)^m)$, for $0 \leq l \leq k$ it follows that $W^q(\mathrm{div}^k; \Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)$ is a Banach space when equipped with $\|\cdot\|_{W^q(\mathrm{div}^k)}$.*

Definition 2.6. *We define, again for $1 \leq q < \infty$,*

$$W_0^q(\mathrm{div}^k; \Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m) = \overline{C_c^\infty(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)}^{\|\cdot\|_{W^q(\mathrm{div}^k)}}.$$

3 Regularization theory

The main purpose of this section is give a quick overview of well known regularization functionals and to introduce the total generalized variation (TGV) functional in a general form, which will be used as regularization term throughout this work. After giving the said overview, we will first define the TGV functional for scalar valued data and state essential properties for arbitrary order $k \in \mathbb{N}$. Then we will introduce its generalization to vector-valued data, show that all desirable properties also generalize, and state some additional results. Again, throughout this subsection, $\Omega \subset \mathbb{R}^d$ with $d \in \mathbb{N}$ will be a domain, if necessary further specified to be a bounded Lipschitz domain.

3.1 An overview

On an abstract level, the problem of recovering an image from compressed or subsampled data can be posed as inversion of a forward operator F, i.e., given some data d, to find u such that $F(u) = d$. This problem of inverse type appears in many areas of applied mathematics. Typically the problem is ill posed, meaning that either no direct inversion is possible or solutions do not depend continuously on the given data. To overcome this, a standard approach is to introduce a regularization function R and perform the inversion by solving

$$\min_u R(u) + G(u,d),$$

where G enforces $F(u) = d$. Crucial for the quality of a resulting reconstruction is the choice of the regularization term, that reflects expected structure of solutions. If one is interested in reconstructing *natural* images from corrupted data, as it is usually the case in image processing, the term R should thus reflect typical properties of such images. From the computational point of view also certain analytical properties, such as convexity, continuity or even differentiability would be helpful.

The task of finding such a regularization term is an active research topic in mathematical image processing, of which we give a brief overview in the following lines. In contrast to a generic test setting such as the denoising problem, i.e. $G(u,d) = \|u-d\|_{L^2}^2$, the application we have in mind requires G to be non-smooth, more specifically to be the indicator function of a convex set, i.e.

$$G(u,d) = \mathcal{I}_{D(d)}(u) = \begin{cases} 0 & \text{if } u \in D(d), \\ \infty & \text{else.} \end{cases}$$

Thus good analytical properties of the regularization term are of even greater importance for numerical solvability of the resulting minimization problem.

As a result, in the subsequent discussion we focus on regularization terms providing a good balance between analytical properties and reconstruction quality, in particular we only discuss convex functionals. We will also assume at least piecewise regularity for images and thus avoid a discussion on how to handle image structures on small scales, usually referred to as texture. In this respect, we refer to [19, 74, 28] and the references therein for interesting approaches in other directions, in particular non-convex curvature based regularization terms.

A classical choice of a regularization term for image processing is the total variation functional [60], defined as the L^1 norm of the first derivative of a function. The underlying space of functions having finite TV value, or equivalently admitting a finite Radon measure as distributional derivative, allows jump discontinuities, a feature that is now widely accepted to be necessary for a realistic image model. Additionally, a simple, convex structure makes a numerical solution of TV regularized problems, without further smoothing of the objective functional, numerically feasible [24]. However, the nature of the TV functional favors piecewise constant reconstructions, a drawback that is well known as the *staircasing effect* [23, 52, 59]. This effect can also be observed in figure 2, image C, where we applied TV regularization for the solution of the JPEG decompression problem of subsection 5.2.

One possible approach to avoid this is to take the L^1 norm of higher order derivatives [58], in particular the L^1 norm of a Hessian, as regularization functional. Natural this avoids a staircasing effect in homogeneous regions, as can be observed in figure 2, image D, but imposes additional regularity assumptions on the images, in particular, jump discontinuities are no longer possible. In practice, this leads to smoothing of edges.

An alternative is to use an *infimal-convolution* based term [25], defined as

$$\begin{aligned} R(u) &= \min_{u=u_1+u_2} \left(\int_\Omega |\nabla u_1| + \int_\Omega |\nabla(\nabla u_2)| \right) \\ &= \min_{u_2} \int_\Omega |\nabla u - \nabla u_2| + \int_\Omega |\nabla(\nabla(u_2))|. \end{aligned} \tag{17}$$

This yields again a convex functional and, as can be observed in figure 2, image E, reduces staircasing effects while edges are kept sharp.

Recently, the total generalized variation functional has been introduced. Also incorporating higher order derivatives, evaluation of the TGV functional can bee seen as an optimal balancing between derivatives up to a certain order. While we refer to subsection 3.2 and the original work [16] for a rigorous definition and derivation of important analytical properties, let us point that the TGV functional is convex and, equivalently to its pre-dual definition[16], can be written as

$$\begin{aligned} R(u) &= \min_{\nabla u=v_1+v_2} \left(\int_\Omega |v_1| + \int_\Omega |\nabla v_2| \right) \\ &= \min_{v_2} \left(\int_\Omega |\nabla u - v_2| + \int_\Omega |\nabla v_2| \right). \end{aligned} \tag{18}$$

Since, in contrast to (17), not u itself but the gradient of u is decomposed for the optimal balancing, v_2 in the second line of (18) is not restricted to gradient fields. This increases the kernel of the functional and, as can be observed in figure 2, image F, leads to a good visual reconstruction quality, in particular not suffering from staircasing phenomenas and maintaining edges sharp. Motivated by that, we will use the TGV functional as regularization term in our image reconstruction setting.

3.2 The total generalized variation functional (TGV)

In this subsection we introduce the total generalized variation functional (TGV). It will serve as regularization term for the image reconstruction problem setting in this work. At first, we consider a version for scalar valued input functions as

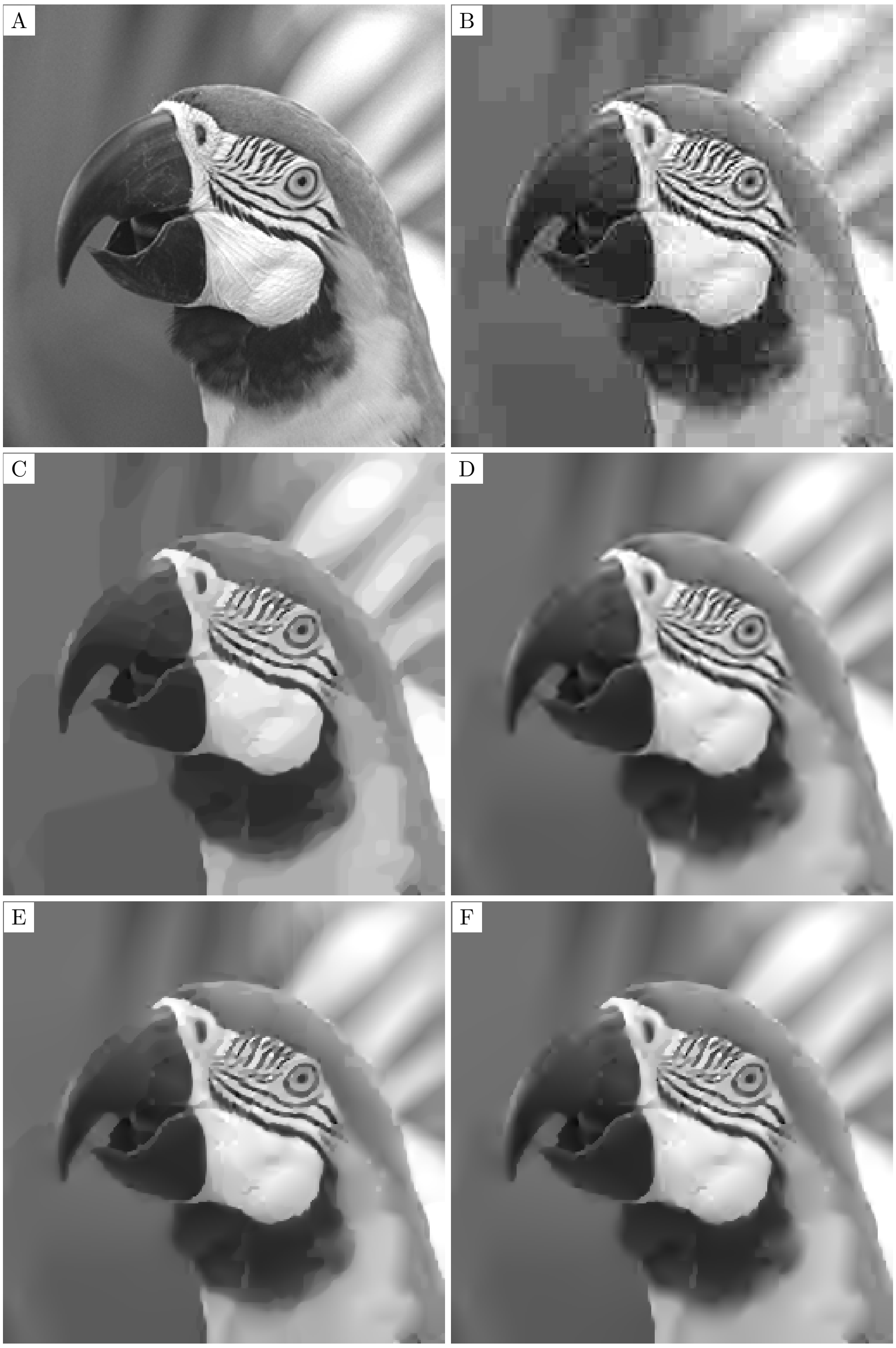


Figure 2: Test of different regularization functionals for JPEG decompression. A: Original image. B: Standard decompression. C: TV based reconstruction. D: Second-order-TV based reconstruction. E: Infimal-Convolution based reconstruction. F: TGV based reconstruction.

originally defined in [16]. We will show how some useful results, obtained in [20] for the second order TGV functional, can be extended to the case of arbitrary order k. This extension is far from being trivial and strongly relies on results presented in [10].

Afterwards, we extend the definition of the TGV functional to vector valued input functions, as done in [9], with the aim of applying this regularization to multichannel images. We also show how the results for the scalar valued case naturally transfer to this setting. At last we provide a density result for a notion of strict convergence with respect to TGV.

3.2.1 The case of scalar valued data

The *Total Generalized Variation (*TGV*)* functional of order $k \in \mathbb{N}$ is, for $u \in L^1_{\text{loc}}(\Omega)$, defined as

$$\mathrm{TGV}^{\mathrm{k}}_{\alpha}(u) = \sup \Bigg\{ \int\limits_{\Omega} u \operatorname{div}^k v \,\mathrm{d}x \,\Big|\, v \in C^k_c(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)), \qquad \| \operatorname{div}^l v\|_\infty \leq \alpha_l,\, l = 0, \dots, k-1 \Bigg\}, \tag{19}$$

and the corresponding space of function of bounded generalized variation as

$$\mathrm{BGV}^k_\alpha(\Omega) = \left\{ u \in L^1(\Omega) \,|\, \mathrm{TGV}^{\mathrm{k}}_{\alpha}(u) < \infty \right\}, \quad \|u\|_{\mathrm{BGV}^k_\alpha} = \|u\|_1 + \mathrm{TGV}^{\mathrm{k}}_{\alpha}(u). \tag{20}$$

At first we summarize some basic properties of the $\mathrm{TGV}^{\mathrm{k}}_{\alpha}$ functional that can be found in [16].

Proposition 3.1. *For any $k \in \mathbb{N}$, $\alpha \in \mathbb{R}^k_{>0}$, the $\mathrm{TGV}^{\mathrm{k}}_{\alpha}$ functional enjoys the following properties*

1. $\mathrm{TGV}^{\mathrm{k}}_{\alpha}$ *is a semi-norm on the normed space $\mathrm{BGV}^k_\alpha(\Omega)$,*
2. $\mathrm{TGV}^{\mathrm{k}}_{\alpha}$ *and $\mathrm{TGV}^k_{\tilde{\alpha}}$ are equivalent for $\tilde{\alpha} \in \mathbb{R}^k_{>0}$,*
3. $\mathrm{BGV}^{\mathrm{k}}{}_\alpha(\Omega)$ *is a Banach space,*
4. $\mathrm{TGV}^{\mathrm{k}}_{\alpha}$ *is proper, convex, lower semi-continuous on each $L^p(\Omega)$, $1 \leq p \leq \infty$,*
5. *For $u \in L^1_{\text{loc}}(\Omega)$, $\mathrm{TGV}^{\mathrm{k}}_{\alpha}(u) = 0$ if and only if u is a polynomial of degree less that k,*
6. $\mathrm{TGV}^{\mathrm{k}}_{\alpha}$ *is rotationally invariant.*
7. $\mathrm{TGV}^{\mathrm{k}}_{\alpha}(u) \leq c\,\mathrm{TV}(u)$ *for any $u \in L^1_{\text{loc}}(\Omega)$.*

Equivalence of the TGV functional for different parameters implies also equivalence of the corresponding spaces of bounded generalized variation, thus we will in the following omit the parameter α and simply write $\mathrm{BGV}^{\mathrm{k}}(\Omega)$. Further we will from now on omit the introduction of $\alpha \in \mathbb{R}^k_{>0}$ as parameter for the $\mathrm{TGV}^{\mathrm{k}}_{\alpha}$ functional.

In [20], some useful properties of the TGV^2_α functional were derived:

Proposition 3.2. *Let Ω be a bounded Lipschitz domain. The* TGV_α^2 *functional enjoys the following properties*

1. $\mathrm{TGV}_\alpha^2(u) = \min_{u \in \mathrm{BD}(\Omega)} \alpha_1 \|\mathrm{D}u - w\|_{\mathcal{M}} + \alpha_0 \|\mathcal{E}(w)\|_{\mathcal{M}}$, *for all* $u \in L^1(\Omega)$,
2. $c\|u\|_{\mathrm{BV}} \leq \|u\|_1 + \mathrm{TGV}_\alpha^2(u) \leq C\|u\|_{\mathrm{BV}}$, *for all* $u \in L^1(\Omega)$,
3. $\|u\|_p \leq \tilde{C}\, \mathrm{TGV}_\alpha^2(u)$ *for all* $u \in \ker P_1 \subset L^p(\Omega)$, $1 \leq p \leq \frac{d}{d-1}$,

for $c, C, \tilde{C} \in \mathbb{R}_{>0}$, *where* $P_1 : L^p(\Omega) \to \mathcal{P}_1(\Omega)$ *a linear, continuous, onto projection to the space of affine functions* $\mathcal{P}_1(\Omega)$.

The above results facilitate and motivate the usage of the TGV_α^2 functional as regularization term for inverse problems: In particular convexity and lower semi continuity, together with the Poincaré type inequality, allow a convex problem setting and, provided the data term possesses suitable coercivity, to obtain existence of a solution.

The norm equivalence of $\mathrm{BGV}^2(\Omega)$ to $\mathrm{BV}(\Omega)$ allows the study of $\mathrm{BGV}^2(\Omega)$ functions in a well known function space setting, e.g. continuous and compact embedding results transfer immediately from $\mathrm{BV}(\Omega)$ to $\mathrm{BGV}^2(\Omega)$. Further, rotational invariance and the minimum representation of TGV_α^2 as in the first point of proposition 3.2 support the usage of TGV_α^2 as regularization parameter for imaging problems.

Motivated by that, our aim is now to extend the results of proposition 3.2 to the $\mathrm{TGV}_\alpha^\mathrm{k}$ functional of arbitrary order $k \in \mathbb{N}$. The techniques to obtain this are also presented in [11].

Minimum representation At first, we extend the minimum representation of proposition 3.2. To this aim, we define a generalization of the $\mathrm{TGV}_\alpha^\mathrm{k}$ functional for symmetric tensor fields:

Definition 3.1. *For* $l \in \mathbb{N}_0, k \in \mathbb{N}$ *we define, for* $v \in L^1_{\mathrm{loc}}(\Omega, \mathrm{Sym}^l(\mathbb{R}^d))$,

$$\mathrm{TGV}_\alpha^{k,l}(v) = \sup \left\{ \int_\Omega v \operatorname{div}^k v \, \mathrm{d}x \,\middle|\, v \in C_c^k(\Omega, \mathrm{Sym}^{l+k}(\mathbb{R}^d)), \right.$$
$$\left. \|\operatorname{div}^i v\|_\infty \leq \alpha_i,\ i = 0, \ldots, k-1 \right\}. \quad (21)$$

The resulting space of functions of bounded generalized variation is then defined is

$$\mathrm{BGV}^k(\Omega, \mathrm{Sym}^l(\mathbb{R}^d)) = \{u \in L^1(\Omega, \mathrm{Sym}^l(\mathbb{R}^d)) \mid \mathrm{TGV}_\alpha^{k,l}(u) < \infty\},$$
$$\|u\|_{\mathrm{BGV}^k} = \|u\|_1 + \mathrm{TGV}_\alpha^{k,l}(u).$$

Note that, similar as in [16, Proposition 3.3], it can easily be seen that, for fixed $l \in \mathbb{N}_0$, $\mathrm{TGV}_\alpha^{k,l}$ and $\mathrm{TGV}_{\tilde{\alpha}}^{k,l}$ are equivalent for any $\alpha, \tilde{\alpha} \in \mathbb{R}^k_{>0}$. Thus indeed the space $\mathrm{BGV}^k(\Omega, \mathrm{Sym}^l(\mathbb{R}^d))$ can be written independently of α. We further need the following basic result on $\mathrm{BGV}^k(\Omega, \mathrm{Sym}^l(\mathbb{R}^d))$.

Proposition 3.3. *For* $k \in \mathbb{N}$, $l \in \mathbb{N}_0$ *it holds that*

1. $\operatorname{TGV}_\alpha^{k,l}$ *is proper, convex and lower semi-continuous in* $L^1(\Omega, \operatorname{Sym}^l(\mathbb{R}^d))$,
2. $\operatorname{TGV}_\alpha^{k,l}$ *is a continuous semi norm on* $\operatorname{BGV}^k(\Omega, \operatorname{Sym}^l(\mathbb{R}^d))$ *with finite-dimensional kernel* $\ker(\mathcal{E}^k)$,
3. *the space* $\operatorname{BGV}^k(\Omega, \operatorname{Sym}^l(\mathbb{R}^d))$ *is a Banach space.*

Proof. Obviously, $\operatorname{TGV}_\alpha^{k,l}$ is proper. Now note that, for each $v \in \mathcal{C}_c^k(\Omega, \operatorname{Sym}^{k+l}(\mathbb{R}^d))$, the map $u \mapsto \int_\Omega u \cdot \operatorname{div}^k v \, \mathrm{d}x$ is a continuous, affine functional on $L^1(\Omega, \operatorname{Sym}^l(\mathbb{R}^d))$. Since $\operatorname{TGV}_\alpha^{k,l}$ is just the pointwise supremum of a family of such functionals, it follows by [37, Proposition I.3.1] that it is convex and lower semi-continuous in $L^1(\Omega, \operatorname{Sym}^l(\mathbb{R}^d))$.

Positive homogeneity of $\operatorname{TGV}_\alpha^{k,l}$ is immediate, hence $\operatorname{TGV}_\alpha^{k,l}$ is a semi-norm and thus $\operatorname{BGV}^k(\Omega, \operatorname{Sym}^l(\mathbb{R}^d))$ a normed space. Since

$$|\operatorname{TGV}_\alpha^{k,l}(u^1) - \operatorname{TGV}_\alpha^{k,l}(u^2)| \leq \operatorname{TGV}_\alpha^{k,l}(u^1 - u^2) \leq \|u^1 - u^2\|_{\mathrm{BGV}}$$

for each $u^1, u^2 \in \operatorname{BGV}^k(\Omega, \operatorname{Sym}^l(\mathbb{R}^d))$, it is also continuous. Finally, for each $v \in \mathcal{C}_c^k \Omega, \operatorname{Sym}^{k+l}(\mathbb{R}^d)$, one can find a $\lambda > 0$ such that $\|\operatorname{div}^i(\lambda v)\|_\infty \leq \alpha_i$ for $i = 0, \ldots, k-1$. Hence, $\operatorname{TGV}_\alpha^{k,l}(u) = 0$ if and only if

$$\int_\Omega u \cdot \operatorname{div}^k v \, \mathrm{d}x = 0 \qquad \text{for each } v \in \mathcal{C}_c^k(\Omega, \operatorname{Sym}^{k+l}(\mathbb{R}^d))$$

which is equivalent to $u \in \ker \mathcal{E}^k$ in the weak sense. As $\ker \mathcal{E}$ considered on $\mathcal{D}(\Omega, \operatorname{Sym}^{l+i}(\mathbb{R}^d))$ is finite-dimensional for each $i = 0, \ldots, k-1$, see proposition 2.1, the latter has to be finite-dimensional.

By standard arguments it can finally be concluded from lower semi continuity of the semi-norm $\operatorname{TGV}_\alpha^{k,l}$ in $L^1(\Omega, \operatorname{Sym}^l(\mathbb{R}^d))$ that $\operatorname{BGV}^k(\Omega, \operatorname{Sym}^l(\mathbb{R}^d))$ is a Banach space when equipped with $\|\cdot\|_{\mathrm{BGV}} = \|\cdot\|_{L^1} + \operatorname{TGV}_\alpha^{k,l}(\cdot)$. □

Also the following assertion will be helpful:

Lemma 3.1. *Let* $k \in \mathbb{N}$ *and* $u_{k-1} \in \mathcal{C}_0^{k-1}(\Omega, \operatorname{Sym}^{k-1}(\mathbb{R}^d))^*$, $u_k \in \mathcal{C}_0^k(\Omega, \operatorname{Sym}^k(\mathbb{R}^d))^*$ *be distributions of order* $k-1$ *and* k, *respectively. Then*

$$\|\mathcal{E}u_{k-1} - u_k\|_{\mathcal{M}} = \sup \{\langle u_{k-1}, \operatorname{div} v_k\rangle + \langle u_k, v_k\rangle \mid v_k \in \mathcal{C}_c^k(\Omega, \operatorname{Sym}^k(\mathbb{R}^d)),\ \|v_k\|_\infty \leq 1\} \tag{22}$$

with the right-hand side being finite if and only if $\mathcal{E}u_{k-1} - u_k \in \mathcal{M}(\Omega, \operatorname{Sym}^k(\mathbb{R}^d))$ *in the distributional sense.*

Proof. Note that in the distributional sense, $\langle u_k - \mathcal{E}u_{k-1}, v_k\rangle = \langle u_{k-1} \operatorname{div} v_k\rangle + \langle u_k, v_k\rangle$ for all $v_k \in \mathcal{C}_c^\infty(\Omega, \operatorname{Sym}^k(\mathbb{R}^d))$. Since $\mathcal{C}_c^\infty(\Omega, \operatorname{Sym}^k(\mathbb{R}^d))$ is dense in $\mathcal{C}_0(\Omega, \operatorname{Sym}^k(\mathbb{R}^d))$, the distribution $u_k - \mathcal{E}u_{k-1}$ can be extended to an element in $\mathcal{C}_0(\Omega, \operatorname{Sym}^k(\mathbb{R}^d))^* = \mathcal{M}(\Omega, \operatorname{Sym}^k(\mathbb{R}^d))$ if and only if the supremum in (22) is finite. In case of finiteness, it coincides with the Radon norm by definition. □

With that, similar to assertion 1 in proposition 3.2, we can now derive a minimum representation of $\operatorname{TGV}_\alpha^{k,l}$:

Theorem 3.1. *Let* Ω *be a bounded Lipschitz domain. For* $\operatorname{TGV}_\alpha^{k,l}$ *defined according to Definition 3.1,* $k \in \mathbb{N}, l \in \mathbb{N}_0$, *we have, for each* $u \in L^1_{\mathrm{loc}}(\Omega, \operatorname{Sym}^l(\mathbb{R}^d))$,

$$\operatorname{TGV}_\alpha^{k,l}(u) = \min_{\substack{w_i \in \operatorname{BD}(\Omega, \operatorname{Sym}^{l+i}(\mathbb{R}^d)),\\ i=0,\ldots,k,\\ w_0 = u,\ w_k = 0}} \sum_{l=1}^k \alpha_{k-l} \|\mathcal{E}w_{l-1} - w_l\|_{\mathcal{M}} \tag{23}$$

with the minimum being finite if and only if $u \in \mathrm{BD}(\Omega, \mathrm{Sym}^l(\mathbb{R}^d))$ *and attained for some* $(w_0, \ldots, w_k)$ *with* $w_i \in \mathrm{BD}(\Omega, \mathrm{Sym}^{l+i}(\mathbb{R}^d))$, $1 \le i < k$, $w_0 = u$ *and* $w_k = 0$ *in case of* $u \in \mathrm{BD}(\Omega, \mathrm{Sym}^l(\mathbb{R}^d))$.

Proof. First, take $u \in L^1_{\mathrm{loc}}(\Omega, \mathrm{Sym}^l(\mathbb{R}^d))$ such that $\mathrm{TGV}_\alpha^{k,l}(u) < \infty$. We will employ Fenchel-Rockafellar duality. For this purpose, we introduce the Banach spaces

$$\begin{aligned} X &= \mathcal{C}_0^1(\Omega, \mathrm{Sym}^{1+l}(\mathbb{R}^d)) \times \ldots \times \mathcal{C}_0^k(\Omega, \mathrm{Sym}^{k+l}(\mathbb{R}^d)), \\ Y &= \mathcal{C}_1^1(\Omega, \mathrm{Sym}^{1+l}(\mathbb{R}^d)) \times \ldots \times \mathcal{C}_0^{k-1}(\Omega, \mathrm{Sym}^{k-1+l}(\mathbb{R}^d)), \end{aligned}$$

the linear operator

$$\Lambda \in \mathcal{L}(X, Y), \qquad \Lambda v = \begin{pmatrix} -v_1 - \operatorname{div} v_2 \\ \cdots \\ -v_{k-1} - \operatorname{div} v_k \end{pmatrix},$$

and the proper, convex and lower semi-continuous functionals

$$\begin{aligned} F &: X \to \,]-\infty, \infty], & F(v) &= \sum_{l=1}^{k} I_{\{\|\cdot\|_\infty \le \alpha_{k-l}\}}(v_l) - \int u \cdot \operatorname{div} v_1, \\ G &: Y \to \,]-\infty, \infty], & G(w) &= I_{\{(0,\ldots,0)\}}(w) \end{aligned}$$

such that

$$\mathrm{TGV}_\alpha^{k,l}(u) = \sup_{v \in X} -F(v) - G(\Lambda v).$$

In order to show the representation of $\mathrm{TGV}_\alpha^{k,l}(u)$ as in (23) we want to obtain

$$\mathrm{TGV}_\alpha^{k,l}(u) = \min_{w \in Y^*} F^*(-\Lambda^* w) + G^*(w). \tag{24}$$

This follows from [4, Corollary 2.3] provided we can show that

$$Y = \bigcup_{\lambda > 0} \lambda(\operatorname{dom} G + \Lambda \operatorname{dom} F).$$

For that purpose, choose $w = (w_1, \ldots, w_{k-1}) \in Y$ arbitrary. Setting $v_k = 0 \in \mathcal{C}_0^k(\Omega, \mathrm{Sym}^{l+k}(\mathbb{R}^d))$ and $v_i = w_i - \operatorname{div} v_{i+1} \in \mathcal{C}_0^i(\Omega, \mathrm{Sym}^{l+i}(\mathbb{R}^d))$ for $i = k-1, \ldots, 1$, we get that $v = (v_1, \ldots, v_k) \in X$ and $-\Lambda v = w$. Choosing $\lambda > 0$ large enough such that $\|\lambda^{-1} v_i\| \le \alpha_{k-i}$, $i = 1, \ldots, k$, it follows that $\lambda^{-1} v \in \mathrm{dom}(F)$ and, since $0 \in \mathrm{dom}(G)$ we can write w as $w = \lambda(0 - \Lambda \lambda^{-1} v)$.

Thus, Equation (24) is satisfied and the minimum is obtained in Y^*. Now Y^* can be written as

$$Y^* = \mathcal{C}_0^1(\Omega, \mathrm{Sym}^{1+l}(\mathbb{R}^d))^* \times \ldots \times \mathcal{C}_0^{k-1}(\Omega, \mathrm{Sym}^{k-1+l}(\mathbb{R}^d))^*,$$

with elements $w = (w_1, \ldots, w_{k-1})$, $w_i \in \mathcal{C}_0^i(\Omega, \mathrm{Sym}^{i+l}(\mathbb{R}^d))^*$, $1 \le i \le k-1$.

Therefore, with $w_0 = u$ and $w_k = 0$, we get

$$\begin{aligned} F^*(-\Lambda w) + G^*(w) &= \sup_{v \in X} \Big(\langle -\Lambda^* w, v\rangle - \sum_{l=1}^{k} I_{\{\|\cdot\|_\infty \leq \alpha_{k-l}\}}(v_l) + \langle u, \operatorname{div} v_1\rangle \Big) \\ &= \sup_{\substack{v \in X, \\ \|v_l\|_\infty \leq \alpha_{k-l}, \\ l=1,\ldots,k}} \Big(\sum_{i=1}^{k-1} \langle w_i, \operatorname{div} v_{i+1}\rangle + \langle w_i, v_i\rangle + \langle u, \operatorname{div} v_1\rangle \Big) \\ &= \sum_{i=1}^{k} \alpha_{k-i} \Big(\sup_{\substack{v_i \in \mathcal{C}_0^i(\Omega, \operatorname{Sym}^{i+l}(\mathbb{R}^d)), \\ \|v_i\|_\infty \leq 1}} \langle w_{i-1}, \operatorname{div} v_i\rangle + \langle w_i, v_i\rangle \Big) \end{aligned}$$

From Lemma 3.1 we know that each supremum is finite and coincides with $\|\mathcal{E}w_{i-1} - w_i\|_{\mathcal{M}}$ if and only if $\mathcal{E}w_{i-1} - w_i \in \mathcal{M}(\Omega, \operatorname{Sym}^{k+i}(\mathbb{R}^d))$ for $i = 1, \ldots, k$. As $w_k = 0$, according to proposition 2.2, finiteness of the minimum already yields $w_{k-1} \in \operatorname{BD}(\Omega, \operatorname{Sym}^{k+l-1}(\mathbb{R}^d))$, in particular $w_{k-1} \in \mathcal{M}(\Omega, \operatorname{Sym}^{k+l-1}(\mathbb{R}^d))$. Proceeding inductively, we see that $w_i \in \operatorname{BD}(\Omega, \operatorname{Sym}^{k+i}(\mathbb{R}^d))$ for each $i = 1, \ldots, k$. Hence, it suffices to take the minimum in (24) over all BD-tensor fields which gives (23).

In addition, the minimum in (23) is finite if $u \in \operatorname{BD}(\Omega, \operatorname{Sym}^l(\mathbb{R}^d))$. Conversely, if $\operatorname{TD}(u) = \infty$, also $\|\mathcal{E}u - w_1\|_{\mathcal{M}} = \infty$ for all $w_1 \in \operatorname{BD}(\Omega, \operatorname{Sym}^{l+1}(\mathbb{R}^d))$. Hence, the minimum in (23) has to be ∞. □

Remark 3.1. *An important step in the proof of proposition 3.1 relies on the fact that, given any $u \in \mathcal{D}(\Omega, \operatorname{Sym}^k(\mathbb{R}^d))$, we have*

$$\mathcal{E}u \in \mathcal{M}(\Omega, \operatorname{Sym}^{k+1}(\mathbb{R}^d)) \quad \Rightarrow \quad u \in \operatorname{BD}(\Omega, \operatorname{Sym}^k(\mathbb{R}^d)).$$

As already mentioned this is a non-trivial generalization of a classical result for $k = 1$ (see e.g. [65, Theorem II.2.3]) and has originally been shown in [11].

In particular, proposition 3.1 now implies the following:

Remark 3.2. *Let Ω be a bounded Lipschitz domain. For all $k \in \mathbb{N}$ and all $u \in L^1_{\mathrm{loc}}(\Omega)$ it holds that*

$$\operatorname{TGV}^k_\alpha(u) = \min_{\substack{w_i \in \operatorname{BD}(\Omega, \operatorname{Sym}^i(\mathbb{R}^d)), \\ i=0,\ldots,k, \\ w_0 = u, \ w_k = 0}} \sum_{l=1}^{k} \alpha_{k-l} \|\mathcal{E}w_{l-1} - w_l\|_{\mathcal{M}} \tag{25}$$

Topological equivalence Our next aim is to get a topological equivalence result for $\operatorname{BGV}^k(\Omega, \operatorname{Sym}^l(\mathbb{R}^d))$ similar as in assertion 2 of proposition 3.2. For that, we need the following preparatory lemma:

Lemma 3.2. *Let Ω be a bounded Lipschitz domain, $k \in \mathbb{N}, l \in \mathbb{N}_0$. There exists a constant $C_1 > 0$, only depending on Ω, k and l, such that, for each $v \in \operatorname{BD}(\Omega, \operatorname{Sym}^l(\mathbb{R}^d))$ and $\overline{w} \in \ker(\operatorname{TGV}^{k,l+1}_\alpha) \subset L^1(\Omega, \operatorname{Sym}^{l+1}(\mathbb{R}^d))$,*

$$\|\mathcal{E}(v)\|_{\mathcal{M}} \leq C_1 \|\mathcal{E}(v) - \overline{w}\|_{\mathcal{M}} + \|v\|_1.$$

Proof. If this is not true, then there exist $(v_n)_{n\in\mathbb{N}}$ and $(\overline{w}_n)_{n\in\mathbb{N}}$ with $v_n \in \mathrm{BD}(\Omega, \mathrm{Sym}^l(\mathbb{R}^d))$ and $\overline{w}_n \in \ker(\mathrm{TGV}_\alpha^{k,l+1})$ such that

$$\|\mathcal{E}(v_n)\|_{\mathcal{M}} = 1 \quad \text{and} \quad \frac{1}{n} \geq \|v_n\|_1 + \|\mathcal{E}(v_n) - \overline{w}_n\|_{\mathcal{M}}.$$

This implies that $(\overline{w}_n)_{n\in\mathbb{N}}$ is bounded in terms of $\|\cdot\|_{\mathcal{M}}$ in the finite dimensional space $\ker \mathrm{TGV}_\alpha^{k,l+1} = \ker \mathcal{E}^k$ (see Proposition 3.3). Consequently, there exists a subsequence, again denoted by $(\overline{w}_n)_{n\in\mathbb{N}}$ and $\overline{w} \in \ker(\mathrm{TGV}_\alpha^{k,l+1})$ such that $\overline{w}_n \to \overline{w}$ with respect to $\|\cdot\|_{L^1}$. Hence, $\mathcal{E}(v_n) \to \overline{w}$. Further we have that $v_n \to 0$ and thus, by closedness of the weak symmetrized gradient, $\mathcal{E}(v_n) \to 0$, which contradicts to $\|\mathcal{E}(v)\|_{\mathcal{M}} = 1$. □

With that, we can show the main result to obtain topological equivalence:

Proposition 3.4. *Let Ω be a bounded Lipschitz domain, $k \in \mathbb{N}, l \in \mathbb{N}_0$. and*

$$R_{l,k} : L^{d/(d-1)}(\Omega, \mathrm{Sym}^l(\mathbb{R}^d)) \to \ker \mathcal{E}^k$$

be a linear, continuous and onto projection. Then, there exists a constant $C > 0$, only depending on k, l, α, Ω and R such that

$$\|\mathcal{E}u\|_{\mathcal{M}} \leq C\big(\|u\|_1 + \mathrm{TGV}_\alpha^{k,l}(u)\big) \qquad \text{as well as} \qquad \|u - Ru\|_{d/(d-1)} \leq C\, \mathrm{TGV}_\alpha^{k,l}(u) \tag{26}$$

for all $u \in L^{d/(d-1)}(\Omega, \mathrm{Sym}^l(\mathbb{R}^d))$.

Proof. We prove the result by induction. In the case $k = 1$, the first inequality is immediate while the second one is equivalent to the Sobolev-Korn inequality in $\mathrm{BD}(\Omega, \mathrm{Sym}^l(\mathbb{R}^d))$ (see proposition 2.1).

Now assume both inequalities hold for a fixed k and perform an induction step with respect to k, i.e., we fix $l \in \mathbb{N}_0$, $(\alpha_0, \dots, \alpha_k)$ with $\alpha_i > 0$, Ω and $R_{l,k+1}$ a linear, continuous onto projection from $L^{d/(d-1)}(\Omega, \mathrm{Sym}^l(\mathbb{R}^d))$ to $\ker \mathcal{E}^{k+1}$ for which we need to show that (26) holds.

Further we assume that assertion (26) holds for $(\alpha_0, \dots, \alpha_{k-1})$, Ω, as well as any $l' \in \mathbb{N}$ and $R_{l',k}$ projecting on $\ker \mathcal{E}^k$.

We will first show the uniform estimate for $\|\mathcal{E}u\|_{\mathcal{M}}$ for which it suffices to consider $u \in \mathrm{BD}(\Omega, \mathrm{Sym}^l(\mathbb{R}^d))$, as otherwise, according to Theorem 3.1, $\mathrm{TGV}_\alpha^{k+1,l}(u) = \infty$. Hence, choose $u \in \mathrm{BD}(\Omega, \mathrm{Sym}^l(\mathbb{R}^d))$ and denote by $R_{l+1,k} : L^{d/(d-1)}(\Omega, \mathrm{Sym}^{l+1}(\mathbb{R}^d)) \to \ker \mathcal{E}^k$ a linear, continuous and onto projection. Recall that such a projection always exists as $\ker \mathcal{E}^k$ is finite-dimensional. Then, with the help of Lemma 3.2, the continuous embeddings

$$\mathrm{BD}(\Omega, \mathrm{Sym}^{l+1}(\mathbb{R}^d)) \hookrightarrow L^{d/(d-1)}(\Omega, \mathrm{Sym}^{l+1}(\mathbb{R}^d)) \hookrightarrow L^1(\Omega, \mathrm{Sym}^{l+1}(\mathbb{R}^d))$$

and the induction hypothesis, we can estimate for arbitrary $w \in \mathrm{BD}(\Omega, \mathrm{Sym}^{l+1}(\mathbb{R}^d))$,

$$\begin{aligned}
\|\mathcal{E}u\|_{\mathcal{M}} &\leq C_1(\|\mathcal{E}u - R_{l+1,k}w\|_{\mathcal{M}} + \|u\|_1) \\
&\leq C_1(\|\mathcal{E}u - w\|_{\mathcal{M}} + \|w - R_{l+1,k}w\|_{d/(d-1)} + \|u\|_1) \\
&\leq C_2(\|\mathcal{E}u - w\|_{\mathcal{M}} + \mathrm{TGV}_\alpha^{k,l+1}(w) + \|u\|_1) \\
&\leq C_3(\alpha_k\|\mathcal{E}u - w\|_{\mathcal{M}} + \mathrm{TGV}_\alpha^{k,l+1}(w) + \|u\|_1)
\end{aligned}$$

for $C_1, C_2, C_3 > 0$ suitable. Taking the minimum over all such $w \in \mathrm{BD}(\Omega, \mathrm{Sym}^{l+1}(\mathbb{R}^d)$ then yields

$$\|\mathcal{E}u\|_{\mathcal{M}} \leq C_3\big(\|u\|_1 + \mathrm{TGV}_\alpha^{k+1,l}(u)\big)$$

by virtue of the minimum respresentation (23).

In order to show the coercivity estimate, assume that the inequality does not hold true. Then, there is a sequence $(u^n)_{n\in\mathbb{N}}$ in $L^{d/(d-1)}(\Omega, \mathrm{Sym}^l(\mathbb{R}^d))$ such that

$$\|u^n - R_{l,k+1}u^n\|_{d/(d-1)} = 1 \quad \text{and} \quad \frac{1}{n} \geq \mathrm{TGV}_\alpha^{k+1,l}(u^n).$$

Note that $\ker \mathrm{TGV}_\alpha^{k+1,l} = \ker \mathcal{E}^{k+1} = \mathrm{Rg}\, R_{l,k+1}$ (confer proposition 3.3), hence $\mathrm{TGV}_\alpha^{k+1,l}(u^n - R_{l,k+1}u^n) = \mathrm{TGV}_\alpha^{k+1,l}(u^n)$ for each n. Thus, since we already know the first estimate in (26) to hold,

$$\|\mathcal{E}(u^n - R_{l,k+1}u^n)\|_{\mathcal{M}} \leq C_3\big(\mathrm{TGV}_\alpha^{k+1,l}(u^n) + \|u^n - R_{l,k+1}u^n\|_1\big), \tag{27}$$

implying, by continuous embedding, that $(u^n - R_{l,k+1}u^n)_{n\in\mathbb{N}}$ is bounded in $\mathrm{BD}(\Omega, \mathrm{Sym}^l(\mathbb{R}^d))$. By compact embedding (see proposition 2.1) we may therefore conclude that $u^n - R_{l,k+1}u^n \to u^*$ in $L^1(\Omega, \mathrm{Sym}^l(\mathbb{R}^d))$ for some subsequence (not relabeled). Moreover, as $R_{l,k+1}(u^n - R_{l,k+1}u^n) = 0$ for all n, the limit has to satisfy $R_{l,k+1}u^* = 0$. On the other hand, by lower semi-continuity (see Proposition 3.3)

$$0 \leq \mathrm{TGV}_\alpha^{k+1,l}(u^*) \leq \liminf_{n\to\infty} \mathrm{TGV}_\alpha^{k+1,l}(u^n) = 0,$$

hence $u^* \in \ker \mathcal{E}^{k+1} = \mathrm{Rg}\, R_{l,k+1}$. Consequently, $\lim_{n\to\infty} u^n - R_{l,k+1}u^n = u^* = R_{l,k+1}u^* = 0$. From (27) it follows that also $\mathcal{E}(u^n - R_{l,k+1}u^n) \to 0$ in $\mathcal{M}(\Omega, \mathrm{Sym}^{l+1}(\mathbb{R}^d))$, so $u^n - R_{l,k+1}u^n \to 0$ in $\mathrm{BD}(\Omega, \mathrm{Sym}^l(\mathbb{R}^d))$ and by continuous embedding also in $L^{d/(d-1)}(\Omega, \mathrm{Sym}^l(\mathbb{R}^d))$. However, this contradicts $\|u^n - R_{l,k+1}u^n\|_{d/(d-1)} = 1$ for all n, and thus the claimed coercivity has to hold. □

From proposition 3.4 we immediately get the following corollary, stating, in particular, topological equivalence of $\mathrm{BGV}^k(\Omega)$ to $\mathrm{BV}(\Omega)$ when equipped with the topologies induced by $\|\cdot\|_{\mathrm{BGV}^k}$ and $\|\cdot\|_{\mathrm{BV}}$, respectively.

Corollary 3.1. *Let Ω be a bounded Lipschitz domain. For all $k \in \mathbb{N}$ there exists $\lambda > 0$ such that, for all $u \in \mathrm{BV}(\Omega)$,*

$$\mathrm{TV}(u) \leq \lambda(\|u\|_1 + \mathrm{TGV}_\alpha^k(u)).$$

In particular, there exist $C, c > 0$ such that, for all $u \in \mathrm{BV}(\Omega)$,

$$c(\|u\|_1 + \mathrm{TGV}_\alpha^k(u)) \leq \|u\|_1 + \mathrm{TV}(u) \leq C(\|u\|_1 + \mathrm{TGV}_\alpha^k(u)).$$

Also an embedding result follows immediately from the embedding of $\mathrm{BV}(\Omega)$ to $L^{d/(d-1)}(\Omega)$:

Corollary 3.2. *Let Ω be a bounded Lipschitz domain. For all $k \in \mathbb{N}$, $1 \leq p \leq \frac{d}{d-1}$ the space $\mathrm{BGV}^k(\Omega)$ is continuously embedded into $L^p(\Omega)$. If, moreover, $1 \leq p < \frac{d}{d-1}$, the embedding is compact.*

Remark 3.3. *Given that, by corollary 3.1, $\mathrm{BGV}^k \simeq \mathrm{BV}(\Omega)$, we will henceforth only use the terminology of $\mathrm{BV}(\Omega)$ to denote $L^1(\Omega)$ functions having bounded total generalized variation.*

3.2.2 The case of vector valued data

Now we generalize the definition of the TGV functional of the previous section to allow vector-valued input functions. As already mentioned, this can be motivated by the future application of this functional as regularization term in the context of multichannel image reconstruction. We will see that the results we obtained in the previous subsection transfer easily to the vector valued data case. We make frequent use of definitions and results obtained for product tensor space valued mappings as stated in subsection 2.2.

Definition 3.2. *For $m \in \mathbb{N}$, we define the vector-input* TGV *functional of order $k \in \mathbb{N}$, for $u \in L^1_{\mathrm{loc}}(\Omega, \mathbb{R}^m)$, as*

$$\mathrm{TGV}^{\mathrm{k}}_{\alpha}(u) = \sup \Bigg\{ \int_{\Omega} u \cdot \operatorname{div}^k \xi \, \mathrm{d}x \,\Big|\, \xi \in C^k_c(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m), \\ \|\operatorname{div}^l \xi\|_\infty \leq \alpha_l,\, l = 0, \dots, k-1 \Bigg\}. \quad (28)$$

Note that we abuse notation by using the same notation as for classical TGV functional.

Remark 3.4. *As already mentioned in subsection 2.2 we restrict ourselves to using only the Euclidean norm on $\mathbb{R}^m$ for the product space. Note, however, that a different norm would influence the exact value of $\mathrm{TGV}^{\mathrm{k}}_{\alpha}(u)$, but not the set of functions for which it is finite. This can be shown easily by equivalence of norms in $\mathbb{R}^m$.*

Similar to the scalar case, we define the space $\mathrm{BGV}^k(\Omega, \mathbb{R}^m)$ as the set of all $L^1(\Omega, \mathbb{R}^m)$ functions such that the total generalized variation functional is finite.

Definition 3.3. *Let $k, m \in \mathbb{N}$. We define*

$$\begin{aligned} \mathrm{BGV}^k(\Omega, \mathbb{R}^m) = & \quad \Big\{ u \in L^1(\Omega, \mathbb{R}^m) \mid \mathrm{TGV}^k_{\alpha}(u) < \infty \Big\} \\ \|u\|_{\mathrm{BGV}^k} = & \quad \|u\|_1 + \mathrm{TGV}^{\mathrm{k}}_{\alpha}(u) \end{aligned} \quad (29)$$

As will be shown in the following proposition, basic properties of the $\mathrm{TGV}^{\mathrm{k}}_{\alpha}$ functional and the space $\mathrm{BGV}^{\mathrm{k}}(\Omega)$, as derived in proposition 3.1, can easily be transfered to the vector-input $\mathrm{TGV}^{\mathrm{k}}_{\alpha}$ functional and the space $\mathrm{BGV}^k(\Omega, \mathbb{R}^m)$. In particular, equivalence of $\mathrm{TGV}^{\mathrm{k}}_{\alpha}$ and $\mathrm{TGV}^k_{\tilde{\alpha}}$ for different $\alpha, \tilde{\alpha} \in \mathbb{R}^k_{>0}$ again justifies the notion of $\mathrm{BGV}^k(\Omega, \mathbb{R}^m)$ independently of α.

Proposition 3.5. *Let $k \in \mathbb{N}$. The following statements hold:*

1. $\mathrm{TGV}^{\mathrm{k}}_{\alpha}$ *is a semi-norm on the normed space* $\mathrm{BGV}^k(\Omega, \mathbb{R}^m)$,
2. $\mathrm{TGV}^{\mathrm{k}}_{\alpha}$ *and* $\mathrm{TGV}^k_{\tilde{\alpha}}$ *are equivalent for* $\tilde{\alpha} \in \mathbb{R}^k_{>0}$,
3. $\mathrm{BGV}^k(\Omega, \mathbb{R}^m)$ *is a Banach space,*
4. $\mathrm{TGV}^{\mathrm{k}}_{\alpha}$ *is proper, convex, lower semi-continuous on each* $L^p(\Omega, \mathbb{R}^m)$, $1 \leq p \leq \infty$,

5. $\mathrm{TGV}^{\mathrm{k}}_{\alpha}(u) = 0$, *for* $u \in L^1_{\mathrm{loc}}(\Omega, \mathbb{R}^n)$, *if and only if each* u_i, $i \in \{1, \dots, m\}$, *is a polynomial of degree less that* k.

Proof. Equivalence of TGV^k_α and $\mathrm{TGV}^k_{\tilde{\alpha}}$ is again immediate. The assertions 1 and 3 follow as in the proof of proposition 3.3. Also assertion 4 can be deduced as in proposition 3.3 and from the continuous embedding of $L^p(\Omega, \mathbb{R}^m)$ into $L^1(\Omega, \mathbb{R}^m)$, $1 \leq p \leq \infty$. To obtain assertion 5, again similar as the proof of proposition 3.3, one can show that $\mathrm{TGV}^{\mathrm{k}}_{\alpha}(u) = 0$ for $u = (u_1, \dots, u_m)$ if and only if

$$\sum_{i=1}^{m} \int\limits_{\Omega} u_i \operatorname{div}^k v_i \; \mathrm{d}x = 0 \quad \text{for each } v = (v_1, \dots, v_m) \in C^k_c(\Omega, \mathrm{Sym}^{k+1}(\mathbb{R}^d)^m).$$

In particular, this is true if and only if $\nabla^k u_i = 0$ for $1 \leq i \leq m$, from which assertion 5 follows.

□

Our next aim is to generalize the minimum representation and the topological equivalence of remark 3.2 and corollary 3.1, respectively, to the vector case. Since norm equivalence can be deduced from corollary 3.1 without a separate proof, there is no need to generalize the vector-input TGV functional to symmetric k-tensor product input. We can thus show the minimum representation directly for the original functional.

Proposition 3.6. *Let* Ω *be a bounded Lipschitz domain. For* $k, m \in \mathbb{N}$ *and any* $u \in L^1(\Omega, \mathbb{R}^m)$ *we have*

$$\mathrm{TGV}^k_\alpha(u) = \min_{\substack{v_i \in \mathrm{BD}(\Omega, \mathrm{Sym}^i(\mathbb{R}^d)^m), \\ i=1\dots k, \\ v_k = 0}} \|\nabla u - v_1\|_{\mathcal{M}} + \sum_{i=2}^{k} \alpha_{k-i} \|\mathcal{E}(v_{i-1}) - v_i\|_{\mathcal{M}} \quad (30)$$

Proof. The proof is very similar to the scalar data case: Defining

$$\begin{aligned} X &= \mathcal{C}^1_0(\Omega, \mathrm{Sym}^1(\mathbb{R}^d)^m) \times \ldots \times \mathcal{C}^k_0(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m), \\ Y &= \mathcal{C}^1_0(\Omega, \mathrm{Sym}^1(\mathbb{R}^d)^m) \times \ldots \times \mathcal{C}^{k-1}_0(\Omega, \mathrm{Sym}^{k-1}(\mathbb{R}^d)^m), \end{aligned}$$

and suitable functions $F : X \to]-\infty, \infty]$ and $G : X \to]-\infty, \infty]$, and $\Lambda \in L(X, Y)$, we can again apply [4, Corollary 2.3] to obtain

$$\mathrm{TGV}^k_\alpha(u) = \min_{w^* \in Y^*} F^*(-\Lambda^* w^*) + G^*(w^*).$$

Rewriting the right term and using that also for the vector valued case (see proposition 2.4)

$$\mathcal{E}w \in \mathcal{M}(\Omega, \mathrm{Sym}^{i+1}(\mathbb{R}^d)^m) \Rightarrow w \in \mathrm{BD}(\Omega, \mathrm{Sym}^i(\mathbb{R}^d)^m), 0 \leq i < k,$$

the assertion follows.

□

The generalization of the topological equivalence follows basically by equivalence of norms.

Proposition 3.7. *Let Ω be a bounded Lipschitz domain. For $k, m \in \mathbb{N}$, let $R : L^{d/(d-1)}(\Omega, \mathbb{R}^m) \to \ker \mathcal{E} \subset L^{d/(d-1)}(\Omega, \mathbb{R}^m)$ be a linear, continuous and onto projection. Then, there exists a constant $C > 0$, only depending on k, m, α, Ω and R such that*

$$\|\mathcal{E}u\|_{\mathcal{M}} \leq C\big(\|u\|_1 + \mathrm{TGV}_\alpha^k(u)\big) \qquad \textit{as well as} \qquad \|u - Ru\|_{d/(d-1)} \leq C\,\mathrm{TGV}_\alpha^k(u) \tag{31}$$

for all $u \in L^{d/(d-1)}(\Omega, \mathbb{R}^m)$.

Proof. Given that there exist $\tilde{c}, \tilde{C} > 0$ such that, for any $u = (u_1, \ldots, u_m) \in L^1(\Omega, \mathbb{R}^m)$,

$$c \sum_{i=1}^{m} \mathrm{TGV}_\alpha^{\mathrm{k}}(u_i) \leq \mathrm{TGV}_\alpha^{\mathrm{k}}(u) \leq C \sum_{i=1}^{m} \mathrm{TGV}_\alpha^{\mathrm{k}}(u_i), \tag{32}$$

as has been shown in [9, Proposition 2], the result follows by equivalence of norms in $\mathbb{R}^m$. □

Again norm equivalence and an embedding result follow trivially.

Corollary 3.3. *Let Ω be a bounded Lipschitz domain. For all $k, m \in \mathbb{N}$ there exists a $\lambda > 0$ such that, for all $u \in \mathrm{BV}(\Omega, \mathbb{R}^m)$,*

$$\mathrm{TV}(u) \leq \lambda(\|u\|_1 + \mathrm{TGV}_\alpha^k(u)).$$

In particular, there exist $C, c > 0$ such that, for all $u \in \mathrm{BV}(\Omega, \mathbb{R}^m)$,

$$c(\|u\|_1 + \mathrm{TGV}_\alpha^k(u)) \leq \|u\|_1 + \mathrm{TV}(u) \leq C(\|u\|_1 + \mathrm{TGV}_\alpha^k(u)).$$

Corollary 3.4. *Let Ω be a bounded Lipschitz domain. For all $k, m \in \mathbb{N}$, $1 \leq p \leq \frac{d}{d-1}$ the space $\mathrm{BGV}^k(\Omega, \mathbb{R}^m)$ is continuously embedded into $L^p(\Omega, \mathbb{R}^m)$. If, moreover, $1 \leq p < \frac{d}{d-1}$, the embedding is compact.*

Thus, since by corollary 3.3 $\mathrm{BV}(\Omega, \mathbb{R}^m) \simeq \mathrm{BGV}(\Omega, \mathbb{R}^m)$ as normed vector spaces, we will in the following only use the notion $\mathrm{BV}(\Omega, \mathbb{R}^m)$.

At last in this subsection, we aim to show that any function $u \in \mathrm{BV}(\Omega, \mathbb{R}^m)$ can be approximated by a sequences of function in $C^\infty(\Omega, \mathbb{R}^m)$ with respect to $\mathrm{TGV}_\alpha^{\mathrm{k}}$-strict convergence in $\mathrm{BV}(\Omega, \mathbb{R}^m)$, as defined in the following :

Definition 3.4. *Let $k, m \in \mathbb{N}$, $u, (u_n)_{n \in \mathbb{N}}$ be a function and a sequence of functions, respectively, in $\mathrm{BV}(\Omega, \mathbb{R}^m)$. We say that $(u_n)_{n \in \mathbb{N}}$ converges to u strictly with respect to $\mathrm{TGV}_\alpha^{\mathrm{k}}$, or equivalently converges $\mathrm{TGV}_\alpha^{\mathrm{k}}$-strictly to u, if and only if*

$$\|u_n - u\|_{L^1} \to 0 \textit{ and } \mathrm{TGV}_\alpha^{\mathrm{k}}(u_n) \to \mathrm{TGV}_\alpha^{\mathrm{k}}(u) \textit{ as } n \to \infty.$$

Note that this is a straightforward generalization of the well known concept of strict convergence in $\mathrm{BV}(\Omega)$. Many results about functions of bounded variation rely on a smooth approximation of any $u \in \mathrm{BV}(\Omega, \mathbb{R}^m)$ with respect to TV-strict convergence. Thus a similar density result for $\mathrm{TGV}_\alpha^{\mathrm{k}}$-strict convergence will be very useful.

Given that the norm topology induced by the $\|\cdot\|_{\mathrm{BGV}^k}$ is equivalent to the one induced by $\|\cdot\|_{\mathrm{BV}}$, one may hope that this is also the case for a topology induced by strict convergence. This would make an additional density proof for $\mathrm{TGV}_{\alpha}^{\mathrm{k}}$-strict convergence superfluous. The following examples, however, show that such an equivalence does not hold. Even worse, neither does TV-strict convergence imply $\mathrm{TGV}_{\alpha}^{\mathrm{k}}$- strict convergence nor does $\mathrm{TGV}_{\alpha}^{\mathrm{k}}$-strict convergence imply TV-strict convergence.

In order to characterize the subdifferential of the L^1 norm in the following example, we first need to define a set valued sign operator [17, Definition 3.4].

Definition 3.5. *Let $\mu \in \mathcal{M}(\Omega)$. Then we define*

$$\begin{aligned}\mathrm{Sgn}(\mu) := \{v \in L^\infty(\Omega) \cap L^\infty(\Omega; |\mu|) \mid \|v\|_\infty \leq 1,\ \|v\|_{\infty,|\mu|} \leq 1, \\ v = \sigma_\mu, |\mu| - \textit{almost everywhere}\}\end{aligned}$$

where σ_μ is the unique density function of μ with respect to $|\mu|$ and $\|v\|_{\infty,|\mu|}$ denotes the $|\mu|$ essential supremum of $|v|$.

Lemma 3.3. *Set $\Omega = (0,1)$ and $\alpha = (\alpha_0, \alpha_1)$ with $\alpha_i > 0$. Then, there exists a sequence $(u_n)_{n\in\mathbb{N}}$ and u in $L^1(\Omega)$ such that $(u_n)_{n\in\mathbb{N}}$ converges strictly to u with respect to TGV_α^2, i.e.*

$$\|u_n - u\|_1 \to 0 \textit{ and } \mathrm{TGV}_\alpha^2(u_n) \to \mathrm{TGV}_\alpha^2(u) \textit{ as } n \to \infty,$$

and $(u_n)_{n\in\mathbb{N}}$ does not converge strictly to u with respect to TV*, in particular there is a $\delta > 0$ and $n_0 \in \mathbb{N}$ such that*

$$|\mathrm{TV}(u_n) - \mathrm{TV}(u)| > \delta \textit{ for all } n \geq n_0.$$

Proof. Without loss of generality we can assume $\alpha = (1, \beta)$. At first we consider the approximating sequence: For any Lipschitz continuous function $u \in \mathcal{C}^{0,1}([0,1])$ we can define

$$u_n(x) = x - \frac{i}{n} + u(\frac{i}{n}) \quad \text{if} \quad x \in \left(\frac{i}{n}, \frac{i+1}{n}\right] \cap (0,1) \tag{33}$$

See figure 3 (red line) for a visualization of u_n in the case $n = 16$ and a fixed $u \in \mathcal{C}^{0,1}([0,1])$. It follows then by pointwise convergence of $(u_n)_{n\in\mathbb{N}}$ to u and Lebesgue's dominated convergence theorem that $\|u_n - u\|_1 \to 0$ as $n \to \infty$. Further, defining $w_0(x) = 1$ the constant function with value one, the weak derivative of u_n is given by

$$u_n' = w_0 - \sum_{i=1}^{n-1} \delta_{\frac{i}{n}} \left(u(\frac{i}{n}) - u(\frac{i-1}{n}) - \frac{1}{n} \right), \tag{34}$$

where the measure $\delta_{x_0}(\lambda)$ is a delta peak at x_0 with height λ. Thus, its Radon norm can be given by

$$\begin{aligned} \|u_n'\|_{\mathcal{M}} &= 1 + \sum_{i=1}^{n-1} \left| u(\frac{i}{n}) - u(\frac{i-1}{n}) - \frac{1}{n} \right| \\ &= 1 + \sum_{i=1}^{n-1} \frac{1}{n} \left| \frac{u(\frac{i}{n}) - u(\frac{i-1}{n})}{1/n} - 1 \right| \\ &= 1 + \int_0^1 |f_n(x)| \, \mathrm{d}x \end{aligned}$$

where

$$f_n(x) = \sum_{i=1}^{n-1} \frac{1}{n}(u(\frac{i}{n}) - u(\frac{i-1}{n}))\chi_{[\frac{i}{n},\frac{i-1}{n})} - w_0.$$

Since u is Lipschitz- and thus absolutely continuous, $|f_n|$ converges to $|u' - w_0|$, pointwise almost everywhere. Further the f_n are bounded uniformly, thus again by Lebesgue's dominated convergence theorem we get

$$\mathrm{TV}(u) = \|u_n'\|_{\mathcal{M}} = 1 + \int\limits_0^1 |f_n(x)|\,\mathrm{d}x \to 1 + \int\limits_0^1 |u'(x) - w_0(x)|\,\mathrm{d}x$$

as $n \to \infty$. Since the delta peaks of u_n' as in (34) cannot be canceled out by an $w \in \mathrm{BD}((0,1))$, we further get

$$\mathrm{TGV}_\alpha^2(u_n) = \beta\|u_n' - w_0\|_{\mathcal{M}} \to \beta\|u' - w_0\|_1$$

as $n \to \infty$. Now for $k \in \mathbb{N}$ even, sufficiently large such that $\frac{1}{k} < \beta$, define $u \in L^1(\Omega)$ such that

$$u(x) = \begin{cases} 2x - \frac{2i}{k} & \text{if } x \in (\frac{2i}{k}, \frac{2i+1}{k}), \\ \frac{2i+2}{k} & \text{if } x \in (\frac{2i+1}{k}, \frac{2i+2}{k}), \end{cases} \qquad 0 \leq i < \frac{k}{2}.$$

Again we refer to figure 3 (blue line) for a visualization of u in the case $k = 4$. We want to show that, for the constant function $w_0(x) = 1$ we have

$$\mathrm{TGV}_\alpha^2(u) = \min_{w \in \mathrm{BD}(0,1)} (\|u' - w\|_1 + \beta\|w'\|_{\mathcal{M}}) = \beta\|u' - w_0\|_1.$$

According to [17, Section 4.1] this is the case if there exists a $v \in H_0^1(\Omega)$ such that

$$-v' \in \mathrm{Sgn}(u' - w_0)$$
$$v \in \beta\,\mathrm{Sgn}(w_0') = \beta\{v \mid \|v\|_\infty \leq 1\}.$$

The condition $-v' \in \mathrm{Sgn}(u' - w_0)$ is satisfied for

$$v(x) = \begin{cases} -x & \text{if } x \in (\frac{2i}{k}, \frac{2i+1}{k}), \\ x & \text{if } x \in (\frac{2i+1}{k}, \frac{2i+2}{k}), \end{cases} \qquad 0 \leq i < \frac{k}{2}.$$

Further we have that $\|v\|_\infty \leq \frac{1}{k} < \beta$, thus w_0 is a minimizer.

If follows that

$$\mathrm{TGV}_\alpha^2(u) = \beta\|u' - w_0\|_1 = \lim_{n\to\infty} \mathrm{TGV}_\alpha^2(u_n),$$

thus $(u_n)_{n\in\mathbb{N}}$ converges strictly with respect to TGV_α^2 to u. But, as an easy calculation shows, $\mathrm{TV}(u) = 1$ and thus

$$|\mathrm{TV}(u_n) - \mathrm{TV}(u)| \to \|u' - w_0\|_1 > 0 \text{ as } n \to \infty.$$

Hence, for $\delta := \|u' - w_0\|_1/2 > 0$ there exists $n_0 \in \mathbb{N}$ such that, for all $n \geq n_0$,

$$|\mathrm{TV}(u_n) - \mathrm{TV}(u)| > \delta$$

and the assertion follows. □

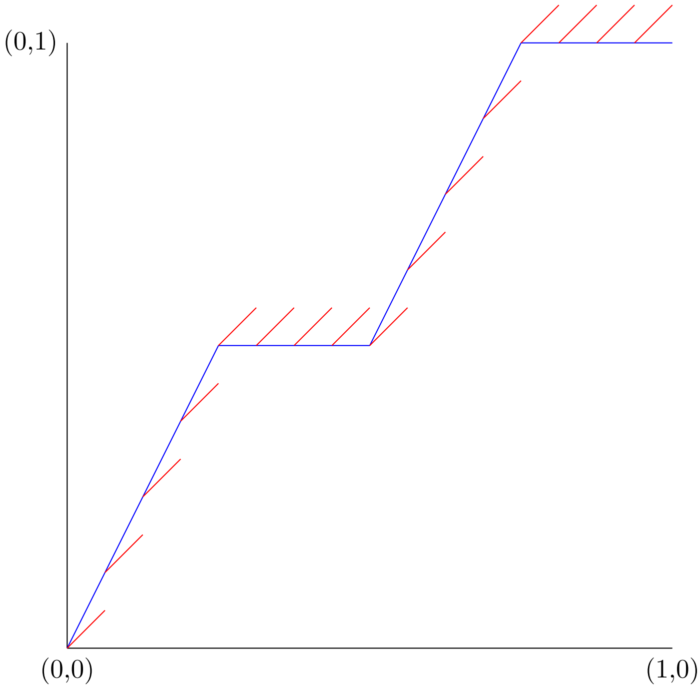


Figure 3: Illustration of the function u (blue line) for $k = 4$ and its approximation u_n (red line) for $n = 16$.

Lemma 3.4. *Set $\Omega = (0, 1)$ and $\alpha = (\alpha_0, \alpha_1)$ with $\alpha_i > 0$. Then, there exists a sequence $(u_n)_{n\in\mathbb{N}}$ and u in $L^1(\Omega)$ such that $(u_n)_{n\in\mathbb{N}}$ converges strictly to u with respect to* TV*, i.e.*

$$\|u_n - u\|_1 \to 0 \text{ and } \mathrm{TV}(u_n) \to \mathrm{TV}(u) \text{ as } n \to \infty,$$

and $(u_n)_{n\in\mathbb{N}}$ does not converge strictly to u with respect to TGV^2_α, in particular there is a $\delta > 0$ and $n_0 \in \mathbb{N}$ such that

$$\left|\mathrm{TGV}^2_\alpha(u_n) - \mathrm{TGV}^2_\alpha(u)\right| > \delta \text{ for all } n \geq n_0.$$

Proof. The construction of a counterexample is similar as in the proof of lemma 3.3: At first we again define, for any Lipschitz continuous $u \in \mathcal{C}^{0,1}((0, 1))$, an approximation sequence similar to (33) as

$$u_n(x) = u(\frac{i}{n}) \quad \text{if} \quad x \in \left(\frac{i}{n}, \frac{i+1}{n}\right] \cap (0, 1). \tag{35}$$

Clearly

$$\|u_n - u\|_1 \to 0$$

as $n \to 0$. Now as in the proof of lemma 3.3 we get that

$$u_n' = \sum_{i=1}^{n-1} \delta_{\frac{i}{n}} \left(u(\frac{i}{n}) - u(\frac{i-1}{n})\right), \tag{36}$$

and, consequently,

$$\begin{aligned}\|u_n'\|_{\mathcal{M}} &= \sum_{i=1}^{n-1} \left| u(\frac{i}{n}) - u(\frac{i-1}{n}) \right| \\ &= \sum_{i=1}^{n-1} \frac{1}{n} \left| \frac{u(\frac{i}{n}) - u(\frac{i-1}{n})}{1/n} \right| \to \int\limits_0^1 |u'(x)| \, \mathrm{d}x = \mathrm{TV}(u)\end{aligned}$$

as $n \to \infty$. Again, since delta peaks cannot be canceled out, we also get

$$\mathrm{TGV}_\alpha^2(u_n) \to \int\limits_0^1 |u'(x)| \, \mathrm{d}x$$

as $n \to \infty$. But choosing now

$$u(x) = x \text{ for } x \in (0,1)$$

we get

$$\mathrm{TGV}_\alpha^2(u) = 0 \neq \int\limits_0^1 |u'(x)| \, \mathrm{d}x,$$

implying that $\mathrm{TGV}_\alpha^2(u_n)$ does not converge to $\mathrm{TGV}_\alpha^2(u)$, hence the assertion follows. □

Thus an additional density result is necessary if we want to rely on a smooth approximation of $\mathrm{BV}(\Omega, \mathbb{R}^m)$ functions with respect to $\mathrm{TGV}_\alpha^\mathrm{k}$-strict convergence for further results. To show this, we remember the following prerequisites:

Definition 3.6. *Let $\xi \in \mathcal{M}(\mathbb{R}^d, \mathrm{Sym}^k(\mathbb{R}^d)^m)$ and $\rho \in \mathcal{C}_c^\infty(\mathbb{R}^d)$ a standard mollifier. We define the convolution of ξ and ρ_ϵ, denoted by $\xi * \rho_\epsilon$, component wise by*

$$\xi * \rho_\epsilon(x) = \int\limits_{\mathbb{R}^d} \rho_\epsilon(x-y) \, \mathrm{d}\xi(y)$$

Note that the convolution is a again a function from $\mathbb{R}^d$ to $\mathrm{Sym}^k(\mathbb{R}^d)^m$. We will need the following identity:

Lemma 3.5. *Let $k, m \in \mathbb{N}$, $\phi \in L^1_{\mathrm{loc}}(\mathbb{R}^d)$, $\xi \in \mathcal{M}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)$ and $\rho \in \mathcal{C}_c^\infty(\mathbb{R}^d)$ be a standard mollifier. Defining $\Omega_\epsilon := \{x \in \Omega | dist(x, \partial\Omega) > \epsilon\}$, we assume that both the support of ϕ and ξ are contained in Ω_ϵ (see [3, Definition 1.64] for the definition of the support of a measure). Then the following identity holds*

$$\int\limits_\Omega (\rho_\epsilon * \phi)(x) \, \mathrm{d}\xi(x) = \int\limits_\Omega (\xi * \rho_\epsilon)(x) \phi(x) \, \mathrm{d}x$$

Proof. Applying Fubini's theorem it follows

$$\begin{aligned}\int\limits_\Omega (\rho_\epsilon * \phi)(x)\,\mathrm{d}\xi(x) = & \int\limits_\Omega\int\limits_\Omega \rho_\epsilon(x-y)\phi(y)\,\mathrm{d}y\,\mathrm{d}\xi(x) \\ = & \int\limits_\Omega \phi(y) \int\limits_\Omega \rho_\epsilon(x-y)\,\mathrm{d}\xi(x)\,\mathrm{d}y \\ = & \int\limits_\Omega (\xi * \rho_\epsilon)(y)\phi(y)\,\mathrm{d}y.\end{aligned} \tag{37}$$

□

Also an estimation on the convolution of a measure will be useful:

Lemma 3.6. *Let again be $k, m \in \mathbb{N}$, $\xi \in \mathcal{M}(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)$ and $\rho \in \mathcal{C}_c^\infty(\mathbb{R}^d)$ be a standard mollifier. Further suppose that $\mathrm{supp}(\xi) \subset \Omega_\epsilon$ with Ω_ϵ as in lemma 3.5. Then*

$$\int\limits_\Omega |\xi * \rho_\epsilon|\,\mathrm{d}x \leq |\xi|(\Omega).$$

Proof. The proof can again similar to [3, Thereom 2.2] be obtained by applying Fubini's theorem. □

Now we can show a result implying the desired smooth approximation.

Proposition 3.8. *Let Ω be a bounded Lipschitz domain, $k \in \mathbb{N}$, $\alpha \in \mathbb{R}^k_{>0}$ be fixed and take $u \in L^1(\Omega, \mathbb{R}^m)$. Then $u \in \mathrm{BV}(\Omega, \mathbb{R}^m)$ if and only if there exists a sequence $(u_n)_{n\in\mathbb{N}}$ in $C^\infty(\Omega, \mathbb{R}^m)$ such that $(u_n)_{n\in\mathbb{N}}$ converges strictly to u with respect to $\mathrm{TGV}^{\mathrm{k}}_\alpha$.*

Proof. Let $k \in \mathbb{N}$ and $\alpha \in \mathbb{R}^k_{>0}$ be fixed. The proof uses techniques from the proofs of [38, Theorem 5.2.2] and [3, Theorem 3.9].

At first suppose that $u \in L^1(\Omega, \mathbb{R}^m)$ can be approximated by $(u_n)_{n\in\mathbb{N}}$ in $\mathrm{BV}(\Omega, \mathbb{R}^m)$ with respect to strict-$\mathrm{TGV}^{\mathrm{k}}_\alpha$ convergence. Using the inequality

$$\|u_n\|_{L^1} + \mathrm{TV}(u_n) \leq C\left(\|u_n\|_{L^1} + \mathrm{TGV}^{\mathrm{k}}_\alpha(u_n)\right)$$

from corollary 3.3, with suitable $C > 0$, we get that the sequence of measures $(\mathrm{D}\,u_n)_{n\in\mathbb{N}}$ is bounded in $\mathcal{M}(\Omega, \mathbb{R}^{m\times d})$, thus possess a subsequence weakly* converging to some $\mu \in \mathcal{M}(\Omega, \mathbb{R}^{m\times d})$. As in [3, Theorem 3.9] this implies, by using the classical integration by parts formula and passing to the limit as $n \to \infty$, that also $\mathrm{D}\,u = \mu \in \mathcal{M}(\Omega, \mathbb{R}^{m\times d})$, thus $u \in \mathrm{BV}(\Omega, \mathbb{R}^m)$.

Conversely, we show that, for any $u \in \mathrm{BV}(\Omega, \mathbb{R}^m)$, $\delta > 0$, there exists $u^\delta \in C^\infty(\Omega, \mathbb{R}^m)$ such that

$$\|u^\delta - u\|_{L^1} < \delta \text{ and } \mathrm{TGV}^{\mathrm{k}}_\alpha(u^\delta) \leq \mathrm{TGV}^{\mathrm{k}}_\alpha(u) + \delta.$$

Lower semi-continuity of $\mathrm{TGV}^{\mathrm{k}}_\alpha$ with respect to $\|\cdot\|_{L^1}$ then implies the result. Take $u \in \mathrm{BV}(\Omega, \mathbb{R}^m)$, $\delta > 0$ fixed.

Then, according to proposition 3.6, we can take $w_1, \dots, w_{k-1}$, with $w_i \in \mathrm{BD}(\Omega, \mathrm{Sym}^i(\mathbb{R}^d))$, $1 \leq i < k$, such that

$$\mathrm{TGV}_\alpha^k(u) = \alpha_{k-1} \| \mathrm{D}\, u - w_1 \|_{\mathcal{M}} + \sum_{l=2}^{k-1} \alpha_{k-l} \| \mathcal{E} w_{l-1} - w_l \|_{\mathcal{M}} + \alpha_0 \| \mathcal{E} w_{k-1} \|_{\mathcal{M}}.$$

Now define a countable number of sets $(\Omega_h)_{h \in \mathbb{N}}$ such that $\Omega = \bigcup_{h \in \mathbb{N}} \Omega_h$, $\overline{\Omega}_h \subset\subset \Omega$ for all $h \in \mathbb{N}$ and any point of Ω belongs to at most four sets Ω_h (cf. [3, Theorem 3.9] for a construction of such sets). Next we choose $(\phi_h)_{h \in \mathbb{N}}$ a partition of unity relative to $(\Omega_h)_{h \in \mathbb{N}}$, i.e. $\phi_h \geq 0$, $\phi_h \in C_c^\infty(\Omega_h)$ for all $h \in \mathbb{N}$ and $\sum_{h \in \mathbb{N}} \phi_h \equiv 1$.

Then, with $\rho \in C_c^\infty(\mathbb{R}^d)$ being a standard mollifier, for any $h \in \mathbb{N}$, we can find $\epsilon_h > 0$ such that $\mathrm{supp}((u\phi_h) * \rho_{\epsilon_h}) \subset \Omega_h$, $\mathrm{supp}((w_i\phi_h) * \rho_{\epsilon_h}) \subset \Omega_h$, for $1 \leq i < k$,

$$\int_\Omega |(u\phi_h) * \rho_{\epsilon_h} - u\phi_h| \,\mathrm{d}x < 2^{-h}\delta,$$

as well as

$$\int_\Omega |(u \otimes \nabla \phi_h) * \rho_{\epsilon_h} - u \otimes \nabla \phi_h| \,\mathrm{d}x < 2^{-h}(\delta/\alpha_{k-1}), \tag{38}$$

and

$$\int_\Omega |(w_i \otimes \nabla \phi_h) * \rho_{\epsilon_h} - w_i \otimes \nabla \phi_h| \,\mathrm{d}x < 2^{-h}(\delta/\alpha_{k-1-i}), \tag{39}$$

for $1 \leq i < k$. The functions $u^\delta = \sum_{h \in \mathbb{N}} (u\phi_h) * \rho_{\epsilon_h}$ and $w_i^\delta = \sum_{h \in \mathbb{N}} (w_i\phi_h) * \rho_{\epsilon_h}$ are smooth because the sums are locally finite. Moreover,

$$\int_\Omega |u^\delta - u| \,\mathrm{d}x \leq \sum_{h \in \mathbb{N}} \int_\Omega |(u\phi_h) * \rho_{\epsilon_h} - u\phi_h| \,\mathrm{d}x < \delta.$$

Now applying lemma 3.5 we get, for $\varphi \in C_c^\infty(\Omega, (\mathbb{R}^d)^m)$ with $\|\varphi\|_\infty \leq 1$,

$$\begin{aligned}
\int_\Omega u^\delta \cdot \mathrm{div}\, \varphi = & \sum_{h \in \mathbb{N}} \int_\Omega ((u\phi_h) * \rho_{\epsilon_h} \cdot \mathrm{div}\, \varphi \\
= & \sum_{h \in \mathbb{N}} \int_\Omega u\phi_h \cdot \mathrm{div}(\rho_{\epsilon_h} * \varphi) \\
= & \sum_{h \in \mathbb{N}} [\int_\Omega u \cdot \mathrm{div}(\phi_h(\varphi * \rho_{\epsilon_h})) - \int_\Omega u \otimes \nabla\, \phi_h \cdot (\rho_{\epsilon_h} * \varphi) \\
= & \sum_{h \in \mathbb{N}} [\int_\Omega -\mathrm{D}(u) \cdot \phi_h(\varphi * \rho_{\epsilon_h}) - \int_\Omega (((u \otimes \nabla\, \phi_h) * \rho_{\epsilon_h}) - u \otimes \nabla\, \phi_h) \cdot \varphi].
\end{aligned}$$

Multiplying by -1, subtracting $\int_\Omega w_1^\delta \cdot \varphi$ on both sides and using the estimate of lemma 3.6 we get

$$\begin{aligned}\int\limits_{\Omega} -u^{\delta}\cdot\operatorname{div}\varphi - w_1^{\delta}\cdot\varphi = \quad & \sum_{h\in\mathbb{N}}[\int\limits_{\Omega}((\mathrm{D}(u)\phi_h * \rho_{\epsilon_h}) - (w_1\phi_h * \rho_{\epsilon_h}))\cdot\varphi \\ & + \int\limits_{\Omega}(((u\otimes\nabla\phi_h) * \rho_{\epsilon_h}) - u\otimes\nabla\phi_h)]\cdot\varphi \\ \leq \quad & \sum_{h\in\mathbb{N}}[\int\limits_{\Omega}|(\mathrm{D}(u)\phi_h - w_1\phi_h) * \rho_{\epsilon_h}| \\ & + \int\limits_{\Omega}|(((u\otimes\nabla\phi_h) * \rho_{\epsilon_h}) - u\otimes\nabla\phi_h)|] \\ \leq \quad & \sum_{h\in\mathbb{N}}[\int\limits_{\Omega}\phi_h|(\mathrm{D}(u) - w_1| + 2^{-h}(\delta/\alpha_{k-1})] \\ \leq \quad & \int\limits_{\Omega}|\,\mathrm{D}(u) - w_1| + (\delta/\alpha_{k-1}),\end{aligned}$$

where we used equation (38). Similar, using equation (39), it follows for $\varphi \in C_c^{\infty}(\Omega, \mathrm{Sym}^{i+1}(\mathbb{R}^d)^m)$ with $\|\varphi\|_{\infty} \leq 1$,

$$\int\limits_{\Omega} -w_i^{\delta}\cdot\operatorname{div}\varphi - w_{i+1}^{\delta}\cdot\varphi \leq \int\limits_{\Omega}|\mathcal{E}w_i - w_{i+1}| + (\delta/\alpha_{k-1-i}),$$

for $1 \leq i < k-1$, and, for $\varphi \in C_c^{\infty}(\Omega, \mathrm{Sym}^{k}(\mathbb{R}^d)^m)$ with $\|\varphi\|_{\infty} \leq 1$,

$$\int\limits_{\Omega} w_{k-1}^{\delta}\cdot\operatorname{div}\varphi \leq \int\limits_{\Omega}|\mathcal{E}w_{k-1}| + (\delta/\alpha_0).$$

Taking the supremum over all such φ we get

$$\begin{aligned}\|\,\mathrm{D}\,u^{\delta} - w_1^{\delta}\|_{\mathcal{M}} &\leq \|\,\mathrm{D}\,u - w_1\|_{\mathcal{M}} + (\delta/\alpha_{k-1}), \\ \|\mathcal{E}w_i^{\delta} - w_{i+1}^{\delta}\|_{\mathcal{M}} &\leq \|\mathcal{E}w_i - w_{i+1}\|_{\mathcal{M}} + (\delta/\alpha_{k-1-i}),\end{aligned}$$

for $1 \leq i < k-1$, and

$$\|\mathcal{E}w_{k-1}^{\delta}\|_{\mathcal{M}} \leq \|\mathcal{E}w_{k-1}\|_{\mathcal{M}} + (\delta/\alpha_0).$$

Thus we have shown that, for $u^{\delta}, w_1^{\delta}, \ldots, w_{k-1}^{\delta}$ suitable,

$$\|u^{\delta} - u\|_{L^1} < \delta$$

and

$$\begin{aligned}\mathrm{TGV}_{\alpha}^{\mathrm{k}}(u^{\delta}) \leq \quad & \alpha_{k-1}\|\,\mathrm{D}\,u^{\delta} - w_1^{\delta}\|_{\mathcal{M}} \\ & + \sum_{i=1}^{k-2}\alpha_{k-1-i}\|\mathcal{E}(w_i^{\delta}) - w_{i+1}^{\delta}\|_{\mathcal{M}} + \alpha_0\|\mathcal{E}(w_{k-1}^{\delta})\|_{\mathcal{M}} \\ < \quad & \delta + \alpha_{k-1}\|\,\mathrm{D}\,u - w_1\|_{\mathcal{M}} \\ & + \sum_{i=1}^{k-2}\alpha_{k-1-i}\|\mathcal{E}(w_i) - w_{i+1}\|_{\mathcal{M}} + \alpha_0\|\mathcal{E}(w_{k-1})\|_{\mathcal{M}} \\ = \quad & \mathrm{TGV}_{\alpha}^{\mathrm{k}}(u) + \delta\end{aligned}$$

From that, the assertion follows. □

Remark 3.5. *Note that, given $u \in \mathrm{BV}(\Omega)$, the above construction yields an approximating sequence $(u_n)_{n\in\mathbb{N}}$ for fixed $k \in \mathbb{N}$, $\alpha \in \mathbb{R}^k_{>0}$. Since the elements of this sequence depend on the minimizers $w_1, \ldots, w_{k-1}$, we cannot assure that $(u_n)_{n\in\mathbb{N}}$ also converges strictly to u with respect to any other choice of k or α (except for the trivial case where $\tilde{\alpha} = \lambda\alpha$ with $\lambda \in \mathbb{R}_{>0}$, i.e. if the minimizers remain the same).*

Part II
A TGV regularized image reconstruction model

The main part of this work deals with the theory and application of a general, regularization based model for image reconstruction. Characteristic about this model is data fidelity, which is realized via the indicator function of a convex set that is described by a basis transformation operator.

The original motivation for this work was the application to artifact-free JPEG decompression. However, due to a very general problem statement, our framework will be applicable to a broad class of problems in mathematical imaging. We will start with a brief motivation followed by a definition and analysis of our model in function space setting. Note that we will, without further comment, make frequent use of the notation introduced in part I.

4 The general reconstruction model

4.1 Problem statement

Before we state the main minimization problem and the generic assumption, we want to briefly sketch the original motivation for our problem setting; a model for artifact free JPEG decompression [2, 13]. For a detailed introduction to this topic we ask for the readers patience until subsection 5.2.

Given any image u, the main step of JPEG compression is a transformation by a blockwise cosine transformation operator (to be defined in (63) later on) followed by quantization to integer. As a result, in the compressed JPEG file, the image u is described by quantized integers of its coefficients for a basis representation with respect to a blockwise cosine basis. Due to quantization, this data does not provide enough information to determine a unique source image of the compression process. But it is possible to define a set of basis coefficient data, whose quantization would coincide with the compressed image data. Denoting this set of basis coefficient data D and the blockwise cosine transformation operator BDCT, we can formally define the (convex) set of possible source images for any given compressed JPEG file by

$$U_D = \{u \mid \mathrm{BDCT}(u) \in D\}.$$

Thus, reconstructing an image from a given compressed JPEG file requires to choose one element of U_D as reconstruction. This choice can be motivated by a predefined image model, which is in our case realized by using the TGV functional as regularization term.

Keeping this application in mind, we now consider the following minimization problem

$$\min_{u \in L^2(\Omega,\mathbb{R}^m)} \mathrm{TGV}_\alpha^{\mathrm{k}}(u) + \mathcal{I}_{U_D}(u), \tag{40}$$

where we mainly use the assumptions

$$
\text{(A)}\begin{cases}
\Omega \subset \mathbb{R}^2 \text{ is a bounded Lipschitz domain,}\\
U_D = \{u \in L^2(\Omega,\mathbb{R}^m) \,|\, Au \in D\}\\
A : L^2(\Omega,\mathbb{R}^m) \to \ell^2, \quad (Au)_n := (u, a_n)_{L^2},\\
(a_n)_{n\in\mathbb{N}} \subset L^2(\Omega,\mathbb{R}^m) \text{ is a Riesz basis,}\\
(\tilde{a}_n)_{n\in\mathbb{N}}, \text{ the dual basis of } (a_n)_{n\in\mathbb{N}}, \text{ is contained in } \mathrm{BV}(\Omega,\mathbb{R}^m),\\
D = \left\{z \in \ell^2 \,|\, z_n \in J_n \;\forall\, n \in \mathbb{N}\right\},\\
(J_n)_{n\in\mathbb{N}} = ([l_n, o_n])_{n\in\mathbb{N}} \text{ is a sequence of non-empty, closed intervals,}\\
\overset{\circ}{\mathrm{U}}_{\mathrm{int}} \neq \emptyset \text{ with } \mathrm{U}_{\mathrm{int}} := \{u \in L^2(\Omega,\mathbb{R}^m) \,|\, (Au)_n \in J_n \,\forall\, n \in \mathbb{N} \setminus W\},\\
W \subset \mathbb{N} \text{ a finite index set,}\\
k, m \in \mathbb{N} \text{ and } \alpha = (\alpha_0, ..., \alpha_{k-1}) \in \mathbb{R}^k_{>0}.
\end{cases}
$$

By $\mathcal{I}_{U_D}$ we denote the convex indicator function of U_D, i.e.,

$$
\mathcal{I}_{U_D}(u) = \begin{cases} 0 & \text{if } u \in U_D,\\ \infty & \text{else.}\end{cases}
$$

Thus, solving the minimization problem (40) amounts to finding a function in U_D minimizing $\mathrm{TGV}^{\mathrm{k}}_{\alpha}$.

Thinking again of JPEG decompression, the operator A in assumption (A) will be the BDCT operator, while $(a_n)_{n\in\mathbb{N}}$ will be an orthonormal blockwise cosine basis. The dimension of the image space, $m \in \mathbb{N}$, reflects the number of image components, typically $m = 3$ for color and $m = 1$ for grayscale images.

The assumption that $\mathrm{U}_{\mathrm{int}}$ has nonempty interior can be seen as a generalization of the assumption that the data set U_D has nonempty interior, which is satisfied in the application to JPEG decompression, and will be needed for the analysis of the model. The more general assumption is motivated by the application of the model to zooming problems, where typically finitely many of the intervals $(J_n)_{n\in\mathbb{N}}$ will consist only of a single point.

At last let us emphasize that in particular the assumption of $(a_n)_{n\in\mathbb{N}}$ being a Riesz basis results in a broad applicability of our framework, as will be discussed in more detail in subsection 5.1. Note also that assumption (A) allows to choose m different scalar valued Riesz bases for $L^2(\Omega)$ and apply all results within assumption (A) to a corresponding Riesz basis of $L^2(\Omega,\mathbb{R}^m)$ as in remark 1.2.

4.2 Existence of a solution

As the name suggests, the purpose of this subsection is to show existence of a solution to the minimization problem (40) under assumption (A). For this purpose, we will need one additional assumption, which is necessary to get a bound on the kernel of the regularization functional $\mathrm{TGV}^{\mathrm{k}}_{\alpha}$. Note that throughout this subsection, we will always assume $\Omega \subset \mathbb{R}^2$ to be a bounded Lipschitz domain.

We start with a technical Lemma.

Lemma 4.1. *Let $(a_n)_{n\in\mathbb{N}}$ be a Riesz basis of $L^2(\Omega,\mathbb{R}^m)$, $m \in \mathbb{N}$. Then, for any $k \in \mathbb{N}$, and any sequence of linear independent functions $(z_i)_{1\leq i\leq k} \subset L^2(\Omega,\mathbb{R}^m)$*

there exist $a_{n_1}, ..., a_{n_k}$ such that the matrix B_k, defined by

$$B_k := \big((z_{k-j}, a_{n_{k-i}})_{L^2}\big)_{0\leq i,j\leq k-1} \in \mathbb{M}^{k\times k}, \tag{41}$$

is invertible.

Proof. We prove the assertion by induction:

$\mathbf{k = 1}$: Since $\mathbf{0} \neq z_1 \in L^2(\Omega, \mathbb{R}^m)$, there exists a_{n_1} such that $(z_1, a_{n_1}) \neq 0$.

$\mathbf{k \to k + 1}$: Given $k+1$ linear independent functions, $(z_i)_{1\leq i\leq k+1}$, suppose that the matrix B_k, generated with $(z_i)_{1\leq i\leq k}$ and suitable $(a_{n_i})_{1\leq i\leq k}$, has full rank and that there does not exist $a_{n_{k+1}}$ such that $\det(B_{k+1}) \neq 0$. Denoting by B^n_{k+1} the matrix generated from B_k by using a_n and z_{k+1} for the additional elements, from Laplace expansion we conclude, for all $n \in \mathbb{N}$,

$$0 = \det(B^n_{k+1}) = (z_{k+1}, a_n)_{L^2} \det(B_k) + \sum_{1\leq i\leq k} c_i(z_i, a_n)_{L^2}$$

with $c_i \in \mathbb{R}$ independent of n. But since $\det(B_k) \neq 0$ this implies

$$(z_{k+1}, a_n)_{L^2} = \sum_{1\leq i\leq k} \tilde{c}_i(z_i, a_n)_{L^2}$$

for all $n \in \mathbb{N}$, with $\tilde{c}_i \in \mathbb{R}$, and thus, using the dual Riesz basis $(\tilde{a}_n)_{n\in\mathbb{N}}$ as in proposition 1.7,

$$z_{k+1} = \sum_{n\in\mathbb{N}}(z_{k+1}, a_n)_{L^2}\tilde{a}_n = \sum_{1\leq i\leq k} \tilde{c}_i \sum_{n\in\mathbb{N}}(z_i, a_n)_{L^2}\tilde{a}_n = \sum_{1\leq i\leq k} \tilde{c}_i z_i$$

which contradicts to linear independence of z_{k+1}. □

Now for fixed $k \in \mathbb{N}$, define $(p_i)_{1\leq i\leq r}$ with $r = \frac{k(k+1)}{2}$ to be a basis for the real valued polynomials over Ω of order less than k, e.g. $p_i(x, y) = x^{w_i}y^{q_i}$ with $w_i, q_i \in \mathbb{N}_0$, $w_i + q_i < k$ suitable. Defining then $(z_i)_{1\leq i\leq mr} = ((z^1_i, \ldots, z^m_i))_{1\leq i\leq mr} \subset L^2(\Omega, \mathbb{R}^m)$ by

$$z^j_i = \begin{cases} p_{i-(j-1)r} & \text{if } (j-1)r + 1 \leq i \leq jr, \\ 0 & \text{else,} \end{cases} \tag{42}$$

to be a component wise basis for the $\mathbb{R}^m$ valued polynomials over Ω, which yields a linear independent sequence in $L^2(\Omega, \mathbb{R}^m)$, we immediately obtain the following corollary, which will be useful to conclude boundedness of a minimizing sequence to the minimization problem (40).

Corollary 4.1. *Take $k, m \in \mathbb{N}$, $(a_n)_{n\in\mathbb{N}}$ to be a Riesz basis of $L^2(\Omega, \mathbb{R}^m)$ and define $(z_i)_{1\leq i\leq mr}$ with $r = \frac{k(k+1)}{2}$ to be a component wise basis for the polynomials of order less than k as in (42). Then there exist $n_1, \ldots, n_{mr} \in \mathbb{N}$ such that the matrix*

$$\begin{pmatrix} (z_{mr}, a_{n_{mr}})_{L^2} & \ldots & (z_1, a_{n_{mr}})_{L^2} \\ \vdots & & \vdots \\ (z_{mr}, a_1)_{L^2} & \ldots & (z_1, a_1)_{L^2} \end{pmatrix}$$

has full rank.

For $(a_n)_{n\in\mathbb{N}} \subset L^2(\Omega,\mathbb{R}^m)$ a Riesz basis, $n_1,\ldots,n_{mr}$, denote by $B((a_{n_i})_{i=1}^{mr})$ the matrix generated with $(a_{n_i})_{i=1}^{mr}$ and a basis of the polynomials as in corollary 4.1. This allows us now to make the following assumption, which is crucial to obtain existence of a solution:

($\mathrm{EX_k}$) Let (A) be satisfied. For $1 \leq i \leq m$ and $r = \frac{k(k+1)}{2}$, choose $n_1,\ldots,n_{mr}$ according to corollary 4.1 such that $B((a_{n_i})_{i=1}^{mr})$ has full rank and suppose that $J_{n_1},\ldots,J_{n_{mr}}$ are bounded.

Remark 4.1. *Note that, given the $k\in\mathbb{N}$ the order of the $\mathrm{TGV}_\alpha^\mathrm{k}$ functional and $m\in\mathbb{N}$ the number of image components, ($\mathrm{EX_k}$) assumes $m\frac{k(k+1)}{2}$ suitable intervals to be bounded. Since this is exactly the dimension of the space of $\mathbb{R}^m$-valued polynomials of degree less or equal to $k-1$, denoted by $\mathcal{P}_{k-1}(\Omega,\mathbb{R}^m)$, we cannot get a bound on a minimizing sequence in general by just requiring a smaller number of interval to be bounded. This can be seen as follows:*

First note that, clearly, any linear map from $\mathcal{P}_{k-1}(\Omega,\mathbb{R}^m)$ to $\mathbb{R}^l$ with $l\in\mathbb{N}$, $l < m\frac{k(k+1)}{2}$, must have a nontrivial kernel. Also, $\mathrm{TGV}_\alpha^\mathrm{k}(u+p) = \mathrm{TGV}_\alpha^\mathrm{k}(u)$ for any $u\in L^2(\Omega,\mathbb{R}^m)$, $p\in\mathcal{P}_{k-1}(\Omega,\mathbb{R}^m)$.

If we assume now that all intervals as in assumption (A), except for l of them, are the whole reals, a modification of the basis transformation operator A to a mapping from $\mathcal{P}_{k-1}(\Omega,\mathbb{R}^m)$ to $\mathbb{R}^m$, by using only the components of A related to the l bounded intervals as image components, must have a nontrivial kernel. Choosing an unbounded sequence in this kernel, we can always add this sequence to any minimizing sequence and, by not affecting the value of $\mathrm{TGV}_\alpha^\mathrm{k}$, obtain an unbounded minimizing sequence as claimed.

We can now use the previous results to proof existence of a solution under weak assumptions:

Proposition 4.1. *Let* (A) *and* ($\mathrm{EX_k}$) *be satisfied. Then there exists a solution to the minimization problem*

$$\min_{u\in L^2(\Omega,\mathbb{R}^m)} \mathrm{TGV}_\alpha^\mathrm{k}(u) + \mathcal{I}_{U_D}(u). \tag{43}$$

Proof. First we show that $F(u) := \mathrm{TGV}_\alpha^\mathrm{k}(u) + \mathcal{I}_{U_D}(u)$ is proper, i.e., F takes nowhere the value $-\infty$ and is finite in at least one point. It is clear that $F(u)\geq 0$, so it remains to find an element $u\in L^2(\Omega,\mathbb{R}^m)$ such that $F(u)<\infty$. This can be constructed as follows: Using that $\overset{\circ}{\mathrm{U}}_{\mathrm{int}} \neq \emptyset$, by density of $C_c^\infty(\Omega,\mathbb{R}^m)$ in $L^2(\Omega,\mathbb{R}^m)$ there exists $u_0\in \mathrm{BV}(\Omega,\mathbb{R}^m)\cap \mathrm{U}_{\mathrm{int}}$.

Now since $(\tilde{a}_n)_{n\geq 0}\subset \mathrm{BV}(\Omega,\mathbb{R}^m)$, for u_0 we can find constants $s_n\in\mathbb{R}$ such that, with

$$u(x,y) := u_0(x,y) + \sum_{n\in W} s_n \tilde{a}_n(x,y),$$

it follows that $u\in \mathrm{BV}(\Omega,\mathbb{R}^m)\cap U_D$.

In order to find a minimizing element, we take a minimizing sequence $(u_n)_{n\geq 0}$, i.e., $\lim_{n\to\infty} F(u_n) = \inf_{u\in L^2(\Omega,\mathbb{R}^m)} F(u)$, for which without loss of generality we can assume that $(u_n)_{n\geq 0}\subset \mathrm{BV}(\Omega,\mathbb{R}^m)\cap U_D$. To show boundedness of $(u_n)_{n\geq 0}$ with respect to $\|\cdot\|_{L^2}$ we estimate

$$\|u_n\|_{L^2} \leq \|u_n - P_{k-1}u_n\|_{L^2} + \|P_{k-1}u_n\|_{L^2} \leq C\,\mathrm{TGV}_\alpha^\mathrm{k}(u_n) + \|P_{k-1}u_n\|_{L^2}$$

with P_{k-1} a linear, continuous onto projection on the $\mathbb{R}^m$-valued polynomials of order less than k, denoted by $\mathcal{P}_{k-1}(\Omega, \mathbb{R}^m)$, and the last estimation holds, with $C > 0$, by proposition 3.7. Using this estimation it suffices to show boundedness of $\|P_{k-1}u_n\|_{L^2}$. Since Ω is connected, it follows from $\nabla^k P_{k-1}u_n = 0$ that

$$u_n^i(x, y) = \sum_{w+q<k} c_{n,w,q}^i x^w y^q$$

with $c_{n,w,q}^i \in \mathbb{R}$, $1 \leq i \leq m$. Hence it is sufficient to show boundedness of $c_{n,w,q}^i$ with respect to n. Denoting by $\overline{c}_n^i \in \mathbb{R}^r$, $r = \frac{k(k+1)}{2}$, the vector containing all $c_{n,w,q}^i$ and by $\overline{c}_n$ the vector obtained by padding all $\overline{c}_n^i$ together, choosing $B((a_{n_i})_{i=1}^{mr}) \in \mathbb{M}^{k \times k}$ as in $(\mathrm{EX_k})$, $u_n \in U_D$ yields

$$B_{mr}\overline{c}_n = z_n$$

with $z_n = (z_n^{mr}, \ldots, z_n^1)$ and z_n^i bounded for $1 \leq i \leq mr$ by $(\mathrm{EX_k})$ and boundedness of $\|u_n - P_{k-1}u_n\|_{L^2}$. Finally invertibility of B_{mr} implies boundedness of $P_{k-1}u_n$ in $L^2(\Omega, \mathbb{R}^m)$.

Hence there exists $\hat{u} \in L^2(\Omega, \mathbb{R}^m)$ and a subsequence of $(u_n)_{n\geq 0}$, denoted by $(u_{n_l})_{l\geq 0}$, weakly converging to $\hat{u}$ in $L^2(\Omega, \mathbb{R}^m)$. At last, convexity and closedness of U_D in $L^2(\Omega, \mathbb{R}^m)$ imply weak closedness (see [37, Section I.1.2]) and thus, from lower semi continuity of $\mathrm{TGV}_\alpha^{\mathrm{k}}$ with respect to the weak L^2 topology (see proposition 3.5), it follows that

$$F(\hat{u}) \quad = \quad \mathrm{TGV}_\alpha^{\mathrm{k}}(\hat{u}) \leq \underline{\lim}_{l\to\infty} \mathrm{TGV}_\alpha^{\mathrm{k}}(u_{n_l}) = \underline{\lim}_{l\to\infty} F(u_{n_l}) = \inf_{u \in L^2(\Omega, \mathbb{R}^m)} F(u)$$

which implies that $\hat{u}$ is a solution to (43). □

Remark 4.2. *Note that, even though we used that* $\overset{\circ}{\mathrm{U}}_{int} \neq \emptyset$ *in the proof of proposition 4.1 to obtain that* $\mathrm{TGV}_\alpha^{\mathrm{k}} + \mathcal{I}_{U_D}$ *is proper, it would have been sufficient to assume that* $U_D \cap \mathrm{BV}(\Omega, \mathbb{R}^m) \neq \emptyset$ *instead, to obtain properness and hence the same existence result. The even stronger assumption* $\overset{\circ}{\mathrm{U}}_{int} \neq \emptyset$ *will play a role when showing additivity of the subdifferential for the derivation of the optimality condition.*

4.3 Optimality condition

Having obtained existence of a solution to (40) we now draw our attention to the derivation of an optimality condition. For this purpose, we will make use of the following obvious identity: Given a function F,

$$u^* = \arg\min_u F u) \quad \Leftrightarrow \quad 0 \in \partial F(u^*).$$

The derivation of an optimality condition will thus be preceded by three main steps:

- Describe $\partial\,\mathrm{TGV}_\alpha^{\mathrm{k}}$, the subdifferential of $\mathrm{TGV}_\alpha^{\mathrm{k}}$.
- Describe $\partial\,\mathcal{I}_{U_D}$, the subdifferential of $\mathcal{I}_{U_D}$.
- Show additivity of the subdifferential operator under assumption (A).

4.3.1 Subdifferential of the TGV functional

Since a description of the subdifferential of the TGV functional is of interest not only for our specific problem setting, we will for a moment leave the context of assumption (A) and, in this subsection, always use the following assumptions:

$$\begin{cases} \Omega \subset \mathbb{R}^d \text{ is a bounded Lipschitz-domain with } d \in \mathbb{N}, d \geq 2 \\ p \in \mathbb{R} \text{ with } 1 < p \leq \frac{d}{d-1} \text{ and } m \in \mathbb{N}. \end{cases}$$

Further, we will denote the conjugate exponent of p by $p' := \frac{p}{p-1}$. Note that the restriction on p is to maintain a continuous embedding of $\mathrm{BV}(\Omega, \mathbb{R}^m)$ to $L^p(\Omega, \mathbb{R}^m)$ (see proposition 1.5).

Also, for this subsection, we always assume $\mathrm{TGV}_\alpha^{\mathrm{k}}$ to be a functional defined on $L^p(\Omega, \mathbb{R}^m)$:

Definition 4.1. *We define*

$$\begin{aligned} \mathrm{TGV}_\alpha^{\mathrm{k}} : L^p(\Omega, \mathbb{R}^m) &\to \overline{\mathbb{R}} \\ u &\mapsto \mathrm{TGV}_\alpha^{\mathrm{k}}(u). \end{aligned}$$

At first, we describe the convex conjugate (or polar) of the $\mathrm{TGV}_\alpha^{\mathrm{k}}$ functional, which will be useful for the subdifferential characterization later on.

Proposition 4.2. *The convex conjugate of* $\mathrm{TGV}_\alpha^{\mathrm{k}}$*, denoted by*

$$\mathrm{TGV}_\alpha^{\mathrm{k}\,*} : L^{p'}(\Omega, \mathbb{R}^m) \to \overline{\mathbb{R}},$$

has the form

$$\mathrm{TGV}_\alpha^{\mathrm{k}\,*}(v) = \mathcal{I}_{\overline{C_\alpha^k}}(v) = \begin{cases} 0 & v \in \overline{C_\alpha^k} \\ \infty & v \notin \overline{C_\alpha^k} \end{cases}$$

where

$$C_\alpha^k := \left\{ \operatorname{div}^k \xi \;\middle|\; \xi \in C_c^k(\Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m), \ \| \operatorname{div}^l \xi \|_\infty \leq \alpha_l, \ l = 0, ..., k-1 \right\}, \tag{44}$$

and the closure is taken with respect to the $L^{p'}$ *norm.*

Proof. This follows easily from convexity and lower semi continuity of $\mathrm{TGV}_\alpha^{\mathrm{k}}$ and $\mathcal{I}_{\overline{C_\alpha^k}}$ since

$$\mathrm{TGV}_\alpha^{\mathrm{k}}(u) = \mathcal{I}_{C_\alpha^k}^*(u)$$

and thus (see [37, Propositions 3.2 and 4.1]),

$$\mathrm{TGV}_\alpha^{\mathrm{k}\,*}(v) = \mathcal{I}_{C_\alpha^k}^{**}(v) = \mathcal{I}_{\overline{C_\alpha^k}}(v).$$

□

A more detailed description of $\mathrm{TGV}_\alpha^{\mathrm{k}\,*}$ follows from a study of $\overline{C_\alpha^k}$:

Proposition 4.3. *With* C_α^k *as in Proposition 4.2, we have*

$$\begin{aligned} \overline{C_\alpha^k} = \Big\{ \operatorname{div}^k g \;\Big|\; & g \in W_0^{p'}(\operatorname{div}^k; \Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m), \\ & \| \operatorname{div}^l g \|_\infty \leq \alpha_l, \ l = 0, ..., k-1 \Big\} := K_\alpha^k, \end{aligned} \tag{45}$$

where the closure is taken with respect to the $L^{p'}$ *norm.*

Proof. In order to show that $\overline{C^k_\alpha} \subset K^k_\alpha$ it is sufficient to show that K^k_α is closed with respect to $\|\cdot\|_{L^{p'}}$. Define

$$W_0^{p',\alpha}(\operatorname{div}^k) := \{g \in W_0^{p'}(\operatorname{div}^k;\Omega,\operatorname{Sym}^k(\mathbb{R}^d)^m) : \|\operatorname{div}^l g\|_\infty \leq \alpha_l,\, l = 0,...,k{-}1\}.$$

Take now $h \in \overline{K^k_\alpha}$. Hence there exists a sequence $(g_n)_{n\geq 0} \subset W_0^{p',\alpha}(\operatorname{div}^k)$ such that $\lim_{n\to\infty} \operatorname{div}^k g_n = h$. If we can show that there exists $g \in W_0^{p',\alpha}(\operatorname{div}^k)$ such that $\operatorname{div}^k g = h$, closedness of K^k_α with respect to $\|\cdot\|_{L^{p'}}$ follows. By boundedness of $\|\operatorname{div}^l g_n\|_\infty$, $0 \leq l < k$, there exist $h^l \in L^{p'}(\Omega,\operatorname{Sym}^{k-l}(\mathbb{R}^d)^m)$ and a set of increasing indices $(n_i)_{i\in\mathbb{N}}$ in $\mathbb{N}$ such that

$$\operatorname{div}^l g_{n_i} \underset{L^{p'}}{\rightharpoonup} h^l \quad \text{as } i \to \infty, \text{ for all } 0 \leq l < k.$$

It follows that, for $0 \leq l < k-1$ and $\phi \in C_c^\infty(\Omega,\operatorname{Sym}^{k-1-l}(\Omega)^m)$,

$$\int_\Omega h^l \cdot \mathcal{E}\phi \;=\; \lim_{i\to\infty}\int_\Omega \operatorname{div}^l g_{n_i}\cdot\mathcal{E}\phi = \lim_{i\to\infty}(-1)\int_\Omega \operatorname{div}^{l+1} g_{n_i}\cdot\phi = (-1)\int_\Omega h^{l+1}\cdot\phi$$

which implies $g := h^0 \in W^{p'}(\operatorname{div}^{k-1};\Omega,\operatorname{Sym}^k(\mathbb{R}^d)^m)$ and $\operatorname{div}^l g = h^l$, $0 \leq l \leq k-1$. Further we have, for any $\phi \in C_c^\infty(\Omega,\mathbb{R}^m)$, that

$$\int_\Omega h^{k-1}\cdot\mathcal{E}\phi = \lim_{i\to\infty}\int_\Omega \operatorname{div}^{k-1} g_{n_i}\cdot\mathcal{E}\phi = \lim_{n\to\infty}(-1)\int_\Omega \operatorname{div}^k g_{n_i}\phi = (-1)\int_\Omega h\phi$$

and thus $g \in W^{p'}(\operatorname{div}^k;\Omega,\operatorname{Sym}^k(\mathbb{R}^d)^m)$ and $\operatorname{div}^k g = h$.

In order to prove that $g \in W_0^{p',\alpha}(\operatorname{div}^k)$ we note that the set

$$\Big\{(z,\operatorname{div} z,\ldots,\operatorname{div}^k z) | z \in W_0^{p'}(\operatorname{div}^k;\Omega,\operatorname{Sym}^k(\mathbb{R}^d)^m),\, \|\operatorname{div}^l z\|_{L^\infty} \leq \alpha_l,\, 0 \leq l < k\Big\}$$

is a convex and closed – and therefore weakly closed – subset of

$$L^{p'}(\Omega,\prod_{l=0}^k \operatorname{Sym}^{k-l}(\mathbb{R}^d)^m).$$

Since the sequence

$$(g_{n_i},\operatorname{div} g_{n_i},\ldots,\operatorname{div}^k g_{n_i})$$

is contained in this set and converges weakly to

$$(g,\operatorname{div} g,\ldots,\operatorname{div}^k g)$$

it follows that $g \in W_0^{p',\alpha}(\operatorname{div}^k)$.

$K^k_\alpha \subset \overline{C^k_\alpha}$: It suffices to show that, for $g \in W_0^{p',\alpha}(\operatorname{div}^k)$ arbitrary, we have

$$\int_\Omega u \operatorname{div}^k g \leq \mathrm{TGV}^{\mathrm{k}}_\alpha(u) \quad \forall u \in \mathrm{BV}(\Omega,\mathbb{R}^m),$$

since this implies that ${\mathrm{TGV}^{\mathrm{k}}_\alpha}^*(\operatorname{div}^k g) = 0$ and hence $\operatorname{div}^k g \in \overline{C^k_\alpha}$. By a straightforward generalization of the result in [16, Remark 3.8] to the vector valued case, it follows that for sufficiently smooth ϕ one can describe $\mathrm{TGV}^{\mathrm{k}}_\alpha(\phi)$ as

$$\mathrm{TGV}^{\mathrm{k}}_\alpha(\phi) = \inf_{\nabla^k\phi=\phi_0+\mathcal{E}(\phi_1)+\ldots+\mathcal{E}^{k-1}(\phi_{k-1})} \sum_{l=0}^{k-1}\alpha_l\|\phi_l\|_{L^1},$$

where $\phi_l \in C^\infty(\Omega, \mathrm{Sym}^{k-l}(\mathbb{R}^d)^m)$. Now take any $\phi \in C^\infty(\Omega, \mathbb{R}^m)$ and $\phi_0, ...\phi_{k-1}$ such that $\nabla^k \phi = \phi_0 + \mathcal{E}(\phi_1) + ... + \mathcal{E}^{k-1}(\phi_{k-1})$. We then have

$$\int_\Omega \phi \operatorname{div}^k g = (-1)^k \sum_{l=0}^{k-1} \int_\Omega \mathcal{E}^l(\phi_l) \cdot g \leq \sum_{l=0}^{k-1} \int_\Omega |\phi_l| |\operatorname{div}^l g| \leq \sum_{l=0}^{k-1} \alpha_l \|\phi_l\|_{L^1}.$$

Taking the infimum over all such decompositions of $\nabla^k \phi$ leads to

$$\int_\Omega \phi \operatorname{div}^k g \leq \mathrm{TGV}_\alpha^\mathrm{k}(\phi).$$

By approximation of $u \in \mathrm{BV}(\Omega, \mathbb{R}^m)$ by a sequence in $C^\infty(\Omega, \mathbb{R}^m)$ with respect to $\mathrm{TGV}_\alpha^\mathrm{k}$-strict convergence (see proposition 3.8) the assertion follows.

□

Having a sufficient description of ${\mathrm{TGV}_\alpha^\mathrm{k}}^*$, we can now characterize its subdifferential. The relation

$$u^* \in \partial \mathrm{TGV}_\alpha^\mathrm{k}(u) \Leftrightarrow \mathrm{TGV}_\alpha^\mathrm{k}(u) + {\mathrm{TGV}_\alpha^\mathrm{k}}^*(u^*) = \langle u, u^* \rangle.$$

(see [37], Proposition I.5.1) together with the description of ${\mathrm{TGV}_\alpha^\mathrm{k}}^*$ immediately implies the following result:

Theorem 4.1. *(Characterization of the subdifferential of* $\mathrm{TGV}_\alpha^\mathrm{k}$*) Let* $u \in L^p(\Omega, \mathbb{R}^m)$, $u^* \in L^{p'}(\Omega, \mathbb{R}^m)$. *Then* $u^* \in \partial \mathrm{TGV}_\alpha^\mathrm{k}(u)$ *if and only if*

> $u \in \mathrm{BV}(\Omega, \mathbb{R}^m)$ *and there exists* $g \in W_0^{p'}(\operatorname{div}^k; \Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)$ *such that* $\|\operatorname{div}^l g\|_\infty \leq \alpha_l$, $l = 0, \ldots, k-1$, $u^* = \operatorname{div}^k g$ *and*
>
> $$\mathrm{TGV}_\alpha^\mathrm{k}(u) = \int_\Omega u \operatorname{div}^k g.$$

4.3.2 $\mathcal{I}_U$ subdifferential

In order to describe $\partial \mathcal{I}_{U_D}$, first note that we can decompose $\mathcal{I}_{U_D} = \mathcal{I}_D \circ A$. Since $\partial \mathcal{I}_D$ can be described quite easily, our aim is to use a chain rule to imply $\partial \mathcal{I}_{U_D} = \partial(\mathcal{I}_D \circ A) = A^* \partial \mathcal{I}_D \circ A$. To ensure this, well known results rely on either orthogonality of A or existence of a point $u \in L^2(\Omega, \mathbb{R}^m)$ such that $\mathcal{I}_D$ is continuous in Au. But none of these assumptions is satisfied in general in within our setting. However, as the following proposition shows, a general result given in [4] can be used to argue that a chain rule holds even if A is only assumed to be a continuous, bijective, linear operator.

Proposition 4.4. *Let* H_1, H_2 *be two Hilbert spaces,* $A : H_1 \to H_2$ *a continuous, bijective, linear operator and* $F : H_2 \to \overline{\mathbb{R}}$ *a proper, convex, lower semi-continuous function. Then*

$$\partial(F \circ A) = A^* \partial F \circ A.$$

Proof. First we define $H : H_1 \times H_2 \to \mathbb{R}$ as

$$H(u, v) = \mathcal{I}_{\{0\}}(Au - v) + F(v).$$

It then follows easily that, given any $v \in H_2$,

$$p \in \partial(F \circ A)(u) \Leftrightarrow (p, 0) \in \partial H(u, v),$$

thus we want to describe ∂H. Defining $G : H_1 \times H_2 \to \overline{\mathbb{R}}$ as $G(u, v) = \mathcal{I}_{\{0\}}(Au - v)$ it follows that

$$G^*(p, q) = \sup_{u \in H_1} (p, u)_{H_1} + (q, Au)_{H_2} = \mathcal{I}_{\{0\}}(p + A^*q)$$

and with that, again by [37], Proposition I.5.1,

$$(p, q) \in \partial G(u, v) \Leftrightarrow Au = v \text{ and } p = -A^*q.$$

Setting $\tilde{F} : H_1 \times H_2 \to \overline{\mathbb{R}}$ to $\tilde{F}(u, v) = F(v)$, it follows easily that $(p, q) \in \partial \tilde{F}(u, v) \Leftrightarrow p = 0$ and $q \in \partial F(v)$. Now to describe $\partial H(u, v) = \partial(G(u, v) + \tilde{F}(u, v))$, we need to establish additivity of ∂ in this case. Following [4, Corollary 2.1], for that it suffices to show

$$\operatorname{dom}(G) - \operatorname{dom}(\tilde{F}) = H_1 \times H_2.$$

Taking $(u, v) \in H_1 \times H_2$ arbitrary, we can write, with $w \in \operatorname{dom}(F)$,

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} A^{-1}(v + w) \\ v + w \end{pmatrix} - \begin{pmatrix} A^{-1}(v + w) - u \\ w \end{pmatrix} \in \operatorname{dom}(G) - \operatorname{dom}(\tilde{F}).$$

Thus $\partial H = \partial G + \partial F$ and with that, $(p, q) \in \partial H(u, v)$ is equivalent to existence of $r \in \partial(F(v))$ such that $(p, q - r) \in \partial G(u, v)$, which is again equivalent to

$$r \in \partial(F(Au)) \text{ and } p = -A^*(q - r).$$

In total we get, for any $p, u \in H_1$ and $v \in H_2$ arbitrary,

$$p \in \partial(F \circ A)(u) \Leftrightarrow (p, 0) \in \partial H(u, v) \Leftrightarrow p = A^*r \text{ with } r \in \partial F(Au),$$

which implies the desired chain rule. □

Remark 4.3. *We want to point out that the result of proposition 4.4 is, independent of our problem formulation, interesting by its own since it provides a chain rule for the subdifferential in a very general setting.*

We will now use this result in order to show a characterization of the subdifferential of the convex indicator function $\mathcal{I}_{U_D}$:

Theorem 4.2. *Let* (A) *be satisfied. Writing* $J_n = [l_n, o_n]$ *we have that*

$$u^* \in \partial \mathcal{I}_{U_D}(u) \Leftrightarrow u \in U_D \textit{ and } u^* = A^*\lambda$$

with $\lambda = (\lambda_n)_{n \in \mathbb{N}} \in \ell^2(\mathbb{N})$ *such that, for every* $n \in \mathbb{N}$,

$$\begin{aligned} &\lambda_n \geq 0 \quad \textit{if} \quad (Au)_n = o_n \neq l_n \\ &\lambda_n \leq 0 \quad \textit{if} \quad (Au)_n = l_n \neq o_n \\ &\lambda_n = 0 \quad \textit{if} \quad (Au)_n \in \overset{\circ}{J}_n \\ &\lambda_n \in \mathbb{R} \quad \textit{if} \quad (Au)_n = l_n = o_n. \end{aligned}$$

Proof. At first we can apply proposition 4.4 to conclude

$$u^* \in \partial \mathcal{I}_{U_D}(u) \Leftrightarrow u^* = A^* \lambda$$

for some $\lambda \in \partial \mathcal{I}_D(Au)$. By a standard result in convex analysis we have

$$\lambda \in \partial \mathcal{I}_D(Au) \Leftrightarrow Au = P_D(Au + \lambda).$$

where, by a straightforward contradiction argument, P_D can be reduced to a component wise projection;

$$Au = P_D(Au + \lambda) \Leftrightarrow (Au)_n = P_{J_n}((Au)_n + \lambda_n) \quad \forall n \in \mathbb{N}.$$

From that, the assertion follows by an easy case study. □

4.3.3 Additivity of the subdifferential

At last, we need to show that $\partial(\mathrm{TGV}_\alpha^\mathrm{k}(u) + \mathcal{I}_{U_D}(u)) = \partial \,\mathrm{TGV}_\alpha^\mathrm{k}(u) + \partial \mathcal{I}_{U_D}(u)$. For that, we first decompose $\mathcal{I}_{U_D} = \mathcal{I}_{\mathrm{U}_\mathrm{int}} + \mathcal{I}_{\mathrm{U}_\mathrm{point}}$ where, based on assumption (A),

$$\mathrm{U}_\mathrm{int} = \{u \in L^2(\Omega, \mathbb{R}^m) \,|\, (Au)_n \in J_n \,\forall\, n \in \mathbb{N} \setminus W\}$$

and

$$\mathrm{U}_\mathrm{point} = \{u \in L^2(\Omega, \mathbb{R}^m) \,|\, (Au)_n \in J_n \,\forall\, n \in W\}$$

for a finite index set $W \subset \mathbb{N}$ such that $\overset{\circ}{\mathrm{U}}_\mathrm{int} \neq \emptyset$.

Theorem 4.3. *Let* (A) *be satisfied. Then, for all* $u \in L^2(\Omega, \mathbb{R}^m)$,

$$\partial\left(\mathrm{TGV}_\alpha^\mathrm{k}(u) + \mathcal{I}_{U_D}(u)\right) \quad = \quad \partial \,\mathrm{TGV}_\alpha^\mathrm{k}(u) + \partial \mathcal{I}_{U_D}(u).$$

Proof. Let $u \in L^2(\Omega, \mathbb{R}^m)$. It is sufficient to show $\partial\left(\mathrm{TGV}_\alpha^\mathrm{k}(u) + \mathcal{I}_{U_D}(u)\right) \subset \partial \,\mathrm{TGV}_\alpha^\mathrm{k}(u) + \partial \mathcal{I}_{U_D}(u)$, since the other inclusion is always satisfied. Continuity of $\mathcal{I}_{\mathrm{U}_\mathrm{int}}$ in at least one point $u \in \mathrm{BV}(\Omega, \mathbb{R}^m) \cap U_D$ allows to apply [37, Proposition I.5.6], and assure that

$$\partial(\mathrm{TGV}_\alpha^\mathrm{k}(u) + \mathcal{I}_{\mathrm{U}_\mathrm{point}}(u) + \mathcal{I}_{\mathrm{U}_\mathrm{int}}(u)) \subset \partial(\mathrm{TGV}_\alpha^\mathrm{k}(u) + \mathcal{I}_{\mathrm{U}_\mathrm{point}}(u)) + \partial(\mathcal{I}_{\mathrm{U}_\mathrm{int}}(u))$$

Similar as in the proof of Theorem 4.2 we now want to use [4, Corollary 2.1] to establish

$$\partial(\mathrm{TGV}_\alpha^\mathrm{k}(u) + \mathcal{I}_{\mathrm{U}_\mathrm{point}}(u)) \subset \partial(\mathrm{TGV}_\alpha^\mathrm{k}(u)) + \partial(\mathcal{I}_{\mathrm{U}_\mathrm{point}}(u)),$$

for which it is sufficient to show that $\mathrm{dom}(\mathrm{TGV}_\alpha^\mathrm{k}) + \mathrm{dom}(\mathcal{I}_{\mathrm{U}_\mathrm{point}}) = L^2(\Omega, \mathbb{R}^m)$. But this is true since, for any $w \in L^2(\Omega, \mathbb{R}^m)$ by taking $j_n \in J_n$ for $n \in W$, we can write

$$w = w_1 - w_2$$

where

$$w_1 = \sum_{n \in W} ((a_n, w)_{L^2} - j_n)\, \tilde{a}_n \in \mathrm{dom}(\mathrm{TGV}_\alpha^\mathrm{k}(u))$$

and

$$w_2 = -\sum_{n \in N \setminus W} (a_n, w)_{L^2} \tilde{a}_n - \sum_{n \in W} j_n \tilde{a}_n \in \mathrm{dom}(\mathcal{I}_{\mathrm{U}_\mathrm{point}}).$$

Again, since

$$\partial(\mathcal{I}_{\mathrm{U}_\mathrm{point}}(u)) + \partial(\mathcal{I}_{\mathrm{U}_\mathrm{int}}(u)) \subset \partial(\mathcal{I}_{\mathrm{U}_\mathrm{point}}(u) + \mathcal{I}_{\mathrm{U}_\mathrm{int}}(u)) = \partial(\mathcal{I}_{U_D}(u))$$

is always satisfied, the assertion is proven. □

4.3.4 Optimality system

The previous results finally allow to derive an optimality system:

Theorem 4.4. *Let* (A) *and* (EX$_\text{k}$) *be satisfied. Then there exists a solution of*

$$\min_{u\in L^2(\Omega,\mathbb{R}^m)} \left(\mathrm{TGV}^{\mathrm{k}}_\alpha(u) + \mathcal{I}_{U_D}(u)\right)$$

and the following are equivalent

1. $\hat{u} = \operatorname*{arg\,min}_{u\in L^2(\Omega,\mathbb{R}^m)} \left(\mathrm{TGV}^{\mathrm{k}}_\alpha(u) + \mathcal{I}_{U_D}(u)\right) = \operatorname*{arg\,min}_{u\in U_D} \mathrm{TGV}^{\mathrm{k}}_\alpha(u),$

2. $\hat{u} \in \mathrm{BV}(\Omega,\mathbb{R}^m)\cap U_D$ *and there exist* $g \in W^2_0(\mathrm{div}^k;\Omega,\mathrm{Sym}^k(\mathbb{R}^2)^m)$ *and* $\lambda = (\lambda_n)_{n\in\mathbb{N}} \in \ell^2$ *satisfying*

   (a) $\|\operatorname{div}^l g\|_\infty \le \alpha_l$, $l = 0, ..., k-1$

   (b) $\mathrm{TGV}^{\mathrm{k}}_\alpha(\hat{u}) = -\int_\Omega \hat{u}\operatorname{div}^k g$

   (c) $\operatorname{div}^k g = \sum_{n\in\mathbb{N}} \lambda_n a_n$, *where*

   $$\begin{cases} \lambda_n \ge 0 \text{ if } (A\hat{u})_n = o_n \ne l_n \\ \lambda_n \le 0 \text{ if } (A\hat{u})_n = l_n \ne o_n \\ \lambda_n = 0 \text{ if } (A\hat{u})_n \in \overset{\circ}{J_n} \end{cases} \qquad \forall n\in\mathbb{N}$$

   (note that, if $J_n = \{j_n\}$*, there is no additional condition on* λ_n*),*

3. $\hat{u} \in \mathrm{BV}(\Omega,\mathbb{R}^m)\cap U_D$ *and there exists* $g \in W^2_0(\mathrm{div}^k;\Omega,\mathrm{Sym}^k(\mathbb{R}^2)^m)$ *satisfying*

   (a) $\|\operatorname{div}^l g\|_\infty \le \alpha_l$, $l = 0, ..., k-1$

   (b) $\mathrm{TGV}^{\mathrm{k}}_\alpha(\hat{u}) = -\int_\Omega \hat{u}\operatorname{div}^k g$

   (c) $$\begin{cases} (\operatorname{div}^k g, \tilde{a}_n)_{L^2} \ge 0 \ \text{ if}(A\hat{u})_n = o_n \ne l_n \\ (\operatorname{div}^k g, \tilde{a}_n)_{L^2} \le 0 \ \text{ if}(A\hat{u})_n = l_n \ne o_n \\ (\operatorname{div}^k g, \tilde{a}_n)_{L^2} = 0 \ \text{ if}(A\hat{u})_n \in \overset{\circ}{J}_n. \end{cases} \qquad \forall n\in\mathbb{N}$$

Proof. Existence of a solution follows from Proposition 4.1. Equivalence of (2) and (3) follows from biorthogonality of $(a_n)_{n\in\mathbb{N}}$ and $(\tilde{a}_n)_{n\in\mathbb{N}}$ (see proposition 1.7), so it is left to show equivalence of (1) and (2):

(1) ⇒ (2) : Let

$$\hat{u} = \operatorname*{arg\,min}_{u\in X} \left(\mathrm{TGV}^{\mathrm{k}}_\alpha(u) + \mathcal{I}_{U_D}(u)\right).$$

Thus $0 \in \partial \left(\mathrm{TGV}_\alpha^{\mathrm{k}}(\hat{u}) + \mathcal{I}_{U_D}(\hat{u})\right)$ and by additivity of the subdifferential for this setting (see Theorem 4.3) we have $0 \in \partial \left(\mathrm{TGV}_\alpha^{\mathrm{k}}(\hat{u})\right) + \partial \left(\mathcal{I}_{U_D}(\hat{u})\right)$. Hence there exist elements $z_1 \in \partial \left(\mathrm{TGV}_\alpha^{\mathrm{k}}(\hat{u})\right)$ and $z_2 \in \partial \left(\mathcal{I}_{U_D}(\hat{u})\right)$ such that $0 = z_1 + z_2$. Now by Theorem 4.1, $\hat{u} \in \mathrm{BV}(\Omega, \mathbb{R}^m)$ and there exists $g \in W_0^2(\mathrm{div}^k; \Omega, \mathrm{Sym}^k(\mathbb{R}^2)^m)$ satisfying (2)-(a) such that $z_1 = -\operatorname{div}^k g$ and $\mathrm{TGV}_\alpha^{\mathrm{k}}(\hat{u}) = -\int_\Omega \hat{u} \operatorname{div}^k g$. Hence we have

$$\operatorname{div}^k g = z_2.$$

By Theorem 4.2 there exists $\lambda = (\lambda_n)_{n \in \mathbb{N}} \in \ell^2$, satisfying the the element wise conditions in (2)-(a), such that $\operatorname{div}^k g = A^*\lambda$. Finally the characterization of A^* (see proposition 1.8) implies $\operatorname{div} g = \sum_{n \in \mathbb{N}} \lambda_n a_n$.

(2) ⇒ (1) : Conditions (a) and (b) together with $\hat{u} \in \mathrm{BV}(\Omega, \mathbb{R}^m)$ imply that $-\operatorname{div}^k g \in \partial \left(\mathrm{TGV}_\alpha^{\mathrm{k}}(\hat{u})\right)$ (Theorem 4.1), while (c) and (d) together with $\hat{u} \in U_D$ imply that $\operatorname{div}^k g \in \partial \left(\mathcal{I}_{U_D}(\hat{u})\right)$ (Theorem 4.2). Hence $0 \in \partial \left(\mathrm{TGV}_\alpha^{\mathrm{k}}(\hat{u})\right) + \partial \left(\mathcal{I}_{U_D}(\hat{u})\right) \subset \partial \left(\mathrm{TGV}_\alpha^{\mathrm{k}}(\hat{u}) + \mathcal{I}_{U_D}(\hat{u})\right)$ and $\hat{u}$ is a minimum.

□

Remark 4.4. *If an optimal solution $\hat{u}$ would be k times continuously differentiable and also $w_1, \ldots, w_{k-1}$ such that*

$$\mathrm{TGV}_\alpha^{\mathrm{k}}(\hat{u}) = \|\mathrm{D}\, u - w_1\|_{\mathcal{M}} + \sum_{l=1}^{k-2} \|\mathcal{E}(w_l) - w_{l+1}\|_{\mathcal{M}} + \|\mathcal{E} w_{k-1}\|_{\mathcal{M}} \tag{46}$$

would be sufficiently smooth, conditions (2)-(b) and (3)-(b) of theorem 4.4 would be equivalent to

- $\operatorname{div}^{k-1} g = \alpha_{k-1} \sigma_{\mathrm{D}\, u - w_1}$,
- $\operatorname{div}^{k-1-l} g = \alpha_{k-1-l} \sigma_{\mathcal{E} w_l - w_{l+1}}$ *for* $l = 1, \ldots k-2$,
- $g = \alpha_0 \sigma_{\mathrm{D}\, w_{k-1}}$,

where $\sigma_\phi = \frac{\phi}{|\phi|}$ whenever ϕ is not equal to zero and $|\sigma_\phi(x)| \leq 1$ for all $x \in \Omega$.

Proof. By iteratively applying a Gauss-Green theorem for $g \in W_0^2(\mathrm{div}; \Omega, \mathrm{Sym}^k(\mathbb{R}^d)^m)$ and adding and subtracting w_l one arrives at

$$\int_\Omega u \operatorname{div}^k g = \int_\Omega (\mathrm{D}\, u - w_1) \operatorname{div}^{k-1} g + \sum_{l=1}^{k-2} \int_\Omega (\mathcal{E} w_l - w_{l+1}) \operatorname{div}^{k-1-l} g + \int_\Omega \mathcal{E} w_{k-1} g.$$

Using the representation of $\mathrm{TGV}_\alpha^{\mathrm{k}}$ as in equation (46) as well as the fact that $\|\operatorname{div}^l g\|_\infty \leq \alpha_l$, $0 \leq l \leq k-1$, the claimed equivalence to (2)-(b) and (3)-(b) of theorem 4.4 follows. □

Remark 4.5. *Note that the equivalent condition given in remark 4.4 suggests that also for the* $\mathrm{TGV}_\alpha^{\mathrm{k}}$ *functional, a point-wise characterization of its subdifferential as obtained in [12] for the* TV *functional may be possible.*

5 Application to data reconstruction

The purpose of this section is to show how various models related to mathematical imaging problems, formulated both in discrete and in function space setting, are covered by the framework as derived in section 4. In the first subsection, we give some remarks about the general class of problems to which the theory of section 4 can be applied. Then, in the succeeding subsections, we will study the specific application to decompression and zooming problems in detail.

5.1 A general class of problems

The aim of this subsection is to describe a class of inverse problems, whose $\mathrm{TGV}_\alpha^\mathrm{k}$ regularization fits into the general framework of section 4: We will show that, under weak additional assumptions, any inverse problem, with the data described by interval restrictions on the coefficients of a operator with close range $B : L^2(\Omega, \mathbb{R}^m) \to \ell^2$, can be reformulated to fit in our general framework. The basis is the following proposition. We will assume that $\Omega \subset \mathbb{R}^2$ is a bounded Lipschitz domain throughout this subsection.

Proposition 5.1. *Let $B : L^2(\Omega, \mathbb{R}^m) \to \ell^2$ have closed range and, with $D \subset \ell^2$ a non-empty, closed, convex set, define*

$$U_D := \{u \in L^2(\Omega, \mathbb{R}^m) \,|\, Bu \in D\}.$$

Then there exist non-empty, closed intervals J_n, $n \in \mathbb{N}$, and a Riesz basis $(a_n)_{n\in\mathbb{N}}$ of $L^2(\Omega, \mathbb{R}^m)$ such that, with $A : L^2(\Omega, \mathbb{R}^m) \to \ell^2$ the basis transformation operator corresponding to $(a_n)_{n\in\mathbb{N}}$, we have

$$U_D = \{u \in L^2(\Omega, \mathbb{R}^m) \,|\, (Au)_i \in J_n \; \forall n \in \mathbb{N}\}.$$

Proof. Choose $(e_i)_{i\in N}$, with $N \subset \mathbb{N}$ an index set, to be an orthonormal basis of $\mathrm{Rg}(B)$. Further choose $(z_i)_{i\in\mathbb{N}\setminus N}$ an orthonormal basis of $\ker(B)$. With that, we define the sequences $(a_n)_{n\in\mathbb{N}}$, $(\tilde{a}_n)_{n\in\mathbb{N}}$ in $L^2(\Omega, \mathbb{R}^m)$ by

$$a_i = \begin{cases} B^* e_i & \text{if } i \in N, \\ z_i & \text{if } i \in \mathbb{N} \setminus N, \end{cases}$$

and

$$\tilde{a}_i = \begin{cases} B^{-1} e_i & \text{if } i \in N, \\ z_i & \text{if } i \in \mathbb{N} \setminus N, \end{cases}$$

where $B^{-1} : \mathrm{Rg}(B) \to \ker(B)^\perp$ denotes the continuous inverse of $B : \ker(B)^\perp \to \mathrm{Rg}(B)$. We want to show that $(a_i)_{i\in\mathbb{N}}$ and $(\tilde{a}_i)_{i\in\mathbb{N}}$ are biorthogonal Riesz bases. For this purpose, according to [72, Theorem 1.9], it suffices to show that $(a_i)_{i\in\mathbb{N}}$ and $(\tilde{a}_i)_{i\in\mathbb{N}}$ are both dense, biorthogonal and that, for any $f \in L^2(\Omega, \mathbb{R}^m)$,

$$\sum_{n\in\mathbb{N}} |(f, a_n)_{L^2}|^2 < \infty, \quad \sum_{n\in\mathbb{N}} |(f, \tilde{a}_n)_{L^2}|^2 < \infty.$$

Concerning density, suppose that, for $w_1, w_2 \in L^2(\Omega, \mathbb{R}^m)$ arbitrary, $(a_n, w_1)_{L^2} = 0$ and $(\tilde{a}_n, w_2)_{L^2} = 0$ for all $n \in \mathbb{N}$. Given that $(z_i)_{i\in\mathbb{N}\setminus N}$ is a basis for $\ker(B)$, this immediately implies $w_1, w_2 \in \ker(B)^\perp$. But $0 = (a_n, w_1)_{L^2} = (e_n, Bw_1)_{\ell^2}$ for all $n \in N$ implies that $Bw_1 = 0$, thus $w_1 \in \ker(B)$ and $w_1 = 0$. Similar, $0 = (\tilde{a}_n, w_2)_{L^2} = (B^{-1}e_n, w_2)_{L^2}$ for all $n \in N$ implies, by surjectivity of $B^{-1} : \operatorname{Rg}(B) \to \ker(B)^\perp$, that also $w_2 = 0$. Thus both sequences are dense. Take now $i, j \in \mathbb{N}$. Then

$$(a_i, \tilde{a}_j)_{L^2} = \left.\begin{cases} (B^* e_i, B^{-1}e_j)_{L^2} & \text{if } i \in N, j \in N \\ (B^* e_i, z_j)_{L^2} & \text{if } i \in N, j \in \mathbb{N} \setminus N \\ (z_i, B^{-1}e_j)_{L^2} & \text{if } i \in \mathbb{N} \setminus N, j \in N \\ (z_i, z_j)_{L^2} & \text{if } i \in \mathbb{N} \setminus N, j \in \mathbb{N} \setminus N \end{cases}\right\} = \delta_{i,j},$$

where we used that $BB^{-1}e_j = e_j$, $Bz_j = 0$, $B^{-1}e_j \in \ker(B)^\perp$ and the fact that $(e_i)_{i \in N}, (z_i)_{i \in \mathbb{N}\setminus N}$ are orthonormal bases. To show the remaining assertion, take any $f \in L^2(\Omega, \mathbb{R}^m)$ and $M \in \mathbb{N}$. Then

$$\sum_{n \in \mathbb{N}}^{M} |(f, a_n)_{L^2}|^2 \le \sum_{n \in N} |(Bf_1, e_n)_{\ell^2}|^2 + \sum_{n \in \mathbb{N}\setminus N} |(f_2, z_n)_{L^2}|^2 \le \|Bf\|_{L^2}^2 + \|f_2\|_{L^2}^2$$

and similar

$$\sum_{n \in \mathbb{N}}^{M} |(f, \tilde{a}_n)_{L^2}|^2 \le \sum_{n \in N} |(B^{-*}f_1, e_n)_{\ell^2}|^2 + \sum_{n \in \mathbb{N}\setminus N} |(f_2, z_n)_{L^2}|^2 \le \|B^{-*}f\|_{L^2} + \|f_2\|_{L^2}$$

with $f = f_1 + f_2 \in \ker(B)^\perp \otimes \ker(B)$. Thus also the limit $M \to \infty$ remains bounded and with that, $(a_n)_{n \in \mathbb{N}}, (\tilde{a}_n)_{n \in \mathbb{N}}$ are biorthogonal Riesz bases.

At last we have to show existence of closed intervals J_n, $n \in \mathbb{N}$, such that, defining

$$\tilde{U}_D = \{u \in L^2(\Omega, \mathbb{R}^m) | (Au)_i \in J_i,\, i \in \mathbb{N}\},$$

provides the identity $\tilde{U}_D = U_D$. Defining $\tilde{D} = A(U_D)$ and $\tilde{U}_D = \{u \in L^2(\Omega, \mathbb{R}^m) \,|\, Au \in \tilde{D}\}$ it follows by injectivity of A that

$$u \in \tilde{U}_D \Leftrightarrow Au \in \tilde{D} \Leftrightarrow \exists v \in U_D : Au = Av \Leftrightarrow u \in U_D.$$

Since U_D is nonempty, closed and convex, also $\tilde{D}$ is a nonempty, closed and convex subset of ℓ^2, thus can be described by intervals as claimed. With that, the assertion is proven.

□

Proposition 5.1 now allows us, under mild assumptions, to reformulate any minimization problem of the form

$$\inf_{u \in L^2(\Omega, \mathbb{R}^m)} \mathrm{TGV}_\alpha^\mathrm{k}(u) + \mathcal{I}_{U_D}(u)$$

where

$$U_D = \{u \in L^2(\Omega, \mathbb{R}^m) | Bu \in D\},$$

with B a continuous, linear operator with closed range and $D \subset \ell^2$ a suitable data set, in a way that it fits to assumption (A). Further, given the operator B, we can directly construct the biorthogonal Riesz basis and this gives the basis transformation operator A. What is left to assure are the assumptions on the data set D and that the dual basis of the resulting Riezs-basis indeed is contained in $\mathrm{BV}(\Omega, \mathbb{R}^m)$.

Since, as already mentioned, we are mainly interested in cases where the Riesz basis and the corresponding transformation operator are defined component wise as in remark 1.2, let us simplify notation as follows:

Given any linear operator $B : L^2(\Omega, \mathbb{R}^m) \to \ell^2$ and operators $B_j : L^2(\Omega) \to \ell^2$, $1 \leq j \leq m$, such that, for $u = (u_1, \ldots, u_m) \in L^2(\Omega, \mathbb{R}^m)$, and any $n \in \mathbb{N}$

$$(Bu)_n = (B_i u_i)_{\tilde{n}} \text{ for suitable } i, \tilde{n} \text{ depending on n},$$

we can always reorder and assume that B is given as $B : L^2(\Omega, \mathbb{R}^m) \to (\ell^2)^m$ with $(Bu)^j_n = (B_j u_j)_n$. In particular, any basis transformation operator A as in remark 1.2 can be assumed to map from $L^2(\Omega, \mathbb{R}^m)$ to $(\ell^2)^m$, and be defined by $(Au)^j_n = (a^j_n, u_j)_{L^2}$, with $(a^j_n)_{n\in\mathbb{N}}$, $1 \leq j \leq m$ the component wise Riesz bases.

Considering this, together with proposition 5.1, the following corollary shows the construction of such a Riesz basis transformation operator from an operator with closed range in a special case.

Corollary 5.1. *Suppose that, in the setting of proposition 5.1, the operator B is given by $B : L^2(\Omega, \mathbb{R}^m) \to (\ell^2)^m$,*

$$(Bu)^j_n = (B_j u_j)_n$$

with $B_j = C_j M_j$, where

$$C_j : L^2(\Omega_j) \to \ell^2$$

are a basis transformation operators related to a orthonormal bases $(c^j_i)_{i\in\mathbb{N}}$ of $L^2(\Omega_j)$, Ω_j are bounded Lipschitz domains and

$$M_j : L^2(\Omega) \to L^2(\Omega_j)$$

are a surjective, continuous linear operators. Then the B_j are also surjective, in particluar have closed range, and the Riesz bases $(a^j_n)_{n\in\mathbb{N}}$, $(\tilde{a}^j_n)_{n\in\mathbb{N}}$ as in Proposition 5.1 can be given by

$$a^j_n = \begin{cases} M^*_j c^j_n & n \in N^j, \\ z^j_n & n \in \mathbb{N} \setminus N^j, \end{cases}$$

and

$$\tilde{a}^j_n = \begin{cases} M^{-1}_j c^j_n & n \in N^j, \\ z^j_n & n \in \mathbb{N} \setminus N^j, \end{cases}$$

with $N^j \subset \mathbb{N}$ suitable index sets and $(z^j_n)_{n\in\mathbb{N}\setminus N^j}$ arbitrary orthonormal bases of $\ker(M_j)$.

The data set U_D, originally defined by

$$U_D := \{u \in L^2(\Omega, \mathbb{R}^m) \,|\, (Bu)_n \in J^j_n \quad \forall 1 \leq j \leq m, n \in \mathbb{N}\},$$

can then be described as

$$U_D = \{u \in L^2(\Omega, \mathbb{R}^m) \,|\, (M^*_j c^j_n, u_j)_{L^2} \in \tilde{J}^j_n \quad \forall\, 1 \leq j \leq m, n \in \mathbb{N}\}. \tag{47}$$

with

$$\tilde{J}_n^j = \begin{cases} J_n^j & n \in N^j, \\ \mathbb{R} & \textit{else}, \end{cases}$$

$1 \leq j \leq m$.

If further $M_j^{-1} = \lambda_j M_j^*$, *with* $\lambda_j \in \mathbb{R}$, *then the* $(a_n^j)_{n\in\mathbb{N}}$, $(\tilde{a}_n^j)_{n\in\mathbb{N}}$ *are orthogonal for all* $1 \leq j \leq m$.

Proof. Given that all B_j are surjective, we can choose $(e_{n_i}^j)_{i\in N^j}$, with N^j countable index sets, to be the standard bases of ℓ^2, i.e. $(e_i^j)_l = \delta_{i,l}$, and immediately obtain the bases $(a_i)_{i\in\mathbb{N}}, (\tilde{a}_i)_{i\in\mathbb{N}}$ as claimed. The representation of U_D as in (47) follows from

$$(Bu)_i^j = (B_j u_j, e_{n_{\pi(i)}}^j)_{\ell^2} = (u_j, M_j^* c_{n_{\pi(i)}}^j)_{L^2},$$

with $\pi : \mathbb{N} \to N$ a suitable, surjective map, for all $1 \leq j \leq m$, $i \in \mathbb{N}$. Orthogonality in the case that $M_j^{-1} = \lambda_j M_j^*$ is immediate. □

The particular case described at the end of corollary 5.1 will later be relevant in the application to JPEG decompression, where the transformation operator is a subsampling operator decomposed with a blockwise cosine transformation.

The following lemma shows that, for this special case, a projection onto the data set can be as easily performed as a projection to a data set solely described by an orthonormal basis transformation operator. This will be very convenient for the numerical solution of the application to JPEG decompression later on.

Lemma 5.1. *Let* H_1, H_2 *be two Hilbert spaces and* $M : H_1 \to H_2$ *a linear operator such that, for all* $v \in H_2$, $MM^*v = cv$ *with* $c \in \mathbb{R}_{>0}$. *With* $A : H_2 \to \ell^2$ *a basis transformation operator and* $I \subset \ell^2$ *a given data set, define*

$$U_d = \{u \in H_1 \,|\, AMu \in I\}$$

and

$$U_A = \{v \in H_2 \,|\, Av \in I\}.$$

Then we get that for all $\hat{u} \in H_1$,

$$P_{U_d}(\hat{u}) = \hat{u} + \frac{1}{c} M^* \left(P_{U_A}(M\hat{u}) - M\hat{u}\right)$$

with P_S *the projection operator on a given set* S.

Proof. Take $\hat{u} \in H_1$. Define $D := \frac{1}{c}M^*$. Then obviously

$$AM(\hat{u} + D(P_{U_A}(M\hat{u}) - M\hat{u}) = A(P_{U_A}(M\hat{u})) \in I,$$

and thus it is left to show that

$$(\hat{u} - \hat{u} - D\left(P_{U_A}(M\hat{u}) - M\hat{u}\right), u - \hat{u} - D\left(P_{U_A}(M\hat{u}) - M\hat{u}\right))_{H_1} \leq 0$$

for all $u \in U_d$. Note that we have

$$(M\hat{u} - P_{U_A}(M\hat{u}), v - P_{U_A}(M\hat{u}))_{H_2} \leq 0 \quad \text{for all } v \in U_A.$$

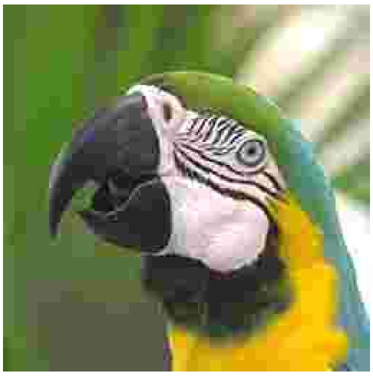


Figure 4: JPEG image with typical blocking and ringing artifacts.

Using this, for all $u \in U_d$, we get

$$\begin{aligned} & (\hat{u} - \hat{u} - D\left(P_{U_A}(M\hat{u}) - M\hat{u}\right), u - \hat{u} - D\left(P_{U_A}(M\hat{u}) - M\hat{u}\right))_{H_1} \\ = \; & \frac{1}{c}(M^*\left(M\hat{u} - P_{U_A}(M\hat{u})\right), u - \hat{u} - D\left(P_{U_A}(M\hat{u}) - M\hat{u}\right))_{H_1} \\ = \; & \frac{1}{c}(M\hat{u} - P_{U_A}(M\hat{u}), Mu - P_{U_A}(M\hat{u}))_{H_2} \leq 0, \end{aligned}$$

since $Mu \in U_A$ for $u \in U_d$. □

Having discussed the general applicability of our model and already given some useful remarks, we now turn to the concrete application to problems in mathematical imaging.

5.2 Color JPEG decompression

As first application, let us consider the problem of artifact-free decompression of JPEG compressed color images. This problem has already been addressed in various publications, of which the TV-based models of [13, 2] are most related. Also, a discrete version of the problem using TGV regularization has already been published in [14]. We further refer to [13, 53, 63, 62] for a short overview of current standard techniques.

We start with a brief explanation of the basic steps of the JPEG compression standard. For further information about our modeling we refer to [13, 14] and for a more detailed explanation of the JPEG compression procedure to [68].

The process of JPEG compression is lossy, which means that typically most of the compression is obtained by loss of data. As a consequence, the original image cannot be restored completely from the compressed object, which causes ringing and blocking artifacts in the reconstructed images, as can be seen for example in figure 4. Figure 5 gives an overview of the basic steps of JPEG compression for color images that are important for our reconstruction framework. In particular, a further lossless coding of integer data is omitted here, since this procedure can be inverted without loss of data.

A color JPEG image is typically processed in the YCbCr color space, where the first (luminance) component essentially contains the brightness information and the second two (chroma) components the color information of the image.

This color space is equivalent to the standard RGB color space and images can be transformed from one to another without significant loss of data. The advantage of using the YCbCr color space is the following: Knowing that the human visual system is less sensitive to color than to brightness oscillations, as first step of JPEG compression, data reduction can be achieved by subsampling the two chroma components.

Next, each component undergoes a discrete cosine transformation on each block of 8×8 pixels, resulting in a local representation of the components as linear combination of different frequencies (see also figure 6). Again, empirical information suggests that the human visual system is less sensitive to high frequency variations than to low frequency variations. Consequently, the image is quantized by pointwise division of each 8×8 pixel block by a predefined quantization matrix reflecting this empirical observation. The resulting data is then rounded to integer and, after further lossless compression, stored in the compressed JPEG object.

In order to reconstruct an image from the compressed file, standard decompression algorithms now simply revert the compression process by dequantization, application of the inverse blockwise cosine transform and color upsampling. It is thereby not taken into account that the data is incomplete, i.e. that it is a result of a rounding procedure, and thus does not uniquely determine a source image, but a set of possible source images. Indeed, since, besides the quantized coefficient data $d = (d^c_{i,j})$, also the quantization matrix $Q = (Q^c_{i,j})$ can be obtained from the compressed file, it is even possible to define a maximal-error interval

$$J^c_{i,j} = \left[Q^c_{i,j}(d^c_{i,j} - \frac{1}{2}), Q^c_{i,j}(d^c_{i,j} + \frac{1}{2})\right] \tag{48}$$

for each quantized coefficient, and with that a convex set of possible source data

$$D = \{(z^c_{i,j}) \,|\, z^c_{i,j} \in J^c_{i,j} \text{ for all } i, j, c\}. \tag{49}$$

With that, D is the set of all coefficients that would, after the quantization and rounding procedure, result in the same data as given by the JPEG compressed file. Note that here, i and j define the vertical and horizontal pixel number, respectively, while $c \in \{1, 2, 3\}$ defines the color component.

Coupling the subsampling S and the cosine transformation operator C, as in figure 5, we will see that with this the set of all possible source images of the compressed JPEG object can be described by D and an (even orthogonal) basis transformation operator. Thus it fits in our image reconstruction framework, where we want to choose one of all possible source images that minimizes the $\mathrm{TGV}^{\mathrm{k}}_{\alpha}$ functional.

5.2.1 Modeling

We consider color images as functions in $L^2(\Omega, \mathbb{R}^3)$, where $\Omega = (0, 8k) \times (0, 8l)$, $k, l \in \mathbb{N}$ is a rectangle domain, in particular a Lipschitz domain.

Subsampled image components are considered as functions in $L^2(\Omega_j)$, where $\Omega_j = (0, 8k_j) \times (0, 8l_j)$ are domains smaller than Ω, i.e. $k_j \leq k$, $l_j \leq l$. With that, the subsampling process can be described color component wise via the operators $S_j : L^2(\Omega) \to L^2(\Omega_j)$, $j \in \{1, 2, 3\}$, given by

$$S_j u(x_1, x_2) = u(s_j x_1, t_j x_2),$$

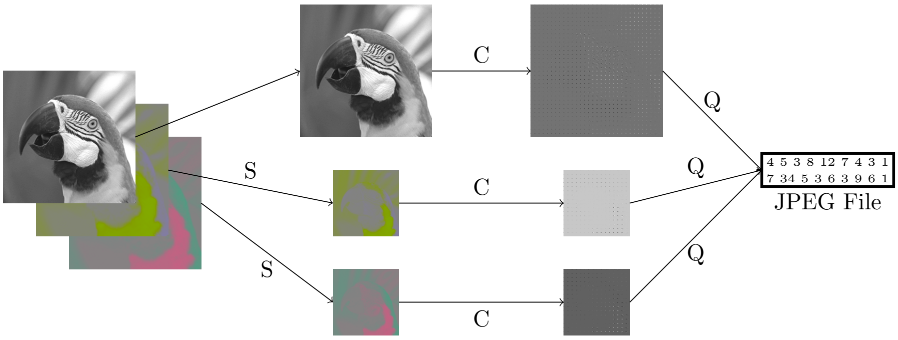


Figure 5: Scheme of JPEG compression procedure. Here, S denotes a subsampling operation, C a blockwise discrete cosine transformation and Q a quantization to integer, i.e. a blockwise devision through a predefined quantization matrix followed by rounding to integer.

where $s_j = \frac{k}{k_j}, t_j = \frac{l}{l_j}$ the subsampling factors.

Remark 5.1. *Note by defining a subsampling operator also for the luminance component we simplify the notation, but also implicitly generalize the model to allow a simple form of image zooming by considering a subsampling factor large than one for the luminance-, and accordingly an increased factor for the chroma components.*

In order to define the blockwise cosine transform, we first need the following definition, which can also be found in [13]:

Definition 5.1 (Blockwise cosine system)**.** *For $t, r \in \mathbb{N}$, set $G = (0, 8t) \times (0, 8r) \subset \mathbb{R}^2$. For $i, j \in \mathbb{N}_0$, $0 \le i < t$, $0 \le j < r$ we define the squares*

$$E_{i,j} = \big([8i, 8i+8) \times [8j, 8j+8) \big) \cap G$$

and

$$\chi_{i,j} = \chi_{E_{i,j}}$$

their characteristic functions. Furthermore, let the standard cosine orthonormal system $(b_{n,m})_{n,m \ge 0} \subset L^2((0,1)^2)$ be defined as

$$b_{n,m}(x,y) = \lambda_n \lambda_m \cos(nx\pi) \cos(my\pi), \tag{50}$$

for $(x,y) \in \mathbb{R}^2$, where

$$\lambda_s = \begin{cases} 1 & \text{if } s = 0, \\ \sqrt{2} & \text{if } s \neq 0. \end{cases}$$

With that, we define the blockwise cosine system $c_{n,m}^{i,j} \in L^2(G)$ as

$$c_{n,m}^{i,j}(x,y) = \frac{1}{8} b_{n,m}\left(\frac{x - 8i}{8}, \frac{y - 8j}{8} \right) \chi_{i,j}(x,y) \tag{51}$$

for $(x,y) \in G$.

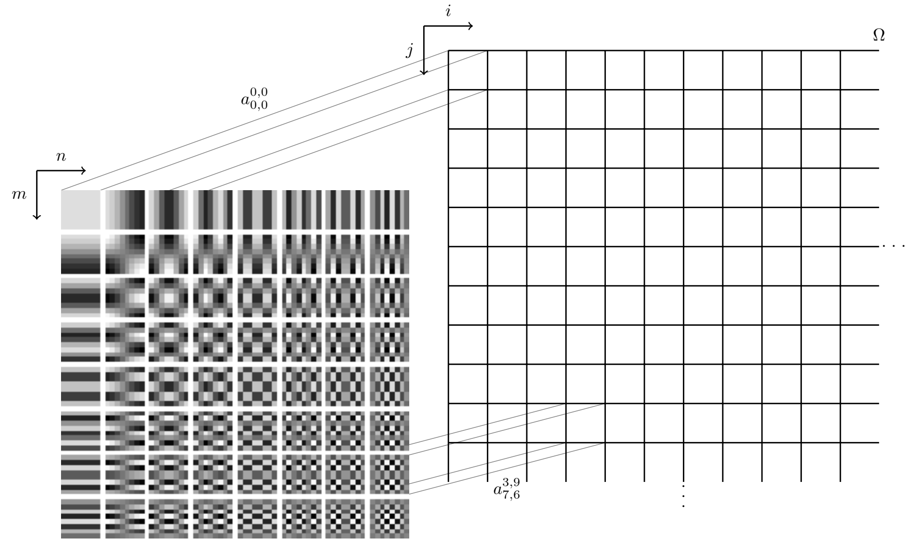


Figure 6: Illustration of the finite dimensional blockwise cosine orthonormal basis used for JPEG decompression. As indicated, the basis function $a^{0,0}_{0,0}$ corresponds to the $(0,0)$ (upper left) 8×8-pixel block of the image and the $(0,0)$ frequency, while the basis function $a^{3,9}_{7,6}$ corresponds to the $(3,9)$ pixel block of the image and the $(7,6)$ frequency.

Remark 5.2. *It follows by reduction to the cosine-orthonormal system* $(b_{n,m})_{n,m\geq 0}$ *that* $\{c^{i,j}_{n,m} \,|\, n,m \in \mathbb{N}_0,\ 0 \leq i < k, 0 \leq j < l\}$ *is a complete orthonormal system in* $L^2(G)$. *Further, one can immediately see that* $\{c^{i,j}_{n,m} \,|\, n,m \in \mathbb{N}_0,\ 0 \leq i < k, 0 \leq j < l\} \subset \mathrm{BV}(G)$.

To illustrate the correspondence of the blockwise cosine system to a linear combination of the image by different frequencies, the finite dimensional equivalent of the system as in definition 5.1 is shown in figure 6.

Denoting, for $j \in \{1,2,3\}$, by $(c^j_{n,m})_{n,m\geq 0}$ a blockwise cosine orthonormal systems of $L^2(\Omega_j)$ as described in definition 5.1 (note that we use a different index notation), the operators $C_j : L^2(\Omega_j) \to \ell^2$ are defined to be their corresponding basis transformation operators, i.e.

$$(C_j v)_{n,m} = (c^j_{n,m}, v)_{L^2},$$

for $v \in L^2(\Omega_j)$.

With these preliminaries, we define the operator modeling the color JPEG compression procedure as $A = (A_1, A_2, A_3)$ where $A_j : L^2(\Omega) \to \ell^2$ is defined as $A_j = C_j S_j$. Given a data set $(J^j_{n,m})_{nm\geq 0}$, where each $J^j_{n,m} = [l^j_{n,m}, o^j_{n,m}]$ is a closed interval reflecting the possible range of the corresponding coefficient (see equation (48)), the set of possible source images can now be described as

$$U_D = \{u \in L^2(\Omega, \mathbb{R}^3) \,|\, (Au)^j_{nm} \in J^j_{nm} \text{ for } n,m \in \mathbb{N}_0, j \in \{1,2,3\}\}. \qquad (52)$$

Clearly, the range of the A_j is closed and since the S_j are bijective with $S_j^{-1} = s_j t_j S_j^*$, according to corollary 5.1, A is a basis transformation operator related to basis elements (a^j_{nm}), that can be given as

$$a^j_{nm}(x_1,x_2) = S^*_j c^j_{nm}(x_1,x_2) = \frac{1}{s_j t_j} c^j_{nm}(\frac{x_1}{s_j}, \frac{x_2}{t_j}), \tag{53}$$

and are orthogonal and contained in $\mathrm{BV}(\Omega,\mathbb{R}^3)$ (see [13, Remark 2]). With that, the continuous minimization problem corresponding to color JPEG decompression is given by

$$\min_{u\in L^2(\Omega,\mathbb{R}^3)} \mathrm{TGV}^{\mathrm{k}}_{\alpha}(u) + \mathcal{I}_{U_D}(u), \tag{54}$$

with U_D defined in equation (52) and $A = (A_1, A_2, A_3)$ the basis transformation operator corresponding the the orthogonal basis as in (53).

For this setting, assumption (A) and (EX_k) are clearly satisfied since, due to the rounding procedure in JPEG compression, all J^j_{nm} are bounded with length greater or equal to one. Thus, existence of a solution and validity of the optimality condition as in theorem 4.4 follows.

5.2.2 Discrete framework

Given the minimization problem (54) for artifact free JPEG decompression, defined in function space setting, the purpose of this subsection is to introduce its discrete equivalent and show how a solution can be obtained numerically.

Some discrete formulations for this problem have already been presented in [14], however, a more extensive treatment of this setting is presented in the following. Since for the rest of this subsection, all formulations are solely in the discrete setting, we abuse notation by using the same notation as for the continuous model.

As a good compromise between computational complexity and obtained image quality, we will now restrict ourselves to the second order $\mathrm{TGV}^{\mathrm{k}}_{\alpha}$ functional, denoted by TGV^2_{α}, knowing that the generalization to arbitrary order is possible within the same framework.

We define $U := \mathbb{R}^{8k\times 8l\times 3}, k, l \in \mathbb{N}$, to be the space of discrete images and $V := U^2, W := U^3$ the spaces for first and second order gradient information. Denoting the elements $u \in U$ by

$$u = (u^c_{i,j})_{\substack{0\leq i,j<8k,8l \\ c\in\{1,2,3\}}} = \left(\begin{pmatrix} u^1 \\ u^2 \\ u^3 \end{pmatrix}_{i,j} \right)_{0\leq i,j<8k,8l},$$

a norm on their entries is defined by

$$\left| \begin{pmatrix} u^1 \\ u^2 \\ u^3 \end{pmatrix} \right|^2_U = (u^1)^2 + (u^2)^2 + (u^3)^2.$$

With that, a norm on U, resulting from an inner product $(\cdot,\cdot)_U$, is given by

$$\|u\|^2_U = (u,u)_U = \sum_{0\leq i,j<8k,8l} |u_{i,j}|^2_U. \tag{55}$$

Similarly, elements $v \in V$ and $w \in W$ are denoted by

$$v = (v_{i,j}^{\lambda,c})_{\substack{0 \leq i,j < 8k,8l \\ \lambda \in \{1,2\} \\ c \in \{1,2,3\}}} = \left(\begin{pmatrix} v^{1,1} & v^{2,1} \\ v^{1,2} & v^{2,2} \\ v^{1,3} & v^{2,3} \end{pmatrix}_{i,j} \right)_{0 \leq i,j < 8k,8l}$$

and

$$w = (w_{i,j}^{\lambda,c})_{\substack{0 \leq i,j < 8k,8l \\ \lambda \in \{1,2,3\} \\ c \in \{1,2,3\}}} = \left(\begin{pmatrix} w^{1,1} & w^{2,1} & w^{3,1} \\ w^{1,2} & w^{2,2} & w^{3,2} \\ w^{1,3} & w^{2,3} & w^{3,3} \end{pmatrix}_{i,j} \right)_{0 \leq i,j < 8k,8l},$$

and norms on their entries by

$$\left| \begin{pmatrix} v^{1,1} & v^{2,1} \\ v^{1,2} & v^{2,2} \\ v^{1,3} & v^{2,3} \end{pmatrix} \right|_V^2 = \sum_{\substack{\lambda \in \{1,2\} \\ c \in \{1,2,3\}}} (v^{\lambda,c})^2$$

and

$$\left| \begin{pmatrix} w^{1,1} & w^{2,1} & w^{3,1} \\ w^{1,2} & w^{2,2} & w^{3,2} \\ w^{1,3} & w^{2,3} & w^{3,3} \end{pmatrix} \right|_W^2 = \sum_{\substack{\lambda \in \{1,2\} \\ c \in \{1,2,3\}}} (w^{\lambda,c})^2 + 2 \sum_{c \in \{1,2,3\}} (w^{3,c})^2.$$

The norms on V and W, also resulting from inner products $(\cdot,\cdot)_V$ and $(\cdot,\cdot)_W$, are then given by

$$\|v\|_V^2 = (v,v)_V = \sum_{0 \leq i,j < 8k,8l} |v_{i,j}|_V^2 \quad \text{and} \quad \|w\|_W^2 = (w,w)_W = \sum_{0 \leq i,j < 8k,8l} |w_{i,j}|_W^2. \tag{56}$$

Based on its equivalent formulation as in proposition 3.6, we define the discrete total generalized variation functional of second order as

$$\mathrm{TGV}_\alpha^2(u) = \inf_{v \in V} \alpha_1 \|\nabla(u) - v\|_1 + \alpha_0 \|\mathcal{E}(v)\|_1, \tag{57}$$

where $\nabla : U \to V$ denotes a discrete, color component wise gradient operator using forward differences and $\mathcal{E} : V \to W$ denotes a discrete, color component wise symmetric gradient operator using backward differences, i.e.,

$$(\nabla u)_{i,j} = \begin{pmatrix} (\delta_{x+}u^1)_{i,j} & (\delta_{y+}u^1)_{i,j} \\ (\delta_{x+}u^2)_{i,j} & (\delta_{y+}u^2)_{i,j} \\ (\delta_{x+}u^3)_{i,j} & (\delta_{y+}u^3)_{i,j} \end{pmatrix}, \tag{58}$$

for $u \in U$, $0 \leq i < 8k, 0 \leq j < 8l$, with

$$\delta_{x+}, \delta_{y+} : \mathbb{R}^{8k \times 8l} \to \mathbb{R}^{8k \times 8l},$$

$$\begin{aligned} (\delta_{x+}z)_{i,j} &= \begin{cases} (z_{i+1,j} - z_{i,j}) & \text{if} \quad 0 \leq i < 8k-1, \\ 0 & \text{if} \quad i = 8k-1, \end{cases} \\ (\delta_{y+}z)_{i,j} &= \begin{cases} (z_{i,j+1} - z_{i,j}) & \text{if} \quad 0 \leq j < 8l-1, \\ 0 & \text{if} \quad j = 8l-1, \end{cases} \end{aligned} \tag{59}$$

and

$$(\mathcal{E}v)_{i,j} = \begin{pmatrix} (\delta_{x-}v^{1,1})_{i,j} & (\delta_{y-}v^{2,1})_{i,j} & (\frac{\delta_{y-}v^{1,1}+\delta_{x-}v^{2,1}}{2})_{i,j} \\ (\delta_{x-}v^{1,2})_{i,j} & (\delta_{y-}v^{2,2})_{i,j} & (\frac{\delta_{y-}v^{1,2}+\delta_{x-}v^{2,2}}{2})_{i,j} \\ (\delta_{x-}v^{1,3})_{i,j} & (\delta_{y-}v^{2,3})_{i,j} & (\frac{\delta_{y-}v^{1,3}+\delta_{x-}v^{2,3}}{2})_{i,j} \end{pmatrix}_{i,j}, \tag{60}$$

for $v \in V$, $0 \leq i < 8k, 0 \leq j < 8l$, with

$$\delta_{x-}, \delta_{y-} : \mathbb{R}^{8k\times 8l} \to \mathbb{R}^{8k\times 8l},$$

$$\begin{aligned} (\delta_{x-}z)_{i,j} &= \begin{cases} -z_{i-1,j} & \text{if} \quad i = 8k-1, \\ (z_{i,j} - z_{i-1,j}) & \text{if} \quad 0 < i < 8k-1, \\ z_{i,j} & \text{if} \quad i = 0, \end{cases} \\ (\delta_{y-}z)_{i,j} &= \begin{cases} -z_{i,j-1} & \text{if} \quad j = 8l-1, \\ (z_{i,j} - z_{i,j-1}) & \text{if} \quad 0 < j < 8l-1, \\ z_{i,j} & \text{if} \quad j = 0. \end{cases} \end{aligned} \tag{61}$$

Note that usage of forward and backward differences for the gradient ∇ and the symmetrized gradient $\mathcal{E}$, respectively, results in second order central differences for $\mathcal{E} \circ \nabla$. Further, in $\mathcal{E}(v)$, the off-diagonal entries need to be stored only once, thus $\mathcal{E}(v) \in W = U^3$. The norms $\|\cdot\|_1$ in (57) are discrete L^1 norms and, depending on their input, defined as

$$\|v\|_1 = \sum_{i,j} |v_{i,j}|_V \quad \|w\|_1 = \sum_{i,j} |w_{i,j}|_V,$$

for $v \in V$, $w \in W$. At this point, as discussed in remark 2.1, again other choices than the Frobenius-type norms $|\cdot|_V$ and $|\cdot|_W$ for the entries $v_{i,j} \in \mathbb{R}^{3\times 2}$ and $w_{i,j} \in \mathbb{R}^{3\times 3}$ are possible, maybe leading to better image quality. But again, as good balance between image quality and computational effort, we will only use this norm in our experiments and refer to [9, Remark 2] and [41] for a discussion on color space norms.

In order to avoid extensive indexing, we will give just a local, component wise definition of the discrete versions of the operators S and C, necessary to describe the discrete data set U_D similar to (52). For subsampling factors f_1, f_2 of the horizontal and vertical directions of one component, the subsampling operator S is defined component wise, locally for $(z_{i,j})_{0\leq i,j<f_1,f_2}$, as

$$Sz = \frac{1}{f_1 f_2} \sum_{m,n=0}^{f_1-1,f_2-1} z_{m,n}, \tag{62}$$

reducing the resolution of the component by factors f_1 and f_2 in the vertical and horizontal direction, respectively. If the resolution of one component is not reduced, as typically for the brightness component, S is the identity for this component. The discrete cosine transformation operator is defined, for each color component, on each disjoint 8×8 block $(z_{i,j})_{0\leq i,j\leq 7}$, as

$$(Cz)_{p,q} \quad = \quad c_p c_q \sum_{n,m=0}^{7} z_{n,m} \cos\left(\frac{\pi(2n+1)p}{16}\right) \cos\left(\frac{\pi(2m+1)q}{16}\right), \tag{63}$$

for $0 \leq p, q \leq 7$ and

$$c_s = \begin{cases} \frac{1}{\sqrt{8}} & \text{if } s = 0, \\ \frac{1}{2} & \text{if } 1 \leq s \leq 7. \end{cases}$$

Having this, the discrete data set U_D can now be defined by

$$U_D = \{u \in U \,|\, CSu \in D\}, \tag{64}$$

where the coefficient data set D depends on the quantization process and can be obtained from the compressed JPEG file as in (49) (see also [13, Section 5]).

With these prerequisites, the finite dimensional optimization problem for artifact-free JPEG decompression reads as

$$\min_{u \in U} \mathrm{TGV}_\alpha^2(u) + \mathcal{I}_{U_D}(u), \tag{65}$$

where

$$\mathcal{I}_{U_D}(u) = \begin{cases} 0 & \text{if } u \in U_D, \\ \infty & \text{else.} \end{cases}$$

Discrete Existence and optimality

This subsection is devoted to argue existence of a solution also for the discrete problem and formulate an equivalent saddle point problem. We define X and Y as

$$X := U \times V, \quad Y := V \times W,$$

equipped with the norms

$$\|(u,v)\|_X^2 = \|u\|_U^2 + \|v\|_V^2, \quad \|(u,v)\|_Y^2 = \|u\|_V^2 + \|v\|_W^2,$$

This leads to inner product spaces X and Y, which will be of relevance in the numerical solution later on.

As first step, the following proposition states existence for the discrete setting and a straightforward equivalence.

Proposition 5.2. *The discrete minimization problem* (65) *related to artifact free JPEG decompression possesses a solution and is equivalent to*

$$\min_{(u,v) \in X} F(K(u,v)) + \mathcal{I}_{U_D}(u), \tag{66}$$

where $K : X \to Y$*,*

$$K = \begin{bmatrix} \nabla & -\mathrm{I} \\ 0 & \mathcal{E} \end{bmatrix} \tag{67}$$

is a gradient-type operator, $F : Y \to \mathbb{R}$*,*

$$F(v,w) = \alpha_1 \|v\|_1 + \alpha_0 \|w\|_1, \tag{68}$$

a discrete L^1*-norm and* $I : V \to V$ *the identity map.*

Proof of proposition 5.2. Clearly the objective functional of (65) is proper, thus we can pick a minimizing sequence, contained in U_D.

Denote $\tilde{u}_n = u_n - P_1(u_n)$, where P_1 is the projection onto discrete functions u such that $\mathcal{E}(\nabla u)) = 0$, i.e. the space of component wise discrete polynomials of order less than 2. At first we bound $\|\tilde{u}_n\|_1$ by showing that there exists a constant $C > 0$ such that

$$\|w\|_1 \leq C \operatorname{TGV}_\alpha^2(w) \quad \text{for all } w \in \ker(P_1).$$

Assume that this is not the case, then there exists sequence $(w_n)_{n\in\mathbb{N}}$ in $\ker(P_1)$ with $\|w_n\|_1 = 1$ such that $1 \geq n \operatorname{TGV}_\alpha^2(w_n)$. Boundedness of $(w_n)_{n\in\mathbb{N}}$ together with continuity of $\operatorname{TGV}_\alpha^2$ thus yields an element $w \in \ker(P_2)$, $\|w\|_1 = 1$ such that $\operatorname{TGV}_\alpha^2(w) = 0$. But this implies $\mathcal{E}(\nabla w) = 0$ and thus $w = 0$, which contradicts that $\|w\|_1 = 1$. Thus $\tilde{u}_n$ is bounded and it is left to bound $P_1(u_n)$. This expression can be written component wise as

$$P_1(u_n)_{i,j}^c = \lambda_n^{0,c} + \lambda_n^{1,c} i + \lambda_n^{2,c} j.$$

Thus it suffices to bound $(\lambda_n^{0,c}, \lambda_n^{1,c}, \lambda_n^{2,c})$ for each $c \in \{1,2,3\}$. Applying the subsampling operator S on $P_1(u_n)$ again yields a component wise affine linear function given by

$$\begin{aligned} S(P_1(u_n))_{i,j}^c = &\quad [\lambda_n^{0,c} + \gamma^c(\lambda_n^{1,c} + \lambda_n^{2,c})] + \lambda_n^{1,c} i + \lambda_n^{2,c} j \\ = &\quad \tilde{\lambda}_n^{0,c} + \lambda_n^{1,c} i + \lambda_n^{2,c} j, \end{aligned}$$

with $\gamma^c \in \mathbb{R}$ independent of n. Further applying the blockwise cosine transformation operator on $S(P_1(u_n))$, by trigonometric identities we get

$$\begin{aligned} CS(P_1(u_n))_{0,0}^c = \mu_1^c(\lambda_n^{1,c} + \lambda_n^{2,c}) + \mu_2^c \tilde{\lambda}_n^{0,c}, \\ CS(P_1(u_n))_{1,0}^c = \mu_3^c \lambda_n^{1,c}, \quad CS(P_1(u_n))_{0,1}^c = \mu_3^c \lambda_n^{2,c}, \end{aligned} \quad c \in \{1,2,3\},$$

with $\mu_1^c, \mu_2^c, \mu_3^c$ not equal to zero, independent of n. Boundedness of D together with boundedness of $(\tilde{u}_n)$ in particular implies that the sequences

$$(CS(P_1(u_n))_{0,0}^c)_{n\in\mathbb{N}},\ (CS(P_1(u_n))_{1,0}^c)_{n\in\mathbb{N}},\ (CS(P_1(u_n))_{0,1}^c)_{n\in\mathbb{N}}$$

and thus the

$$(\tilde{\lambda}_n^{0,c})_{n\in\mathbb{N}},\ (\lambda_n^{1,c})_{n\in\mathbb{N}},\ (\lambda_n^{2,c})_{n\in\mathbb{N}},\ c \in \{1,2,3\}$$

are bounded. Consequently also $\lambda_n^{0,c}$ and with that $P_1(u_n)$ is bounded. Existence of a - not relabeled - subsequence $(u_n)_{n\in\mathbb{N}}$ converging to $u \in U_D$ follows. Continuity of $\operatorname{TGV}_\alpha^2$ finally implies that u is a minimizer. Equivalence to (66) follows immediately. □

Next we show equivalence of the problem (66), that we will call the primal problem, to a dual and a saddle point problem. This will be the basis for the numerical solution of the discrete minimization problem. For that purpose, we introduce $G : X \to \overline{\mathbb{R}}$ as $G(u,v) = \mathcal{I}_{U_D}(u)$.

Proposition 5.3. *There exists a solution to the dual problem of* (66)*, characterized by*

$$\max_{y\in Y} -\mathcal{G}^*(-K^*y) - F^*(y), \tag{69}$$

as well as to the saddle point problem, given by

$$\min_{x\in X}\max_{y\in Y}(Kx,y)_Y - F^*(y) + G(x). \tag{70}$$

Further, $\hat{x}, \hat{y}$ are solutions to the primal and dual problem, respectively, if and only if $(\hat{x}, \hat{y})$ solves the saddle point problem.

Proof. Given that F is continuous, existence of a solution to the dual problem as well as equality of the primal and the dual problem at optimal points follow immediately from [37, Theorem III.4.1]. From this, together with proposition 5.2, equivalence of the saddle point problem follows from [37, Proposition III.3.1]. □

Note that, as can be easily shown, G^* and F^* are given by

$$G^*(u^*, v^*) = \sup_{u\in U_D}(u^*, u)_U + \mathcal{I}_{\{0\}}(v^*),$$

where

$$\mathcal{I}_{\{0\}}(v) = \begin{cases} 0 & \text{if } v = 0, \\ \infty & \text{else,} \end{cases}$$

and

$$F^*(p,q) = \mathcal{I}_{\|\cdot\|\le\alpha_1}(p) + \mathcal{I}_{\|\cdot\|\le\alpha_0}(q),$$

where

$$\mathcal{I}_{\|\cdot\|_\infty\le\alpha_1}(p) = \begin{cases} 0 & \text{if } \|p\|_\infty \le \alpha_1, \\ \infty & \text{else,} \end{cases} \quad \mathcal{I}_{\|\cdot\|_\infty\le\alpha_0}(q) = \begin{cases} 0 & \text{if } \|q\|_\infty \le \alpha_0, \\ \infty & \text{else,} \end{cases}$$

and

$$\|p\|_\infty = \sup_{i,j}\{|p_{i,j}|_V\}, \quad \|q\|_\infty = \sup_{i,j}\{|q_{i,j}|_W\}.$$

Further, the operator K^* denotes the adjoint of K and is given by

$$K^* = \begin{bmatrix} -\operatorname{div} & 0 \\ -1 & -\operatorname{div} \end{bmatrix},$$

with, abusing notation, $\operatorname{div} = -\nabla^*$ and $\operatorname{div} = -\mathcal{E}^*$ denoting discrete divergence operators depending on their domain of definition, i.e.

$$(\operatorname{div}(p))_{i,j} = \begin{pmatrix} (\delta_{x-}p^{1,1})_{i,j} + (\delta_{y-}p^{2,1})_{i,j} \\ (\delta_{x-}p^{1,2})_{i,j} + (\delta_{y-}p^{2,2})_{i,j} \\ (\delta_{x-}p^{1,3})_{i,j} + (\delta_{y-}p^{2,3})_{i,j} \end{pmatrix}$$

for $p = \begin{pmatrix} (p^{1,1}, p^{2,1}) \\ (p^{1,2}, p^{2,2}) \\ (p^{1,3}, p^{2,3}) \end{pmatrix}$, $0 \le i < 8k$, $0 \le j < 8l$, and

$$(\operatorname{div} q)_{i,j} = \begin{pmatrix} ((\delta_{x+}q^{1,1})_{i,j} + (\delta_{y+}q^{3,1})_{i,j}, (\delta_{x+}q^{3,1})_{i,j} + (\delta_{y+}q^{2,1})_{i,j}) \\ ((\delta_{x+}q^{1,2})_{i,j} + (\delta_{y+}q^{3,2})_{i,j}, (\delta_{x+}q^{3,2})_{i,j} + (\delta_{y+}q^{2,2})_{i,j}) \\ ((\delta_{x+}q^{1,3})_{i,j} + (\delta_{y+}q^{3,3})_{i,j}, (\delta_{x+}q^{3,3})_{i,j} + (\delta_{y+}q^{2,3})_{i,j}) \end{pmatrix}$$

for $q \in W$, where δ_{x+}, δ_{y+} and δ_{x-}, δ_{y-} are defined in equations (59) and (61), respectively.

At last in this subsection, we give an optimality condition for the discrete saddle point problem:

Proposition 5.4. *There exists a solution to* (70) *and* $\hat{x} = (\hat{u}, \hat{v}, \hat{p}, \hat{q})$ *being optimal is equivalent to*

- $\hat{p}_{i,j} = \alpha_1 \frac{(\nabla \hat{u} - \hat{v})_{i,j}}{|(\nabla \hat{u} - \hat{v})_{i,j}|_V}$ *if* $|(\nabla \hat{u} - \hat{v})_{i,j}|_V \neq 0$*, and* $|\hat{p}_{i,j}|_V \leq \alpha_1$ *else.*
- $\hat{q}_{i,j} = \alpha_0 \frac{(\mathcal{E}\hat{v})_{i,j}}{|(\mathcal{E}\hat{v})_{i,j}|_W}$ *if* $|(\mathcal{E}\hat{v})_{i,j}|_W \neq 0$*, and* $|\hat{q}_{i,j}|_W \leq \alpha_1$ *else.*
- $\hat{p} + \operatorname{div} \hat{q} = 0$
- $\hat{u} \in U_D$ *and* $\operatorname{div} \hat{p} = S^* C^* \hat{w}$ *with* $\hat{w} \in U$ *such that*

$$\begin{cases} \hat{w}^c_{i,j} \in \mathbb{R} & \textit{if } ((CS)^c \hat{u}^c)_{i,j} = l^c_{i,j} = o^c_{i,j}, \\ \hat{w}^c_{i,j} \geq 0 & \textit{if } ((CS)^c \hat{u}^c)_{i,j} = o^c_{i,j} \neq l^c_{i,j}, \\ \hat{w}^c_{i,j} \leq 0 & \textit{if } ((CS)^c \hat{u}^c)_{i,j} = l^c_{i,j} \neq o^c_{i,j}, \\ \hat{w}^c_{i,j} = 0 & \textit{if } ((CS)^c \hat{u}^c)_{i,j} \in \overset{\circ}{[l^c_{i,j}, o^c_{i,j}]}, \end{cases}$$

if the data intervals describing D *are given by* $J^c_{i,j} = [l^c_{i,j}, o^c_{i,j}]$.

Proof. The proof is rather short, using that the objective functional is convex and concave in the primal and dual direction, respectively, and thus $(\hat{x}, \hat{y})$ is optimal if and only if

$$0 \in \partial_x \left((\hat{x}, K^* \hat{y})_Y - F^*(\hat{y}) + G(\hat{x}) \right)$$

and

$$0 \in \partial_y \left(-(\hat{x}, K^* \hat{y})_Y + F^*(\hat{y}) - G(\hat{x}) \right).$$

By a standard continuity arguments ([37, Proposition I.5.6]) we get additivity of the subdifferential in the primal direction. Thus, zero being in the subdifferential with respect to the primal variable is equivalent to $\hat{p} + \operatorname{div} \hat{q} = 0$ as well as $\operatorname{div} \hat{p} \in \partial(\mathcal{I}_D \circ (CS))(\hat{u})$. Again, since there exists a point x_0 where $\mathcal{I}_D \circ (CS)$ is finite and continuous, we can apply a chain rule [37, Proposition I.5.7] and get that

$$\operatorname{div} \hat{p} \in \partial(\mathcal{I}_D \circ (CS))(\hat{u}) \quad \Leftrightarrow \quad \operatorname{div} \hat{p} \in S^* C^* \, \partial \mathcal{I}_D(CS\hat{u}),$$

from which the last assertion follows by a case study. Again by a continuity argument, additivity of the subdifferential in the dual direction is satisfied, implying that $K\hat{x} \in \partial F^*(\hat{y})$ is equivalent to zero being in the subdifferential with respect to the dual direction. A component wise evaluation of this further yields equivalence to the first two assertions. □

5.2.3 Practical implementation

We solve the saddle point problem (70), which is equivalent to the original, discrete minimization problem (57), using a primal dual algorithm presented in [26]. Global convergence of this algorithm can be ensured and it is well suited for our minimization problem because, as we will see, all necessary steps during one iteration reduce to simple arithmetic operations and the evaluation of a forward and inverse blockwise cosine transformation, for which highly optimized code already exists. This makes the algorithm fast and also easy to implement on

Algorithm 1 Abstract primal dual algorithm for JPEG decompression

```
• Initialization: Choose $\tau, \sigma > 0$ such that $\|K\|^2\tau\sigma < 1$, $(x^0, y^0) \in X \times Y$
  and set $\overline{x}^0 = x^0$

• Iterations ($n \geq 0$): Update $x^n, y^n, \overline{x}^n$ as follows:

  $$\begin{cases} y^{n+1} = & (I + \sigma\,\partial\,F^*)^{-1}(y^n + \sigma K\overline{x}^n) \\ x^{n+1} = & (I + \tau\,\partial\,G)^{-1}(x^n - \tau K^* y^{n+1}) \\ \overline{x}^{n+1} = & 2x^{n+1} - x^n \end{cases}$$
```

the GPU. Also, as we will see later in this subsection, we can obtain a suitable stopping rule ensuring optimality.

The primal dual strategy for finding saddle points presented in [26] amounts to performing the abstract iteration shown in Algorithm 1. For the practical implementation of this algorithm, it is thus left to find an explicit assignment for $(I + \sigma\,\partial\,F^*)^{-1}$ and $(I + \tau\,\partial\,G)^{-1}$ and to estimate $\|K\|$, the norm of the operator K as linear mapping from X to Y.

Using standard arguments from convex analysis, it can be shown that the resolvent-type operator $(I + \sigma\,\partial\,F^*)^{-1}$ takes the following form:

$$(I + \sigma\partial F^*)^{-1}(v, w) = \big(\mathrm{proj}_{\alpha_1}(v), \mathrm{proj}_{\alpha_0}(w)\big)\,,$$

where

$$\begin{aligned} \mathrm{proj}_{\alpha_1}(v) &= \frac{v}{\max(1, \frac{\|v\|_\infty}{\alpha_1})}, \\ \mathrm{proj}_{\alpha_0}(w) &= \frac{w}{\max(1, \frac{\|w\|_\infty}{\alpha_0})}. \end{aligned} \tag{71}$$

Similar, the evaluation of $(I + \tau\,\partial\,G)^{-1}$ is equivalent to a projection on the data set

$$U_D = \{u \in U \,|\, CSu \in D\}$$

of u in $x = (u, v)$. As lemma 5.1 shows this projection can be reduced to a projection on

$$U_C = \{u \in \tilde{U} \,|\, Cu \in D\}$$

by

$$P_{U_D}(u) = u + S^{-1}\left(P_{U_C}(Su) - Su\right),$$

where $\tilde{U}$ is the low-resolution image space and $\tilde{S}$ denotes the upsampling operator (left inverse) associated with S, given locally by replication of $z \in \mathbb{R}$, i.e.,

$$\tilde{S}z = \begin{pmatrix} z & \dots & z \\ \vdots & & \vdots \\ z & \dots & z \end{pmatrix}.$$

Orthogonality of C allows to reduce $(I + \sigma\,\partial\,G)^{-1}$ to

$$(I + \tau\,\partial\,G)^{-1}(u, v) = u + \tilde{S}\big(\mathrm{proj}_{U_C}(Su) - Su\big),$$

where

$$\operatorname{proj}_{U_C}(u) = C^* z,$$

with

$$z^c_{i,j} = \begin{cases} u^c_{i,j} & \text{if } (Cu)^c_{i,j} \in J^c_{i,j} = [l^c_{i,j}, o^c_{i,j}], \\ o^c_{i,j} & \text{if } (Cu)^c_{i,j} > o^c_{i,j}, \\ l^c_{i,j} & \text{if } (Cu)^c_{i,j} < l^c_{i,j}, \end{cases}$$

$C^* = C^{-1}$ the adjoint of C and $J^c_{i,j} = [l^c_{i,j}, o^c_{i,j}]$ the data interval corresponding to the (i,j)'th pixel of the c'th color component.

Altogether, the concrete implementation of the primal dual algorithm for JPEG decompression can be given in Algorithm 2. Note that the step-size restriction $\sigma\tau \leq \frac{1}{12}$ results from the straightforward estimate $\|K\|^2 < 12$. As one can see, all steps of Algorithm 2 can be evaluated by simple, mostly pixel-wise operations making each iteration step fast.

Remark 5.3. *Actually, the estimate $\|K\|^2 < 12$ can even be slightly improved: First we estimate $\|\nabla\|^2 \leq 8$ and $\|\mathcal{E}u\|^2 \leq 8$. With that, we can estimate $\|K(u,v)\|^2 \leq 8\|u\|^2_U + 2\sqrt{8}\|u\|_U\|v\|_V + 9\|v\|^2_V$. By Youngs inequality we can estimate, for any $a, b \in \mathbb{R}$, $2ab \leq a^2 + b^2$. Now to distribute $\sqrt{8}$ optimally to 8 and 9, we impose*

$$ab = \sqrt{8} \quad a^2 = b^2 + 1,$$

which results in $a^2 = \frac{\sqrt{33}+1}{2}, b^2 = \frac{\sqrt{33}-1}{2}$ and we obtain the estimate

$$\|K\|^2 \leq (8 + \frac{\sqrt{33}+1}{2}) \approx 11.3723 < 12.$$

Stopping criterion

In order to validate our numerical solution, we seek for a suitable stopping rule. As has been shown in the proof of proposition 5.3, an optimal solution $(\hat{x}, \hat{y})$ of the saddle point problem (70) satisfies

$$F(K\hat{x}) + G(\hat{x}) = \quad -(G^*(-K^*\hat{y}) + F^*(\hat{y})).$$

This allows, for any $(x, y) \in X \times Y$ with $x = (u, v)$, $y = (p, q)$, the estimate

$$\begin{aligned} 0 \leq F(Kx) - F(K\hat{x}) = & \qquad F(Kx) + G(\hat{x}) + G^*(-K^*\hat{y}) + F^*(\hat{y}) \\ \leq & \qquad F(Kx) + G(\hat{x}) + G^*(-K^*y) + F^*(y) \\ = & \quad F(Kx) + \sup_{u \in U_D} (\operatorname{div} p, u)_U + \mathcal{I}_{\{0\}}(p + \operatorname{div} q,) + F^*(y). \end{aligned} \tag{72}$$

Plugging in the iterates of algorithm 2, this yields

$$0 \leq F(Kx) - F(K\hat{x}) \leq F(Kx_n) + \sup_{u \in U_D} (\operatorname{div} p_n, u)_U + \mathcal{I}_{\{0\}}(p_n + \operatorname{div} q_n,), \tag{73}$$

which would allow a good stopping rule if we can suitably bound $\sup_{u \in U_D} (\operatorname{div} p_n, u)_U + \mathcal{I}_{\{0\}}(p_n + \operatorname{div} q_n)$.

At first we focus on $\mathcal{I}_{\{0\}}(p_n + \operatorname{div} q_n)$:

Algorithm 2 Scheme of implementation for JPEG decompression

```
function TGV-JPEG($J_{\mathrm{comp}}$)
    $(d, Q) \leftarrow$ Decoding of JPEG-Object $J_{\mathrm{comp}}$
    $d \leftarrow d \cdot Q$
    $u \leftarrow S^{-1}(C^*(d))$
    $v \leftarrow 0, \overline{u} \leftarrow u, \overline{v} \leftarrow 0, p \leftarrow 0, q \leftarrow 0$
    choose $\sigma, \tau > 0$ such that $\sigma\tau \leq 1/12$
    repeat
        $p \leftarrow \mathrm{proj}_{\alpha_1}\left(p + \sigma(\nabla \overline{u} - \overline{v})\right)$
        $q \leftarrow \mathrm{proj}_{\alpha_0}(q + \sigma(\mathcal{E}(\overline{v}))$
        $u_+ \leftarrow u + \tau(\operatorname{div} p)$
        $v_+ \leftarrow v + \tau(p + \operatorname{div} q)$
        $u_+ \leftarrow u_+ + \tilde{S}(\mathrm{proj}_{U_C}(Su_+) - Su_+)$
        $\overline{u} \leftarrow (2u_+ - u)$, $\overline{v} \leftarrow (2v_+ - v)$
        $u \leftarrow u_+$, $v \leftarrow v_+$
    until Stopping criterion fulfilled
    return $u_+$
end function
```

Proposition 5.5. *Let $(x_n, y_n) = ((u_n, v_n, p_n, q_n))$ be the iterates of Algorithm 2 and $(\hat{x}, \hat{y})$ a saddle point of* (70). *We define*

$$\beta_n := \frac{\alpha_1}{\max(\alpha_1, \|\operatorname{div} q_n\|_\infty)}$$

and

$$\tilde{q}_n := \beta_n q_n.$$

Then

$$F(Kx_n) - F(K\hat{x}) \leq F(Kx_n) + \sup_{u \in U_D} (u, -\operatorname{div}^2 \tilde{q}_n)_U. \tag{74}$$

Proof. Using the estimate (72) the inequality follows immediately, given that $\|\tilde{q}_n\|_\infty \leq \alpha_0, \|\operatorname{div} \tilde{q}_n\|_\infty \leq \alpha_1$. □

In the case of grayscale images, or more general, if no subsampling is performed, this already yields a practicable stopping rule since the data set U_D is bounded and thus the right hand side of (74) converges to zero. But in general, unboundedness of the kernel of the subsampling operator prevents us from controlling $\sup\limits_{u \in U_D} (u, -\operatorname{div}^2 \tilde{q}_n)_U$. A solution to this is the following proposition, which is partly motivated by [67, Subsection 4.2].

Proposition 5.6. *Let again $(x_n, y_n) = ((u_n, v_n, p_n, q_n))$ be the iterates of Algorithm 2, $(\hat{x}, \hat{y}) = (\hat{u}, \hat{v}, \hat{p}, \hat{q})$ a saddle point of* (70) *and $(\tilde{q}_n)$ as in proposition 5.5. Then, with $\tilde{S}^*$ the adjoint of the upsampling operator $\tilde{S}$, we have*

$$\begin{aligned} F(Kx_n) - F(K\hat{x}) \leq \quad & F(Kx_n) + \sup_{u \in D} (u, -C^* \tilde{S}^* \operatorname{div}^2 \tilde{q}_n)_U \\ & + T \| \operatorname{div}^2 \tilde{q}_n - S^* \tilde{S}^* \operatorname{div}^2 \tilde{q}_n \|_{p'} \end{aligned} \tag{75}$$

for any $1 \leq p \leq \infty$ *and* $T \geq 0$ *such that* $\|\hat{u} - \tilde{S}^* S\hat{u}\|_p \leq T$*. Additionally, the last quantity converges to zero as* $n \to \infty$*.*

Remark 5.4. *Note that by* $\|\cdot\|_p$ *we denote a discrete* L^p *norm on* U*, i.e.,*

$$\|u\|_p = \begin{cases} \left(\sum_{i,j} |u_{i,j}|_U^p\right)^{1/p} & \text{if } p < \infty, \\ \max_{i,j}\{|u_{i,j}|_U\} & \text{if } p = \infty, \end{cases}$$

for $u \in U$*, and by* p' *the conjugate exponent of* p*, i.e.,* $p' = \frac{p}{p-1}$*.*

Proof of proposition 5.6. First note that $\tilde{S}^* S^*$ and $S\tilde{S}$ both are the identity. Thus, for any $u \in U$, we can estimate

$$\begin{aligned} & (u, -\operatorname{div}^2 \tilde{q}_n)_U \\ = & \left[(u - \tilde{S}Su + \tilde{S}Su, -\operatorname{div}^2 \tilde{q}_n + S^*\tilde{S}^* \operatorname{div}^2 \tilde{q}_n - S^*\tilde{S}^* \operatorname{div}^2 \tilde{q}_n)_U\right] \\ = & \left[(u - \tilde{S}Su, -\operatorname{div}^2 \tilde{q}_n + S^*\tilde{S}^* \operatorname{div}^2 \tilde{q}_n)_U + (\tilde{S}Su, -S^*\tilde{S}^* \operatorname{div}^2 \tilde{q}_n)_U\right] \\ \leq & \|u - \tilde{S}Su\|_p \|\operatorname{div}^2 \tilde{q}_n - S^*\tilde{S}^* \operatorname{div}^2 \tilde{q}_n\|_{p'} + (Su, -\tilde{S}^* \operatorname{div}^2 \tilde{q}_n)_U. \end{aligned}$$

Now given any optimal solution $(\hat{u}, \hat{v})$ of the primal problem, we can define a modified objective functional as in in (66) by adding $\mathcal{I}_B(u)$ with

$$B = \{u \in U | \|u - \tilde{S}Su\|_p \leq T\}$$

and any $T \geq \|\hat{u} - \tilde{S}S\hat{u}\|_p$ to the original objective functional as in (66). As a result, $(\hat{u}, \hat{v})$ is also a solution to the modified problem and estimations like (72) and (74) for the modified problem result in

$$0 \leq F(Kx_n) - F(K\hat{x}) \leq F(Kx_n) + \sup_{\substack{u \in U_D \\ \|u - \tilde{S}Su\|_p \leq T}} (u, -\operatorname{div}^2 \tilde{q}_n)_U. \tag{76}$$

Combining this with our previous estimations yields

$$\begin{aligned} F(Kx_n) - F(K\hat{x}) \leq \; & F(Kx_n) + \sup_{u \in U_D} (Su, -\tilde{S}^* \operatorname{div}^2 \tilde{q}_n)_U \\ & + T\|\operatorname{div}^2 \tilde{q}_n - S^*\tilde{S}^* \operatorname{div}^2 \tilde{q}_n\|_q \\ \leq \; & F(Kx_n) + \sup_{v \in D} (v, -C\tilde{S}^* \operatorname{div}^2 \tilde{q}_n)_U \\ & + T\|\operatorname{div}^2 \tilde{q}_n - S^*\tilde{S}^* \operatorname{div}^2 \tilde{q}_n\|_q, \end{aligned} \tag{77}$$

which is the desired estimation. Finally, convergence of $\tilde{q}_n$ to $\hat{q}$, together with $\operatorname{div} \hat{q} + \hat{p} = 0$ and $\operatorname{div}^2 \hat{q} = -\operatorname{div} \hat{p} = -S^*C^*w$ for some $w \in U$, as shown in proposition 5.4, implies the claimed convergence to zero. □

In practice a bound on $\|\hat{u} - \tilde{S}S\hat{u}\|_p$ is hard to obtain. We can, however, use $\gamma\|u_n - \tilde{S}Su_n\|_p$ with $\gamma > 1$ to estimate T and with that obtain the desired estimation on $F(Kx_n) - F(K\hat{x})$ at least in the limit.

Given that D is bounded, we have thus obtained a practicable stopping criterion, which can be summed up in the following

Proposition 5.7. *Take* $1 < \gamma$, $1 \leq p \leq \infty$, $(x_n, y_n) = ((u_n, v_n, p_n, q_n))$ *to be the iterates of Algorithm 2 and* $(\hat{x}, \hat{y}) = (\hat{u}, \hat{v}, \hat{p}, \hat{q})$ *a saddle point of* (70). *Define*

$$\begin{aligned} \mathcal{G}(x_n, y_n) = \quad & F(Kx_n) + T_n \| \operatorname{div}^2 \tilde{q}_n - S^* \tilde{S}^* \operatorname{div}^2 \tilde{q}_n \|_{p'} \\ & + \sum_{i,j,c} \left(\frac{o^c_{i,j} + l^c_{i,j}}{2} (-C\tilde{S}^* \operatorname{div}^2 \tilde{q}_n)^c_{i,j} \right. \\ & \left. + \frac{o^c_{i,j} - l^c_{i,j}}{2} |(-C\tilde{S}^* \operatorname{div}^2 \tilde{q}_n)^c_{i,j}| \right), \end{aligned} \tag{78}$$

where $\tilde{q}_n$ *is defined as in proposition 5.5,* $l^c_{i,j}, o^c_{i,j}$ *are such that* $J^c_{i,j} = [l^c_{i,j}, o^c_{i,j}]$ *and* $T_n = \gamma \| u_n - \tilde{S}Su_n \|_p$. *Then,*

$$\mathcal{G}(x_n, y_n) \to 0 \text{ as } n \to \infty$$

and, additionally,

$$\mathcal{G}(x_n, y_n) \geq F(Kx_n) - F(K\hat{x}) \geq 0 \tag{79}$$

whenever $T_n \geq \| \hat{u} - \tilde{S}S\hat{u} \|_p$.

This allows, for given $\epsilon > 0$, to use $\mathcal{G}(x_n, y_n) < \epsilon$ as stopping criterion.

5.2.4 Numerical experiments

In this subsection we present the numerical evaluation of the proposed method for artifact free JPEG decompression. We tested the method for several lossy JPEG compressed images, where the memory requirement of each JPEG compressed image (including lossless compression) is given in *bits per pixel (bpp)*. Note that we consider 8 bit grayscale- and 24 bit true color images, i.e. an uncompressed color images requires 24 bpp.

We fixed the ratio between α_0 and α_1 for the evaluation of the TGV^2_α functional (cf. equation (57)) as $\frac{\alpha_0}{\alpha_1} = \sqrt{2}$ based on some empirical observations. However, this choice can certainly be improved by more extensive numerical testing.

As stopping criterion we use either the iteration number, if this is necessary for comparability, or a normalized primal dual gap. This normalized primal dual gap is given by

$$\overline{\mathcal{G}}(x_i, y_i) = \frac{1}{nm} \mathcal{G}(x_i, y_i) \tag{80}$$

with $\mathcal{G}(x_i, y_i)$ as in equation (78), (x_i, y_i) the current iterates of algorithm 2 and $n \in \mathbb{N}$ and $m \in \mathbb{N}$ are the vertical and horizontal number of pixels, respectively. The reason for normalizing the primal dual gap is to make it image size independent and, taking into account estimation (79), to to get an estimation on the average pixel-wise difference

$$\left[|(\nabla u_i - v_i)_{i,j}|_V + |(\mathcal{E}(v_i))_{i,j}|_W \right] - \left[|(\nabla \hat{u} - \hat{v})_{i,j}|_V + |(\mathcal{E}(\hat{v}))_{i,j}|_W \right],$$

where again $(\hat{u}, \hat{v})$ denote optimal solutions of the primal problem (66). Considering this estimation, it is important that in the numerical computations we process our images with range [0,255], thus a pixel wise image error less than 1 would assure, due to truncation to integer when visualizing the image, a maximal visible error of one pixel-step for each color component. This theoretical

Table 1: Parameter setting for JPEG decompression

Algorithm 2		**TGV** as in (57)		**Gap** as in (78)	
σ	$\sqrt{1/12}$	$\frac{\alpha_0}{\alpha_1}$	$\sqrt{2}$	γ	1.001
τ	$\sqrt{1/12}$			p	2

bound cannot be improved, since even an arbitrary small pixel-wise error could result in a pixel-wise error of one after truncation to integer.

At first, figure 7 shows three JPEG compressed images, corrupted by blocking and ringing artifacts, and their improved decompression obtained with our TGV based method. As stopping criterion we require the normalized primal dual gap to be below $\epsilon = 10^{-1}$. As one can see, indeed all blocking and ringing artifacts could be removed, while edges are kept sharp, leading to more natural and visually more appealing images.

Note that for the primal dual gap stopping criterion as in proposition 5.7 we still have some parameter choices, namely $\gamma > 0$ and p (resulting in a p') for the norm estimation. For the images of figure 7 we chose $\gamma = 1.001$ and $p = 2$. This can be motivated as follows:

Figure 8 compares the primal dual gap for different parameters as well as the difference of the TGV value of the current iterate and the “true” solution, with respect to iteration number. The first 1000 iterations are plotted for comparison, while the true solution is obtained by 6000 iterations. We have observed that, even for the choice $\gamma = 0$, which is not consistent with (79), the primal dual gap always remains above the TGV difference. As can be seen in figure 8 at the bottom, different parameters γ and p result in only marginal changes of the primal dual gap. However, in order to maintain a theoretical error bound, we choose $\gamma = 1.001$ in all experiments and also for the plot in figure 8 at the top. Numerical observations further suggest that the choice $p = 2$ results in the lowest primal dual gap, hence this is used for our experiments. For convenience we summarize the final choice of parameters for JPEG decompression in table 1.

Next we compare the standard reconstruction of JPEG images with a TV based reconstruction as in [13] and the proposed TGV based reconstruction. As can be seen, in particular in the surface plots, the TV-based method is also able to remove blocking artifacts and maintain sharp edges. However, it leads to staircasing effects in regions that should be smooth. In contrast to that, the TGV based method yields a more natural and visually more appealing reconstruction.

Figure 10 then shows a more detailed comparison of the standard and TGV based reconstruction for a color test image. The original, uncompressed image is shown on the top, while the standard and TGV based reconstruction of a JPEG compressed version are shown on the left and right, respectively. It can be seen that, in particular for this example, the TGV based method yields a good reconstruction that visually appears very close to the original image. To further confirm this visual impression we show the difference between the standard and TGV based reconstruction, respectively, to the original image at a logarithmic scale. Note that, even though the overall energy of the error may not be improved, it is of a different, less visible structure as being almost

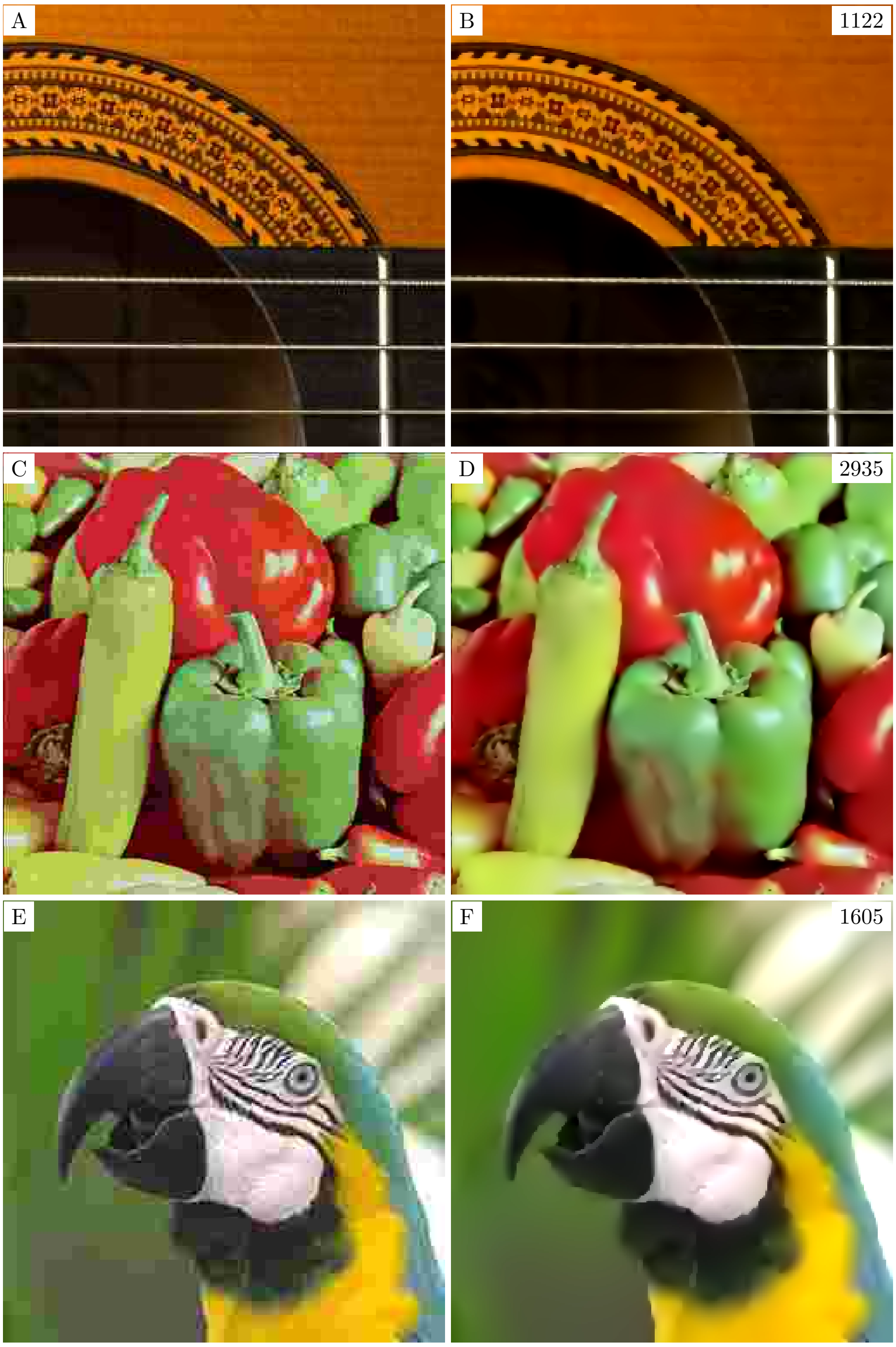


Figure 7: On the left: Standard decompression. On the right: TGV-based reconstruction obtained with normalized primal dual gap below 10^{-1} as stopping criterion (Number of iterations on top-right). A-B: Guitar image at 1.06 bpp (256×256 pixels). C-D: Peppers image at 0.15 bpp (512×512 pixels). E-F: Parrot image at 0.3 bpp (256×256 pixels).

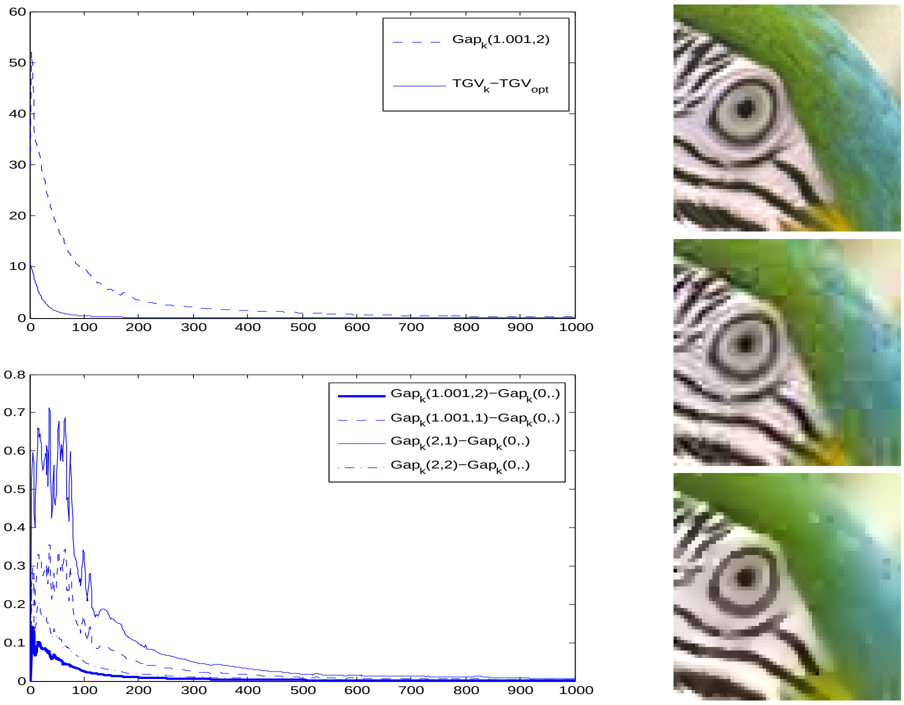


Figure 8: On the left: Top: Plot of the normalized primal dual gap as in (80) compared to difference in TGV value of current iterate and true solution (obtained at 6000 iterations). Bottom: Difference of normalized primal dual gap for different values of γ and p to a normalized primal dual gap for $\gamma = 0$. On the right: From top to bottom: Uncompressed version of the parrots eye image used for the plot, standard reconstruction of a JPEG compressed version at 1.55 bpp, TGV based reconstruction (6000 iterations) of the same JPEG compressed version..

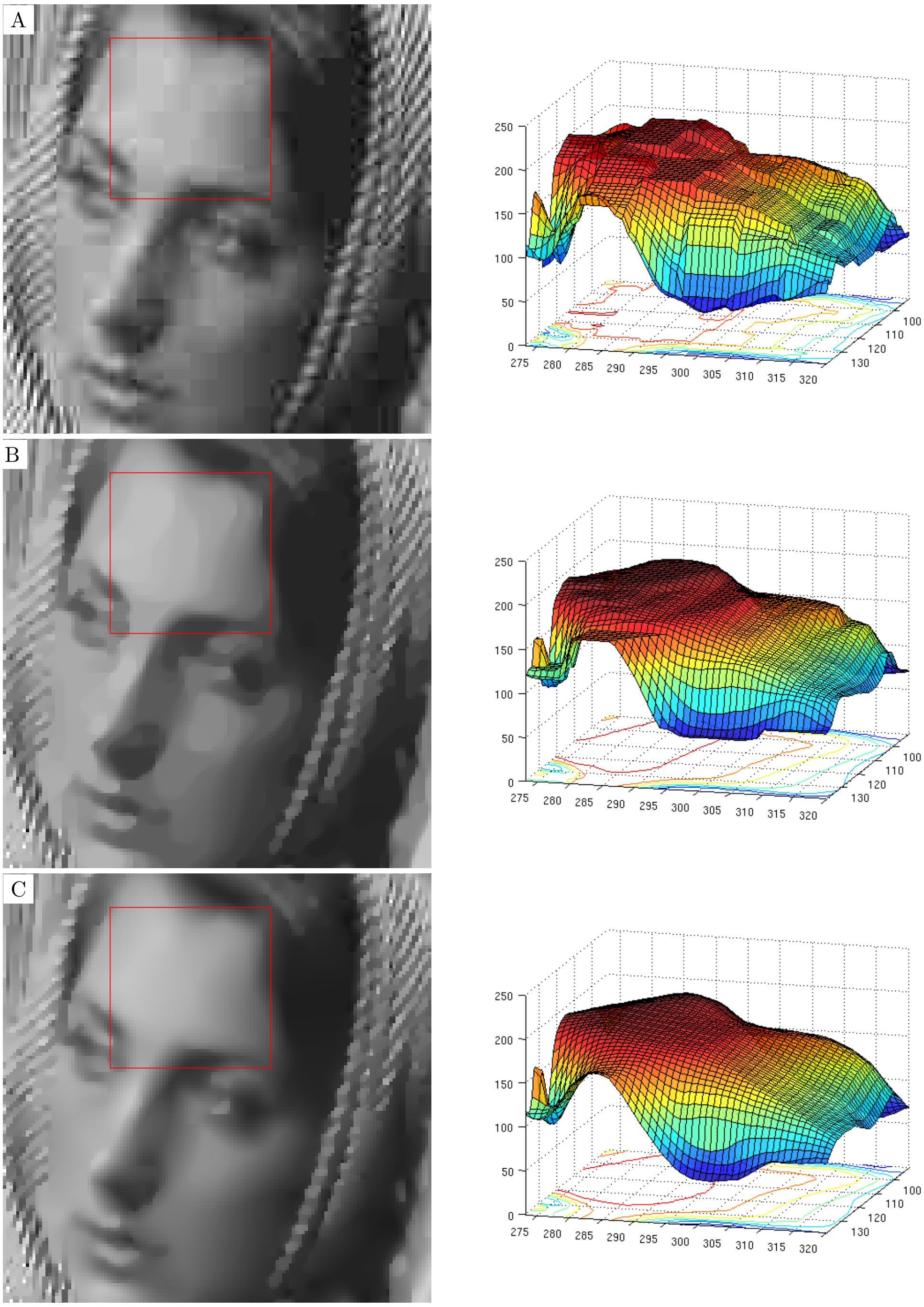


Figure 9: Close-up of Barbara image at 0.4 bpp at 1000 iterations. The marked region on the left is plotted as surface on the right. A: Standard decompression. B: TV-based reconstruction. C: TGV-based reconstruction.

Table 2: Computation times in seconds to perform 1000 iterations for different devices and image sizes. CPU: AMD Phenom 9950. GPUs: Nvidia Quadro FX 3700 (compute capability 1.1), Nvidia GTX 280 (compute capability 1.3), Nvidia GTX 580 (compute capability 2.0). Note that on the Quadro FX 3700 and GTX 280, not enough memory was available to perform the algorithm for the 3200×2400 pixel image.

Device	512×512	1600×1200	3200×2400
CPU Single-core	53.22	672.51	1613.44
CPU Quad-core	28.32	263.70	812.18
GPU Quadro FX 3700	4.92	35.52	-
GPU Nvidia GTX 280	2.2	10.22	-
GPU Nvidia GTX 580	1.2	6.6	25.70

constant over different image patches. Also, as can be seen in particular a little above the center of the image, some errors at images edges have been completely removed with our method. Another interesting observation is that on the wave structure at the top of the image, peaks have been reconstructed accurately with the TGV based method, while there is some error in the transitions. This may be due to a linear approximation of an image area with polynomial structure, and at this point, a higher order TGV functional for regularization may lead to further improvement.

5.2.5 A GPU implementation

As already discussed in [14], we also developed a parallel implementation of the reconstruction method for multi-core CPUs and GPUs, using OpenMP [57] and Nvidia's Cuda [55], respectively. For the GPU implementation we partly used kernel functions adapted to the compute capability of the device. The blockwise DCT was performed on the CPU and the GPU using FFTW [39] and a block-DCT kernel provided by the Cuda SDK, respectively. Computation times of those implementations for multiple image sizes are given in table 2, taken from [14]. The relative time cost of particular iteration steps is compared in table 3. As one can see, especially the GPU implementation yields a high acceleration and makes the method suitable for practical applications. The given computation times correspond to the computation of 1000 iterations, which is motivated by the number of iterations resulting from a normalized primal dual gap below 10^{-1} as stopping criterion. Let us remark however that, since the decrease of the TGV-value of the image is typically very high especially during the first iterations of the algorithm (see figure 8), and since $u_n \in U_D$ can be ensured for any iteration step image u_n, one can also use the images obtained after only a few iterations as (intermediate or final) reconstruction. This yields a practicable method that allows to improve given JPEG images in almost real time.

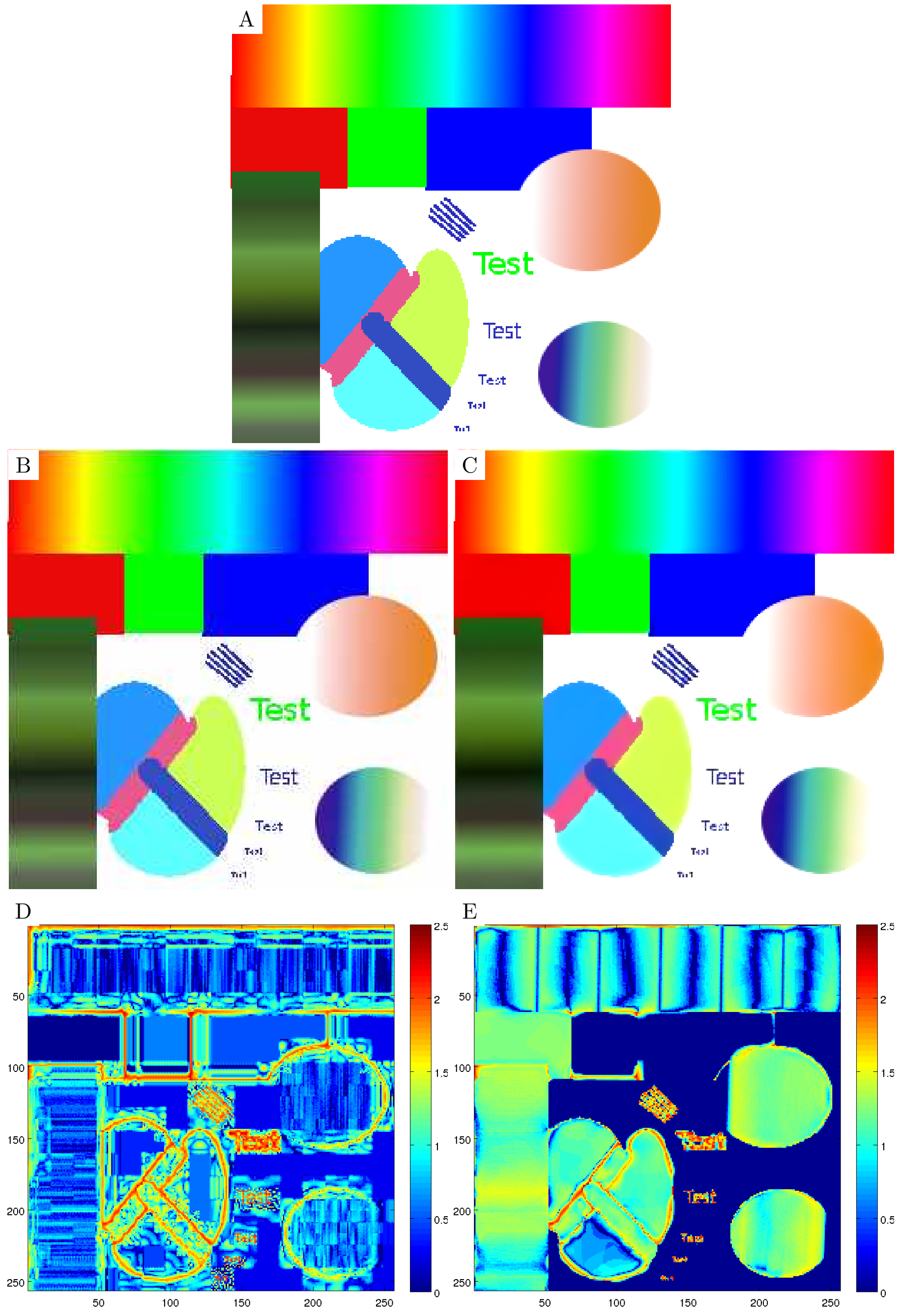


Figure 10: A: Uncompressed color test image (256 × 256 pixels). B: Standard reconstruction of JPEG compression at 0.76 bpp. C: TGV based reconstruction obtained as $\overline{\mathcal{G}}(x_i, y_i) < 10^{-1}$ (1017 iterations). D: Visualization of the pointwise reconstruction error of the standard reconstruction (logarithmic scale). E: Visualization of the pointwise reconstruction error of the TGV based reconstruction (logarithmic scale).

Table 3: Relative computation times in percent for different iteration steps. The data was obtained by reconstructing a 1600 × 1200 pixel image with 1000 iterations.

Iteration step	CPU	GPU
$\left\{\begin{array}{l} p \leftarrow \mathrm{proj}_{\alpha_1}(p + \sigma(\nabla \overline{u} - \overline{v})) \\ q \leftarrow \mathrm{proj}_{\alpha_0}(q + \sigma(\mathcal{E}(\overline{v})) \end{array}\right\}$	52%	32%
$\left\{\begin{array}{l} u_+ \leftarrow u + \tau(\operatorname{div} p) \\ v_+ \leftarrow v + \tau(p + \operatorname{div} q) \end{array}\right\}$	19%	29%
$\left\{ u_+ \leftarrow u_+ + \tilde{S}(\mathrm{proj}_{U_C}(Su_+) - Su_+) \right\}$	20%	21%
$\left\{\begin{array}{l} \overline{u} \leftarrow (2u_+ - u) \\ \overline{v} \leftarrow (2v_+ - v) \end{array}\right\}$	9%	18%

5.2.6 Extension for DCT based zooming

As already mentioned, by defining the subsampling operation also on the brightness component of the images, the model of subsection 5.2 yields a method that allows for combined decompression and zooming of the JPEG compressed images, that will further be called zooming by upsampling. This can be realized without further modification and hence all results of subsection 5.2, in particular existence of a solution, convergence of the primal dual algorithm as well as the stopping rule estimation, apply.

But, as discussed in [14], also a slightly different extension of the model of subsection 5.2 allows for combined decompression and zooming. We will now briefly discuss this extension and compare the obtained results with zooming by upsampling and other zooming methods.

The model formulation is straightforward: Given again a compressed JPEG file, as described at the beginning of subsection 5.2, we can obtain, for each color component $c \in \{1, 2, 3\}$, bounded closed intervals $(J^c_{i,j})$ describing the set of possible source data D. The data size for the brightness component describes the actual image size, i.e. given $8k \times 8k$ data intervals $(J^1_{i,j})_{0 \leq i,j < 8k}$, the discrete image is given in $\mathbb{R}^{8k \times 8k \times 3}$. Note that, for simplicity, we assume the images to be quadratic. The horizontal and vertical number of pixels, $8k$, is a multiple of the subsampled data size of the color components, i.e. $8k = s8k_2$, $8k = s8k_3$, $s \in \mathbb{N}$, where k_2, k_3 are such that $(J^2_{i,j})_{0 \leq i,j < 8k_2}$ and $(J^3_{i,j})_{0 \leq i,j < 8k_3}$ are given. Note that, again for simplicity, we assume the same horizontal and vertical subsampling factor for both, the Cb and Cr color component.

Now to obtain a high resolution decompression, increased by a factor $f \in \mathbb{N}$, we assume $(8fk)^2$ data intervals $(\tilde{J}^c_{i,j})_{0 \leq i,j < 8fk}$ to be given for each color component $c \in \{1, 2, 3\}$. These data intervals are defined blockwise by adding intervals containing all of $\mathbb{R}$ to the original 8×8 data intervals of each block, up to a certain blocksize $N \times N$. The blocksize N is defined to be $8f$ and $8sf$ for the brightness and color components, respectively. Further, for normalization purposes, the original data intervals, describing the first 8×8 coefficients of each block, are enlarged by multiplying them with $8f$ and $8sf$, respectively.

Extending now the component-, blockwise DCT operator defined locally for each component in (63) up to a block size $8f$ for the brightness- and $8sf$ for the color components, and denoting this extension again by

$$C : U \to U, \quad U := \mathbb{R}^{8fk \times 8fk \times 3},$$

the set of high resolution source images for a given JPEG compressed low resolution file can be defined as

$$U_D = \{u \in U \,|\, (Cu)^c_{i,j} \in \tilde{J}^c_{i,j},\, c \in \{1,2,3\}, 0 \leq i,j < 8fk\}. \tag{81}$$

Note that in the equivalent setting in function space, we find ourselves thus with the situation that only finitely many intervals are bounded, which again justifies the general problem formulation of section 4.

Given any compressed, low resolution JPEG image, we can now formulate the discrete minimization problem for combined decompression and DCT-based zooming as

$$\min_{u \in U} \mathrm{TGV}^2_\alpha(u) + \mathcal{I}_{U_D}(u). \tag{82}$$

Existence of a solution to this problem can be shown as follows:

Proposition 5.8. *With U_D defined as in* (81)*, there exists a solution to* (82) *and this problem is equivalent to solving*

$$\min_{(u,v) \in U \times V} F(K(u,v)) + \mathcal{I}_{U_D}(u), \tag{83}$$

with $V = U \times U$, K and F as defined in (67) *and* (68)*, respectively.*

Proof. Since problem (82) only differs to the pure decompression problem (65) by the fact that some intervals may be unbounded, we can use the same techniques as in the proof of proposition 5.2 and obtain existence of a solution, provided that the coefficients

$$(C(P_1(u_n)))^c_{0,0}, (C(P_1(u_n)))^c_{1,0}, (C(P_1(u_n)))^c_{0,1}, \quad c \in \{1,2,3\},$$

where C is the component wise block cosine operator, are bounded. But, due to the modeling of U_D as described in (81) and above, the first 8×8 intervals of each block, and in particular the intervals $\tilde{J}^c_{0,0}, \tilde{J}^c_{1,0}, \tilde{J}^c_{0,1}$, $c \in \{1,2,3\}$ are bounded. Thus boundedness of $u_n - P_1(u_n)$ implies boundedness of said coefficients and hence existence of a solution to (82). The equivalence to (83) again is immediate. □

Due to the generality of the proofs of propositions 5.3 and 5.4, their results, with the subsampling operator being the identity but some intervals containing all of $\mathbb{R}$, apply, and we thus know that (82) possesses a solution, any solution $(\hat{u}, \hat{v}, \hat{p}, \hat{q})$ of the related saddle point problem,

$$\min_{x \in X} \max_{y \in Y} (Kx, y)_Y - F^*(y) + G(x), \tag{84}$$

yields solutions $(\hat{u}, \hat{v})$ of the primal problem (83), and $(\hat{p}, \hat{q})$ of its dual problem, defined by

$$\max_{y \in Y} -G^*(-K^*y) - F^*(y), \tag{85}$$

the primal and the dual problem coincide at $(\hat{u},\hat{v}),(\hat{p},\hat{q})$ and for (84) existence of a solution as well as the optimality condition as in proposition 5.4 applies. Also algorithm 2 can be applied to solve (84) and global convergence can be assured.

Only the stopping rule of proposition 5.7 needs some amendments: For the setting of DCT based zooming, the subsampling operator S is just the identity, thus unboundedness of its kernel no longer is an issue, but unboundedness of the data D comes as additional feature. Taking this into account, we can define the following modified primal dual gap:

Proposition 5.9. *Take* $1 < \gamma$, $1 \leq p \leq \infty$, $(x_n, y_n) = ((u_n, v_n, p_n, q_n))$ *to be the iterates of Algorithm 2 with the data set* U_D *as in* (81) *and* $(\hat{x},\hat{y}) = (\hat{u},\hat{v},\hat{p},\hat{q})$ *a saddle point of* (84). *With* $I_1 \subset \mathbb{N}\times\mathbb{N}\times\{1,2,3\}$ *the set of all indices* (i,j,c) *such that* $\tilde{J}^c_{i,j} = \mathbb{R}$ *and* I_2 *the set of all indices* (i,j,c) *such that* $J^c_{i,j} = [l^c_{i,j}, r^c_{i,j}]$, *with* $l^c_{i,j}, r^c_{i,j} \in \mathbb{R}$, *define*

$$\begin{aligned}\mathcal{G}(x_n,y_n) = \quad & F(Kx_n) + T_n \|(C(\operatorname{div}^2 \tilde{q}_n))|_{I_1}\|_{p'} \\ & + \sum_{(i,j,c)\in I_2} \left(\frac{o^c_{i,j} + l^c_{i,j}}{2} (-(C \operatorname{div}^2 \tilde{q}_n)|_{I_2})^c_{i,j} \right. \\ & \left. + \frac{o^c_{i,j} - l^c_{i,j}}{2} |(-(C \operatorname{div}^2 \tilde{q}_n)|_{I_2})^c_{i,j}| \right),\end{aligned} \tag{86}$$

where $\tilde{q}_n$ *is defined as in proposition 5.5 and* $T_n = \gamma \|(Cu_n)|_{I_1}\|_p$. *Then,*

$$\mathcal{G}(x_n,y_n) \to 0 \textit{ as } n \to \infty$$

and, additionally,

$$\mathcal{G}(x_n,y_n) \geq F(Kx_n) - F(K\hat{x}) \geq 0 \tag{87}$$

whenever $T_n \geq \|\hat{u}\|_p$.

Proof. This can be shown similar as in proposition 5.7 since $\operatorname{div}^2 \tilde{q_n} \to -\operatorname{div}(\hat{p})$ as $n \to \infty$ and, by proposition 5.4, $(C(\operatorname{div}(\hat{p}))^c_{i,j} = 0$ for $(i,j,c) \in I_1$. □

Again this allows, for given $\epsilon > 0$, to use $\mathcal{G}(x_n,y_n) < \epsilon$ as stopping criterion.

In practice, we again use a normalization of $\mathcal{G}$, denoted by $\overline{\mathcal{G}}$, as in (80). We have plotted the iterative development of this normalized gap in in figure 11. As one can see there, again the normalized primal dual gap corresponding to the parameters $\gamma = 1.001$, $p = p' = 2$ is always above the difference of the current TGV value to the optimal TGV value. However, as shown in the lower plot of figure 11, since, as consequence of the 8 times magnification, the unbounded data part has more overall influence, the normalized primal dual gaps for different parameters differ more strongly than they do in figure 8.

Figure 12 then finally allows to compare our combined decompression method to other methods for 8 times magnification of a 64×64 color image. We compared the previously described zooming technique to cubic interpolation, cubic interpolation of a low resolution TGV reconstruction of the small JPEG image and TGV based combined decompression and zooming by upsampling. Note that for the combined decompression and zooming, we fixed the ratio between α_0 and α_1 for the evaluation of TGV^2_α to 4. As one can see, the method of extending the data set by unbounded intervals performs best in terms of visual

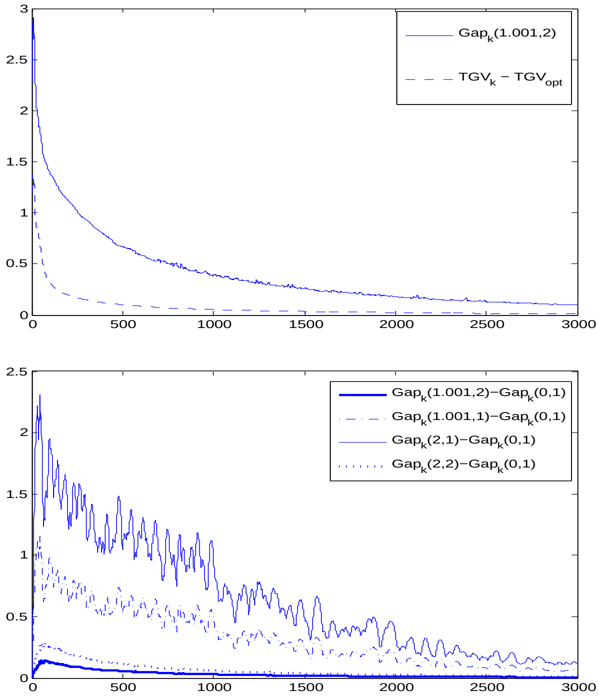


Figure 11: Top: Plot of the normalized primal dual gap and the difference in TGV value of current iterate and true solution (obtained at 9000 iterations). Bottom: Difference of normalized primal dual gap for different values of γ and p and normalized primal dual gap for $\gamma = 0$. The data was obtained by simultaneous TGV based decompression and zooming of the hand image as in figure 12

image quality. Note, however, that the downsampling was done by applying a block DCT, leaving out high frequency information and then applying JPEG compression to the resulting low resolution image, thus the source data fits to the model assumption of DCT based decompression and zooming.

5.3 Color JPEG 2000 decompression

As second application, we want to use the reconstruction model of section 4 for the improved reconstruction of JPEG 2000 color images, where the coding is essentially based on a biorthogonal wavelet transform. For the sake of self-containedness we will briefly explain basic features of JPEG 2000 compression that are necessary to understand the modeling. It will turn out that, again, the set of possible source images can be described by interval restriction on the coefficients of the transformed image and thus fits to our reconstruction model of section 4. However, due to the coding process, it will not be possible to restrict every coefficient by a bounded interval.

After presenting an overview of JPEG 2000 compression, we will define and discuss the minimization problem for artifact free JPEG 2000 decompression and, subsequently, its discretization and numerical solution.

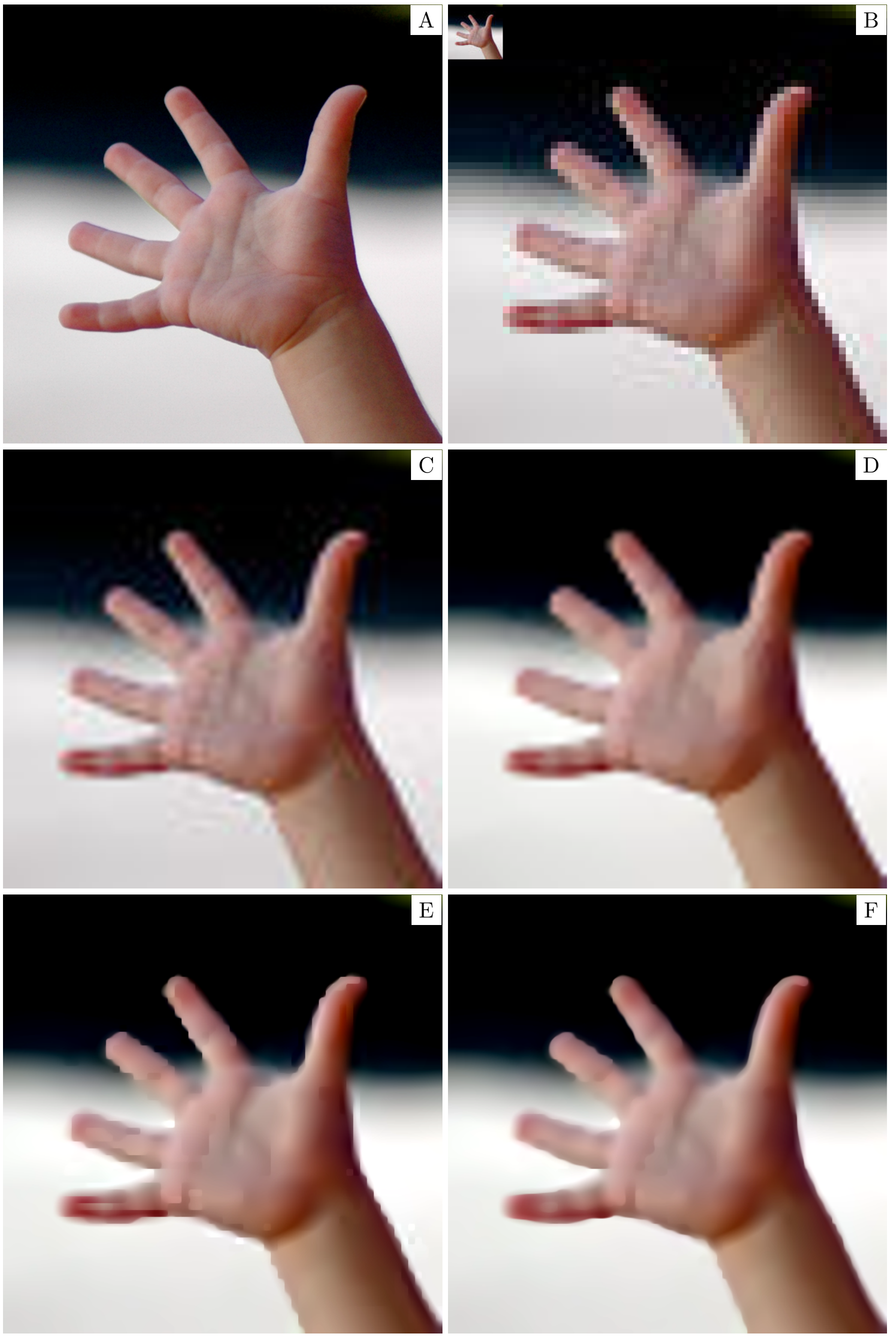


Figure 12: A: Original-sized, uncompressed image (512×512 pixels). B: Downsampled, JPEG compressed image (64×64 pixels, 2.96 bpp) together with 8 times magnification by pixel repetition. C: 8 times magnification by cubic interpolation. D: 8 times magnification of a TGV based decompression by cubic interpolation. E: Simultaneous TGV based decompression and factor 8 zooming by upsampling. F: Simultaneous TGV based decompression and factor 8 DCT based zooming. All TGV based decompressions were obtained with the normalized primal dual gap being below 10^{-1} as stopping rule. Image by [47], licensed under CC-BY-2.0 (http://creativecommons.org/licenses/by/2.0/).

But at first, let us discuss previous attempts to improve the reconstruction quality of the JPEG 2000 standard. To the best knowledge of the author, in contrast to the JPEG decompression model, there does not exist any model or method designed in particular for improved JPEG 2000 decompression that is related to the present one. However, even if not designated to improve the JPEG 2000 compression/decompression procedure, some works on wavelet inpainting aim to solve a very similar task: Assuming that, due to transmission or storage error, some coefficients of the wavelet representation of an image are lost, the aim is to reconstruct an image that fits to the known coefficients and minimizes the TV functional. In our terminology, given a suitable basis $(\phi_n)_{n\in\mathbb{N}}$ of $L^2(\Omega)$ and a source image u_0, this means to solve

$$\min_{u\in L^2(\Omega)} \mathrm{TV}(u) + \mathcal{I}_V(u)$$

with

$$V = \{u \in L^2(\Omega) \,|\, (u,\phi_n)_{L^2} = (u_0,\phi_n)_{L^2} \text{ for all } n \in M\}$$

and M being the index set of known coefficients. In [30] existence of a solution for this problem was established in function space setting under the assumption that $\Omega = \mathbb{R}^2$ and only finitely many coefficients are unknown. Numerical solution strategies for this, and a similar model with L^2 data fit, were presented in [30, 27, 71, 56]. In [73] the same model using non-local TV regularization was considered. In [36] the authors present the statement and numerical solution of a TV - wavelet denoising scheme whose formulation is also quite similar to those methods: Motivated by denoising with wavelet thresholding, the authors propose to minimize the TV functional subject to equality constraints on all wavelet coefficients with absolute value above a certain threshold.

However, even when considered solely as method for wavelet constraint optimization, our work differs significantly from the ones cited above. First of all we use the total generalized variation functional of arbitrary order as regularization, which is a non trivial generalization of the TV based models. Also, we are able to establish existence of a solution and optimality conditions in the case Ω is a *bounded Lipschitz domain* and using *natural boundary extension* also in function space setting. Additionally, we allow *infinitely many* wavelet coefficients to be unbounded and possible *interval constraints*. We also formulate the model for general biorthogonal wavelet bases from the very beginning and our numerical solution scheme is different to the ones of previous works. Let us point out however, that the assumptions of our work include the problem of wavelet inpainting, thus it can also be seen as a generalization of methods of [30, 36] using arbitrary order TGV regularization. Also our numerical solution scheme can be applied to these problems as is.

Further methods mainly focused on the concealment of wavelet data error due to transmission are the works of [48, 5, 31]. In [69] the aim is the reduction of artifacts due to tile separation of the image. We also refer to [54] for a post processing method that attempts to improve reconstruction quality by reapplication of JPEG 2000 on shifted versions of the image.

The JPEG 2000 standard

We will now briefly discuss the JPEG 2000 compression procedure. For more information, we refer to [51, 64, 66, 42] and the references therein.

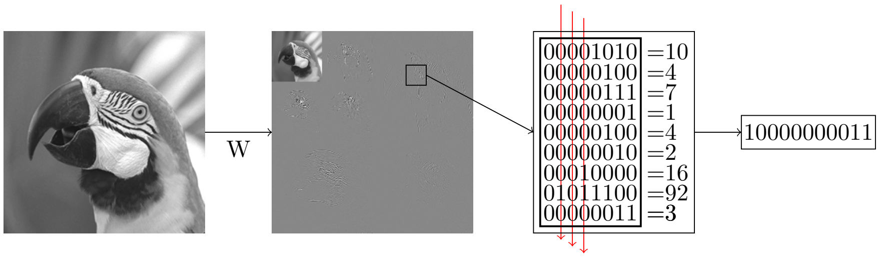
Figure 13 gives a schematic overview of some main steps for JPEG 2000 compression that will be discussed in the following. As first step, the image is split into color components and further into subunits (tiles), where each subunit undergoes the same compression process. Next, a discrete wavelet transformation of arbitrary order is applied to each subunit. Two types of wavelet transformation are possible within the standard, the *Cohen-Daubechies-Feauveau (CDF) 9/7* and the *LeGall 5/3* wavelet transform (see [64, 42]). The numbers 9/7 and 5/3 indicate the support length of related filters. The resulting coefficients are then quantized depending on their importance for visual image quality. The values used for quantization are uniform on each subband, i.e., on each direction dependent part of each resolution level of each subunit, and they can be obtained from the compressed code stream.

The quantized coefficients are then further split into different kinds of subunits, resulting finally in a set of code blocks. Each of these code blocks then undergoes a bit-level encoding consisting of three different passes. Starting from the highest nonzero bit-level, these three passes are repeated until the lowest bit-level has been encoded. This generates, for each code block, an independent bit-stream. According to mean-squared error estimations with respect to the original image, truncation points are then defined for each code block. Finally, the data is reorganized into a bit-stream that can be truncated at various lengths. The point at which this bit-stream is finally truncated determines the compression rate of the image. Due to the mean-squared error estimations at the definition of the truncation points, the truncation of the final bit-stream is expected to be optimal in terms of PSNR (see [51, Section 10.5.2],[42, Section J.10]).

In the compressed JPEG 2000 file, the amount of information available in the bit stream of one code block depends on the importance of the information in the code block for the PSNR rate. Thus, if, due to truncation, for one code block no bit-level information is left at all, the only information we can determine is that skipping information about its coefficients resulted in a better estimated PSNR rate than for other code blocks. However, since the original image is not known, we cannot use this information to obtain any estimation on its coefficients. Each individual coefficient could have been arbitrary high, as long as the overall information of the code block was less important for the PSNR value.

However, if at least one bit of coefficient information is left for a given code block, we can determine a bounded error interval for each of its coefficients as follows: As already explained, during compression, each code block is transformed into a bit stream by repeating three passes, the *significance propagation pass*, the *magnitude refinement pass* and the *cleanup pass* (see [42, Annex D]). Starting at the highest bit level, each pass follows predefined rules whether it encodes a particular bit or not. Thus, extracting the truncated, nonempty bit stream and information about which pass has been performed last before truncation from the compressed file, we can determine, for each coefficient of the code block, up to which bit level it has been encoded, i.e. its precision.

Using this knowledge, we can define a source value together with a (bounded) error interval for each coefficient of the code block, exactly how we could do for a JPEG compressed image. Thus, given any code block with non-zero information, we can define the set of its possible source coefficients again by bounded interval restrictions. As we will see in our numerical experiments, this is possi-



Uncompressed image Transformed image Bit-level coding JPEG2000 file

Figure 13: Selected steps of JPEG 2000 compression including bit-level coding for one image component. See also [51, Figures 10.16,10.17].

ble for sufficiently many code blocks to keep the set of possible source images small.

Note that, in contrast to JPEG compression, the JPEG 2000 standard does not include explicit color subsampling. However, since due to the wavelet transformation the image is composed into a low-resolution and a detail-part, subsampling is still possible by skipping the rank one detail coefficients for the color components.

5.3.1 Modeling

When considering a model to improve the decompression process of any given JPEG 2000 image, the important information the reader should remember from the sketch of the JPEG 2000 compression procedure above, is that, given any JPEG 2000 file, we can again define a maximal error interval for each of its coefficients for the representation with respect to a wavelet basis. Due to the specifications of the standard, in contrast to the JPEG model, these intervals may now also be unbounded. However, we will see that we can still obtain a bound on sufficiently many of these coefficients. Thus set of possible coefficient data, described by

$$D = \{z^c_{i,j} \,|\, z^c_{i,j} \in J^c_{i,j}\},$$

with $J^c_{i,j}$ nonempty, closed, but not necessarily bounded intervals, is sufficiently small. Given now W to be an accurate wavelet transformation operator, the set of possible source images for any given JPEG 2000 compressed file can thus be described as

$$U_D = \{u \,|\, Wu \in D\}.$$

Thus the problem setting of section 4 is applicable, provided that the wavelet operator modeling the wavelet transform of the JPEG 2000 compression procedure is indeed given by Riesz bases with its dual basis contained in $\mathrm{BV}(\Omega, \mathbb{R}^3)$. It will be the aim of the following pages to show that this is the case: We will sketch the construction of two biorthogonal wavelet bases modeling the two possible types of wavelet transform used within the JPEG 2000 standard. Besides that, we also want to make the reader familiar with wavelet transforms, as it is necessary to understand the JPEG 2000 decompression model.

Wavelet transformation for JPEG 2000 compression

The aim of this paragraph is to give a brief introduction to wavelet transforms and to construct biorthogonal wavelet bases that model the wavelet transform within JPEG 2000 compression. Let us point out, however, that in the introduction we will solely focus on the steps necessary to understand and construct the wavelet bases we need for our application and by no means consider this a complete introduction to wavelet analysis. For this purpose, we refer to [34, 51, 49]. Further we refer to [32, 33] for the theoretical results that will be presented and used.

The paragraph is divided into five parts:

- Introduction to wavelet transforms
- Conditions for constructing biorthogonal wavelet bases
- Construction of the wavelet bases for JPEG 2000 coding
- Two dimensional wavelet transform and extension
- Regularity

Introduction to wavelet transforms Given a suitable function ψ, the *mother wavelet*, one can construct an orthonormal basis $(\psi_{j,k})_{j,k\in\mathbb{Z}}$ by translations and dilations of ψ as

$$\psi_{j,k}(x) = 2^{-j/2}\psi(2^{-j}x - k), \quad j,k \in \mathbb{Z}. \tag{88}$$

See figure 14, top row, for a visualization. The basic idea is now to choose ψ such that the representation of image information with respect to this basis is sparse, thus allowing a high compression rate by further encoding. Now if the dilation factor j is fixed, one can think of the coefficients $(u, \psi_{j,k})_{k\in\mathbb{Z}}$ as sparse representation of the image details of u at a certain scale. This leads to the concept of multiscale analysis, where one defines a sequence of spaces $\ldots V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \subset \ldots$, that correspond to different scales. One requirement (among others) is that the space V_0 contains a function ϕ, often called *scaling function*, such that the functions $(\phi_{0,k})_{k\in\mathbb{Z}}$,

$$\phi_{0,k}(x) = \phi(x-k),$$

constitute an orthonormal basis of V_0. Additional conditions on the space V_j then imply that, for fixed $j \in \mathbb{Z}$, the functions $(\phi_{j,k})_{k\in\mathbb{Z}}$,

$$\phi_{j,k}(x) = 2^{-j/2}\phi(2^{-j}x - k), \tag{89}$$

build an orthonormal basis of V_j.

The other way around, given a fixed, suitable scaling function ϕ and a resulting sequence of spaces $V_j := \overline{\text{span}\{\phi_{j,k} \,|\, k \in \mathbb{Z}\}}$, one can construct a mother wavelet ψ such that the functions $(\psi_{j,k})_{k\in\mathbb{Z}}$ constitute an orthonormal basis of W_j, defined as the orthogonal complement of V_j in V_{j-1}, i.e.

$$V_{j-1} = V_j \otimes W_j, \quad V_j \perp W_j.$$

Having defined the resulting scaling and wavelet bases

$$(\phi_{j,k})_{j,k\in\mathbb{Z}} \text{ and } (\psi_{j,k})_{j,k\in\mathbb{Z}},$$

as in (89),(88), given any function $f \in L^2(\mathbb{R})$, and any $R \in \mathbb{Z}$, one can write

$$\begin{aligned} f = & \sum_{j,k\in\mathbb{Z}} (\psi_{j,k}, f)\psi_{j,k} \\ = & \sum_{j>R,k\in\mathbb{Z}} (\psi_{j,k}, f)\psi_{j,k} + \sum_{j\leq R,k\in\mathbb{Z}} (\psi_{j,k}, f)\psi_{j,k} \\ = & \sum_{k\in\mathbb{Z}} (\phi_{R,k}, f)\phi_{R,k} + \sum_{j\leq R,k\in\mathbb{Z}} (\psi_{j,k}, f)\psi_{j,k} \\ = & f_1 + f_2, \end{aligned} \tag{90}$$

where one can think of $f_1 = P_{V_R}(f)$ as a representation of f at a certain scale (see figure 14, middle row) and f_2 containing all the detail informations of f at finer scales. It follows directly that

$$(\phi_{R,k})_{k\in\mathbb{Z}} \cup (\psi_{j,k})_{j\leq R,k\in\mathbb{Z}} \tag{91}$$

is an orthonormal bases of $L^2(\mathbb{R})$. In the application to image processing, this basis allows to decompose an image u, given at a certain resolution, as a low resolution image u_0 and a (hopefully) sparse linear combination of wavelet functions $\psi_{j,k}$ containing the high resolution information (see figure 14, bottom). However, to realize this idea in practice, several additional features of the wavelet basis are desirable, such as vanishing moments, compact support or symmetry. To achieve this, an efficient approach, as proposed in [32], is to loosen orthonormality restrictions on the wavelet basis. We will later on state sufficient conditions to construct such a general wavelet basis.

Let us first continue this brief motivation by explaining how a decomposition as in equation (90) can be calculated efficiently. Now since $\phi \in V_0 \subset V_{-1}$, there exist $(h_k)_{k\in\mathbb{Z}}$ such that

$$\phi(x) = \sum_{k\in\mathbb{Z}} h_k\sqrt{2}\phi(2x - k). \tag{92}$$

These coefficients (also called filters) $(h_k)_{k\in\mathbb{Z}}$ form the basis of the numerical calculation of the wavelet transform. Given a signal f, the coefficients for the decomposition of f at a certain scale j can be obtained recursively from the coefficients of a representation of f with respect to $(\phi_{j-1,k})_{k\in\mathbb{Z}}$ at a finer scale, i.e.

$$(f, \phi_{j,k}) = \sum_{l\in\mathbb{Z}} h_l(f, \phi_{j-1,l+2k}), \quad (f, \psi_{j,k}) = \sum_{k\in\mathbb{Z}} (-1)^l h_{1-l}(f, \phi_{j-1,l+2k}), \tag{93}$$

for all $k \in \mathbb{Z}$. Similarly, the representation of f at a fine scale $j - 1$ can be obtained from its wavelet decomposition at the coarser scale j as

$$(f, \phi_{j-1,m}) = \sum_{k\in\mathbb{Z}} \left[h_{m-2k}(f, \phi_{j,k}) + (-1)^{m-2k}h_{1-(m-2k)}(f, \psi_{j,k})\right], \tag{94}$$

$m \in \mathbb{Z}$.

Now in the discrete setting, a function f is assumed to be given at a certain scale, i.e. $f \in V_{R_0}$ for $R_0 \in \mathbb{Z}$, thus the coefficients $(f, \phi_{R_0,k})_{k\in\mathbb{Z}}$ are known. Then, since $W_j \perp V_{R_0}$ for all $j \leq R_0$, according to equation (90), f can be decomposed to a lower resolution scale $R_1 > R_0$, $R_1 \in \mathbb{Z}$, as

$$f = \sum_{k\in\mathbb{Z}} (\phi_{R_1,k}, f)\phi_{R_1,k} + \sum_{R_0<j\leq R_0, k\in\mathbb{Z}} (\psi_{j,k}, f)\psi_{j,k}.$$

Note that due to (93) this decomposition requires only the filter $(h_k)_{k\in\mathbb{Z}}$. Similarly, f can be reconstructed from the low-scale data

$$(\phi_{R_1,k}, f), (\psi_{j,k}, f) \quad R_0 < j \leq R_1, k \in \mathbb{Z}, \tag{95}$$

using equation (94), again just by knowing the filters $(h_k)_{k\in\mathbb{Z}}$.

Thus, when applying the wavelet decomposition and reconstruction to a signal, only the filters $(h_k)_{k\in\mathbb{Z}}$ are needed, but not the corresponding multiresolution basis functions as in (91).

Indeed, a typical way to construct a wavelet basis is to determine suitable filters $(h_k)_{k\in\mathbb{Z}}$ satisfying certain properties ensuring that the corresponding scaling function ϕ and mother wavelet ψ allow to define an orthonormal bases. If necessary, these function then be constructed by the identities

$$\hat{\phi}(\xi) = (2\pi)^{-1/2} \prod_{j=1}^{\infty} m_0(2^{-j}\xi), \tag{96}$$

where $\hat{\phi}$ is the Fourier-transform of ϕ,

$$m_0(\xi) = 2^{-1/2} \sum_{n\in\mathbb{Z}} h_n \exp(-in\xi) \tag{97}$$

and

$$\psi(x) = \sqrt{2} \sum_{n\in\mathbb{Z}} (-1)^n h_{-n+1} \phi(2x - n). \tag{98}$$

We will now focus on conditions for filter banks ensuring that their corresponding scaling and wavelet functions as in equations (96) and (98) constitute not an orthonormal but a Riesz basis.

Note that for simplicity we will henceforth write $(h_n)_{n\in\mathbb{Z}}$ or even $(h_n)_n$ to denote filter banks but, in any case, assume that only finitely many of the h_n are nonzero. This is the case for most wavelets considered in practice, in particular for the LeGall 5/3 and CDF 9/7 wavelets used for JPEG 2000 coding.

Conditions for constructing biorthogonal wavelet bases As shown in [32], replacing the orthnormality restriction on the bases $(\phi_{j,k})_{j,k\in\mathbb{Z}}$, $(\psi_{j,k})_{j,k\in\mathbb{Z}}$ by requiring them only to be a Riesz bases, still generates a multiscale decomposition framework as introduced in the previous subsection. This generalization allows more flexibility in the construction of the filters $(h_k)_{k\in\mathbb{Z}}$ and, for example, to obtain a symmetric, compactly supported wavelet basis $(\psi_{j,k})_{j,k\in\mathbb{Z}}$ with an arbitrary amount of vanishing moments. However, in this case the reconstruction cannot be performed with the same filter as the decomposition, implying the necessity to construct dual filters $(\tilde{h}_k)_{k\in\mathbb{Z}}$. These dual filters then

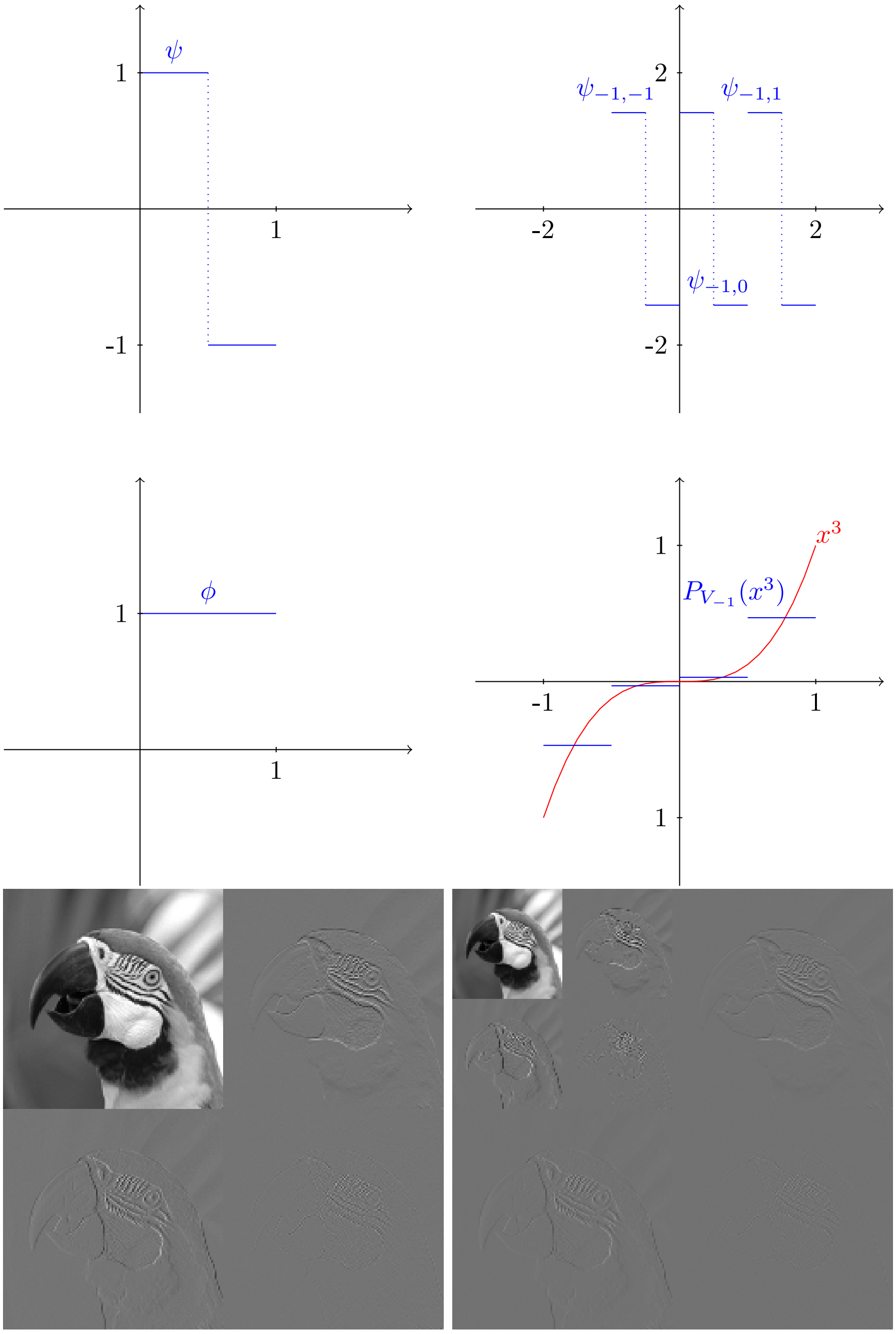


Figure 14: Multiresolution examples for Haar wavelet. Top: Haar wavelet and some wavelet basis elements. Middle: Haar scaling function and Projection onto V_{-1} of the function $x \mapsto x^3$. Bottom: Wavelet decomposition of Bird image using one and two resolution levels.

allow to define biorthogonal functions to $(\phi_{j,k})_{j,k\in\mathbb{Z}}$ and $(\psi_{j,k})_{j,k\in\mathbb{Z}}$ denoted by $(\tilde{\phi}_{j,k})_{j,k\in\mathbb{Z}}$ and $(\tilde{\psi}_{j,k})_{j,k\in\mathbb{Z}}$, respectively, just by replacing $(h_k)_{k\in\mathbb{Z}}$ by $(\tilde{h}_k)_{k\in\mathbb{Z}}$ in the equations (96),(97),(98).

Let us at first state the basic theorem given in [32, Theorem 3.8] which ensures that given filters $(h_k)_{k\in\mathbb{Z}}, (\tilde{h}_k)_{k\in\mathbb{Z}}$ allow to construct biorthogonal wavelet bases $(\psi_{j,k})_{j,k\in\mathbb{Z}}, (\tilde{\psi}_{j,k})_{j,k\in\mathbb{Z}}$.

Proposition 5.10. *Let $(h_k)_{k\in\mathbb{Z}}, (\tilde{h}_k)_{k\in\mathbb{Z}}$ be finite real sequences satisfying*

$$\sum_{n\in\mathbb{Z}} h_n \tilde{h}_{n+2k} = \delta_{k,0}. \tag{99}$$

Define

$$\begin{aligned} m_0(\xi) = &\ \ 2^{-1/2}\sum_{n\in\mathbb{Z}} h_n \exp(-in\xi), \\ \tilde{m}_0(\xi) = &\ \ 2^{-1/2}\sum_{n\in\mathbb{Z}} \tilde{h}_n \exp(-in\xi), \end{aligned} \tag{100}$$

$$\begin{aligned} \hat{\phi}(\xi) = &\ \ (2\pi)^{-1/2}\prod_{j=1}^{\infty} m_0(2^{-j}\xi), \\ \hat{\tilde{\phi}}(\xi) = &\ \ (2\pi)^{-1/2}\prod_{j=1}^{\infty} \tilde{m}_0(2^{-j}\xi). \end{aligned} \tag{101}$$

Suppose that, for some $C > 0$ and $\epsilon > 0$

$$\begin{aligned} |\hat{\phi}(\xi)| \leq &\ \ C(1+|\xi|)^{-1/2-\epsilon} \\ |\hat{\tilde{\phi}}(\xi)| \leq &\ \ C(1+|\xi|)^{-1/2-\epsilon}. \end{aligned} \tag{102}$$

Define ψ as in equation (98) and $\tilde{\psi}$ accordingly with $(h_k)_{k\in\mathbb{Z}}$ and ϕ replaced by $(\tilde{h}_k)_{k\in\mathbb{Z}}$ and $\tilde{\phi}$, respectively. Then the functions $\psi_{j,k} = 2^{-j/2}\psi(2^{-j}x-k), j, k \in \mathbb{Z}$ constitute a frame in $L^2(\mathbb{R})$, i.e. there exists constants $A, B > 0$ such that

$$A\|u\|_{L^2}^2 \leq \sum_{j,k\in\mathbb{Z}} |(u,\psi_{j,k})|^2 \leq B\|u\|_{L^2}^2.$$

Further, also $(\tilde{\psi})_{j,k\in\mathbb{Z}}$ constitutes a frame and for any $f \in L^2(\mathbb{R})$,

$$f = \sum_{j,k\in\mathbb{Z}} (f,\tilde{\psi})\psi_{j,k} = \sum_{j,k\in\mathbb{Z}} (f,\psi_{j,k})\tilde{\psi}_{j,k},$$

where the series converges strongly. Moreover, the $(\psi)_{j,k\in\mathbb{Z}}, (\tilde{\psi})_{j,k\in\mathbb{Z}}$ constitute dual Riesz bases of $L^2(\mathbb{R})$ with

$$(\psi_{j,k}, \tilde{\psi}_{j',k'}) = \delta_{j,j'}\delta_{k,k'}$$

if and only if

$$\int_{\mathbb{R}} \phi(x)\tilde{\phi}(x-k) = \delta_{k,0}. \tag{103}$$

It follows easily from the definition of ϕ and $\tilde{\phi}$ as in (101), that both satisfy a scaling equation as (92). This allows to construct a multiresolution framework from both ϕ and $\tilde{\phi}$ (see also [34, Chapter 8.3.2] and [51, Chapter 7]). Using estimations obtained during the proof of Proposition 5.10 as in [32, Theorem 3.8], one finally gets the following result:

Corollary 5.2. *Assume that $(h_k)_{k\in\mathbb{Z}}, (\tilde{h}_k)_{k\in\mathbb{Z}}, \phi, \tilde{\phi}, \psi, \tilde{\psi}$ are given as in proposition 5.10 and that conditions* (99),(102) *and* (103) *are satisfied. With $\phi_{j,k}(x) = 2^{-j/2}\phi(2^{-j}x-k)$ and $\tilde{\phi}_{j,k}(x) = 2^{-j/2}\tilde{\phi}(2^{-j}x-k)$, define*

$$V_j = \overline{\operatorname{span}\{\phi_{j,k} | k \in \mathbb{Z}\}}, \quad \tilde{V}_j = \overline{\operatorname{span}\{\tilde{\phi}_{j,k} | k \in \mathbb{Z}\}}.$$

Then, for each $R \in \mathbb{Z}$, the sequences $(\phi_{R,k})_{k\in\mathbb{Z}}, (\phi_{R,k})_{k\in\mathbb{Z}}$ are a Riesz basis of $V_R, \tilde{V}_R$, respectively. Further the sequences $(\phi_{R,k})_{k\in\mathbb{Z}} \cup (\psi_{j,k})_{j\leq R, k\in\mathbb{Z}}$ and $(\tilde{\phi}_{R,k})_{k\in\mathbb{Z}} \cup (\tilde{\psi}_{j,k})_{j\leq R, k\in\mathbb{Z}}$ both are a Riesz basis of $L^2(\mathbb{R})$.

Thus, when trying to construct filters for a multiresolution decomposition framework, it is sufficient to check conditions (99),(102),(103). It then follows, similarly to (90), that any $f \in L^2(\mathbb{R})$ can be decomposed as

$$\begin{aligned} f = & \sum_{k\in\mathbb{Z}} (\tilde{\phi}_{R,k}, f)\phi_{R,k} + \sum_{j\leq R, k\in\mathbb{Z}} (\tilde{\psi}_{j,k}, f)\psi_{j,k} \\ = & \sum_{k\in\mathbb{Z}} (\phi_{R,k}, f)\tilde{\phi}_{R,k} + \sum_{j\leq R, k\in\mathbb{Z}} (\psi_{j,k}, f)\tilde{\psi}_{j,k}. \end{aligned} \tag{104}$$

for any $R \in \mathbb{Z}$. Again the decomposition and reconstruction process can be performed just in terms of the filters $(h_k)_{k\in\mathbb{Z}}, (\tilde{h}_k)_{k\in\mathbb{Z}}$ similar to equations (93),(94) by replacing $(h_k)_{k\in\mathbb{Z}}$ by $(\tilde{h}_k)_{k\in\mathbb{Z}}$ in (94).

As already mentioned, we only consider finite filters $(h_n)_{n\in\mathbb{Z}}$, $(\tilde{h}_n)_{n\in\mathbb{Z}}$. This affects also our scaling functions $\phi, \tilde{\phi}$:

Remark 5.5. *With the assumptions of proposition 5.10, take $N, \tilde{N} \in \mathbb{N}$ such that $h = (h_i)_{|i|\leq N}$, $\tilde{h} = (\tilde{h}_i)_{|i|\leq \tilde{N}}$. Then both ϕ and $\tilde{\phi}$ have compact support contained in $[-N, N]$ and $[-\tilde{N}, \tilde{N}]$, respectively.*

Proof. This follows from the definition of ϕ and $\tilde{\phi}$ as in proposition 5.10 and [32, Lemma 3.1]. □

Thus, the shorter the filter length, the smaller the support of the resulting scaling function. Since the mother wavelets ψ and $\tilde{\psi}$ are linear combinations of the scaling functions as in equation (98), they are also compactly supported with their support length depending on the filter length.

The following result from [32, Corollary 5.2] shows necessity of a minimal filter support length given that one wants to obtain a particular smoothness of the wavelet basis:

Proposition 5.11. *Let $\psi, \tilde{\psi}$ be defined as in Proposition 5.10. If $(\psi)_{j,k\in\mathbb{Z}}, (\tilde{\psi})_{j,k\in\mathbb{Z}}$ constitute dual Riesz bases, then*

$$\begin{aligned} \psi \in \mathcal{C}^L(\mathbb{R}) \Rightarrow m_0(\xi) \textit{ is divisible by } & \quad (1+\exp(-i\xi))^{L+1} \\ \tilde{\psi} \in \mathcal{C}^L(\mathbb{R}) \Rightarrow \tilde{m}_0(\xi) \textit{ is divisible by } & \quad (1+\exp(-i\xi))^{L+1}. \end{aligned} \tag{105}$$

When it comes to the implementation of the wavelet transform, naturally, one is interested in filters having only finitely many - and moreover as few as possible - elements different from zero, which correspond to the basis functions having compact support. The above result shows that a given order of smoothness L for one of the wavelet functions implies at least $L+1$ elements of the corresponding filter to be nonequal to zero.

Propositions 5.10 and 5.11 will now be the basis for the construction of the JPEG 2000 wavelets.

Construction of the wavelet bases for JPEG 2000 coding As mentioned before, we are now seeking for suitable filters $(h_k)_{k\in\mathbb{Z}}, (\tilde{h}_k)_{k\in\mathbb{Z}}$ that allow the construction of a wavelet basis possessing additional regularity properties. Therefore, we want the filters to satisfy conditions (99),(102),(103). As one can easily check by equating coefficients, condition (99) is equivalent to

$$m_0(\xi)\tilde{m}_0(\xi) + m_0(\xi+\pi)\tilde{m}_0(\xi+\pi) = 1. \tag{106}$$

If we now require our scaling function to be symmetric, it follows that $m_0(-\xi) = m_0(\xi)$ and thus $m_0(\xi)$ has to be a polynomial in $\cos(\xi)$. Further, [32, Propsition 6.1] shows that in this case we can also assume $\tilde{m}_0$ to be symmetric and thus a polynomial in $\cos(\xi)$. If, additionally, we assume m_0 and $\tilde{m}_0$ to be divisible by $(1+\exp(-i\xi))^k$ and $(1+\exp(-i\xi))^{\tilde{k}}$, respectively (see condition (105)), this implies [32, Propsition 6.4] that all such solutions of (106) can be written as

$$\begin{aligned} m_0(\xi) &= (\cos(\xi/2))^{2k} p_0(cos(\xi)),\\ \tilde{m}_0(\xi) &= (\cos(\xi/2))^{2\tilde{k}} \tilde{p}_0(cos(\xi)), \end{aligned} \tag{107}$$

where

$$p_0(\cos(\xi))\tilde{p}_0(\cos(\xi)) = \sum_{n=0}^{L-1} \binom{L-1+n}{n} \sin(\xi/2)^{2n} + \sin(\xi/2)^{2L} P(\cos(\xi)), \tag{108}$$

and P is an odd polynomial and $L = k + \tilde{k}$. Our aim is thus to find functions $m_0, \tilde{m}_0$ satisfying (107), knowing that any such functions satisfy condition (99). Note that we have free choice of $k, \tilde{k}$, P and how to factorize the resulting polynomial in (108) to obtain $p_0, \tilde{p}_0$.

A first and easy choice would be $P \equiv 0$. Then, equation (108) reduces to

$$p_0(\cos(\xi))\tilde{p}_0(\cos(\xi)) = \sum_{n=0}^{L-1} \binom{L-1+n}{n} \sin(\xi/2)^{2n}. \tag{109}$$

and the corresponding filters contain the class of *spline filters* [32, Section 6.A]. By further restricting ourselves to $k = \tilde{k} = 1$ and $\tilde{p}_0 \equiv 1$, equation (109) implies

$$p_0 = \left(1 + 2\sin(\xi/2)^2\right) = (2 - \cos(\xi)).$$

With that, the polynomials $m_0, \tilde{m}_0$ can be written as

$$\tilde{m}_0 = \frac{1}{2}(1+\cos\xi) = \left(\frac{1}{4}\cos(-\xi) + \frac{1}{2} + \frac{1}{4}\cos(\xi)\right) \tag{110}$$

and

$$\begin{aligned} m_0 = & \quad \frac{1}{2}\left(1+\cos(\xi)\right)\left(2-\cos(\xi)\right) \\ = & \quad -\frac{1}{8}\cos(-2\xi)+\frac{1}{4}\cos(-\xi)+\frac{3}{4}+\frac{1}{4}\cos(\xi)-\frac{1}{8}\cos(2\xi). \end{aligned} \tag{111}$$

Thus we have $2^{-1/2}\tilde{h}_0 = \frac{1}{2}$, $2^{-1/2}\tilde{h}_{\pm 1} = \frac{1}{4}$ and $2^{-1/2}h_0 = \frac{3}{4}$, $2^{-1/2}h_{\pm 1} = \frac{1}{4}$, $2^{-1/2}h_{\pm 2} = -\frac{1}{8}$ (cf. [32, Table 6.1]).

Remark 5.6. *Note that, for the numerical realization, the $(h_n)_n$ are multiplied by $2^{-1/2}$ while the $(\tilde{h}_n)_n$ are multiplied by $2^{1/2}$ (this amounts the choice $\tilde{p}_0 = 2^{1/2}$ in the previous construction) to get an implementation friendly form. This choice yields the same biorthogonal wavelet framework as constructed here and corresponds to the LeGall 5/3 filters used in JPEG 2000 as lossless wavelet transformation (cf. [64]).*

Another possibility would be to again choose $P \equiv 0$ in condition (108), and then try to get a nontrivial factorization of p_0 and $\tilde{p}_0$ corresponding to condition (109) in order to obtain filters $(h_k)_{k\in\mathbb{Z}}, (\tilde{h}_k)_{k\in\mathbb{Z}}$ with similar length. Since we want our filters to have real coefficients, and thus $p_0, \tilde{p}_0$ to be real polynomials, $L = 4$ is the least number allowing a nontrivial factorization of (109) in such a way.

Factorizing (109) amounts to find the zeros of the polynomial $20x^3 + 10x^2 + 4x + 1$. Since this polynomial possesses only one real root, there is only one nontrivial factorization possible. Denoting $\hat{x}$ its real root, we can write

$$p_0(\cos(\xi))\tilde{p}_0(\cos(\xi)) = \quad C(\sin^2(\xi/2) - \hat{x})q(\sin^2(\xi/2)), \tag{112}$$

where $C \in \mathbb{R}$ and q is a polynomial of order 2 possessing no real roots. It is now left to choose either k or $\tilde{k}$ and which of the two factors in (112) to assign to p_0 or $\tilde{p}_0$. As mentioned we want the length of the filters as close as possible, thus we choose $k = 2$ implying $\tilde{k} = 2$. If we set $\tilde{p}_0 = C(\sin^2(\xi/2) - \hat{x})$ where $C \approx 5.841$ we obtain exactly the second class of filters used in the JPEG 2000 standard (cf. [64]), which are multiples of the filters corresponding to Cohen-Daubechies-Wavelets (CDF) 9/7 wavelets (see [32, Table 6.2]).

For both the LeGall 5/3 and the CDF 9/7 wavelets, one still has to check that conditions (102),(103) are satisfied. This can be done either directly or by using one of the equivalent conditions as given in [32, Theorems 4.3,4.9]. However, we refer to [32, Section 6] for a discussion on that.

Remark 5.7. *Note that for both filter types used to construct the LeGall 5/3 and CDF 9/7 wavelets we have $\sum_n h_n = \sum_n \tilde{h}_n = \sqrt{2}$ which, by construction, implies that $m_0(0) = \tilde{m}_0(0) = 1$ and, consequently,*

$$\int_{\mathbb{R}} \phi(x)\,\mathrm{d}x = (2\pi)^{1/2}\hat{\phi}(0) = 1$$

as well as

$$\int_{\mathbb{R}} \tilde{\phi}(x)\,\mathrm{d}x = (2\pi)^{1/2}\hat{\tilde{\phi}}(0) = 1.$$

Accordingly, the implementation friendly, modified filters for the JPEG 2000 standard satisfy

$$\sum_n h_n = 1$$

as well as

$$\sum_n \tilde{h}_n = 2.$$

Two dimensional wavelet transform and extension When applying the wavelet transform corresponding to the CDF 9/7 or LeGall 5/3 filters, as derived in the previous passage, to a function u modeling an image, two issues arise. First of all, the function u will not be defined on the whole space, but rather on a bounded domain. Thus one needs to restrict the wavelet decomposition. Secondly, the function u will be defined on a subset of $\mathbb{R}^2$, thus one needs to extend the wavelet decomposition to the two dimensional setting. The aim of this passage is to adapt the previously defined wavelets accordingly.

First we consider the restriction to a bounded domain, in particular the unit interval $[0,1]$. Intuitively, one wants to extend a given function $u \in L^2([0,1])$ to $L^2(\mathbb{R})$ in order to apply the wavelet decomposition. There are several ways to do so. A straightforward choice would be to extend the function by 0. However, this introduces artificial jump discontinuities at the boundary of $[0,1]$ resulting in large wavelet coefficients close to the boundary, even if the function u is smooth. Periodic extension suffers from similar drawbacks. Perhaps a more natural way to extend the function beyond $[0,1]$ is to consider symmetric extension by reflection on the boundary. This is a very common choice of boundary extension in mathematical imaging, since it does not introduce additional discontinuities at the boundary. Extending any given function $u \in L^2([0,1])$ that way, we thus seek basis functions $(\eta_k)_{k\in\mathbb{Z}}$ of $L^2([0,1])$ satisfying

$$\int\limits_{[0,1]} u(x)\eta_k(x)\,\mathrm{d}x = \int\limits_{\mathbb{R}} u(x)\chi_k(x)\,\mathrm{d}x$$

where $(\chi_k)_{k\in\mathbb{Z}}$ is a Riesz bases associated with one of the JPEG 2000 wavelets and a certain scale. It has been shown in [33, Section 2] that this indeed can be done: Given any reasonably decaying function f on $\mathbb{R}$, its *folded* version can be defined as

$$f^{\text{fold}}(x) = \sum_{n\in\mathbb{Z}} [f(x-2n) + f(2n-x)]. \tag{113}$$

Note that this folded version now can be seen as a symmetric extension of a function defined solely on $[0,1]$, i.e. it is determined just by its values in $[0,1]$ and satisfies, for all $x \in \mathbb{R}, k \in \mathbb{Z}$,

$$\begin{aligned} f^{\text{fold}}(-x) &= \quad f^{\text{fold}}(x), \\ f^{\text{fold}}(x+2k) &= \quad f^{\text{fold}}(x). \end{aligned} \tag{114}$$

Further, for any function $f \in L^2(\mathbb{R})$ decaying sufficiently fast and g the symmetric extension of a function in $L^2([0,1])$, it follows

$$\int\limits_{[0,1]} f^{\text{fold}}(x)g(x)\,\mathrm{d}x = \int\limits_{\mathbb{R}} f(x)g(x)\,\mathrm{d}x. \tag{115}$$

Given $(\phi_{j,k})_{j,k\in\mathbb{Z}}$, $(\psi_{j,k})_{j,k\in\mathbb{Z}}$ and $(\tilde{\phi}_{j,k})_{j,k\in\mathbb{Z}}$, $(\tilde{\psi}_{j,k})_{j,k\in\mathbb{Z}}$ to be basis functions for a multiresolution framework and their duals, respectively, they can be folded as in equation (113) to obtain $(\phi^{\text{fold}}_{j,k})_{j,k\in\mathbb{Z}}$, $(\psi^{\text{fold}}_{j,k})_{j,k\in\mathbb{Z}}$, $(\tilde{\phi}^{\text{fold}}_{j,k})_{j,k\in\mathbb{Z}}$, $(\tilde{\psi}^{\text{fold}}_{j,k})_{j,k\in\mathbb{Z}}$. Following [33, Section 2] one can see that

$$\begin{aligned} \psi^{\text{fold}}_{j,k} &= \quad 0 \quad \text{if} \quad & j \geq 2, k \in \mathbb{Z}, \\ \psi^{\text{fold}}_{1,k} &= \quad \psi^{\text{fold}}_{1,0} \quad \text{for all} \quad & k \in \mathbb{Z}, \\ \phi^{\text{fold}}_{j,k} &= \quad 2^{j/2} \quad \text{if} \quad & j \geq 1, k \in \mathbb{Z}, \end{aligned} \tag{116}$$

and due to the symmetry properties of $\phi^{\text{fold}}, \psi^{\text{fold}}$ that

$$\begin{aligned} \phi^{\text{fold}}_{-j,k+2^{j+1}m} &= \phi^{\text{fold}}_{-j,k} \quad \text{and} \quad & \phi^{\text{fold}}_{-j,2^{j+1}-k} &= \phi^{\text{fold}}_{-j,k}, \\ \psi^{\text{fold}}_{-j,k+2^{j+1}m} &= \psi^{\text{fold}}_{-j,k} \quad \text{and} \quad & \psi^{\text{fold}}_{-j,2^{j+1}-k-1} &= \psi^{\text{fold}}_{-j,k}, \end{aligned} \tag{117}$$

for $j \geq 0$, and the same for $\tilde{\phi}^{\text{fold}}, \tilde{\psi}^{\text{fold}}$.

Thus for a basis representation with respect to $(\phi^{\text{fold}}_{j,k})_{j,k\in\mathbb{Z}}$ and $(\psi^{\text{fold}}_{j,k})_{j,k\in\mathbb{Z}}$ it suffices to consider $\{\phi^{\text{fold}}_{1,0}\} \cup (\phi^{\text{fold}}_{\substack{j\leq 0\\ 0\leq k\leq 2^j}})_{j,k\in\mathbb{Z}}$ and $\{\psi^{\text{fold}}_{1,0}\} \cup (\psi^{\text{fold}}_{j,k})_{\substack{j\leq 0\\ 0\leq k\leq 2^j-1}}$, respectively.

Indeed, normalizing the functions $(\phi^{\text{fold}}_{j,k})_{j,k}$ by defining $(\phi^{\text{fold,n}}_{j,k})_{j,k}$ as

$$\phi^{\text{fold,n}}_{-j,0} = \frac{1}{\sqrt{2}}\phi^{\text{fold}}_{-j,0}, \quad \phi^{\text{fold,n}}_{-j,2^j} = \frac{1}{\sqrt{2}}\phi^{\text{fold}}_{-j,2^j}, \quad \phi^{\text{fold,n}}_{1,0} = 1 \tag{118}$$

and the same for $(\tilde{\phi}^{\text{fold}}_{j,k})_{j,k}$, it has been shown in [33, Section 2] that $(\psi^{\text{fold}}_{j,k})$, $(\tilde{\psi}^{\text{fold}}_{j,k})$ again are dual Riesz bases and together with $(\phi^{\text{fold,n}}_{j,k})$, $(\tilde{\phi}^{\text{fold,n}}_{j,k})$ constitute a multiresolution framework.

In particular, for any decomposition level $R \leq 1$, the functions $(\phi^{\text{fold,n}}_{R,k})_{k\in\mathbb{Z}} \cup (\psi^{\text{fold}}_{j,k})_{j\leq R,k\in\mathbb{Z}}$ and $(\tilde{\phi}^{\text{fold,n}}_{R,k})_{k\in\mathbb{Z}} \cup (\tilde{\psi}^{\text{fold}}_{j,k})_{j\leq R,k\in\mathbb{Z}}$ are a Riesz basis of $L^2([0,1])$ and any $f \in L^2([0,1])$ can be written as

$$\begin{aligned} f = \quad & (f, \tilde{\phi}^{\text{fold,n}}_{1,0})\phi^{\text{fold,n}}_{1,0} + (f, \tilde{\psi}^{\text{fold}}_{1,0})\psi^{\text{fold}}_{1,0} + \sum_{j=0}^{-\infty} \sum_{k=0}^{2^{-j}-1} (f, \tilde{\psi}^{\text{fold}}_{j,k})\psi^{\text{fold}}_{j,k} \\ = \quad & \sum_{k=0}^{2^{-R}} (f, \tilde{\phi}^{\text{fold}}_{R,k})\phi^{\text{fold}}_{R,k} + \sum_{j=R}^{-\infty} \sum_{k=0}^{2^{-j}-1} (f, \tilde{\psi}^{\text{fold}}_{j,k})\psi^{\text{fold}}_{j,k}, \text{ if } R \leq 0. \end{aligned} \tag{119}$$

Also the filters for the recursive computation of the coefficients for the folded functions can be obtained from the original ones by folding them at the boundary, see again [33, Section 2]

Now given a wavelet basis of $L^2([0,1])$ and the corresponding filters for recursive computation of the coefficients, the next step is to obtain corresponding wavelet bases for $L^2([0,1]^2)$. The construction is based on [34, Section 10.1], where orthonormal bases of wavelets are considered. Let $(\phi_{j,k})_{\substack{j\leq 1\\ k\in M_j}}$, $(\psi_{j,k})_{\substack{j\leq 1\\ k\in L_j}}$ and $(\tilde{\phi}_{j,k})_{\substack{j\leq 1\\ k\in M_j}}$, $(\tilde{\psi}_{j,k})_{\substack{j\leq 1\\ k\in L_j}}$ be the folded scaling and wavelet functions for a multiresolution on $L^2([0,1])$, with $M_1 = \{0\}$, $L_1 = \{0\}$ and $M_j = \{0, \ldots, 2^{-j}\}$,

$L_j = \{0, \dots, 2^{-j} - 1\}$ for $j \leq 0$. We can define the functions

$$\begin{aligned} \Phi_{j,k_1,k_2}(x,y) = &\quad \phi_{j,k_1}(x)\phi_{j,k_2}(y) \\ \tilde{\Phi}_{j,k_1,k_2}(x,y) = &\quad \tilde{\phi}_{j,k_1}(x)\tilde{\phi}_{j,k_2}(y) \end{aligned} \tag{120}$$

which are, for given $j \leq 1$, tensor products of the basis functions of $V_j = \operatorname{span}\{\phi_{j,k} | k \in M_j\}$ and $\tilde{V}_j = \operatorname{span}\{\tilde{\phi}_{j,k} | k \in M_j\}$, respectively. It can thus be shown easily by standard results on tensor product spaces [70, Section 3.4], that $(\Phi_{j,k_1,k_1})_{k_1,k_2 \in M_j}$ and $(\tilde{\Phi}_{j,k_1,k_1})_{k_1,k_2 \in M_j}$ constitute Riesz bases for $V_j \otimes V_j$ and $\tilde{V}_j \otimes \tilde{V}_j$, respectively.

Now given any $f \in L^2([0,1]^2)$ we can approximate it by $f_1 f_2$ where $f_1, f_2 \in L^2([0,1])$. Taking into account a decomposition of f_1 and f_2 to an arbitrary level $R_0 \leq 1$ as in equation (104), for any $\epsilon > 0$ we can find R_1 sufficiently small and coefficients $\lambda^i_k, \mu^i_{j,k}$ such that, with

$$g_i = \sum_{k \in M_{R_0}} \lambda^i_k \phi_{R_0,k} + \sum_{\substack{R_1 < j \leq R_0 \\ k \in L_j}} \mu^i_{j,k} \psi_{j,k}, \quad i \in \{1,2\} \tag{121}$$

we have

$$\|f_1 f_2 - g_1 g_2\|_{L^2} \leq \|f_2\|_{L^2} \|f_1 - g_1\|_{L^2} + \|g_1\|_{L^2} \|f_2 - g_2\|_{L^2} < \epsilon. \tag{122}$$

Note that we sum only over a finite number of elements, thus a reordering as follows can be done without additional convergence analysis. Reordering the sums in $g_1 g_2$ and exploiting the fact that $V_m \subset V_{m-1}, W_m \subset V_{m-1}$, any element of V_m or W_m can be re-written as linear combination of elements in V_{m-1}, allows to write $g_1 g_2$ as

$$\begin{aligned} g_1(x)g_2(y) = &\quad \sum_{\mathbf{k} \in \mathbf{M}_{R_0}} \tilde{\lambda}_k \phi_{R_0,k_1}(x)\phi_{R_0,k_2}(y) + \sum_{\substack{R_1 < j \leq R_0 \\ \mathbf{k} \in \mathbf{L}^h_j}} \tilde{\mu}^1_{j,k} \phi_{j,k_1}(x)\psi_{j,k_2}(y) \\ + &\quad \sum_{\substack{R_1 < j \leq R_0 \\ \mathbf{k} \in \mathbf{L}^v_j}} \tilde{\mu}^2_{j,k} \psi_{j,k_1}(x)\phi_{j,k_2}(y) + \sum_{\substack{R_1 < j \leq R_0 \\ \mathbf{k} \in \mathbf{L}^d_j}} \tilde{\mu}^3_{j,k} \psi_{j,k_1}(x)\psi_{j,k_2}(y), \end{aligned} \tag{123}$$

where $\mathbf{k} = (k_1, k_2)$ and $\mathbf{M}_j = M_j \times M_j$, $\mathbf{L}^h_j = M_j \times L_j$, $\mathbf{L}^v_j = L_j \times M_j$, $\mathbf{L}^d_j = L_j \times L_j$ for $j \leq 1$. Thus, defining

$$\begin{aligned} \Psi^h_{j,\mathbf{k}}(x,y) = &\quad \phi_{j,k_1}(x)\psi_{j,k_2}(y) \\ \Psi^v_{j,\mathbf{k}}(x,y) = &\quad \psi_{j,k_1}(x)\phi_{j,k_2}(y) \\ \Psi^d_{j,\mathbf{k}}(x,y) = &\quad \psi_{j,k_1}(x)\psi_{j,k_2}(y) \end{aligned} \tag{124}$$

it follows that $g_1 g_2$ can be written as a linear combination of elements as in equation (124) and of $\Phi_{j,\mathbf{k}}$ as in equation (120). Thus, for any given scale $R_0 \leq 1$, the linear span of

$$(\Phi^h_{R_0,\mathbf{k}})_{\mathbf{k} \in \mathbf{M}_{R_0}} \cup (\Psi^h_{j,\mathbf{k}})_{\substack{j \leq R_0 \\ \mathbf{k} \in \mathbf{L}^h_j}} \cup (\Psi^v_{j,\mathbf{k}})_{\substack{j \leq R_0 \\ \mathbf{k} \in \mathbf{L}^v_j}} \cup (\Psi^d_{j,\mathbf{k}})_{\substack{j \leq R_0 \\ \mathbf{k} \in \mathbf{L}^d_j}}$$

is dense in $L^2([0,1]^2)$. The same construction can be done for the dual functions $(\tilde{\phi}_{j,k})_{\substack{j \leq 1 \\ k \in M_j}}, (\tilde{\psi}_{j,k})_{\substack{j \leq 1 \\ k \in L_j}}$. It remains to show that both basis form dual Riesz bases. But this can be deduced again with a density argument, and from their one dimensional equivalents being dual Riesz bases, by showing the equivalent characterization (5) of [72, Theorem I.9].

A recursive filter-based calculation scheme for the coefficients of the four two dimensional basis elements can be obtained from the corresponding one dimensional schemes by stepwise filtering first across the horizontal and then the vertical direction.

Regularity Since the general model we aim to apply for JPEG 2000 decompression requires some weak regularity assumptions on the Riesz basis (i.e. the dual basis must be contained in $\mathrm{BV}((0,1)^2)$), regularity of the wavelet basis functions is important. In particular, the Riesz bases corresponding to synthesis, denoted by $(\tilde{\Phi}^h_{R_0,\mathbf{k}})_{\mathbf{k}\in\mathbf{M}_{R_0}} \cup (\tilde{\Psi}^h_{j,\mathbf{k}})_{\substack{j\le R_0\\ \mathbf{k}\in\mathbf{L}^h_j}} \cup (\tilde{\Psi}^v_{j,\mathbf{k}})_{\substack{j\le R_0\\ \mathbf{k}\in\mathbf{L}^v_j}} \cup (\tilde{\Psi}^d_{j,\mathbf{k}})_{\substack{j\le R_0\\ \mathbf{k}\in\mathbf{L}^d_j}}$, with $R_0 \le 1$, are required to possess a certain amount of regularity. Let us first focus on the one dimensional basis functions defined one the whole reals.

Since all basis elements are finite linear combinations of translated, scaled versions of the scaling function $\tilde{\phi}$, it suffices to show regularity of $\tilde{\phi}$. In the case of LeGall 5/3 filters the scaling function $\tilde{\phi}$ is just a piecewise linear spline (see [32, Section 6.A] and note that there, ϕ and $\tilde{\phi}$ are switched), thus it is contained in $W^{1,1}(\mathbb{R})$. In the case of CDF 9/7 filters it has been shown in [66] that the scaling function corresponding to synthesis possesses a Sobolev regularity of more than 2, in particular is also contained in $W^{1,1}(\mathbb{R})$.

Now given that both one dimensional basis functions defined on $\mathbb{R}$ are sufficiently regular, it is straightforward that also their folded versions, as being again finite linear combinations of shifted versions of the original functions, and their tensor products are contained in $\mathrm{BV}((0,1)^2)$.

The resulting minimization problem

In this paragraph we finally formulate the process of artifact-free JPEG 2000 decompression according to our general modeling framework. As already presented in this subsection, for a given resolution level $R \le 1$ and each color component $c \in \{1,2,3\}$, we can construct the CDF 9/7 or LeGall 5/3 wavelet basis of $L^2((0,1)^2) \simeq L^2([0,1]^2)$, which we will denote by

$$({}_c\Phi^h_{R,\mathbf{k}})_{\mathbf{k}\in\mathbf{M}_R} \cup \big({}_c\Psi^h_{j,\mathbf{k}}\big)_{\substack{j\le R\\ \mathbf{k}\in\mathbf{L}^h_j}} \cup \big({}_c\Psi^v_{j,\mathbf{k}}\big)_{\substack{j\le R\\ \mathbf{k}\in\mathbf{L}^v_j}} \cup \big({}_c\Psi^d_{j,\mathbf{k}}\big)_{\substack{j\le R\\ \mathbf{k}\in\mathbf{L}^d_j}}. \tag{125}$$

Further we can define its dual basis by

$$({}_c\tilde{\Phi}^h_{R,\mathbf{k}})_{\mathbf{k}\in\mathbf{M}_R} \cup \big({}_c\tilde{\Psi}^h_{j,\mathbf{k}}\big)_{\substack{j\le R\\ \mathbf{k}\in\mathbf{L}^h_j}} \cup \big({}_c\tilde{\Psi}^v_{j,\mathbf{k}}\big)_{\substack{j\le R\\ \mathbf{k}\in\mathbf{L}^v_j}} \cup \big({}_c\tilde{\Psi}^d_{j,\mathbf{k}}\big)_{\substack{j\le R\\ \mathbf{k}\in\mathbf{L}^d_j}}, \tag{126}$$

that is contained in $\mathrm{BV}((0,1)^2)$. Note that we omit writing R color component dependent, even though this is clearly possible. Now, as shown in remark 1.2, we can use these component wise bases to define a related vector valued Riesz basis of $L^2((0,1)^2,\mathbb{R}^3)$, denoted by

$$(a_n)_{n\in\mathbb{N}},$$

and its dual basis, denoted by

$$(\tilde{a}_n)_{n\in\mathbb{N}},$$

that is contained in $\mathrm{BV}((0,1)^2,\mathbb{R}^3)$. Also, as explained at the beginning of this subsection, we can, again for each color component $c \in \{1,2,3\}$, obtain data intervals

$$({}_cJ_{R,\mathbf{k}})_{\mathbf{k}\in\mathbf{M}_j},\ ({}_cJ^h_{j,\mathbf{k}})_{\substack{j\le R\\ \mathbf{k}\in\mathbf{L}^h_j}},\ ({}_cJ^v_{j,\mathbf{k}})_{\substack{j\le R\\ \mathbf{k}\in\mathbf{L}^d_j}},\ ({}_cJ^v_{j,\mathbf{k}})_{\substack{j\le R\\ \mathbf{k}\in\mathbf{L}^d_j}} \tag{127}$$

from a given compressed JPEG 20000 file. We will also sum up these data intervals by

$$(J_n)_{n\in\mathbb{N}}$$

such that they fit to the $\mathbb{R}^3$ valued Riesz basis $(a_n)_{n\in\mathbb{N}}$.

With that, given a JPEG 2000 compressed image, we can again define its set of possible source images as

$$U_D = \{u \in L^2((0,1)^2,\mathbb{R}^3) \,|\, (a_n,u)_{L^2} \in J_n \text{ for all } n\in\mathbb{N}\},$$

or, more explicitly,

$$\begin{aligned} U_D = \Big\{ u \in L^2((0,1)^2,\mathbb{R}^3) \,\Big|\, ({}_c\Phi_{R,\mathbf{k}}, u^c)_{L^2} \in {}_cJ_{R,\mathbf{k}},\ \text{for all } k\in\mathbf{M}_{R_0}, \\ ({}_c\Psi^h_{j,\mathbf{k}}, u^c)_{L^2} \in {}_cJ^h_{j,\mathbf{k}},\ \text{for all } k\in\mathbf{L}^h_j,\quad ({}_c\Psi^v_{j,\mathbf{k}}, u^c)_{L^2} \in {}_cJ^v_{j,\mathbf{k}},\ \text{for all } k\in\mathbf{L}^v_j, \\ ({}_c\Psi^d_{j,\mathbf{k}}, u^c)_{L^2} \in {}_cJ^d_{j,\mathbf{k}},\ \text{for all } k\in\mathbf{L}^d_j,\quad \text{for all } j\le R,\ c\in\{1,2,3\}\Big\}. \end{aligned} \tag{128}$$

Having defined the source image set U_D, the infinite-dimensional problem setting related to JPEG 2000 decompression is to solve

$$\min_{u\in L^2((0,1)^2,\mathbb{R}^3)} \mathrm{TGV}^{\mathrm{k}}_{\alpha}(u) + \mathcal{I}_{U_D}(u). \tag{129}$$

Since we know already that the dual Riesz basis $(\tilde{a}_n)_{n\in\mathbb{N}}$ is contained in $\mathrm{BV}((0,1)^2,\mathbb{R}^3)$, and since, due to a rounding step as part of lossy JPEG 2000 compression, the data set U_D again has nonempty interior, in order to ensure the framework of section 4 to be applicable, it is only left to show boundedness of suitable intervals $J_{n_l}, 0 \le l \le \frac{k(k+1)}{2} - 1$. For that, first consider the following preparatory lemma:

Lemma 5.2. *Let $h \in L^1(\mathbb{R}^2)$ with compact support and $\int_{\mathbb{R}^2} h \neq 0$. Then, for any $k\in\mathbb{N}$, any nonzero constant $\delta\in\mathbb{R}$ and any polynomial $p:\Omega\to\mathbb{R}$, $p(\mathbf{x}) = \sum_{|\alpha|_1<k}\lambda_\alpha\mathbf{x}^\alpha$, of degree $k-1$,*

$$(h*p)(\delta\mathbf{l}) = \int_{\mathbb{R}^2} h(\mathbf{x}-\delta\mathbf{l})p(\mathbf{x})\,\mathrm{d}x = 0 \quad \textit{for all } \mathbf{l}\in\mathbb{N}_0^2, |\mathbf{l}|_1 < k \tag{130}$$

implies $p\equiv 0$.

Remark. *Note that by bold letters we denote elements of $\mathbb{R}^2$, i.e. $\mathbf{x} = (x_1,x_2) \in \mathbb{R}^2$, and we use the multi-index notation, i.e. $\mathbf{x}^\alpha = x_1^{\alpha_1}x_2^{\alpha_2}$ for $\alpha\in\mathbb{R}^2$, with $|\alpha|_1 = |(\alpha_1,\alpha_2)|_1 = \alpha_1+\alpha_2$.*

Proof of lemma 5.2. As will become clear during the proof, we can without loss of generality assume $\delta = 1$. Since h has compact support, the convolution is indeed well defined. Suppose that p is a polynomial of degree $k-1$ satisfying (130). Then we can assume that $\lambda_{\tilde{\alpha}} \neq 0$ for at least one $|\tilde{\alpha}|_1 = k-1$.

Now note that, due to $\partial_\alpha(h*p) = h*\partial_\alpha p$, also $h*p$ is a polynomial of degree $k-1$. It is a standard result in the context of finite elements that in this case, $(h*p)(\mathbf{l}) = 0$, for all $|\mathbf{l}|_1 < k$, implies $h*p \equiv 0$:

Denoting $q(\mathbf{x}) = (h*p)(\mathbf{x})$ it follows that $q(\cdot, 0) : \mathbb{R} \to \mathbb{R}$ is a polynomial of degree $k-1$ on $\mathbb{R}$ vanishing on k points $0, \ldots, k-1$, thus q is zero on all the hyperplane $H(x_1, x_2) = x_1$. Following [21, Lemma 3.1.10] this implies that we can write $q(x_1, x_2) = x_2 q_1(x_1, x_2)$ with $q_1 : \mathbb{R}^2 \to \mathbb{R}$ being a polynomial of degree $k-2$. Proceeding like this we finally can write $q(x_1, x_2) = C \prod_{i=0}^{k-2}(x_2 - i)$ with $C \in \mathbb{R}$, thus constant in the second coordinate. With that, $q(0, k-1) = 0$ implies that $C = 0$ and finally that $q = h*p \equiv 0$.

But this yields

$$0 = \partial_\alpha(h*p)(x) = \lambda_\alpha \int\limits_{\mathbb{R}^2} h(y) \,\mathrm{d}y$$

and, with $\int_{\mathbb{R}^2} h \neq 0$, $\lambda_\alpha = 0$ for all $|\alpha|_1 = k-1$ and thus a contradiction to $\lambda_{\tilde{\alpha}} \neq 0$. □

Our aim is now to show existence of a solution to problem (129) for a boundedness assumption at least as general as assuming that all intervals corresponding to the basis coefficients obtained as inner products with the scaling functions, i.e. corresponding to

$$(u^c,\ {}_c\Phi_{R,\mathbf{k}})_{L^2},$$

are bounded. This assumption is reasonable since in practice these coefficients represent a low-resolution version of the image, thus are important for reconstruction quality in terms of PSNR, and we can always get a bound on these coefficients from the compressed JPEG 2000 data (cf. the explanation at the beginning of this subsection). The generality on the other hand is also necessary since, in contrast to the $(u^c,\ {}_c\Phi_{R,\mathbf{k}})_{L^2}$, we cannot expect to get a bound on any particular $(u^c,\ {}_c\Psi^\lambda_{j,\mathbf{k}})_{L^2}$, $\lambda \in \{h, v, d\}$.

Additionally, in the wavelet based zooming model that we present later on, the intervals corresponding to the $(u^c,\ {}_c\Phi_{R,\mathbf{k}})_{L^2}$ will be bounded while the others will include all of $\mathbb{R}$.

To show existence, according to proposition 4.1, it is sufficient to find suitable elements $({}_c\Phi_{R,\mathbf{k}_i})_{0 \leq i \leq r}$, $r \in \mathbb{N}$ sufficiently large, such that the matrix as in corollary 4.1 has full rank. Since the elements $({}_c\Phi_{R,\mathbf{k}})_k$ are obtained from $({}_c\Phi_{R,\mathbf{0}})$ by shifting, lemma 5.2 will be very useful for this purpose. An existence result can thus be given as follows:

Proposition 5.12 (Existence for Wavelets)**.** *Set $k \in \mathbb{N}$ to be the order of the $\mathrm{TGV}^{\mathrm{k}}_\alpha$ functional. Given any biorthogonal Riesz basis $(a_n)_{n\in\mathbb{N}}$ of $L^2((0,1)^2, \mathbb{R}^3)$, explicitly defined by a component wise basis as in* (125)*, assume that R is small enough such that there exists $\mathbf{k}_0 = (k_0^1, k_0^2) \in \mathbf{M}_R$ with $\operatorname{supp}({}_c\Phi_{R,k_0^1+i,k_0^2+j}) \subset (0,1)^2$ for all $c \in \{1,2,3\}$, $0 \leq i, j < k$. Then, if the corresponding intervals*

$${}_cJ_{R,k_0^1+i,k_0^2+j},\ 0 \leq i,j < k, c \in \{1,2,3\} \textit{ are bounded,}$$

there exists a solution to (129)*.*

Proof. First note that $(a_n)_{n\in\mathbb{N}}$ is constructed from a component wise bases as in remark 1.2 and we will use component wise polynomials as described before corollary 4.1. The matrix with its entries being the inner products of elements of $(a_n)_{n\in\mathbb{N}}$, corresponding to the $({}_c\Phi_{R,k_0^1+i,k_0^2+j})$, with the component wise polynomials, is thus a block matrix of the form

$$B_k = \begin{pmatrix} M_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & M_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & M_3 \end{pmatrix}.$$

Consequently it has full rank if each M_c, $c \in \{1,2,3\}$, has full rank, with M_c being a matrix with elements

$$(({}_c\Phi_{R,k_0^1+i,k_0^2+j}, p_{l_1,l_2})_{L^2})_{0\le i,j,l_1,l_2<k},$$

where i,j varies along the rows, l_1, l_2 along the columns and $p_{l_1,l_2}(x,y) = x^{l_1}y^{l_2}$. To show that M_c, $c \in \{1,2,3\}$ has full rank, we suppose that a linear combination of its columns, with coefficients λ_{l_1,l_2}, is zero. Now, due to the support restriction,

$${}_c\Phi_{R,k_0^1+l_1,k_0^2+l_2}(x,y) = {}_c\Phi_{R,k_0^1,k_0^2}(x-2^{-R}l_1, y-2^{-R}l_2)$$

is satisfied. This implies that

$$({}_c\Phi_{R,k_0^1,k_0^2} * p)(2^{-R}\mathbf{l}) = \int\limits_{\mathbb{R}^2} {}_c\Phi_{R,k_0^1,k_0^2}(\mathbf{x}-2^{-R}\mathbf{l})p(\mathbf{x})\,\mathrm{d}x = 0 \text{ for all } \mathbf{l}\in\mathbb{N}_0^2, |\mathbf{l}|_1 < k,$$

where $p(x,y) = \sum_{|(l_1,l_2)|_1<k} \lambda_{l_1,l_2}x^{l_1}y^{l_2}$ is a polynomial of degree $k-1$. Further we have $\int_{\mathbb{R}^2} {}_c\Phi_{R,k_0^1,k_0^2} = 1 \neq 0$ (see remark 5.7), thus we can apply lemma 5.2 to get $p \equiv 0$. Consequently $\lambda_{l_1,l_2} = 0$ for all $|(l_1,l_2)|_1 < k$ and hence linear independence of the columns and thus full rank of the matrix M_c follows. □

Note that, since the original scaling function ϕ has compact support (cf. remark 5.5), indeed the support of $({}_c\Phi_{R,\mathbf{k}})$ gets smaller as R decreases, thus the assumption of proposition 5.12 is reasonable.

Remark 5.8. *In practice, the support restriction on ${}_c\Phi_{R,\mathbf{k}}$ corresponds to the lowest resolution image having more horizontal and vertical pixels than the filter length of the low-pass filter plus the order of the $\mathrm{TGV}_\alpha^{\mathrm{k}}$ functional minus one, e.g. having more that $(9+1)\times(9+1) = 10\times 10$ pixels in the case of CDF 9/7 filters and TGV_α^2. Compare also to proposition 5.13 later on.*

Remark 5.9. *Note that, as already discussed in remark 4.1, the boundedness assumptions on the suitable subclass of intervals ${}_cJ_{R,k_0^1+i,k_0^2+j}$ is as general as possible, if we want to bound the kernel of the $\mathrm{TGV}_\alpha^{\mathrm{k}}$ functional.*

With that, (A) and (EX_k) are assured for the infinite dimensional model corresponding to JPEG 2000 decompression and the results of section 4, in particular the optimality condition, apply.

5.3.2 Discrete framework

In this section we deal with the discrete implementation of the decompression process for JPEG 2000 images. For the sake of simplicity, we consider only quadratic images defined on the finite dimensional space $U = \mathbb{R}^{N\times N\times 3}$ equipped with the norm $\|\cdot\|_U$ as in (55). Further we define $V = U^2$ and $W = U^3$ with norms $\|\cdot\|_V$ and $\|\cdot\|_W$ as in (56).

Given the scaling function ϕ corresponding to the CDF 9/7 or LeGall 5/3 wavelets, and possibly shifting the resolution levels, we assume any discrete image $u \in U$ is described by the coefficients

$$(u^c, {}_c\Phi_{0,\mathbf{k}})_{0\leq \mathbf{k}<N}, \quad c \in \{1,2,3\},$$

where, for $\mathbf{k} = (k_1,k_2) \in \mathbb{N}^2$, $0 \leq \mathbf{k} < N$ can be understood componentwise

Now as described at the beginning of this section, given any JPEG 2000 compressed image $u \in U$, we can obtain a resolution level $R \in \mathbb{N}$ and, for each color component $c \in \{1,2,3\}$, data intervals

$$({}_cJ^0_{R,\mathbf{l}})_{0\leq \mathbf{l}<L(R)}, ({}_cJ^h_{j,\mathbf{l}})_{\substack{0<j\leq R\\ 0\leq \mathbf{l}<\mathbf{K}^h(j)}}, ({}_cJ^v_{j,\mathbf{l}})_{\substack{0<j\leq R\\ 0\leq \mathbf{l}<\mathbf{K}^v(j)}}, ({}_cJ^d_{j,\mathbf{l}})_{\substack{0<j\leq R\\ 0\leq \mathbf{l}<\mathbf{K}^d(j)}}, \tag{131}$$

where $\mathbf{l} = (l_1,l_2) \in \mathbb{N}^2$, $L(0) = N$ and

$$\begin{aligned}
L(j) &= \lceil L(j-1)/2 \rceil,\\
\mathbf{K}^h(j) &= L(j) \times (L(j-1) - L(j)),\\
\mathbf{K}^v(j) &= (L(j-1) - L(j)) \times L(j),\\
\mathbf{K}^d(j) &= (L(j-1) - L(j)) \times (L(j-1) - L(j)),
\end{aligned}$$

for $1 \leq j \leq R$. The restriction $0 \leq \mathbf{l} < \mathbf{K}$ is again meant component wise.

The ${}_cJ^0_{R,\mathbf{l}}$ provide low resolution- and the ${}_cJ^\lambda_{j,\mathbf{l}}$, $\lambda \in \{h,v,d\}$, detail information about the image to decompress.

Note that we consider only the case where the whole image is processed as one tile. For multiple tiles, data fidelity for each tile can be obtained independently as for separate images, while the TGV functional will be evaluated globally over all tiles. We further assume the same resolution level R for each color component. A generalization of our model to a color dependent resolution level $R(c)$, as it is possible within the JPEG 2000 standard, is straightforward but will be omitted for the sake of simplicity.

We can now define, for the resolution level $R \in \mathbb{N}$, the component wise wavelet transform operator of order R, $W = (W^1,W^2,W^3) : U \to U$, which can be evaluated by repeatedly filtering each component with the finite filters $(h_n)_n, (g_n)_n$ and allows to obtain the inner products

$$(u^c, {}_c\Phi_{R,\mathbf{k}})_{0\leq \mathbf{k}<L(R)}, (u^c, {}_c\Psi^h_{j,k})_{\substack{0<j\leq R\\ 0\leq \mathbf{k}<\mathbf{K}^h(j)}}, \tag{132}$$

$$(u^c, {}_c\Psi^v_{j,k})_{\substack{0<j\leq R\\ 0\leq \mathbf{l}<\mathbf{K}^v(j)}}, (u^c, {}_c\Psi^d_{j,k})_{\substack{0<j\leq R\\ 0\leq \mathbf{k}<\mathbf{K}^d(j)}}, \tag{133}$$

$c \in \{1,2,3\}$, from $(u^c, {}_c\Phi_{0,k})_{0\leq \mathbf{k}<N}$, $u \in U$. The filters $(h_n)_n$ correspond to the scaling function for a CDF 9/7 or LeGall 5/3 wavelet basis, and can be obtained as described in subsection 5.3.1. The filters $(g_n)_n$ are defined by

$$g_n = (-1)^n h_{1-n}$$

and correspond to the mother wavelet of the CDF 9/7 or LeGall 5/3 wavelet basis (cf. equations (93),(94)).

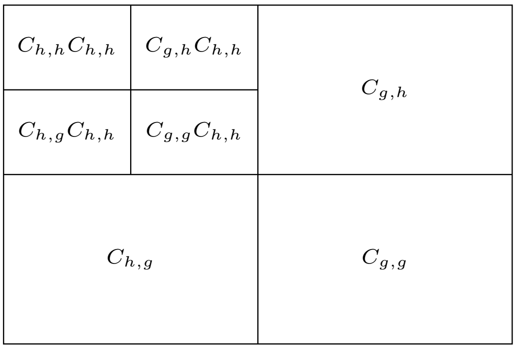


Figure 15: Visualization of the discrete wavelet operator W^c in the case $R = 2$

Note that, for simplicity, we again omit the (finite) filter length of $(h_n)_n$, $(g_n)_n$, do not write W resolution depended and also assume the same filters for each component. However, a generalization to different filters for each component is easily possible. Each operator W^c can then be described as

$$(W^c u)_{i,j} = \begin{cases} ((C_{h,h})^R u)_{i,j} & \text{if } (i,j) \in I_R \\ (C_{g,h}(C_{h,h})^{s-1}u)_{i,j-L(s+1)} & \text{if } (i,j) \in I_s^h \\ (C_{h,g}(C_{h,h})^{s-1}u)_{i-L(s+1),j} & \text{if } (i,j) \in I_s^v \\ (C_{g,g}(C_{h,h})^{s-1}u)_{i-L(s+1),j-L(s+1)} & \text{if } (i,j) \in I_s^d \end{cases} \quad 0 < s \leq R \tag{134}$$

where

$$\begin{aligned} I_s &= \big[0, L(s)\big) \times \big[0, L(s)\big), \\ I_s^h &= \big[0, L(s)\big) \times \big[L(s), L(s-1)\big), \\ I_s^v &= \big[L(s), L(s-1)\big) \times \big[0, L(s)\big), \\ I_s^d &= \big[L(s), L(s-1)\big) \times \big[L(s), L(s-1)\big), \end{aligned} \tag{135}$$

for $0 < s \leq R$. See also Figure 15 for a visualization in the case $R = 2$. The operators C_{h^1,h^1} correspond to filtering followed by downsampling, using h^1 and h^2 as filters in the horizontal and vertical direction, respectively, with the image symmetrically extended. For an input signal $w = (w_{i,j})_{\substack{0 \leq i < N_1 \\ 0 \leq j < N_2}}$ they can be given as

$$C_{h^1,h^2} w = C^H_{h^1} C^V_{h^2} w, \tag{136}$$

with

$$(C^H_{h^1} w)_{i,j} = \sum_m h^1_m w_{i,2j+m+1-\beta}, \quad 0 \leq i < N_1, 0 \leq j < \lfloor N_2/2 + \beta \rfloor, \tag{137}$$

$$(C^V_{h^2} w)_{i,j} = \sum_m h^2_m w_{2i+m+1-\beta,j}, \quad 0 \leq i < \lfloor N_1/2 + \beta \rfloor, 0 \leq j < N_2 \tag{138}$$

where $\beta = 1$ in the case $h_i = h$ and $\beta = 0$ if $h_i = g$ and, as already mentioned, the signal w is extended symmetrically, i.e. $w_{i,-j} = w_{i,j}$, $w_{i,N_2-1+j} = w_{i,N_2-1-j}$ for $j \in \mathbb{N}$ and similar for the vertical direction.

Using the dual filters $(\tilde{h}_n)_n$ and $(\tilde{g}_n)_n$, which correspond to the scaling function and mother wavelet, respectively, of the dual wavelet basis as described in subsection 5.3.1, the inverse wavelet transform operator $W^{-1} =$

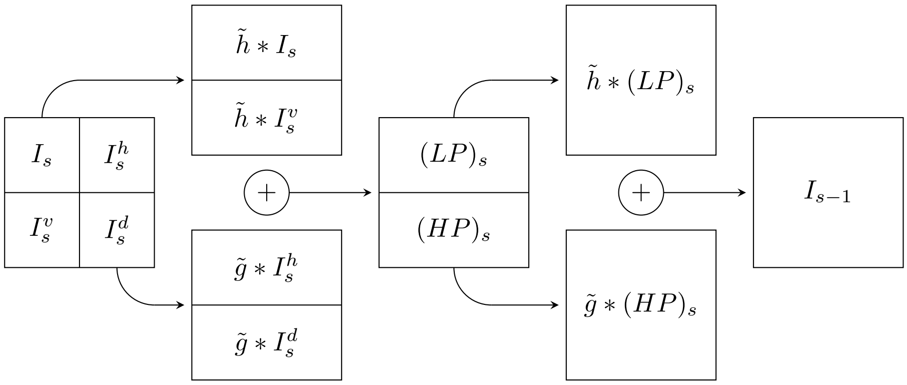


Figure 16: Visualization of the inverse discrete wavelet operator $(W^c)^{-1}$ acting on I_{s-1}

$((W^1)^{-1}, (W^2)^{-1}, (W^3)^{-1})$ can be defined component wise by

$$(W^c)^{-1}u = \left(\prod_{s=R}^{1} V_s H_s \right) u. \tag{139}$$

for a color component $u \in \mathbb{R}^{N \times N}$. The operators H_s are defined as

$$(H_s u)_{i,j} = \begin{cases} \sum_l \tilde{h}_{2l-j}(u^\beta|_{I_s \cup I_s^v})_{i,l} + \tilde{g}_{2l+1-j}(u|^\beta_{I_s^h \cup I_s^d})_{i,l} & (i,j) \in I_{s-1}, \\ u_{i,j} & (i,j) \in I_0 \setminus I_{s-1}, \end{cases} \tag{140}$$

where the restrictions $u|_I^\beta$, for $\beta \in \{0,1\}$, means to take only elements of w with indices in I, renumber them starting by zero and extend the signal symmetrically as $w_{-j} = w_{j-\beta}$, $w_{M-1+j} = w_{M-1-j+\beta}$, for $(w_i)_{0 \leq i < M}$ the restricted signal.

The operators V_s are then defined as

$$(V_s u)_{i,j} = \begin{cases} \sum_l \tilde{h}_{2l-i}(u|^\beta_{I_s \cup I_s^h})_{l,j} + \tilde{g}_{2l+1-i}(u|^\beta_{I_s^v \cup I_s^d})_{l,j} & (i,j) \in I_{s-1}, \\ u_{i,j} & (i,j) \in I_0 \setminus I_{s-1}. \end{cases} \tag{141}$$

Suitably summing up all data intervals as defined in (131) to

$$(J^c_{i,j})_{\substack{0 \leq i,j < N, \\ c \in \{1,2,3\}}}$$

the discrete data set for the JPEG 2000 decompression process can then be described as

$$U_D = \{u \in U \,|\, (W^c u^c)_{i,j} \in J^c_{i,j}, c \in \{1,2,3\}, 0 \leq i,j < N\}. \tag{142}$$

As in Subsection 5.2.2 we consider a discrete version of the TGV functional of order two defined on U as in equation (57). With that, the discrete minimization problem for artifact free JPEG 2000 decompression can be written as

$$\min_{u \in U} \mathrm{TGV}^2_\alpha(u) + \mathcal{I}_{U_D}(u). \tag{143}$$

Remember that, due to the JPEG 2000 compression procedure as described at the beginning of subsection 5.3, we cannot assume each of the intervals $J^c_{i,j}$, resulting from the data intervals as in (131), to be bounded. However, as already explained in the paragraph before proposition 5.12, it is reasonable and matches with our experiments, to assume each low-resolution interval ${}_cJ^0_{R,1}$ from (131) to be bounded, which will, as for the continuous setting, be sufficient to obtain existence of a solution.

But before, let us discuss how to solve the minimization problem for artifact free JPEG 2000 decompression numerically: In order to obtain a solution to problem (143) it would be convenient to again use the primal dual algorithm as presented in [26]. However, using exactly the same setting as for JPEG decompression, just by replacing the data set U_D, would amount to calculate a projection onto U_D in each iteration. Since our transformation operator is now no longer orthogonal in general, this would require a minimization problem itself, resulting in unacceptable performance decrease. Our solution to this is to introduce an additional variable for the algorithm as follows:

Denoting $D = \{w \in U | w^c_{i,j} \in J^c_{i,j}\}$, the convex indicator function for the data set U_D can be written as $\mathcal{I}_{U_D}(u) = \mathcal{I}_D(Ww)$. Now defining $\mathcal{N} \subset \{1,2,3\} \times \mathbb{N} \times \mathbb{N}$ to be the set of indices (c,i,j) such that the intervals $J^c_{i,j}$ are bounded, we denote by $l = (l^c_{i,j})_{(c,i,j)\in\mathcal{N}}$ and $o = (o^c_{i,j})_{(c,i,j)\in\mathcal{N}}$ the vectors of lower and upper bounds such that $J^c_{i,j} = [l^c_{i,j}, o^c_{i,j}]$ for $(c,i,j) \in \mathcal{N}$. Further splitting each variable $w \in U$ into $w = (w_1, w_2)$, where $w_1 = ((w_1)^c_{i,j})_{(c,i,j)\in\mathcal{N}}$ and $w_2 = ((w_1)^c_{i,j})_{(c,i,j)\notin\mathcal{N}}$, an easy calculation shows that the dual of $\mathcal{I}_D$ can be written as

$$\mathcal{I}^*_D(w) = \sup_{v\in D}(v,w) = (\frac{l+o}{2}, w_1) + \|\frac{o-l}{2} \cdot w_1\|_{\ell^1} + \mathcal{I}_{\{0\}}(w_2), \tag{144}$$

where the dot product $\cdot$ means a component wise product and the indicator function $I_{\{0\}}(w_2)$ is zero only if all components of w_2 are zero and infinity else. With that, the function $\mathcal{I}_{U_D}$ can be written as

$$\begin{aligned}\mathcal{I}_{U_D}(u) = & \qquad \mathcal{I}_D(Wu) = \sup_{w\in U}(Wu,w) - \mathcal{I}^*_D(w) \\ = & \sup_{w=(w_1,w_2)} (Wu,w) - (\frac{l+o}{2}, w_1) - \|\frac{o-l}{2} \cdot w_1\|_{\ell^1} - \mathcal{I}_{\{0\}}(w_2) \\ := & \qquad \sup_{w=(w_1,w_2)} (Wu,w) - G^*_1(w_1) - G^*_2(w_2).\end{aligned} \tag{145}$$

Realizing the discrete version of the TGV functional again by

$$x \mapsto F(Kx)$$

with

$$F : Y \to \mathbb{R} \text{ and } K : X \to Y$$

as in (68) and (67), respectively, and $X := U \times V$, $Y := V \times W$, equipped with the standard Euclidean norm on its product components, the minimization problem

(143) can formally be reformulated as

$$\begin{aligned} \min_{\substack{x\in X\\ x=(u,v)}} F(Kx)+\mathcal{I}_{U_D}(u) &\Leftrightarrow \min_{\substack{x\in X\\ x=(u,v)}} \max_{y\in Y}(Kx,y)-F^*(y)+\mathcal{I}_{U_D}(u)\\ \Leftrightarrow \min_{\substack{x\in X\\ x=(u,v)}} \max_{y\in Y}(Kx,y)-F^*(y)&+\max_{w\in U}(Wu,w)-G_1^*(w_1)-G_2^*(w_2)\\ \Leftrightarrow \min_{\substack{x\in X\\ x=(u,v)}} \max_{\substack{y\in Y\\ w\in U}}((Kx,Wu),(y,w))&-F^*(y)-G_1^*(w_1)-G_2^*(w_2)\\ &\Leftrightarrow \min_{x\in X}\max_{z\in Z}(\mathbf{K}x,z)-\mathbf{F}^*(z), \end{aligned} \tag{146}$$

where $Z := Y \times U$, again equipped with the standard Euclidean norm on its product components, $\mathbf{F} : Z \to \mathbb{R}$ and $\mathbf{K} : X \to Z$ are defined by

$$\mathbf{F}(z) = \mathbf{F}(x,w) = F(x) + \mathcal{I}_D(w)$$

and

$$\mathbf{K} = \begin{bmatrix} \nabla & -I \\ 0 & \mathcal{E} \\ W & 0 \end{bmatrix}$$

with $I : V \to V$ again the identity. Note that the definition of $\mathbf{F}$ implies that

$$\mathbf{F}^*(z) = \mathbf{F}^*((y,w_1,w_2)) = F^*(y) + G_1^*(w_1) + G_2^*(w_2)$$

as expected. Let us show below that (146) is a saddle point problem equivalent to (143) and study the discrete problems in detail.

Discrete existence and optimality

First note that the original problem (143) possesses a solution and is equivalent to minimizing both u and v simultaneously:

Proposition 5.13. *Given any discrete lowpass wavelet filter $(h_n)_{n\in\mathbb{Z}}$ with odd length $2L-1$, $L \in \mathbb{N}$, and $\sum_l h_l = 1$, assume that the lowest resolution of the JPEG 2000 compressed image is at least described by $2L \times 2L$ coefficients and that ${}_cJ^0_{R,L+i,L+j}$ is bounded for $0 \le i + j < 2$, $i, j \in \mathbb{N}$. Then, there exists a solution to (143) and $\hat{u}$ is an optimal solution to (143) if and only if there exists $\hat{v}$ such that*

$$(\hat{u},\hat{v}) = \operatorname*{arg\,min}_{\substack{x\in X\\ x=(u,v)}} \mathbf{F}(\mathbf{K}x). \tag{147}$$

Proof. Clearly, the functional $\mathrm{TGV}^2_\alpha + \mathcal{I}_{U_D}$ is proper. Let $(u_n)_{n\in\mathbb{N}}$ be a minimizing sequence for (143). Denote $\tilde{u}_n = u_n - P_1(u_n)$, where P_1 is the projection onto the functions u such that $\mathcal{E}(\nabla u) = 0$. We have already shown in the proof of proposition 5.2 that there exists a constant $C > 0$ such that

$$\|\tilde{u}_n\|_1 \le C\,\mathrm{TGV}^2_\alpha(\tilde{u}_n) \quad \text{for all } n \in \mathbb{N},$$

thus $\|\tilde{u}_n\|_1$ is bounded.

It remains to bound $\|P_1(u_n)\|_1$. For that, remember that $P_1(u_n)$ is a discrete component wise polynomial of degree less or equal to 1, i.e., $(P_1(u_n))^c_{i,j} = {}_c\lambda^n_0 + {}_c\lambda^n_1 i + {}_c\lambda^n_2 j$ for $c \in \{1,2,3\}$. Since subsampling followed by linear filtering with $(h_n)_{n\in\mathbb{Z}}$ does not destroy the linear structure of $P_1(u_n)^c$, we have, for $c \in \{1,2,3\}$,

$$(W^c P_1(u_n)^c)_{L+i,L+j} = {}_c\tilde{\lambda}_0^n + {}_c\tilde{\lambda}_1^n(L+i) + {}_c\tilde{\lambda}_2^n(L+j).$$

Note that this is only true since, due to the choice of $L+i, L+j$, with $0 \leq i+j < 2$, and the assumption that the lowest resolution level is described by at least $2L \times 2L$ coefficients, for the transform W^c at $(L+i, L+j)$ no boundary extension has to be considered, implying indeed the linear structure.

An explicit calculation shows that, after filtering a linear signal, say $F = (F_i)_i$, $F_i = f_0 + f_1 i$, with a filter $(\tilde{h}_n)_n$ the coefficients of the resulting linear signal, denoted by $\tilde{F} = (\tilde{F}_i)_i$, $\tilde{F}_i = \tilde{f}_0 + \tilde{f}_1 i$, are such that

$$\tilde{f}_0 = f_0(\sum_i \tilde{h}_i) + f_1(\sum_i \tilde{h}_i i), \quad \tilde{f}_1 = -f_1(\sum_i \tilde{h}_i).$$

Thus, if

$$\sum_i \tilde{h}_i \neq 0,$$

which is by remark 5.7 satisfied for the JPEG 2000 filters, boundedness of f_0, f_1 follows from boundedness of $\tilde{f}_0, \tilde{f}_1$. By this argument, since in our case indeed $\sum_n h_n = 1$, it suffices to show boundedness of ${}_c\tilde{\lambda}_0^n$, ${}_c\tilde{\lambda}_1^n$, ${}_c\tilde{\lambda}_2^n$ to conclude boundedness ${}_c\lambda_0^n$, ${}_c\lambda_1^n$, ${}_c\lambda_2^n$.

Using boundedness of ${}_cJ^0_{R,L,L}, {}_cJ^0_{R,L+1,L}, {}_cJ^0_{R,L,L+1}$ together with boundedness of $W^c((u - P_2(u_n))^c)$ we get that all rows of

$$\begin{gathered} {}_c\tilde{\lambda}_0^n + {}_c\tilde{\lambda}_1^n(L) + {}_c\tilde{\lambda}_2^n(L) \\ {}_c\tilde{\lambda}_0^n + {}_c\tilde{\lambda}_1^n(L+1) + {}_c\tilde{\lambda}_2^n(L) \\ {}_c\tilde{\lambda}_0^n + {}_c\tilde{\lambda}_1^n(L) + {}_c\tilde{\lambda}_2^n(L+1), \end{gathered}$$

and with that, all linear combinations of these rows are bounded. Thus, solving this system, yields ${}_c\tilde{\lambda}_0^n$, ${}_c\tilde{\lambda}_1^n$, ${}_c\tilde{\lambda}_2^n$, with that ${}_c\lambda_0^n$, ${}_c\lambda_1^n$, ${}_c\lambda_2^n$ and finally $P_1(u_n)$ to be bounded.

Existence of a subsequence of $(u_n)_{n\in\mathbb{N}}$ converging to an element $u \in U_D$ together with continuity of TGV^2_α finally implies existence of a solution to (143). The claimed equivalence of (143) to (147) is again immediate. □

Remark 5.10. *Note that the argumentation presented in the proof of proposition 5.13 above essentially carries out the arguments of the existence proof for the continuous setting (cf. proposition 5.12) in a discrete setting.*

The next result aims to summarize the primal problem (147), a dual and a saddle point problem and to argue existence as well as equivalence for all of them.

Proposition 5.14. *With the assumptions of proposition 5.13, there exists a solution to the primal problem* (147)*, its dual, given by*

$$\max_{z\in Z} -\mathcal{I}_{\{0\}}(-\mathbf{K}^* z) - \mathbf{F}^*(z), \tag{148}$$

and the saddle point problem

$$\min_{x\in X} \max_{z\in Z} (\mathbf{K}x, z) - \mathbf{F}^*(z). \tag{149}$$

Further, $\hat{x}, \hat{z}$ are solutions to the primal problem (147) *and the dual problem* (148)*, respectively, if and only if* $(\hat{x}, \hat{z})$ *solves the saddle point problem* (149).

Proof. Given that existence of a solution to the primal problem is already ensured, following [37, Proposition III.3.1] it suffices to show existence of a solution to the dual problem and equality of the primal and dual problem at optimal points, to conclude all claims. Using [4, Corollary 2.3] this follows provided that

$$\bigcup_{\lambda \geq 0} \lambda[\mathrm{dom}(\mathbf{F}) - \mathbf{K}(\mathrm{dom}(\mathbf{0}))] = Z,$$

where $\mathbf{0} : X \to 0$ is the zero mapping. But this is satisfied since, given any $w_0 \in U_D$ fixed, any (p, q, w) can be written as

$$\begin{pmatrix} p \\ q \\ w \end{pmatrix} = \begin{pmatrix} p + \nabla W^{-1}(w_0 - w) \\ q \\ w_0 \end{pmatrix} - \mathbf{K} \begin{pmatrix} W^{-1}(w_0 - w) \\ 0 \end{pmatrix} \in \mathrm{dom}(\mathbf{F}) - \mathbf{K}(\mathrm{dom}(\mathbf{0})).$$

□

Remark 5.11. *By standard results (i.e. [37, Theorem III.4.1]), existence of a point $x_0 \in X$ such that $\mathbf{F}(\mathbf{K}x_0)$ is finite and $\mathbf{F} \circ \mathbf{K}$ is continuous at x_0, which is satisfied in the application to JPEG 2000 decompression, would also imply the above result. However, using the more general result from [4] allows to apply the same theory also for the case that $\overset{\circ}{U_D} = \emptyset$, which is needed for the application to wavelet based zooming later on.*

Knowing about existence of a solution, we are now able to formulate the optimality conditions for the discrete saddle point problem.

Proposition 5.15. *Let the assumptions of proposition 5.13 be satisfied. Then, there exists a solution to* (149) *and* $(\hat{x}, \hat{z}) = (\hat{u}, \hat{v}, \hat{p}, \hat{q}, \hat{w})$ *solves* (149) *if and only if*

- $\begin{pmatrix} -\operatorname{div} \hat{p} + W^*\hat{w} \\ -\hat{p} - \operatorname{div} \hat{q} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$
- $\hat{p}_{i,j} = \alpha_1 \frac{(\nabla \hat{u} - \hat{v})_{i,j}}{|(\nabla \hat{u} - \hat{v})_{i,j}|_V}$ *if* $|(\nabla \hat{u} - \hat{v})_{i,j}|_V \neq 0$ *and* $|\hat{p}_{i,j}|_V \leq \alpha_0$ *else.*
- $\hat{q}_{i,j} = \alpha_0 \frac{(\mathcal{E}\hat{v})_{i,j}}{|(\mathcal{E}\hat{v})_{i,j}|_W}$ *if* $|(\mathcal{E}\hat{v})_{i,j}|_W \neq 0$ *and* $|\hat{q}_{i,j}|_W \leq \alpha_1$ *else.*
- $\hat{u} \in U_D$ *and*

$$\begin{cases} \hat{w}^c_{i,j} \in \mathbb{R} & \textit{if } (W^c\hat{u}^c)_{i,j} = l^c_{i,j} = o^c_{i,j} \\ \hat{w}^c_{i,j} \geq 0 & \textit{if } (W^c\hat{u}^c)_{i,j} = o^c_{i,j} \neq l^c_{i,j} \\ \hat{w}^c_{i,j} \leq 0 & \textit{if } (W^c\hat{u}^c)_{i,j} = l^c_{i,j} \neq o^c_{i,j} \\ \hat{w}^c_{i,j} = 0 & \textit{if } (W^c\hat{u}^c)_{i,j} \in \overset{\circ}{[l^c_{i,j}, o^c_{i,j}]} \end{cases}$$

Proof. The proof is rather short, using that the objective functional is convex and concave in the primal and dual direction, respectively, and thus $(\hat{x}, \hat{z})$ is optimal if and only

$$0 \in \partial_x \left((\hat{x}, \mathbf{K}^*\hat{z})_Z - \mathbf{F}^*(\hat{z}) \right)$$

Algorithm 3 Abstract primal dual algorithm for JPEG 2000 decompression

- Initialization: Choose $\tau, \sigma > 0$ such that $\|\mathbf{K}\|^2\tau\sigma < 1$, $(x^0, z^0) \in X \times Z$ and set $\overline{x}^0 = x^0$
- Iterations ($n \geq 0$): Update $x^n, z^n, \overline{x}^n$ as follows:

$$\begin{cases} z^{n+1} = & (I + \sigma\,\partial\,\mathbf{F}^*)^{-1}(z^n + \sigma\mathbf{K}\overline{x}^n) \\ x^{n+1} = & x^n - \tau\mathbf{K}^* z^{n+1} \\ \overline{x}^{n+1} = & 2x^{n+1} - x^n \end{cases}$$

and

$$0 \in \partial_z\left((-\mathbf{K}\hat{x}, \hat{z})_Z + \mathbf{F}^*(\hat{z})\right).$$

Using a continuity argument to assure additivity of the subdifferential as in proposition 5.4, zero being in the subdifferential with respect to the primal variable is equivalent to the first assertion of the optimality condition. Also by additivity of ∂, $\mathbf{K}\hat{x} \in \partial\,\mathbf{F}^*(\hat{z})$ is equivalent to zero being in the subdifferential with respect to the dual direction. A component wise evaluation of this further yields equivalence to the second two assertions as well as $W\hat{u} \in \partial\mathcal{I}_D^*(\hat{w})$. But this is equivalent to $\hat{w} \in \partial\mathcal{I}_D(W\hat{u})$ which, again by component wise evaluation, is equivalent to the last assertion. □

5.3.3 Practical implementation

Now to solve (149) we can formulate the abstract primal dual algorithm from [26] in algorithm 3, for which global convergence can be assured.

In order to get a practical implementation it is left to estimate $\|\mathbf{K}\|$ and to get an explicit form of W^* and $(I + \sigma\,\partial\,\mathbf{F}^*)^{-1}$.

In order to estimate $\|\mathbf{K}\|$ first note that (see also remark 5.3), for $(u, v) \in X$,

$$\|\mathbf{K}(u,v)\|^2 \leq (8 + \|W\|^2)\|u\|_U^2 + 2\sqrt{8}\|u\|_U\|v\|_V + 9\|v\|_V^2, \tag{150}$$

so this reduces to estimate $\|W\|$. This will be done in the following lemma.

Lemma 5.3. *With $W : U \to U$, defined component wise as in* (134)*, depending on a resolution level $R \in \mathbb{N}$ and filters $h = (h_n)_n$, $g = (g_n)_n$, we have*

$$\|W\| \leq 2\|g\|_1^2(2\|h\|_1^2)^{R-1} = \begin{cases} 8(\frac{9}{2})^{R-1} & \textit{for LeGall 5/3 filters,} \\ 13.4708(3.8107)^{R-1} & \textit{for CDF 9/7 filters.} \end{cases} \tag{151}$$

Proof. First note that for an estimate on $\|W^c\|$, $c \in \{1, 2, 3\}$, it suffices to estimate $\|W\|$. Due to the special form of W^c as in (134), we can further reduce this to estimating $\|C_{h^1,h^2}\|$ for two filters h^1, h^2, which again reduces to an estimate on $\|C_{h^1}^H\|$ and $\|C_{h^2}^V\|$.

For that, note that both operants $C_{h^1}^H$, $C_{h^2}^V$are just sub-sampling followed by a symmetrically extended convolution with a given filter $\tilde{h}$, where only the central points are considered. Since the norm of the subsampling process is one, we finally arrive at the task of estimating the norm of this symmetrically extended convolution to obtain an estimate on $\|W\|$.

Using Youngs-inequality for convolutions [8, Theorem 3.9.4], any convolution $\mathrm{conv}_{\tilde{h}}$ of a discrete function w defined on all of $\mathbb{R}$ with a filter $\tilde{h}$ can be estimated by

$$\|\operatorname{conv}_{\tilde{h}}(w)\|_2 \leq \|\tilde{h}\|_1 \|w\|_2,$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ are discrete ℓ^1 and ℓ^2 norms, respectively.

Now given any discrete signal of finite support, $(w_i)_{0\leq i<N}$, the center points of its symmetrically extended convolution with a filter $\tilde{h} = (\tilde{h}_i)_{-L\leq i\leq L}$, $2L+1 < N$, can be written as

$$(C_{\tilde{h}}w)_i = \sum_{j=-L}^{L} \tilde{h}_j (E^s w)_{i+j} \quad 0 \leq i < N,$$

where $E^s w$ is a symmetric extension of the signal $(w_i)_{0\leq i<N}$ to $((E^s w)_i)_{-L\leq i<N+L}$. Denoting by $E^0(E^s w)$ a zero extension of $E^s w$ to $((E^0 E^s w)_i)_{i\in\mathbb{Z}}$, we can estimate

$$\begin{aligned} \|C_{\tilde{h}}w\|_2 \leq & \quad \|\operatorname{conv}_{\tilde{h}}(E^0 E^s w)\|_2 \leq \|\tilde{h}\|_1 \|E^s w)\|_2 \\ = & \quad \|\tilde{h}\|_1 \sqrt{\|w\|_2^2 + \|(w_i)_{i=1}^{L}\|_2^2 + \|(w_i)_{i=N-L-1}^{N-2}\|^2} \\ \leq & \quad \sqrt{2}\|\tilde{h}\|_1 \|w\|_2. \end{aligned} \tag{152}$$

For two filters h^1, h^2, we have the estimate

$$\|C_{h^1,h^2}\| \leq 2\|h^1\|_1 \|h^2\|_1.$$

In the case of LeGall 5/3 filters, we have $\|h\|_1 = \frac{3}{2}$ and $\|g\|_1 = 2 > \|h\|_1$ and in the case of CDF 9/7 filters we have $\|h\|_1 \approx 1.3803$ and $\|g\|_1 \approx 2.5953 > \|h\|_1$. So in both cases $\|g\|_1 > \|h\|_1$. With that, we get

$$\|W^c u\|_2 \leq \max_{\substack{\tilde{h}^1,\tilde{h}^2\in\{h,g\} \\ 0\leq s<R}} \|C_{\tilde{h}^1,\tilde{h}^2} C^s_{h,h}\| \|u\|_2 \leq \|C_{g,g}\| \|C_{h,h}\|^{R-1} \|u\|_2$$

and consequently the estimate as claimed follows.

If we use the resulting estimate on $\|W\|$ now in equation (150), and solve $ab = \sqrt{8}, a^2 = b^2 + 1 - \|W\|^2$, similarly as in remark 5.3, an estimate for (150) can be given as

$$\|\mathbf{K}\|^2 \leq 8 + \|W\|^2 + \left(\frac{1 - \|W\|^2}{2} + \sqrt{\frac{(1 - \|W\|^2)^2}{4} + 8} \right)^2 \tag{153}$$

which, for example if $R = 5$, amounts $\|K\| \leq 2839.97$ in the case of CDF 9/7 filters and $\|K\| \leq 3280.5$ in the case of LeGall 5/3. Choosing equal stepsizes in the primal and dual direction, this results in $\sigma = \tau \approx 0.000352$ for example for CDF 9/7 filters, which is quite small compared for example to the JPEG case where $\sigma = \tau \approx 0.2965$ and thus yields slow convergence of the algorithm.

Obviously, the estimation of $\|C_{\tilde{h}}\|$ in equation (152) was not optimal. However, as the choice $w = (1, 1)$, $\tilde{h} = (1, 1, 1)$ shows, we cannot expect to get any better than

$$\|C_{\tilde{h}}w\|_2 \leq \|\tilde{h}\|_1 \|w\|_2,$$

which still yields $\|W^c\| = 88.7$ and, consequently, the small stepsizes

$$\sigma = \tau = 0.0113.$$

Also the estimate in equation (151) can be shown to be tight by choosing $R = 1$ and a function u that is in the kernel of both C_h^H and C_h^V. Thus there is not much hope to improve the theoretical bound on $\|W\|$ and with that, justify a larger stepsize.

We propose two possibilities to resolve this issue, one of which is a balancing technique we explain next:

Remember that the derivation of the JPEG 2000 algorithm was based on writing $\mathcal{I}_{U_D}$ as

$$\mathcal{I}_{U_D}(u) = \mathcal{I}_D(Wu) = \sup_{w\in U}(Wu, w) - \mathcal{I}_D^*(w),$$

where $\mathcal{I}_D^*$ is known explicitly. Now with $\Lambda : U \to U$, $(\Lambda u)_{i,j}^c := \lambda_{i,j}^c u_{i,j}^c$, we can define $\tilde{W} = \Lambda \circ W$ and write

$$\mathcal{I}_{U_D}(u) = \sup_{w\in U}(\tilde{W}u, w) - \mathcal{I}_{\tilde{D}}^*(w)$$

with $\tilde{D} = \{u \in X | u_{i,j}^c \in \lambda_{i,j}^c J_{i,j}^c\}$. Note that $\mathcal{I}_{\tilde{D}}$ can still be given explicitly by

$$\mathcal{I}_{\tilde{D}}^*(w) = \sup_{v\in U}(v, w) = ((\Lambda|_{\mathcal{N}})(\frac{l+o}{2}), w_1) + \|(\Lambda|_{\mathcal{N}})(\frac{o-l}{2}) \cdot w_1\|_{\ell^1} + \mathcal{I}_{\{0\}}(w_2).$$

Remember that $\mathcal{N}$ is the index set such that $J_{i,j}^c$ is bounded for $(c, i, j) \in \mathcal{N}$. But now we can choose

$$\lambda_{i,j}^c = \gamma \cdot \begin{cases} (2^R\|h\|_1^{2R})^{-1} & \text{if } (i,j) \in I_R, \\ (2^s\|g\|_1\|h\|_1\|h\|_1^{2s-2})^{-1} & \text{if } (i,j) \in I_s^h, \\ (2^s\|g\|_1\|h\|_1\|h\|_1^{2s-2})^{-1} & \text{if } (i,j) \in I_s^v, \\ (2^s\|g\|_1^2\|h\|_1^{2s-2})^{-1} & \text{if } (i,j) \in I_s^d, \end{cases} \quad 0 < s \leq R,\ c \in \{1, 2, 3\}, \tag{154}$$

This ensures $\|\tilde{W}\| \leq \gamma$ for $\gamma > 0$ arbitrary, i.e. an arbitrary low bound on $\|W\|$, independent of the resolution level R. Choosing for example $\gamma = 1$, and thus $\|\tilde{W}\| \leq 1$, results in $\|K\|^2 \leq 11.8284$ and thus allows $\sigma = \tau = 0.2908$, i.e., a sufficiently large stepsize for which convergence can be assured. However, we will see in subsection 5.3.4, that choosing γ too small reduces data influence during the iterations and may also lead to slow convergence. Thus the choice of gamma needs to be balanced between stepsize and data influence.

Remark 5.12. *Note that this modification of W to $\tilde{W} = \Lambda \circ W$ only affects the evaluation of $\tilde{W}$, $\tilde{W}^*$, $\tilde{W}^{-1}$, the stepsize estimation and the definition of the data set $\tilde{D}$. Since all elements of Λ are nonequal to zero, its inverse as well as its adjoint are well defined as element-wise multiplication with $(\lambda_{i,j}^c)^{-1}$ and $(\lambda_{i,j}^c)$, respectively, and thus also the inverse and adjoint of $\tilde{W}$ can directly be obtained from W by $\tilde{W}^{-1} = W^{-1}\Lambda^{-1}$ and $\tilde{W}^* = W^*\Lambda^*$, respectively.*

As a result we will continue writing W and return to $\tilde{W}$, if needed, only for the numerical experiments in subsection 5.3.4.

Another possibility to improve convergence speed would be to choose the stepsize adaptively: As one can observe in the proof of the algorithm in [26, Theorem 1], it is possible to violate $\sigma\tau\|\mathbf{K}\|^2 < 1$ but still guarantee convergence, provided that

$$\|\mathbf{K}(x^n - x^{n-1})\|_Z < \frac{1}{\sqrt{\sigma\tau}}\|x^n - x^{n-1}\|_X \tag{155}$$

is satisfied for all $n \in \mathbb{N}$.

Now for the estimation of the theoretical bound on $\|W\|$, and with that on $\|\mathbf{K}\|$, one has to take into account all special cases, such as x being non-zero only close to the boundary or being in the kernel of horizontal- and vertical high pass filtering. Even if done in an optimal way, this may result in a worst-case estimation of $\|Wx\|$ that is irrelevant in practice.

Thus, by basically using only the subset of the iterates of the algorithm for the estimation of $\|W\|$, there is hope that, even for large stepsizes σ, τ, convergence can still be guaranteed. However, due to the nature of the proof of convergence in [26, Theorem 1], we cannot adapt the stepsize in each iteration. We thus formulate the following criterion for adaptive stepsize choice, that still allows to ensure convergence of the primal dual algorithm:

Given $\theta, \delta \in (0, 1)$, in each iteration of algorithm 3, set τ_{n+1}, σ_{n+1}, where

$$\begin{aligned} z^{n+1} &= (I + \sigma_{n+1}\,\partial\,\mathbf{F}^*)^{-1}(z^n + \sigma_{n+1}\mathbf{K}\overline{x}^n) \\ x^{n+1} &= x^n - \tau_{n+1}\mathbf{K}^* z^{n+1}, \end{aligned}$$

such that

$$\sigma_{n+1}\tau_{n+1} = \begin{cases} \frac{\delta\|\mathbf{K}(x^n - x^{n-1})\|_Z}{\|x^n - x^{n-1}\|_X} & \text{if } \theta\sigma_n\tau_n \geq \frac{\|\mathbf{K}(x^n - x^{n-1})\|_Z}{\|x^n - x^{n-1}\|_X} \\ \theta\sigma_n\tau_n & \text{if } \sigma_n\tau_n \geq \frac{\|\mathbf{K}(x^n - x^{n-1})\|_Z}{\|x^n - x^{n-1}\|_X} > \theta\sigma_n\tau_n \\ \sigma_n\tau_n & \text{if } \sigma_n\tau_n < \frac{\|\mathbf{K}(x^n - x^{n-1})\|_Z}{\|x^n - x^{n-1}\|_X}. \end{cases} \tag{156}$$

This choice guarantees equation (155) to hold, but also a sufficient decay of $(\sigma_n)_{n\in\mathbb{N}}, (\tau_n)_{n\in\mathbb{N}}$.

With that, there exists an $n_0 \in \mathbb{N}$ such that either condition (155) holds for all $n \geq n_0$, even though $\sigma_n\tau_n\|\mathbf{K}\|^2 \geq 1$, or $\sigma_n\tau_n\|\mathbf{K}\| < 1$ for all $n \geq n_0$. In both cases, this means $\sigma_n = \sigma_{n_0}$ and $\tau_n = \tau_{n_0}$ for all $n \geq n_0$ and with that, convergence of algorithm 3 can be assured.

Of course, the algorithm can be arbitrary slow, but, as we will see in subsection 5.3.4, in practice this method allows the same stepsize as for the JPEG algorithm and thus yields satisfactory convergence properties.

Having now a suitable choice for the stepsize of the primal dual algorithm, we draw our attention to getting an explicit form of W^*. For $u \in U$, the calculation of Wu amounts essentially filtering with the filters $(h_n)_n, (g_n)_n$, thus the adjoint of W can essentially be given by applying W^{-1} but with flipped versions of the filters $(h_n)_n, (g_n)_n$ instead of $(\tilde{h}_n)_n, (\tilde{g}_n)_n$. Due to symmetry, the flipping is obsolete. What is left is to correctly handle boundary indices. Taking this into account, W^* can be defined component wise as follows:

$$(W^c)^* u = \left(\prod_{s=R}^{1}(\tilde{V}_s + B_s^{\updownarrow})(\tilde{H}_s + B_s^{\leftrightarrow})\right)u, \tag{157}$$

where, similar as for $(W^c)^{-1}$, the operators $\tilde{H}_s$ are defined as

$$(\tilde{H}_s u)_{i,j} = \begin{cases} \sum_l h_{2l-j}(u|_{I_s \cup I_s^v})_{i,l} + g_{2l+1-j}(u|_{I_s^h \cup I_s^d})_{i,l} & (i,j) \in I_{s-1} \\ u_{i,j} & (i,j) \in I_0 \setminus I_{s-1}, \end{cases} \tag{158}$$

only that this time the restriction $w|_I$ is used, which means to take only elements of w with indices in I, renumber them starting by zero and use zero extension at the boundary. The boundary effects in the horizontal direction are compensated by $B_s^{\leftrightarrow}$, which is defined as

$$(B_s^{\leftrightarrow} u)_{i,j} = \begin{cases} b_h^{\leftrightarrow}(S_1(u|_{I_s \cup I_s^v})) + b_g^{\leftrightarrow}(S_0(u|_{I_s^h \cup I_s^d})) & \text{if } (i,j) \in I_{s-1}, \\ 0 & \text{if } (i,j) \in I_0 \setminus I_{s-1}, \end{cases} \tag{159}$$

where S_β, for $\beta \in \{0,1\}$, means to upsample a signal by inserting zero before each entry, starting with the β-entry and, for any given filter $\tilde{h}$ with odd length $2L+1 \in \mathbb{N}$ and any signal $w = (w_{i,j})_{\substack{0 \leq i < N_1 \\ 0 \leq j < N_2}}$, $2l+1 < N_2$, $b_{\tilde{h}}^{\leftrightarrow}(w)$ is defined as

$$(b_{\tilde{h}}^{\leftrightarrow} w)_{i,j} = \begin{cases} \sum_{l=L}^{j} h_l w_{i,l-j} & \text{if } 1 \leq j \leq L, \\ \sum_{l=-L}^{j-N_2+1} h_l w_{i,l+N_2} & \text{if } N_2 - L - 1 \leq j \leq N_2 - 2, \\ 0 & \text{else.} \end{cases} \tag{160}$$

Note that $b_{\tilde{h}}^{\leftrightarrow}(w)$ does not evaluate outside the boundaries of w. The operators $\tilde{V}_s$ are then defined as

$$(\tilde{V}_s u)_{i,j} = \begin{cases} \sum_l h_{2l-i}(u|_{I_s \cup I_s^h})_{l,j} + g_{2l+1-i}(u|_{I_s^v \cup I_s^d})_{l,j} & (i,j) \in I_{s-1} \\ u_{i,j} & (i,j) \in I_0 \setminus I_{s-1} \end{cases} \tag{161}$$

and the $B_s^{\updownarrow}$ are defined as

$$(B_s^{\updownarrow} u)_{i,j} = \begin{cases} b_h^{\updownarrow}(S_1(u|_{I_s \cup I_s^h})) + b_g^{\updownarrow}(S_0(u|_{I_s^v \cup I_s^d})) & \text{if } (i,j) \in I_{s-1}, \\ 0 & \text{if } (i,j) \in I_0 \setminus I_{s-1}, \end{cases} \tag{162}$$

with, again for any given filter $\tilde{h}$ with odd length $2L+1 \in \mathbb{N}$ and any signal $w = (w_{i,j})_{\substack{0 \leq i < N_1 \\ 0 \leq j < N_2}}$, $2l+1 < N_1$,

$$(b_{\tilde{h}}^{\updownarrow} w)_{i,j} = \begin{cases} \sum_{l=L}^{i} h_l w_{l-i,j} & \text{if } 1 \leq i \leq L, \\ \sum_{l=-L}^{i-N_1+1} h_l w_{l+N_2,j} & \text{if } N_1 - L - 1 \leq i \leq N_1 - 2, \\ 0 & \text{else.} \end{cases} \tag{163}$$

Having determined W^* it is only left to get an explicit form of $(I + \sigma\,\partial\,\mathbf{F}^*)^{-1}$ to be finally able to implement algorithm 3. Now since $\mathbf{F}^*(z)$ can be decomposed as $F^*(y) + G_1^*(w_1) + G_2^*(w_2)$, with y, w_1, w_2 independent variables, $(I + \sigma\,\partial\,\mathbf{F}^*)^{-1}(z)$ can be written as

$$(I + \sigma\,\partial\,\mathbf{F}^*)^{-1}(z) = \begin{pmatrix} (I + \sigma\,\partial\,F^*)^{-1}(y) \\ (I + \sigma\,\partial\,G_1^*)^{-1}(w_1) \\ (I + \sigma\,\partial\,G_2^*)^{-1}(w_2) \end{pmatrix}.$$

We have already seen in subsection 5.2.2 that $(I + \sigma\,\partial\,F^*)^{-1}(y)$ can be written as

$$(I + \sigma\,\partial\,F^*)^{-1}(y) = (I + \sigma\,\partial\,F^*)^{-1}(v, w) = \big(\mathrm{proj}_{\alpha_1}(v), \mathrm{proj}_{\alpha_0}(w)\big) \tag{164}$$

with $\mathrm{proj}_{\alpha_1}, \mathrm{proj}_{\alpha_0}$ as in equation (71). It is also easy to see that

$$(I + \sigma\,\partial\,G_2^*)^{-1}(w_2) = P_{\{\mathbf{0}\}}(w_2). \tag{165}$$

Thus it is left to get an explicit form of $(I + \sigma\,\partial\,G_1^*)^{-1}(w_1)$, which can be done as follows:

Defining the linear operator L as $Lw_1 = \frac{o-l}{2} \cdot w_1$, it follows that

$$\begin{aligned} r^* \in \partial\,G_1^*(r) \Leftrightarrow \qquad & r^* - \frac{l+o}{2} \in \partial(\|\cdot\|_{\ell^1} \circ L)(r) = L^*\,\partial\,\|Lr\|_{\ell^1} \\ \Leftrightarrow \quad & r^* - \frac{l+o}{2} = L^* w^* \text{ with } \|w^*\|_\infty \leq 1, \|Lr\|_{\ell^1} = (w^*, Lr). \end{aligned}$$

Thus, using that $L^* = L$, we have

$$\begin{aligned} & r \in (I + \sigma\,\partial\,G_1^*)^{-1}(r^*) \Leftrightarrow \frac{r^* - r}{\sigma} \in \partial\,G_1^*(r) \\ & \Leftrightarrow \frac{r^* - r}{\sigma} - \frac{l+o}{2} = Lw^* : \|w^*\|_\infty \leq 1 \text{ and } \|Lw^*\|_{\ell^1} = (w^*, Lr) \\ & \Leftrightarrow \forall i : \begin{cases} r_i = r_i^* - \sigma\frac{l_i + o_i}{2} & \text{if } L_i = 0, \\ |s_i - L_i^{-1} r_i| \leq \sigma \text{ and } |L_i r_i| = \frac{1}{\sigma}(s_i - L_i^{-1} r_i, L_i r_i) & \text{if } L_i \neq 0, \end{cases} \end{aligned}$$

where $s_i = L_i^{-1} r_i^* - \sigma\frac{l_i + o_i}{2L_i}$ and we renumbered the elements $r_{i,j}^c$, $(c, i, j) \in \mathcal{N}$. The first case already gives us an explicit assignment for r_i in the case that $L_i = 0$, i.e. J_i is a point-interval. The second case is further equivalent to

$$\begin{cases} L_i^{-1} r_i = s_i - \sigma & \text{if } s_i > \sigma, \\ L_i^{-1} r_i = s_i + \sigma & \text{if } s_i < -\sigma, \\ r_i = 0 & \text{if } |s_i| \leq \sigma, \end{cases}$$

and plugging in s_i, noting that $L_i \geq 0$ for all i, in all cases $r \in (I + \sigma\,\partial\,G_1^*)^{-1}(r^*)$ is equivalent to

$$\begin{cases} r_i = r_i^* - \sigma\frac{l_i + o_i}{2} - L_i\sigma & \text{if } r_i^* - \sigma\frac{l_i + o_i}{2} > L_i\sigma, \\ r_i = r_i^* - \sigma\frac{l_i + o_i}{2} + L_i\sigma & \text{if } r_i^* - \sigma\frac{l_i + o_i}{2} < -L_i\sigma, \\ r_i = 0 & \text{if } |r_i^* - \sigma\frac{l_i + o_i}{2}| \leq L_i\sigma. \end{cases} \tag{166}$$

Thus an assignment operator

$$(w_1, w_2) \mapsto \begin{pmatrix} (I + \sigma G_1^*)^{-1}(w_1) \\ (I + \sigma G_2^*)^{-1}(w_2) \end{pmatrix}$$

can be defined as $\mathrm{assign}_{J,\mathcal{N}}$ by

$$(\mathrm{assign}_{J,\mathcal{N}}(w))_{i,j}^c = \begin{cases} \begin{cases} \tilde{s}_{i,j}^c - L_{i,j}^c\sigma & \text{if } \tilde{s}_{i,j}^c > L_{i,j}^c\sigma, \\ \tilde{s}_{i,j}^c + L_{i,j}^c\sigma & \text{if } \tilde{s}_{i,j}^c < -L_{i,j}^c\sigma, \\ 0 & \text{if } |\tilde{s}_{i,j}^c| \leq L_{i,j}^c\sigma, \end{cases} & \text{if } (c, i, j) \in N, \\ 0 & \text{else,} \end{cases} \tag{167}$$

Algorithm 4 Scheme of implementation for JPEG 2000 decompression

```
function TGV-JP2($J_{\text{comp}}$)
    $(d, J) \leftarrow$ Decoding of JPEG 2000 object $J_{\text{comp}}$
    $\mathcal{N} \leftarrow J$
    $u \leftarrow W^{-1}(d)$
    $v \leftarrow 0, \overline{u} \leftarrow u, \overline{v} \leftarrow 0, p \leftarrow 0, q \leftarrow 0, w \leftarrow 0$
    choose $\sigma, \tau > 0$ such that $\sigma\tau < (1/\|\mathbf{K}\|)^2$
    repeat
        $p \leftarrow \operatorname{proj}_{\alpha_1}(p + \sigma(\nabla\overline{u} - \overline{v}))$
        $q \leftarrow \operatorname{proj}_{\alpha_0}(q + \sigma(\mathcal{E}(\overline{v}))$
        $w_+ \leftarrow \operatorname{assign}_{J,N}(w + \sigma(W(\overline{u}))$
        $u_+ \leftarrow u - \tau(-\operatorname{div} p + W^* w_+)$
        $v_+ \leftarrow v - \tau(-p - \operatorname{div}_2 q)$
        $\overline{u} \leftarrow (2u_+ - u)$, $\overline{v} \leftarrow (2v_+ - v)$
        $u \leftarrow u_+$, $v \leftarrow v_+$
    until Stopping criterion fulfilled
    return $u_+$
end function
```

where $\tilde{s}^c_{i,j} = w^c_{i,j} - \sigma\frac{l^c_{i,j}+o^c_{i,j}}{2}$.

Bringing equations (164),(165),(167) together finally gives us an explicit representation of $(I+\sigma\,\partial\,\mathbf{F}^*)^{-1}(z)$. A scheme of implementation for the JPEG 2000 decompression process can thus finally be given in algorithm 4. Note that there, J denotes the collection of all data intervals and d the matrix where each entry is the midpoint of the corresponding interval (0 if the interval is $\mathbb{R}$). Further $\nabla, \mathcal{E}, \operatorname{div}, \operatorname{div}$ are again the component wise gradient and divergence operators as defined in subsection 5.2.2. Note also that, based on the stepsize discussion earlier this section, in practice we relax the stepsize constraint $\sigma\tau < (1/\|\mathbf{K}\|)^2$ either by renormalizing W or using an adaptive stepsize, but still can guarantee convergence.

Stopping rule

In order to validate our numerical solution, we again seek for a suitable stopping rule.

Similar as in subsection 5.2.3, we can use that, as shown in the proof of proposition 5.14, at a saddle point $(\hat{x}, \hat{z}) = (\hat{u}, \hat{v}, \hat{p}, \hat{q}, \hat{w})$, the primal problem (147) and the dual problem (148) coincide, i.e.

$$\mathbf{F}(\mathbf{K}\hat{x}) = -\mathcal{I}_{\{0\}}(-\mathbf{K}^*\hat{z}) - \mathbf{F}^*(\hat{z}),$$

to estimate, for any $(x, z) = (u, v, p, q, w) \in X \times Z$,

$$\begin{aligned} 0 \le \mathbf{F}(Kx) - \mathbf{F}(\mathbf{K}\hat{x}) = \quad & \mathbf{F}(\mathbf{K}x) + \mathcal{I}_{\{0\}}(-\mathbf{K}^*\hat{z}) + \mathbf{F}^*(\hat{z}) \\ \le \quad & \mathbf{F}(\mathbf{K}x) + \mathcal{I}_{\{0\}}(-\mathbf{K}^*z) + \mathbf{F}^*(z). \end{aligned} \tag{168}$$

Now plugging in the definitions of $\mathbf{F}, \mathbf{K}$ (note that $G^1(w_1) + G^2(w_2) = \mathcal{I}_D(Ww)$) yields

$$\begin{aligned}
0 \leq & \qquad \|\nabla u - v\|_1 + \|\mathcal{E}v\|_1 + \mathcal{I}_D(Wu) - \|\nabla \hat{u} - \hat{v}\|_1 - \|\mathcal{E}\hat{v}\|_1 \\
\leq & \quad \|\nabla u - v\|_1 + \|\mathcal{E}v\|_1 + \mathcal{I}_D(Wu) + \sup_{u',v'} ((\operatorname{div} p - W^* w, u') + (p + \operatorname{div} q, v')) \\
& \qquad + \mathcal{I}_{\|\cdot\|_\infty \leq \alpha_1}(p) + \mathcal{I}_{\|\cdot\|_\infty \leq \alpha_0}(q) + (\frac{l+o}{2}, w_1) + \|\frac{o-l}{2} \cdot w_1\|_{\ell^1} + \mathcal{I}_{\{0\}}(w_2).
\end{aligned}$$

Using the iterates of the primal dual algorithm 4, denoted by $x_n = (u_n, v_n)$, $z_n = (p_n, q_n, r_n)$, instead of arbitrary x, z simplifies this estimation to

$$\begin{aligned}
0 \leq & \qquad \|\nabla u_n - v_n\|_1 + \|\mathcal{E}v_n\|_1 + \mathcal{I}_D(Wu_n) - \|\nabla \hat{u} - \hat{v}\|_1 - \|\mathcal{E}\hat{v}\|_1 \\
\leq & \quad \|\nabla u_n - v_n\|_1 + \|\mathcal{E}v_n\|_1 + \mathcal{I}_D(Wu_n) + \sup_{u' \in U} (\operatorname{div} p_n - W^* w_n, u') \\
& \qquad + \sup_{v' \in V} (p_n + \operatorname{div} q_n, v') + (\frac{l+o}{2}, (w_n)_1) + \|\frac{o-l}{2} \cdot (w_n)_1\|_{\ell^1}.
\end{aligned}$$

But still, since for the iterates $(u_n, v_n), (p_n, q_n, r_n)$ in general neither $Wu_n \in D$ nor $\operatorname{div} p_n = W^* w_n$ nor $p_n = -\operatorname{div} q_n$ is satisfied, we cannot use this estimations without further modifications.

Note that we cannot simply define $\tilde{w_n} = (W^*)^{-1} \operatorname{div}^2 \tilde{q}_n$, with a modified $\tilde{q}_n$ as for the stopping rule for the JPEG algorithm, since $\mathcal{I}_{\{0\}}((\tilde{w}_n)_2) \neq \infty$ has to be ensured. Solving the equation $\operatorname{div}^2 q_n - W^* w_n$ for q_n and again modifying this q_n would resolve this issue. But this requires the solve of a PDE in each iteration step, yielding unpractical computation times, and still would not get rid of $\mathcal{I}_D(Wu)$. A projection of u_n to U_D would resolve also this issue, but introduce an additional computational expensive equation solve in each iteration.

We propose a different approach: Our aim is to bound

- $\mathcal{I}_D(Wu_n)$,
- $\sup_{u' \in U} (\operatorname{div} p_n - W^* w_n, u')$,
- $\sup_{v' \in V} (p_n + \operatorname{div} q_n, v')$.

We start with $\mathcal{I}_D(Wu_n)$: Considering the optimality condition for the discrete saddle point problem as derived in proposition 5.15, the data term $\mathcal{I}_D \circ W$ of the original problem implies that any optimal solution $(\hat{u}, \hat{v}), (\hat{p}, \hat{q}, \hat{w})$ must satisfy

$$\begin{cases}
\hat{w}_i \geq 0 & \text{if } (W\hat{u})_i = o_i, \\
\hat{w}_i \leq 0 & \text{if } (W\hat{u})_i = l_i, \\
\hat{w}_i = 0 & \text{if } (W\hat{u})_i \in \overset{\circ}{[l_i, o_i]}, \\
\hat{w}_i \in \mathbb{R} & \text{if } (W\hat{u})_i = l_i = o_i,
\end{cases} \tag{169}$$

for all $0 \leq i < 3N^2$, where we reordered indices such that $(w_i)_{0 \leq i < 3N^2} \simeq (w^c_{i,j})_{\substack{0 \leq i,j < N \\ c \in \{1,2,3\}}}$. If we now use $I_{D_C} \circ W$ instead of $\mathcal{I}_D \circ W$ as data term, where $I_{D_C}(w) = \sum_{i=0}^{3N^2-1} I^i_{D_C}(w_i)$ with

$$I^i_{D_C}(w_i) = \begin{cases}
0 & \text{if } w_i \in [l_i, o_i], \\
C_i(w_i - o_i) & \text{if } w_i > o_i, \\
C_i(l_i - w_i) & \text{if } w_i < l_i,
\end{cases}$$

for fixed $(C_i)_{0\leq i<3N^2}$, we arrive at a modified optimization problem. The optimality condition for this modified problem then requires, in addition to (169), that

$$|w_i| \leq C_i \quad \text{for all } 0 \leq i < 3N^2.$$

Note that the C_i are nothing but Lagrange multipliers for the data constraint. Choosing now $\hat{C} = (\hat{C}_i)_{0\leq i<3N^2}$ with $\hat{C}_i = |\hat{w}_i|$, where $(\hat{u},\hat{v}),(\hat{p},\hat{q},\hat{w})$ is an optimal solution of the original problem, we get that $(\hat{u},\hat{v}),(\hat{p},\hat{q},\hat{w})$ is also a solution of the modified problem (and again equality of the resulting modified primal- and dual problem holds). Estimating the primal dual gap for the modified problem thus yields, for $(x,z) = (u,v,p,q,w) \in X \times Z$ arbitrary,

$$\begin{aligned} 0 \leq \quad & \|\nabla u - v\|_1 + \|\mathcal{E}v\|_1 + I_{D_{\hat{C}}}(Wu) - \|\nabla\hat{u} - \hat{v}\|_1 - \|\mathcal{E}\hat{v}\|_1 \\ \leq \quad & \|\nabla u - v\|_1 + \|\mathcal{E}v\|_1 + I_{D_{\hat{C}}}(Wu) + \sup_{u'}(\operatorname{div} p - W^*w, u') \\ & + \sup_{v'}(p + \operatorname{div} q, v') + \mathcal{I}_{\|\cdot\|_\infty\leq\alpha_1}(p) + \mathcal{I}_{\|\cdot\|_\infty\leq\alpha_0}(q) + \mathcal{I}^*_{D_{\hat{C}}}(w), \end{aligned} \tag{170}$$

which requires an explicit form of $I^*_{D_{\hat{C}}}$. This can be given as follows: First note that

$$I^*_{D_C}(w^*) = \sup_{w\in U}(w^*,w) - I_{D_C}(w) = \sum_i \sup_\lambda (w_i^*,\lambda) - I^i_{D_C}(\lambda).$$

Now in the case that $|w_i^*| = C_i + \epsilon$, $\epsilon > 0$, we get

$$\sup_\lambda (w_i^*,\lambda) - I^i_{D_C}(\lambda) = \begin{cases} \sup_{\lambda\geq o_i}(C_i+\epsilon,\lambda) - (C_i,\lambda - o_i) = \infty & \text{if } w_i^* = C+\epsilon, \\ \sup_{\lambda\leq l_i}(-C_i-\epsilon,\lambda) - (C_i, l_i - \lambda) = \infty & \text{if } w_i^* = -C-\epsilon. \end{cases}$$

In the case $|w_i^*| \leq C_i$ it is straightforward to show that $\sup_{\lambda\in\mathbb{R}}(w^*,\lambda) - I^i_{D_C}(\lambda) = \sup_{\lambda\in J_i}(w^*,\lambda)$ and thus, in the case $|w_i^*| \leq C_i$ for all i, $I^*_{D_C}(w^*)$ takes the form

$$I^*_{D_C}(w^*) = (\frac{l+o}{2}, w_1) + \|\frac{o-l}{2}\cdot w_1\|_{\ell^1} + \mathcal{I}_{\{0\}}(w_2)$$

coincides with $\mathcal{I}^*_D$ in (144). In total, with $S(C) = \{z \in U \| |z_i| \leq C_i\}$, we can thus write

$$I^*_{D_C}(w^*) = \mathcal{I}_{S(C)}(w) + (\frac{l+o}{2}, w_1) + \|\frac{o-l}{2}\cdot w_1\|_{\ell^1} + \mathcal{I}_{\{0\}}(w_2).$$

Using the primal dual gap estimation of the modified problem as in (170), but for the iterates of algorithm 4, thus allows us to control $\mathcal{I}_D(Wu_n)$.

Next consider $\sup_u(\operatorname{div} p_n - W^*w_n, u)$: Setting $\hat{T}$ to be an upper bound of the L^p norm of an optimal solution $\hat{u}$ of the original problem, $1 \leq p \leq \infty$, we can, similar as before, also add $\mathcal{I}_{\|\cdot\|_p\leq\hat{T}}$ to the modified saddlepoint problem. With that, we finally get that the optimal solution $(\hat{u},\hat{v}),(\hat{p},\hat{q},\hat{w})$ of the original problem also solves

$$\begin{aligned} \min_{x=(u,v)} \max_{z=(p,q,w)} (\mathbf{K}x,z) - F^*(p,q) - \mathcal{I}_{S(\hat{C})}(w) - (\frac{l+o}{2}, w_1) - \|\frac{o-l}{2}\cdot w_1\|_{\ell^1} \\ -\mathcal{I}_{\{0\}}(w_2) + \mathcal{I}_{\|\cdot\|_p\leq\hat{T}}(u). \end{aligned} \tag{171}$$

Using again the primal dual gap estimation for (171) and plugging in the iterates of algorithm 4 for the original problem, yields

$$\begin{aligned} 0 \leq & \quad \|\nabla u_n - v_n\|_1 + \|\mathcal{E}v_n\|_1 + I_{D_{\hat{C}}}(Wu_n) - \|\nabla \hat{u_n} - \hat{v_n}\|_1 - \|\mathcal{E}\hat{v_n}\|_1 \\ \leq & \quad \|\nabla u_n - v_n\|_1 + \|\mathcal{E}v_n\|_1 + I_{D_{\hat{C}}}(Wu_n) + \hat{T}\|\operatorname{div} p_n - W^* w_n\|_{p'} \\ & + \sup_v (p_n + \operatorname{div} q_n, v) + \mathcal{I}_{\|\cdot\|_\infty \leq \alpha_1}(p_n) + \mathcal{I}_{\|\cdot\|_\infty \leq \alpha_0}(q_n) + \mathcal{I}_{S(\hat{C})}(w_n) \\ & + (\frac{l+o}{2}, (w_1)_n) + \|\frac{o-l}{2} \cdot (w_1)_n\|_{\ell^1}, \end{aligned}$$

with $p' = \frac{p}{p-1}$. Now it is only left to estimate $\sup_v(p_n + \operatorname{div} q_n, v)$: As already done for the JPEG related minimization problem, we can define $\tilde{q}_n = \beta_n q_n$ with $\beta_n := \frac{\alpha_1}{\max(\alpha_1, \|\operatorname{div} q_n\|_\infty)}$,

$$\tilde{p}_n = -\operatorname{div} \tilde{q}_n \tag{172}$$

and finally get

$$\begin{aligned} 0 \leq & \quad F(Kx_n) + I_{D_{\hat{C}}}(Wu_n) - F(K\hat{x}) \\ \leq & \quad F(Kx_n) + I_{D_{\hat{C}}}(Wu_n) + \hat{T}\|\operatorname{div} \tilde{p}_n - W^* w_n\|_{p'} \\ & + \mathcal{I}_{S(\hat{C})}(w_n) + (\frac{l+o}{2}, (w_n)_1) + \|\frac{o-l}{2} \cdot (w_n)_1\|_{\ell^1} \end{aligned}$$

where $I_{D_{\hat{C}}}$ can be written in a compact form as

$$I_{D_{\hat{C}}}(r) = \sum_{(r)_i \notin [l_i, o_i]} \hat{C}_i \max\{(r)_i - o_i, l_i - (r)_i\}.$$

for $r \in U$. Provided that $w_n \in S(\hat{C})$, this gives us a computational feasible estimation of the error in terms of functional values, if we know $\hat{T}$ and $(\hat{C}_i)_{0 \leq i < 3N^2}$, which depend on the optimal solution. Further, due to continuity of all terms and our specific choice of $(\tilde{p}_n, \tilde{q}_n, w_n)$, the last term converges to zeros as $(u_n, v_n), (p_n, q_n, w_n)$ converge to an optimal solution.

Using the current iterates to estimate $\hat{T}$ and $(\hat{C}_i)_{0 \leq i < 3N^2}$, we can finally sum up the results above and the proposed stopping criterion by the following proposition:

Proposition 5.16. *Let $\gamma > 1$, $1 \leq p \leq \infty$ $x_n = (u_n, v_n), z_n = (p_n, q_n, w_n)$ be the iterates of algorithm 4 and $(\hat{u}, \hat{v}), (\hat{p}, \hat{q}, \hat{w})$ be an optimal solution to* (149). *Then, defining*

$$\begin{aligned} \mathcal{G}(x_n, z_n) := & \quad F(Kx_n) + I_{D_{C_n}}(Wu_n) + T_n\|\operatorname{div} \tilde{p}_n - W^* w_n\|_{p'} \\ & + (\frac{l+o}{2}, (w_n)_1) + \|\frac{o-l}{2} \cdot (w_n)_1\|_{\ell^1} \end{aligned} \tag{173}$$

with $I_{D_{C_n}}$ as

$$I_{D_{C_n}}(r) = \sum_{(r)_i \notin [l_i, o_i]} (C_n)_i \max\{(r)_i - o_i, l_i - (r)_i\}, \tag{174}$$

where $(C_n)_i = \gamma |(w_n)_i|$, $T_n := \gamma \|u_n\|_p$ and $\tilde{p}_n$ as in (172), *we get that*

$$\mathcal{G}(x_n, z_n) \to 0 \text{ as } n \to \infty$$

and, additionally,

$$\begin{aligned} 0 \leq & \quad F(Kx_n) + I_{D_{\hat{C}}}(Wu_n) - F(K\hat{x}) \\ \leq & \quad F(Kx_n) + I_{D_{C_n}}(Wu_n) + T_n \| \operatorname{div} \tilde{p}_n - W^* w_n \|_{p'} \\ & \quad +(\frac{l+o}{2}, (w_n)_1) + \| \frac{o-l}{2} \cdot (w_n)_1 \|_{\ell^1} \end{aligned} \tag{175}$$

whenever $T_n \geq \|\hat{u}\|_p$ *and* $(C_n)_i \geq |\hat{w}_i|$ *for all* $0 \leq i < 3N^2$.

Proof. The prove can be done as described above, where the last estimation follows since $I_{D_{C_n}}(r) \geq I_{D_{\hat{C}}}(r)$ in the case that $(C_n)_i \geq |\hat{w}_i|$ for all $0 \leq i < 3N^2$. □

This allows, for given $\epsilon > 0$, to use $\mathcal{G}(x_n, y_n) < \epsilon$ as stopping criterion and provides, at least in the limit, a suitable estimate of the error in terms of functional values.

Note that we cannot expect to get the estimate

$$\mathcal{G}(x_n, z_n) \geq F(Kx_n) - F(K\hat{x}) \geq 0$$

since the iterates $(u_n)_{n\in\mathbb{N}}$ are only contained in the data set U_D in the limit and thus it is possible that $F(Kx_n) < F(K\hat{x})$. This was observed also in numerical experiments.

5.3.4 Numerical experiments

The aim of this subsection is to provide numerical results obtained with our framework for improved decompression of JPEG 2000 images. As we will see, the method performs very well but, in contrast to JPEG decompression, some prior considerations are needed for efficient usage of the proposed algorithm.

For our experiments we use again three component color images and grayscale images with component range $[0, 255]$. Except for the combined decompression and zooming examples, we always fix the ratio $\frac{\alpha_0}{\alpha_1}$ for definition of the TGV_α^2 functional to $\sqrt{2}$.

We will use several iteration dependent quantities to measure convergence and reconstruction quality in the subsequent experiments. Besides the maximal and average data error, i.e., the average and maximal distance of the wavelet coefficients to the data intervals (only for those how are bounded), we also use $\mathrm{TGV}(u_n) + I_{D_{C_n}}(Wu_n)$ and $\overline{\mathcal{G}}(x_n, y_n)$, where $I_{D_{C_n}}$ is defined as in (174) and $\overline{\mathcal{G}}$ again is a normalization of $\mathcal{G}$ as in (173), defined by

$$\overline{\mathcal{G}}(x_n, y_n) = \frac{\mathcal{G}(x_n, y_n)}{N^2}, \tag{176}$$

with N^2 the number of image pixels. As for the application to JPEG decompression, the normalization is motivated by making $\mathcal{G}$ image size independent and getting an estimation on an average pixel error; remember that, in the limit, we have the estimation

$$0 \leq \mathrm{TGV}(u_n) + I_{D_{C_n}}(Wu_n) - \mathrm{TGV}(\hat{u}) \leq \overline{G}(x_n, y_n).$$

Note also that for the evaluation of $\overline{\mathcal{G}}(x_n, y_n)$ it is left to choose γ and p, and with that fixing $p' = \frac{p}{p-1}$ of (173). We will discuss a suitable choice later on, but for the moment those parameters are fixed to $\gamma = 1.001$, $p = 2$.

In contrast to the JPEG decompression setting, we did not implement CPU or GPU optimized code for the TGV based JPEG 2000 decompression algorithm. Considering computation times, however, one could expect that each iteration step of an optimized JPEG 2000 decompression scheme would take about as long as for the JPEG decompression setting. Indeed, from the computational viewpoint, the only considerable difference is that instead of a Block-DCT transform, a wavelet transform is used.

Let us now start by discussing a suitable stepsize choice:

Stepsize choice

As we have seen in subsection 5.3.3, the choice of a suitable stepsize for the primal dual algorithm as in algorithm 4 is more difficult than for the case of JPEG decompression. A potentially very high value of $\|W\|$, and, consequently, of $\|\mathbf{K}\|$, yields very small stepsizes σ, τ when respecting the restriction $\sigma\tau\|\mathbf{K}\|^2 < 1$ without further modification.

Renormalization: One possibility to resolve this issue is to renormalize the wavelet transform operator W as in (154) and modify the data accordingly. In theory, this allows to set $\|W\|$ arbitrary low. In practice however, choosing $\|W\|$ too small reduces data influence and with that again convergence speed. In figure 18 we have plotted several iteration depended quantities for the decompression of the birds eye images as shown in figure 17 for different choices of $\|W\|$. The top plots show the development of $\text{TGV}(u_n) + I_{D_{C_n}}(Wu_n)$ and $\overline{\mathcal{G}}(x_n, y_n)$, while the bottom plots of figure 18 show the maximal and average error in the wavelet data. Remember that, in contrast to JPEG decompression, we cannot assure at every iteration step that the current image is contained in the set of possible source data, but only that the optimal solution is.

As we can observe in all four plots of figure 18, setting $\|W\|$ too small leads to oscillations while setting it too high yields slow convergence. In particular the choice $\|W\| = 1$ yields high oscillations in the objective functional as well as the data error and as a result, even after 5000 iterations, the algorithm is far from the optimal solution. In contrast to that, setting $\|W\| = 50$ yields stable but very slow convergence. For the setting of figure 18 the choice $\|W\| = 20$ seems best.

This however, depends on the image size and, more importantly, the level of wavelet decomposition of the image to decompress: The birds eye image used for figure 18 was compressed by using two levels of wavelet decomposition. Since our estimate on $\|W\|$ can be expected to be too conservative, and since it highly increases with the level of wavelet decomposition, in order to ensure $\|W\| = 20$ for wavelet level 5, W has to be divided by a much higher quantity than for wavelet level 2, leading again to oscillatory behaviour. This can be observed in figure 20, where again the development of iteration dependent quantities is shown for the same image but different levels of wavelet decomposition. As we can see, in particular for 5 levels of wavelet decomposition, the choice $\|W\| = 20$ yields a high data error even after 20 000 iterations. This effect is much weaker for the case of 2 levels of wavelet decompositions, but still, as one can observe

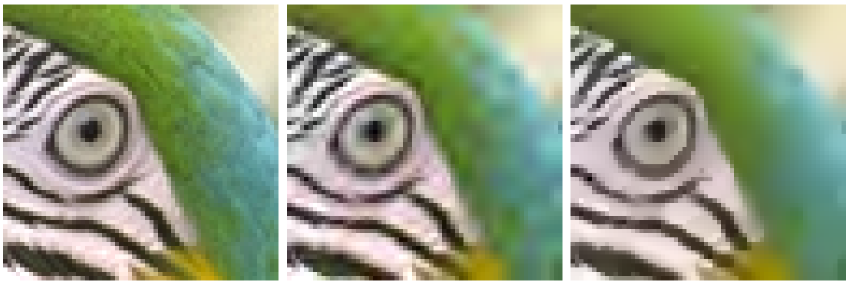


Figure 17: Birds eye image (64×64) used in the numerical experiments for figures 18,22. From left to right: Original, uncompressed image. JPEG 2000 compressed image at 0.19 bpp. True solution of TGV based decompression, i.e., images obtained with adaptive stepsize after 15000 iterations.

in figure 19, the reconstructions for $\|W\| = 20$ are overregularized and the reconstruction for $\|W\| = 160$ look better. In particular, brightness oscillations in the background are better preserved in the case $\|W\| = 160$. This suggests that the choice of $\|W\|$ has to be increased depending on the level of wavelet decomposition, e.g. for the most common case of five levels of wavelet decomposition, motivated by our experiments, on could choose $\|W\| = 160$. While this indeed allows a good reconstruction quality, in practice it takes the large number of about 10^5 iterations to obtain this reconstruction.

Summing up, we can numerically confirm the convergence of the algorithm with renormalized W. Also, the renormalization significantly improves convergence speed in a situation where, without renormalization, the stepsize would be impracticable small. This at least allows to obtain an optimal solution. But still, the necessary iteration number is far to high for a reasonable applicability of the proposed method. Thus an alternative is needed, which will be discussed in the following.

Adaptive stepsize: To improve convergence speed, the second suggestion in subsection 5.3.3 was, instead of renormalizing W, to adaptively choose the stepsizes σ and τ as in (156). This idea was motivated by the fact that we observed convergence of algorithm 4 even for infeasible large stepsize choices such as $\sigma = \tau = \frac{1}{3}$. In figure 21 we show some iteration dependent quantities for different, also adaptive, stepsize choices.

As one can see, fixing the stepsize to $\sigma = \tau = \frac{1}{\sqrt{8}} =\approx 0.3536$ does not lead to convergence, while starting with $\sigma = \tau = \frac{1}{\sqrt{8}}$ but allowing the algorithm to adapt the stepsize does. We have observed that for this choice, the stepsize gets reduced only once to $\sigma = \tau \approx 0.3359$ leading to convergence of the algorithm. For the choice $\sigma = \tau = \frac{1}{3}$ as initial stepsize the algorithm shows similar convergence properties, but the stepsize gets reduced not even once. We observed the same behaviour in all our numerical experments, thus we will use $\sigma = \tau = \frac{1}{3}$ in all future experiments, allowing the algorithm to reduce the stepsize. The resulting decompression of the birds image and other examples can be seen in figures 23,24 later on. As can also be seen in figure 21, normalizing $\|W\|$ to 160 and fixing the stepsize accordingly yields very slow convergence compared to the adaptive choice.

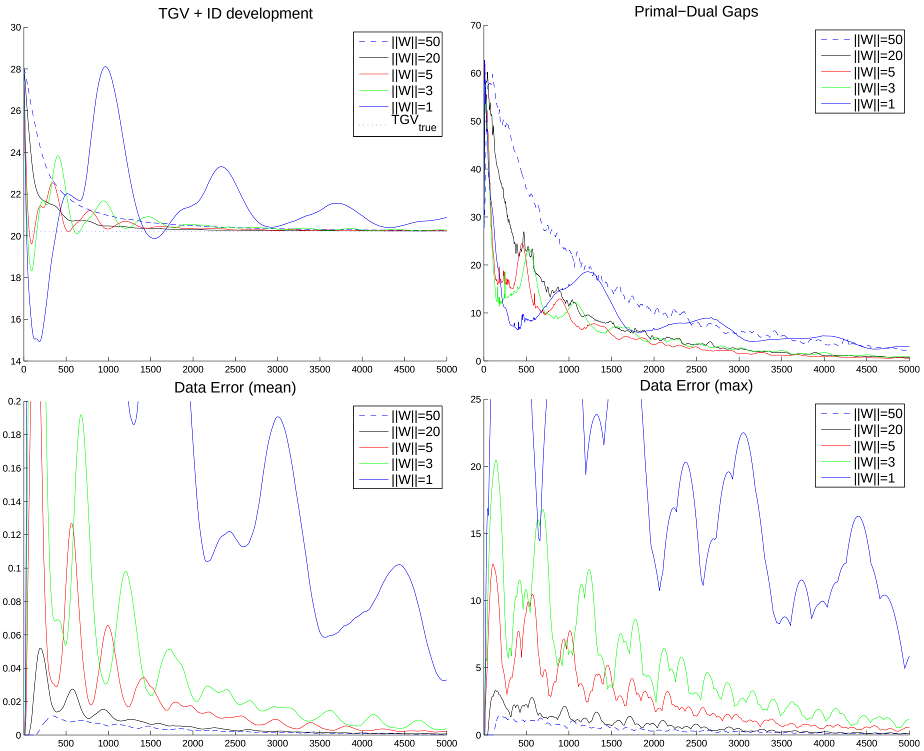


Figure 18: Comparison of iteration dependent quantities while decompressing the parrots eye image of figure 17 for different renormalization of W.

Primal dual gap parameters

As for the application to JPEG decompression, we again have free choice of the parameters γ and p for evaluation of the primal dual gap. In previous experiments we already chose $\gamma = 1.001$ and $p = 2$. We will now see that this indeed is a reasonable choice: Figure 22 evaluates different choices of γ and p. As can be seen in the right plot, in contrast to JPEG decompression, this time the choice $\gamma = 0$ violates the estimation (175), but the choice $\gamma = 1.001$, $p = 2$ does not. This can be explained by the fact that this time the unbounded data part, and with that the choice of γ, has much more influence. As can be seen in figure 22 on the left, even not violating estimation (175), the choices $\gamma = 1.001$ and $p = 2$ results again in the lowest primal dual gap. Thus we will use this choice for all future experiments.

Experimental results for decompression

Having fixed all remaining parameters for the JPEG 2000 decompression process, we can now discuss the reconstruction quality in numerical experiments. For convenience we sum up the proposed parameter choices for JPEG 2000 decompression in table 4

Figure 23 compares the standard reconstruction with the TGV based reconstruction. As one can see, in the TGV based reconstruction, all wavelet artifacts

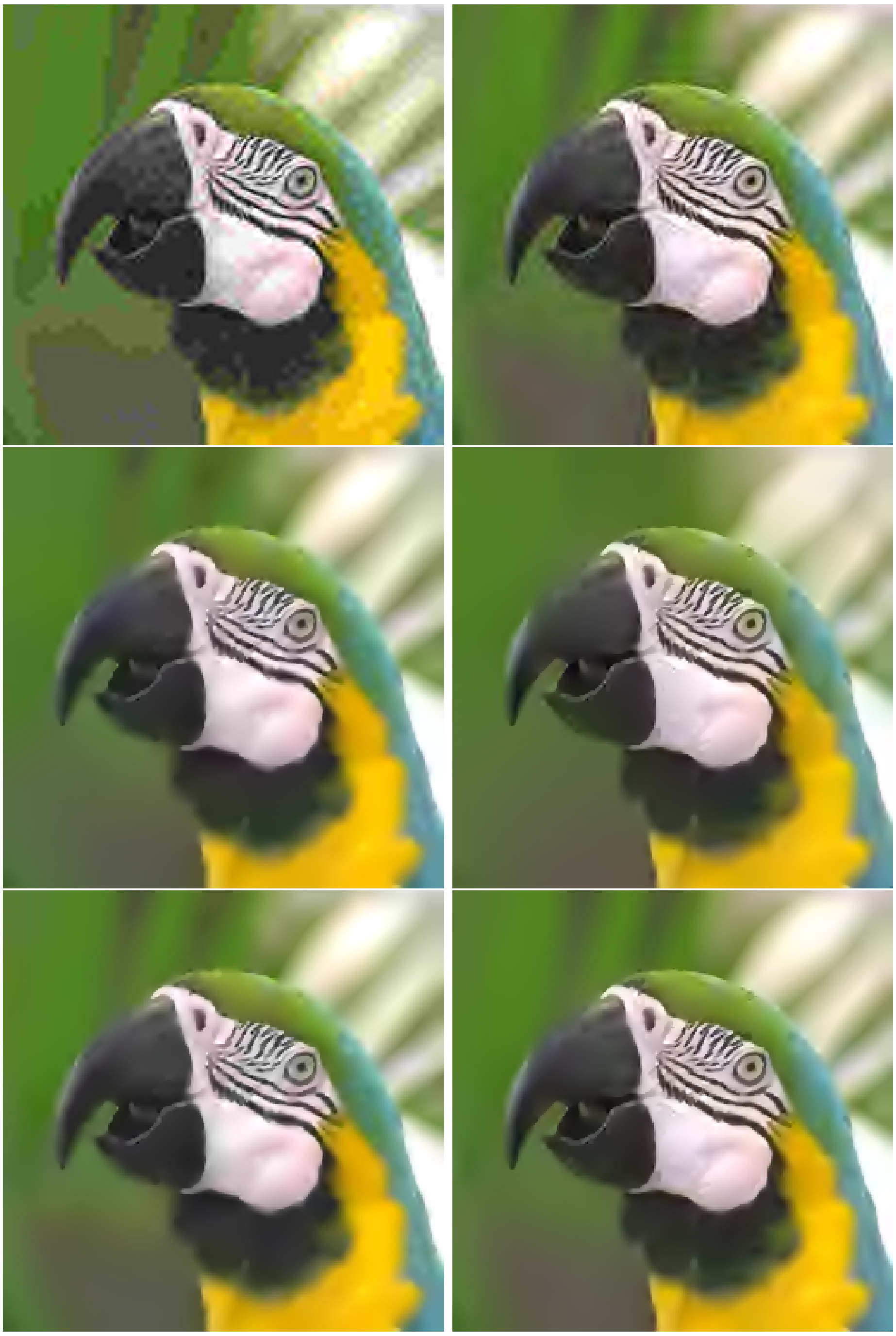


Figure 19: Bird image (256×256) used for the numerical experiments of figures 20 and 21. Left, from top to bottom: JPEG 2000 compressed image using 2 wavelet levels (0.3 bpp), TGV based reconstruction with $\|W\|$ set to 20, TGV based reconstruction with $\|W\|$ set to 160. Right, from top to bottom: JPEG 2000 compressed image using 5 wavelet levels (0.3 bpp), TGV based reconstruction with $\|W\|$ set to 20, TGV based reconstruction with $\|W\|$ set to 160. For the choice $\|W\| = 20$ and 5 wavelet levels, brightness oscillations in the background are not recovered even after 20 000 iterations.

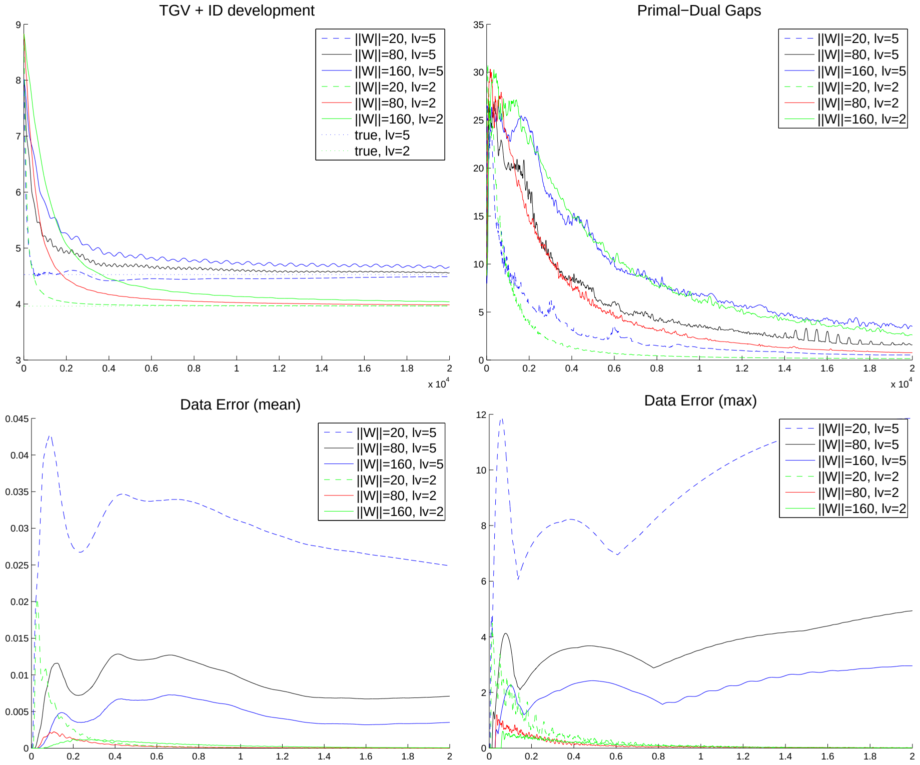


Figure 20: Comparison of iteration dependent quantities while decompression the images of figure 19 for different renormalizations of W. For five levels of wavelet decomposition, the choice $\|W\| = 160$ seem best in terms of data error.

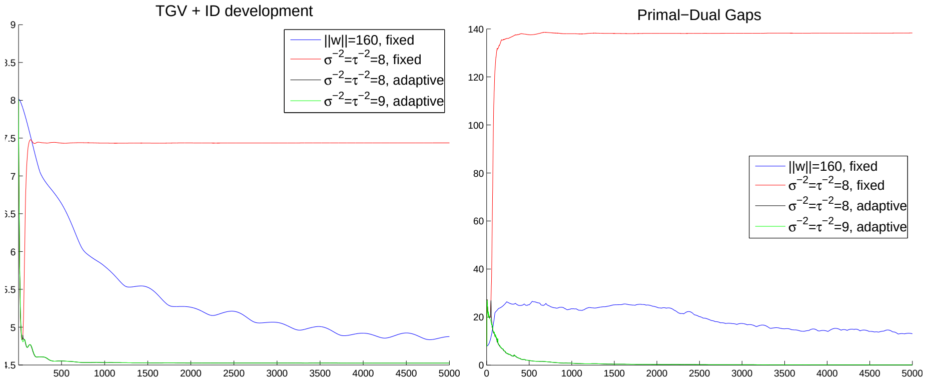


Figure 21: Comparison of TGV +ID and $\overline{\mathcal{G}}$ for different stepsize choices. The infeasible choice $\sigma^{-2} = \tau^{-2} = 8$ yields convergence if the stepsize is choosen adaptively.

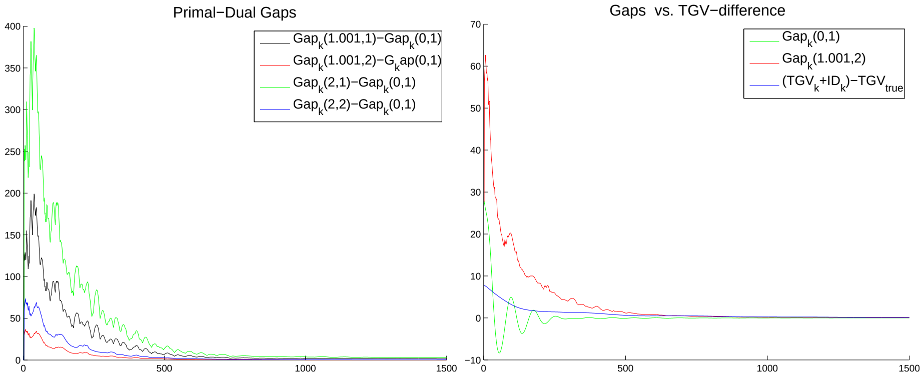


Figure 22: Left: Difference of primal dual gap for different choices of γ, p to primal dual gap obtained with $\gamma = 0$. Right: Comparison of primal dual gap for $\gamma = 1.001$, $p = 2$ and $\gamma = 0$ to difference of $\text{TGV}(u_k) + \mathcal{I}_{D_{C_k}}(u_k)$ and optimal TGV value (obtained with adaptive stepsize 15000 iterations). As one can see, the primal dual gap with $\gamma = 0$ violates the estimation $\mathcal{G}(x_k, y_k) \geq \text{TGV}(u_k) + \text{ID}(u_k) - \text{TGV}(\hat{u})$.

Table 4: Parameter setting for JPEG 2000 decompression

Algorithm 4		**TGV** as in (57)		**Gap** as in (173)	
σ	adaptive, initialization 1/3	$\frac{\alpha_0}{\alpha_1}$	$\sqrt{2}$	γ	1.001
τ	adaptive, initialization 1/3			p	2

have been removed while edges are kept sharp. In general, with our method, the reconstruction quality is again highly improved and yields more natural and visually more appealing images. For figure 23, we compressed the same parrot image once using 4 tiles and once without tiling. Even at the same bit rate, the image with tiling looks worse and the tile boundaries are clearly visible. In the improved reconstruction, these artificial edges as well as the stronger wavelet artifacts have been removed completely. But also the improved reconstruction quality is slightly worse when using multiple tiles. We have again used a normalized primal dual gap below 10^{-1} as stopping rule to obtain the images in figure 23. The larger number of iterations, compared to JPEG reconstruction, necessary to achieve this bound may be due to the unboundedness of some data intervals for JPEG 2000 decompression. However, as we will see in figure 25 later on, this bound is merely important to ensure optimality, in practical applications a much lower iteration number is sufficient to obtain a reconstruction visually almost indistinguishable from the optimal one.

Figure 24 then allows a more detailed comparison of the standard and TGV based reconstruction for two different images. On the top, the close up of Barbara's face shows again a more natural reconstruction, while the lower two lines show the standard and TGV based reconstruction and an error visualization. As one can see, the total energy of the error has been significantly reduced with the TGV based reconstruction and, again, the error is of a less visible type.

Figure 25 now allows to compare reconstruction quality for the parrot image by performing smaller iterations numbers. It compares the standard decom-

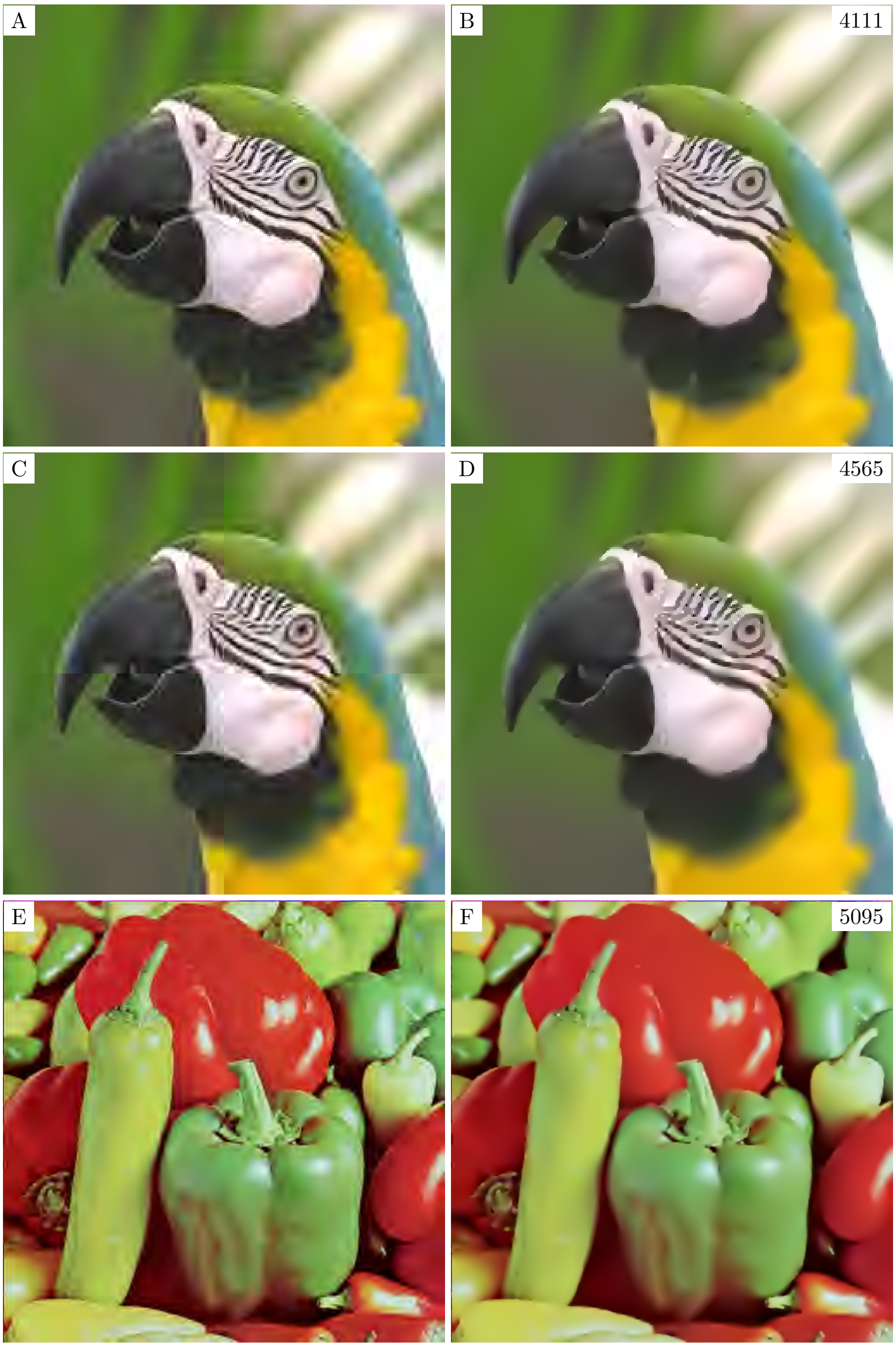


Figure 23: On the left: Standard decompression. On the right: TGVbased reconstruction obtained with normalized primal dual gap below 10^{-1} as stopping criterion (number of iterations on top-right). A-B: Parrot image at 0.3 bpp (256×256 pixels). C-D: Parrot image with tiling at 0.3 bpp (256×256 pixels). E-F: Peppers image at 0.15 bpp (512×512 pixels).

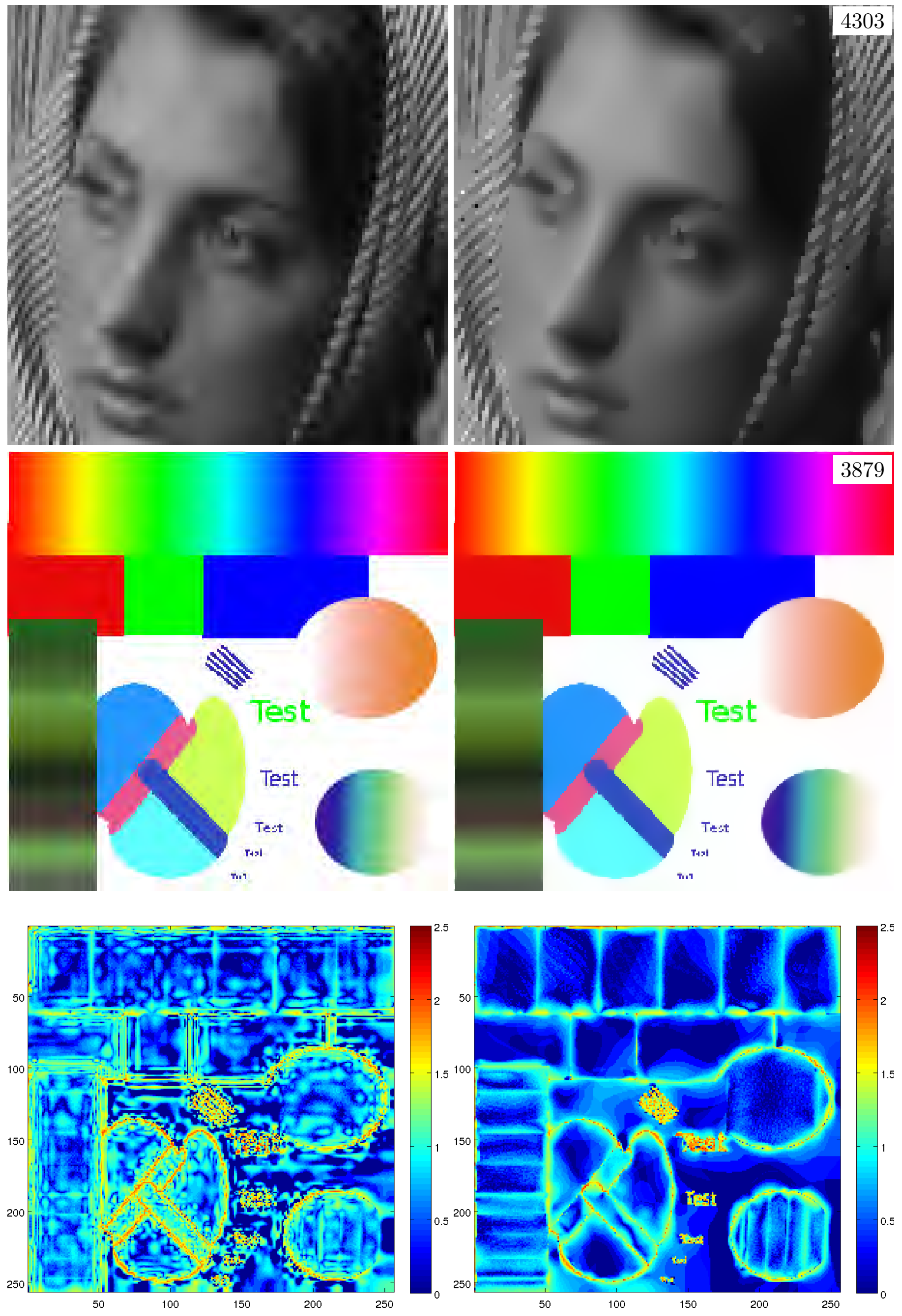


Figure 24: Left: Standard decompression, right: TGV based reconstruction. Top: Close up of Barbara images at 0.4 bpp (512 × 512 pixels) Middle: Color test image at 0.6 bpp (256×256 pixels). Bottom: Visualization of the pointwise reconstruction error (logarithmic scale). The numbers in the image show the iteration number until $\overline{\mathcal{G}}(x_n, y_n) < 10^{-1}$ was satisfied.

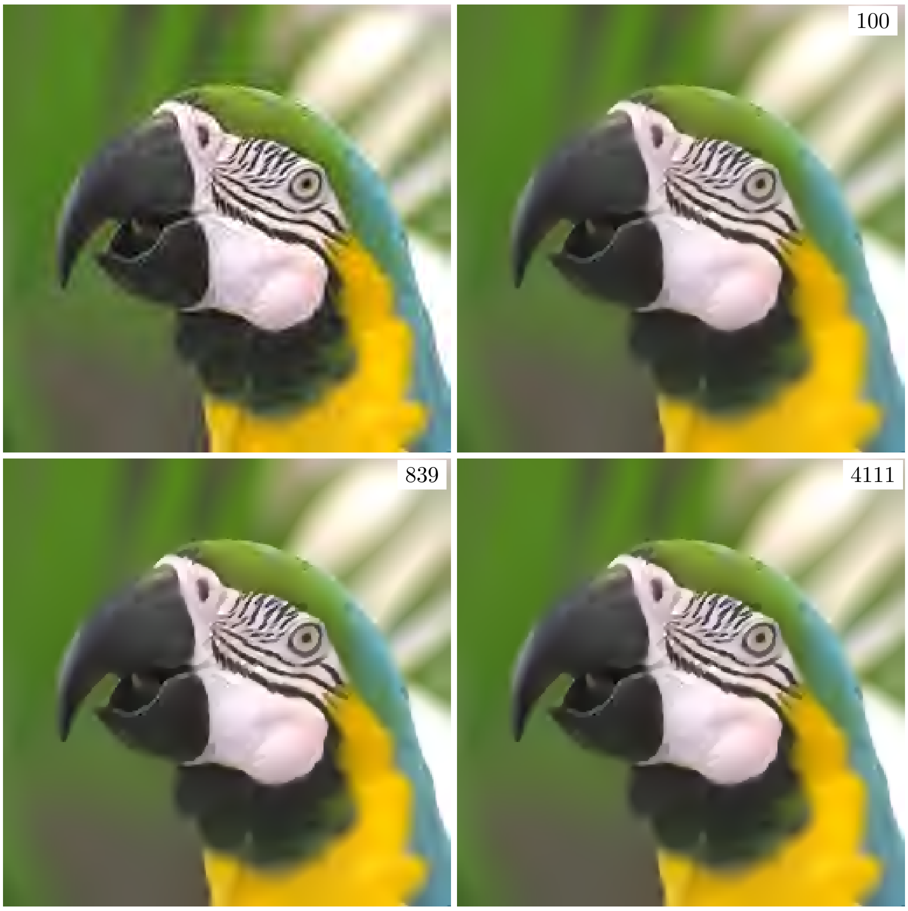


Figure 25: Decompression of parrot image (0.3 bpp, 256 × 256 pixels) using only a few number of iterations. The standard decompression is shown in the top left corner. The iteration number for the improved reconstructions is shown in the top right of the image. The bottom images were obtained using $\overline{\mathcal{G}}(x_n, y_n) < 1$ and $\overline{\mathcal{G}}(x_n, y_n) < 0.1$ as stopping rule, respectively.

pression with reconstructions obtained by using 100 iterations-, $\overline{\mathcal{G}}(x_n, y_n) < 1$ (839 iterations) and again $\overline{\mathcal{G}}(x_n, y_n) < 0.1$ (4111 iterations) as stopping rule. As one can see, even after 100 iterations the JPEG 2000 artifacts are already removed and the image is comparable to the optimal solution at 4111 iterations. However, in contrast to JPEG decompression, for such suboptimal reconstructions a perfect fit to data cannot be ensured and, as we have already seen, is not satisfied in practice.

Unbounded data intervals

As already mentioned in subsection 5.3.1, to define the set of possible source images for JPEG 2000 decompression, we cannot bound every coefficient in the wavelet domain, i.e., some of the data intervals for the pointwise restriction of the wavelet coefficients contain all of $\mathbb{R}$. Figure 26 visualizes the wavelet

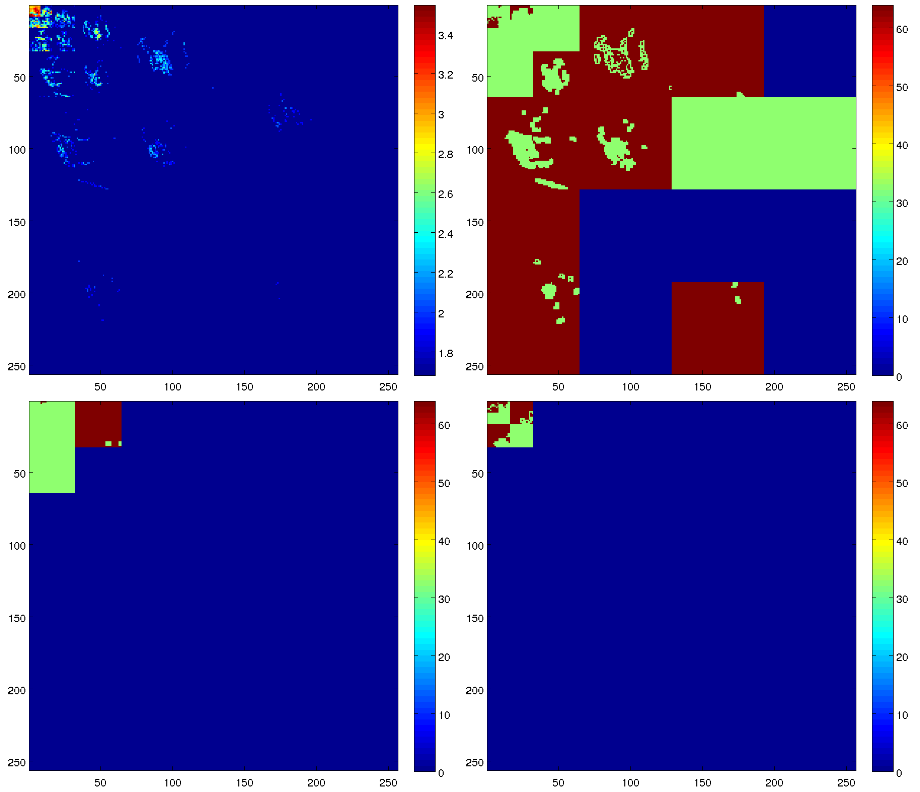


Figure 26: Top left: Wavelet coefficients of the brightness component of the birds images (logarithmic scale). Top right: Size of data intervals for the wavelet coefficients of the brightness component. Bottom: Size of data intervals for the Cb and Cr component. Note that 0, i.e. dark blue, indicates that the data is unbounded.

decomposition of the brightness component as well as the bounds on the wavelet coefficients for all color components for the parrot image without tiling of figure 23. As one can see, details such as the cheek of the parrot yield a better bound on the wavelet coefficients. Also, there is much less boundedness information left for the color components, which can be seen as implicit color subsampling.

Given the many unbounded data intervals, one could think of somehow penalizing the coefficients corresponding to these intervals instead of leaving them unpenalized. This is possible within our framework and figure 27 shows some experiments with different data penalization, again for the parrots image of figure 23. We tested four different situations: Leaving both the bounded and unbounded coefficients unpenalized within their bounds, fixing the unbounded coefficients to zero and letting the others unpenalized within their bounds, fixing the bounded coefficients to the midpoints of the data intervals and allowing the others to vary unpenalized and also penalizing the unbounded coefficients with an L^1 norm and again letting the others unpenalized. The resulting decompressions are shown in figure 27, with a plot of the TGV value and the L^1 norm of the unbounded coefficients for all four situations on the bottom. For the L^1 bounded situation, we used $\lambda = 0.5$ as regularization parameter. As one can see, penalization of the unbounded coefficients does not have a high influence on visual reconstruction quality, while, as could be expected, fixing the bounded coefficients decreases reconstruction quality since the wavelet artifacts cannot be completely removed. In terms of PSNR error, the situation where both types of coefficients can vary unpenalized within their bounds results in the best reconstruction quality.

Extension to combined decompression and zooming

As for the application to JPEG decompression, we can also extend the TGV based JPEG 2000 framework to combined decompression and zooming. Indeed, since the modeling of subsection 5.3.1 and the practical implementation of subsection 5.3.3 already allows all data intervals corresponding to the wavelet basis function ψ to be unbounded, the generalization is straightforward:

Given a JPEG 2000 compressed image, the resulting set of data intervals, denoted by

$$(J_{i,j}^c)_{\substack{c\in\{1,2,3\},\\ 0\leq i,j<N}}$$

$N \in \mathbb{N}$, can be extended by intervals containing all of $\mathbb{R}$ up to a size $2^f N$, denoted by

$$(\tilde{J}_{i,j}^c)_{\substack{c\in\{1,2,3\}\\ 0\leq i,j<2^f N}}.$$

Defining now the set of possible source images as

$$U_D = \{u \in \mathbb{R}^{2^f N\times 2^f N\times 3} \,|\, (Wu)_{i,j}^c \in \tilde{J}_{i,j}^c\},$$

where the decomposition level of the wavelet transform operator W is $f + d$, with d the decomposition level of for the original, JPEG 2000 compressed file, the solution of

$$\min_u \mathrm{TGV}(u) + \mathcal{I}_{U_D}(u)$$

yields an improved decompression of the JPEG 2000 compressed image whose resolution has been increased by a factor 2^f. As result of the generality in the modeling of JPEG 2000 decompression, all results such as existence of a solution and the optimality conditions apply and the same primal dual algorithm as shown in 4 can be used to obtain a solution. In particular, the same considerations with respect to stepsize apply and we can use the adaptive stepsize choice as in (156). Figure 28 compares the results of combined decompression and zooming by a factor 8 of a JPEG 2000 compressed version of the hand image to standard techniques. The wavelet upsampled image was obtained as standard decompression for the extended data set, i.e., all unbounded coefficients have been set to zero.

5.4 Wavelet-based zooming

Apart from the data decompression models of subsections 5.2 and 5.3, we now consider the problem of obtaining a high resolution image from uncompressed, low resolution data. The following method is also presented in [15]. Image zooming can be seen as an inverse problem, where the objective is the inversion

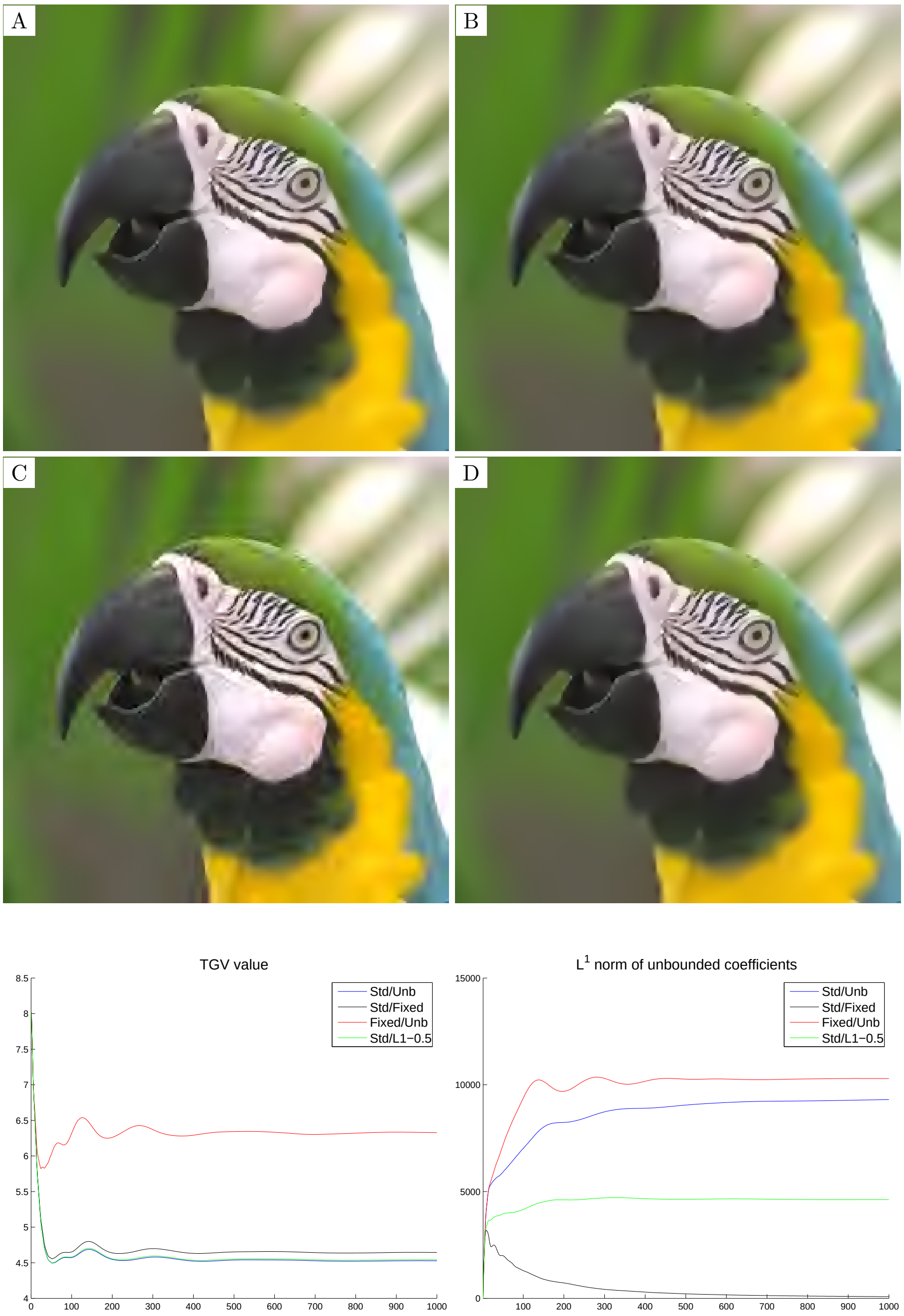


Figure 27: Reconstructions of the JPEG 2000 compressed parrot image without tiling of figure 23 for different data penalization types at 5000 iterations. A: No additional penalization. B: Fixing the unbounded coefficients to zero. C: Fixing the bounded coefficients to the midpoints of the intervals. D: Using 0.5 times the L^1 norm of the unbounded coefficients as penalization.

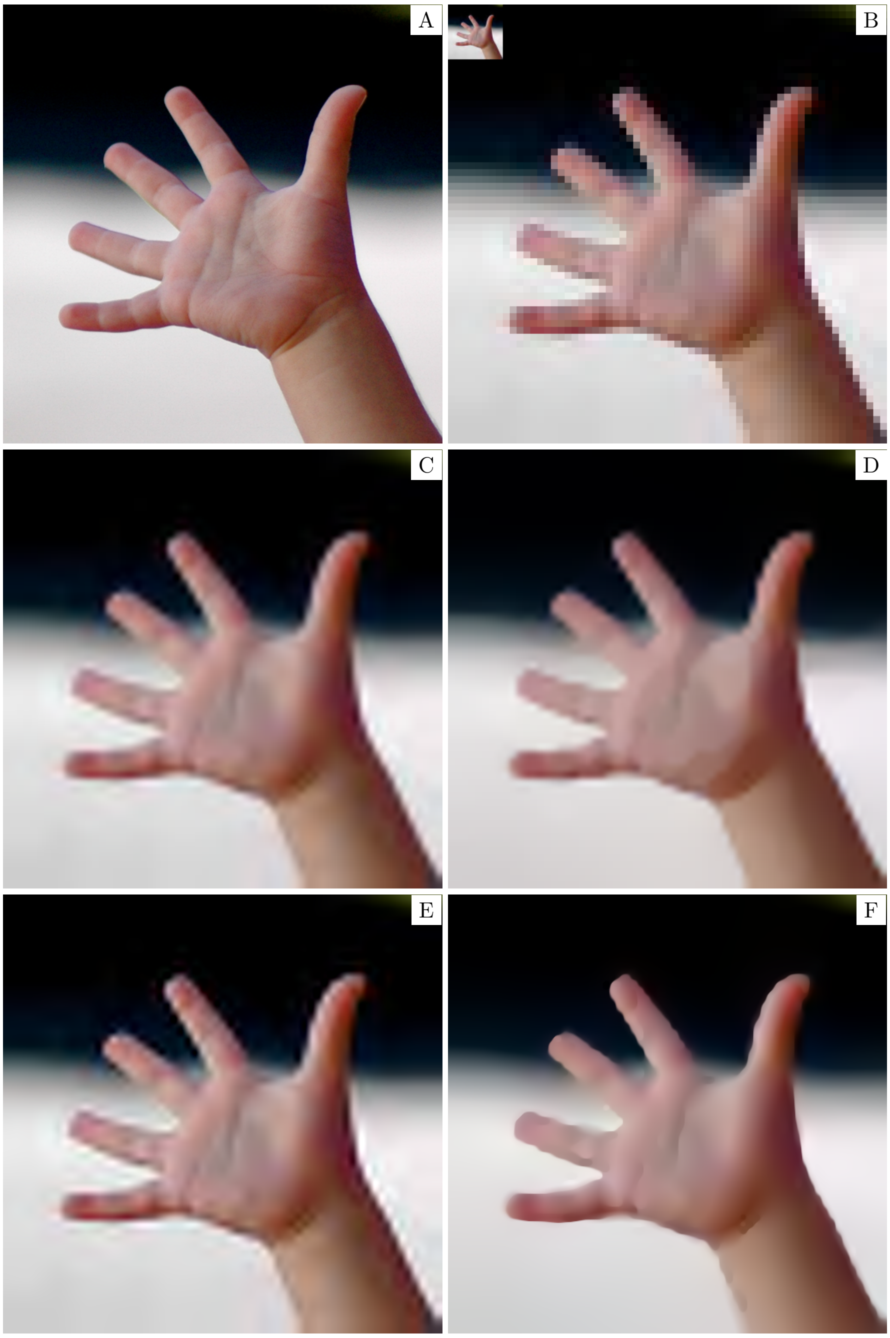


Figure 28: A: Original-sized, uncompressed image (512×512 pixels). B: Downsampled, JPEG2000 compressed image (64×64 pixels, 2.96 bpp) together 8 times magnification by pixel repetition. C: 8 times magnification by cubic interpolation. D: 8 times magnification of a TGV based decompression by cubic interpolation. E: Wavelet based upsampling F: Simultaneous TGV based decompression and factor 8 zooming. All TGV based decompressions were obtained with the normalized primal dual gap being below 10^{-1} as stopping rule. Image by [47], licensed under CC-BY-2.0 (http://creativecommons.org/licenses/by/2.0/).

of a downsampling operator denoted by A. The problem is ill-posed since the kernel of A is large. To get a method of stable inversion, and hence obtain an image zooming method, we aim to again use a TGV regularized model fitting to the general problem setting of section 4. For that purpose, we first have to make an appropriate choice of how to define the downsampling operator A, i.e. the process of obtaining discrete pixel values from an image u defined, for instance, on the unit square.

A straightforward approach, that can also be motivated by data acquisition in digital cameras, is to consider averaging over each pixel as downsampling procedure. A right inverse of the resulting downsampling operator is then an upsampling operator using pixel repetition. Though this downsampling method makes sense from the physical perspective, both downsampling by averaging and upsampling by pixel repetition do not yield satisfactory results when applied to discrete images.

An alternative to obtain a suitable downsampling procedure would be to use a left inverse of a linear interpolation procedure, a technique that is widely applied for image zooming and known to result in visually more appealing images.

The multiresolution approach of wavelet bases provides a framework to describe downsampling procedures: In a simple, one dimensional setting, given orthogonal scaling and wavelet functions $(\phi_{j,k})_{j,k\in\mathbb{Z}}$ and $(\psi_{j,k})_{j,k\in\mathbb{Z}}$, respectively, remember that any signal $u \in L^2(\mathbb{R})$ can be fully described by the L^2- inner products

$$(u, \phi_{R,k})_{L^2}, \qquad (u, \psi_{j,k})_{L^2}, \qquad \text{for } j,k \in \mathbb{Z}, j \leq R,$$

and any $R \in \mathbb{Z}$ fixed. The inner products $((u, \phi_{R,k})_{L^2})_{k\in\mathbb{Z}}$ are interpreted to be the values of the signal u at resolution R, while the inner products $((u, \psi_{j,k})_{L^2})_{j\leq R,k\in\mathbb{Z}}$ contain all remaining detail information. Thus, the mapping that asserts to any given signal u the inner products $((u, \phi_{R,k})_{L^2})_{k\in\mathbb{Z}}$ can be seen as subsampling operation to a resolution R. Since this multiresolution framework can be considered for any choice of wavelet basis, even for non-orthogonal Riesz bases, it allows a general approach to a downsampling operator for the zooming problem. Given the multiresolution framework of any wavelet basis and fixed a resolution level $R \in \mathbb{Z}$, we consider the linear operator

$$u \mapsto ((u, \phi_{R,k})_{L^2})_{k\in\mathbb{Z}}$$

as downsampling operator. As we will see, this approach includes the setting where downsampling is modeled as left inverse of a pixel repetition- or linear interpolation operator.

If we now assume a function, or image, u_0 to be given at the scale R, i.e. the coefficients $((u, \phi_{R,j})_{L^2})_{j\in\mathbb{Z}}$ are known, in order to obtain its high resolution reconstruction we need to determine $((u, \psi_{R,j})_{L^2})_{k\leq R,j\in\mathbb{Z}}$.

As we will see in subsection 5.4.1, this can be done by using the TGV related image model and solving an optimization problem very similar to the one for JPEG 2000 decompression.

The assumptions for this minimization problem are again captured by our general model assumption and also the framework for its solution in a discrete setting is already available. In the numerical experiments section we will compare this zooming method to standard zooming methods based on linear filters, for different choices of wavelet bases.

The idea of image zooming by interpolating wavelet coefficients is not new; we refer to [46] and the references therein for an overview. However, the crucial point of these approaches is how to obtain the missing detail coefficients. In contrast to the methods in [46], we propose to use a variational technique, in particular TGV regularization, to resolve this issue.

Not modeling subsampling by a wavelet transform, variational methods, using the TV functional as regularization term, have already been used in [50, 24] for image zooming.

5.4.1 Modeling

As already mentioned, the modeling for the problem of high resolution image reconstruction is very similar to the modeling for JPEG 2000 decompression: Set $\Omega = (0,1) \times (0,1)$,

$$(\Phi^h_{R,\mathbf{k}})_{\substack{j\leq 1\\ \mathbf{k}\in\mathbf{M}_j}}, \quad (\Psi^h_{j,\mathbf{k}})_{\substack{j\leq 1\\ \mathbf{k}\in\mathbf{L}^h_j}}, \quad (\Psi^v_{j,\mathbf{k}})_{\substack{j\leq 1\\ \mathbf{k}\in\mathbf{L}^v_j}}, \quad (\Psi^d_{j,\mathbf{k}})_{\substack{j\leq 1\\ \mathbf{k}\in\mathbf{L}^d_j}} \tag{177}$$

to be given scaling and wavelet functions contained in $L^2(\Omega)$ and

$$(\tilde{\Phi}^h_{R,\mathbf{k}})_{\substack{j\leq 1\\ \mathbf{k}\in\mathbf{M}_R}}, \quad (\tilde{\Psi}^h_{j,\mathbf{k}})_{\substack{j\leq 1\\ \mathbf{k}\in\mathbf{L}^h_j}}, \quad (\tilde{\Psi}^v_{j,\mathbf{k}})_{\substack{j\leq 1\\ \mathbf{k}\in\mathbf{L}^v_j}}, \quad (\tilde{\Psi}^d_{j,\mathbf{k}})_{\substack{j\leq 1\\ \mathbf{k}\in\mathbf{L}^d_j}} \tag{178}$$

to be their dual functions. Remember that the $\mathbf{M}_j$ and $\mathbf{L}^h_j, \mathbf{L}^v_j, \mathbf{L}^d_j$ denote finite index sets in $\mathbb{N}^2$. Note also that for the wavelet based zooming model, for simplicity, we consider only grayscale images while the generalization to color images can be done exactly as for the JPEG 2000 decompression model of subsection 5.3.

For a fixed resolution level $R \in \mathbb{Z}$, the functions in (177) and (178) constitute biorthogonal Riesz bases of $L^2(\Omega)$ and we can assume a low resolution image $u_0 \in \operatorname{span}\{\tilde{\phi}_{R,j} | j \in \mathbb{Z}\}$ to be given by

$$((u_0, \Phi_{R,\mathbf{k}})_{L^2})_{\mathbf{k}\in\mathbf{M}_R},$$

(see figure 29 for a visualization).

Reconstructing a high resolution image from this given data, that minimizes the $\mathrm{TGV}^{\mathrm{k}}_\alpha$ functional, for $k \in \mathbb{N}$, $\alpha \in \mathbb{R}^k_{>0}$, amounts solving

$$\min_{u\in L^2(\Omega)} \mathrm{TGV}^{\mathrm{k}}_\alpha(u) + \mathcal{I}_{U_D}(u),$$

where

$$U_D = \{u \in L^2(\Omega) \,|\, (u, \Phi_{R,\mathbf{k}})_{L^2} = (u_0, \Phi_{R,k})_{L^2} \;\forall \mathbf{k} \in \mathbf{M}_R\}.$$

Defining data intervals $(J^0_{R,\mathbf{k}})$ and $J^h_{j,\mathbf{k}}, J^v_{j,\mathbf{k}}, J^d_{j,\mathbf{k}}$ corresponding to $\Phi_{R,\mathbf{k}}$ and $\Psi^h_{j,\mathbf{k},j}, \Psi^v_{j,\mathbf{k},j}, \Psi^d_{j,\mathbf{k},j}$ by

$$J^0_{R,\mathbf{k}} = \{(u_0, \Phi_{R,\mathbf{k}})_{L^2}\} \quad \text{and} \quad J^r_{j\mathbf{k}} = \mathbb{R}, \quad r \in \{h, v, d\},$$

suitably summing up these intervals by $(J_n)_{n\in\mathbb{N}}$, and setting $W : L^2(\Omega) \to \ell^2$ to be the basis transformation operator for the primal Riesz basis in (177), U_D can equivalently be written as

$$U_D = \{u \in L^2(\Omega) \,|\, (Wu)_n \in J_n \;\forall n \in \mathbb{N}\}$$

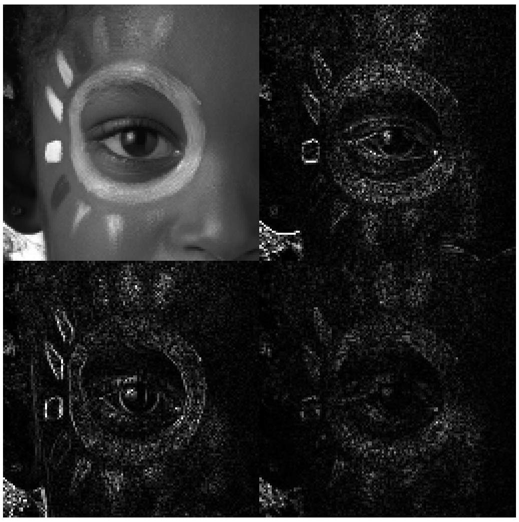


Figure 29: Visualization of the low resolution image and the detail coefficients obtained from a high resolution image with one level of wavelet decomposition. Note that the wavelet coefficients have been rescaled for better visibility.

and with that fits into the general problem assumption (A). Since all intervals corresponding to $\Phi_{R,\mathbf{k}}$ are point intervals, and hence in particular bounded, the existence assumption (EX_{k}) can be verified exactly as in proposition 5.12 by using the same support assumption as written there. Thus, all results of section 4, in particular existence of a solution and the optimality condition, apply.

Choice of scaling and wavelet functions

Note that, since no longer determined by the compression standard, we have now free choice of a scaling and wavelet basis. For a simple interpretation of a resulting subsampling operator, we will, for the rest of this subsection, go back to the unconstrained, one dimensional setting. Assuming a discrete signal to be given by $((u, \phi_{R,k})_2)_{k\in\mathbb{Z}}$, for fixed $R \in \mathbb{Z}$, i.e.

$$u \in \operatorname{span}\{\tilde{\phi}_{R,k} \,|\, k \in \mathbb{Z}\} \subset L^2(\mathbb{R}),$$

its projection onto the smaller, low resolution subspace $\operatorname{span}\{\tilde{\phi}_{R+1,k} \,|\, k \in \mathbb{Z}\}$ is described by

$$(u, \phi_{R+1,k}) = \sum_{l\in\mathbb{Z}} h_l (u, \phi_{R,l+2k}), \quad k \in \mathbb{N}, \tag{179}$$

(cf. (93)), i.e. linear filtering followed by subsampling, where the filters can be constructed from ϕ. Similar, obtaining a higher resolution representation from low resolution data amounts setting

$$(u, \phi_{R,m}) = \sum_{k\in\mathbb{Z}} \left[\tilde{h}_{m-2k}(u, \phi_{R+1,k}) + (-1)^{m-2k} \tilde{h}_{1-(m-2k)}(u, \psi_{R+1,k}) \right],$$

i.e. upsampling followed by linear filtering, where again the filters can be constructed from ϕ. Not knowing the coefficients $(u, \psi_{R+1,k})$, a straightforward upsampling can be obtained by assuming them to be zero, thus

$$(u, \phi_{R,m}) \approx \sum_{k\in\mathbb{Z}} \tilde{h}_{m-2k}(u, \phi_{R+1,k}). \tag{180}$$

We will now interpret this expressions for different choices of scaling functions.

Haar wavelet: A first, intuitive choice of one dimensional scaling function, from which the two dimensional scaling and wavelet functions can be obtained, would be to define

$$\tilde{\phi}(x) = \begin{cases} 1 & 0 \leq x < 1, \\ 0 & \text{else.} \end{cases}$$

This yields the well known Haar wavelet (cf. [32, Section 6.A]), and the filters associated with ϕ and $\tilde{\phi}$ are given by

$$2^{-1/2}h_0 = 2^{-1/2}\tilde{h}_0 = \frac{1}{2}, \quad 2^{-1/2}h_1 = 2^{-1/2}\tilde{h}_1 = \frac{1}{2}.$$

Thus, the down- and upsampling as in equations (179),(180) is given by

$$2^{-1/2}(u, \phi_{R+1,k}) = \frac{1}{2}\left[(u, \phi_{R,2k}) + (u, \phi_{R,1+2k})\right]$$

and

$$2^{-1/2}(u, \phi_{R,2l}) = \frac{1}{2}(u, \phi_{R+1,l}), \quad 2^{-1/2}(u, \phi_{R,2l+1}) = \frac{1}{2}(u, \phi_{R+1,l}).$$

This corresponds to downsampling by averaging and upsampling by pixel repetition. Note that, in contrast to the Le Gall and CDF wavelet, the Haar wavelet basis is orthogonal. Further, due to the specific form of the wavelet and the scaling function, no boundary extension has to be considered when restricting the resulting basis to $(0, 1)$. Indeed, for all resolution levels less or equal to 0 the support of each basis element is either contained in $(0, 1)$ or $\mathbb{R} \setminus (0, 1)$. For resolution levels above 0 the basis functions are constant on the whole $(0, 1)$.

Le Gall wavelet: Another choice is to define

$$\tilde{\phi}(x) = \begin{cases} 1 + x & -1 \leq x < 0 \\ 1 - x & 0 \leq x \leq 1 \\ 0 & \text{else,} \end{cases}$$

i.e. a piecewise linear scaling function. This yields the LeGall wavelet used for lossless coding in JPEG 2000 compression (cf. [32, Section 6.A]), and the filters associated with ϕ and $\tilde{\phi}$ are given by

$$2^{-1/2}\tilde{h}_0 = \frac{1}{2}, \quad 2^{-1/2}\tilde{h}_{\pm 1} = \frac{1}{4}$$

and

$$2^{-1/2}h_0 = \frac{3}{4}, \quad 2^{-1/2}h_{\pm 1} = \frac{1}{4}, \quad 2^{-1/2}h_{\pm 2} = -\frac{1}{8}.$$

The down- and upsampling as in equations (179),(180) can then be given by

$$2^{-1/2}(u, \phi_{R+1,k}) = \frac{3}{4}(u, \phi_{R,2k}) + \frac{1}{4}\sum_{l=\pm 1}(u, \phi_{R,2k+l}) - \frac{1}{8}\sum_{l=\pm 2}(u, \phi_{R,2k+l})$$

and

$$2^{-1/2}(u, \phi_{R,2l}) = \frac{1}{2}(u, \phi_{R+1,l}), \quad 2^{-1/2}(u, \phi_{R,2l+1}) = \frac{1}{4}(u, \phi_{R+1,l-1}) + \frac{1}{4}(u, \phi_{R+1,l}).$$

This corresponds to upsampling by linear interpolation.

CDF 9/7 wavelet: At last we can again use the CDF 9/7 wavelets as described already in subsection 5.3.1, whose filters can be found in [32, Table 6.2]. Again, the upsampling process can be seen as linear filtering, but we do not have an direct interpretation.

5.4.2 Discrete Setting

For the discrete setting, we define $U = \mathbb{R}^{N \times N}$, $N \in \mathbb{N}$, to be the space of discrete, high resolution images, equipped with $\|\cdot\|_U$ as in (55). We proceed as in subsection 5.3.2: Given scaling functions $(\Phi_{j,\mathbf{k}})_{j,\mathbf{k}}$, we assume that the pixels of any discrete image $u \in U$ can be described by the coefficients

$$(u, \Phi_{0,\mathbf{k}})_{L^2}, \quad 0 \leq \mathbf{k} < N,$$

where the inequalities are again meant component wise. A low resolution image $\tilde{v}_0 \in \tilde{U} := \mathbb{R}^{(2^{-R}N)\times(2^{-R}N)}$ can then be obtained from $v_0 \in U$ by applying the wavelet transform operator $W : U \to U$, as defined in (134), with decomposition level $R \in \mathbb{N}$, and taking

$$(Wv_0)_{\mathbf{k}} = (v_0, \Phi_{R,\mathbf{k}})_{L^2}, \quad 0 \leq \mathbf{k} < 2^{-R}N$$

to be its pixel values. The other way around, assuming $\tilde{v}_0$ to be given, one aims to find $v_0 \in U$ such that

$$(Wv_0)_{\mathbf{k}} = (\tilde{v}_0)_{\mathbf{k}},\ 0 \leq \mathbf{k} < 2^{-R}N,$$

i.e. an image $v_0 \in U$ such that v_0 yields $\tilde{v}_0$ when subsampled using the wavelet transform.

Thus, given a discrete image $u_0 \in \tilde{U} = \mathbb{R}^{(2^{-R}N)\times(2^{-R}N)}$ with $R \in \mathbb{N}$, the wavelet transform operator W corresponding to scaling functions $(\Phi_{j,\mathbf{k}})_{j,\mathbf{k}}$, the resulting wavelet and dual functions, and the discrete version of TGV_α^2 as in (57), we can write the discrete minimization problem for wavelet based zooming as

$$\min_{u \in U} \mathrm{TGV}_\alpha^2(u) + \mathcal{I}_{U_D}(u), \tag{181}$$

where

$$\begin{aligned} U_D = & \quad \{u \in U \,|\, (Wu)_{\mathbf{k}} = (u_0)_{\mathbf{k}}, \text{ for all } 0 \leq \mathbf{k} < 2^{-R}N\} \\ = & \qquad\quad \{u \in U \,|\, (Wu)_{\mathbf{k}} \in J_{\mathbf{k}}, \text{ for all } 0 \leq \mathbf{k} < N\}, \end{aligned} \tag{182}$$

with $J_{\mathbf{k}} = (u_0)_{\mathbf{k}}$ for $0 \leq \mathbf{k} < 2^{-R}N$ and $J_{\mathbf{k}} = \mathbb{R}$ else.

Table 5: Parameter setting for wavelet-based zooming

Algorithm 4		**TGV** as in (57)		**Gap** as in (184)	
σ	adaptive, initialization 1/3	$\frac{\alpha_0}{\alpha_1}$	4	γ	1.001
τ	adaptive, initialization 1/3			p	2

Again, since the assumptions we took for the discrete minimization problem for JPEG 2000 decompression did not exclude the case of point intervals, all results, such as existence of a solution and the optimality condition, apply. Consequently, also the saddle point problem

$$\min_{x\in X}\max_{z\in Z}(\mathbf{K}x,z)-\mathbf{F}^*(z), \tag{183}$$

with X, Z, K, F defined as in subsection 5.3.2 , where U_D is now given as in (182), is equivalent to (181) and can be solved by the primal dual algorithm 4. Therefore, since we now have only point intervals or intervals containing all of $\mathbb{R}$, the operation $\operatorname{assign}_{J,N}(w)$ reduces to

$$\operatorname{assign}_J(w)_{i,j}=\begin{cases} w_{i,j}-\sigma\lambda_{i,j} & \text{if } 0\leq i,j<2^{-R}N\\ 0 & \text{else,}\end{cases}$$

where $\lambda_{i,j}=(u_0,\Phi_{R,i,j})_{L^2}$. Also the considerations concerning the norm of $\mathbf{K}$ and with that the stepsizes σ, τ apply directly to the wavelet based zooming model. Finally, also the estimations on the primal dual gap $\mathcal{G}$ defined in (173) are valid, but $\mathcal{G}$ simplifies to

$$\mathcal{G}(x_n,z_n)=F(Kx_n)+I_{D_{C_n}}(Wu_n)+T_n\|\operatorname{div}\tilde{p}_n-W^*w_n\|_{p'}+\sum_{i,j=0}^{2^{-R}N}\lambda_{i,j}w_{i,j}. \tag{184}$$

5.4.3 Numerical experiments

Again, we now evaluate and compare numerical results obtained with the TGV based wavelet zooming algorithm. We will see that the method performs well and leads to highly improved results compared to standard zooming methods, such as, for example, bilinear or bicubic interpolation.

Based on the discussion in subsection 5.3.4, we use the adaptive stepsizes. The parameters for the wavelet based zooming algorithm are summarized in table 5. Note in particular, that we now fix the ratio between α_0 and α_1 for evaluation of the TGV functional to 4 rather that $\sqrt{2}$ since, as we experienced, this choice improves reconstruction quality in zooming applications.

As stopping rule we again used the normalized primal dual gap below 10^{-1} for all experiments. We tested three different wavelets, the Haar, Le Gall and CDF 9/7 wavelet as described in subsection 5.4. Note that, since, in contrast to the Le Gall and CDF 9/7 wavelet, the Haar wavelet leads to an orthogonal basis we can directly project on the given data set. Thus for this wavelet we used the algorithm similar as for JPEG decompression, where no additional variable has been introduced and a fit to data is ensured in each iteration step.

We first consider a four times magnification of a patch of the Barbara image, containing a stripe structure. For better comparability, we used the original image rather than a downsampled version. Thus the downsampling procedure is not known and cannot favor any particular method, but also no original data is available. The results, using the three different wavelet types as well as bilinear-, bicubic- and interpolation with a Lancos 2 filter are shown in figure 30. As one can see, the linear filter based zooming leads to blurring of the stripes while our method yields a reconstruction appearing much sharper. Using the CDF 9/7 wavelets results in the best reconstruction quality. In particular, we observe that not only the edges are preserved, but also the geometrical information is extended in a natural manner for the CDF 9/7 wavelet (as opposed to the Haar wavelet, where "geometrical staircasing" occurs).

In figure 31 we then show the four times magnification of a test image. Again the subsampling method is unknown. In the top left the straightforward upsampling by pixel repetition is shown, while the top right image shows again zooming by cubic interpolation. The lower images show the results with our method and the Haar and CDF 9/7 wavelet, respectively. As one can see, usage of both wavelets results in a sharper reconstruction than cubic interpolation. Using the Haar wavelet conserves the blocky structure of the image and leads to sharp edges, while homogeneous regions are smoothed. The CDF 9/7 wavelet based reconstruction comes closer to a realistic reconstruction: It smooths homogeneous regions nicely and, in opposition to the Haar wavelet, again emphasizes also non-horizontal and non-vertical edges. As minor drawback some edges are slightly less sharper than with the Haar wavelet.

At last, figure 32 compares the TGV based zooming method for the Haar and CDF 9/7 wavelet in the situation where the subsampling process is known and fits to the model assumption, i.e. the subsampling was done by applying a wavelet decomposition on the original image and throwing away the high resolution detail coefficients. On the left of the figure, we show the subsampled version of the image and its upsampling by setting the unknown coefficients to zero. This is the initial image for our TGV based method. On the right, we show the outcome of our method as the primal dual gap is below 10^{-1}. This allows to compare the effect of TGV regularization independent of the wavelet basis. As one can see, indeed the reconstruction quality is clearly improved when TGV based regularization is applied. This justifies the application of TGV regularization instead of simple wavelet based upsampling.

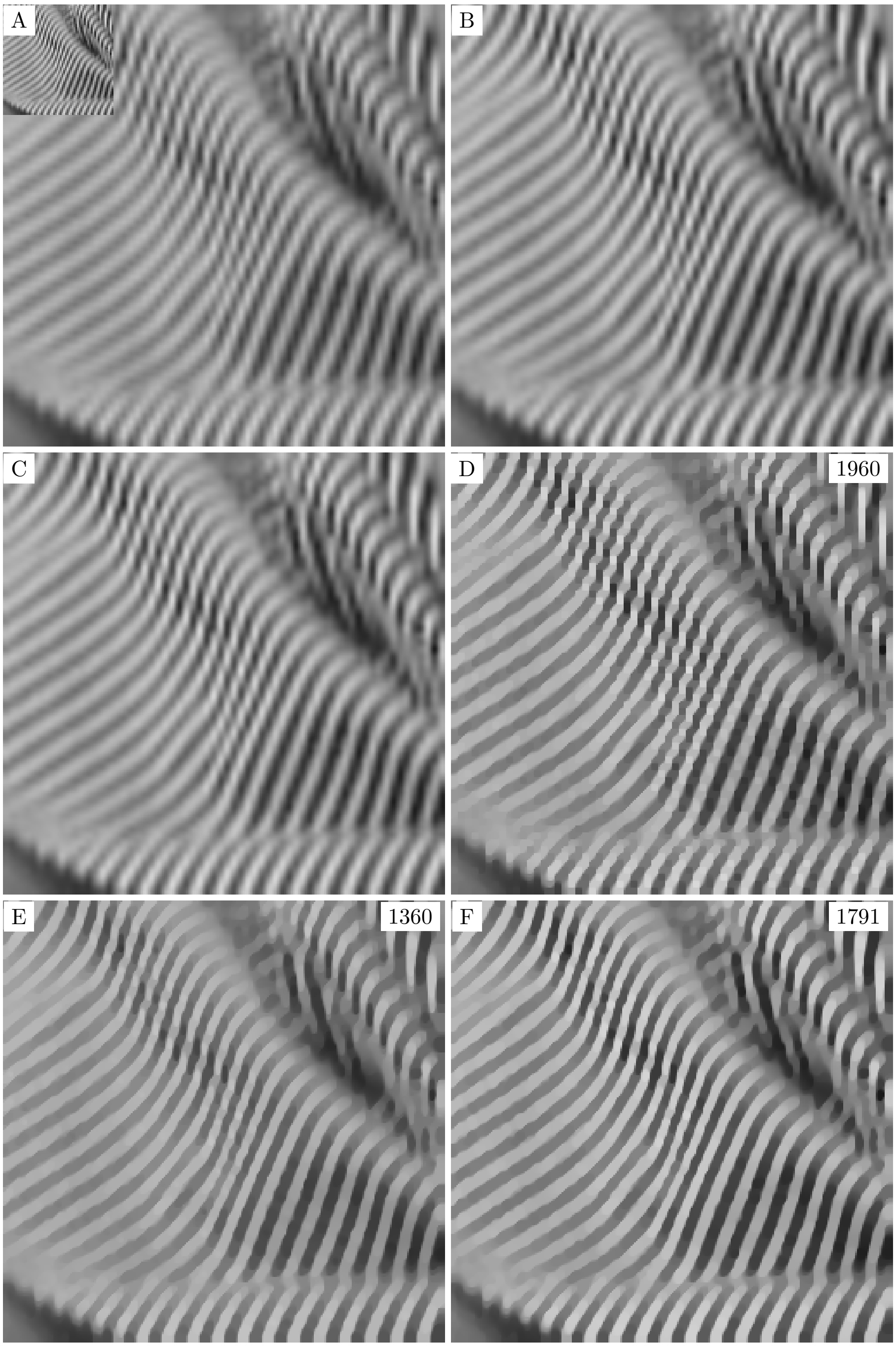


Figure 30: A-C: 4 times magnification by linear filtering with bilinear-, bicubic and Lancos 2 filter. D-F: 4 times magnification by TGV based wavelet zooming using the Haar, Le Gall and CDF 9/7 wavelet, with iteration number on top right. The stopping rule was $\overline{\mathcal{G}}(x_n, y_n) < 10^{-1}$.

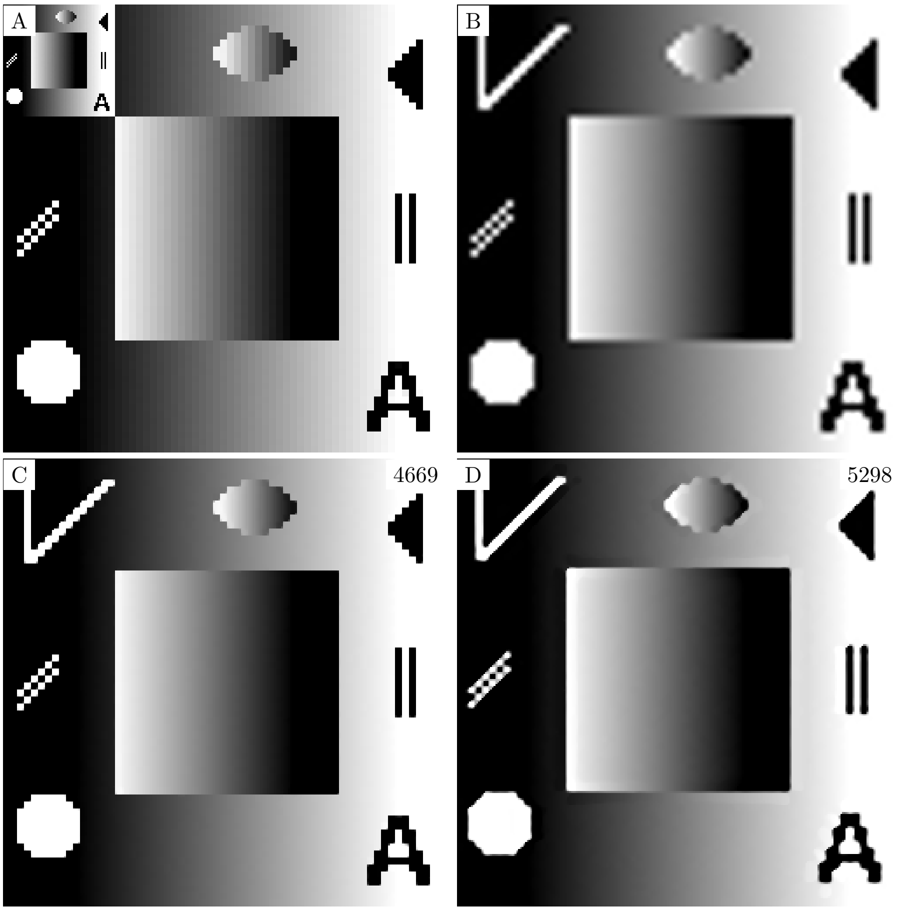


Figure 31: A-B: 4 times magnification by linear filtering with box and bicubic filter. C-D: 4 times magnification by TGV based wavelet zooming using the Haar and CDF 9/7 wavelet, with iteration number on top right. The stopping rule was $\overline{\mathcal{G}}(x_n, y_n) < 10^{-1}$.

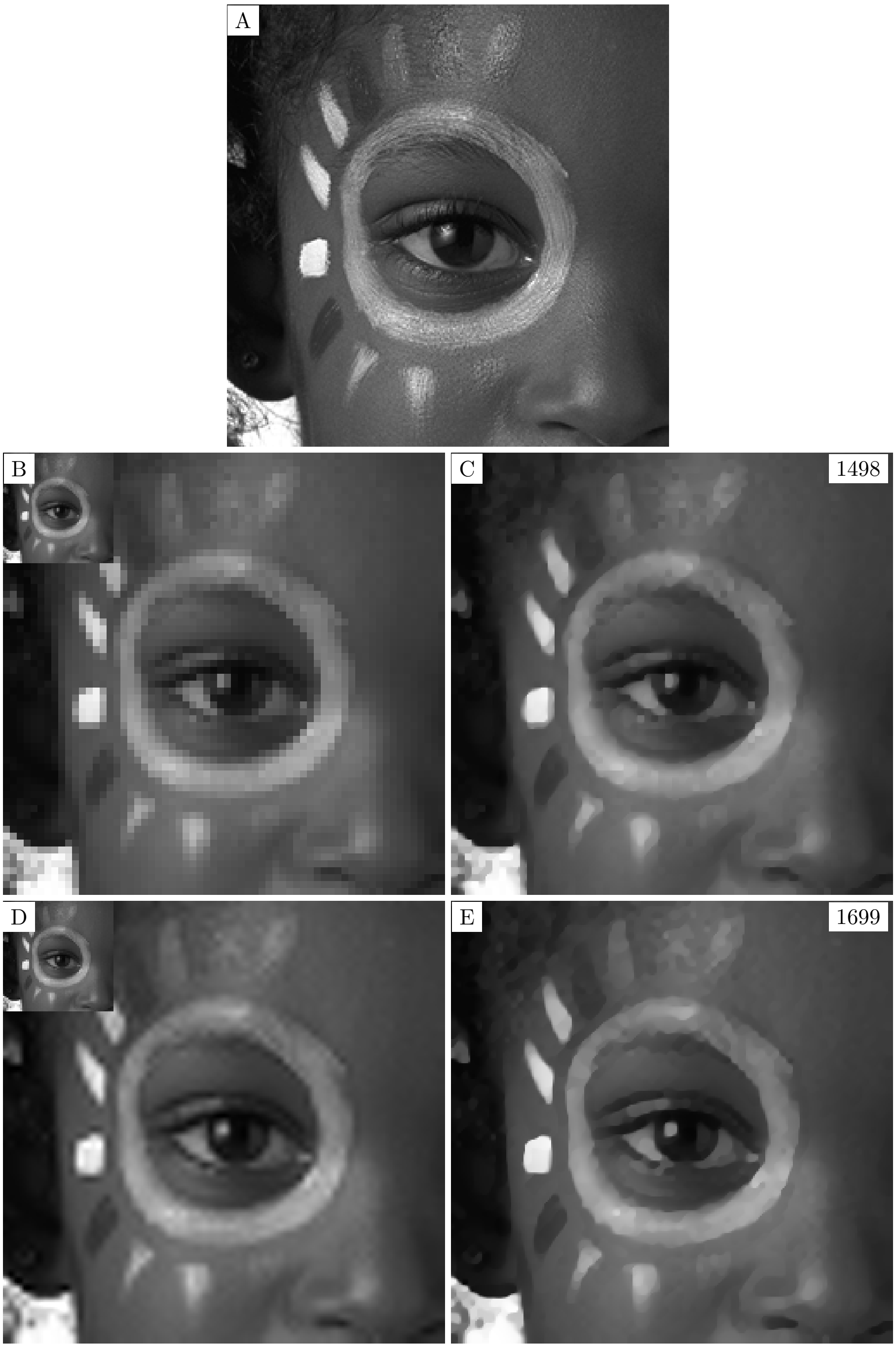


Figure 32: A: Original girls-eye image (256×256 pixels). B: Wavelet upsampled image (without TGV regularization) and downsampled version in top left, up- and downsampling with Haar wavelet. C: TGV based wavelet upsampling using Haar wavelet. D: Wavelet upsampled image (without TGV regularization) and downsampled version in top left, up- and downsampling with CDF 9/7 wavelet. D: TGV based wavelet upsampling using CDF 9/7 wavelet.

References

[1] R. Acar and C. R. Vogel. Analysis of bounded variation penalty methods for ill-posed problems. *Inverse Problems*, 10:1217–1229, 1994.

[2] F. Alter, S. Durand, and J. Froment. Adapted total variation for artifact free decompression of JPEG images. *Journal of Mathematical Imaging and Vision*, 23:199–211, 2005.

[3] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems.* Oxford University Press, 2000.

[4] H. Attouch and H. Brezis. Duality for the sum of convex functions in general Banach spaces. *Aspects of Mathematics and its Applications*, 34:125–133, 1986.

[5] L. Atzori. JPEG2000-coded image error concealment exploiting convex sets projections. *Image Processing, IEEE Transactions on*, 14(4):487–498, 2005.

[6] G. Auber and P. Kornprobst. *Mathematical Problems in Image Processing.* Springer, 2006.

[7] R. L. Bishop and S. I. Goldberg. *Tensor Analysis on Manifolds.* Macmillan, New York, 1968.

[8] V. I. Bogachev. *Measure Theory.* Springer, 2007.

[9] K. Bredies. Recovering piecewise smooth multichannel images by minimization of convex functionals with total generalized variation penalty. *Submitted*, 2012. http://math.uni-graz.at/mobis/publications.html.

[10] K. Bredies. Symmetric tensor fields of bounded deformation. *Annali di Matematica Pura ed Applicata*, pages 1–37, 2012.

[11] K. Bredies and M. Holler. TGV_α^k regularized inverse problems. To be published at http://math.uni-graz.at/mobis/publications.html.

[12] K. Bredies and M. Holler. A pointwise characterization of the subdifferential of the total variation functional. *Submitted*, 2012. http://math.uni-graz.at/mobis/publications.html.

[13] K. Bredies and M. Holler. A total variation–based JPEG decompression model. *SIAM Journal on Imaging Sciences*, 5(1):366–393, 2012.

[14] K. Bredies and M. Holler. Artifact-free decompression and zooming of JPEG compressed images with total generalized variation. In *Computer Vision, Imaging and Computer Graphics. Theory and Application*, volume 359, pages 242–258. Springer, 2013.

[15] K. Bredies and M. Holler. A TGV regularized wavelet based zooming model. In *Scale Space and Variational Methods in Computer Vision*, volume 7893, pages 149–160. Springer, 2013.

[16] K. Bredies, K. Kunisch, and T. Pock. Total generalized variation. *SIAM Journal on Imaging Sciences*, 3(3):492–526, 2010.

[17] K. Bredies, K. Kunisch, and T. Valkonen. Properties of $L^1 - \mathrm{TGV}^2$: The one-dimensional case. *Journal of Mathematical Analysis and Applications*, 398(1):438–454, 2013.

[18] K. Bredies and D. Lorenz. *Mathematische Bildverarbeitung*. Vieweg+Teubner, 2011.

[19] K. Bredies, T. Pock, and B. Wirth. Convex relaxation of a class of vertex penalizing functionals. *Journal of Mathematical Imaging and Vision*, pages 1–25, 2012.

[20] K. Bredies and T. Valkonen. Inverse problems with second-order total generalized variation constraints. In *Proceedings of SampTA 2011 - 9th International Conference on Sampling Theory and Applications*, Singapore, 2011.

[21] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer, 2008.

[22] E. Casas, K. Kunisch, and C. Pola. Regularization by functions of bounded variation and application to image enhancement. *Applied Mathematics and Optimization*, 40:229–257, 1999.

[23] V. Caselles, A. Chambolle, and M. Novaga. The discontinuity set of solutions of the TV denoising problem and some extensions. *Multiscale Modeling and Simulation*, 6:879–894, 2007.

[24] A. Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20:88–97, 2004.

[25] A. Chambolle and P.-L. Lions. Image recovery via total variation minimization and related problems. *Numerische Mathematik*, 76(2):167–188, 1997.

[26] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40:120–145, 2011.

[27] R. H. Chan, J. Yang, and X. Yuan. Alternating direction method for image inpainting in wavelet domains. *SIAM Journal on Imaging Sciences*, 4(3):807–826, 2010.

[28] T. Chan, A. Marquina, and P. Mulet. Higher order total variation-based image restoration. *SIAM Journal on Scientific Computing*, 22:503–516, 2000.

[29] T. F. Chan and S. Esedoḡlu. Aspects of total variation regularized L^1 function approximation. *SIAM Journal on Applied Mathematics*, 65:1817–1837, 2005.

[30] T. F. Chan, J. Shen, and H.-M. Zhou. Total variation wavelet inpainting. *Journal of Mathematical Imaging and Vision*, 25(1):107–125, 2006.

[31] P. C. Chung. A JPEG 2000 error resilience method using uneven block-sized information included markers. *Circuits and Systems for Video Technology, IEEE Transactions on*, 15(3):420–424, 2005.

[32] A. Cohen, I. Daubechies, and J.-C. Feauveau. Biorthogonal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 45(5):485–560, 1992.

[33] A. Cohen, I. Daubechies, and P. Vial. Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis*, 1(1):54 – 81, 1993.

[34] I. Daubechies. *Ten Lectures on Wavelets.* Number 61 in CBMS-NSF Lecture Notes. SIAM, 1992.

[35] Y. Dong, M. Hintermüller, and M. Neri. An efficient primal-dual method for L^1 TV image restoration. *SIAM Journal on Imaging Science*, 2:1168–1189, 2009.

[36] S. Durand and J. Froment. Reconstruction of wavelet coefficients using total variation minimization. *SIAM Journal on Scientific Computing*, 24(5):1754–1767, 2003.

[37] I. Ekeland and R. Témam. *Convex Analysis and Variational Problems.* SIAM, 1999.

[38] L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions.* CRC Press, 1992.

[39] M. Frigo and S. G. Johnson. The design and implementation of FFTW3. *Proceedings of the IEEE*, 93(2):216–231, 2005. Special issue on "Program Generation, Optimization, and Platform Adaptation".

[40] V. Girault and P.-A. Raviart. *Finite Element Method for Navier-Stokes Equation.* Springer, 1986.

[41] B. Goldluecke, E. Strekalovskiy, and D. Cremers. The natural total variation which arises from geometric measure theory. *SIAM Journal on Imaging Sciences*, 5(2):537–563, 2012.

[42] Joint Bilevel Image Experts Group and Joint Photographic Experts Group. *JPEG 2000 image coding system*, 2000. ISO/IEC 15444-1.

[43] P. R. Halmos. *Measure Theory.* Springer, 1974.

[44] M. Hintermüller and K. Kunisch. Total bounded variation as bilaterally constrained optimization problem. *SIAM Journal on Applied Mathematics*, 64:1311–1333, 2004.

[45] M. Hintermüller and M. M. Rincon-Camacho. Expected absolute value estimators for a spatially adapted regularization parameter choice rule in L^1-TV-based image restoration. *Inverse Problems*, 26(8):085005, 30, 2010.

[46] N. Kaulgud and U. B. Desai. Image zooming: Use of wavelets. In *The International Series in Engineering and Computer Science*, volume 632, pages 21–44. Springer, 2002.

[47] J. Kubina. http://flickr.com/photos/kubina/42275122, 2008.

[48] P. J. Lee. Error concealment algorithm using interested direction for JPEG 2000 image transmission. *Consumer Electronics, IEEE Transactions on*, 49(4):1395–1401, 2003.

[49] A. K. Louis, P. Maass, and A. Rieder. *Wavelets.* John Wiley & Sons, 1997.

[50] F. Malgouyres and F. Guichard. Edge direction preserving image zooming: A mathematical and numerical analysis. *SIAM Journal on Numerical Analysis*, 39:1–37, 2001.

[51] S. Mallat. *A Wavelet Tour of Signal Processing.* Elsevier, 2009.

[52] M. Nikolova. Local strong homogeneity of a regularized estimator. *SIAM Journal on Applied Mathematics*, 61:633–658, 2000.

[53] A. Nosratinia. Enhancement of JPEG-compressed images by re-application of JPEG. *The Journal of VLSI Signal Processing*, 27:69–79, 2001.

[54] A. Nosratinia. Postprocessing of JPEG-2000 images to remove compression artifacts. *Signal Processing Letter, IEEE*, 10(10):296–299, 2003.

[55] NVIDIA. NVIDIA CUDA programming guide 2.0. NVIDIA Cooperation, 2008.

[56] R. Oktem. Regularization-based error concealment in JPEG 2000 coding scheme. *IEEE Signal Processing Letters*, 14(12):956–959, 2001.

[57] OpenMP Architecture Review Board. Openmp application program interface, version 3.1. http://www.openmp.org, 2011.

[58] C. Pöschl and O. Scherzer. Characterization of minimizers of convex regularization functionals. In *Frames and operator theory in analysis and signal processing*, volume 451 of *Contemp. Math.*, pages 219–248. Amer. Math. Soc., Providence, RI, 2008.

[59] W. Ring. Structural properties of solutions to total variation regularization problems. *ESAIM: Mathematical Modelling and Numerical Analysis*, 34:799–810, 2000.

[60] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

[61] V. A. Sharafutdinov. *Integral Geometry of Tensor Fields.* VSP, Utrecht, 1994.

[62] M.-Y. Shen and C.-C. Jay Kuo. Review of postprocessing techniques for compression artifact removal. *Journal of Visual Communication and Image Representation*, 9(1):2–14, 1998.

[63] S. Singh, V. Kumar, and H. K. Verma. Reduction of blocking artifacts in JPEG compressed images. *Digital Signal Processing*, 17(1):225–243, 2007.

[64] A. Skodras, C. Christopoulos, and T. Ebrahimi. The jpeg 2000 still image compression standard. *IEEE Signal processing Magazine*, 18:36–58, 2001.

[65] R. Témam. *Mathematical Problems in Plasticity.* Gauthier-Villars, Paris, 1985.

[66] M. Unser and T. Blu. Mathematical properties of the jpeg2000 wavelet filters. *Image Processing, IEEE Transactions on*, 12:1080–1090, 2003.

[67] T. Valkonen, K. Bredies, and F. Knoll. Total generalized variation in diffusion tensor imaging. *Submitted*, 2012. http://math.uni-graz.at/mobis/publications.html.

[68] G. K. Wallace. The JPEG still picture compression standard. *Communications of the ACM*, 34(4):30–44, 1991.

[69] J. Wei, M. Pickering, M. Frator, J. Arnold, J. Boman, and W. Zeng. Boundary artefact reduction using odd tile length and low pass first convention (OTLPF). *Image Processing, IEEE Transactions on*, 14(8):1033–1042, 2001.

[70] J. Weidmann. *Linear Operators in Hilbert Spaces.* Springer, 1980.

[71] Y.-W. Wen, R. H. Chan, and A. M. Yip. A primal-dual method for total variaton-based wavelet domain inpainting. *Image Processing, IEEE Transactions on*, 21(1):106–114, 2011.

[72] R. M. Young. *An Introduction to Nonharmonic Fourier Series.* Academic Press, 2001.

[73] X. Zhang and T. F. Chan. Wavelet inpainting by nonlocal total variation. *Inverse Problems and Imaging*, 4(1):191–210, 2010.

[74] W. Zhu and T. Chan. Image denoising using mean curvature of image surface. *SIAM Journal on Imaging Sciences*, 5(1):1–32, 2012.

[75] W. P. Ziemer. *Weakly Differentiable Functions.* Springer, 1989.