SpezialForschungsBereich F 32

Karl–Franzens Universität Graz
Technische Universität Graz
Medizinische Universität Graz

A Bilevel Optimization Approach for Parameter Learning in Variational Models

K. Kunisch T. Pock

SFB-Report No. 2013-007 May 2013

A–8010 GRAZ, HEINRICHSTRASSE 36, AUSTRIA

Supported by the
Austrian Science Fund (FWF)

SFB sponsors:

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

SIAM J. IMAGING SCIENCES
Vol. 0, No. 0, pp. 000–000

A Bilevel Optimization Approach for Parameter Learning in Variational Models*

Karl Kunisch† and Thomas Pock‡

Abstract. In this work we consider the problem of parameter learning for variational image denoising models. The learning problem is formulated as a bilevel optimization problem, where the lower-level problem is given by the variational model and the higher-level problem is expressed by means of a loss function that penalizes errors between the solution of the lower-level problem and the ground truth data. We consider a class of image denoising models incorporating ℓ_p-norm–based analysis priors using a fixed set of linear operators. We devise semismooth Newton methods for solving the resulting nonsmooth bilevel optimization problems and show that the optimized image denoising models can achieve state-of-the-art performance.

Key words. regularization parameter, image denoising, learning theory, nondifferentiable optimization, bilevel optimization, semismooth Newton algorithm

AMS subject classifications. 49J52, 49N45, 68U10

DOI. 10.1137/120882706

1. Introduction.

Variational approaches have had great success in solving inverse problems in imaging, such as image restoration, optical flow, and stereo vision. The fundamental principle behind these approaches is to devise the solution of the inverse problem as the minimizer of an energy functional, which is designed such that its minimum-energy state reflects the characteristic properties of the solution. For example, popular priors assume that the solution is piecewise constant or piecewise smooth.

Usually, variational models incorporate a number of free parameters. These parameters are used, for example, to trade off between regularization and data fidelity or to locally adapt the variational model to the input data. Selecting optimal parameters is by far not trivial. A possible procedure for determining these free parameters is to evaluate the performance of the variational model on some test data with a known optimal solution by performing an exhaustive search over a range of useful parameter settings. This is tedious and already becomes infeasible for more than two or three parameters.

In this work a systematic approach for the above procedure will be provided. We cast parameter selection as a learning problem. Given a certain variational model, the task consists in learning the parameters such that the variational model minimizes a certain loss functional on a training database. This naturally leads to a bilevel optimization problem of the following

*Received by the editors June 27, 2012; accepted for publication (in revised form) March 25, 2013; published electronically DATE.
http://www.siam.org/journals/siims/x-x/88270.html
†Institute of Mathematics and Scientific Computing, University of Graz, A-8010 Graz, Austria (karl.kunisch@uni-graz.at).
‡Institute for Computer Graphics and Vision, Graz University of Technology, A-8010 Graz, Austria (pock@icg.tugraz.at).

form:

$$\begin{cases} \min_{\vartheta \geq 0} \mathcal{E}(x(\vartheta)) \\ \text{subject to } x(\vartheta) \in \arg\min_{x} \mathcal{F}(x, \vartheta). \end{cases} \tag{1.1}$$

The bilevel problem consists of a lower-level optimization whose solution $x(\vartheta)$ is an argument of the higher-level minimization problem. The aim of the bilevel problem is then to find a parameter vector ϑ such that $\mathcal{E}(x(\vartheta))$ attains a minimum value.

Concerning the choice of regularization parameters the literature typically distinguishes between a posteriori and a priori parameter rules, as well as error-free parameter choice rules; see, e.g., [10, 12] and the references cited therein. The discrepancy principle is a prominent example of an a posteriori rule, where the regularization parameter is determined such that the data fidelity term at the optimum equals the size of the noise level. Here we require knowledge of the noise level as well as the noisy data. A priori rules determine the regularization parameter solely from knowledge of the noise level. The class of parameter-free methods includes generalized cross validation and balancing principles between the error in the fidelity and the regularization terms. Most of the work on parameter choice techniques addresses the case of a single scalar parameter.

Bilevel optimization problems are an active research area in their own right; see, e.g., [2] and the references cited therein. Here we analyze only the specific bilevel problem (1.1) to the extent that is required to propose and investigate numerical methods for its solution. In this work the functional $\mathcal{E}$ of the upper-level problem will be smooth, while for the lower-level problem we distinguish between a smooth quadratic case and the nonsmooth ℓ_1 and $\ell_{\frac{1}{2}}$ cases.

For the application of image restoration, bilevel optimization has been used by Tappen et al. in [29, 28, 27] to learn the parameters of different Markov random field models. In particular, they showed that bilevel optimization provides an effective learning method, as it overcomes the typical problems of classical probabilistic learning methods that require one to compute the partition function of the underlying probability density function. However, while Tappen et al. used gradient methods for learning that do not come along with any convergence guarantees, we propose fast Newton methods that come along with locally superlinear convergence. It will turn out that our proposed Newton algorithms not only provide an effective learning framework but also lead to image restoration results superior to those reported in [27]. We attribute this fact mainly to the ability of our proposed algorithms to be more successful in finding a (local) minimizer of the bilevel optimization problems than the gradient methods used in [27]. In [24], a bilevel learning approach was proposed for sparse analysis prior learning using an ℓ_1 model. The approach is similar to that in [27], as it uses implicit differentiation to compute the gradient of the higher-level problem with respect to the learning parameters.

Let us give a brief summary of the contents of the following sections. In section 2 we present the precise problem statement and provide some preliminaries. The smooth case with a single prior as well as multiple priors is analyzed in section 3. We investigate aspects of the geometry of the value functional $\mathcal{E}$ and develop a Newton algorithm for the solution of the inequality constrained problem (1.1). Section 4 is devoted to the existence of (1.1) and the derivation of an optimality condition by means of a regularization procedure for the case

when the lower-level problem is nonsmooth. The regularized problems are semismooth, and thus we propose a semismooth Newton algorithm for their solution. Numerical experiments for a wide variety of priors and for images of different qualitative features are presented in section 5.

2. Preliminaries.

In this work we put our emphasis on the following class of problems:

$$\begin{cases} \min\limits_{\vartheta \geq 0} \mathcal{E}(x(\vartheta)) = \|x(\vartheta) - g\|_2^2 \\ \text{subject to } x(\vartheta) = \arg\min\limits_{x} \mathcal{F}(x, \vartheta) = \dfrac{1}{p} \displaystyle\sum_{k=1}^{q} \vartheta_k \|K_k x\|_p^p + \frac{1}{2}\|x - f\|_2^2. \end{cases} \tag{2.1}$$

The lower-level optimization problems $\mathcal{F}(x, \vartheta)$ consists of a data and a regularization term. The data term penalizes the squared ℓ_2-norm of the discrepancy between the noisy image $f \in \mathbb{R}^n$ and the unknown image $x \in \mathbb{R}^n$. The regularization term is a sum of $q \geq 1$ so-called analysis-based priors (see, e.g., [31]), penalizing the ℓ_p^p-norms

$$\|K_k x\|_p^p = \sum_{i=1}^{n} |(K_k x)_i|^p$$

of the result of applying linear operators $K_k \in \mathbb{R}^{m \times n}$, $1 \leq k \leq q$, to x. We shall consider primarily the cases $p \in \{1, 2\}$, and in numerical experiments also $p = \frac{1}{2}$. The importance of the priors $\|K_k x\|_p^p$, $1 \leq k \leq q$ are weighted by parameters $\vartheta_k \geq 0$, which are assembled in a parameter vector $\vartheta = (\vartheta_1, \ldots, \vartheta_q)$. Observe that (2.1) admits a unique solution, which follows from the strict convexity of the data term. The case of a nonstrictly convex data term, for example a data term of the form $\frac{1}{2}\|Ax - f\|_2^2$, where A is a singular linear operator, is left for future work.

The higher-level optimization problem $\mathcal{E}(x(\vartheta))$ penalizes the discrepancy between the minimizer of the lower-level optimization problem $x(\vartheta)$ and given ground truth data $g \in \mathbb{R}^n$ by means of the squared 2-norm. In some situations we will eliminate x, which leads to a reduced single-level optimization problem $\mathcal{E}(\vartheta)$, as opposed to the bilevel optimization problem $\mathcal{E}(x(\vartheta))$.

We frequently make use of a standard inner product on $\mathbb{R}^n$ denoted by $\langle \cdot, \cdot \rangle$, which induces the 2-norm $\|\cdot\|_2 = \langle \cdot, \cdot \rangle^{\frac{1}{2}}$. We further denote by $\ker(K) = \{x \in \mathbb{R}^n : Kx = 0\}$ the kernel of K and by $\operatorname{ran}(K) = \{Kx : x \in \mathbb{R}^n\}$ the range or column space of K. The operation max on a vector $x \in \mathbb{R}^n$ is understood to operate elementwise, i.e.,

$$\max(0, x) = (\max(0, x_1), \ldots, \max(0, x_n)).$$

To obtain some insight into the cost functional $\mathcal{E}$ associated with (1.1) let us investigate the scalar-valued case, i.e., $x, f, g \in \mathbb{R}$, $q = 1$, and $K_1 = 1$. For $p = 2$ and by combining the lower-level problem with the higher-level problem, we arrive at the single level problem

$$\min_{\vartheta \geq 0} \mathcal{E}_{\ell_2}(\vartheta) = \left(\frac{f}{1 + \vartheta} - g \right)^2.$$

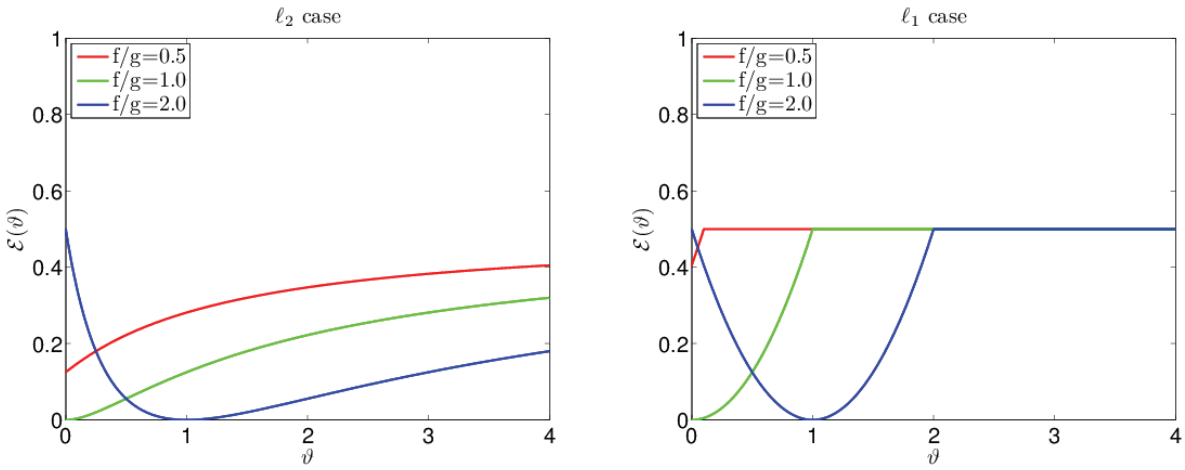


Figure 1. *Shape of the reduced single-level problems for the ℓ_2 and ℓ_1 cases.*

We plot its graph for the scalar-valued case in Figure 1(a) for various choices of the ratio f/g. It is easy to show that all sublevel sets of $\mathcal{E}_{\ell_2}(\vartheta)$ are convex and hence that $\mathcal{E}_{\ell_2}(\vartheta)$ is quasi-convex. In the case of $p = 1$ the single-level problem becomes

$$\min_{\vartheta \geq 0} \mathcal{E}_{\ell_1}(\vartheta) = \left(\max(0, |f| - \vartheta)\,\mathrm{sgn}(f) - g\right)^2,$$

which is nonsmooth since the solution to the lower-level problem coincides with f for all ϑ larger than a threshold value. Figure 1(b) shows $\mathcal{E}_{\ell_1}(\vartheta)$ again for various choices of the ratio f/g. Again, it can be shown that $\mathcal{E}_{\ell_1}(\vartheta)$ is quasi-convex. The quasi convexity is of interest since it improves the chance that optimization algorithms find the optimal regularization parameters of the models. In the following section a sufficient condition is found that guarantees this property also for the multidimensional, single prior case for ℓ^2 models.

Note that (2.1) does not explicitly consider the case of multiple training images. However, as we will see in section 5, this does not put any restrictions on our formulation. Indeed, we can always combine multiple images to form one large image and consider linear operators that operate on the combined image.

3. The ℓ_2 model.

3.1. Single prior. Let us first consider the most simple instance of (2.1), where we set $p = 2$ and $q = 1$, which corresponds to computing the optimal regularization parameter in a classical Tikhonov regularization functional:

$$\begin{cases} \min_{\vartheta \geq 0} \mathcal{E}(x(\vartheta)) = \|x(\vartheta) - g\|_2^2 \\ \text{subject to } x(\vartheta) = \arg\min_x \dfrac{\vartheta}{2}\|Kx\|_2^2 + \dfrac{1}{2}\|x - f\|_2^2. \end{cases} \tag{3.1}$$

Solving the lower-level optimization problem we find that $x(\vartheta) = (I + \vartheta K^T K)^{-1} f$, and hence (3.1) is equivalent to

$$\min_{\vartheta \geq 0} \mathcal{E}(\vartheta) = \|(I + \vartheta K^T K)^{-1} f - g\|_2^2. \tag{3.2}$$

It will be convenient to introduce $\mathcal{K} = K^T K \in \mathbb{R}^{n \times n}$. Every element $x \in \mathbb{R}^n$ can be uniquely decomposed as

$$x = x^N + x^\perp \in \ker(\mathcal{K}) \oplus \operatorname{ran}(\mathcal{K}).$$

In our first result, we give a condition which ensures the existence of a minimizer of (3.2).

Proposition 3.1. *If* $\|f^\perp - g^\perp\|_2 < \|g^\perp\|_2$, *then* (3.2) *admits a solution* $\vartheta^* \geq 0$. *If, moreover,* $\langle \mathcal{K} f, f - g \rangle > 0$, *then* $\vartheta^* > 0$.

Proof. Let $\{\vartheta_n\}_{n=1}^\infty$, with $\vartheta_n \geq 0$ being a minimizing sequence, i.e.,

$$\lim_{n \to \infty} \mathcal{E}(\vartheta_n) = \inf_{\vartheta \geq 0} \mathcal{E}(\vartheta). \tag{3.3}$$

We argue that $\lim_{n\to\infty} \vartheta_n = \infty$ is impossible. In fact,

$$\|(I + \vartheta_n \mathcal{K})^{-1} f - g\|_2^2 = \|(I + \vartheta_n \mathcal{K})^{-1} f^\perp - g^\perp\|_2^2 + \|(I + \vartheta_n \mathcal{K})^{-1} f^N - g^N\|_2^2,$$

and hence, if $\lim_{n\to\infty} \vartheta_n = \infty$, then

$$\lim_{n \to \infty} \|(I + \vartheta_n \mathcal{K})^{-1} f - g\|_2^2 = \|g^\perp\|_2^2 + \|f^N - g^N\|_2^2. \tag{3.4}$$

From (3.3), (3.4), and the assumptions on $f^\perp$ and $g^\perp$ we have

$$\lim_{n \to \infty} \mathcal{E}(\vartheta_n) = \|g^\perp\|_2^2 + \|f^N - g^N\|_2^2 > \|f^\perp - g^\perp\|_2^2 + \|f^N - g^N\|_2^2 = \mathcal{E}(0),$$

which is a contradiction, and thus $\{\vartheta_n\}$ is bounded. It follows that there exist a convergent subsequence and an accumulation point $\vartheta^* \in [0, \infty)$. Since $\vartheta \to \mathcal{E}(\vartheta)$ is continuous, it follows from (3.3) that every accumulation point is a solution to (3.2).

Now we assume that $\vartheta^* = 0$ and note that

$$\mathcal{E}'(\vartheta) = -\left\langle (I + \vartheta \mathcal{K})^{-2} \mathcal{K} f, (I + \vartheta \mathcal{K})^{-1} f - g \right\rangle. \tag{3.5}$$

We find that $\mathcal{E}'(0) = -\langle \mathcal{K} f, f - g \rangle$, which by assumption is strictly negative. This contradicts that 0 is a minimum, and hence $\vartheta^* \in (0, \infty)$. ∎

Remark 3.2. If $K = I$ and $\|f\|_2 = \|g\|_2 = 1$, then the condition $\langle \mathcal{K} f, f - g \rangle > 0$ becomes $1 > \langle f, g \rangle$, which is equivalent to assuming that $f \neq g$.

We next turn to investigating some of the properties of $\mathcal{E}(\vartheta)$. We shall use that

$$\mathcal{E}''(\vartheta) = 3 \left\langle (I + \vartheta \mathcal{K})^{-4} \mathcal{K} f, \mathcal{K} f \right\rangle - 2 \left\langle (I + \vartheta \mathcal{K})^{-3} \mathcal{K} f, \mathcal{K} g \right\rangle. \tag{3.6}$$

Since $\mathcal{K} \geq 0$ is symmetric, every element $x \in \mathbb{R}^n$ can be expressed as

$$x = \sum_{i=1}^r x_i e_i + x^N,$$

where $\{e_i\}$ are the normalized eigenvectors of $\mathcal{K}$ corresponding to nontrivial eigenvalues $0 < \lambda_1 \leq \cdots \leq \lambda_r$, $r \leq n$ of $\mathcal{K}$. If $r = n$, then $\ker(\mathcal{K}) = \{0\}$. We shall express $f^\perp = \sum_{i=1}^r f_i e_i$ and $g^\perp = \sum_{i=1}^r g_i e_i$.

3.1.1. The noise-free case. In this and the following subsections we investigate qualitative properties of $\mathcal{E}(\vartheta)$. We first consider the special case $f^\perp = g^\perp$, which we refer to as the noise-free case. While this case is less relevant in practice, it is important to understand the properties of the function $\mathcal{E}(\vartheta)$ also in the noise-free situation.

Proposition 3.3. *Assume that $f^\perp = g^\perp$ and $f^\perp \neq 0$:*

(a) *Then $\vartheta^* = 0$ is the unique global solution to (3.2). Moreover, $\vartheta \to \mathcal{E}(\vartheta)$ is strictly increasing from $\|f^N - g^N\|_2^2$ to $\|g^\perp\|_2^2 + \|f^N - g^N\|_2^2$, and it is strictly convex for $\vartheta \in [0, \frac{1}{2\lambda_r})$ and concave for $\vartheta \in (\frac{1}{2\lambda_1}, \infty)$.*

(b) *If $\lambda_r \leq 2\lambda_1$, then there exists a unique $\tilde{\vartheta} \in [\frac{1}{2\lambda_r}, \frac{1}{2\lambda_1}]$ such that $\mathcal{E}(\vartheta)$ is convex for $\vartheta \in [0, \tilde{\vartheta})$ and concave for $\vartheta \in (\tilde{\vartheta}, \infty)$.*

Proof.

(a) Note that $\mathcal{E}(0) = \|f^N - g^N\|_2^2$ and

$$\begin{aligned}\lim_{\vartheta\to\infty} \|(I + \vartheta\mathcal{K})^{-1} f - g\|_2^2 &= \lim_{\vartheta\to\infty} \|(I + \vartheta\mathcal{K})^{-1}(f^N + f^\perp) - g\|_2^2 \\ &= \|g^\perp\|_2^2 + \|f^N - g^N\|_2^2 .\end{aligned}$$

By (3.5) and since $f^\perp = g^\perp$ we have

$$\mathcal{E}'(\vartheta) = -\sum_{i=1}^r \left(\frac{\lambda_i}{(1+\lambda_i\vartheta)^3} - \frac{\lambda_i}{(1+\lambda_i\vartheta)^2} \right) f_i^2 = \sum_{i=1}^r \frac{\lambda_i^2\vartheta}{(1+\lambda_i\vartheta)^3} f_i^2.$$

Therefore $\mathcal{E}'(0) = 0$ and $\mathcal{E}'(\vartheta) > 0$ for $\vartheta > 0$, where we use that $f^\perp \neq 0$. Hence $\mathcal{E}$ is strictly increasing from $\|f^N - g^N\|_2^2$ to $\|g^\perp\|_2^2 + \|f^N - g^N\|_2^2$. Similarly we find that

$$\mathcal{E}''(\vartheta) = \sum_{i=1}^r \left(\frac{3\lambda_i^2}{(1+\lambda_i\vartheta)^4} - \frac{2\lambda_i^2}{(1+\lambda_i\vartheta)^3} \right) f_i^2 = \sum_{i=1}^r \frac{\lambda_i^2}{(1+\lambda_i\vartheta)^4} (1 - 2\lambda_i\vartheta) f_i^2.$$

Hence $\mathcal{E}''$ is strictly convex for $\vartheta \in [0, \frac{1}{2\lambda_r})$ and strictly concave for $\vartheta \in (\frac{1}{2\lambda_1}, \infty)$.

(b) We express $\mathcal{E}''(\vartheta) = \sum_{i=1}^r h_i$, where

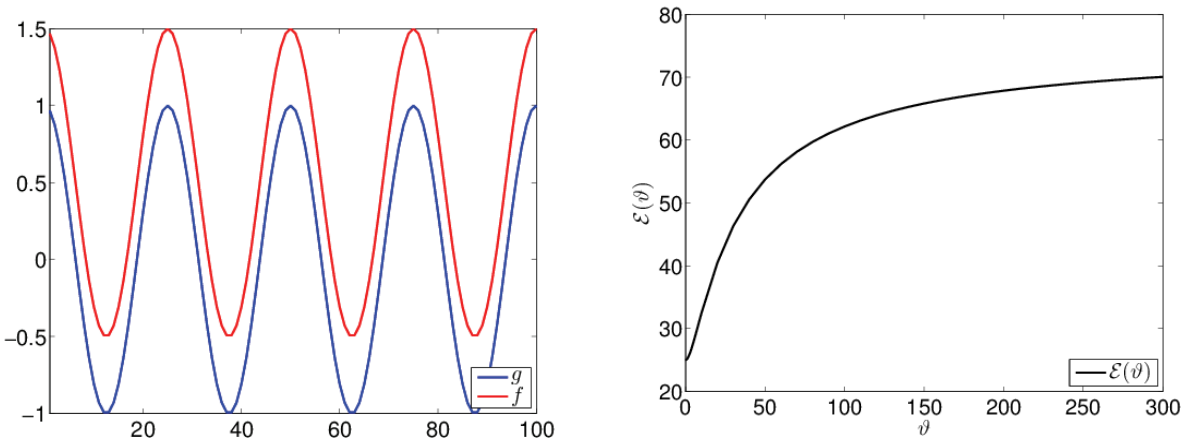
$$h_i = \frac{\lambda_i^2}{(1+\lambda_i\vartheta)^4}(1 - 2\lambda_i\vartheta) f_i^2 \quad \text{and} \quad h_i' = \frac{-6\lambda_i^3}{(1+\lambda_i\vartheta)^5}(1 - \lambda_i\vartheta) f_i^2.$$

We note that h_i is strictly monotonically decreasing on $[0, \frac{1}{\lambda_i})$ for $i = 1, \ldots, r$, and hence $\mathcal{E}''$ is strictly decreasing on $[0, \frac{1}{\lambda_r})$. We have that

$$\mathcal{E}''(\vartheta) > 0 \ \text{ for } \ \vartheta \in \left[0, \frac{1}{2\lambda_r}\right) \ \text{ and } \ \mathcal{E}''(\vartheta) < 0 \ \text{ for } \ \vartheta \in \left(\frac{1}{2\lambda_1}, \infty\right).$$

Together with $\lambda_r \leq 2\lambda_1$ these observations imply the claim. ∎

Example 3.4. Let $g = (g_1, \ldots, g_n)$ be a discrete cosine defined by $g_i = \cos(8\pi i/n)$, $1 \leq i \leq n$, and let $f = (f_1, \ldots, f_n)$ be a shifted version computed as $f_i = g_i + c$, $c \in \mathbb{R}$. Figure 2(a) plots the signal g for $n = 100$ together with its shifted version f, where $c = 1/2$. Furthermore, let K be a finite difference approximation of a one-dimensional gradient operator; i.e., $(Kx)(i) =$



(a) (b)

Figure 2. *The noise-free case.* (a) *shows the discrete cosine signals* g *and* f*, where we used an offset value of* $c = 1/2$*.* (b) *shows the function values of* $\mathcal{E}(\vartheta)$ *depending on the parameter* ϑ*.*

$x(i+1) - x(i)$ if $1 \leq i < n$ and $(Kx)(n) = 0$. Note that since $(c, \dots, c)^T \in \ker(\mathcal{K})$, $c \in \mathbb{R}$, we have that $g^\perp = f^\perp$. The nontrivial eigenvalues of $\mathcal{K}$ are given in ascending order by

$$\lambda_i = 4\sin^2((i\pi)/(2n)), \quad i = 1, \dots, n-1.$$

According to Proposition 3.3 we find that $\mathcal{E}$ is strictly convex for $\vartheta \in [0, 0.125)$ and strictly concave for $\vartheta \in (506.648, \infty)$. See also Figure 2.

3.1.2. The noisy case. The following result provides sufficient conditions for convexity and concavity of $\mathcal{E}$, for the case where $f^\perp$ may differ from $g^\perp$.

Proposition 3.5 (convexity/concavity).

(a) *If* $\|\mathcal{K}g\|_2 < \frac{3}{2}\|\mathcal{K}f\|_2$*, then* $\mathcal{E}$ *is strictly convex on*

$$\left(0, \frac{1}{\|\mathcal{K}\|_2}\left(\sqrt{\frac{3}{2}\frac{\|\mathcal{K}f\|_2}{\|\mathcal{K}g\|_2}} - 1\right)\right).$$

(b) *If* $f^\perp \neq 0$*, then* $\mathcal{E}$ *is strictly convex on* $(0, \underline{\vartheta})$*, where* $\underline{\vartheta} = \min_{f_i g_i > 0} \frac{1}{\lambda_i}(\frac{3f_i^2}{2f_i g_i} - 1)$*. If* $f_i g_i \leq 0$ *for* $i = 1, \dots, m$*, then* $\underline{\vartheta} = \infty$*.*

(c) *If* $\sum_{i=1}^r \frac{1}{\lambda_i^2} f_i g_i > 0$*, then there exists* $\tilde{\vartheta}$ *such that* $\mathcal{E}$ *is strictly concave on* $(\tilde{\vartheta}, \infty)$*.*

Proof.

(a) We have

$$1 = \|(I + \vartheta\mathcal{K})^{-1}(I + \vartheta\mathcal{K})\|_2 \leq \|(I + \vartheta\mathcal{K})^{-1}\|_2 \|(I + \vartheta\mathcal{K})\|_2,$$

from which together with $\|(I + \vartheta\mathcal{K})\|_2 \leq 1 + \vartheta\|\mathcal{K}\|_2$ it follows that

$$\frac{1}{1 + \vartheta\|\mathcal{K}\|_2} \leq \frac{1}{\|(I + \vartheta\mathcal{K})\|_2} \leq \|(I + \vartheta\mathcal{K})^{-1}\|_2 \leq 1,$$

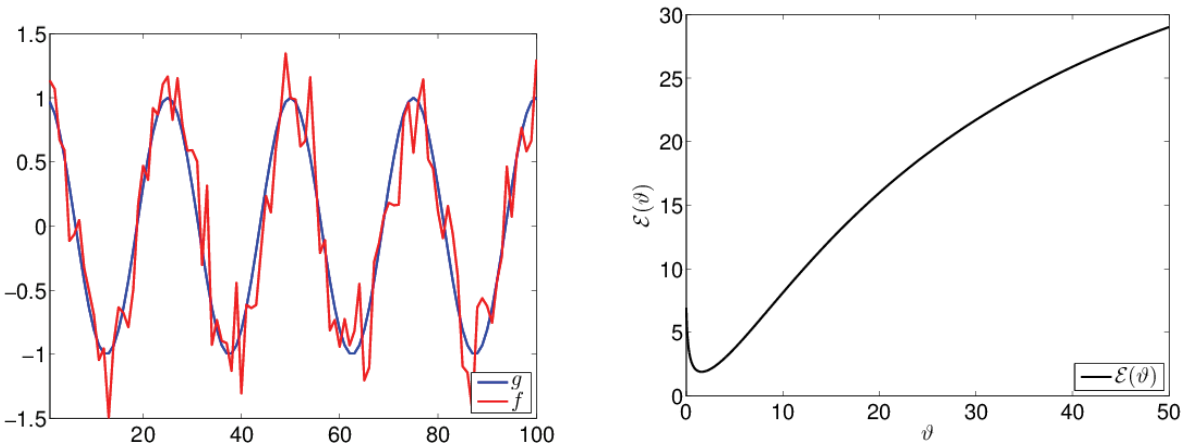


Figure 3. *The noisy case.* (a) *shows the discrete cosine signal* g *and its noisy version* f *with additive Gaussian noise with a standard deviation of* $\sigma = 1/4$. (b) *shows the function values of* $\mathcal{E}(\vartheta)$ *depending on the parameter* ϑ *together with the bound of strict convexity which is computed according to Proposition* 3.5(a).

where the upper bound follows from the fact that $\mathcal{K}$ is positive definite. From (3.6) we have

$$\begin{aligned}\mathcal{E}''(\vartheta) &\geq \|(I+\vartheta\mathcal{K})^{-2}\mathcal{K}f\|_2 \left(3\|(I+\vartheta\mathcal{K})^{-2}\mathcal{K}f\|_2 - 2\|(I+\vartheta\mathcal{K})^{-1}\mathcal{K}g\|_2\right)\\ &\geq \|(I+\vartheta\mathcal{K})^{-2}\mathcal{K}f\|_2 \left(\frac{3}{(1+\vartheta\|\mathcal{K}\|_2)^2}\|\mathcal{K}f\|_2 - 2\|\mathcal{K}g\|_2\right) > 0,\end{aligned}$$

provided that $\vartheta \in (0, \frac{1}{\|\mathcal{K}\|_2}(\sqrt{\frac{3}{2}\frac{\|\mathcal{K}f\|_2}{\|\mathcal{K}g\|_2}} - 1))$.

(b) Let $P = \{i \in \{1,\dots,m\} : f_i g_i > 0\}$. Utilizing (3.6) we find that

$$\begin{aligned}\mathcal{E}''(\vartheta) &= 3\sum_{i=1}^{m} \frac{\lambda_i^2}{(1+\vartheta\lambda_i)^4} f_i^2 - 2\sum_{i=1}^{m} \frac{\lambda_i^2}{(1+\vartheta\lambda_i)^3} f_i g_i\\ &\geq 3\sum_{i=1, i\notin P}^{m} \frac{\lambda_i^2}{(1+\vartheta\lambda_i)^4} f_i^2 + \sum_{i\in P}^{m} \frac{\lambda_i^2}{(1+\vartheta\lambda_i)^4}(3f_i^2 - 2f_i g_i(1+\vartheta\lambda_i)) > 0\end{aligned} \tag{3.7}$$

for $\vartheta \in (0, \underline{\vartheta})$. Here we also use that $f^{\perp} \neq 0$.

(c) For $\vartheta \geq \frac{1}{\lambda_1}$ we have

$$\mathcal{E}''(\vartheta) \leq \frac{3}{\vartheta^4}\sum_{i=1}^{r}\frac{1}{\lambda_i^2} f_i^2 - \frac{2}{\vartheta^3}\sum_{i=1}^{r}\frac{\lambda_i}{(\frac{1}{\vartheta}+\lambda_i)^3} f_i g_i \leq \frac{3}{\vartheta^4}\frac{1}{\lambda_1^2}\|f\|_2^2 - \frac{1}{4\vartheta^3}\sum_{i=1}^{r}\frac{1}{\lambda_i^2} f_i g_i,$$

and the claim follows. ∎

Example 3.6. Let g and K be as defined in Example 3.4, but now let f be a noisy version of g, where we add zero-mean Gaussian noise with $\sigma = 1/4$. Figure 3(a) plots the cosine

signal g for $n = 100$ together with its noisy version f. According to Proposition 3.5(a), we get that $\mathcal{E}$ is strictly convex on $\vartheta \in (0, \tilde{\vartheta})$, where $\tilde{\vartheta} = \frac{1}{\|\mathcal{K}\|_2}(\sqrt{\frac{3}{2}\frac{\|\mathcal{K}f\|_2}{\|\mathcal{K}g\|_2}} - 1)$ is computed as $\tilde{\vartheta} = 0.8932$. See Figure 3, where the typical quasi-convex behavior of the learning functional $\mathcal{E}$ can be observed.

3.1.3. A remark on the infinite-dimensional case. Let K be a closed densely defined linear operator between Hilbert spaces H and Y, with H separable. Then $\mathcal{K} = K^*K$ is a self-adjoint nonnegative operator in H with dense domain that we denote by $\mathrm{dom}(\mathcal{K})$; see, e.g., [18, page 326]. Moreover, for every λ with $\mathrm{Re}\,\lambda > 0$, the resolvent $(\mathcal{K} + \lambda I)^{-1}$ exists as a bounded linear operator on H; see, e.g., [18, page 279]. Within this setting we consider for $g \in H$, $f \in H$, and $\vartheta \geq 0$

$$\begin{cases} \min\limits_{\vartheta \geq 0} \mathcal{E}(x(\vartheta)) = \|x(\vartheta) - g\|_H^2 \\ \text{subject to } x(\vartheta) = \arg\min\limits_x \dfrac{\vartheta}{2}\|Kx\|_Y^2 + \dfrac{1}{2}\|x - f\|_H^2. \end{cases} \tag{3.8}$$

The necessary and sufficient optimality condition for the lower-level problem is given by

$$(I + \vartheta\mathcal{K})x = f. \tag{3.9}$$

It has a unique solution $x(\vartheta) \in H$ for each $\vartheta \geq 0$. If $\vartheta > 0$, then $x(\vartheta) \in \mathrm{dom}(\mathcal{K})$. Again, we have an equivalent reduced problem

$$\min_{\vartheta \geq 0} \mathcal{E}(\vartheta) = \|(I + \vartheta\mathcal{K})^{-1} f - g\|_H^2 \tag{3.10}$$

and the orthogonal decomposition

$$x = x^N + x^\perp \in \ker(\mathcal{K}) \oplus \overline{\mathrm{ran}(\mathcal{K})},$$

where the closure is taken in H. We assume that $(I + \vartheta\mathcal{K})^{-1}$ is compact for some (or, equivalently, all) $\vartheta > 0$. Then the spectrum of $\mathcal{K}$ consists entirely of isolated eigenvalues $0 < \lambda_1 \leq \lambda_2 \ldots$ of finite multiplicity plus possibly the eigenvalue 0, and every $x \in H$ can be expressed as $x = \sum_{i=1}^\infty x_i e_i + x^N$, with $x^N \in \ker(\mathcal{K})$ and e_i eigenvectors of $\mathcal{K}$ associated with the eigenvalues $\neq 0$. We have the analogue of Proposition 3.1.

Proposition 3.7. *If $\|f^\perp - g^\perp\|_H < \|g^\perp\|_H$, then (3.10) admits a solution $\vartheta^* \geq 0$. If, moreover, $f \in \mathrm{dom}\mathcal{K}$ and $\langle \mathcal{K}f, f - g\rangle_H > 0$, then $\vartheta^* > 0$.*

Proof. Using the fact that $(I + \vartheta\mathcal{K})^{-1}$ leaves $\ker(\mathcal{K})$ and $(\ker(\mathcal{K})^\perp$ invariant we can proceed as in the proof of Proposition 3.1 to get the first part of the result. Note that $\lim_{\vartheta \to 0^+}(I + \vartheta\mathcal{K})v = v$ for all $v \in H$. Consequently $\mathcal{E}'(\vartheta)$ is continuous on $[0, \infty)$ if $f \in \mathrm{dom}(\mathcal{K})$. The proof of the second part now follows as in that of Proposition 3.1. ∎

3.2. Multiple priors. In this section we study the ℓ_2 model with multiple priors, i.e., $p = 2$ and $q \geq 1$. It is defined as

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\sum_{k=1}^q \vartheta_k \|K_k x\|_2^2 + \frac{1}{2}\|x - f\|_2^2, \tag{3.11}$$

with the parameter vector $\vartheta = (\vartheta_1, \dots, \vartheta_q) \geq 0$. The minimum of the above problems is characterized by

$$x + \sum_{k=1}^{q} \vartheta_k \mathcal{K}_k x = f,$$

or, equivalently, $x = (I + \sum_{k=1}^{q} \vartheta_k \mathcal{K}_k)^{-1} f$. The reduced quadratic learning functional is then given by

$$\min_{\vartheta \geq 0} \mathcal{E}(\vartheta) = \frac{1}{2} \left\| \left(I + \sum_{k=1}^{q} \vartheta_k \mathcal{K}_k \right)^{-1} f - g \right\|^2 . \tag{3.12}$$

For convenience we introduce the symmetric positive definite matrix

$$\mathcal{R} = \left(I + \sum_{k=1}^{q} \vartheta_k \mathcal{K}_k \right)^{-1} .$$

To guarantee existence the following condition will be used:

$$\inf\{\|\tilde{x} - g\|_2 : \tilde{x} \in \ker(K_k) \ \text{ for some } \ k = 1, \dots, q\} > \|f - g\|_2. \tag{3.13}$$

We observe that in case $\ker(K_k) = \{0\}$ for all k, condition (3.13) amounts to $\|g\|_2 > \|f - g\|_2$. If $q = 1$, then (3.13) is equivalent to assuming that $\|g^{\perp}\|_2 > \|f^{\perp} - g^{\perp}\|_2$. This condition was already used for the single-parameter case in Proposition 3.1.

Proposition 3.8. *If* (3.13) *holds and* $\ker K_k \cap \ker K_l = \{0\}$ *for all* $k \neq l$, *then* (3.12) *admits a solution.*

Proof. Let $\{\vartheta^n\}_{n=1}^{\infty}$ denote a minimizing sequence, and suppose that $\lim_{n\to\infty} \|\vartheta^n\|_2 = \infty$. Then there exist index sets $\mathcal{J} \subseteq \{1, \dots, q\}$, $\overline{\mathcal{J}} = \{1, \dots, q\} \setminus \mathcal{J}$ and a constant κ_1 such that

$$\lim_{n\to\infty} \vartheta_k^n = \infty \quad \text{for } k \in \mathcal{J} \quad \text{and} \quad |\vartheta_k^n| \leq \kappa_1 \quad \text{for } k \in \overline{\mathcal{J}} \text{ and all } n. \tag{3.14}$$

We set

$$x^n = \left(I + \sum_{k=1}^{q} \vartheta_k^n \mathcal{K}_k \right)^{-1} f. \tag{3.15}$$

Clearly $\{x^n\}$ is bounded and hence, on a subsequence, denoted by the same index, $\lim_{n\to\infty} x^n = \hat{x}$ for some $\hat{x} \in \mathbb{R}^n$. From (3.15)

$$\sum_{k\in\mathcal{J}} \vartheta_k^n \mathcal{K}_k x^n = f - \left(x_n + \sum_{k\in\overline{\mathcal{J}}} \vartheta_k^n \mathcal{K}_k x^n \right).$$

Taking the inner product with x^n and observing that the right-hand side is bounded,

$$\min_{k\in\mathcal{J}} \vartheta_k^n \sum_{k\in\mathcal{J}} \|K_k x^n\|_2^2 \leq \kappa_2$$

for a constant κ_2 independent of n. Since $\min_{k\in\mathcal{J}} \vartheta_k^n \to \infty$ for $n \to \infty$, we find that $\lim_{n\to\infty} K_k x^n = K_k \hat{x} = 0$ for all, meaning that $\hat{x} \in \ker(K_k)$ for all $k \in \mathcal{J}$. Since $\{\vartheta^n\}$ was chosen as the minimizing sequence, we obtain

$$\inf_{\vartheta\geq 0} \mathcal{E}(\vartheta) = \lim_{n\to\infty} \|x^n - g\|_2^2 = \|\hat{x} - g\|_2^2 > \|f - g\|_2^2 = \mathcal{E}(0),$$

where we used (3.13). This is a contradiction, and hence every minimizing sequence is bounded. Since $\vartheta \to \mathcal{E}(\vartheta)$ is continuous, the claim follows. ∎

The partial derivatives of $\mathcal{E}$ with respect to ϑ_k are given by

$$(\nabla\mathcal{E}(\vartheta))_k = -\langle \mathcal{R}f - g, \mathcal{R}\mathcal{K}_k\mathcal{R}f\rangle \qquad \text{for } k = 1,\dots,q,$$

where $\mathcal{R}$ is evaluated at ϑ. Taking into account the inequality constraint $\vartheta \geq 0$ in (3.12), the first-order necessary condition is given by

$$\nabla\mathcal{E}(\vartheta^*) - \mu = 0, \quad \mu \geq 0,\ \vartheta^* \geq 0,\ \langle \mu, \vartheta^*\rangle = 0, \tag{3.16}$$

where $\mu \in \mathbb{R}^q$ is the Lagrange multiplier associated with the constraint $\vartheta \geq 0$. It can be checked that the three last conditions can be equivalently expressed as

$$\mu - \max(0, \mu - \vartheta^*) = 0.$$

For the Hessian of $\mathcal{E}$ we obtain for $k = 1,\dots,q$ and $l = 1,\dots,q$ the expression

$$\nabla^2\mathcal{E}(\vartheta) = M_1 + M_2,$$

where

$$(M_1)_{k,l} = \langle \mathcal{R}\mathcal{K}_k\mathcal{R}f, \mathcal{R}\mathcal{K}_l\mathcal{R}f\rangle \qquad \text{and} \qquad (M_2)_{k,l} = \langle \mathcal{R}f - g, \mathcal{R}\mathcal{K}_k\mathcal{R}\mathcal{K}_l\mathcal{R}f + \mathcal{R}\mathcal{K}_l\mathcal{R}\mathcal{K}_k\mathcal{R}f\rangle$$

are symmetric matrices.

Let $\mathcal{A} = \{k \in \{1,\dots,q\} : (\vartheta^*)_k = 0\}$ denote the set of active constraints for some local solution ϑ^* of (3.11). Then, the second-order necessary optimality condition implies that

$$\nabla^2\mathcal{E}(\vartheta^*) \text{ is semidefinite on } T, \tag{3.17}$$

where T is the tangent space of the active constraints $T = \{\vartheta \in \mathbb{R}^q : \vartheta_k = 0 \text{ for all } \ k \in \mathcal{A}\}$. Note that M_1 is a Gram matrix corresponding to the vectors $\{\mathcal{R}\mathcal{K}_k\mathcal{R}f\}_{k=1}^q$. We assume that

$$\{\mathcal{K}_k\mathcal{R}f\}_{k=1}^q \text{ is linearly independent.} \tag{3.18}$$

Since $\mathcal{R}$ is positive definite, $\{\mathcal{R}\mathcal{K}_k\mathcal{R}f\}_{k=1}^m$ is linearly independent and M_1 is nonsingular. If $\|\mathcal{R}f - g\|_2$ is sufficiently small, then $M_1 + M_2$ is nonsingular as well. This implies that $\nabla^2\mathcal{E}(\vartheta^*) > 0$ on $\mathbb{R}^q$. We summarize our discussion in a theorem.

Theorem 3.9. *Assume that (3.18) is satisfied, and let ϑ^* be a local solution of (3.11). Then, if*

$$\left\| \left(I + \sum_{k=1}^q \vartheta_k^* \mathcal{K}_k \right)^{-1} f - g \right\| \text{ is sufficiently small,} \tag{3.19}$$

the second-order sufficient optimality condition is satisfied at ϑ^*; in particular, ϑ^* is a locally unique minimum.

Note that for $f = g$ we have $\vartheta^* = 0$ as global solution. Therefore (3.19) can be interpreted as a smallness condition on the error in the data.

Algorithm 3.1. Newton Learning for ℓ_2 (NL-ℓ_2).

(i) Choose $(\vartheta^0, \mu^0) \in \mathbb{R}^r \times \mathbb{R}^q$, and set $n = 0$.
(ii) Determine $\mathcal{A}^n = \{k : \mu_k^n - \vartheta_k^n \geq 0\}$, $\mathcal{I}^n = \{k : \mu_k^n - \vartheta_k^n < 0\}$.
(iii) Assign $M = \nabla^2 \mathcal{E}(\vartheta^n)$, $P = \operatorname{diag}(p_k)$, $Q = \operatorname{diag}(q_k)$, where

$$p_k = \begin{cases} 1 \text{ if } k \in \mathcal{A}^n, \\ 0 \text{ if } k \in \mathcal{I}^n, \end{cases} \qquad q_k = \begin{cases} 0 \text{ if } k \in \mathcal{A}^n, \\ 1 \text{ if } k \in \mathcal{I}^n. \end{cases}$$

(iv) Solve for $(\delta\vartheta, \delta\mu)$:

$$\begin{pmatrix} M & -I \\ P & Q \end{pmatrix} \begin{pmatrix} \delta\vartheta \\ \delta\mu \end{pmatrix} = - \begin{pmatrix} \nabla \mathcal{E}(\vartheta^n) - \mu^n \\ \mu^n - \max(0,\, \mu^n - \vartheta^n) \end{pmatrix}. \tag{3.20}$$

(v) Update $(\vartheta^{n+1}, \mu^{n+1}) = (\vartheta^n, \mu^n) + (\delta\vartheta, \delta\mu)$, set $n = n + 1$, and goto (ii).

3.3. Newton algorithm. We propose and analyze a semismooth Newton scheme for solving (3.12). For this purpose we express the necessary optimality condition (3.16) in the form

$$\begin{cases} \nabla \mathcal{E}(\vartheta^*) - \mu = 0, \\ \mu - \max(0,\, \mu - \vartheta) = 0. \end{cases} \tag{3.21}$$

To solve (3.21) we utilize a semismooth Newton algorithm which is outlined in Algorithm 3.1.

To analyze this algorithm the vectors $\delta\vartheta$ and $\delta\mu$ are decomposed into inactive and active components $(\delta\vartheta)_\mathcal{I}, (\delta\vartheta)_\mathcal{A}$ and $(\delta\mu)_\mathcal{I}, (\delta\mu)_\mathcal{A}$, respectively, and M is partitioned accordingly:

$$M = \begin{pmatrix} M_{\mathcal{I}\mathcal{I}} & M_{\mathcal{I}\mathcal{A}} \\ M_{\mathcal{A}\mathcal{I}} & M_{\mathcal{A}\mathcal{A}} \end{pmatrix}.$$

Here, for notational convenience the unknowns are ordered in such a manner that the inactive coordinates appear first and the active ones last, and the iteration index for the sets $\mathcal{A}^n$ and $\mathcal{I}^n$ is dropped. From the second equation in (3.20) we obtain

$$(\delta\vartheta)_\mathcal{A} = -\vartheta_\mathcal{A}^n, \quad (\delta\mu)_\mathcal{I} = -\mu_\mathcal{I}^n, \quad \vartheta_\mathcal{A}^{n+1} = 0, \quad \mu_\mathcal{I}^{n+1} = 0. \tag{3.22}$$

Turning to the first equation in (3.20) we first solve for the inactive components of $\delta\vartheta$ by

$$M_{\mathcal{I}\mathcal{I}}(\delta\vartheta)_\mathcal{I} = -M_{\mathcal{I}\mathcal{A}}(\delta\vartheta)_\mathcal{A} - (\nabla \mathcal{E}(\vartheta^n))_\mathcal{I} \tag{3.23}$$

and then assign

$$(\delta\mu)_\mathcal{A} = M_{\mathcal{A}\mathcal{I}}(\delta\vartheta)_\mathcal{I} + M_{\mathcal{A}\mathcal{A}}(\delta\vartheta)_\mathcal{A} + (\nabla \mathcal{E}(\vartheta^n))_\mathcal{A} - \mu_\mathcal{A}^n.$$

Note that while (3.20) is asymmetric, system (3.23), which is of the dimension of the inactive set, is symmetric.

Theorem 3.10. *Let ϑ^* be a local solution of (3.12) with associated Lagrange multiplier μ^*, and suppose that (3.18) and (3.19) hold. Then, if $\|(\vartheta^0, \mu^0) - (\vartheta^*, \mu^*)\|_2$ is sufficiently small, the iterations of Algorithm 3.1 converge superlinearly to (ϑ^*, μ^*).*

Proof. We verify here the requirements for superlinear convergence of the semismooth Newton method as given in, e.g., [17, page 238]. The max-operation is well known to be semismooth (see, e.g., [17, 30] and the references cited therein), and $D\max(0, x) = \chi_{\{x\geq 0\}}$ is a generalized or Newton derivative. Here $(\chi_{\{x\geq 0\}})_i = 1$ if $x_i \geq 0$ and $(\chi_{\{x\geq 0\}})_i = 0$ otherwise. This choice of generalized derivative determines step (iii) of Algorithm 3.1. The proof will be complete if we argue that the system matrices

$$H(\vartheta, \mu) = \begin{pmatrix} M(\vartheta) & -I \\ P(\vartheta, \mu) & Q(\vartheta, \mu) \end{pmatrix}$$

are invertible with uniformly bounded inverses in a neighborhood $B_\rho(\vartheta^*, \mu^*)$ of (ϑ^*, μ^*) for some radius $\rho > 0$. The notation $H(\vartheta, \mu)$ emphasizes the dependence of M, P, and Q on ϑ and μ. The discussion before Theorem 3.10 implies that $\nabla^2\mathcal{E}(\vartheta^*) = M(\vartheta^*) > 0$. Hence there exists a neighborhood $B_\rho(\vartheta^*)$, with $\rho > 0$, and $\kappa > 0$ such that $\|M^{-1}(\vartheta)\|_2 \leq \kappa$ for all $\vartheta \in B_\rho(\vartheta^*)$. In particular, this implies that $\|M(\vartheta)^{-1}_{\mathcal{I}\mathcal{I}}\|_2 \leq \kappa$ for all $\vartheta \in B_\rho(\vartheta^*)$ and any combination of $\mathcal{I} \in \{1, \dots, q\}$. Now consider for $\vartheta \in B_\rho(\vartheta^*)$, $\mu \in \mathbb{R}^q$, and $(y, z) \in \mathbb{R}^{2q}$

$$H(\vartheta, \mu) \begin{pmatrix} \delta\vartheta \\ \delta\mu \end{pmatrix} = \begin{pmatrix} y \\ z \end{pmatrix}. \tag{3.24}$$

As in the computation before the statement of the theorem we find that

$$(\delta\vartheta)_{\mathcal{A}} = \frac{1}{c} z_{\mathcal{A}}, \quad (\delta\mu)_{\mathcal{I}} = z_{\mathcal{I}}.$$

From the first equation in (3.24) we find that

$$\begin{aligned} M_{\mathcal{I}\mathcal{I}}(\vartheta)(\delta\vartheta)_{\mathcal{I}} &= -\frac{1}{c} M_{\mathcal{I}\mathcal{A}}(\vartheta) z_{\mathcal{A}} + z_{\mathcal{I}} + y_{\mathcal{I}}, \\ (\delta\mu)_{\mathcal{A}} &= M_{\mathcal{A}\mathcal{I}}(\delta\vartheta)_{\mathcal{I}} + M_{\mathcal{A}\mathcal{A}}(\delta\vartheta)_{\mathcal{A}} - y_{\mathcal{A}}. \end{aligned}$$

Combining these equalities, the invertibility of $H(\vartheta, \mu)$ with uniformly bounded inverses for (ϑ, μ) in a neighborhood of (ϑ^*, μ^*) follows. ∎

4. The ℓ_1 model. In this section we analyze variational models with ℓ_1 and hence nondifferentiable regularization terms. This type of model has great impact in signal processing, in particular in imaging and compressed sensing. First, we investigate the existence of a solution of the bilevel optimization problems. Then, we derive the optimality conditions by using a regularization approach and passing the regularization parameter to zero. Finally, we will develop superlinearly converging semismooth Newton algorithms that solve the regularized problems.

4.1. Problem formulation and existence.

$$\begin{cases} \min\limits_{\vartheta \geq 0} \mathcal{E}(x(\vartheta)) = \|x(\vartheta) - g\|_2^2 \\ \text{subject to } x(\vartheta) = \arg\min\limits_x \displaystyle\sum_{k=1}^{q} \vartheta_k \|K_k x\|_1 + \frac{1}{2}\|x - f\|_2^2 . \end{cases} \tag{4.1}$$

The lower-level problem (4.1) admits a unique solution $x = x(\vartheta)$. Its optimality condition is given by

$$\begin{cases} \sum_{k=1}^{q} \vartheta_k K_k^* \lambda^k + x = f, \\ \lambda_i^k \in \begin{cases} \operatorname{sgn}(K_k x)_i \text{ if } (K_k x)_i \neq 0, \\ [-1, 1] \text{ if } (K_k x)_i = 0. \end{cases} \end{cases} \tag{4.2}$$

We have the following existence result analogous to Proposition 3.1.

Proposition 4.1. *If* (3.13) *holds, then* (4.1) *admits a solution* $\vartheta^* \geq 0$.

Proof. We first argue that $\vartheta \to x(\vartheta)$, with $x(\vartheta)$ the solution to the lower-level problem, is continuous. Let $\vartheta^n \to \overline{\vartheta}$ and $x_n = x(\vartheta^n)$. Since

$$\sum_{k=1}^{q} \vartheta_k^n \|K_k x_n\|_1 + \frac{1}{2}\|x_n - f\|_2^2 \leq \frac{1}{2}\|f\|_2^2,$$

the sequence $\{x_n\}$ is bounded, and hence it admits a convergent subsequence $x_{n_k} \to \overline{x}$. We need to argue that $\overline{x} = x(\overline{\vartheta})$. For this purpose we note that

$$\sum_{k=1}^{q} \vartheta_k^n \|K_k x_n\|_1 + \frac{1}{2}\|x_n - f\|_2^2 \leq \sum_{k=1}^{q} \vartheta_k^n \|K_k x\|_1 + \frac{1}{2}\|x - f\|_2^2 \quad \text{for all } x \in \mathbb{R}^n$$

implies that

$$\sum_{k=1}^{q} \overline{\vartheta_k} \|K_k \overline{x}\|_1 + \frac{1}{2}\|\overline{x} - f\|_2^2 \leq \sum_{k=1}^{q} \overline{\vartheta_k} \|K_k x\|_1 + \frac{1}{2}\|x - f\|_2^2 \quad \text{for all } x \in \mathbb{R}^n,$$

and hence $\overline{x} = x(\overline{\vartheta})$, since the solution to the lower-level problem is unique.

Next, let $\{\vartheta^n\}_{n=1}^{\infty}$ be a minimizing sequence, and abbreviate $x_n = x(\vartheta^n)$. If $\lim_{n\to\infty} \|\vartheta^n\|_2 = \infty$, determine $\mathcal{J}$ as in (3.14). Since

$$\sum_{k=1}^{q} \vartheta_k^n \|K_k x_n\|_1 + \frac{1}{2}\|x_n - f\|_2^2 \leq \frac{1}{2}\|f\|_2^2,$$

we deduce that $\{x_n\}_{n=1}^{\infty}$ is bounded and that $\lim_{n\to\infty} \|K_i x_n\|_1 = 0$ for all $i \in \mathcal{J}$. Hence there exist a subsequence, denoted by the same symbol, and $\hat{x}$ such that $\lim_{n\to\infty} x_n = \hat{x}$ and $K_i \hat{x} = 0$ for all $i \in \mathcal{J}$. In particular, $\hat{x}$ is contained in the kernel of at least one operator K_i, and thus by (3.13)

$$\inf_{\vartheta \geq 0} \mathcal{E}(x(\vartheta)) = \lim_{n\to\infty} \mathcal{E}(x_n) = \lim_{n\to\infty} \|x_n - g\|_2^2 = \|\hat{x} - g\|_2^2 > \|f - g\|_2^2 = \mathcal{E}(x(0)),$$

which contradicts the choice of $\{\vartheta_n\}_{n=1}^{\infty}$ as the minimizing sequence. Hence $\{\vartheta_n\}_{n=1}^{\infty}$is bounded in $\mathbb{R}^q$. Consequently there exist another subsequence, denoted by the same symbol, and $\vartheta^* \in [0, \infty)$ such that $\lim_{n\to\infty} \vartheta_n = \vartheta^*$. Since $\vartheta \to x(\vartheta)$, and hence $\vartheta \to \mathcal{E}(x(\vartheta))$ are

continuous, it follows that every accumulation point ϑ^* of $\{\vartheta_n\}$ is a solution to (4.1), and $x^* = x(\vartheta^*)$. ∎

Remark 4.2. In case of only one prior, we can give a sufficient condition to exclude the case that $\vartheta^* = 0$. For this purpose we assume that

$$(Kg)_i = 0 \text{ if } (Kf)_i = 0 \quad \text{and} \quad \left\langle Kf - g, \frac{Kf}{|Kf|} \right\rangle > 0, \tag{4.3}$$

where $\frac{Kf}{|Kf|}$ is interpreted componentwise as $\frac{(Kf)_i}{|(Kf)_i|}$ if $(Kf)_i \neq 0$ and $\frac{(Kf)_i}{|(Kf)_i|}$ is interpreted as some element in $[-1, 1]$ if $(Kf)_i = 0$. We now exclude that $\vartheta^* = 0$ is the minimum. For this purpose we argue that $\frac{d}{d\vartheta}\mathcal{E}(x(\vartheta))|_{\vartheta=0^+}$ exists and is negative. We have

$$\begin{aligned}\mathcal{E}(x(\vartheta)) - \mathcal{E}(x(0)) &= \langle x(\vartheta) + x(0) - 2g, x(\vartheta) - x(0)\rangle \\ &= -\vartheta \langle x(\vartheta) + f - 2g, K^*\lambda(\vartheta)\rangle = -\vartheta \langle K(x(\vartheta) + f - 2g), \lambda(\vartheta)\rangle ,\end{aligned}$$

where we use that $x(0) = f$.

Let $\mathcal{I} = \{i : (Kx(0))_i \neq 0\}$. Then $(Kx(\vartheta))_i \neq 0$ for all $i \in \mathcal{I}$ and all $\vartheta > 0$ sufficiently small. For these i and ϑ we have

$$\lambda_i(\vartheta) = \frac{(Kx(\vartheta))_i}{|(Kx(\vartheta))_i|} \to \frac{(Kx(0))_i}{|(Kx(0))_i|}$$

as $\vartheta \to 0^+$. For $i \notin \mathcal{I}$ we have $\lambda_i(\vartheta) \in [-1, 1]$ and $(K(x(\vartheta) + f - 2g))_i \to 0$ for $\vartheta \to 0^+$, where we use that $\lim_{\vartheta\to 0^+} x(\vartheta)_i = f_i$ and (4.3). Therefore

$$\lim_{\vartheta\to 0^+} \frac{1}{\vartheta}(\mathcal{E}(x(\vartheta)) - \mathcal{E}(x(0))) = -2\left\langle K(f - g), \frac{Kf}{|Kf|}\right\rangle ,$$

and $\vartheta \to \mathcal{E}(x(\vartheta))$ is differentiable at $\vartheta = 0^+$. By (4.3) we have $\frac{d}{dt}\mathcal{E}(x(\vartheta))|_{\vartheta=0^+} < 0$, and hence $\vartheta = 0$ cannot be a solution to (4.1). We note that the condition $\langle K(f - g), \frac{Kf}{|Kf|}\rangle > 0$ can be expressed equally well as $\langle K(f - g), \lambda(0)\rangle > 0$ for any Lagrange multiplier $\lambda(0)$ associated with $\vartheta = 0$.

4.2. Optimality system. To derive an optimality system for (4.1) we use a regularization approach and consider

$$\begin{cases} \min\limits_{\vartheta\geq 0} \mathcal{E}(x(\vartheta)) = \|x(\vartheta) - g\|_2^2 \\ \text{subject to } x(\vartheta) = \arg\min\limits_x \displaystyle\sum_{k=1}^{q} \vartheta_k \sum_{j=1}^{m} n_\varepsilon((K_k x)_j) + \frac{1}{2}\|x - f\|_2^2, \end{cases} \tag{4.4}$$

where, for $\varepsilon > 0$,

$$n_\varepsilon(t) = \begin{cases} -\dfrac{1}{8\varepsilon^3}t^4 + \dfrac{3}{4\varepsilon}t^2 + \dfrac{3\varepsilon}{8} & \text{if } |t| < \varepsilon, \\ |t| & \text{else.} \end{cases} \tag{4.5}$$

The following properties of n_ε will be used repeatedly:

$$\begin{cases} n_\varepsilon \in C^2(\mathbb{R},\mathbb{R}),\ n_\varepsilon(\varepsilon) = \pm\varepsilon,\ n_\varepsilon'(\pm\varepsilon) = \pm 1,\ n_\varepsilon''(\pm\varepsilon) = 0,\ n_\varepsilon'(t) \in [-1,1], \\ n_\varepsilon''(t) \in \left[0, \dfrac{3}{2\varepsilon}\right],\ n_\varepsilon(t) \geq t, \text{ for all } t \in \mathbb{R}. \end{cases} \tag{4.6}$$

Furthermore, we have

$$n_\varepsilon'(t) = \begin{cases} -\frac{1}{2\varepsilon^3}t^3 + \frac{3}{2\varepsilon}t & \text{if} \quad |t| < \varepsilon, \\ \operatorname{sgn}(t) & \text{else} \end{cases}$$

$$n_\varepsilon''(t) = \begin{cases} -\frac{3}{2\varepsilon^3}t^2 + \frac{3}{2\varepsilon} & \text{if} \quad |t| < \varepsilon, \\ 0 & \text{else} \end{cases}$$

$$n_\varepsilon'''(t) = \begin{cases} -\frac{3}{\varepsilon^3}t & \text{if} \quad |t| < \varepsilon, \\ \left\{0, \frac{3}{\varepsilon^2}\right\} & \text{if} \quad t = -\varepsilon, \\ \left\{-\frac{3}{\varepsilon^2}, 0\right\} & \text{if} \quad t = \varepsilon, \\ 0 & \text{else.} \end{cases}$$

At $t = \pm\varepsilon$ we consider, for the time being, the third derivative to be multivalued, consisting of the right and left directional derivatives. It is simple to argue the existence of a unique lower-level solution $x_\varepsilon(\vartheta)$ for each $\varepsilon > 0$. It is characterized as the solution $x = x(\vartheta)$ to

$$x + \sum_{k=1}^{q} \vartheta_k K_k^T N_\varepsilon'(K_k x) = f, \tag{4.7}$$

where

$$N_\varepsilon'(K_k x) = (n_\varepsilon'((K_k x)_1), \dots, n_\varepsilon'((K_k x)_m))^T \in \mathbb{R}^m.$$

Since $t \to n_\varepsilon'(t)$ is monotone, the operator $x \to x + \sum_{k=1}^q \vartheta_k K_k^T N_\varepsilon'(K_k x)$ is strictly monotone, and hence the solution to (4.7) is unique. Using (4.7) it follows that $\vartheta \to x_\varepsilon(\vartheta)$ is differentiable on $[0,\infty)^q$ for each $\varepsilon > 0$, with the sensitivity equation given by

$$D_\vartheta x + [K_k^T N_\varepsilon'(K_k x)] + \sum_{k=1}^{q} \vartheta_k K_k^T N_\varepsilon''(K_k x) K_k D_\vartheta x = 0, \tag{4.8}$$

where

$$D_\vartheta x \in \mathbb{R}^{n\times q}, \quad [K_k^T N_\varepsilon'(K_k x)] = (K_1^T N_\varepsilon'(K_1 x), \dots, K_q^T N_\varepsilon'(K_q x)) \in \mathbb{R}^{n\times q}$$

and

$$N_\varepsilon''(K_k x) = \operatorname{diag}(n_\varepsilon''((K_k x)_1), \dots, n_\varepsilon''((K_k x)_m)) \in \mathbb{R}^{m\times m}.$$

Let ϑ_ε denote a solution to (4.4), which exists under the assumption of Proposition 4.1. Then, the first-order optimality condition for (4.4) is given by

$$D_\vartheta \mathcal{E}(x_\varepsilon(\vartheta_\varepsilon))(\vartheta - \vartheta_\varepsilon) = 2\left\langle x_\varepsilon(\vartheta_\varepsilon) - g, D_\vartheta x_\varepsilon(\vartheta_\varepsilon)(\vartheta - \vartheta_\varepsilon)\right\rangle \geq 0 \quad \text{for all } \vartheta \geq 0. \tag{4.9}$$

To eliminate $D_\vartheta x_\varepsilon$ from the first-order condition (4.9) we introduce the adjoint equation

$$p + \sum_{k=1}^{q} \vartheta_k K_k^T N_\varepsilon''(K_k x) K_k p = -(x_\varepsilon(\vartheta_\varepsilon) - g). \tag{4.10}$$

Since $n_\varepsilon'' \geq 0$, the adjoint equation admits a unique solution. Taking the inner product of (4.8) with p and of (4.10) with $D_\vartheta x(\vartheta_\varepsilon)$ we obtain

$$D_\vartheta \mathcal{E}(x_\varepsilon(\vartheta_\varepsilon))(\vartheta - \vartheta_\varepsilon) = 2\left\langle p, [K_k^T N_\varepsilon'(K_k x_\varepsilon(\vartheta_\varepsilon))](\vartheta - \vartheta_\varepsilon)\right\rangle \geq 0 \quad \text{for all } \vartheta \geq 0 \tag{4.11}$$

or, equivalently,

$$\left\langle N_\varepsilon'(K_k x_\varepsilon(\vartheta_\varepsilon)), K_k p\right\rangle (\vartheta_k - \vartheta_{\varepsilon,k}) \geq 0 \quad \text{for all } \vartheta_k > 0,\ k = 1, \dots, q. \tag{4.12}$$

Summarizing, the necessary optimality condition for the regularized problem is given by

$$\begin{cases} x_\varepsilon + \sum_{k=1}^{q} \vartheta_{\varepsilon,k} K_k^T N_\varepsilon'(K_k x_\varepsilon) = f \quad \text{(primal equation)}\ , \\ p_\varepsilon + \sum_{k=1}^{q} \vartheta_{\varepsilon,k} K_k^T N_\varepsilon''(K_k x_\varepsilon) K_k p_\varepsilon = -(x_\varepsilon - g) \quad \text{(adjoint equation)}\ , \\ \left\langle N_\varepsilon'(K_k x_\varepsilon), K_k p_\varepsilon\right\rangle (\vartheta_k - \vartheta_{\varepsilon,k}) \geq 0 \quad \text{for all } \vartheta_k \geq 0,\ k = 1, \dots, q \quad \text{(optimality)}. \end{cases} \tag{4.13}$$

The last expression in (4.13) can be expressed equally well as $N_\varepsilon'(K_k x)^T K_k p \in -\partial I_{\mathbb{R}^+}(\vartheta_k^*)$, where $I_{\mathbb{R}^+}$ is the indicator function of $\mathbb{R}^+$ and $\partial I_{\mathbb{R}^+}(\vartheta_k^*)$ denotes the subdifferential evaluated at ϑ_k^*, $k = 1, \dots, q$. To obtain an optimality system for the original problem (4.1) we shall pass to the limit $\varepsilon \to 0^+$ in (4.13). A similar procedure was used in [9] in the context of optimal control of a Bingham fluid; in that case the minimization variable appeared as affine, rather than as a multiplicative term like in our case, and a different type of regularization was used. Alternatively a first-order condition can be obtained by using the Mordukhovich calculus; compare [23] for mathematical programming problems with equilibrium constraints.

Theorem 4.3. *Let $\vartheta^* \geq 0$ denote a solution to (4.1) with associated state $x^* = x(\vartheta^*)$. Then there exist an adjoint state $p \in \mathbb{R}^n$ and multipliers $\lambda_k \in \mathbb{R}^m$, $k = 1, \dots, q$, and $\xi \in \mathbb{R}^n$ satisfying the following optimality system:*

$$\begin{cases} x^* + \sum_{k=1}^{q} \vartheta_k^* K_k^T \lambda_k = f, \\ (\lambda_k)_i \in \begin{cases} \operatorname{sgn}(K_k x^*)_i & \text{if } (K_k x^*)_i \neq 0, \\ [-1, 1] & \text{if } (K_k x^*)_i = 0, \end{cases} \\ p + \xi = -(x^* - g), \\ \langle \lambda_k, K_k p\rangle \in -\partial I_{\mathbb{R}^+}(\vartheta_k^*), \\ \langle \xi, p\rangle \geq 0, \\ \langle x^* - g, p\rangle \leq 0, \\ \langle \xi, x^*\rangle = 0, \\ (K_k p)_i = 0 \quad \text{if } |(\lambda_k)_i| < 1 \quad \text{for } k = 1, \dots, q,\ i = 1, \dots, m. \end{cases} \tag{4.14}$$

Proof. The proof of Theorem 4.3 is given in the appendix. ∎

Remark 4.4. Before closing this subsection we comment on the chosen regularization n_ε of the norm function in (4.5) by comparing to other choices that were made in related cases. The optimality condition for the lower-level problem in (4.4) with $q = 1$ is given by

$$x + \vartheta K^T \lambda = f,$$

where $\lambda_i \in \partial(|(Kx)_i|)$, which can also be expressed as

$$\begin{cases} x + \vartheta K^T \lambda = f, \\ |Kx| \otimes \lambda = Kx,\ |\lambda|_\infty \leq 1, \end{cases} \tag{4.15}$$

where $a \otimes b = (a_1 b_1, \dots, a_n b_n)$. The same system is obtained by Fenchel dualization of the lower-level problem in (4.4) with λ chosen as the dual variable. A regularization of this primal-dual formulation is obtained by replacing coordinatewise the norm operation $|t|$ in (4.15) by $\tilde{n}_\varepsilon(t) = \sqrt[2]{t^2 + \varepsilon}$. Such an approach was used for TV-regularized problems in [5], and it is also related to the taut string algorithm, as pointed out in [13]. Alternatively a localized regularization can be chosen by setting

$$\hat{n}_\varepsilon(t) = \begin{cases} \dfrac{1}{2\varepsilon} t^2 + \dfrac{\varepsilon}{2} & \text{if } |t| < \varepsilon, \\ |t| & \text{else,} \end{cases} \tag{4.16}$$

as used in [9, 14], for example. Let us compute to which kind of regularized primal formulation the primal-dual formulation (4.15) regularized by (4.16) would lead; i.e., we replace the generalized derivative $\lambda \in \partial(|Kx|)$ by $\frac{Kx}{\hat{n}_\varepsilon(Kx)}$, coordinatewise, and compute the antiderivatives to obtain a new regularization $\tilde{n}_\varepsilon(|Kx|)$. Carrying this out coordinatewise we obtain

$$\tilde{n}_\varepsilon(t) = \begin{cases} \varepsilon \left[\log\left(\dfrac{\varepsilon}{2} + \dfrac{t^2}{2\varepsilon} \right) - \log\left(\dfrac{\varepsilon}{2} \right) \right] & \text{if } |t| < \varepsilon, \\ (|t| + \varepsilon(\log(2) - 1)) & \text{else.} \end{cases} \tag{4.17}$$

This regularization of $n(t)$ is again C^2-regular with monotone derivative $\tilde{n}'_\varepsilon$, which is essential for the solvability of the necessary condition associated with the lower-level problem. Differently from (4.4), $\tilde{n}_\varepsilon(t)$ acts globally and the expressions for the derivations are rational functions rather than polynomials. Thus we prefer (4.4) over (4.17).

4.3. Necessary second-order optimality condition. Here we derive a second-order necessary condition for local solutions of (4.4). Beyond the intrinsic relevance for describing the structure of the second-order necessary condition, its discussion is motivated by the fact that we introduce a second-order sufficient condition in the following subsection in order to analyze a semismooth Newton method for solving (4.13). Of course, it is desirable that the gap between the necessary and sufficient optimality conditions is small. We henceforth drop the dependence of (ϑ, x, p), a solution to (4.13), on $\varepsilon > 0$.

In principle, the derivation of the second-order conditions is quite standard; see, e.g., [21, section 10.5]. Our situation, however, is complicated due to the lack of second-order

smoothness of the equality constraint in (4.4). Second-order conditions for general semismooth optimization problems were investigated, for instance, in [6]. Our situation here is somewhat different, however. First, only the constraints lack sufficient regularity, while the objective functional is regular, and, second, the null-space representation of the linearized equality has a special structure since the variables $x \in \mathbb{R}^n$ can be represented in terms of $\vartheta \in \mathbb{R}^n$. It is therefore appropriate to give an independent derivation.

Let $\overline{\vartheta}$ denote a local solution to (4.4) with associated state $\overline{x} = x(\overline{\vartheta})$ (i.e., the dependence of the solution on $\varepsilon > 0$ is dropped here). We denote the set of strongly active indices by

$$\overline{\mathcal{A}}_S = \{k : \langle N_\varepsilon'(K_k\overline{x}), K_k p\rangle > 0\}.$$

On this set $\overline{\vartheta}_k = 0$ is determined by the necessary conditions. The critical cone for the necessary second-order condition is defined by

$$\overline{\mathcal{C}} = \{\vartheta \in \mathbb{R}^q : \vartheta_k = 0 \text{ for } k \in \overline{\mathcal{A}}_S,\ \vartheta_k \geq 0 \text{ if } \overline{\vartheta}_k = 0\}.$$

For any $\hat{\vartheta} \in \overline{\mathcal{C}}$ we have $\overline{\vartheta} + t\hat{\vartheta} \geq 0$ for all $t \geq 0$ sufficiently small. For convenience we also recall the primal equation

$$x + \sum_{k=1}^{q} \vartheta_k K_k^T N_\varepsilon'(K_k x) = f. \tag{4.18}$$

The directional derivative of x with respect to ϑ at $\overline{\vartheta}$ in direction $\hat{\vartheta}$ is denoted by $\dot{x} \in \mathbb{R}^n$. It satisfies

$$L_1\dot{x} + L_2\hat{\vartheta} = 0. \tag{4.19}$$

Here $L_1 \in \mathbb{R}^{n\times n}$ and $L_2 \in R^{n\times q}$ are given by

$$L_1 = I + \sum_{k=1}^{q} \overline{\vartheta}_k K_k^T N_\varepsilon''(K_k\overline{x})K_k, \quad L_2 = (K_q^T N'(K_q\overline{x}), \dots, K_q^T N'(K_q\overline{x})).$$

We shall need the third derivatives of $t \to n_\varepsilon((K_k x(\bar{\vartheta}+t\hat{\vartheta}))_i)$ at $t = 0$, which requires attention in case $|(K_k x(\overline{\vartheta}))_i| = \varepsilon$. If $(K_k x(\vartheta^*))_i = \varepsilon$ and $\frac{d}{dt}((K_k x(\overline{\vartheta} + t\hat{\vartheta}))_i)|_{t=0} = (K_k\dot{x})_i > 0$, then by the formulas above (4.7) we have that the third-order directional derivative is 0; if, on the other hand, $(K_k\dot{x})_i < 0$, then the third-order right directional derivative is $n_\varepsilon'''((K_k x(\overline{\vartheta}))_i) = -\frac{3}{\varepsilon^3}(K_k x(\overline{\vartheta}))_i$. Finally, if $(K_k x(\dot{\vartheta}))_i = 0$, then the third-order right directional derivative is multivalued with values in $\{0, -\frac{3}{\varepsilon^3}, (K_k x(\overline{\vartheta}))_i\}$. Summarizing, if $n_\varepsilon((K_k x(\overline{\vartheta}))_i) = \varepsilon$, then we denote the third-order directional derivative of $t \to n_\varepsilon((K_k x(\overline{\vartheta} + t\hat{\vartheta}))_i)|_{t=0}$ by $n'''_{\varepsilon,\hat{\vartheta}}(K_k x(\overline{\vartheta}))_i$, and it is given by

$$n'''_{\varepsilon,\hat{\vartheta}}((K_k x(\overline{\vartheta}))_i) \in \begin{cases} 0 \text{ if } (K_k\dot{x})_i > 0, (K_k x(\overline{\vartheta}))_i = \varepsilon, \\ -\dfrac{3}{\varepsilon^3}(K_k x(\overline{\vartheta}))_i \text{ if } (K_k\dot{x})_i < 0, (K_k x(\overline{\vartheta}))_i = \varepsilon, \\ \left\{0, -\dfrac{3}{\varepsilon^3}(K_k x(\overline{\vartheta}))_i\right\} \text{ if } (K_k\dot{x})_i = 0, (K_k x(\overline{\vartheta}))_i = \varepsilon. \end{cases}$$

We shall see that the expression $n'''_{\varepsilon,\hat\vartheta}(K_k x(\overline\vartheta))_i$ always appears as a factor with $(K_k\dot x(\overline\vartheta))_i$, and we note that the expression $n'''_{\varepsilon,\hat\vartheta}((K_k x(\overline\vartheta))_i)(K_k\dot x(\overline\vartheta))_i$ is single valued. The expressions for the case $(K_k x(\overline\vartheta))_i = -\varepsilon$ can be derived in a completely analogous manner. If $|(K_k x(\overline\vartheta))_i| \neq \varepsilon$, then the third derivative of $t \to n_\varepsilon(K_k x(\overline\vartheta + t\hat\vartheta))_i|_{t=0}$ is clearly well defined. It is again denoted by $n'''_{\varepsilon,\hat\vartheta}((K_k x(\overline\vartheta))_i)$. The expression for the third-order right directional derivative of $t \to N_{\varepsilon,\hat\vartheta}(K_k x(\overline\vartheta))$ is obtained from

$$N'''_{\varepsilon,\hat\vartheta}(K_k x(\overline\vartheta)) = \operatorname{diag}\big(n'''_{\varepsilon,\hat\vartheta}((K_k x(\overline\vartheta))_1), \dots, n'''_{\varepsilon,\hat\vartheta}((K_k x(\overline\vartheta))_m)\big).$$

Associated with the local solution $(\overline x, \overline\vartheta)$ we recall the adjoint equation, which we now express as

$$L_1 p = -(\overline x - g). \tag{4.20}$$

Finally we introduce the Lagrangian associated with (4.1)

$$\hat{\mathcal{L}}(x,\vartheta,p) = \frac{1}{2}\|x-g\|_2^2 + \left\langle p, x + \sum_{k=1}^q \vartheta_k K_k^T N_\varepsilon'(K_k x) - f \right\rangle.$$

We are now prepared to establish a second necessary condition for (4.1) at $(\overline x, \overline\vartheta)$.

Theorem 4.5 (second-order necessary condition). *With the notation for $N'''_{\varepsilon,\hat\vartheta}$ introduced above we have*

$$0 \le (\dot x^T, \hat\vartheta^T) \begin{pmatrix} I + \sum_{i=1}^q \overline\vartheta_i K_i^T N'''_{\varepsilon,\hat\vartheta}(K_i\overline x)\operatorname{diag}(Kp)K_i & R \\ R^T & 0 \end{pmatrix} \begin{pmatrix} \dot x \\ \hat\vartheta \end{pmatrix}$$

for each $\overline\vartheta \in \overline{\mathcal{C}}$. Here $R \in \mathbb{R}^{n\times q}$ is given by

$$R = (K_1^T N_\varepsilon''(K_1\overline x)K_1 p, \dots, K_q^T N_\varepsilon''(K_q\overline x)K_q p)$$

and $\dot x$ satisfies (4.19).

Proof. Let $\hat\vartheta \in \overline{\mathcal{C}}$, set $\vartheta(t) = \overline\vartheta + t\hat\vartheta$, and let $x = x(t)$ denote the solution to

$$x + \sum_{k=1}^q (\overline\vartheta_k + t\hat\vartheta_k) K_k^T N_\varepsilon'(K_k x) = f, \tag{4.21}$$

where $t > 0$, and $\dot x$ is the solution to (4.19). In the following computation it is assumed that t is sufficiently small so that $\overline\vartheta + t\hat\vartheta \ge 0$ and such that $\mathcal{E}(\vartheta(t)) \ge \mathcal{E}(\overline\vartheta)$. Consequently we have

$$0 \le \hat{\mathcal{L}}(x(t),\vartheta(t),p) - \hat{\mathcal{L}}(\overline x,\overline\vartheta,p), \tag{4.22}$$

and, moreover,

$$\nabla_x \hat{\mathcal{L}}(\overline x,\overline\vartheta,p) = 0.$$

Therefore we find, using $|a|^2 - |b|^2 - 2\langle a-b, b\rangle = |a-b|^2$, and $\hat{\vartheta} \in \overline{\mathcal{C}}$ in the first equality below, that

$$
\begin{aligned}
0 &\leq \hat{\mathcal{L}}(x(t), \vartheta(t), p) - \hat{\mathcal{L}}(\overline{x}, \overline{\vartheta}, p) - \left\langle \nabla_x \hat{\mathcal{L}}(\overline{x}, \overline{\vartheta}, p), x(t) - \overline{x} \right\rangle \\
&= \frac{1}{2}||x(t) - g||_2^2 - \frac{1}{2}||\overline{x} - g||_2^2 - \langle x(t) - \overline{x}, \overline{x} - g\rangle \\
&\quad + \left\langle p, \sum_{k=1}^{q} (\overline{\vartheta}_k + t\hat{\vartheta}_k) K_k^T N_\varepsilon'(K_k x(t)) - \sum_{k=1}^{q} \overline{\vartheta}_k K_k^T (N_\varepsilon'(K_k \overline{x})) \right. \\
&\qquad \left. - \sum_{k=1}^{q} t\hat{\vartheta}_k K_k^T N_\varepsilon'(K_k \overline{x}) - \sum_{k=1}^{q} \overline{\vartheta}_k K_k^T N''(K_k \overline{x}) K_k (x(t) - \overline{x}) \right\rangle \\
&= \frac{1}{2}||x(t) - \overline{x}||^2 + \left\langle p, \sum_{k=1}^{q} \overline{\vartheta}_k K_k^T (N_\varepsilon'(K_k x(t)) - N_\varepsilon'(K_k \overline{x}) - N_\varepsilon''(K\overline{x}) K_k (x(t) - \overline{x})) \right\rangle \\
&\qquad + \left\langle p, \sum_{k=1}^{q} t\hat{\vartheta}_k K_k^T (N_\varepsilon'(K_k x(t)) - N_\varepsilon'(K_k \overline{x})) \right\rangle.
\end{aligned}
$$

By the discussion preceding the statement of the theorem we obtain that

$$
\begin{aligned}
&\lim_{t \to 0^+} \frac{1}{t^2} (N_\varepsilon'(K_k x(t)) - N_\varepsilon'(K_k \overline{x}) - N_\varepsilon''(K_k \overline{x}) K_k (x(t) - x)) \\
&\quad = \lim_{t \to 0^+} \frac{1}{t^2} \big(N_\varepsilon'(K_k x(t)) - N_\varepsilon'(K_k \overline{x}) - N_\varepsilon''(K_k \overline{x}) K_k (x(t) - \overline{x}) \\
&\qquad - \frac{1}{2} N_\varepsilon'''(K_k \overline{x}) \mathrm{diag}(K_k (x(t) - \overline{x})) K_k (x(t) - \overline{x}) \\
&\qquad\quad + \frac{1}{2} N_{\varepsilon,\hat{\vartheta}}'''(K_k \overline{x}) \mathrm{diag}(K_k (x(t) - \overline{x})) K_k (x(t) - \overline{x}) \big) \\
&\quad = \frac{1}{2} N_{\varepsilon,\hat{\vartheta}}'''(K_k \overline{x}) \mathrm{diag}(K_k \dot{x}) K_k \dot{x}.
\end{aligned}
$$

As a consequence we have

$$
\begin{aligned}
0 &\leq \frac{1}{2}||\dot{x}||^2 + \frac{1}{2}\left\langle p, \sum_{k=1}^{q} \overline{\vartheta}_k K_k^T N_{\varepsilon,\hat{\vartheta}}'''(K_k \overline{x}) \mathrm{diag}(K_k \dot{x}) K_k \dot{x} \right\rangle + \left\langle p, \sum_{k=1}^{q} \hat{\vartheta}_k K_k^T N_\varepsilon''(K_k \overline{x}) K_k \dot{x} \right\rangle \\
&= \frac{1}{2}||\dot{x}||^2 + \frac{1}{2} \sum_{k=1}^{q} \overline{\vartheta}_k \left\langle \mathrm{diag}(K_k p) N_{\varepsilon,\hat{\vartheta}}'''(K_k \overline{x})(K_k \dot{x}), K_k \dot{x} \right\rangle \\
&\quad + \sum_{k=1}^{q} \overline{\vartheta}_k \langle K_k^T N_\varepsilon''(K_k \overline{x}) K_k p, \overline{x} \rangle,
\end{aligned}
$$

which can be expressed as

$$
0 \leq (\dot{x}^T, \hat{\vartheta}^T) \begin{pmatrix} I + \sum_{k=1}^{q} \overline{\vartheta}_k K_k^T N_{\varepsilon,\hat{\vartheta}}'''(K_k \overline{x}) \mathrm{diag}\,(K_k p) K_k & R \\ R^T & 0 \end{pmatrix} \begin{pmatrix} \dot{x} \\ \hat{\vartheta} \end{pmatrix},
$$

as desired. ∎

From the discussion before Theorem 4.5 we recall that the coordinates of $N'''_{\varepsilon,\hat{\vartheta}}(K_i\hat{x})K_i\dot{x}$ and hence of $N'''_{\varepsilon,\hat{\vartheta}}(K_i\hat{x})\mathrm{diag}(Kp)K_i\dot{x}$ are single valued.

4.4. Semismooth Newton algorithms. In this subsection a semismooth Newton method for solving the necessary optimality system (4.13) for the regularized problem (4.4) is developed and analyzed. Clearly, due to the regularization of the ℓ_1-norm, we obtain only an approximate solution of the original problem, but convergence for $\varepsilon \to 0$ to the ℓ^1 problem was established above. An exact error analysis in terms of ε is left for future work. Convergence of the regularized problem to the original one was already studied in Theorem 4.3. We utilize that the optimality condition in (4.11) can equivalently be expressed by means of the complementarity formulation

$$\begin{aligned}(\langle N'_\varepsilon(K_kx), K_kp\rangle)_{q\times 1} - \mu &= 0,\\ \mu - \max(0, \mu - \vartheta) &= 0,\end{aligned}$$

where

$$(\langle N'_\varepsilon(K_kx), K_kp\rangle)_{q\times 1} = (\langle N'_\varepsilon(K_1x), K_1p\rangle, \dots, \langle N'_\varepsilon(K_rx), K_rp\rangle),$$

and max operates coordinatewise.

System (4.11) can therefore be expressed equivalently as

$$G(x, \vartheta, p, \mu) = 0, \tag{4.23}$$

where

$$G(x,\vartheta,p,\mu) = \begin{pmatrix} p + \sum_{k=1}^{q} \vartheta_k K_k^T N''_\varepsilon(K_kx)K_kp + x - g \\ (\langle N'_\varepsilon(K_kx),\, K_kp\rangle)_{q\times 1} - \mu \\ \sum_{k=1}^{q} \vartheta_k K_k^T N'_\varepsilon(K_kx) + x - f \\ \mu - \max(0, \mu - \vartheta) \end{pmatrix}.$$

The reason for exchanging the order of the equations is to create symmetries in the generalized Jacobian $J(x, \vartheta, p, \mu)$ of $G(x, \vartheta, p, \mu)$ to be specified below. In what follows we specify the value of $n'''(t)$ at $t = \pm\varepsilon$ to be 0. We could equally well take $\mp\frac{3}{\varepsilon^2}$. For $(x, \vartheta, p, \mu) \in \mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^n \times \mathbb{R}^q$ we define

$$\begin{aligned} &L_1(x,\vartheta) = I + \sum_{k=1}^{q} \vartheta_k K_k^T N''_\varepsilon(K_kx)K_k, \quad L_2(x) = (K_1^T N'_\varepsilon(K_1x), \dots, K_q^T N'_\varepsilon(K_qx)),\\ &R(x) = (K_1^T N''_\varepsilon(K_1x)K_1p, \dots, K_q^T N''_\varepsilon(K_qx)K_qp) \in \mathbb{R}^{n\times q},\\ &\mathrm{Max}'(0,\mu) = \mathrm{diag}(\max'(0,\mu_1), \dots, \max'(0,\mu_q)),\end{aligned}$$

where

$$\max'(0,\mu_k) = \begin{cases} 1 \text{ if } \mu_k > 0,\\ 0 \text{ if } \mu_k \le 0.\end{cases}$$

We note that there exists a neighborhood U. We find that the generalized Jacobian of G is given by

$$J(x,\vartheta,p,\mu) = \begin{pmatrix} Q(x,\vartheta,p) & R(x) & L_1(x,\vartheta) & 0 \\ R(x)^T & 0 & L_2^T(x) & -I \\ L_1(x,\vartheta) & L_2(x) & 0 & 0 \\ 0 & \mathrm{Max}'(0,\mu-\vartheta) & 0 & I-\mathrm{Max}'(0,\mu-\vartheta) \end{pmatrix}, \tag{4.24}$$

where

$$Q(x,\vartheta,p) = I + \sum_{k=1}^{q} \vartheta_k K_k^T N_\varepsilon'''(K_k x)\,\mathrm{diag}(K_k p)\,K_k.$$

A positive definiteness assumption of the upper 2×2 block of $J(x,\vartheta,p,\mu)$ will be required. Let $(\overline{x},\overline{\vartheta},\overline{p},\overline{\mu})$ denote a solution to $G(x,\vartheta,p,\mu)=0$. Furthermore, let $U = U(\overline{x},\overline{\vartheta},\overline{p},\overline{\mu})$ denote an open neighborhood of $(\overline{x},\overline{\vartheta},\overline{p},\overline{\mu})$, and set

$$\mathcal{A}(\vartheta,\mu) = \{k : \mu_k - \vartheta_k > 0\}, \tag{4.25}$$

and

$$\mathcal{C} = \{\delta\vartheta \in \mathbb{R}^q : \delta\vartheta_k = 0 \text{ if } k \in \mathcal{A}(\vartheta,\mu)\}.$$

Note that at the solution we have $\overline{\vartheta}_k\overline{\mu}_k = 0$, and $\overline{\mu}_k \geq 0, \overline{\vartheta}_k \geq 0$, and hence $\mathcal{A}(\overline{\vartheta},\overline{\mu}) = \{k : \overline{\mu}_k > 0\}$ coincides with the strongly active set of section 3.3. We shall utilize the following assumption:

(H1)
$$\begin{cases} \text{There exists a bounded neighborhood } U = U(\overline{x},\overline{\vartheta},\overline{p},\overline{\mu}) \\ \text{such that for all } (x,\vartheta,p,\mu) \in U \\ (\delta x^T, \delta\vartheta^T)\begin{pmatrix} Q(x,\vartheta,p) & R(x) \\ R(x)^T & 0 \end{pmatrix}\begin{pmatrix}\delta x \\ \delta\vartheta\end{pmatrix} > 0 \\ \text{for all } \ \vartheta \in \mathcal{C} \ \text{ and } \ L_1(x,\vartheta)\delta x + L_2(x)\delta\vartheta = 0. \end{cases}$$

Note that μ does not appear in the positive definite condition, but boundedness of the μ-coordinates in U will be used below.

Comparing (H1) to the second-order necessary condition of Theorem 4.5 we note that (H1) requires positive definiteness in a neighborhood of $(\overline{x},\overline{\vartheta},\overline{p})$, that perturbations of the active set need to be admitted, and that the values of n_ε''' at $t = \pm\varepsilon$ are fixed, whereas in Theorem 4.5 they appear as directional derivatives. For the purpose of this subsection the choice of $\max'(0,\mu_k)$ could be 0 at $\mu_k = 0$. This would change $\mathcal{A}(\vartheta,\mu) = \{k : \mu_k - \vartheta_k \geq 0\}$, but the following convergence result would remain unchanged.

By (H1) and the fact that $t \to n_\varepsilon'''(t)$ has only finitely many discontinuities, there exists $\kappa > 0$ such that

$$(\delta x^T, \delta\vartheta^T)\begin{pmatrix} Q(x,\vartheta,p) & R(x) \\ R^T(x) & 0 \end{pmatrix}\begin{pmatrix}\delta x \\ \delta\vartheta\end{pmatrix} \geq \kappa \left\| \begin{pmatrix}\delta x \\ \delta\vartheta\end{pmatrix} \right\|_2^2$$

for all $\vartheta \in \mathcal{C},\ L_1(x,\vartheta)\delta x + L_2(x)\delta\vartheta = 0,$ and $(x,\vartheta,p,\mu) \in U$.

Proposition 4.6. *If* (H1) *holds, then* $J(x,\vartheta,p,\mu)$ *is regular for each* $(x,\vartheta,p,\mu)\in U$ *and the inverses are uniformly bounded.*

Proof. Let $(x,\vartheta,p,\mu)\in U$, set $\mathcal{A}=\mathcal{A}(\vartheta,\mu)$ as defined above, and let $\mathcal{I}=\{1,\dots,r\}\setminus\mathcal{A}$. We show that $J(x,\vartheta,p,\mu)$ is injective. Let $(\delta x,\delta\vartheta,\delta p,\delta\mu)\in\mathbb{R}^n\times\mathbb{R}^q\times\mathbb{R}^n\times\mathbb{R}^q$, and assume that

$$J(x,\vartheta,p,\mu)\begin{pmatrix}\delta x\\ \delta\vartheta\\ \delta p\\ \delta\mu\end{pmatrix}=0. \tag{4.26}$$

We partition $\delta\vartheta$ into coordinates associated with inactive $\delta\vartheta_{\mathcal{I}}$ and active $\delta\vartheta_{\mathcal{A}}$ coordinates, and similarly for $\delta\mu$. The columns of $R(x)$ corresponding to inactive coordinates are denoted by $R(x)_{\mathcal{I}}$, and analogously for $L_2(x)_{\mathcal{I}}$. Thus $R(x)_{\mathcal{I}}$ is of dimension $n\times\#(\mathcal{I})$, where $\#(\mathcal{I})$ denotes the cardinality of $\mathcal{I}$. From the last equation in (4.26) we have

$$\delta\vartheta_{\mathcal{A}}=0\quad\text{ and }\quad\delta\mu_{\mathcal{I}}=0.$$

From the third equation in (4.26) we have

$$L_1(x,\vartheta)\delta x+L_2(x)\delta\vartheta=L_1(x,\vartheta)\delta x+L_2(x)_{\mathcal{I}}(\delta\vartheta)_{\mathcal{I}}=0. \tag{4.27}$$

Now from the first and second equations of (4.26)

$$\begin{aligned}Q(x,\vartheta,p)\delta x+R(x)_{\mathcal{I}}\delta\vartheta_{\mathcal{I}}&=-L_1^T(x,\vartheta)p,\\ (R(x)_{\mathcal{I}})^T\delta x&=-(L_2(x)_{\mathcal{I}})^Tp,\end{aligned} \tag{4.28}$$

where we use that $(\delta\mu)_{\mathcal{I}}=0$. Since by (H1) the matrix $\left(\begin{smallmatrix}Q & R_{\mathcal{I}}\\ R_{\mathcal{I}}^T & 0\end{smallmatrix}\right)$ is positive definite on $\ker\,(L_1(x,\vartheta),L_2(x)_{\mathcal{I}})$ and the right-hand side is in its orthogonal complement, we find that $\delta x=0$ and $\delta\vartheta_{\mathcal{I}}=0$. From the first equation in (4.26) we deduce that $\delta p=0$. The third equation, evaluated for the $\mathcal{A}$-coordinates, implies that $\delta\mu_{\mathcal{A}}=0$, and hence $J(x,\vartheta,p,\mu)$ is a regular matrix. Since U is bounded, its closure is compact. Now, by a compactness argument and the fact that $J(x,\vartheta,p,\mu)$ has at most finitely many discontinuities in x, uniform boundedness of the inverses follows. ∎

A semismooth Newton algorithm for solving $G(x,\vartheta,p,\mu)=0$ can now be specified.

Algorithm 4.1. Newton Learning for ℓ_1 (NL-ℓ_1).

(i) Choose $(x^0,\vartheta^0,p^0,\mu^0)\in\mathbb{R}^n\times\mathbb{R}^q\times\mathbb{R}^n\times\mathbb{R}^q$, and set $n=0$.

(ii) Solve $J(x^n,\vartheta^n,p^n,\mu^n)\begin{pmatrix}\delta x\\ \delta\vartheta\\ \delta p\\ \delta\mu\end{pmatrix}=-G(x^n,\vartheta^n,p^n,\mu^n)$.

(iii) Update $(x^{n+1},\vartheta^{n+1},p^{n+1},\mu^{n+1})=(x^n,\vartheta^n,p^n,\mu^n)+(\delta x,\delta\vartheta,\delta p,\delta\mu)$, set $n=n+1$, and goto (ii).

Theorem 4.7. *Let $(\overline{x}, \overline{\vartheta}, \overline{p}, \overline{\mu})$ denote a solution of $G(x, \vartheta, p, \mu) = 0$, and assume that (H1) holds. Then, Algorithm 4.1 converges locally superlinearly.*

Proof. Using well-known characterizations for semismooth functions (see, e.g., [30, page 27], it can be argued that G is semismooth. Together with uniform boundedness of the generalized Jacobians $J(x, \vartheta, p, \mu)$ in a neighborhood of $(\overline{x}, \overline{\vartheta}, \overline{p}, \overline{\mu})$, the claim follows; see, e.g., [30, page 29]. ∎

Algorithm 4.1, which arises as a Newton algorithm applied to the optimality condition, is frequently referred to as the sequential quadratic programming (SQP) algorithm. The algorithm can equivalently be obtained by iteratively minimizing a quadratic approximation to the cost and a linear (in x and ϑ) approximation to the constraining equation, which in our case is the necessary optimality condition to the lower-level problem. Algorithm 4.1 is closely related to applying a Newton algorithm to the reduced functional $\vartheta \to \mathcal{E}(x(\vartheta))$, where $x(\vartheta)$ satisfies the nonlinear constraining equation. They differ by the property that the primal updates of the Newton algorithm applied to the reduced functional $\vartheta \to \mathcal{E}(x(\vartheta))$ live on the nonlinear constraining manifold, while iterates of Algorithm 4.1 are contained in the tangent space to the constraint at the current iterate. It is well known that the former of these two algorithms can be obtained from the latter by introducing feasibility steps [16]. For the current problem this is given in Algorithm 4.2.

Algorithm 4.2. Reduced Newton Learning for ℓ_1 (RNL-ℓ_1).

(i) Choose $(\vartheta^0, \mu^0) \in \mathbb{R}^q \times \mathbb{R}^q$, and set $n = 0$.
(ii) Solve $x + \sum_{k=1}^{q} \vartheta^n K_k^T N_\varepsilon'(K_k x) = f$ for x^n (primal feasibility step).
(iii) Solve $L_1(x^n, \vartheta^n)p = -(x^n - g)$ for p^n (dual feasibility step).
(iv) Solve $J(x^n, \vartheta^n, p^n, \mu^n) \begin{pmatrix} \delta x \\ \delta \vartheta \\ \delta p \\ \delta \mu \end{pmatrix} = -G(x^n, \vartheta^n, p^n, \mu^n)$.
(v) Update $(\vartheta^{n+1}, \mu^{n+1}) = (\vartheta^n, \mu^n) + (\delta\vartheta, \delta\mu)$, set $n = n + 1$, and goto (ii).

Due to the feasibility steps the right-hand side of step (iii) in Algorithm 4.2 has the form

$$G(x, \vartheta, p, \mu) = \begin{pmatrix} 0 \\ (\langle N_\varepsilon'(K_k x),\, K_k p\rangle)_{q\times 1} - \mu \\ 0 \\ \mu - \max(0, \mu - \vartheta) \end{pmatrix}.$$

For any solution $(\overline{x}, \overline{\vartheta}, \overline{p}, \overline{\mu})$ to $G(x, \vartheta, p, \mu) = 0$ we have $\bar{\vartheta} \geq 0$. Hence $L_1(\bar{x}, \bar{\vartheta})$ is positive definite, and the implicit function theorem implies the existence of neighborhoods $U(\overline{x}) \times U(\overline{\vartheta})$ such that for each $\vartheta \in U(\overline{\vartheta})$ there exists a solution $x = x(\vartheta) \in U(\bar{x})$ satisfying

$$x + \sum_{k=1}^{q} \vartheta_k K_k^T N_\varepsilon'(K_k x) - f = 0. \tag{4.29}$$

Moreover, $\vartheta \to x(\vartheta)$ is continuously differentiable from $U(\overline{\vartheta})$ to $U(\bar{x})$. The solution in step (ii) of Algorithm 4.2 is chosen to satisfy $x^n \in U(\bar{x})$. Without loss of generality we may assume that $\{(x,\vartheta) : (x,\vartheta,p,\mu) \in U((\bar{x},\bar{\vartheta},\bar{p},\bar{\mu}))\} \subset U(\bar{x}) \times U(\bar{\vartheta})$, where $U(\bar{x},\bar{\vartheta},\bar{p},\bar{\mu})$ was introduced in (H1), and that $L_1(x,\vartheta)$ is regular with uniformly bounded inverses for all $(x,\vartheta) \in U(\bar{x}) \times U(\bar{\vartheta})$. For nonnegative ϑ this property is obviously satisfied for all x.

In numerical practice we switched from Algorithm 4.1 to Algorithm 4.2 for small values of epsilon (e.g., $\varepsilon \leq 10^{-2}$). Moreover, we used a reduced form of the system in step (iii) which will be detailed after addressing convergence for Algorithm 4.1.

Theorem 4.8. *Let $(\overline{x},\overline{\vartheta},\overline{p},\overline{\mu})$ be a solution to $G(x,\vartheta,p,\mu) = 0$. If (H1) holds, then the iterates of Algorithm 4.1 converge locally superlinearly.*

Proof. The proof can be given by standard arguments, and hence it suffices to give the main steps.

The iteration can be characterized by

$$z^n \to \hat{z}^{n+1} = z^n + \delta z \to z^{n+1} = (x^{n+1}, \vartheta^n + \delta\vartheta, p^{n+1}, \mu^n + \delta\mu), \tag{4.30}$$

where $z^n = (x^n, \vartheta^n, p^n, \mu^n)$, and $\delta z = (\delta x, \delta\vartheta, \delta p, \delta\mu)$ is the solution to the system in step (iii) in Algorithm 4.2. The first step in (4.30) is a semismooth Newton step, and hence

$$\|z^n + \delta z - \bar{z}\| = o(\|z^n - \bar{z}\|), \tag{4.31}$$

where $\bar{z} = (\bar{x},\bar{\vartheta},\bar{p},\bar{\mu})$, and the norm $\|\cdot\|$ is taken in $\mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^n \times \mathbb{R}^q$. Arguing iteratively, by (4.31) together with the Lipschitz estimates below, we find that the iterates $z^n \in U(\bar{x},\bar{\vartheta},\bar{p},\bar{\mu})$ if $\|(x^0,\vartheta^0) - (\bar{x},\bar{\vartheta})\|$ is sufficiently small.

Since $\vartheta \to x(\vartheta)$ is C^1 on $U(\bar{\vartheta})$, there exists a constant K_1 such that $\|x(\vartheta) - \bar{x}\| \leq K_1\|\vartheta - \bar{\vartheta}\|$ for all $\vartheta \in U(\vartheta)$, and in particular

$$\|x(\vartheta^{n+1}) - \bar{x}\| = \|x^{n+1} - \bar{x}\| \leq K_1\|\vartheta^{n+1} - \bar{\vartheta}\| = o(\|\vartheta^n - \bar{\vartheta}\|). \tag{4.32}$$

Moreover, we find that

$$L_1(x^{n+1},\vartheta^{n+1})(p^{n+1} - \bar{p}) = -(L_1(x^{n+1},\vartheta^{n+1}) - L_1(\bar{x},\bar{\vartheta}))\,\bar{p} + \bar{x} - x^{n+1},$$

and therefore there exists a constant K_2, independent of n, such that

$$\|p^{n+1} - \bar{p}\| \leq K_2\|(x^{n+1},\vartheta^{n+1}) - (\bar{x},\bar{\vartheta})\| = o(\|\vartheta^n - \bar{\vartheta}\|). \tag{4.33}$$

Combining (4.31)–(4.33) the claim follows. ∎

We next express step (iv) of Algorithm 4.2 in terms of the variables (ϑ,μ). From the first and third equations of (iv) we derive

$$\delta p = -L_1^{-1}\left(Q\,\delta x + R\,\delta\vartheta\right),$$
$$\delta x = -L_1^{-1}L_2\,\delta\vartheta.$$

Here and below we drop the dependence of L_1, L_2, R, and Q on the current iterate (x^n,ϑ^n,p^n). The second equation of (iv) gives

$$R^T\,\delta x + L_2^T\,\delta p - \delta\mu = -G_3. \tag{4.34}$$

Introducing the symmetric matrix

$$P(x^n, \vartheta^n, p^n) = L_2^T L_1^{-1} Q L_1^{-1} L_2 - R^T L_1^{-1} L_2 - L_2^T L_1^{-1} R$$

and δx and $\delta\mu$ in terms of $\delta\vartheta$ in (4.34) we obtain

$$P(x^n, \vartheta^n, p^n)\,\delta\vartheta - \delta\mu = -G_2(x^n, p^n, \mu^n).$$

Combined with the fourth equation in (iii) we obtain the asymmetric system

$$\begin{pmatrix} P(x^n, \vartheta^n, p^n) & -I \\ \mathrm{Max}'(0, \mu^n - \vartheta^n) & I - \mathrm{Max}'(0, \mu^n - \vartheta^n) \end{pmatrix} \begin{pmatrix} \delta\vartheta \\ \delta\mu \end{pmatrix} = \begin{pmatrix} -G_2(x^n, p^n, \mu^n) \\ -G_4(\vartheta^n, \mu^n) \end{pmatrix}. \tag{4.35}$$

The second equality in (4.35) can be expressed as

$$\mathrm{Max}'(0, \mu^n - \vartheta^n)\,\delta\vartheta + (I - \mathrm{Max}'(0, \mu^n - \vartheta^n))\,\delta\mu + \mu^n - \max(0, \mu^n - \vartheta^n) = 0.$$

This implies that

$$\vartheta_{\mathcal{A}}^{n+1} = 0 \quad \text{and} \quad \mu_{\mathcal{I}}^{n+1} = 0, \tag{4.36}$$

where $\mathcal{A} = \mathcal{A}(\vartheta^n, \mu^n)$ is defined in (4.25) with (ϑ, μ) replaced by (ϑ^n, μ^n) and the subscript $\mathcal{A}$ with $\vartheta_{\mathcal{A}}^{n+1}$ is defined in the proof of Proposition 4.6.

Finally we partition the coordinates into active and inactive ones, so that, after possible reordering, $x = (x_{\mathcal{I}}, x_{\mathcal{A}})$. Accordingly $P(x^n, \vartheta^n, p^n)$ is split into block matrices

$$P(x^n, \vartheta^n, p^n) = \begin{pmatrix} P(x^n, \vartheta^n, p^n)_{\mathcal{I}} & P(x^n, \vartheta^n, p^n)_{\mathcal{I},\mathcal{A}} \\ P(x^n, \vartheta^n, p^n)_{\mathcal{A},\mathcal{I}} & P(x^n, \vartheta^n, p^n)_{\mathcal{A}} \end{pmatrix}.$$

Thus (4.35) is equivalent to solving the symmetric system

$$P(x^n, \vartheta^n, p^n)_{\mathcal{I}}\,\delta\vartheta_{\mathcal{I}} = -\langle N_\varepsilon'(K_i x),\, K_i p\rangle_{\mathcal{I}} + P(x^n, \vartheta^n, p^n)_{\mathcal{I},\mathcal{A}}\vartheta_{\mathcal{A}}^n,$$

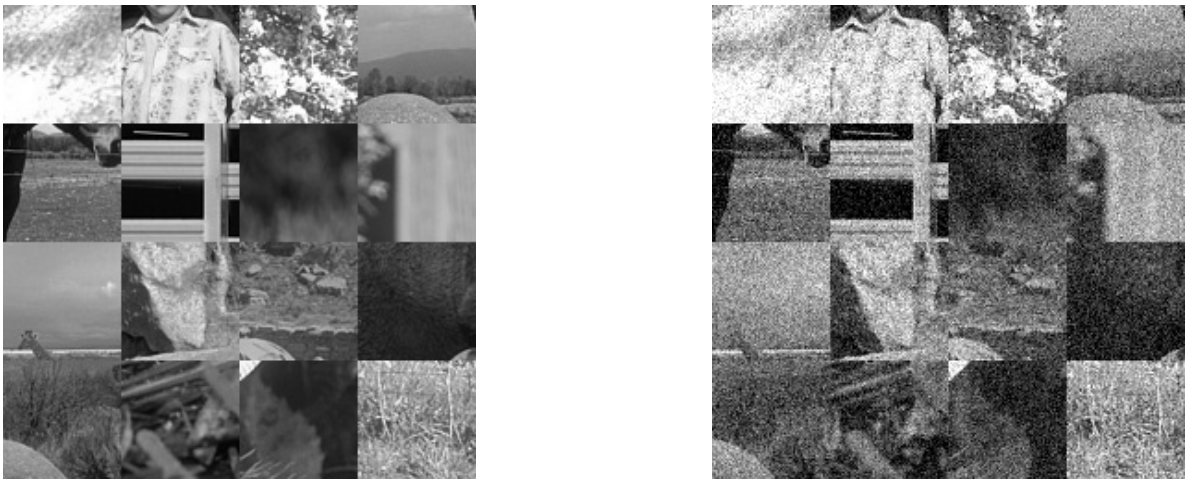
where we use that $\delta\vartheta_{\mathcal{A}} = -\vartheta_{\mathcal{A}}$, and assigning

$$\mu_{\mathcal{A}}^{n+1} = (P\,\delta\vartheta)_{\mathcal{A}} + \langle N_\varepsilon'(K_i x),\, K_i p\rangle_{\mathcal{A}},$$

and $\vartheta_{\mathcal{A}}^{n+1}$, $\mu_{\mathcal{I}}^{n+1} = 0$ according to (4.36).

5. Numerical realization. In our numerical experiments, we consider the problem of learning the optimal regularization parameters for the ℓ_1 model with multiple priors. For this purpose, we consider a training data base consisting of a set of clean images and their noisy versions that are generated by adding noise of different strengths to the clean images. Then, if the training data base is large enough, we can expect that we can compute some "universally optimal" parameters that lead to good denoising results, at least for images that are in the same class as the images in the training data base (e.g., natural images or medical images).

In section 5.1 we will present the results of learning the optimal parameters for different noise levels using different algorithms. Once we have computed the optimal parameters, the



(a) (b)

Figure 4. *Subset of the ground truth data g extracted from the BSDS500 database [1] and the noisy data f using a noise level of $\sigma = 25$.*

models can be tested on previously unseen images. Results for testing will be presented in section 5.2.

Having a set of training data (g_i, f_i), $i = 1 \dots, N$, the bilevel problem we are aiming to solve is given by

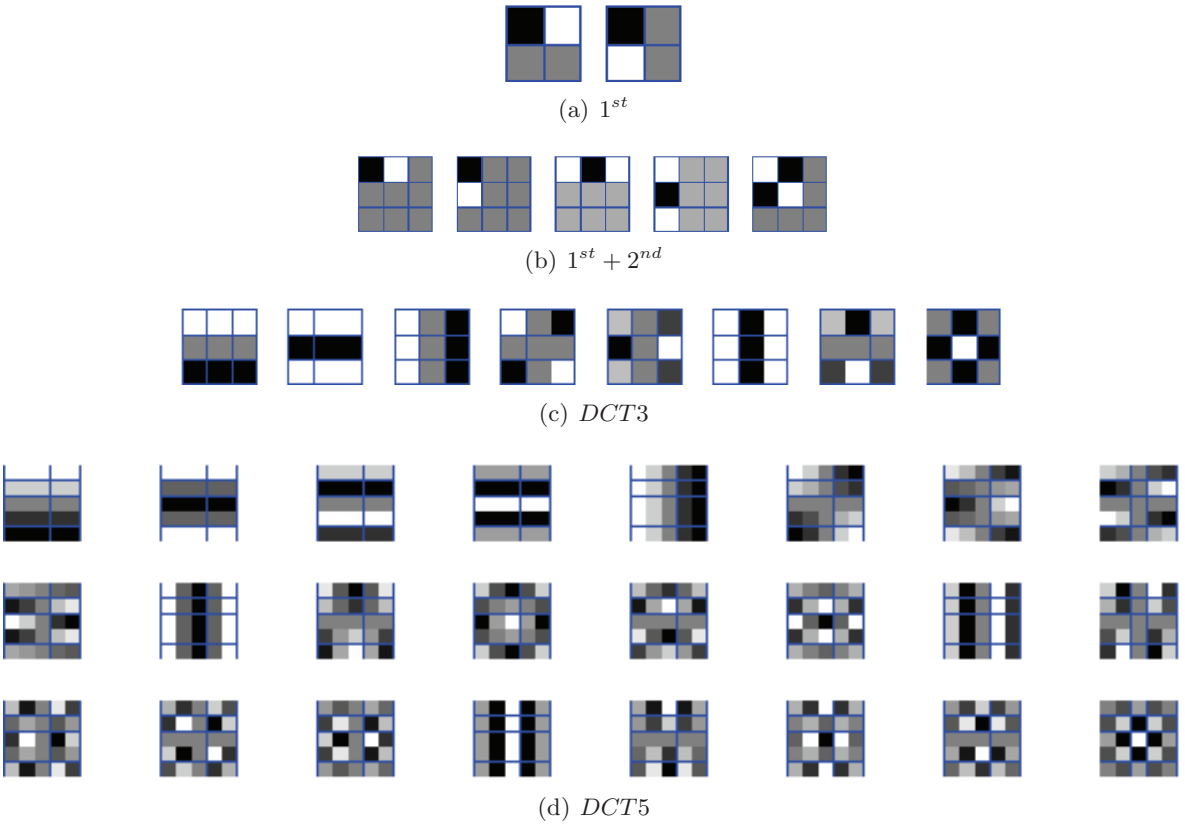
$$\begin{cases} \min\limits_{\vartheta \geq 0} \mathcal{E}(x(\vartheta)) = \sum\limits_{i=1}^{N} \|x_i(\vartheta) - g_i\|_2^2 \\ \text{subject to } x_i(\vartheta) = \operatorname*{argmin}\limits_{x} \sum\limits_{k=1}^{q} \vartheta_k g \|K_k x\|_1 + \dfrac{1}{2}\|x - f_i\|_2^2. \end{cases}$$

To generate the training data, we first randomly sample $N = 64$ patches of size $w \times h = 64 \times 64$ from the BSDS500 image segmentation database [1] and store them in vectors $g_i \in \mathbb{R}^{wh}$. The reason for sampling random patches in a large database is to generate samples of a large diversity by simultaneously minimizing the amount of training data. Then, we generate the noisy versions $f_i \in \mathbb{R}^{wh}$ by adding Gaussian noise with different standard deviations $\sigma \in \{15, 25, 50\}$ to g_i. Figure 4 shows an exemplary subset of the training data together with a noisy version.

In previous sections, we did not consider the case of multiple training data (g_i, f_i). However, we can easily convert the learning problem for multiple training data to the form (4.1) by stacking up all g_i and f_i to large vectors, i.e., $\tilde{g} = (g_1, \dots, g_N)$ and $\tilde{f} = (f_1, \dots, f_N)$, and by defining the linear operators $\widetilde{K}_k$ as the $N \times N$ block-diagonal matrices

$$\widetilde{K}_k = \operatorname{diag}(\underbrace{K_k, \dots, K_k}_{N \text{ times}}). \tag{5.1}$$

Then, we can treat the multiple training data problem as a single training data problem with



(a) 1^{st}

(b) $1^{st} + 2^{nd}$

(c) $DCT3$

(d) $DCT5$

Figure 5. *Different sets of filters used in the experiments.*

a modified linear operator $\widetilde{K}$ and the analysis carried out in the previous sections can be applied.

The linear operators $K_k \in \mathbb{R}^{m \times n}$ we consider in our experiments are generated from local filter kernels $\kappa_k \in \mathbb{R}^{\mu \times \nu}$ such that the matrix-vector product $K_k x$ is equal to the two-dimensional convolution of the two-dimensional image x with the filter kernel κ_k, i.e.,

$$K_k x = x * \kappa_k,$$

where $*$ denotes the two-dimensional convolution operation. Note that for the matrix-vector product $K_k x$, the image is treated as a column vector, whereas for the two-dimensional convolution with the filter kernels κ_k, the image is treated as a two-dimensional array.

For the filter kernels we consider various choices, e.g., standard finite difference approximations of first- and second-order derivatives or higher-order linear operators obtained from the basis vectors of the two-dimensional discrete cosine transform (DCT). Figure 5 shows the filter kernels we used in our experiments. For the boundary conditions, we modify the linear operators in a way such that the image data is reflected at the boundaries.

5.1. Learning. In the following sections we show how to learn the optimal regularization parameters ϑ in the multiple prior ℓ_1 model. We shall study two approaches: a first approach

that reduces the ℓ_1 learning problem to a sequence of reweighted ℓ_2 learning problems which will be solved using Algorithm 3.1 and a second approach that directly solves the ℓ_1 learning problem using the reduced Newton algorithm, Algorithm 4.2. We will compare the performances of both approaches and finally show preliminary extensions to solving the optimal parameters of a nonconvex $\ell_{\frac{1}{2}}$ model. All algorithms are implemented in MATLAB and are executed on a 2.60GHz i5 CPU running a 64bit Linux system.

5.1.1. Iteratively reweighted ℓ_2 learning. Motivated by the fixed-point algorithm for solving the lower-level ℓ_1 problems [32, 8], we consider a sequence of reweighted ℓ_2 problems for learning the optimal regularization parameters of the ℓ_1 problem.

Let n_ε be the ε regularized Huber-ℓ_1 norm

$$n_\varepsilon(t) = \begin{cases} \frac{t^2}{2\varepsilon} + \frac{\varepsilon}{2} & \text{if } |t| \leq \varepsilon, \\ |t| & \text{else.} \end{cases} \tag{5.2}$$

Given a point $\hat{t}$, we can bound $n_\varepsilon(t)$ from above via the quadratic function [3]

$$n_\varepsilon(t) \leq \frac{1}{2}\left(\frac{t^2}{\max(\varepsilon, |\hat{t}|)} + \max(\varepsilon, |\hat{t}|)\right).$$

Now, assume we are given an $\hat{x}$ which is sufficiently close to the optimal solution of the lower-level problem. We can then approximate the ℓ_1 bilevel learning problem as a quadratic single level problem

$$\min_{\vartheta \geq 0} \mathcal{E}(\vartheta) = \frac{1}{2}\left\| \left(I + \sum_{k=1}^{q} \vartheta_k \mathcal{K}_k(\hat{x})\right)^{-1} f - g \right\|^2,$$

where

$$\mathcal{K}_k(\hat{x}) = K_k^T \operatorname{diag}\left(\frac{1}{\max(\varepsilon, |(K_k\hat{x})_1|)}, \ldots, \frac{1}{\max(\varepsilon, |(K_k\hat{x})_m|)}\right) K_k$$

is the weighted linear operator. This motivates an iterative algorithm which starts with an initial estimate of $\hat{x}$ and then solves a sequence of quadratic single-level problems with iteratively updated versions of $\hat{x}$. The outline of the algorithm is presented in Algorithm 5.1. The most involved step in the algorithm is computing the solution of the weighted ℓ_2 single-level problem which is carried out by using the semismooth Newton algorithm, Algorithm 3.1. In our experiments, we observe that the Hessian matrix M involved in the Newton equation (3.20) can have negative eigenvalues, which means that the Newton direction is not a descent direction. In view of the higher-level function $\mathcal{E}(\vartheta)$ as depicted in Figure 1, this comes as no surprise given the concave behavior of $\mathcal{E}(\vartheta)$ away from zero. In this case we use a positive definite approximation of M by flipping the signs of the negative eigenvalues (see [22] for more details). It is important to point out that M is always positive definite when the iterate becomes sufficiently close to the optimal solution, which enables local superlinear convergence of the algorithm. During the iterations of Algorithm 3.1, we always perform full steps in μ and in ϑ_i for all $i \in \mathcal{A}^n$ and carry out an Armijo-type linesearch in ϑ_i for all $i \in \mathcal{I}^n$ using the function value of the higher-level optimization problem as the merit function. We set $\varepsilon = 10^{-3}$ in the Huber-regularized $|\cdot|$ function in (5.2). The iterations of the inner algorithms, Algorithm 3.1,

are stopped when a maximum number of inner iterations $\texttt{maxiter}_1 = 100$ is reached or the normalized residual, i.e., the ℓ_2-norm of the right-hand side of (3.20) divided by its number of elements, is less than a tolerance of $\texttt{tol}_1 = 10^{-6}$. We stop the iterates of the outer algorithms, Algorithm 5.1, when a maximum number of outer iterations $\texttt{maxiter}_2 = 100$ is reached or the normalized outer residual, i.e., the ℓ_2-norm of (3.20) using ϑ^n and recomputing $\hat{x}$, is below a tolerance of $\texttt{tol}_2 = 10^{-3}$.

Algorithm 5.1. Iteratively Reweighted Learning for ℓ_2 (IRL-ℓ_2).

(i) Set $n = 0, \vartheta = 0$, $\hat{x} = f$.

(ii) Compute $\mathcal{K}_k(\hat{x}) = K_k^T \operatorname{diag}(\frac{1}{\max(\varepsilon, |(K_k\hat{x})_1|)}, \dots, \frac{1}{\max(\varepsilon, |(K_k\hat{x})_m|)}) K_k$.

(iii) Solve

$$\vartheta^n = \arg\min_{\vartheta \geq 0} \mathcal{E}(\vartheta) = \frac{1}{2} \left\| \left(I + \sum_{k=1}^{q} \vartheta_k \mathcal{K}_k(\hat{x}) \right)^{-1} f - g \right\|^2$$

using Algorithm 3.1.

(iv) Compute $\hat{x} = (I + \sum_{k=1}^{q} \vartheta_k^n \mathcal{K}_k(\hat{x}))^{-1} f$.

(v) Set $n = n + 1$, and goto (ii).

5.1.2. Direct ℓ_1 learning. Next we discuss the reduced Newton learning algorithm for ℓ_1 problems as presented in Algorithm 4.2. In step (ii) of the algorithm, we need to perform the primal feasibility step, which amounts to computing the minimizer of the lower-level problem. For this, we use a standard primal Newton algorithm with an Armijo-type backtracking linesearch which takes on average 10–20 iterations to bring the normalized residual of the primal equation below a threshold of $\texttt{tol}_1 = 10^{-9}$. In step (iii) of the algorithm we need to compute the dual feasibility step, which we solve by using the MATLAB `mldivide` command. We again use a positive definite approximation of the matrix P in (4.35) in case it has negative eigenvalues by flipping the signs of the negative eigenvalues. Furthermore, we perform full steps on μ and ϑ_i for all $i \in \mathcal{A}^n$ and an Armijo-type backtracking linesearch on ϑ_i for all $i \in \mathcal{I}^n$ using the higher-level problem $\mathcal{E}(\vartheta)$ as the merit function. We set $\varepsilon = 10^{-3}$ in the fourth-order polynomial approximation of the $|\cdot|$ function in (4.5). We stop the algorithm when a maximum number of iterations $\texttt{maxiter} = 100$ is reached or the normalized residual, i.e., the ℓ_2-norm of the right-hand side of the Newton equation in step (iv) divided by its number of elements, is less than a tolerance of $\texttt{tol}_2 = 10^{-3}$.

5.1.3. Results. Table 1 shows the result of learning the optimal regularization parameters on natural images for various linear operators and noise levels using the iteratively reweighted ℓ_2 learning algorithm (IRL-ℓ_2) and the reduced Newton ℓ_1 learning algorithm (RNL-ℓ_1).

In general, one can see that the energy of the higher-level problem $\mathcal{E}(\vartheta)$ decreases with the diversity of the filter banks and, equivalently, that the quality of the ℓ_1 models increases with the diversity of the differentiation order included in the filter banks. Observe that the largest performance increase comes through adding second-order filters to the first-order derivative filters. We also performed experiments where we added first-order derivative filters to the DCT filter banks, and it happened that the weights of the first-order filters were set

Table 1

Results for the ℓ_1 learning algorithms on natural images. The table shows the number of Newton steps and the value of the higher-level problem $\mathcal{E}(\vartheta)$.

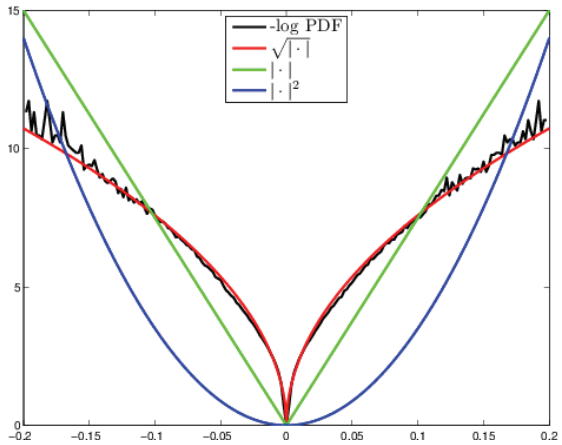
Algorithm	IRL-ℓ_2					
	$\sigma = 15$		$\sigma = 25$		$\sigma = 50$	
Filters	k	$\mathcal{E}(\vartheta)$	k	$\mathcal{E}(\vartheta)$	k	$\mathcal{E}(\vartheta)$
1^{st}	107	163.19	145	303.21	191	602.83
$1^{\text{st}} + 2^{\text{nd}}$	119	152.98	190	282.91	174	563.86
DCT3	132	148.79	141	272.60	183	545.46
DCT5	101	147.67	150	268.12	506	529.83
Algorithm	RNL-ℓ_1					
	$\sigma = 15$		$\sigma = 25$		$\sigma = 50$	
Filters	k	$\mathcal{E}(\vartheta)$	k	$\mathcal{E}(\vartheta)$	k	$\mathcal{E}(\vartheta)$
1^{st}	8	162.87	24	302.69	16	601,88
$1^{\text{st}} + 2^{\text{nd}}$	18	152.45	33	282.02	43	562.44
DCT3	12	147.55	20	270.62	37	542.90
DCT5	16	144.69	44	265.41	100	525.97

to zero by the learning algorithm. This experiment suggests that the first-order filters and hence the classical total variation prior are not very suitable for natural images. In contrast, we observed that on randomly generated piecewise constant images the learning algorithm always preferred first-order filters over any additional higher-order filter, which suggests that for piecewise constant images, the total variation is already a very good prior.

Comparing the results of the IRL-ℓ_2 and RNL-ℓ_1 algorithms, one can clearly see that RNL-ℓ_1 needs far fewer Newton steps to converge. This can be explained by the fact that the IRL-ℓ_2 algorithm performs a fixed-point iteration by solving a sequence of reweighted ℓ_2 learning problems, and hence the overall algorithm is in principle a first-order algorithm. In contrast, the RNL-ℓ_1 algorithm is a full Newton algorithm on the original ℓ_1 learning problem and hence exhibits superlinear convergence.

Furthermore, one can see that the RNL-ℓ_1 algorithm stops at slightly smaller energies. This is explained by the fact that for a fixed value of ε, the function (4.5) utilized in the RNL-ℓ_1 algorithm is a better approximation to the true $|\cdot|$ function than the Huber function (5.2) utilized in the IRL-ℓ_2 algorithm. We also tried to use a smaller ε in the IRL-ℓ_2 algorithm, which, however, led to convergence problems.

5.1.4. Learning of a nonconvex $\ell_{\frac{1}{2}}$ model. It is well known that the probability density function (PDF) of the responses of zero mean linear filters (e.g., DCT filters) on natural images has heavily tailed distributions [15]. Figure 6 plots the negative log PDF of the last DCT5 filter shown in Figure 5 applied to natural images together with different model fits. One can clearly see that the $|\cdot|^2$ function provides a bad fit to the negative log PDF, which is consistent with the inferior performance of quadratic energies for image regularization. Although the $|\cdot|$ function provides a much better fit than the $|\cdot|^2$ function, the $\sqrt{|\cdot|}$ function represents an almost perfect model. However, while the $|\cdot|$ function is still convex, the $\sqrt{|\cdot|}$ is nonconvex,



(a)

Figure 6. *Negative log PDF of the filter response of a DCT5 filter applied to natural images. Note that the* $\sqrt{|\cdot|}$ *function provides the best fit to the heavy-tailed shape of the true density function.*

which makes the lower problem much harder to solve.

Our aim is now to show that we can utilize the algorithms developed in this paper to learn the optimal regularization parameters of a model involving the nonconvex $\ell_{\frac{1}{2}}$ quasi norm. We shall see that this simple nonconvex model achieves excellent image denoising results very close to state-of-the-art algorithms.

The bilevel learning problem involving the nonconvex $\ell_{\frac{1}{2}}$ quasi norm is given by

$$\begin{cases} \min\limits_{\vartheta \geq 0} \mathcal{E}(x(\vartheta)) = \sum\limits_{i=1}^{l} \|x_i(\vartheta) - g_i\|_2^2 \\ \text{subject to } x_i(\vartheta) = \operatorname*{argmin}\limits_{x} \sum\limits_{k=1}^{q} \vartheta_k \|K_k x\|_{\frac{1}{2}}^{\frac{1}{2}} + \dfrac{1}{2}\|x - f_i\|_2^2, \end{cases}$$

where $\|K_k x\|_{\frac{1}{2}}^{\frac{1}{2}} = \sum_{i=1}^{n} \sqrt{|(K_k x)_i|}$. In order to apply our learning algorithms, we need to regularize the above problem. Similar to (4.5) we use a locally regularized approximation of the $\sqrt{|\cdot|}$ function:

$$n_\varepsilon(t) = \begin{cases} -\dfrac{3t^4}{32\sqrt{\varepsilon^7}} + \dfrac{7t^2}{16\sqrt{\varepsilon^3}} + \dfrac{21\sqrt{\varepsilon}}{32} & \text{if } |t| < \varepsilon, \\ \sqrt{|t|} & \text{else,} \end{cases} \tag{5.3}$$

with derivatives

$$n_\varepsilon'(t) = \begin{cases} -\frac{3t^3}{8\sqrt{\varepsilon^7}} + \frac{7t}{8\sqrt{\varepsilon^3}} & \text{if } |t| < \varepsilon, \\ \frac{t}{2\sqrt{|t|^3}} & \text{else,} \end{cases}$$

$$n_\varepsilon''(t) = \begin{cases} -\frac{9t^2}{8\sqrt{\varepsilon^7}} + \frac{7}{8\sqrt{\varepsilon^3}} & \text{if} \quad |t| < \varepsilon, \\ -\frac{1}{4\sqrt{|t|^3}} & \text{else}, \end{cases}$$

$$n_\varepsilon'''(t) = \begin{cases} -\frac{9t}{4\sqrt{\varepsilon^7}} & \text{if} \quad |t| < \varepsilon, \\ \frac{3t}{8\sqrt{|t|^7}} & \text{else}. \end{cases}$$

For learning, we use the reduced Newton algorithm, Algorithm 4.2, which can be easily adapted to the $\ell_{\frac{1}{2}}$ setting by replacing the regularized ℓ_1-norm with the regularized $\ell_{\frac{1}{2}}$–quasi norm. We term the resulting algorithm the reduced Newton $\ell_{\frac{1}{2}}$ learning algorithm (RNL-$\ell_{\frac{1}{2}}$).

In our experiments we observe that the Hessian matrix in the $\ell_{\frac{1}{2}}$ case can have strongly negative eigenvalues and that computing a positive definite approximation of the Hessian by simply flipping the signs of the negative eigenvalues does not always lead to a very good second-order approximation. This results in a worse convergence behavior of the algorithm. We stop the algorithm after the normalized residual is below a tolerance of $\texttt{tol} = 10^{-3}$ or a maximum number of iterations $\texttt{maxiter} = 100$ is reached. The investigation of an improved Newton algorithm, for example the development of a trust region Newton method, is subject to future work. For computing the primal feasibility step, we use the limited memory BFGS quasi-Newton method [20], where again for convergence reasons, we set $\varepsilon = 10^{-2}$ in the ε regularized $\sqrt{|\cdot|}$ function (5.3). The development of an algorithm that can handle smaller ε is left to future work. Table 2 shows the results of applying the RNL-$\ell_{\frac{1}{2}}$ to natural images using DCT3 and DCT5 filter banks and various noise levels. Observe that the RNL-$\ell_{\frac{1}{2}}$ algorithm takes significantly more iterations than the RNL-ℓ_1 algorithm. However, as already said, our predominant aim is to show the potential of the nonconvex $\ell_{\frac{1}{2}}$ model and hence also the limitations of the convex ℓ_1 model. Comparing the function values of $\mathcal{E}(\vartheta)$ using the $\ell_{\frac{1}{2}}$ models to the function values when using ℓ_1 models, as shown in Table 1, we can see that the nonconvex $\ell_{\frac{1}{2}}$ models lead to significantly lower function values, which means that the $\ell_{\frac{1}{2}}$ model can recover images which are closer to the ground truth images.

Table 2

Results for the $\ell_{\frac{1}{2}}$ learning algorithm on natural images. The table shows the number of Newton steps and the value of the higher level problem $\mathcal{E}(\vartheta)$.

Algorithm	RNL-$\ell_{\frac{1}{2}}$					
	$\sigma = 15$		$\sigma = 25$		$\sigma = 50$	
Filters	k	$\mathcal{E}(\vartheta)$	k	$\mathcal{E}(\vartheta)$	k	$\mathcal{E}(\vartheta)$
DCT3	47	134.02	100	253.35	100	527.13
DCT5	13	128.63	100	240.63	100	500.83

5.2. Testing. In this section we use the learned models from the last section to evaluate their denoising performance on unseen images from the BSDS500 database [1]. Furthermore, we will show comparisons to related methods as well as state-of-the-art algorithms.

In this work, we inherently assumed that the noise level of the images is known in advance. We point out that this assumption is reasonable also for practical problems since in many cases, the noise level can be computed from the image acquisition process, can be specified by the user, or can be estimated by separate algorithms [19].

Having given the noise level, we compute the solution of the lower-level problems using the first-order primal-dual algorithm [4] with the preconditioning described in [25] in the case of the ℓ_1 models and using the limited memory BFGS quasi-Newton method [20] in the case of the $\ell_{\frac{1}{2}}$ models. Note that for testing we require only a moderate accuracy of the minimizers of the lower-level problems, and hence we stopped the algorithms after the change of the function value was below a threshold of $\texttt{tol} = 10^{-3}$.

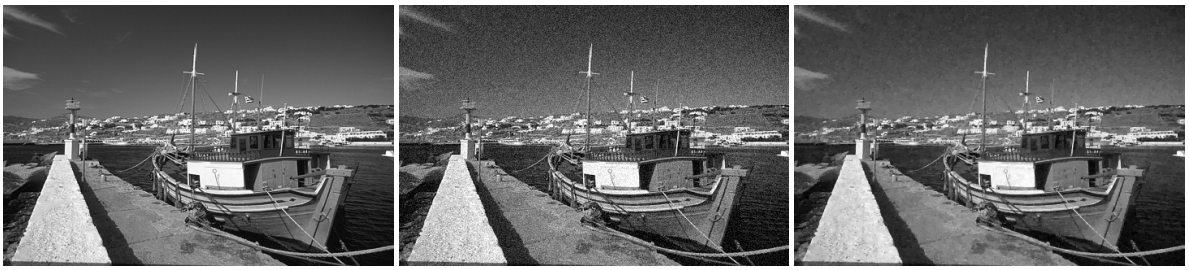
5.2.1. Results of the ℓ_1 model. Figures 7, 8, and 9 show the denoising results of the learned ℓ_1 models on natural images containing zero-mean Gaussian noise of various standard deviations, $\sigma \in \{15, 25, 50\}$. One can observe that larger filter banks consistently lead to a better image denoising performance, where, in particular, the DCT filters are much better in recovering textured areas. Furthermore, one can see that while the first-order filters lead to cartoon-like images (see the detail views in the last rows of the figures), the higher-order filters lead to much more naturally appearing results.

From the experiments we can observe an interesting limitation of the ℓ_1 models. While the step from simple first-order priors (i.e., the total variation) to higher-order priors (e.g., second-order derivatives or DCT3) gives the largest performance increase, the performance seems to saturate when further increasing the diversity of the filter banks (e.g., from DCT3 to DCT5), and hence we expect that the performance of ℓ_1 models cannot be improved much more by keep adding priors to the model. We do not think that this is due to a wrong set of priors (we also experimented with dictionary priors such as SVD and ICA priors) but is an inherent limitation of the ℓ_1 model. Indeed, we will see that switching from the convex ℓ_1 model to the nonconvex $\ell_{\frac{1}{2}}$ model will overcome this limitation.

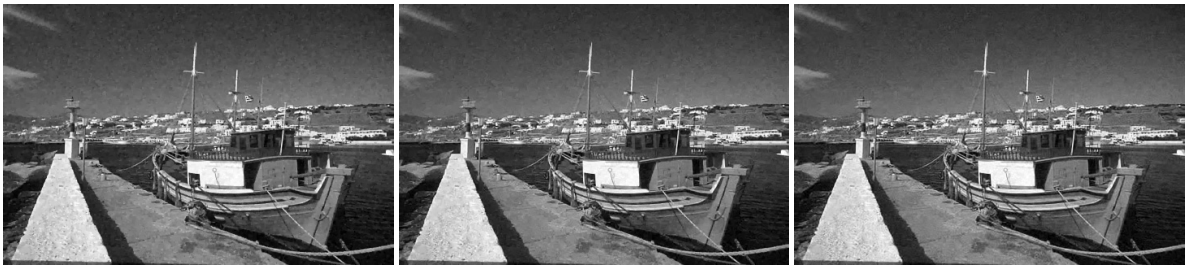
5.2.2. Comparison between the ℓ_1 model and the $\ell_{\frac{1}{2}}$ model. Figure 10 shows a comparison between the convex ℓ_1 model and the nonconvex $\ell_{\frac{1}{2}}$ model using the DCT5 filter bank for different noise levels. One can clearly see that the nonconvex $\ell_{\frac{1}{2}}$ model leads to significantly better denoising results and the difference is higher for smaller noise levels. We can characterize the qualitative differences between the ℓ_1 model and the $\ell_{\frac{1}{2}}$ model as follows:

(i) The $\ell_{\frac{1}{2}}$ model leads to a better preservation of the contrast in the reconstructed image than the ℓ_1 model. Let us interpret both models in terms of a shrinking process. It is known that the ℓ_1 model performs in principle a soft shrinkage of the coefficient which shrinks the coefficients independently of their strength. The $\ell_{\frac{1}{2}}$, however, performs a stronger shrinkage of smaller coefficients and a weaker shrinkage of larger coefficients which result in a better preservation of the contrast.

(ii) The ℓ_1 model is not very successful in recovering homogeneous areas, although it preserves textured regions very well. This effect comes from the convexity of the ℓ_1-norm, which cannot distinguish very well between homogeneous regions and textured regions. In contrast, the concave shape of the $\ell_{\frac{1}{2}}$-norm is much more successful in distinguishing textured and nontextured areas and hence gives better results.



(a) clean (b) noisy (24.61) (c) 1^{st}(29.12)



(d) $1^{\text{st}} + 2^{\text{nd}}$(29.31) (e) DCT3(29.50) (f) DCT5(29.49)

(g) clean (h) noisy (i) 1^{st} (j) $1^{\text{st}} + 2^{\text{nd}}$ (k) DCT3 (l) DCT5

Figure 7. *Image denoising performance of the trained ℓ_1 model for a natural image and $\sigma = 15$. The numbers shown in the brackets refer to PSNR values with respect to the clean image.*

(a) clean (b) noisy (20.17) (c) 1^{st}(28.17)

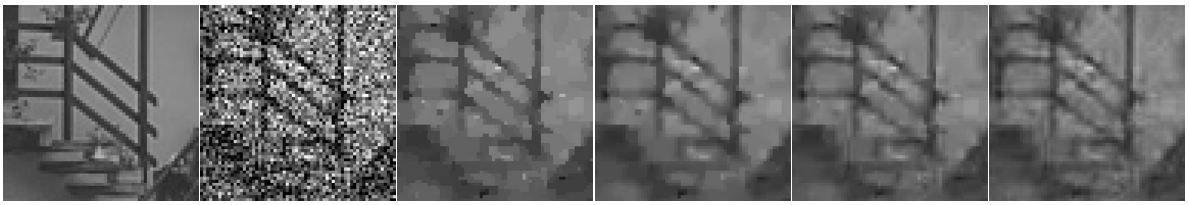
(d) $1^{\text{st}} + 2^{\text{nd}}$(28.73) (e) DCT3(28.81) (f) DCT5(28.93)

(g) clean (h) noisy (i) 1^{st} (j) $1^{\text{st}} + 2^{\text{nd}}$ (k) DCT3 (l) DCT5

Figure 8. *Image denoising performance of the trained ℓ_1 model for a natural image and $\sigma = 25$. The numbers shown in the brackets refer to PSNR values with respect to the clean image.*

(a) clean (b) noisy(14.15) (c) 1^{st}(22.97)

(d) $1^{\text{st}} + 2^{\text{nd}}$(23.15) (e) DCT3(23.28) (f) DCT5(23.31)



(g) clean (h) noisy (i) 1^{st} (j) $1^{\text{st}} + 2^{\text{nd}}$ (k) DCT3 (l) DCT5

Figure 9. *Image denoising performance of the trained ℓ_1 model for a natural image and $\sigma = 50$. The numbers shown in the brackets refer to PSNR values with respect to the clean image.*

(iii) As already pointed out above, further increasing the diversity of the ℓ_1 model does not improve the denoising performance. In contrast, the performance of the $\ell_{\frac{1}{2}}$ model can be further improved by increasing the diversity of the filter bank (see also Table 2).

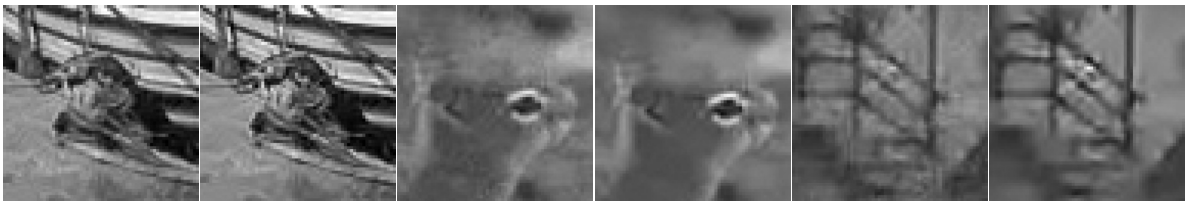
In [27], Samuel and Tappen proposed a bilevel learning algorithm for learning the optimal filters (comparable to a dictionary) of the so-called Fields of Experts (FoE) model of Roth and Black [26]. The FoE model uses a sum of priors involving nonconvex potential functions related to the negative log density of a Student-t distribution. The optimization algorithm is a plain gradient descent algorithm, where the gradients are computed using implicit differentiation. Since the FoE model has many more degrees of freedom than our simple models, one would expect that the FoE model would lead to better results. However, it turns out that our simple convex ℓ_1 model leads to comparable results and our nonconvex $\ell_{\frac{1}{2}}$ model leads to significantly better results (see Figure 11 for an example). We do not exactly know the reason for the improved performance of our simpler model, but possibly our Newton algorithms are distinctly more accurate in approximating (locally) optimal solutions than the gradient descent methods that are used in [27]. This fact justifies the use of Newton algorithms for this kind of learning problem.

5.2.3. Comparison to state-of-the-art methods. In our last experiment, we compare the results of our ℓ_1 and $\ell_{\frac{1}{2}}$ models to state-of-the art algorithms. Figure 12 shows a comparison of the proposed models to the FoE model of Roth and Black [26], the KSVD dictionary learning algorithm of Elad and Aharon [11], the recently proposed Gaussian mixture model (GMM) of Zoran and Weiss [33], and the BM3D algorithm of Dabov et al. [7], which define the current state-of-the-art in image denoising. One can see that while the convex ℓ_1 model cannot compete with the current state-of-the-art, the nonconvex $\ell_{\frac{1}{2}}$ model is clearly state-of-the-art. Note that the two methods GMM and BM3D, which are superior to our $\ell_{\frac{1}{2}}$ model, are much more involved. For example, the GMM method uses a generic image prior consisting of a GMM with 200 components, each of them specified by a 64×64 covariance matrix. Decomposing these covariance matrices into its eigenvectors, we end up with a total of 12800 filters, whereas our $\ell_{\frac{1}{2}}$ model uses only 24 DCT5 filters. The BM3D method is still the best method in this example, although it can also lead to strange artifacts, as can be seen from the overemphasis of the stripe-like texture in the detail view in Figure 12.

6. Conclusion and outlook. In this paper we have proposed semismooth Newton methods for learning the optimal regularization parameters in variational image denoising models, including the smooth ℓ_2-norm as well as the nonsmooth ℓ_1-norm. The parameters are learned in a way such that the minimizers of the variational models give the best approximation to given ground truth solutions. This naturally leads to a bilevel optimization approach with the higher-level problem being a loss function that minimizes the error between the solution of the lower-level optimization problem (the variational model) and the given ground truth data. We have analyzed the structure of the arising bilevel optimization problems, and in the case of an ℓ_2 model with a single prior we were able to show that the problem is quasi-convex in the regularization parameter.

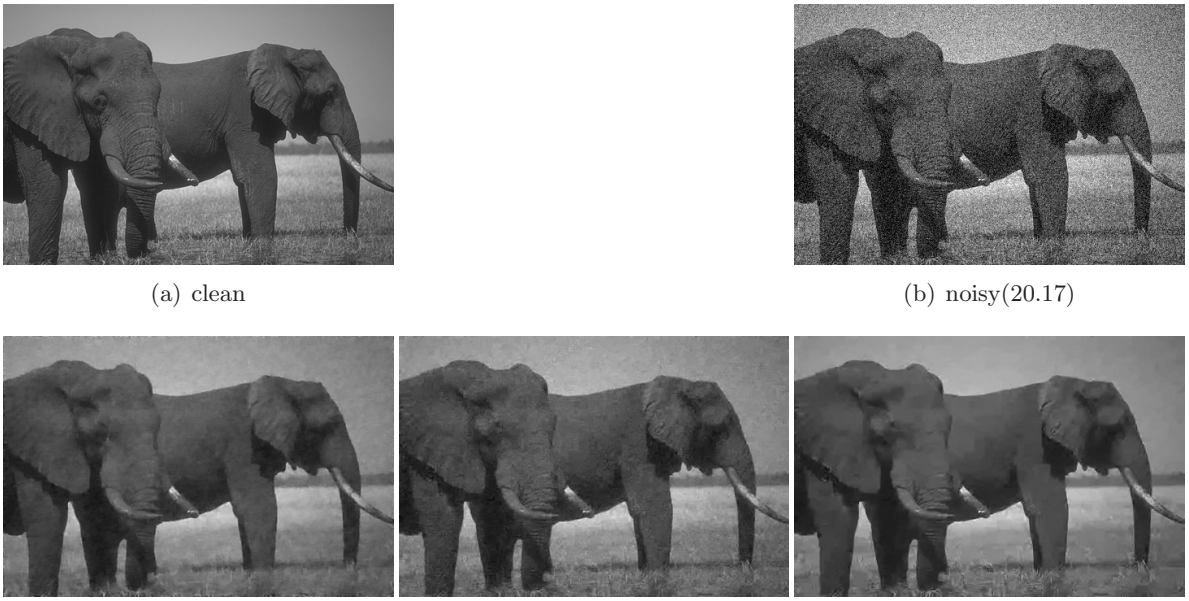
(a) $\ell_1(29.49)$ (b) $\ell_1(28.93)$ (c) $\ell_1(23.31)$

(d) $\ell_{\frac{1}{2}}(30.24)$ (e) $\ell_{\frac{1}{2}}(29.19)$ (f) $\ell_{\frac{1}{2}}(23.42)$



(g) ℓ_1 (h) $\ell_{\frac{1}{2}}$ (i) ℓ_1 (j) $\ell_{\frac{1}{2}}$ (k) ℓ_1 (l) $\ell_{\frac{1}{2}}$

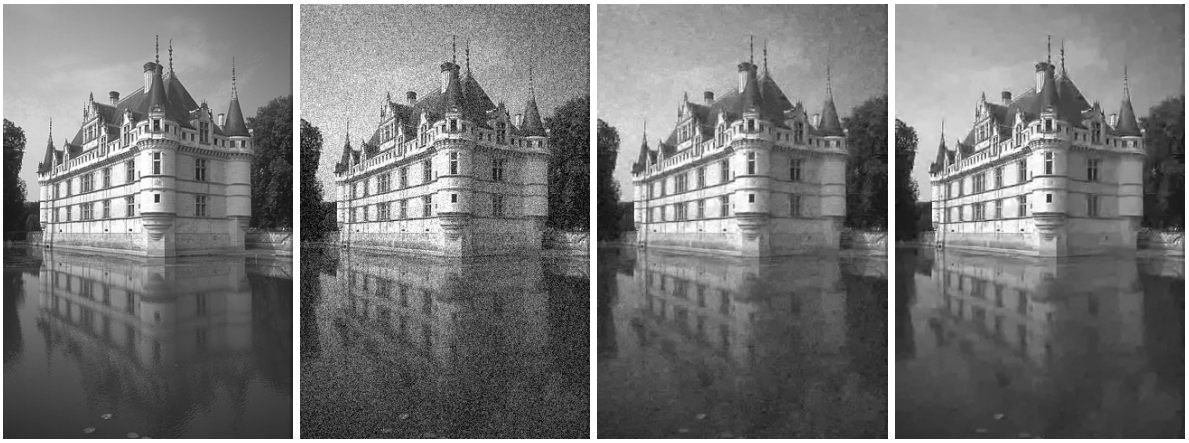
Figure 10. *Comparison between the convex ℓ_1 model and the nonconvex $\ell_{\frac{1}{2}}$ model for different noise levels and using DCT5 filters. The numbers shown in the brackets refer to PSNR values with respect to the clean image.*



(a) clean (b) noisy(20.17)

(c) FoE-bi (29.33) (d) ℓ_1(29.33) (e) $\ell_{\frac{1}{2}}$(29.70)

Figure 11. *Comparison between our ℓ_1 and $\ell_{\frac{1}{2}}$ models using DCT5 filters and the bilevel-optimized Fields of Experts (FoE-bi) model* [27]. *The numbers shown in the brackets refer to PSNR values with respect to the clean image.*



(a) clean (b) noisy(20.17) (c) ℓ_1(28.46) (d) $\ell_{\frac{1}{2}}$(29.26)

(e) FoE(28.72) (f) KSVD(29.19) (g) GMM(29.48) (h) BM3D(29.53)

(i) clean (j) noisy (k) ℓ_1 (l) $\ell_{\frac{1}{2}}$ (m) FoE (n) KSVD (o) GMM (p) BM3D

Figure 12. *Comparison between the proposed ℓ_1 and $\ell_{\frac{1}{2}}$ models to the FoE model of Roth and Black [26], the KSVD dictionary learning algorithm of Elad and Aharon [11], the recently proposed GMM of Zoran and Weiss [33], and the BM3D algorithm of Dabov et al. [7]. The numbers shown in the brackets refer to PSNR values with respect to the clean image.*

We have proposed and analyzed semismooth Newton methods that lead to efficient learning algorithms with guaranteed locally superlinear convergence. We tested the algorithms on natural image denoising problems using different noise levels and different sets of regularization priors. We have demonstrated that our proposed Newton algorithms can efficiently find optimal regularization parameters requiring approximately 20 Newton iterations on average.

Furthermore, we have presented preliminary results on applying the bilevel learning framework to variational models, including the nonsmooth and nonconvex $\ell_{\frac{1}{2}}$-norms. In particular, we have shown that switching from the ℓ_1-norm to the $\ell_{\frac{1}{2}}$-norm consistently improved the denoising performance over the ℓ_1 models.

Future work should include the investigation of data fidelity terms different from quadratic ones and a further analysis of models incorporating the nonconvex $\ell_{\frac{1}{2}}$-norm.

Appendix A. Proof of Theorem 4.3. The proof is given in several steps.

(i) First we need to address convergence of the solutions ϑ^ε to (4.4) as $\varepsilon \to 0^+$. It is not difficult to see that convergent subsequences of ϑ^ε converge to a solution of (4.1), but since the solutions to (4.1) are not unique, this may not be the desired one. For this reason we adapt Barbu's trick and introduce (only for the purpose of deriving the optimality condition) the auxiliary problem

$$\text{(A.1)} \qquad \begin{cases} \min\limits_{\vartheta \geq 0} \quad \|x(\vartheta) - g\|_2^2 + \|\vartheta - \vartheta^*\|_2^2 \\ \text{subject to } x(\vartheta) = \arg\min\limits_x \sum\limits_{k=1}^q \vartheta_k \|K_k x\|_1 + \dfrac{1}{2}\|x - f\|_2^2 \end{cases}$$

and the auxiliary regularized problem

$$\text{(A.2)} \qquad \begin{cases} \min\limits_{\vartheta \geq 0} \quad \|x(\vartheta) - g\|_2^2 + \|\vartheta - \vartheta^*\|_2^2 \\ \text{subject to } x(\vartheta) = \arg\min\limits_x \sum\limits_{k=1}^q \vartheta_k \sum\limits_{i=1}^m n_\varepsilon((K_k x)_i) + \dfrac{1}{2}\|x - f\|_2^2. \end{cases}$$

Adding the term $\|\vartheta - \vartheta^*\|_2^2$ to the cost has no effect on the discussion preceding the statement of the theorem. Problem (A.1) has ϑ^* as a unique solution. The necessarily optimality condition for (A.2) consists of the first two equations in (4.13) and

$$\text{(A.3)} \quad \left(\left\langle N_\varepsilon'(K_k x_\varepsilon), K_k p_\varepsilon \right\rangle + 2(\vartheta_{\varepsilon,k} - \vartheta_k^*)\right)(\vartheta_k - \vartheta_{\varepsilon,k}) \geq 0 \quad \text{for all } \vartheta_k \geq 0,\ k = 1, \ldots, q.$$

Let $\{\vartheta_\varepsilon\}_{\varepsilon>0}$ denote a family of solutions to (A.2). Since ϑ^* is suboptimal for (A.2), we obtain that

$$\|\vartheta_\varepsilon - \vartheta^*\|_2 \leq \|x(\vartheta^*) - g\|_2 + \|\vartheta^*\|_2,$$

and therefore $\{\vartheta_\varepsilon\}_{\varepsilon>0}$ is bounded. By the first equation in (4.13) the family $x_\varepsilon = x(\vartheta_\varepsilon)$ is bounded as well. Hence there exist a subsequence, denoted by the same symbol, and $\overline{\vartheta} \in \mathbb{R}^q$ such that $\lim_{\varepsilon\to 0^+} \vartheta_\varepsilon = \overline{\vartheta}$ and $\lim_{\varepsilon\to 0^+} x_\varepsilon(\vartheta_\varepsilon) = x(\overline{\vartheta})$. Taking the limit $\varepsilon \to 0^+$ in

$$\|x_\varepsilon(\vartheta_\varepsilon) - g\|_2^2 + \|\vartheta_\varepsilon - \vartheta^*\|_2^2 \leq \|x_\varepsilon(\vartheta) - g\|_2^2 + \|\vartheta - \vartheta^*\|_2^2,$$

where $x_\varepsilon(\vartheta) = \arg\min_x \sum_{k=1}^{q} \vartheta_k \sum_{i=1}^{m} n_\varepsilon((K_k x)_i) + \frac{1}{2}\|x - f\|_2^2$,
we obtain

$$\|x(\overline{\vartheta}) - g\|_2^2 + \|\overline{\vartheta} - \vartheta^*\|_2^2 \leq \|x(\vartheta) - g\|_2 + \|\vartheta - \vartheta^*\|_2^2$$

for all $\vartheta \geq 0$, where $x(\vartheta) = \arg\min_x \sum_{k=1}^{q} \vartheta_k \|K_k x\|_1 + \frac{1}{2}\|x - f\|_2^2$. By construction this implies that $\vartheta = \vartheta^*$.

Let us henceforth set $(\lambda_\varepsilon)_k = N_\varepsilon'(K_k x_\varepsilon)$ and $\xi_\varepsilon = \sum_{k=1}^{q} \vartheta_{\varepsilon,k} K_k^T N_\varepsilon''(K_k x_\varepsilon) K_k p_\varepsilon$.

(ii) By (4.13) and the properties of n_ε the families $\{p_\varepsilon\}_{\varepsilon>0}$, $\{(\lambda_\varepsilon)_k\}_{\varepsilon>0}$, and $\{\xi_\varepsilon\}_{\varepsilon>0}$ are bounded. Note that the boundedness of $\{\xi_\varepsilon\}_{\varepsilon>0}$ follows from the adjoint equation since $\{p_\varepsilon\}_{\varepsilon>0}$ is bounded. Hence, possibly after taking another subsequence, there exist p, λ_k, $k = 1, \ldots, q$ and ξ such that

$$(p_\varepsilon, \lambda_{\varepsilon,k}, \xi_\varepsilon) \longrightarrow (p, \lambda, \xi) \quad \text{as } \varepsilon \to 0^+.$$

We can now pass to the limit in the first and second equations of (4.13) and in (A.3) to obtain the first, third, and fourth equations of (4.14). Moreover, $0 \leq \lim_{\varepsilon\to 0^+} \langle \xi_\varepsilon, p_\varepsilon \rangle = \langle \xi, p \rangle$, which gives the fifth assertion in (4.14). Passing to the limit in λ_ε we find the second assertion of (4.14). Taking the inner product of the adjoint equation with p_ε and passing to the limit we obtain the sixth equation in (4.14).

(iii) To verify the last two claims, we note at first that by the adjoint equation

$$\|p_\varepsilon\|_2^2 + \left| \sum_{k=1}^{q} \vartheta_{\varepsilon,k} \left\langle N_\varepsilon''(K_k x_\varepsilon) K_k p_\varepsilon, K_k p_\varepsilon \right\rangle \right| \leq \|x_\varepsilon - g\| \, \|p_\varepsilon\|_2.$$

Consequently $\{\sum_{k=1}^{q} \vartheta_{\varepsilon,k} \| \sqrt{N_\varepsilon''(K_k x_\varepsilon)} K_k p_\varepsilon\|_2^2\}_{\varepsilon>0}$ is bounded. We have

$$\begin{aligned}
|\langle \xi_\varepsilon, x_\varepsilon \rangle| &= \left| \sum_{k=1}^{q} \vartheta_{\varepsilon,k} \left\langle N_\varepsilon''(K_k x_\varepsilon) K_k p_\varepsilon, K_k x_\varepsilon \right\rangle \right| \\
&\leq \sum_{i=1}^{q} \vartheta_{\varepsilon,k} \| \sqrt{N_\varepsilon''(K_k x_\varepsilon)} K_k p_\varepsilon\|_2 \, \| \sqrt{N_\varepsilon''(K_k x_\varepsilon)} K_k x_\varepsilon\|_2 \\
&\leq \left(\sum_{k=1}^{q} \vartheta_{\varepsilon,k} \| \sqrt{N_\varepsilon''(K_k x_\varepsilon)} K_k p_\varepsilon\|_2^2 \right)^{\frac{1}{2}} \left(\sum_{k=1}^{q} \vartheta_{\varepsilon,k} \| \sqrt{N_\varepsilon''(K_k x_\varepsilon)} K_k x_\varepsilon\|_2^2 \right)^{\frac{1}{2}} \overset{\varepsilon\to 0^+}{\to} 0
\end{aligned}$$

by the properties of n_ε''. Therefore $|\langle \xi, x^* \rangle| = \lim_{\varepsilon\to 0^+} |\langle \xi_\varepsilon, x_\varepsilon \rangle| = 0$, which is the seventh claim in (4.14). To verify the last one we set $\mathcal{I}_{\varepsilon,k} = \{i : |(K_k x_\varepsilon)_i| < \varepsilon\}$ and

find that

$$
\begin{aligned}
0 &\leq \sum_{i=1}^{m} |(K_k p_\varepsilon)_i| \, (1 - |(\lambda_{\varepsilon,k})_i|) \\
&= \sum_{i \in \mathcal{I}_{\varepsilon,k}}^{m} |(K_k p_\varepsilon)_i| \left(1 - \left| \frac{3}{2\varepsilon} (K_k x)_i - \frac{1}{2\varepsilon^3} (K_k x)_i^3 \right| \right) \\
&\leq \sum_{i \in \mathcal{I}_{\varepsilon,k}}^{m} |\sqrt{n''(K_k x_\varepsilon)} (K_k p_\varepsilon)_i| \, \frac{1}{\sqrt{n''((K_k x_\varepsilon)_i)}} \left(1 - \left| \frac{3}{2\varepsilon} (K_k x_\varepsilon)_i - \frac{1}{2\varepsilon^3} (K_k x_\varepsilon)_i^3 \right| \right) \\
&\leq \| \sqrt{n''(K_k x_\varepsilon)} \, (K_k p_\varepsilon) \|_2 \left(\sum_{i \in \mathcal{I}_{\varepsilon,k}} \frac{1}{n''((K_k x_\varepsilon)_i)} \left(1 - \left| \frac{3}{2\varepsilon} (K_k x_\varepsilon)_i - \frac{1}{2\varepsilon^3} (K_k x_\varepsilon)_i^3 \right| \right)^2 \right)^{\frac{1}{2}}.
\end{aligned}
$$

Utilizing that $n''(t) = -\frac{3}{2\varepsilon^3} t^2 + \frac{3}{2\varepsilon}$, for $|t| < \varepsilon$ one argues that $\lim_{\varepsilon \to 0^+} \sup_{|t| \leq \varepsilon} \frac{1}{n''(t)} (1 - |\frac{3}{2\varepsilon} t - \frac{1}{2\varepsilon^3} t^3|)^2 = 0$, and hence $\sum_{i=1}^{\mu} (K_k p)_i (1 - |(\lambda_i)_i|) = 0$, as desired.

REFERENCES

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, *Contour detection and hierarchical image segmentation*, IEEE Trans. Pattern Anal. Mach. Intell., 33 (2011), pp. 898–916.

[2] J. Bard, *Practical Bilevel Optimization*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[3] A. Chambolle and P.-L. Lions, *Image recovery via total variation minimization and related problems*, Numer. Math., 76 (1997), pp. 167–188.

[4] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging Vision, 40 (2010), pp. 120–145.

[5] 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.

[6] R. W. Chaney, *Second-order necessary conditions in constrained semismooth optimization*, SIAM J. Control Optim., 25 (1987), pp. 1072–1081.

[7] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, *Image restoration by sparse 3d transform-domain collaborative filtering*, in Proceedings of SPIE Electronic Imaging, 2008.

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

[9] J. C. De los Reyes, *Optimal control of a class of variational inequalities of the second kind.*, SIAM J. Control Optim., 49 (2011), pp. 1629–1658.

[10] A. Doicu, T. Trautmann, and T. Schreier, *Numerical Regularization for Atmospheric Inverse Problems*, Springer-Verlag, Heidelberg, 2010.

[11] M. Elad and M. Aharon, *Image denoising via sparse and redundant representations over learned dictionaries*, IEEE Trans. Image Process., 15 (2006), pp. 3736–3745.

[12] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Math. Appl. 375, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[13] W. Hinterberger, M. Hintermüller, K. Kunisch, M. Von Oehsen, and O. Scherzer, *Tube methods for BV regularization*, J. Math. Imaging Vision, 19 (2003), pp. 219–235.

[14] 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.

[15] J. Huang and D. Mumford, *Statistics of natural images and models*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR1999), Fort Collins, CO, 1999, pp. 541–547.

[16] K. Ito and K. Kunisch, *Newton's method for a class of weakly singular optimal control problems*, SIAM J. Optim., 10 (2000), pp. 896–916.

[17] K. Ito and K. Kunisch, *Lagrange Multiplier Approach to Variational Problems and Applications*, SIAM, Philadelphia, 2008.
[18] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, Berlin, 1976.
[19] C. Liu, W. T. Freeman, R. Szeliski, and S. B. Kang, *Noise estimation from a single image*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR2006), 2006, pp. 901–908.
[20] D. C. Liu and J. Nocedal, *On the limited memory BFGS method for large scale optimization*, Math. Programming, 45 (1989), pp. 503–528.
[21] D. Luenberger and Y. Ye, *Linear and Nonlinear Programming*, 3rd ed., Springer, New York, 2008.
[22] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, 2000.
[23] J. V. Outrata, *A generalized mathematical program with equilibrium constraints*, SIAM J. Control Optim., 38 (2000), p. 1623–1638.
[24] G. Peyré and J. Fadili, *Learning analysis sparsity priors*, in Proceedings of the International Conference on Sampling Theory and Applications (Sampta'11), 2011.
[25] T. Pock and A. Chambolle, *Diagonal preconditioning for first order primal-dual algorithms*, in Proceedings of the IEEE International Conference on Computer Vision (ICCV 2011), 2011.
[26] S. Roth and M. J. Black, *Fields of experts*, Int. J. Comput. Vis., 82 (2009), pp. 205–229.
[27] K. G. G. Samuel and M. Tappen, *Learning optimized map estimates in continuously-valued MRF models*,in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR2009), 2009.
[28] M. F. Tappen, *Utilizing variational optimization to learn Markov random fields*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR2007), 2007.
[29] M. F. Tappen, C. Liu, E. H. Adelson, and W. T. Freeman, *Learning Gaussian conditional random fields for low-level vision*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR2007), 2007.
[30] M. Ulbrich, *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, MOS-SIAM Ser. Optim. 11, SIAM, Philadelphia, 2011.
[31] S. Vaiter, G. Peyré, C. Dossal, and J. Fadili, *Robust Sparse Analysis Regularization*, preprint, CoRR, abs/1109.6222, 2011.
[32] C. R. Vogel and M. E. Oman, *Iterative methods for total variation denoising*, SIAM J. Sci. Comput., 17 (1996), pp. 227–238.
[33] D. Zoran and Y. Weiss, *From learning models of natural image patches to whole image restoration*, in Proceedings of the IEEE International Conference on Computer Vision (ICCV 2011), 2011, pp. 479–486.