SpezialForschungsBereich F 32

Karl-Franzens Universität Graz
Technische Universität Graz
Medizinische Universität Graz

Nonconvex TV^q-Models in Image Restoration: Analysis and a Trust-Region Regularization Based Superlinearly Convergent Solver

Michael Hintermüller Tao Wu

SFB-Report No. 2011–026 November 2011

A–8010 GRAZ, HEINRICHSTRASSE 36, AUSTRIA

Supported by the
Austrian Science Fund (FWF)

Der Wissenschaftsfonds.

SFB sponsors:

- **Austrian Science Fund** (FWF)
- **University of Graz**
- **Graz University of Technology**
- **Medical University of Graz**
- **Government of Styria**
- **City of Graz**

Nonconvex TV^q-Models in Image Restoration: Analysis and a Trust-Region Regularization Based Superlinearly Convergent Solver*

Michael Hintermüller† Tao Wu‡

First Draft: November 9, 2011

Abstract

A nonconvex variational model is introduced which contains the ℓ^q-"norm", $q \in (0,1)$, of the gradient of the image to be reconstructed as the regularization term together with a least-squares type data fidelity term which may depend on a possibly spatially dependent weighting parameter. Hence, the regularization term in this functional is a nonconvex compromise between the minimization of the support of the reconstruction and the classical convex total variation model. In the discrete setting, existence of a minimizer is proven, a Newton-type solution algorithm is introduced and its global as well as locally superlinear convergence is established. The potential indefiniteness (or negative definiteness) of the Hessian of the objective during the iteration is handled by a trust-region based regularization scheme. The performance of the new algorithm is studied by means of a series of numerical tests. For the associated infinite dimensional model an existence result based on the weakly lower semicontinuous envelope is established and its relation to the original problem is discussed.

1 Introduction

In many applications of signal and image recovery one is interested in obtaining solutions with the sparsest or smallest support set, either of the signal directly or of a related quantity of interest (such as the gradient of an image for instance), from a limited number of measurements. This topic is at the core of *compressed sensing* (see, e.g., [2, 16, 17, 12] and the references therein) or *basis pursuit* (see, e.g., [8]) and has sparked significant research activities in the recent past. Mathematically, finding the smallest support set of a signal or an image requires to minimize the ℓ_0-norm, i.e. the number of nonzero entries in the solution vector or the related quantity of interest, subject to a constraint reflecting data fidelity. This problem is of combinatorial nature and it is well-known that it is essentially NP-hard [31]. Thus, for practical purposes the ℓ_0-norm minimization problem is usually replaced by a convex relaxation leading to the minimization of the ℓ_1-norm which can be solved efficiently; see the discussion in [12] and, for instance, [38] and the references therein for further algorithmic developments.

*This research was supported by the Austrian Science Fund (FWF) through START project Y305 "Interfaces and Free Boundaries" and through SFB project F3204 "Mathematical Optimization and Applications in Biomedical Sciences".

†Department of Mathematics, Humboldt-University of Berlin, Unter den Linden 6, 10099 Berlin, Germany (hint@math.hu-berlin.de).

‡Institute for Mathematics and Scientific Computing, Karl-Franzens-University of Graz, Heinrichstrasse 36, A-8010 Graz, Austria (tao.wu@edu.uni-graz.at).

In image processing one typically aims at recovering an image from noisy data while still keeping edges in the image. The latter requirement is responsible for the tremendous success of total variation based image restoration [37]. In connection with the sparsity requirement alluded to above, this implies to compute a restoration result with gradient-sparsity, i.e. a piecewise constant image with a small number of patches. Hence, rather than minimizing the support of the image directly, one is interested in minimizing the support of the gradient of the recovered image. In the context of the convex relaxation mentioned above this means to minimize the ℓ_1-norm of the gradient of the image subject to data fidelity.

There is evidence [5, 34] that replacing the ℓ_1-norm by the nonconvex and nondifferentiable function $\|v\|_{\ell_q}^q = \sum_i |v_i|^q$ with $q \in (0,1)$, which for the ease of reference we refer to as ℓ_q-norm in what follows, promotes gradient-sparsity even better. Moreover, the ℓ_q-norm allows possibly a smaller number of measurements than the ℓ_1-norm in compressed sensing. In [32] (see also the more recent paper [34]) it was shown that nonconvex regularization terms in total variation based image restoration yield even better edge preservation when compared to the convex ℓ_1-type regularization. Moreover, it appears that the ℓ_q-norm regularization is also more robust with respect to noise.

Nonconvex and nonsmooth regularization in image restoration (and more generally in inverse problems) poses significant challenges with respect to both, the existence of solutions of associated minimization problems and, in particular, the development of efficient (i.e. locally more than the linearly convergent) solution algorithms. Linearly convergent gradient projection type methods for compressed sensing problems minimizing the ℓ_q-norm can be found in [5]. In [7] the latter solver was replaced by a regularized iteratively reweighted least squares (IRLS) technique. Based on [21], Chartrand extends in [6] the Bregman iteration which relies on a variable splitting approach combined with a q-shrinkage operation to ℓ_q-norm minimization. The resulting method typically has a linear convergence behavior. In [12], the iteratively reweighted least squares solver for compressed sensing with the ℓ_q-norm is shown to converge locally superlinearly. The result depends on a q-null-space-condition, the sparsity of the solution and a locality requirement of the initial guess. A different perspective was taken in [34] where, under certain conditions, more general nonconvex regularization functionals are considered. Concerning the solver development, a technique based on interior point method is proposed. The authors point out the interesting observation that, under the stated conditions, the nonsmooth and nonconvex regularization functional may be decomposed as the sum of a nonconvex but smooth part plus a convex and nonsmooth part. Increasing the variable space and rewriting the problem then yields the minimization of a nonconvex and smooth function subject to linear or affine equality constraints and nonnegativity constraints, which is equivalent to the original problem. The reformulated problem may now be tackled by interior point methods [41], which were very recently shown to compute a local minimizer in compressed sensing in polynomial time [20]. Clearly, the increase of the variable space and the computational effort implied by the interior point methods might be considered as disadvantages. In the follow-up work [33] the interior point solver is replaced by variable splitting techniques resulting in alternating minimization methods which converge linearly. Unfortunately, the conditions required for the success of the algorithms proposed in [34] and [33] rule out the ℓ_q-norm minimization and also the modified version of this problem considered in this paper. We also mention the development of a smoothing nonlinear conjugate gradient solver in [9] which is based on [34].

In this paper we are interested in expanding the scope of solvers for ℓ_q-norm-based regularization of the gradient of the image to be recovered (we refer to this regularization as the TV^q-regularization as it combines the edge preservation property of total variation regularization with the sparsity-promoting ℓ_q-norm). In particular we are interested in locally superlinearly

convergent methods which are robust with respect to noise. In order to achieve this, our proposed method combines a reweighting technique for handling the nonconvexity with primal-dual semismooth Newton methods for image restoration [3, 23, 24], which exhibit a fast (local) convergence. For stabilizing the Newton solver in the presence of indefiniteness due to the involved nonconvexity, a specific regularization scheme is applied which modifies the (generalized) Hessian of the underlying variational problem based on a trust-region technique [11]. The latter technique has the advantage of allowing a transition of the modified (generalized) Hessian to the true Hessian as the solution is approached and, thus, enabling the locally superlinear convergence properties of the underlying Newton iteration. We point out that in contrast to the IRLS solver of [12] we guarantee global convergence. Moreover, locally superlinear convergence is established without requiring conditions like the q-null-space property or sparsity conditions concerning the solution.

The rest of the paper is organized as follows. In Section 2 we introduce our TV^q-model problem and discuss its regularization by a Huber-type function. The primal-dual Newton solver is the subject of Section 3. In this core section of the present paper, we introduce the stabilization of Newton's method (which we call R-regularization) together with the associated trust-region scheme for deciding on the amount of R-regularization required. Furthermore, the overall algorithm is defined and its global as well as locally superlinear convergence is established. Section 4 is devoted to numerical tests showing the efficiency of our new method. Finally, in Section 5 we address the function space setting of the underlying variational problem and discuss the associated difficulties including a warning example.

2 TV^q variational model and its Huberization

We consider the following variational problem

$$\min_{u\in\mathbb{R}^{|\Omega|}} f_0(u) := \sum_{(i,j)\in\Omega} \left(\frac{\mu}{2}|(\nabla u)_{ij}|^2 + \frac{\alpha}{q}|(\nabla u)_{ij}|^q + \frac{\lambda_{ij}}{2}|(Ku-z)_{ij}|^2 \right), \tag{1}$$

where Ω is a two-dimensional index set representing the image domain. By $|\Omega|$ we denote its cardinality. We have $\alpha > 0$, $0 < q < 1$, $0 < \mu \ll \alpha$ as the given model parameters. The matrix $K \in \mathbb{R}^{|\Omega|\times|\Omega|}$ is assumed to not annihilate a constant vector, e.g. K might be a blurring matrix. The vector $z \in \mathbb{R}^{|\Omega|}$ stands for the given noisy data, and $u \in \mathbb{R}^{|\Omega|}$ is the image to be restored. Despite the fact that we refer to $u \in \mathbb{R}^{|\Omega|}$ as a vector, we denote the elements of u by u_{ij} with $(i,j) \in \Omega$. This appears natural as the image domain is given as a two-dimensional array of pixels. Analogously one has to understand the action of the blurring operator (matrix) K. Subsequently we allow situations where the fidelity coefficient $\lambda \in \mathbb{R}^{|\Omega|}$ is possibly spatially dependent (see, e.g., [14, 15]) such that $\lambda_{ij} > 0$ for all $(i,j) \in \Omega$ and $\sum_{(i,j)\in\Omega} \lambda_{ij} = |\Omega|$. The discrete gradient operator ∇ is decomposed as $\nabla = (\nabla_x \ \nabla_y)$ such that $(\nabla u)_{ij} = ((\nabla_x u)_{ij} \ (\nabla_y u)_{ij})$, where $\nabla_x \in \mathbb{R}^{|\Omega|\times|\Omega|}$ is the discrete derivative in x-direction and $\nabla_y \in \mathbb{R}^{|\Omega|\times|\Omega|}$ is the discrete derivative in y-direction, respectively. The Euclidean norm of $(\nabla u)_{ij}$ in $\mathbb{R}^2$ is denoted by $|(\nabla u)_{ij}|$. For elements $\vec{p} \in (\mathbb{R}^{|\Omega|})^2$, $\vec{p}_x$ denotes components corresponding to the x-direction in the above sense and $\vec{p}_y$ components belonging to the y-direction. The discrete Laplacian Δ is defined as $\Delta := -\nabla_x^\top \nabla_x - \nabla_y^\top \nabla_y$. The multiplication of vectors is understood in the pointwise sense, i.e. $(uv)_{ij} = u_{ij}v_{ij}$ for $u, v \in \mathbb{R}^{|\Omega|}$ and $(u\vec{p})_{ij} = (u_{ij}(\vec{p}_x)_{ij} \ u_{ij}(\vec{p}_y)_{ij})$ for $u \in \mathbb{R}^{|\Omega|}$, $\vec{p} \in (\mathbb{R}^{|\Omega|})^2$. Similarly, for $u \in \mathbb{R}^{|\Omega|}$ and $q \in \mathbb{R}$, the power u^q is a vector in $\mathbb{R}^{|\Omega|}$ such that $(u^q)_{ij} = u_{ij}^q$. For $u, v \in \mathbb{R}^{|\Omega|}$ and $\gamma \in \mathbb{R}$, the max-operation is understood in a componentwise sense, i.e. $(\max(u,\gamma))_{ij} = \max(u_{ij}, \gamma)$ and $(\max(u,v))_{ij} = \max(u_{ij}, v_{ij})$. A diagonal matrix

with its diagonal elements given by the vector u is denoted by $D(u)$. The characteristic vector $\chi_{\mathcal{A}}$ of the set $\mathcal{A} \subset \Omega$ is defined as $(\chi_{\mathcal{A}})_{ij} = 1$ if $(i,j) \in \mathcal{A}$ and $(\chi_{\mathcal{A}})_{ij} = 0$ otherwise. The identity vector $\vec{e} \in (\mathbb{R}^{|\Omega|})^2$ is defined as $\vec{e}_{ij} = (1\ 1)$ for all $(i,j) \in \Omega$. We use $\|\cdot\|$ to denote the 2-norm for vectors in $\mathbb{R}^{|\Omega|}$ and the spectral norm for matrices in $\mathbb{R}^{|\Omega|\times|\Omega|}$. The symbols $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ represent the maximal eigenvalue and the minimal eigenvalue of a matrix, respectively. The constant C may take different values at different occasions.

We start our investigations of (1) by establishing the existence of a solution.

Theorem 2.1 (Existence of solution)**.** *Assume that $\mu \geq 0,\ \alpha > 0,\ q > 0,\ \lambda_{ij} > 0$ for all $(i,j) \in \Omega$, and*

$$\mathrm{Ker}\nabla \cap \mathrm{Ker}K = \{0\}. \tag{2}$$

Then there exists a global minimizer for the variational problem (1).

Proof. Since f_0 is bounded from below, it suffices to show that f_0 is coercive, i.e. $|f_0(u^k)| \to +\infty$ whenever $\|u^k\| \to +\infty$ for some sequence (u^k) in $\mathbb{R}^{|\Omega|}$. We prove this by contradiction. For this purpose, assume that $\|u^k\| \to +\infty$ and that there exists $C > 0$ such that $|f_0(u^k)| \leq C$ for all k. Note that the function $s \mapsto |s|^q$ with $q > 0$ is strictly increasing on $\mathbb{R}_+$. Thus we must have $u^k \in \mathrm{Ker}\nabla \cap \mathrm{Ker}K$ and $u^k \neq 0$, for all sufficiently large k. This contradicts our hypothesis (2). □

In order to characterize an optimal solution u, we define the active set $\mathcal{A}_0(u) := \{(i,j) \in \Omega : |(\nabla u)_{ij}| \neq 0\}$ and the inactive set $\mathcal{I}_0(u) := \Omega \backslash \mathcal{A}_0(u)$. Due to the occurrence of the term involving q in (1) with $0 < q < 1$ (which we call the TV^q-term from now on), the objective f_0 (which we refer to as the TV^q-model) is nondifferentiable on $\mathcal{I}_0(u)$. Therefore, the Euler-Lagrange equation for characterizing a stationary point is separately posed on $\mathcal{A}_0(u)$ and on $\mathcal{I}_0(u)$, i.e.

$$\begin{cases} -\mu\Delta u + K^\top\lambda(Ku - z) + \alpha\nabla^\top(|\nabla u|^{q-2}\nabla u) = 0, & \text{if } (i,j) \in \mathcal{A}_0(u); \\ \nabla u = 0, & \text{if } (i,j) \in \mathcal{I}_0(u). \end{cases} \tag{3}$$

Since the objective f_0 is nonconvex, the solution to (3) is in general not unique.

In order to make the problem numerically tractable, we locally smooth the TV^q-term by a *Huber function* φ_γ defined by

$$\varphi_\gamma(s) := \begin{cases} \frac{1}{q}s^q - (\frac{1}{q} - \frac{1}{2})\gamma^q & \text{if } s > \gamma, \\ \frac{1}{2}\gamma^{q-2}s^2 & \text{if } 0 \leq s \leq \gamma. \end{cases}$$

Correspondingly, the *Huberized* variational problem is written as

$$\min_{u\in\mathbb{R}^{|\Omega|}} f_\gamma(u) := \sum_{(i,j)\in\Omega} \left(\frac{\mu}{2}|(\nabla u)_{ij}|^2 + \alpha\varphi_\gamma(|(\nabla u)_{ij}|) + \frac{\lambda_{ij}}{2}|(Ku - z)_{ij}|^2 \right). \tag{4}$$

Note that the Huberized functional f_γ is continuously differentiable, and the Euler-Lagrange equation associated with (4) is given by

$$\nabla f_\gamma(u) = -\mu\Delta u + K^\top\lambda(Ku - z) + \alpha\nabla^\top\left(\max(|\nabla u|, \gamma)^{q-2}\nabla u\right) = 0. \tag{5}$$

The Huber function [25], as a tool of local smoothing, has been successfully applied previously in (convex) nondifferentiable variational models in image processing; see, e.g., [39, 24, 13]. Next we study the behavior of our Huberization of the nonconvex TV^q-model for vanishing Huber parameter, i.e. for $\gamma \to 0^+$.

Theorem 2.2 (Consistency of Huberization)**.** *Let the assumptions of Theorem 2.1 hold true, and* (u^k) *be a sequence such that* $\gamma^k \to 0^+$ *and each* u^k *is a global minimizer of the Huberized functional* f_{γ^k}*. Then there exists an accumulation point* u^* *of* (u^k) *such that* u^* *satisfies the original Euler-Lagrange equation (3).*

Proof. With an analogous argument to the proof of Theorem 2.1, we have the coercivity of all (f_{γ^k}), uniformly with respect to γ^k. Therefore the sequence (u^k) is uniformly bounded. By compactness, there exists a subsequence of (u^k), say $(u^{k'})$, such that $(u^{k'})$ converges to some u^* as $k' \to +\infty$.

Next we show that u^* is a solution to (3). Since each $(u^{k'})$ satisfies the Huberized Euler-Lagrange equation (5), we have

$$-\mu\Delta u^{k'} + K^\top\lambda(Ku^{k'} - z) + \alpha\nabla^\top\left(\max(|\nabla u^{k'}|, \gamma^{k'})^{q-2}\nabla u^{k'}\right) = 0. \tag{6}$$

Let $k' \to +\infty$ so that $\gamma^{k'} \to 0^+$. On the active set $\mathcal{A}_0(u^*)$ where $|\nabla u^*| > 0$, the first argument of the max-function in (6) is taken in the limit, i.e.

$$-\mu\Delta u^* + K^\top\lambda(Ku^* - z) + \alpha\nabla^\top\left(|\nabla u^*|^{q-2}\nabla u^*\right) = 0, \text{ for } (i,j) \in \mathcal{A}_0(u^*).$$

On the inactive set $\mathcal{I}_0(u^*)$, we have $\nabla u^* = 0$ by definition. Thus we conclude that u^* satisfies the Euler-Lagrange equation (3). □

We note that finding global minimizers for nonconvex problems often represents a challenging (if not impossible) task. Therefore, the remainder of this section is devoted to designing and analyzing an algorithm for numerically finding (local) minimizers of (4). For the ease of notation, we shall frequently dismiss the subscript of f_γ without causing ambiguity.

We start by noting that the gradient mapping in (5), i.e. $\nabla f : \mathbb{R}^{|\Omega|} \to \mathbb{R}^{|\Omega|}$, is locally Lipschitz. According to Rademacher's Theorem, ∇f is differentiable almost everywhere. Then the generalized Hessian of f at u [10], denoted by $\partial^2 f(u)$, is defined as the convex hull of $\partial_B^2 f(u)$, where $\partial_B^2 f(u)$ consists of all matrices in $\mathbb{R}^{|\Omega|\times|\Omega|}$ that are limits of sequences of the form $\nabla^2 f(u^k)$ with $u^k \to u$ and ∇f differentiable at all u^k, i.e.

$$\partial_B^2 f(u) := \{\lim \nabla^2 f(u^k) : u^k \to u,\ \nabla f \text{ is differentiable at } u^k\}.$$

Moreover, the gradient mapping $\nabla f : \mathbb{R}^{|\Omega|} \to \mathbb{R}^{|\Omega|}$ is semismooth at any u, i.e.

$$\lim_{\substack{V(u+td') \in \partial^2 f(u+td'),\\ d' \to d,\ t \to 0^+}} V(u+td')d' \text{ exists for all } d \in \mathbb{R}^{|\Omega|};$$

see [36]. Due to Theorem 2.3 in [36], f is directionally differentiable at any u, and for any $V(u+d) \in \partial^2 f(u+d)$,

$$V(u+d)d - \nabla f(u;d) = o(\|d\|),\ \text{as } \|d\| \to 0,$$

where $o(t)/t \to 0$ as $t \to 0^+$, and $\nabla f(u;d)$ denotes the directional derivative of f at u in direction d. Thus, for any $V(u+d) \in \partial_B^2 f(u+d)$ we have

$$\|\nabla f(u+d) - \nabla f(u) - V(u+d)d\| = o(\|d\|),\ \text{as } \|d\| \to 0. \tag{7}$$

In our subsequently defined algorithm, we are in particular interested in the elements of the (possibly) set-valued mapping $\partial_B^2 f$ at u, which can be written explicitly as follows:

$$\nabla_B^2 f(u) := -\mu\Delta + K^\top\lambda K + \alpha\nabla^\top D\left(\max(|\nabla u|, \gamma)^{q-2}(I - (2-q)\chi_{\mathcal{A}}(u)|\nabla u|^{-2}(\nabla u)(\nabla u)^\top)\right)\nabla,$$

where $\chi_{\mathcal{A}}(u)$ is defined by

$$(\chi_{\mathcal{A}}(u))_{ij} := \begin{cases} 1, & \text{if } |(\nabla u)_{ij}| > \gamma, \\ 0, & \text{otherwise.} \end{cases}$$

We shall refer to $\nabla_B^2 f(u)$ as the *B-Hessian* of f at u.

Due to its favorable local convergence properties, we are interested in applying a generalized version of Newton's method for solving (5). In variational image processing it has turned out that primal-dual Newton schemes are typically superior to purely primal or dual iterations; see, e.g., [3, 23, 24]. Hence, we reformulate the Euler-Lagrange equation (5) by introducing a new variable $\vec{p} \in (\mathbb{R}^{|\Omega|})^2$, which plays the role of a dual variable, i.e.

$$\begin{cases} -\mu\Delta u + K^\top \lambda(Ku - z) + \alpha\nabla^\top \vec{p} = 0, \\ \max(|\nabla u|, \gamma)^{2-q}\vec{p} = \nabla u. \end{cases} \tag{8}$$

This system is the starting point for developing our generalized Newton scheme in the next section.

3 Primal-dual Newton method

3.1 Regularized Newton via reweighted Euler-Lagrange equation

In order to handle the nonlinear diffusion term (which contains the $(q-2)$-th power of the max-term) in the Euler-Lagrange equation (5), we invoke an approach relying on reweighting. Similar techniques were previously considered in [40, 4, 7, 12]. In fact, let u^k be our current approximation of a solution to (5). Then the reweighted Euler-Lagrange equation is given by

$$-\mu\Delta u + K^\top \lambda(Ku - z) + \alpha\nabla^\top \left(w^k \max(|\nabla u|, \gamma)^{-r} \nabla u \right) = 0, \tag{9}$$

with $0 \le r \le 2 - q$ and the weight w^k defined by

$$w^k := \max(|\nabla u^k|, \gamma)^{q+r-2}.$$

We further introduce a reweighted dual variable

$$\vec{p} = w^k \max(|\nabla u|, \gamma)^{-r} \nabla u.$$

As a result, the equation (8) may be written as

$$\begin{cases} -\mu\Delta u + K^\top \lambda(Ku - z) + \alpha\nabla^\top \vec{p} = 0, \\ (w^k)^{-1} \max(|\nabla u|, \gamma)^r \vec{p} = \nabla u. \end{cases} \tag{10}$$

Next, at u^k we define the active set $\mathcal{A}^k := \{(i,j) \in \Omega : |(\nabla u^k)_{ij}| > \gamma\}$. Given a current approximation $(u^k, \vec{p}^k)$, we apply a generalized linearization to (10) and obtain the generalized Newton system

$$\begin{bmatrix} -\mu\Delta + K^\top \lambda K & \alpha\nabla^\top \\ -\tilde{C}^k(r)\nabla & D((m^k)^{2-q}\vec{e}) \end{bmatrix} \begin{bmatrix} \delta u^{k+1} \\ \delta \vec{p}^{\,k+1} \end{bmatrix} = \begin{bmatrix} \mu\Delta u^k - K^\top \lambda(Ku^k - z) - \alpha\nabla^\top \vec{p}^{\,k} \\ \nabla u^k - (m^k)^{2-q}\vec{p}^{\,k} \end{bmatrix}, \tag{11}$$

where

$$m^k := \max(|\nabla u^k|, \gamma), \tag{12}$$

$$\tilde{C}^k(r) := I - rD(\chi_{\mathcal{A}^k}(m^k)^{-q}\vec{p}^k) \begin{bmatrix} D(\nabla_x u^k) & D(\nabla_y u^k) \\ D(\nabla_x u^k) & D(\nabla_y u^k) \end{bmatrix}. \tag{13}$$

After eliminating $\delta\vec{p}^{k+1}$, we are left with the linear system

$$\tilde{H}^k(r)\delta u^{k+1} = -g^k, \tag{14}$$

where

$$\tilde{H}^k(r) := -\mu\Delta + K^\top\lambda K + \alpha\nabla^\top D((m^k)^{q-2}\vec{e})\tilde{C}^k(r)\nabla, \tag{15}$$

$$g^k := -\mu\Delta u^k + K^\top\lambda(Ku^k - z) + \alpha\nabla^\top((m^k)^{q-2}\nabla u^k). \tag{16}$$

Note that $g^k = \nabla f(u^k)$ in (5). Upon solving (14) for δu^{k+1}, we compute $\delta\vec{p}^{k+1}$ according to (11), i.e.

$$\delta\vec{p}^{k+1} = (m^k)^{q-2}(\nabla u^k + \tilde{C}^k(r)\nabla\delta u^{k+1}) - \vec{p}^k. \tag{17}$$

Assuming that δu^{k+1} is a descent direction, i.e. $(g^k)^\top\delta u^{k+1} < 0$, we update $u^{k+1} := u^k + a^k\delta u^{k+1}$ and $\vec{p}^{k+1} := \vec{p}^k + a^k\delta\vec{p}^{k+1}$ with a suitable step size a^k, and then go to the next Newton iteration.

Note that $H^k := \tilde{H}^k(2-q)$ is the B-Hessian in the non-reweighted primal-dual Newton method [40, 24]. We observe that the reweighting procedure is, in fact, equivalent to a *R-regularization* of the B-Hessian of the non-reweighting approach. In order to see this, let

$$R^k := \alpha\nabla^\top D(\chi_{\mathcal{A}^k}(m^k)^{-2}\vec{p}^k) \begin{bmatrix} D(\nabla_x u^k) & D(\nabla_y u^k) \\ D(\nabla_x u^k) & D(\nabla_y u^k) \end{bmatrix} \nabla.$$

Then the Newton update (14) becomes

$$(H^k + \beta R^k)\delta u^{k+1} = -g^k, \tag{18}$$

with $\beta = 2 - q - r$.

Subsequently we consider variable β, i.e. $\beta = \beta^k$, and a slight modification of the R-matrix to guarantee (i) well-definedness of the Newton iteration defined below, (ii) the aforementioned descent property and (iii) ultimately the locally superlinear convergence of our overall algorithmic scheme. For the latter, we show in the proof of Theorem 3.10 that $\lim_{k\to+\infty}\beta^k = 0$. Thus, the R-regularization vanishes for $k \to +\infty$.

3.2 Infeasible Newton method

We next study feasibility properties of the iterates of a generalized Newton method relying on (14) and definiteness of $\tilde{H}^k(r)$. For this discussion, we return to the reweighted Euler-Lagrange equation (9) with $0 \le r \le 1$ (or $1 - q \le \beta \le 2 - q$). In particular, assuming that $\vec{p}^k = |\nabla u^k|^{q-2}\nabla u^k$ on $\mathcal{A}^k$, we have that

$$\tilde{C}^k(r) = I - rD(\chi_{\mathcal{A}^k}(m^k)^{-2}\vec{e}) \begin{bmatrix} D(|\nabla_x u^k|^2) & D(\nabla_x u^k\nabla_y u^k) \\ D(\nabla_x u^k\nabla_y u^k) & D(|\nabla_y u^k|^2) \end{bmatrix} \succeq 0, \tag{19}$$

where "$\succeq$" indicates positive semidefiniteness of a matrix. Therefore, we conclude

$$\tilde{H}^k(r) = -\mu\Delta + K^\top\lambda K + \alpha\nabla^\top D((m^k)^{q-2}\vec{e})\tilde{C}^k(r)\nabla \succ 0,$$

i.e. $\tilde{H}^k(r)$ is positive definite, since $-\mu\Delta + K^\top \lambda K \succ 0$ under the hypothesis (2). In general, however, $\tilde{H}^k(r)$ may be indefinite during generalized Newton iterations.

In the following, we derive a sufficient condition for r (or β) such that the system matrix $\tilde{H}^k(r)$ is positive definite; see Theorem 3.2 below. This property of $\tilde{H}^k(r)$ is useful to guarantee that a descent direction δu^k is computed in each Newton iteration. Moreover, it constitutes an iteration dependent regularization scheme.

For this purpose, we propose two modifications of the system matrix $\tilde{H}^k(r)$. First, we replace $\vec{p}^k$ by $\vec{p}^k_+$, where

$$\vec{p}^k_+ := \frac{\chi_{\mathcal{A}^k}(m^k)^{q-1}\vec{p}^k}{\max((m^k)^{q-1}, |\vec{p}^k|)} + (1 - \chi_{\mathcal{A}^k})\vec{p}^k. \tag{20}$$

Note that the modified $\vec{p}^k_+$ satisfies its feasibility condition on $\mathcal{A}^k$, i.e.

$$|(\vec{p}^k_+)_{ij}| \le |(\nabla u^k)_{ij}|^{q-1}, \quad \text{whenever } (i,j) \in \mathcal{A}^k. \tag{21}$$

Secondly, we replace $\tilde{C}^k(r)$ by its symmetrization denoted by $\tilde{C}^k_+(r)$, i.e.

$$\begin{aligned}\tilde{C}^k_+(r) :=& \frac{\tilde{C}^k(r) + \tilde{C}^k(r)^\top}{2} = I - rD(\chi_{\mathcal{A}^k}(m^k)^{-q})\cdot \\ &\cdot \begin{bmatrix} D((\vec{p}^k_+)_x \nabla_x u^k) & D(\frac{1}{2}((\vec{p}^k_+)_x \nabla_y u^k + (\vec{p}^k_+)_y \nabla_x u^k)) \\ D(\frac{1}{2}((\vec{p}^k_+)_x \nabla_y u^k + (\vec{p}^k_+)_y \nabla_x u^k)) & D((\vec{p}^k_+)_y \nabla_y u^k) \end{bmatrix}.\end{aligned} \tag{22}$$

Accordingly, the system matrix $\tilde{H}^k(r)$ in (14) is replaced by $\tilde{H}^k_+(r)$ with

$$\tilde{H}^k_+(r) := -\mu\Delta + K^\top \lambda K + \alpha \nabla^\top D((m^k)^{q-2}\vec{e})\tilde{C}^k_+(r)\nabla, \tag{23}$$

and the regularizer R^k is replaced by R^k_+ with

$$\begin{aligned}R^k_+ :=& \alpha \nabla^\top D(\chi_{\mathcal{A}^k}(m^k)^{-2})\cdot \\ &\cdot \begin{bmatrix} D((\vec{p}^k_+)_x \nabla_x u^k) & D(\frac{1}{2}((\vec{p}^k_+)_x \nabla_y u^k + (\vec{p}^k_+)_y \nabla_x u^k)) \\ D(\frac{1}{2}((\vec{p}^k_+)_x \nabla_y u^k + (\vec{p}^k_+)_y \nabla_x u^k)) & D((\vec{p}^k_+)_y \nabla_y u^k) \end{bmatrix} \nabla.\end{aligned} \tag{24}$$

Lemma 3.1. *Assume that $0 \le r \le 1$ (or $1 - q \le \beta \le 2 - q$) and the feasibility condition (21) holds true. Then the matrix $\tilde{C}^k_+(r)$ given in (22) is positive semidefinite.*

Proof. By reordering, it suffices to show that each 2-by-2 block

$$[\tilde{C}^k_+(r)]_{ij} = I - r\chi_{\mathcal{A}^k}(m^k)^{-q} \begin{bmatrix} (\vec{p}^k_+)_x \nabla_x u^k & \frac{1}{2}((\vec{p}^k_+)_x \nabla_y u^k + (\vec{p}^k_+)_y \nabla_x u^k) \\ \frac{1}{2}((\vec{p}^k_+)_x \nabla_y u^k + (\vec{p}^k_+)_y \nabla_x u^k) & (\vec{p}^k_+)_y \nabla_y u^k \end{bmatrix}$$

is positive semidefinite. For the ease of notation, the subscripts ij are frequently omitted for the remainder of this proof.

We distinguish two cases with respect to (i,j). First, consider the case where $(i,j) \notin \mathcal{A}^k$. Then we have $[\tilde{C}^k_+(r)]_{ij} = I$ and the assertion holds immediately.

In the second case where $(i,j) \in \mathcal{A}^k$, we have

$$[\tilde{C}^k_+(r)]_{ij} = \begin{bmatrix} 1 - r|\nabla u^k|^{-q}(\vec{p}^k_+)_x \nabla_x u^k & -\frac{r}{2}|\nabla u^k|^{-q}((\vec{p}^k_+)_x \nabla_y u^k + (\vec{p}^k_+)_y \nabla_x u^k) \\ -\frac{r}{2}|\nabla u^k|^{-q}((\vec{p}^k_+)_x \nabla_y u^k + (\vec{p}^k_+)_y \nabla_x u^k) & 1 - r|\nabla u^k|^{-q}(\vec{p}^k_+)_y \nabla_y u^k \end{bmatrix}.$$

This 2-by-2 block has nonnegative eigenvalues, since its diagonal elements are nonnegative and its determinant satisfies

$$
\begin{aligned}
&(1 - r|\nabla u^k|^{-q}(\bar{p}_+^k)_x \nabla_x u^k)(1 - r|\nabla u^k|^{-q}(\bar{p}_+^k)_y \nabla_y u^k) - \frac{r^2}{4}|\nabla u^k|^{-2q}|(\bar{p}_+^k)_x \nabla_y u^k + (\bar{p}_+^k)_y \nabla_x u^k|^2 \\
=&1 - r|\nabla u^k|^{-q}((\bar{p}_+^k)_x \nabla_x u^k + (\bar{p}_+^k)_y \nabla_y u^k) - \frac{r^2}{4}|\nabla u^k|^{-2q}|(\bar{p}_+^k)_x \nabla_y u^k - (\bar{p}_+^k)_y \nabla_x u^k|^2 \\
=&1 - r|\nabla u^k|^{-q}((\bar{p}_+^k)_x \nabla_x u^k + (\bar{p}_+^k)_y \nabla_y u^k) - \frac{r^2}{4}|\nabla u^k|^{-2q}. \\
&\quad \cdot \Big[(|(\bar{p}_+^k)_x|^2 + |(\bar{p}_+^k)_y|^2)(|\nabla_x u^k|^2 + |\nabla_y u^k|^2) - |(\bar{p}_+^k)_x \nabla_x u^k + (\bar{p}_+^k)_y \nabla_y u^k|^2\Big] \\
=&-\frac{r^2}{4}|\nabla u^k|^{2-2q}|\bar{p}_+^k|^2 + \Big[1 - \frac{r}{2}|\nabla u^k|^{-q}((\bar{p}_+^k)_x \nabla_x u^k + (\bar{p}_+^k)_y \nabla_y u^k)\Big]^2 \\
\geq&-\frac{r^2}{4}|\nabla u^k|^{2-2q}|\bar{p}_+^k|^2 + \Big[1 - \frac{r}{2}|\nabla u^k|^{1-q}|\bar{p}_+^k|\Big]^2 = 1 - r|\nabla u^k|^{1-q}|\bar{p}_+^k| \geq 0.
\end{aligned}
$$

In deriving the above inequalities, we have used the assumption that $0 \leq r \leq 1$, the feasibility condition (21), and the Cauchy-Schwarz inequality. □

The following theorem is an immediate consequence of Lemma 3.1 and the structure of $\tilde{H}_+^k(r)$.

Theorem 3.2. *Suppose the same assumptions of Lemma 3.1 are satisfied. Then the following statements hold true:*

1. *The matrix $\tilde{H}_+^k(r)$ is positive definite.*
2. *We have the following estimate on the spectrum of $\tilde{H}_+^k(r)$:*

$$
\begin{aligned}
\lambda_{\min}(\tilde{H}_+^k(r)) &\geq \lambda_{\min}(-\mu\Delta + K^\top \lambda K), \\
\lambda_{\max}(\tilde{H}_+^k(r)) &\leq \lambda_{\max}(-(\mu + 3\alpha\gamma^{q-2})\Delta + K^\top \lambda K).
\end{aligned}
$$

3. *We obtain from (14) a descent direction δu^{k+1} satisfying*

$$
-\frac{(g^k)^\top \delta u^{k+1}}{\|g^k\|\|\delta u^{k+1}\|} \geq \frac{\lambda_{\min}(\tilde{H}_+^k(r))}{\lambda_{\max}(\tilde{H}_+^k(r))} \geq \bar{\epsilon}_d := \frac{\lambda_{\min}(-\mu\Delta + K^\top \lambda K)}{\lambda_{\max}(-(\mu + 3\alpha\gamma^{q-2})\Delta + K^\top \lambda K)}.
$$

3.3 Superlinear convergence by adaptive regularization

Using the results in [40, 4], one readily finds that the R-regularized version of the generalized Newton method with fixed β, which results in the reweighting approach, is linearly convergent. This is also confirmed by our numerical test in Section 4; see Figure 5.

In this section, we propose a new adaptively R-regularized version of the generalized Newton method that attains superlinear local convergence. This requires an appropriate update strategy for $\beta > 0$. For this purpose, we propose a trust-region-type scheme; see, e.g., [35, 11] for comprehensive discussions of trust-region methods. Given a current iterate u^k, these methods typically model f locally by a quadratic function $h^k : \mathbb{R}^{|\Omega|} \to \mathbb{R}$ with

$$
h^k(d) := f(u^k) + (g^k)^\top d + \frac{1}{2} d^\top H_+^k d. \tag{25}
$$

Here we let $H_+^k := \tilde{H}_+^k(2-q)$; see (23). Consider now the minimization of h^k subject to the trust-region constraint, i.e.

$$\text{minimize } h^k(d) \quad \text{over } d \in \mathbb{R}^{|\Omega|} \tag{26}$$

$$\text{subject to } \frac{1}{2} d^\top R_{+,\varepsilon}^k d \le \frac{1}{2}(\sigma^k)^2. \tag{27}$$

Here $\sigma^k > 0$ represents the trust-region radius, and

$$R_{+,\varepsilon}^k := R_+^k + \varepsilon I, \tag{28}$$

is defined with an arbitrarily fixed regularization parameter $0 < \varepsilon \ll 1$. The existence of a solution to (26)–(27) hinges on the interplay of H_+^k and $R_{+,\varepsilon}^k$.

Lemma 3.3. *The matrix H_+^k is positive definite on $\{d \in \mathbb{R}^{|\Omega|} : d^\top R_{+,\varepsilon}^k d \le 0\}$.*

Proof. Suppose $d \in \mathbb{R}^{|\Omega|}$ satisfies $d \neq 0$ and $d^\top R_+^k d \le -\varepsilon \|d\|^2 < 0$. Then we have

$$d^\top H_+^k d = d^\top(-\mu\Delta + K^\top \lambda K)d + \alpha(\nabla d)^\top D((m^k)^{q-2}\vec{e})\nabla d - (2-q)d^\top R_+^k d > 0,$$

which proves the assertion. □

Theorem 3.4. *There exists a solution to the trust-region subproblem (26)–(27).*

Proof. Note that the objective is at most quadratic and the feasible set is nonempty and closed. It suffices to show that $h^k(d^l) \to +\infty$ for any feasible sequence (d^l) with $\|d^l\| \to +\infty$. We shall prove this by contradiction. Let such a sequence (d^l) be given, and assume oppositely that $(h^k(d^l))$ is uniformly bounded from above. For each l, we write $d^l = s^l v^l$ such that $s^l \in \mathbb{R}$, $v^l \in \mathbb{R}^{|\Omega|}$, and $\|v^l\| = 1$. By compactness, there exists a subsequence of (v^l), say $(v^{l'})$, such that $v^{l'} \to v^*$ for some $v^* \in \mathbb{R}^{|\Omega|}$. The constraint (27) yields that $(v^{l'})^\top R_{+,\varepsilon}^k v^{l'} \le (\sigma^k)^2/(s^{l'})^2$. Letting $l' \to +\infty$, we get $(v^*)^\top R_{+,\varepsilon}^k v^* \le 0$. It follows from Lemma 3.3 that $(v^*)^\top H_+^k v^* > 0$. Thus we must have $h^k(d^{l'}) \to +\infty$ as $l' \to +\infty$, which contradicts our assumption. □

Given the current iterate u^k, we aim to determine a search direction d^k by approximately solving the trust-region subproblem. A classical argument in the convergence analysis of trust-region methods requires that the search direction d^k yields a reduction in the model function h^k proportional to the decrease implied by the *Cauchy point* [11].

The Cauchy point is defined by $d_C^k := -t^k g^k$, where t^k minimizes the one-dimensional problem

$$t^k := \arg\min\{h^k(-tg^k) : \ t^2 (g^k)^\top R_{+,\varepsilon}^k g^k \le (\sigma^k)^2 \wedge t \ge 0\}.$$

Let $t_*^k := \|g^k\|^2/((g^k)^\top H_+^k g^k)$ be the critical point presumed it exists. The Cauchy point can be explicitly computed through the following three cases:

1. Suppose $(g^k)^\top H_+^k g^k \le 0$. By Lemma 3.3, we have $(g^k)^\top R_{+,\varepsilon}^k g^k > 0$. The Cauchy point lies on the boundary of the trust region, i.e. $d_C^k = -\left(\sigma^k / \sqrt{(g^k)^\top R_{+,\varepsilon}^k g^k}\right) g^k$, and

$$h^k(0) - h^k(d^k) = \frac{\sigma^k \|g^k\|^2}{\sqrt{(g^k)^\top R_{+,\varepsilon}^k g^k}} - \frac{(\sigma^k)^2 (g^k)^\top H_+^k g^k}{2(g^k)^\top R_{+,\varepsilon}^k g^k} \ge \frac{\sigma^k \|g^k\|^2}{\sqrt{(g^k)^\top R_{+,\varepsilon}^k g^k}}. \tag{29}$$

2. Suppose $(g^k)^\top H_+^k g^k > 0$ and $(t_*^k)^2 (g^k)^\top R_{+,\varepsilon}^k g^k \le (\sigma^k)^2$. Then we have $d_C^k = -t_*^k g^k = -\left(\|g^k\|^2/((g^k)^\top H_+^k g^k)\right) g^k$, and

$$h^k(0) - h^k(d^k) = \frac{\|g^k\|^4}{2(g^k)^\top H_+^k g^k} \ge \frac{\|g^k\|^2}{2\lambda_{\max}(H_+^k)}. \tag{30}$$

3. Suppose $(g^k)^\top H_+^k g^k > 0$ and $(t_*^k)^2 (g^k)^\top R_{+,\varepsilon}^k g^k > (\sigma^k)^2$. Then similar to the first case, we have $d_C^k = -\left(\sigma^k / \sqrt{(g^k)^\top R_{+,\varepsilon}^k g^k}\right) g^k$. In particular, $\sigma^k((g^k)^\top H_+^k g^k)/\sqrt{(g^k)^\top R_{+,\varepsilon}^k g^k} < \|g^k\|^2$. Therefore, we have

$$h^k(0) - h^k(d^k) = \frac{\sigma^k \|g^k\|^2}{\sqrt{(g^k)^\top R_{+,\varepsilon}^k g^k}} - \frac{(\sigma^k)^2 (g^k)^\top H_+^k g^k}{2(g^k)^\top R_{+,\varepsilon}^k g^k} \ge \frac{\sigma^k \|g^k\|^2}{2\sqrt{(g^k)^\top R_{+,\varepsilon}^k g^k}}. \tag{31}$$

The search direction d^k is said to satisfy the *Cauchy-point-based model reduction criterion* if

$$h^k(0) - h^k(d^k) \ge C \|g^k\|^2 \eta^k, \tag{32}$$

for some constant $C > 0$, where

$$\eta^k := \begin{cases} \dfrac{\sigma^k}{\sqrt{(g^k)^\top R_{+,\varepsilon}^k g^k}}, & \text{if } (g^k)^\top H_+^k g^k \le 0, \\ \dfrac{1}{\lambda_{\max}(H_+^k)}, & \text{if } (g^k)^\top R_{+,\varepsilon}^k g^k \le 0, \\ \min\left(\dfrac{\sigma^k}{\sqrt{(g^k)^\top R_{+,\varepsilon}^k g^k}}, \dfrac{1}{\lambda_{\max}(H_+^k)} \right), & \text{otherwise.} \end{cases} \tag{33}$$

Due to Lemma 3.3, η^k is well-defined. It is easily seen that (29)–(31) satisfy the criterion (32) with $C = 1/2$.

Now we turn to the computation of an approximate solution to the trust-region subproblem (26)–(27). In the forthcoming Theorem 3.5, we shall characterize this solution d_*^k by

$$(H_+^k + \beta_*^k R_{+,\varepsilon}^k) d_*^k = -g^k, \tag{34}$$

$$\beta_*^k \left((d_*^k)^\top R_{+,\varepsilon}^k d_*^k - (\sigma^k)^2 \right) = 0, \tag{35}$$

$$H_+^k + \beta_*^k R_{+,\varepsilon}^k \succeq 0, \tag{36}$$

for some $\beta_*^k \ge 0$. Its proof essentially adopts that of [35, Theorem 4.1] under our context.

Theorem 3.5. *The trust-region subproblem (26)–(27) has a global solution d_*^k if and only if d_*^k is feasible and there exists a scalar $\beta_*^k \ge 0$ such that (34)–(36) are satisfied.*

Proof. (if part) Suppose there exists $\beta_*^k \ge 0$ such that (34)–(36) hold. Then by Lemma 4.7 in [35], d_*^k minimizes $\hat{h}^k : \mathbb{R}^{|\Omega|} \to \mathbb{R}$, where

$$\hat{h}^k(d^k) = (g^k)^\top d^k + \frac{1}{2}(d^k)^\top (H_+^k + \beta_*^k R_{+,\varepsilon}^k) d^k = h^k(d^k) + \frac{\beta_*^k}{2}(d^k)^\top R_{+,\varepsilon}^k d^k - f(u^k).$$

If follows from $\hat{h}^k(d^k) \geq \hat{h}^k(d^k_*)$ that

$$\begin{aligned} h^k(d^k) \geq h^k(d^k_*) + \frac{\beta^k_*}{2}((d^k_*)^\top R^k_{+,\varepsilon} d^k_* - (d^k)^\top R^k_{+,\varepsilon} d^k) \\ = h^k(d^k_*) + \frac{\beta^k_*}{2}((\sigma^k)^2 - (d^k)^\top R^k_{+,\varepsilon} d^k) \geq h^k(d^k_*). \end{aligned}$$

Since d^k is arbitrary but feasible, the assertion follows.

(only-if part) Suppose now that d^k_* is the global solution of the trust-region subproblem (26)–(27).

- Case 1: $(d^k_*)^\top R^k_{+,\varepsilon} d^k_* < (\sigma^k)^2$. The second-order necessary conditions of the unconstrained problem imply that

$$\begin{aligned} \nabla h^k(d^k_*) &= H^k_+ d^k_* + g^k = 0, \\ \nabla^2 h^k(d^k_*) &= H^k_+ \succeq 0. \end{aligned}$$

  We get the desired conclusion with $\beta^k_* = 0$.

- Case 2: $(d^k_*)^\top R^k_{+,\varepsilon} d^k_* = (\sigma^k)^2$. In particular we have $R^k_{+,\varepsilon} d^k_* \neq 0$, and therefore the linear independence constraint qualification (see, e.g., [35]) is fulfilled at d^k_*. By the second-order necessary condition, there exists $\beta^k_* \geq 0$ such that

$$H^k_+ d^k_* + g^k + \beta^k_* R^k_{+,\varepsilon} d^k_* = 0, \tag{37}$$

  and

$$v^\top (H^k_+ + \beta^k_* R^k_{+,\varepsilon}) v \geq 0, \tag{38}$$

  for any nonzero vector $v \in \mathbb{R}^{|\Omega|}$ with $v^\top R^k_{+,\varepsilon} d^k_* = 0$.

  It remains to show (38) for any nonzero vector v with $v^\top R^k_{+,\varepsilon} d^k_* \neq 0$. Let such a vector v be given. In particular we have $R^k_{+,\varepsilon} v \neq 0$. Define

$$d^k := d^k_* - \frac{2 v^\top R^k_{+,\varepsilon} d^k_*}{v^\top R^k_{+,\varepsilon} v} v. \tag{39}$$

  Then it is easy to check that $(d^k)^\top R^k_{+,\varepsilon} d^k = (\sigma^k)^2$. Since $h^k(d^k) \geq h^k(d^k_*)$, we have

$$h^k(d^k) \geq h^k(d^k_*) + \frac{\beta^k_*}{2}((d^k_*)^\top R^k_{+,\varepsilon} d^k_* - (d^k)^\top R^k_{+,\varepsilon} d^k).$$

  From this and (37), we infer

$$\frac{1}{2}(d^k - d^k_*)^\top (H^k_+ + \beta^k_* R^k_{+,\varepsilon})(d^k - d^k_*) \geq 0.$$

  Thus in view of (39) we have shown (38) for any nonzero vector v with $v^\top R^k_{+,\varepsilon} d^k_* \neq 0$, which completes the proof.

□

Based on the above observation concerning h^k and using a complementarity function; see, e.g., [22], we can equivalently formulate (34)–(36), with an arbitrarily fixed scalar $c > 0$, as follows:

$$\begin{aligned} (H_+^k + \beta_*^k R_{+,\varepsilon}^k) d_*^k &= -g^k, \\ \beta_*^k - \max\left(\beta_*^k + \frac{1}{2c}((d_*^k)^\top R_{+,\varepsilon}^k d_*^k - (\sigma^k)^2), 0\right) &= 0, \\ H_+^k + \beta_*^k R_{+,\varepsilon}^k &\succeq 0. \end{aligned}$$

From this formulation, we propose an adaptively regularized Newton iteration which converges globally and locally at a superlinear rate.

Algorithm 3.6 (Adaptively regularized Newton method)**.**
Choose parameters: $1 - q \leq \beta_{\max} \leq 2 - q$, $c > 0$, $0 < \rho_1 \leq \rho_2 < 1$, $0 < \kappa_1 < 1 < \kappa_2$, $0 < \varepsilon \ll 1$, $0 < \epsilon_d \leq \bar{\epsilon}_d$. *Then iterate as follows:*

0. *Initialize the primal and dual variables* $(u^0, \vec{p}^0)$, *the regularization scalar* $\beta^0 \geq 0$, *and the trust-region radius* $\sigma^0 > 0$. *Set* $k := 0$.
1. *Generate* H_+^k, $R_{+,\varepsilon}^k$, *and* g^k.
2. *Solve* $(H_+^k + \beta^k R_{+,\varepsilon}^k) d^k = -g^k$ *for* d^k.
3. *If* $-(g^k)^\top d^k / (\|g^k\| \|d^k\|) < \epsilon_d$, *then set* $\beta^k := \beta_{\max}$ *and return to Step 2.*
4. *If* $\beta^k = \beta_{\max}$ *and* $(d^k)^\top R_{+,\varepsilon}^k d^k > (\sigma^k)^2$, *then set* $\sigma^k := \sqrt{(d^k)^\top R_{+,\varepsilon}^k d^k}$ *and go to Step 8.*
5. *Update* $\beta^k := \beta^k + ((d^k)^\top R_{+,\varepsilon}^k d^k - (\sigma^k)^2)/(2c)$.
6. *Projection onto the feasible interval:* $\beta^k := \max(\min(\beta^k, \beta_{\max}), 0)$.
7. *If the stopping criterion (for the inner loop) is not fulfilled, then return to Step 2.*
8. *Evaluate the ratio* $\rho^k := [f(u^k) - f(u^k + d^k)] / [f(u^k) - (f(u^k) + (g^k)^\top d^k + (d^k)^\top H_+^k d^k / 2)]$.
9. *If* $\rho^k < \rho_1$, *then set* $\sigma^{k+1} := \kappa_1 \sigma^k$; *else if* $\rho^k > \rho_2$, *then set* $\sigma^{k+1} := \kappa_2 \sigma^k$; *else* $\sigma^{k+1} := \sigma^k$.
10. *Set* $\delta u^{k+1} := d^k$, *and compute* $\delta \vec{p}^{k+1}$ *according to (17).*
11. *Perform a Wolfe-Powell line search along the direction* d^k *to determine the step size* a^k. *Update* $u^{k+1} := u^k + a^k \delta u^{k+1}$ *and* $\vec{p}^{k+1} := \vec{p}^k + a^k \delta \vec{p}^{k+1}$.
12. *If the stopping criterion (for the outer loop) is not fulfilled, then set* $\beta^{k+1} := \beta^k$, $k := k+1$, *and return to Step 1.*

We observe that the above algorithm combines a trust-region technique for adjusting the weight β in the R-regularization (Steps 2–7) with a line search method for updating the iterate along the direction obtained from the approximately weighted R-regularized problem (Step 11). We emphasize, however, that the classical trust-region approach might be used instead of the line search procedure for globalizing Newton's method. In Algorithm 3.6, the global convergence is guaranteed by the Wolfe-Powell line search while the trust-region-type framework is utilized to guarantee that d^k is a descent direction for f at u^k and to retain the locally superlinear

convergence of Newton's method. Based on our numerical experience we prefer the Wolfe-Powell line search over other, possibly simpler, rules as it appears to better resolve the line search problem for our nonconvex objective.

Note that our objective f is bounded from below and continuously differentiable. Moreover, its gradient $\nabla f(\cdot)$ is Lipschitz continuous on an open set containing the level set $\{u \in \mathbb{R}^{|\Omega|} : f(u) \leq f(u^0)\}$. Thus Zoutendijk's theorem (see, e.g., [35]) can be applied in order to derive global convergence of Algorithm 3.6.

Theorem 3.7 (Global Convergence)**.** *Let $u^{k+1} = u^k + a^k d^k$ such that the Wolfe-Powell conditions are satisfied, i.e.*

$$f(u^{k+1}) \leq f(u^k) + \tau_1 a^k \nabla f(u^k)^\top d^k, \tag{40}$$

$$\nabla f(u^{k+1})^\top d^k \geq \tau_2 \nabla f(u^k)^\top d^k, \tag{41}$$

with $0 < \tau_1 < 1/2$ and $\tau_1 < \tau_2 < 1$. Then we have $\lim_{k\to+\infty} \|\nabla f(u^k)\| = 0$.

Proof. By Theorem 3.2 in [35], we have $\sum_{k=0}^{+\infty} \cos^2\theta^k \|g^k\|^2 < +\infty$, where

$$\cos\theta^k := -\frac{(g^k)^\top d^k}{\|g^k\|\|d^k\|}.$$

Since $\cos\theta^k \geq \epsilon_d$ holds true for every k due to Step 3 of Algorithm 3.6 and Theorem 3.2, we conclude that $\lim_{k\to+\infty} \|\nabla f(u^k)\| = 0$. □

Next we study the locally superlinear convergence of Algorithm 3.6; see Theorem 3.10 below. For this purpose, we need to understand the approximation properties of (H_+^k) with respect to $(\nabla_B^2 f(u^k))$, and the definiteness properties of $(R_{+,\varepsilon}^k)$. This is the subject of the following auxiliary result.

Lemma 3.8. *Assume that the primal-dual sequence $(u^k, \vec{p}^k)$ converges to some $(u^*, \vec{p}^*)$ satisfying the Euler-Lagrange system (8). Then the following statements hold true:*

1. *The modified system matrix H_+^k approaches asymptotically the B-Hessian $\nabla_B^2 f(u^k)$, i.e. $\lim_{k\to+\infty} \|H_+^k - \nabla_B^2 f(u^k)\| = 0$.*
2. *For all sufficiently large k, the matrix $R_{+,\varepsilon}^k$ is strictly positive definite and its minimal eigenvalue satisfies $\lambda_{\min}(R_{+,\varepsilon}^k) > \varepsilon/2$.*

Proof. (Proof of 1.) Let $C^k := \tilde{C}^k(2-q)$ in (19) and $C_+^k := \tilde{C}_+^k(2-q)$ in (22). For $k \to +\infty$ we have $(u^k, \vec{p}^k) \to (u^*, \vec{p}^*)$ with the latter satisfying the Euler-Lagrange equation (8). Further, for all $(i,j) \in \Omega$ we have

$$\begin{aligned} |\vec{p}_+^k - \vec{p}^k| &\leq |\vec{p}^k| \left| \frac{(m^k)^{q-1}}{\max((m^k)^{q-1}, |\vec{p}^k|)} - 1 \right| \to |\vec{p}^*| \left| \frac{\max(|\nabla u^*|, \gamma)^{q-1}}{\max(\max(|\nabla u^*|, \gamma)^{q-1}, |\vec{p}^*|)} - 1 \right| \\ &= |\vec{p}^*| \left| \frac{\max(|\nabla u^*|, \gamma)^{q-1}}{|\vec{p}^*| \max(|\nabla u^*|, \gamma)/|\nabla u^*|} - 1 \right| = 0 \end{aligned} \tag{42}$$

as $k \to \infty$. Moreover, C^k will converge to a symmetric matrix, and therefore $C_+^k = (C^k + (C^k)^\top)/2$ approaches asymptotically C^k, i.e. $\lim_{k\to+\infty} \|C_+^k - C^k\| = 0$. Thus, due to the structures of H^k and H_+^k, we have $\lim_{k\to+\infty} \|H_+^k - H^k\| = 0$.

Finally, as $(u^k, \vec{p}^k) \to (u^*, \vec{p}^*)$, it is easy to see that both H^k and $\nabla_B^2 f(u^k)$ converge to $\nabla_B^2 f(u^*)$, which yields $\lim_{k\to+\infty} \|H^k - \nabla_B^2 f(u^k)\| = 0$. Thus we conclude that $\lim_{k\to+\infty} \|H_+^k - \nabla_B^2 f(u^k)\| = 0$ as desired.

(Proof of 2.) Our proof again utilizes the reordered system as in Lemma 3.1. In view of the definition of $R_{+,\varepsilon}^k$, see (28), and the structure of R_+^k, see (24), it suffices to show that for all $(i,j) \in \Omega$, the minimal eigenvalue of the 2-by-2 block

$$\chi_{\mathcal{A}^k}(m^k)^{-2} \begin{bmatrix} (\vec{p}_+^k)_x \nabla_x u^k & \frac{1}{2}((\vec{p}_+^k)_x \nabla_y u^k + (\vec{p}_+^k)_y \nabla_x u^k) \\ \frac{1}{2}((\vec{p}_+^k)_x \nabla_y u^k + (\vec{p}_+^k)_y \nabla_x u^k) & (\vec{p}_+^k)_y \nabla_y u^k \end{bmatrix} \tag{43}$$

goes to zero as $k \to +\infty$. The characteristic equation of the 2-by-2 block (43) without the factor $\chi_{\mathcal{A}^k}$ is given by

$$\begin{aligned} &t^2 - (m^k)^{-2}((\vec{p}_+^k)_x \nabla_x u^k + (\vec{p}_+^k)_y \nabla_y u^k)t \\ &\quad + (m^k)^{-4}\left((\vec{p}_+^k)_x \nabla_x u^k (\vec{p}_+^k)_y \nabla_y u^k - \frac{1}{4}|(\vec{p}_+^k)_x \nabla_y u^k + (\vec{p}_+^k)_y \nabla_x u^k|^2\right) = 0. \end{aligned}$$

Note that due to (42) we have $\lim_{k\to+\infty} \vec{p}_+^k = \vec{p}^*$ such that $(u^*, \vec{p}^*)$ satisfies (8). Therefore, as $k \to +\infty$, we have

$$\begin{aligned} &(m^k)^{-4}\left((\vec{p}_+^k)_x \nabla_x u^k (\vec{p}_+^k)_y \nabla_y u^k - \frac{1}{4}|(\vec{p}_+^k)_x \nabla_y u^k + (\vec{p}_+^k)_y \nabla_x u^k|^2\right) \\ =& -\frac{(m^k)^{-4}}{4}\left[|(\vec{p}_+^k)_x \nabla_y u^k|^2 + |(\vec{p}_+^k)_y \nabla_x u^k|^2 - 2(\vec{p}_+^k)_x (\vec{p}_+^k)_y \nabla_x u^k \nabla_y u^k\right] \\ =& \frac{(m^k)^{-4}}{4}\left[|\vec{p}_+^k|^2 |\nabla u^k|^2 - |(\vec{p}_+^k)_x \nabla_x u^k + (\vec{p}_+^k)_y \nabla_y u^k|^2\right] \to 0, \end{aligned} \tag{44}$$

and

$$(m^k)^{-2}((\vec{p}_+^k)_x \nabla_x u^k + (\vec{p}_+^k)_y \nabla_y u^k) \to \max(|\nabla u^*|, \gamma)^{q-4} |\nabla u^*|^2 > 0. \tag{45}$$

From (44) and (45), we conclude that the minimal eigenvalue of the 2-by-2 block (43) without the factor $\chi_{\mathcal{A}^k}$ goes to zero as $k \to +\infty$. Since $(\chi_{\mathcal{A}^k})$ is uniformly bounded, the minimal eigenvalue of (43) goes to zero as $(u^k, \vec{p}^k) \to (u^*, \vec{p}^*)$, which completes the proof. □

The inner loop, i.e. Steps 2–7 in Algorithm 3.6, is studied next.

Lemma 3.9. *Assume that H_+^k and $R_{+,\varepsilon}^k$ are both positive definite, and*

$$0 < \|g^k\| < \sqrt{\frac{c(\lambda_{\min}(H_+^k))^3}{(\lambda_{\max}(R_{+,\varepsilon}^k))^2}}.$$

Then the sequence $(\beta_l^k, d_l^k)_{l\in\mathbb{N}}$ generated by Steps 2–7 of Algorithm 3.6 converges to some (β_*^k, d_*^k) satisfying the optimality characterization of the trust-region subproblem; see (34)–(36).

Proof. By our assumption, the definiteness condition (36) is automatically satisfied. In the case where Step 4 of Algorithm 3.6 is active, the inner iterations terminate with a modified σ^k such that the conditions (34)–(35) are satisfied. Hence, in what follows we assume that Step 4 is inactive all along the sequence $(\beta_l^k, d_l^k)_{l\in\mathbb{N}}$.

We define the function $\phi : [0, \beta_{\max}] \to \mathbb{R}$ by

$$\phi(\beta) = \beta + \frac{((H_+^k + \beta R_{+,\varepsilon}^k)^{-1} g^k)^\top R_{+,\varepsilon}^k (H_+^k + \beta R_{+,\varepsilon}^k)^{-1} g^k - (\sigma^k)^2}{2c}.$$

Then by eliminating d^k by $d^k = -(H_+^k + \beta^k R_{+,\varepsilon}^k)^{-1} g^k$ in step 5 of Algorithm 3.6, we have the update rule (steps 5-6)

$$\beta_{l+1}^k = \max\left(\min\left(\phi(\beta_l^k), \beta_{\max}\right), 0\right).$$

Note that ϕ is continuously differentiable, and its derivative is given by

$$\phi'(\beta) = 1 - \frac{1}{c}(g^k)^\top (H_+^k + \beta R_{+,\varepsilon}^k)^{-1} (R_{+,\varepsilon}^k (H_+^k + \beta R_{+,\varepsilon}^k)^{-1})^2 g^k.$$

It follows from our assumptions that

$$\begin{aligned}&\left|\frac{1}{c}(g^k)^\top (H_+^k + \beta R_{+,\varepsilon}^k)^{-1} (R_{+,\varepsilon}^k (H_+^k + \beta R_{+,\varepsilon}^k)^{-1})^2 g^k\right| \le \frac{(\lambda_{\max}(R_{+,\varepsilon}^k))^2 \|g^k\|^2}{c\lambda_{\min}(H_+^k + \beta R_{+,\varepsilon}^k)^3} \\ \le &\frac{(\lambda_{\max}(R_{+,\varepsilon}^k))^2 \|g^k\|^2}{c(\lambda_{\min}(H_+^k))^3} < 1.\end{aligned}$$

By the above inequality and the mean value theorem, there exists a constant $C \in (0,1)$ such that for any $\beta_1, \beta_2 \in [0, \beta_{\max}]$,

$$|\phi(\beta_1) - \phi(\beta_2)| \le |\beta_1 - \beta_2| \sup_{\beta \in [0,\beta_{\max}]} |\phi'(\beta)| \le C|\beta_1 - \beta_2|,$$

i.e. ϕ is a *contractive* mapping. As a consequence, the mapping $\beta \mapsto \max\left(\min\left(\phi(\beta), \beta_{\max}\right), 0\right)$ is also contractive. Thus by the Banach fixed-point theorem (see, e.g., [42, Theorem 1.A]), we have $\beta_l^k \to \beta_*^k$ as $l \to +\infty$ for some $\beta_*^k \in [0, \beta_{\max}]$. Accordingly, $d_l^k \to d_*^k = -(H_+^k + \beta_*^k R_{+,\varepsilon}^k)^{-1} g^k$ as $l \to +\infty$. This completes the proof. □

Now we are in the position to prove our local convergence result.

Theorem 3.10 (Local Convergence)**.** *Let (d^k) be generated by Algorithm 3.6, and let the sequence (u^k) converge to some u^* satisfying $\nabla f(u^*) = 0$. Assume that all elements in $\partial_B^2 f(u^*)$ are strictly positive definite. Then Algorithm 3.6 is locally superlinearly convergent, i.e. for sufficiently large k we have*

$$\|u^{k+1} - u^*\| = o(\|u^k - u^*\|) \quad \text{for } k \to \infty. \tag{46}$$

Proof. Throughout the proof we argue only for sufficiently large k. From our assumption that all elements of $\partial_B^2 f(u^*)$ are strictly positive definite, it follows that all elements in $\partial_B^2 f(u^k)$, including $\nabla_B^2 f(u^k)$, are strictly positive definite with uniformly bounded (in norm) inverses; see [29, 19, 26]. Furthermore, due to Lemma 3.8 we have that H_+^k is also strictly positive definite.

Since $R_{+,\varepsilon}^k \succ 0$ according to Lemma 3.8, we have

$$-(d^k)^\top g^k = (d^k)^\top H_+^k d^k + \beta^k (d^k)^\top R_{+,\varepsilon}^k d^k \ge \lambda_{\min}(H_+^k)\|d^k\|^2 \ge 0.$$

Letting $k \to +\infty$, we have $\|d^k\| \to 0$ since $\|g^k\| \to 0$ by Theorem 3.7.

Next, we show that $\lim_{k\to+\infty} \beta^k = 0$. From the semismoothness property (7) and Lemma 3.8, we have that as $k \to +\infty$,

$$\begin{aligned}&|(f(u^k) - f(u^k + d^k)) - (h^k(0) - h^k(d^k))| = \left|f(u^k + d^k) - f(u^k) - (g^k)^\top d^k - \frac{1}{2}(d^k)^\top H_+^k d^k\right| \\ \le &\left|f(u^k + d^k) - f(u^k) - (g^k)^\top d^k - \frac{1}{2}(d^k)^\top \nabla_B^2 f(u^k) d^k\right| + \left|\frac{1}{2}(d^k)^\top (\nabla_B^2 f(u^k) - H_+^k) d^k\right| \\ = &o(\|d^k\|^2).\end{aligned} \tag{47}$$

For sufficiently large k, all assumptions in Lemma 3.9 hold true. Therefore, we have that

$$(d^k)^\top R^k_{+,\varepsilon} d^k \le \nu^2 (\sigma^k)^2,$$

for some constant $\nu > 0$, and that d^k satisfies the Cauchy-point-based model reduction criterion (32). In fact, only the last case in (33) may occur. So as $k \to +\infty$, we have

$$\begin{aligned}
&h^k(0) - h^k(d^k) \ge C\|g^k\|^2 \min\left(\frac{\sigma^k}{\sqrt{(g^k)^\top R^k_{+,\varepsilon} g^k}}, \frac{1}{\lambda_{\max}(H^k_+)}\right) \\
\ge& C\|g^k\| \min\left(\frac{\|g^k\|\sqrt{(d^k)^\top R^k_{+,\varepsilon} d^k}}{\nu\sqrt{(g^k)^\top R^k_{+,\varepsilon} g^k}}, \frac{\|g^k\|}{\lambda_{\max}(H^k_+)}\right) \ge C\|g^k\| \min\left(\frac{\sqrt{\lambda_{\min}(R^k_{+,\varepsilon})}\|d^k\|}{\nu\sqrt{\lambda_{\max}(R^k_{+,\varepsilon})}}, \frac{\|g^k\|}{\lambda_{\max}(H^k_+)}\right) \\
\ge& C\lambda_{\min}(H^k_+) \min\left(\frac{\sqrt{\lambda_{\min}(R^k_{+,\varepsilon})}}{\nu\sqrt{\lambda_{\max}(R^k_{+,\varepsilon})}}, \frac{\lambda_{\min}(H^k_+)}{\lambda_{\max}(H^k_+)}\right)\|d^k\|^2. \qquad (48)
\end{aligned}$$

Combining (47) and (48), we have that as $k \to +\infty$

$$|\rho^k - 1| = \frac{|(f(u^k) - f(u^k + d^k)) - (h^k(0) - h^k(d^k))|}{|h^k(0) - h^k(d^k)|} \le o(1) \to 0.$$

Thus the sequence (σ^k) is uniformly bounded away from 0. Consequently, $\lim_{k\to+\infty} \beta^k = 0$, and the Dennis-Moré condition [27] is satisfied, i.e. as $k \to +\infty$,

$$\frac{\|(H^k_+ + \beta^k R^k_{+,\varepsilon})d^k - \nabla^2_B f(u^*)d^k\|}{\|d^k\|} \le \|H^k_+ - \nabla^2_B f(u^*)\| + \beta^k \lambda_{\max}(R^k_{+,\varepsilon}) \to 0,$$

as the sequence $(\lambda_{\max}(R^k_{+,\varepsilon}))$ is uniformly bounded.

It follows from the semismoothness property (7) that

$$\begin{aligned}
&f(u^k + d^k) - f(u^k) - \tau_1 \nabla f(u^k)^\top d^k = (1 - \tau_1)\nabla f(u^k)^\top d^k + \frac{1}{2}(d^k)^\top \nabla^2_B f(u^k) d^k + o(\|d^k\|^2) \\
=&(d^k)^\top [(\tau_1 - 1)(H^k_+ + \beta^k R^k_{+,\varepsilon}) + \frac{1}{2}\nabla^2_B f(u^k)]d^k + o(\|d^k\|^2) \\
=&(\tau_1 - \frac{1}{2})(d^k)^\top \nabla^2_B f(u^k) d^k + o(\|d^k\|^2) \le 0,
\end{aligned}$$

and

$$\begin{aligned}
&\nabla f(u^k + d^k)^\top d^k - \tau_2 \nabla f(u^k)^\top d^k = (d^k)^\top \nabla^2_B f(u^k) d^k + (1 - \tau_2)\nabla f(u^k)^\top d^k + o(\|d^k\|^2) \\
=&(d^k)^\top [(\tau_2 - 1)(H^k_+ + \beta^k R^k_{+,\varepsilon}) + \nabla^2_B f(u^k)]d^k + o(\|d^k\|^2) = \tau_2 (d^k)^\top \nabla^2_B f(u^k) d^k + o(\|d^k\|^2) \ge 0,
\end{aligned}$$

for sufficiently large k since $\|d^k\| \to 0$ as $k \to +\infty$. Hence the Wolfe-Powell conditions (40)–(41) are satisfied for $a^k = 1$, i.e. $u^{k+1} = u^k + d^k$, for all sufficiently large k.

Let $d^k_N := -\nabla^2_B f(u^k)^{-1} g^k$. Note that

$$\begin{aligned}
&\|d^k - d^k_N\| = \|\nabla^2_B f(u^k)^{-1}(\nabla^2_B f(u^k) d^k + g^k)\| \\
=&\|\nabla^2_B f(u^k)^{-1}(\nabla^2_B f(u^k) - (H^k_+ + \beta^k R^k_+))d^k\| \le \|\nabla^2_B f(u^k)^{-1}\| o(\|d^k\|) = o(\|d^k\|),
\end{aligned}$$

since $(\|\nabla_B^2 f(u^k)^{-1}\|)$ is uniformly bounded as $u^k \to u^*$ for $k \to +\infty$. As a consequence, we have

$$\|u^{k+1} - u^*\| = \|u^k + d^k - u^*\| \leq \|u^k + d_N^k - u^*\| + \|d^k - d_N^k\| = o(\|u^k - u^*\|).$$

We have used that $\|u^k + d_N^k - u^*\| = o(\|u^k - u^*\|)$ (see, e.g., [26, Theorem 8.5]), and that $\|d^k\| = O(\|u^k - u^*\|)$. From this we conclude that Algorithm 3.6 is locally superlinearly convergent. □

The assumption of Theorem 3.10 relates to second-order sufficient optimality conditions for smooth problems. Although such assumptions typically occur in the optimization literature (also in the context of nonsmooth problems), they are difficult to check in an algorithm.

4 Numerics

In this section we present numerical results obtained by our primal-dual Newton method. Throughout this section, Ω denotes the m-by-n pixel-domain, i.e. $\Omega = \{(i,j) \in \mathbb{Z}^2 : 1 \leq i \leq m,\ 1 \leq j \leq n\}$. We discretize the gradient operator by

$$(\nabla u)_{ij} = \begin{pmatrix} \dfrac{u_{i+1,j} - u_{i,j}}{\omega} & \dfrac{u_{i,j+1} - u_{i,j}}{\omega} \end{pmatrix}$$

with $\omega := \sqrt{1/|\Omega|}$, assuming a square domain. We set $u_{ij} = 0$ whenever $(i,j) \notin \Omega$. Unless otherwise specified, the following parameters are used in our experiments:

$$\begin{aligned} &q = 0.75,\ \lambda = \chi_\Omega,\ \mu = 0,\ \gamma = 0.1,\ \beta_{\max} = 1.2 - q,\ c = 1,\ \varepsilon = 10^{-4}\alpha, \\ &\rho_1 = 0.25,\ \rho_2 = 0.75,\ \kappa_1 = 0.25,\ \kappa_2 = 2,\ \tau_1 = 0.1,\ \tau_2 = 0.9,\ \epsilon_d = 10^{-8}. \end{aligned} \tag{49}$$

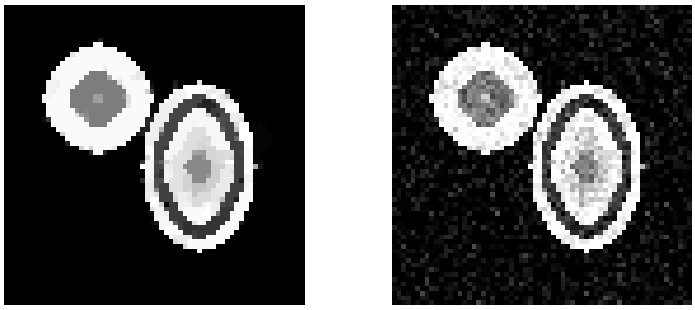
The trust-region subproblem (26)–(27) is solved only approximately. In fact, from our numerical experience it appears that only one (inner) iteration on β^k is most efficient in practice, i.e. Step 7 of Algorithm 3.6 is dismissed. The (outer) Newton iteration is terminated once the residual norm $\|\nabla f(u^k)\|$, see formula (5), has been reduced by a factor of 10^{-7}. All experiments are performed under MATLAB R2009b on a 2.66 GHz Intel Core Laptop with 4 GB RAM.

4.1 Test on "Two Circles" image

The 64-by-64 image "Two Circles" in [34] is used as our first test example, see Figure 1(a), in the context of a denoising problem, i.e. $K = I$. This image is corrupted by white Gaussian noise of zero mean and 0.1 standard deviation as shown in Figure 1(b). We choose the regularization parameter $\alpha = 2 \times 10^{-3}$ in the experiments.

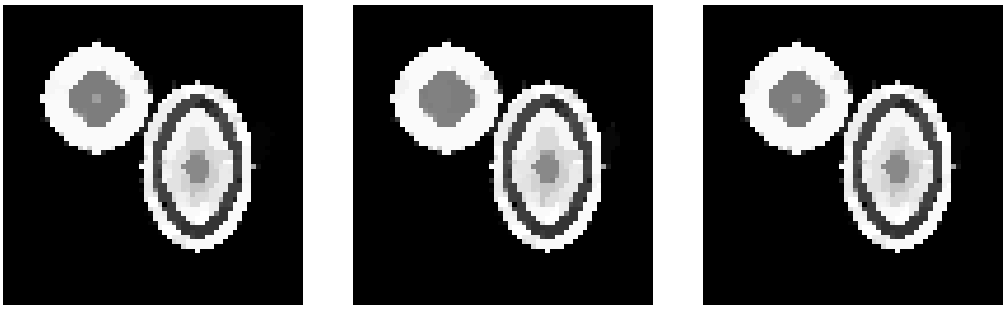
Dependence on initial guess. Three different choices of initial guesses are considered, namely the observed data, the zero vector, and a randomly chosen initial guess. The corresponding restored images are displayed in Figure 2, and the corresponding statistics are given in Table 1. We observe that the convergence behavior is rather insensitive with respect to the choice of the initial guess, in terms of both convergence speed and restoration quality. Due to the nonconvex nature of the problem, our iterative algorithm is expected to terminate at a local minimizer. In our experiments, the qualities of the obtained local minimizers are almost equally good with respect to objective value and PSNR (peak signal noise ratio). In the sequel, we shall choose the observed data as our initial guess if not otherwise specified.

Dependence on Huber parameter γ. In the discrete variational model (4), the nondifferentiable TV^q-penalty term is locally smoothed by the Huber function φ_γ with Huber parameter γ. In Table 2, we show the results of numerical tests for four different choices of γ. It is observed



(a) Original image. (b) Degraded image.

Figure 1: "Two Circles" image.



(a) $u^0 = z$. (b) $u^0 = 0$. (c) Randomly chosen u^0.

Figure 2: Dependence on initial guess.

Initial guess	# Newton iter.	Objective value	PSNR
$u^0 = z$	47	41.9868	29.1755
$u^0 = 0$	48	42.0064	28.9662
random u^0	48	41.9973	29.1687

Table 1: Dependence on initial guess.

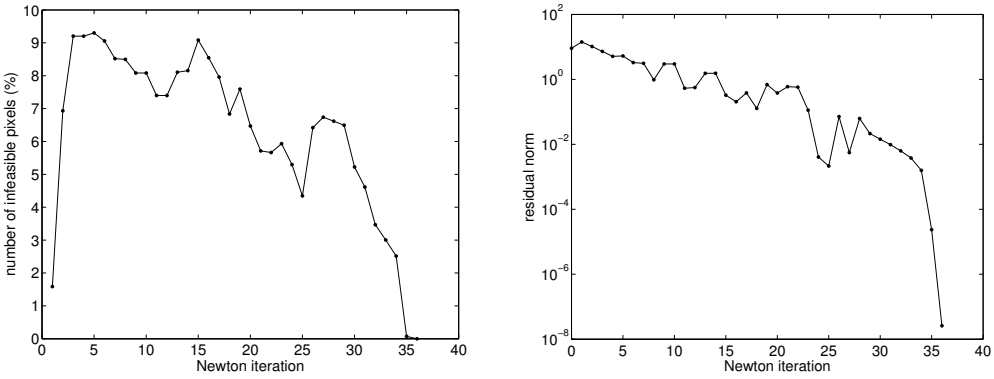
that the convergence behavior of our algorithm is insensitive with respect to the choice of γ, once γ is sufficiently small. Clearly, with respect to γ there is a tradeoff between the convergence speed and the restoration quality. As γ goes to zero, one obtains higher restoration quality but at the same time the computational cost increases. From our experience, $\gamma = 0.1$ is practically a reasonable choice in general.

Huber parameter γ	1e1	1e0	1e-1	1e-2
# Newton iter.	5	30	38	45
PSNR	25.5992	30.4193	31.0231	31.0321

Table 2: Dependence on Huber parameter γ.

Infeasible Newton iteration. We note that in contrast to the primal-dual algorithm (for $q = 1$) in, e.g., [3], our algorithm allows violations of the feasibility condition (21) during the Newton iterations. Yet towards the convergence of the algorithm we expect the feasibility condition (21) to hold true for $(u^k, \vec{p}^k)$, as established in the proof in Lemma 3.8. This is illustrated

in Figure 3. In plot (a) the number of infeasible pixels $(i,j) \in \Omega$, where $|(\nabla u^k)_{ij}| > \gamma$ and $|(\bar{p}^k)_{ij}||(\nabla u^k)_{ij}|^{1-q} \geq 1 + \epsilon_p$, is plotted for each Newton iteration. Here $\epsilon_p = 10^{-6}$ is introduced to compensate roundoff errors. In plot (b), the residual norm $\|\nabla f(u^k)\|$ is shown for each Newton iteration. It is observed that the number of infeasible pixels decreases to zero as the algorithm converges.



(a) Number of infeasible pixels. (b) Residual norm.

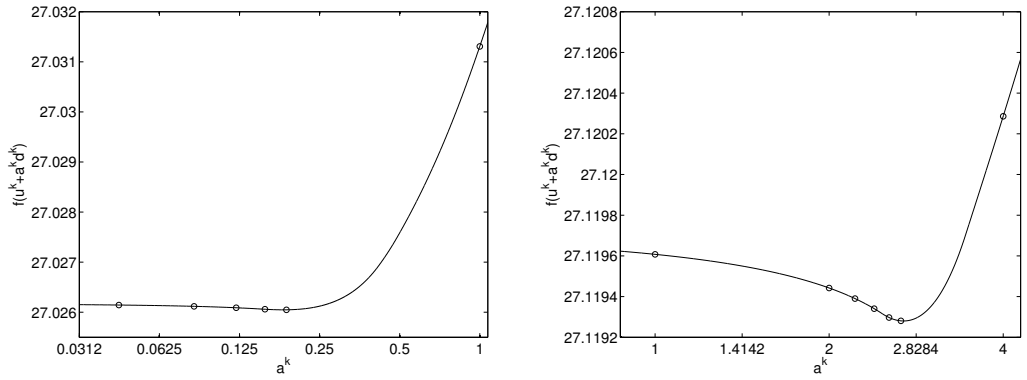
Figure 3: Infeasible Newton iteration.

Globalization by Wolfe-Powell line search. In Algorithm 3.6, after the search direction d^k is computed, the Wolfe-Powell line search is performed, where we aim to find an approximation of the solution to the one-dimensional problem $f_*^k := \min_{a^k>0} f(u^k + a^k d^k)$. Here we utilize an implementation according to [35, Algorithm 3.5–3.6]. Essentially, we begin with an initial step size a^k equal to 1. If either this step size is acceptable or the interval $[0, 1]$ contains an acceptable step size (which we refer to as Case 1), we directly proceed to the *zoom* procedure [35], which successively reduces the size of the interval until an acceptable step size is found. Otherwise (which we refer to as Case 2), we keep increasing a^k until we find either an acceptable step size or a *solution interval* that contains the acceptable step size. Once the solution interval is found, we proceed to the zoom procedure as in Case 1. In Table 3 and Figure 4, we provide an example of the Wolfe-Powell line search for each of the two cases: zoom only (see the upper part of Table 3), and first increase a^k and then zoom (see the lower part of Table 3). We remark that backtracking-only line search rules, e.g. the backtracking Armijo rule (see, e.g., [35]), do not perform well in our context. A backtracking-only line search rule would terminate with $a^k = 1$ in the example for Case 2, which poorly resolves the line search problem and therefore causes more (outer) Newton iterations.

Case 1: zoom only							
a^k	1	0.04	0.084	0.122	0.156	0.188	
$f(u^k + a^k d^k) - f_*^k$	5.26e-3	9.57e-5	6.97e-5	4.13e-5	1.14e-5	4.06e-7	
Case 2: increase a^k and then zoom							
a^k	1	2	4	2.217	2.393	2.539	2.662
$f(u^k + a^k d^k) - f_*^k$	3.27e-4	1.62e-4	1.01e-3	1.1e-4	6.05e-5	1.7e-5	5.89e-7

Table 3: Wolfe-Powell line search history.

Comparison with constantly regularized Newton method. In Section 3.1, we have seen the equivalence between the reweighted Euler-Lagrange approach and the constantly (R-)regularized



(a) Case 1: zoom only.

(b) Case 2: increase a^k and then zoom.

Figure 4: Wolfe-Powell line search. In each figure, the solid line is a plot of the function $a^k \mapsto f(u^k + a^k d^k)$, and the circled points are plots of the data in Table 3.

Newton method. Now we compare the constantly regularized Newton method and the adaptively regularized one proposed in this paper. For the constantly regularized Newton method we fix $\beta^k := \beta_{\max} = 1.2 - q$ (or equivalently $r := 0.8$). The history of the residual norm $\|\nabla f(u^k)\|$ is plotted in Figure 5 for both methods. It is observed that the constantly regularized Newton method is only linearly convergent as expected; see [4]. In contrast, our adaptively regularized Newton method attains a locally rapid superlinear convergence, indicated by a fast decrease in residual norm towards to the end of iterations.

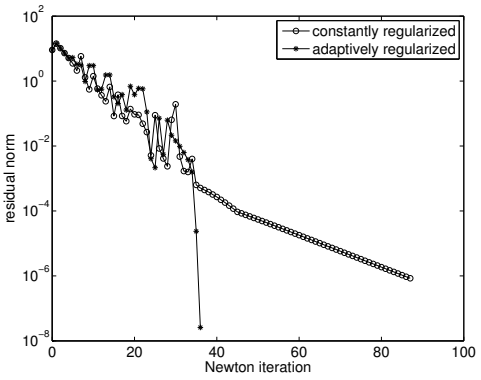


Figure 5: Comparison with constantly regularized Newton method. The residual norm history of the the constantly regularized Newton method is marked by circles, and the one of the adaptively regularized Newton method is the star-marked curve.

4.2 Test on "Shepp-Logan Phantom"

Our second testing image is the "Shepp-Logan Phantom" contaminated by white Gaussian noise of zero mean and 0.05 standard deviation.

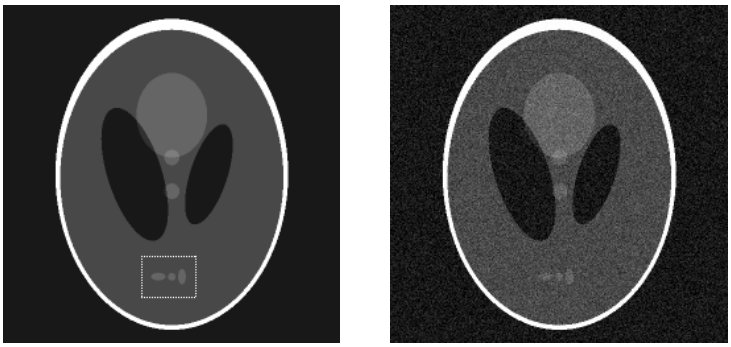
Dependence on image resolution. Our algorithm is implemented to restore the "Shepp-Logan Phantom" images of different resolutions, namely 64-by-64, 128-by-128, and 256-by-256. The regularization parameter $\alpha = 4 \times 10^{-4}$ is fixed in all three restorations. The algorithm terminates after 60, 57, and 58 Newton iterations for restoring images of resolutions 64-by-64,

q	PSNR	$\lvert\mathcal{A}_\gamma(\hat{u}_q)\rvert/\lvert\Omega\rvert$	$f_q(\hat{u}_q)$	$f_q(\hat{u}_1)$
1	37.5709	0.196	-	-
0.75	41.0039	0.0578	134.3021	137.6719
0.5	41.0191	0.0531	116.6721	125.6655
0.25	39.9259	0.0503	113.9953	133.6758

Table 4: Performances of TV^q-models.

128-by-128, and 256-by-256, respectively. This indicates that our algorithm is stable with the image resolution.

Performance of the TV^q-model for different q-values. We compare the performance of our TV^q-model for q =1, 0.75, 0.5, and 0.25 for denoising the 256-by-256 Shepp-Logan Phantom; see Figure 6. For each q, the parameter α is manually chosen in order to obtain the best PSNR value. The restored images $\hat{u}_q$ are shown in Figure 7 for each q. It is observed from the rescaled zoom-in views that the TV^q-models provide better contrast in restoration as q becomes smaller. The performance of the TV^q-model for different q-choices is also compared quantitatively; see Table 4. The PSNR values of the restoration from the nonconvex TV^q-models (with $0 < q < 1$) are significantly higher than those from the TV-model (with $q = 1$). In addition, we measure the gradient sparsity by $|\mathcal{A}_\gamma(\hat{u}_q)|/|\Omega|$, where $\mathcal{A}_\gamma(u) := \{(i,j) \in \Omega : |(\nabla u)_{ij}| > \gamma\}$. It is observed that in comparison with the conventional TV-model, the sparsity is well enhanced under the nonconvex TV^q-regularizations. Note that the gradient sparsity of the true image is 0.0503. Furthermore, we compare each solution of the TV^q-model, denoted by $\hat{u}_q$, with the solution of the TV-model, denoted by $\hat{u}_1$, by plugging both solutions into the objective of the nonconvex TV^q-problem. We find that $\hat{u}_1$ is far from being optimal with respect to the objective value. This phenomenon is more distinct as q becomes smaller.

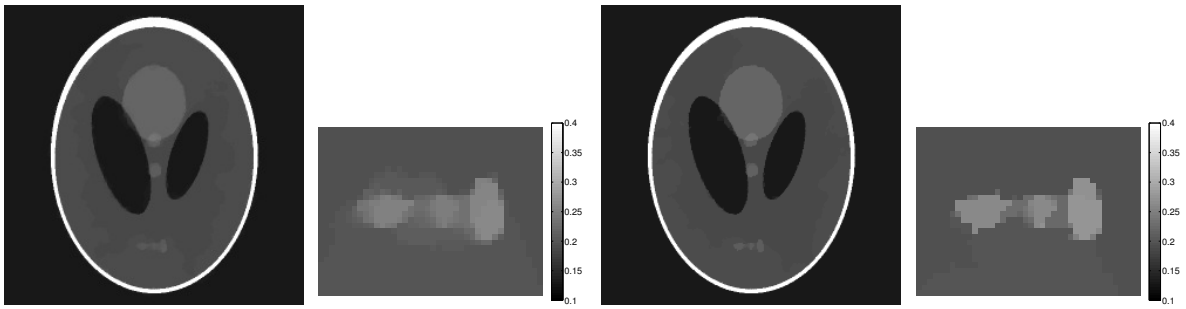


(a) Original image. (b) Degraded image.

Figure 6: 256-by-256 Shepp-Logan Phantom. The dash-boxed region in (a) is zoomed in for restored images in forthcoming Figure 7.

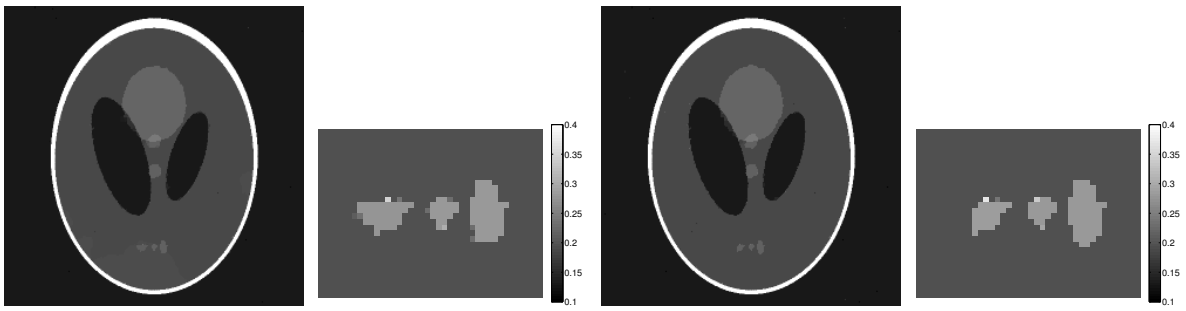
4.3 Test on simultaneously blurred and noisy images

Now we apply our algorithm for simultaneously blurring and denoising the text image "TV^q-model" (see Figure 8) and the image "Cameraman" (see Figure 9). For both images, the blurring



(a) $q = 1,\ \alpha = 3 \times 10^{-4}$.

(b) $q = 0.75,\ \alpha = 4 \times 10^{-4}$.



(c) $q = 0.5,\ \alpha = 6 \times 10^{-4}$.

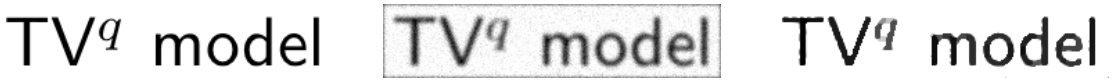
(d) $q = 0.25,\ \alpha = 1.2 \times 10^{-3}$.

Figure 7: Restoration via TV^q-models. In each group, the left figure is the restored image $\hat{u}_q$, and the right figure is the rescaled zoom-in of the restored image on the dash-boxed region of Figure 6(a).

kernel is chosen to be a two-dimensional truncated Gaussian kernel, i.e.

$$(Ku)_{ij} = \sum_{|i'|\le 3,\ |j'|\le 3} \exp\left(-\frac{|i'|^2 + |j'|^2}{2|\sigma_K|^2}\right) u_{i-i',j-j'}.$$

with $\sigma_K = 1.5$. After blurring, white Gaussian noise of zero mean and 0.05 standard deviation is added. The restored images are shown in the corresponding figures. It is visually observed that the nonconvex TV^q-model promotes piecewise constant images in the restoration results. This is expected because $q \to 0$ results in the problem of minimizing the support of the image intensity.



(a) Original image. (b) Observed image. (c) Restored image.

Figure 8: "TV^q-model" text image: restoration with $\alpha = 5 \times 10^{-4},\ \mu = 10^{-4}\alpha$.

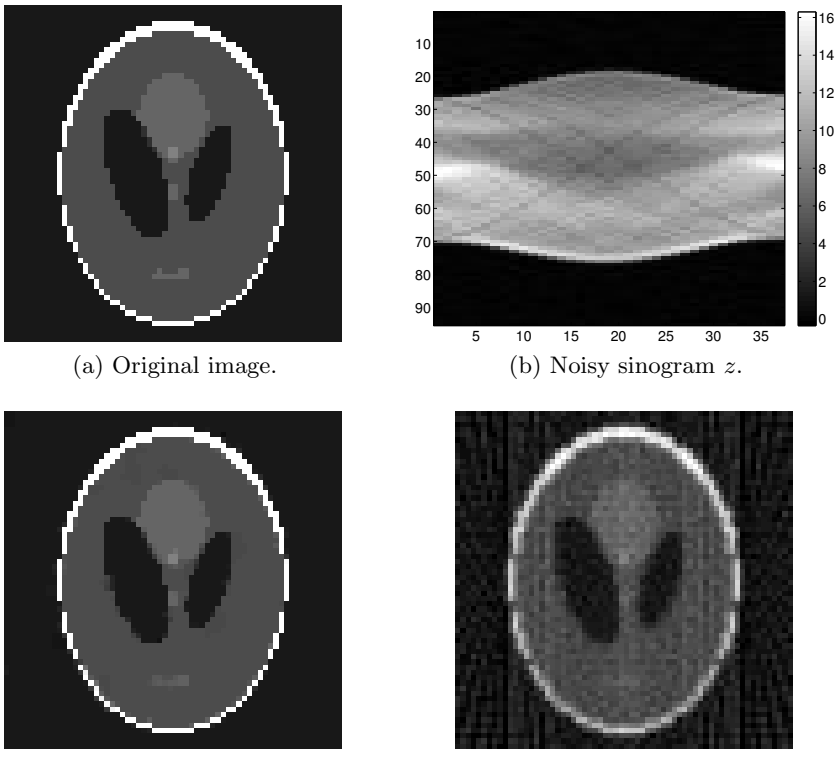
4.4 Test on tomographic data

Our algorithm can be applied to restoring images from possibly noisy tomographic data. In Figure 10, the 64-by-64 Shepp-Logan Phantom (plot (a)) is used as test example. The tomographic data, or the *sinogram*, of size 95-by-37 is obtained from applying the 2D Radon transform [28]

(a) Original image. (b) Observed image. (c) Restored image.

Figure 9: "Cameraman" image: restoration with $\alpha = 2 \times 10^{-4},\ \mu = 10^{-4}\alpha$.

along the angles 0, 5, 10, ..., 180 degrees. Then white Gaussian noise of zero mean and 0.1 standard deviation is added to the sinogram. The resulting observed data is shown in plot (b). Note that the matrix K is the discrete Radon transform of size 3515-by-4096. We choose $u^0 = 0$ as our initial guess. The Newton algorithm is implemented with $\alpha = 8 \times 10^{-3}$ and $\mu = 10^{-4}\alpha$. The restored image is displayed in (c). We compare this result with the restoration obtained by the frequently used filtered backprojection [28], which is shown in (d). We find that the restoration in (c) is almost perfect, whereas the one in (d) contains considerable artifacts.



(a) Original image. (b) Noisy sinogram z.

(c) Restored by TV^q-model. (d) Restored by filtered backprojection.

Figure 10: Restoration from Radon transformed data.

5 TV^q-models in function space: a partial result and a warning example

Often one aims at studying the variational problem in its original function space setting. In our context, the infinite dimensional version associated with (1) reads

$$\inf_{u\in H_0^1(\Omega)} f(u) = \int_\Omega F(x,u,\nabla u)dx = \int_\Omega \left(\frac{\mu}{2}|\nabla u|^2 + \frac{\alpha}{q}|\nabla u|^q + \frac{\lambda}{2}|Ku-z|^2\right) dx, \qquad (50)$$

where $\alpha > 0$, $0 < q < 1$, $0 < \mu \ll \alpha$, $z \in L^2(\Omega)$, $\lambda \in L^\infty(\Omega)$ such that $\lambda(x) > 0$ a.e. on Ω and $\int_\Omega \lambda(x)dx = \text{Area}(\Omega)$, and $K \in \mathcal{L}(L^2(\Omega))$, i.e. it is a linear and continuous operator from $L^2(\Omega)$ to $L^2(\Omega)$, such that $K\chi_\Omega \neq 0$.

Obviously, f is coercive, i.e. $f(u) \to \infty$ as $\|u\|_{H_0^1(\Omega)} \to \infty$. Note that the integrand $F(x,u,\xi)$ is nonconvex in ξ. It is known [1, Theorem 2.1.3] that f is weakly lower semicontinuous on $H_0^1(\Omega)$ if and only if F is convex in ξ. As a consequence, $f : H_0^1(\Omega) \to \mathbb{R}$ in (50) is not weakly lower semicontinuous, a usual prerequisite for proving existence of minimizers. Hence, the direct methods of the calculus of variations cannot be applied.

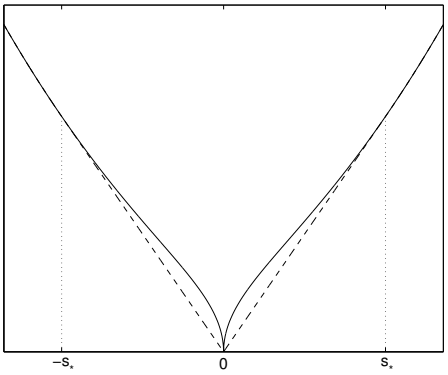


Figure 11: The function $s \mapsto \frac{\mu}{2}|s|^2 + \frac{\alpha}{q}|s|^q$ (in solid line), and its envelope (in dashed line).

Nevertheless, there exists a minimizer for a relaxed version of the problem (50). For this purpose, we construct a relaxed functional by taking the bipolar [18] of $F(x,u,\xi)$ with respect to ξ, i.e.

$$\bar{F}(x,u,\xi) := F^{**}(x,u,\xi) = \begin{cases} (\alpha s_*^{q-1} + \mu s_*)|\xi| + \frac{\lambda}{2}|Ku-z|^2, & \text{if } |\xi| < s_*, \\ F(x,u,\xi), & \text{if } |\xi| \geq s_*, \end{cases}$$

where the convexity threshold s_* is given by

$$s_*(q,\mu,\alpha) := \left(\frac{\alpha(1/q-1)}{\mu/2}\right)^{1/(2-q)}.$$

We define $\bar{f}(u) := \int_\Omega \bar{F}(x,u,\nabla u)dx$. It turns out that $\bar{f}$ represents the weakly lower semicontinuous envelope of $f(u)$ under the weak $H_0^1(\Omega)$-topology [30, pp. 34], i.e.

$$\bar{f}(u) = \sup\{\tilde{f}(u) : \tilde{f}(u) \leq f(u)\ \forall u \in H_0^1(\Omega),\ \tilde{f} \text{ is weakly lower semicontinuous on } H_0^1(\Omega)\}.$$

Concerning the existence of minimizers in $H_0^1(\Omega)$ for $\bar{f}$ and their relations to f, we provide the following two theorems, which can be found in [1]; see Theorem 2.1.5 and Theorem 2.1.6 in this reference.

Theorem 5.1 (Characterization). *The relaxed functional $\bar{f}$ is characterized by the following properties:*

1. *For every sequence (u^k) that weakly converges to u in $H_0^1(\Omega)$, we have $\bar{f}(u) \leq \liminf f(u^k)$.*
2. *For every $u \in H_0^1(\Omega)$, there exists a sequence (u^k) that weakly converges to u in $H_0^1(\Omega)$ and $\bar{f}(u) \geq \limsup f(u^k)$.*

Theorem 5.2 (Main properties). *Suppose $f : H_0^1(\Omega) \to \mathbb{R}$ is coercive. Then the following properties hold:*

1. *$\bar{f}$ is coercive and weakly lower semicontinuous on $H_0^1(\Omega)$.*
2. *$\bar{f}$ has a minimizer in $H_0^1(\Omega)$.*
3. *$\min_{u \in H_0^1(\Omega)} \bar{f}(u) = \inf_{u \in H_0^1(\Omega)} f(u)$.*
4. *Every accumulation point of an infimizing sequence for f is a minimizer for $\bar{f}$ under the weak $H_0^1(\Omega)$-topology.*
5. *Every minimizer for $\bar{f}$ is the limit of an infimizing sequence for f under the weak $H_0^1(\Omega)$-topology.*

In a nutshell, we associate the original nonconvex problem, for which no minimizer may exist, with a relaxed problem, which admits the existence of a minimizer. However, the minimizer of the relaxed problem may be far from optimal for the original problem with respect to the objective value. This is illustrated by the following example; see [1, pp. 36] for a related example. Note that this example shares the nonconvexity in the ξ-variable with our TV^q-model, but otherwise has a different structure in the term involving the derivative.

Example 5.3. Let $F(x, u, \xi) := u^2 + (|\xi| - 1)^2$. The Bolza problem is

$$\inf\{f(u) := \int_0^1 \left((|u'| - 1)^2 + u^2\right) dx : u \in H_0^1(0,1)\}.$$

The integrand $F(x, u, \xi)$ is nonconvex in ξ. We claim that $\inf f = 0$. Indeed, consider the sequence (u^k) defined by

$$u^k(x) = \begin{cases} x - \frac{l}{k} & \text{if } x \in \left(\frac{l}{k}, \frac{2l+1}{2k}\right) \\ -x + \frac{l+1}{k} & \text{if } x \in \left(\frac{2l+1}{2k}, \frac{l+1}{k}\right) \end{cases}, \text{ for } l = 0, 1, 2, ..., n-1.$$

Then $u^k \in W_0^{1,\infty}(0,1)$ such that $0 \leq u^k(x) \leq \frac{1}{2k}, \forall x \in (0,1)$, and $|(u^k)'(x)| = 1$ a.e. in $(0,1)$. Therefore, we have $0 \leq \inf_u f(u) \leq f(u^k) \leq \frac{1}{4k^2}$. Thus the claim is verified. However, there exists no function $u \in H_0^1(0,1)$ such that $f(u) = 0$. Hence there exists no solution to the Bolza problem.

Nevertheless, the Bolza problem can be relaxed, using the weakly lower semicontinuous envelope of f, as follows:

$$\min\{\bar{f}(u) := \int_0^1 \left((\max(|u'| - 1, 0))^2 + u^2\right) dx : u \in H_0^1(0,1)\}.$$

The relaxed problem admits a unique solution $u^* = 0$. Obviously the set $\{x \in (0,1) : |(u^*)'(x)| < 1\}$ is of positive Lebesgue measure; otherwise u^* would be a minimizer for $\inf_u f(u)$. Finally, we notice that $f(u^*) = 1$. This indicates that u^* is far from optimal for the original problem.

6 Conclusion

Nonconvex regularization still represents a significant analytical as well as numerical challenge, but appears to yield results superior to the ones obtained by ℓ_1-regularization. The latter is typically chosen as the "closest" convex relaxation of ℓ_0-norm problems. On the numerical side of this work, the proposed semismooth Newton based solver combines a trust-region technique for automatic stabilization whenever required due to the involved nonconvexity. In this context and from our point of view, the result contained in Lemma 3.3 offers an interesting extension of the available trust-region literature and is a consequence of the structural specifics of the variational model under consideration. Nevertheless it appears to carry the potential for being extended to more general problem classes. The global as well as locally superlinear convergence of our overall algorithm is obtained without further conditions on the problem and its solutions with respect to null space properties of the involved operators (aside from the minimal requirement for guaranteeing existence of a solution) or specific sparsity properties of the solution. Our numerical tests support our theoretical findings and show the effectiveness of our algorithm also in cases where K represents the Radon transform, which typically occurs in computerized tomography (CT).

In contrast to the classical convex total variation regularization according to Rudin, Osher and Fatemi [37], the function space analysis of the TV^q-model is troubled by the nonconvexity. This fact prevents existence of solutions in function space, in general. It is also demonstrated that the usual way of computing envelops of the objective under consideration and then analyzing the enveloped problem guarantees existence, but the solution to the enveloped problem may be far away from optimal for the original problem.

References

[1] G. Aubert and P. Kornprobst, *Mathematical Problems in Image Processing*, Springer, New York, 2002.

[2] E. J. Candès and T. Tao, *Near optimal signal recovery from random projections: Universal encoding strategies?*, IEEE Trans. Inform. Theory, 52 (2006), pp. 5406–5425.

[3] T. F. Chan, G. H. Golub, and P. Mulet, *A nonlinear primal-dual method for total variation-based image restoration*, SIAM J. Sci. Comput., 20 (1999), pp. 1964–1977.

[4] T. F. Chan and P. Mulet, *On the convergence of the lagged diffusivity fixed point method in total variation image restoration*, SIAM J. Numer. Anal., 36 (1999), pp. 354–367.

[5] R. Chartrand, *Exact reconstruction of sparse signals via nonconvex minimization*, IEEE Signal Process. Lett., 14 (2007), pp. 707–710.

[6] ——, *Fast algorithms for nonconvex compressive sensing: MRI reconstruction from very few data*, in IEEE International Symposium on Biomedical Imaging, 2009.

[7] R. Chartrand and W. Yin, *Iteratively reweighted algorithms for compressive sensing*, in 33rd IEEE International Conference on Acoustics, Speech and Signal Processing, 2008.

[8] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput., 20 (1998), pp. 33–61.

[9] X. Chen and W. Zhou, *Smoothing nonlinear conjugate gradient method for image restoration using nonsmooth nonconvex minimization*, SIAM J. Imaging Sciences, 3 (2010), pp. 765–790.

[10] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, 1983.

[11] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust-Region Methods*, SIAM, Philadelphia, 2000.

[12] I. Daubechies, R. DeVore, M. Fornasier, and C. Gunturk, *Iteratively reweighted least squares minimization for sparse recovery*, Comm. Pure Appl. Math., 63 (2010), pp. 1–38.

[13] Y. Dong, M. Hintermüller, and M. Neri, *An efficient primal-dual method for L^1TV image restoration*, SIAM J. Imaging Sciences, 2 (2009), pp. 1168–1189.

[14] Y. Dong, M. Hintermüller, and M. M. Rincon-Camacho, *Automated regularization parameter selection in a multi-scale total variation model for image restoration*, J. Math. Imaging Vis., 40 (2011), pp. 82–104.

[15] ——, *A multi-scale vectorial L^τ-TV framework for color image restoration*, Int. J. Comput. Vis., 92 (2011), pp. 296–307.

[16] D. L. Donoho and B. F. Logan, *Signal recovery and the large sieve*, SIAM J. Appl. Math., 52 (1992), pp. 577–591.

[17] D. L. Donoho and P. B. Stark, *Uncertainty principles and signal recovery*, SIAM J. Appl. Math., 49 (1989), pp. 906–931.

[18] I. Ekeland and R. Témam, *Convex Analysis and Variational Problems*, SIAM, Philadelphia, 1999.

[19] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems: Vol. 2*, Springer, New York, 2003.

[20] D. Ge, X. Jiang, and Y. Ye, *A note on the complexity of L_p minimization*, Math. Program., Ser. B, 129 (2011), pp. 285–299.

[21] T. Goldstein and S. Osher, *The split Bregman method for L1-regularized problems*, SIAM J. Imaging Sciences, 2 (2009), pp. 323–343.

[22] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth*, SIAM J. Optim., 13 (2003), pp. 865–888.

[23] M. Hintermüller and K. Kunisch, *Total bounded variation regularization as a bilaterally constrained optimization problem*, SIAM J. Appl. Math., 64 (2004), pp. 1311–1333.

[24] M. Hintermüller and G. Stadler, *An infeasible primal-dual algorithm for total bounded variation-based inf-convolution-type image restoration*, SIAM J. Sci. Comput., 28 (2006), pp. 1–23.

[25] P. J. Huber, *Robust estimation of a location parameter*, Ann. Math. Statist., 53 (1964), pp. 73–101.

[26] K. Ito and K. Kunisch, *Lagrange Multiplier Approach to Variational Problems and Applications*, SIAM, Philadelphia, 2008.

[27] J. E. Dennis Jr. and J. J. Moré, *Quasi-Newton methods, motivation and theory*, SIAM Rev., 19 (1977), pp. 46–89.

[28] A. C. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging*, SIAM, Philadelphia, 2001.

[29] D. Klatte and B. Kummer, *Nonsmooth Equations in Optimization: Regularity, Calculus, Methods and Applications*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002.

[30] G. Dal Maso, *An Introduction to* Γ*-Convergence*, Birkhäuser, Boston, 1993.

[31] B. K. Natarajan, *Sparse approximate solutions to linear systems*, SIAM J. Comput., 24 (1995), pp. 227–234.

[32] M. Nikolova, *Minimizers of cost-functions involving nonsmooth data-fidelity terms: Application to the processing of outliers*, SIAM J. Numer. Anal., 40 (2002), pp. 965–994.

[33] M. Nikolova, M. K. Ng, and C.-P. Tam, *Fast nonconvex nonsmooth minimization methods for image restoration and reconstruction*, IEEE Trans. Image Process., 19 (2010), pp. 3073–3088.

[34] M. Nikolova, M. K. Ng, S. Zhang, and W.-K. Ching, *Efficient reconstruction of piecewise constant images using nonsmooth nonconvex minimization*, SIAM J. Imaging Sciences, 1 (2008), pp. 2–25.

[35] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, New York, 2nd ed., 2006.

[36] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Program., 58 (1993), pp. 353–367.

[37] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D, 60 (1992), pp. 259–268.

[38] J. A. Tropp and S. J. Wright, *Computational methods for sparse solution of linear inverse problems*, in Proceedings of the IEEE, 2010.

[39] C. R. Vogel, *Computational Methods for Inverse Problems*, SIAM, Philadelphia, 2002.

[40] C. R. Vogel and M. E. Oman, *Iterative methods for total variation denoising*, SIAM J. Sci. Comput., 17 (1996), pp. 227–238.

[41] S. J. Wright, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, 1997.

[42] E. Zeidler, *Nonlinear Functional Analysis and Its Applications, Pt. 1: Fixed-point Theorems*, Springer, New York, 1986.