

BILEVEL PARAMETER LEARNING FOR HIGHER-ORDER TOTAL VARIATION REGULARISATION MODELS*

J.C. DE LOS REYES¹, C.-B. SCHÖNLIEB² AND T. VALKONEN²

ABSTRACT. We consider a bilevel optimisation approach for parameter learning in higher-order total variation image reconstruction models. Apart from the least squares cost functional, naturally used in bilevel learning, we propose and analyse an alternative cost, based on a Huber regularised TV-seminorm. Differentiability properties of the solution operator are verified and a first-order optimality system is derived. Based on the adjoint information, a quasi-Newton algorithm is proposed for the numerical solution of the bilevel problems. Numerical experiments are carried out to show the suitability of our approach and the improved performance of the new cost functional. Thanks to the bilevel optimisation framework, also a detailed comparison between TGV² and ICTV is carried out, showing the advantages and shortcomings of both regularisers, depending on the structure of the processed images and their noise level.

1. INTRODUCTION

In this paper we propose a bilevel optimisation approach for parameter learning in higher-order total variation regularisation models for image restoration. The reconstruction of an image from imperfect measurements is essential for all research which relies on the analysis and interpretation of image content. Mathematical image reconstruction approaches aim to maximise the information gain from acquired image data by intelligent modelling and mathematical analysis.

A variational image reconstruction model can be formalised as follows. Given data f which is related to an image (or to certain image information, e.g. a segmented or edge detected image) u through a generic forward operator (or function) K the task is to retrieve u from f . In most realistic situations this retrieval is complicated by the ill-posedness of K as well as random noise in f . A widely accepted method that approximates this ill-posed problem by a well-posed one and counteracts the noise is the method of Tikhonov regularisation. That is, an approximation to the true image is computed as a minimiser of

$$(1.1) \quad \alpha R(u) + d(K(u), f),$$

where R is a regularising energy that models a-priori knowledge about the image u , $d(\cdot, \cdot)$ is a suitable distance function that models the relation of the data f to the unknown u , and $\alpha > 0$ is a parameter that balances our trust in the forward model

¹RESEARCH CENTER ON MATHEMATICAL MODELLING (MODEMAT), ESCUELA POLITÉCNICA NACIONAL, QUITO, ECUADOR.

²DEPARTMENT OF APPLIED MATHEMATICS AND THEORETICAL PHYSICS, UNIVERSITY OF CAMBRIDGE, CAMBRIDGE, UNITED KINGDOM.

*This research has been supported by King Abdullah University of Science and Technology (KAUST) Award No. KUK-I1-007-43, EPSRC grants Nr. EP/J009539/1 “Sparse & Higher-order Image Restoration” and Nr. EP/M00483X/1 “Efficient computational tools for inverse imaging problems”, Escuela Politécnica Nacional de Quito Award No. PIS 12-14 and MATHAmSud project SOCDE “Sparse Optimal Control of Differential Equations”. While in Quito, T. Valkonen has moreover been supported by SENESCYT (Ecuadorian Ministry of Higher Education, Science, Technology and Innovation) under a Prometeo Fellowship.

against the need of regularisation. The parameter α in particular, depends on the amount of ill-posedness in the operator K and the amount (amplitude) of the noise present in f . A key issue in imaging inverse problems is the correct choice of α , image priors (regularisation functionals R), fidelity terms d and (if applicable) the choice of what to measure (the linear or nonlinear operator K). Depending on this choice, different reconstruction results are obtained.

While functional modelling (1.1) constitutes a mathematically rigorous and physical way of setting up the reconstruction of an image – providing reconstruction guarantees in terms of error and stability estimates – it is limited with respect to its adaptivity for real data. On the other hand, data-based modelling of reconstruction approaches is set up to produce results which are optimal with respect to the given data. However, in general it neither offers insights into the structural properties of the model nor provides comprehensible reconstruction guarantees. Indeed, we believe that for the development of reliable, comprehensible and at the same time effective models (1.1) it is essential to aim for a unified approach that seeks tailor-made regularisation and data models by combining model- and data-based approaches.

To do so we focus on a bilevel optimisation strategy for finding an optimal setup of variational regularisation models (1.1). That is, for a given training pair of noisy and original clean images (f, f_0) , respectively, we consider a learning problem of the form

$$(1.2) \quad \min F(u^*) = \text{cost}(u^*, f_0) \quad \text{subject to} \quad u^* \in \arg \min_u \{ \alpha R(u) + d(K(u), f) \},$$

where F is a generic cost functional that measures the fitness of u^* to the original image f_0 . The argument of the minimisation problem will depend on the specific setup (i.e. the degrees of freedom) in the constraint problem (1.1). In particular, we propose a bilevel optimisation approach for learning optimal parameters in higher-order total variation regularisation models for image reconstruction in which the arguments of the optimisation constitute parameters in front of the first- and higher-order regularisation terms. Rather than working on the discrete problem, as is done in standard parameter learning and model optimisation methods, we optimise the regularisation models in infinite dimensional function space. We will explain this approach in more detail in the next section. Before, let us give an account to the state of the art of bilevel optimisation for model learning. In machine learning bilevel optimisation is well established. It is a semi-supervised learning method that optimally adapts itself to a given dataset of measurements and desirable solutions. In [34, 18, 14], for instance the authors consider bilevel optimization for finite dimensional Markov random field models. In inverse problems the optimal inversion and experimental acquisition setup is discussed in the context of optimal model design in works by Haber, Horesh and Tenorio [20, 21], as well as Ghattas et al. [8, 3]. Recently parameter learning in the context of functional variational regularisation models (1.1) also entered the image processing community with works by the authors [16, 9], Kunisch, Pock and co-workers [26, 13], Chung et al. [15] and Hintermüller et al. [24].

Apart from the work of the authors [16, 9], all approaches so far are formulated and optimised in the discrete setting. Our subsequent modelling, analysis and optimisation will be carried out in function space rather than on a discretisation of (1.1). While digitally acquired image data is of course discrete, the aim of high resolution image reconstruction and processing is always to compute an image that is close to the real (analogue, infinite dimensional) world. Hence, it makes sense to seek images which have certain properties in an infinite dimensional function space. That is, we aim for a processing method that accentuates and preserves qualitative properties in images independent of the resolution of the image itself [36]. Moreover, optimisation

methods conceived in function space potentially result in numerical iterative schemes which are resolution and mesh-independent upon discretisation [23].

Higher-order total variation regularisation has been introduced as an extension of the standard total variation regulariser in image processing. As the **Total Variation (TV)** [32] and many more contributions in the image processing community have proven, a non-smooth first-order regularisation procedure results in a nonlinear smoothing of the image, smoothing more in homogeneous areas of the image domain and preserving characteristic structures such as edges. In particular, the TV regulariser is tuned towards the preservation of edges and performs very well if the reconstructed image is piecewise constant. The drawback of such a regularisation procedure becomes apparent as soon as images or signals (in 1D) are considered which do not only consist of constant regions and jumps, but also possess more complicated, higher-order structures, e.g. piecewise linear parts. The artefact introduced by TV regularisation in this case is called staircasing [31]. One possibility to counteract such artefacts is the introduction of higher-order derivatives in the image regularisation. Chambolle and Lions [10], for instance, propose a higher order method by means of an infimal convolution of the TV and the TV of the image gradient called **Infimal-Convolution Total Variation (ICTV)** model. Other approaches to combine first and second order regularisation originate, for instance, from Chan, Marquina, and Mulet [11] who consider total variation minimisation together with weighted versions of the Laplacian, the Euler-elastica functional [29, 12] which combines total variation regularization with curvature penalisation, and many more [27, 30] just to name a few. Recently Bredies et al. have proposed **Total Generalized Variation (TGV)** [4] as a higher-order variant of TV regularisation.

In this work we mainly concentrate on two second-order total variation models: the recently proposed TGV [4] and the ICTV model of Chambolle and Lions [10]. We focus on second-order TV regularisation only since this is the one which seems to be most relevant in imaging applications [25, 5]. For $\Omega \subset \mathbb{R}^2$ open and bounded and $u \in BV(\Omega)$, the ICTV regulariser reads

$$(1.3) \quad \text{ICTV}_{\alpha,\beta}(u) := \min_{v \in W^{1,1}(\Omega), \nabla v \in BV(\Omega)} \alpha \|Du - \nabla v\|_{\mathcal{M}(\Omega;\mathbb{R}^2)} + \beta \|D\nabla v\|_{\mathcal{M}(\Omega;\mathbb{R}^{2 \times 2})}.$$

On the other hand, second-order TGV [7, 6] for $u \in BV(\Omega)$ reads

$$(1.4) \quad \text{TGV}_{\alpha,\beta}^2(u) := \min_{w \in BD(\Omega)} \alpha \|Du - w\|_{\mathcal{M}(\Omega;\mathbb{R}^2)} + \beta \|Ew\|_{\mathcal{M}(\Omega;\text{Sym}^2(\mathbb{R}^2))}.$$

Here $BD(\Omega) := \{w \in L^1(\Omega;\mathbb{R}^n) \mid \|Ew\|_{\mathcal{M}(\Omega;\mathbb{R}^{n \times n})} < \infty\}$ is the space of vector fields of bounded deformation on Ω , E denotes the *symmetrised gradient* and $\text{Sym}^2(\mathbb{R}^2)$ the space of symmetric tensors of order 2 with arguments in \mathbb{R}^2 . The parameters α, β are fixed positive parameters and will constitute the arguments in the special learning problem á la (1.2) we consider in this paper. The main difference between (1.3) and (1.4) is that we do not generally have that $w = \nabla v$ for any function v . That results in some qualitative differences of ICTV and TGV regularisation, compare for instance [1]. Substituting $\alpha R(u)$ in (1.1) by $\text{TGV}_{\alpha,\beta}^2(u)$ or $\text{ICTV}_{\alpha,\beta}(u)$ gives the TGV image reconstruction model and the ICTV image reconstruction model, respectively. In this paper we only consider the case $K = Id$ identity and $d(u, f) = \|u - f\|_{L^2(\Omega)}^2$ in (1.1) which corresponds to an image de-noising model for removing Gaussian noise. With our choice of regulariser the former scalar α in (1.1) has been replaced by a vector (α, β) of two parameters in (1.4) and (1.3). The choice of the entries in this vector now do not only determine the overall strength of the regularisation (depending on the properties of K and the noise level) but those parameters also balance between the different orders of regularity of the function u , and their choice is indeed crucial for the image reconstruction result. Large β will give regularised solutions that are

close to TV regularised reconstructions, compare Figure 1. Large α will result in TV^2 type solutions, that is solutions that are regularised with TV of the gradient [22, 30], compare Figure 2. With our approach described in the next section we propose a learning approach for choosing those parameters optimally, in particular optimally for particular types of images.

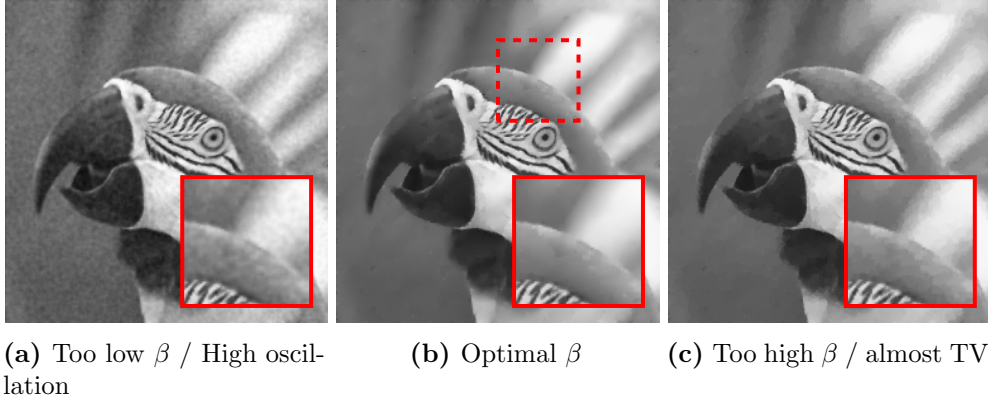


Figure 1. Effect of β on TGV^2 denoising with optimal α

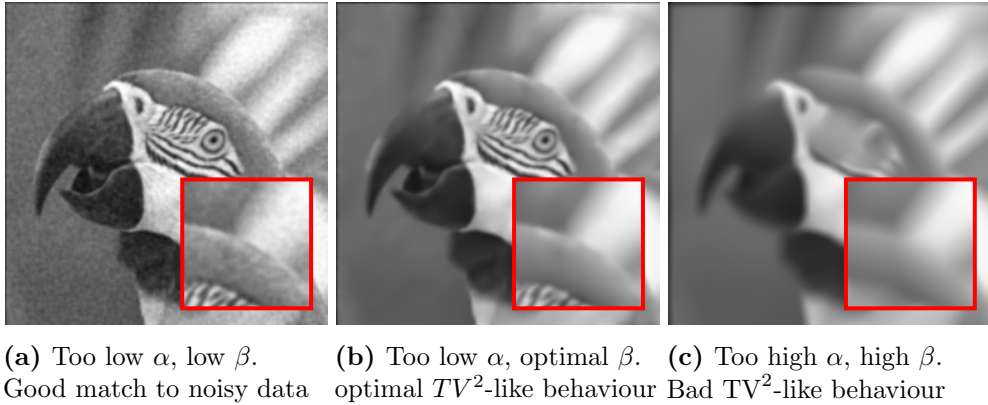


Figure 2. Effect of choosing α too large in TGV^2 denoising

For the existence analysis of an optimal solution as well as for the derivation of an optimality system for the corresponding learning problem (1.2) we will consider a smoothed version of the constraint problem (1.1) – which is the one in fact used in the numerics. That is, we replace $R(u)$ – being TV, TGV or ICTV in this paper – by a Huber regularised version and add an H^1 regularisation with a small weight to (1.1). In this setting and under the special assumption of box constraints on α and β we provide a simple existence proof for an optimal solution. A more general existence result that holds also for the original non-smooth problem and does not require box constraints is derived in [17] and we refer the reader to this paper for a more sophisticated analysis on the structure of solutions.

A main challenge in the setup of such a learning approach is to decide what is the best way to measure fitness (optimality) of the model. In our setting this amounts to choosing an appropriate distance F in (1.2) that measures the fitness of reconstructed images to the ‘perfect’, noise-free images in an appropriate training set. We have to formalise what we mean by an optimal reconstruction model. Classically, the

difference between the original, noise-free image f_0 and its regularised version $u_{\alpha,\beta}$ is computed with an L_2^2 cost functional

$$(1.5) \quad F_{L_2^2}(u_{\alpha,\beta}) = \|u_{\alpha,\beta} - f_0\|_{L^2(\Omega)}^2,$$

which is closely related to the PSNR quality measure. Apart from this, we propose in this paper an alternative cost functional based on a Huberised total variation cost

$$(1.6) \quad F_{L_{\eta}^1 \nabla}(u_{\alpha,\beta}) := \int_{\Omega} |D(u_{\alpha,\beta} - f_0)|_{\gamma} dx,$$

where the Huber regularisation $|\cdot|_{\gamma}$ will be defined later on in Definition 2.1. We will see that the choice of this cost functional is indeed crucial for the qualitative properties of the reconstructed image.

The proposed bilevel approach has an important indirect consequence: It establishes a basis for the comparison of the different total variation regularisers employed in image denoising tasks. In the last part of the paper we exhaustively compare the performance of TV, TGV² and ICTV for various image datasets. The parameters are chosen optimally, according to the proposed bilevel approach, and different quality measures (like PSNR and SSIM) are considered for the comparison. The obtained results are enlightening about when to use each one of the considered regularisers. In particular, ICTV appears to behave better for images with arbitrary structure and moderate noise levels, whereas TGV² behaves better for images with large smooth areas.

Outline of the paper In Section 2 we state the bilevel learning problem for the two higher-order total variation regularisation models, TGV and ICTV, and prove existence of an optimal parameter pair α, β . The bilevel optimization problem is analysed in Section 3, where existence of Lagrange multipliers is proved and an optimality system, as well as a gradient formula, are derived. Based on the optimality condition, a BFGS algorithm for the bilevel learning problem is devised in Section 5.1. For the numerical solution of each denoising problem an infeasible semi-smooth Newton method is considered. Finally, we discuss the performance of the parameter learning method by means of several examples for the denoising of natural photographs in Section 5. Therein, we also present a statistical analysis on how TV, ICTV and TGV regularisation compare in terms of returned image quality, carried out on 200 images from the Berkeley segmentation dataset BSDS300.

2. PROBLEM STATEMENT AND EXISTENCE ANALYSIS

We strive to develop a parameter learning method for higher-order total variation regularisation models that maximises the fit of the reconstructed images to training images simulated for an application at hand. For a given noisy image $f \in L^2(\Omega)$, $\Omega \subset \mathbb{R}^2$ open and bounded, we consider

$$(2.1) \quad \min_u \left\{ R_{\alpha,\beta}(u) + \frac{1}{2} \|u - f\|_{L^2(\Omega)}^2 \right\}.$$

where, $\alpha, \beta \in \mathbb{R}$. We focus on TGV² and ICTV image denoising:

$$R_{\alpha,\beta}(u) = \text{TGV}_{\alpha,\beta}^2(u) := \min_{w \in BD(\Omega)} \|\alpha (Du - w)\|_{\mathcal{M}(\Omega; \mathbb{R}^2)} + \|\beta Ew\|_{\mathcal{M}(\Omega; \text{Sym}^2(\mathbb{R}^2))}.$$

and (1.3) with spatial dependence

$$R_{\alpha,\beta}(u) = \text{ICTV}_{\alpha,\beta}(u) := \min_{\substack{v \in W^{1,1}(\Omega) \\ \nabla v \in BV(\Omega)}} \|\alpha (Du - \nabla v)\|_{\mathcal{M}(\Omega; \mathbb{R}^2)} \\ + \|\beta D\nabla v\|_{\mathcal{M}(\Omega; \mathbb{R}^{2 \times 2})},$$

for $u \in BV(\Omega)$. For this model, we want to determine the optimal choice of α, β , given a particular type of images and a fixed noise level. More precisely, we consider a training pair (f, f_0) , where f is a noisy image corrupted by normally distributed noise with a fixed variation, and the image f_0 represents the ground truth or an image that approximates the ground truth within a desirable tolerance. Then, we determine the optimal choice of α, β by solving the following problem

$$(2.2) \quad \min_{(\alpha,\beta) \in \mathbb{R}^2} F(u_{\alpha,\beta}) \quad \text{s.t. } \alpha, \beta \geq 0,$$

where F equals the L^2_2 cost (1.5) or the Huberised TV cost (1.6) and $u_{\alpha,\beta}$ for a given f solves a regularised version of the minimization problem (2.1) that will be specified in the next section, compare problem (2.3b). This regularisation of the problem is a technical requirement for solving the bilevel problem that will be discussed in the sequel. In contrast to learning α, β in (2.1) in finite dimensional parameter spaces (as is the case in machine learning) we aim for novel optimisation techniques in infinite dimensional function spaces.

2.1. Formal statement. Let $\Omega \subset \mathbb{R}^n$ be an open bounded domain with Lipschitz boundary. This will be our image domain. Usually $\Omega = (0, w) \times (0, h)$ for w and h the width and height of a two-dimensional image, although no such assumptions are made in this work. Our data f and f_0 are assumed to lie in $L^2(\Omega)$.

In our learning problem, we look for parameters (α, β) that for some cost functional $F : H^1(\Omega) \rightarrow \mathbb{R}$ solve the problem

$$(2.3a) \quad \min_{(\alpha,\beta) \in \mathbb{R}^2} F(u_{\alpha,\beta})$$

subject to

$$(2.3b) \quad u_{\alpha,\beta} \in \arg \min_{u \in H^1(\Omega)} J^{\gamma,\mu}(u; \alpha, \beta)$$

$$(2.3c) \quad \alpha, \beta \geq 0,$$

where

$$J^{\gamma,\mu}(u; \alpha, \beta) := \frac{1}{2} \|u - f\|_{L^2(\Omega)}^2 + R_{\alpha,\beta}^{\gamma,\mu}(u).$$

Here $J^{\gamma,\mu}(\cdot; \alpha, \beta)$ is the regularised denoising functional that amends the regularisation term in (2.1) by a Huber regularised version of it with parameter $\gamma > 0$, and an elliptic regularisation term with parameter $\mu > 0$. In the case of TGV^2 the modified regularisation term $R_{\alpha,\beta}^{\gamma,\mu}(u)$ then reads for $u \in H^1(\Omega)$

$$\text{TGV}_{\alpha,\beta}^{2,\gamma,\mu}(u) := \min_{w \in H^1(\Omega)} \int_{\Omega} \alpha |Du - w|_{\gamma} dx \\ + \int_{\Omega} \beta |Ew|_{\gamma} dx + \frac{\mu}{2} \left(\|u\|_{H^1(\Omega)}^2 + \|w\|_{\mathbb{H}^1(\Omega)}^2 \right)$$

and in the case of ICTV we have

$$\begin{aligned} \text{ICTV}_{\alpha,\beta}^{\gamma,\mu}(u) := & \min_{\substack{v \in W^{1,1}(\Omega) \\ \nabla v \in \text{BV}(\Omega, \mathbb{R}^n) \cap \mathbb{H}^1(\Omega)}} \int_{\Omega} \alpha |Du - \nabla v|_{\gamma} dx \\ & + \int_{\Omega} \beta |D\nabla v|_{\gamma} dx + \frac{\mu}{2} \left(\|u\|_{H^1(\Omega)}^2 + \|\nabla v\|_{\mathbb{H}^1(\Omega)}^2 \right). \end{aligned}$$

Here, $\mathbb{H}^1(\Omega) = H^1(\Omega; \mathbb{R}^n)$ and the Huber regularisation $|\cdot|_{\gamma}$ is defined as follows.

Definition 2.1. Given $\gamma \in (0, \infty]$, we define for the norm $\|\cdot\|_2$ on \mathbb{R}^m , the Huber regularisation

$$|g|_{\gamma} = \begin{cases} \|g\|_2 - \frac{1}{2\gamma}, & \|g\|_2 \geq 1/\gamma, \\ \frac{\gamma}{2} \|g\|_2^2, & \|g\|_2 < 1/\gamma. \end{cases}$$

For the cost functional F , given noise-free data $f_0 \in L^2(\Omega)$ and a regularised solution $u \in H^1(\Omega)$, we consider in particular the L^2 cost

$$F_{L^2}(u) := \frac{1}{2} \|f_0 - u\|_{L^2(\Omega; \mathbb{R}^d)}^2,$$

as well as the Huberised total variation cost

$$F_{L^1_{\nabla}}(u) := \int_{\Omega} |D(f_0 - u)|_{\gamma} dx$$

with noise-free data $f_0 \in \text{BV}(\Omega)$.

2.2. Existence of an optimal solution. The existence of an optimal solution for the learning problem (2.3) is a special case of the class of bilevel problems considered in [17], where existence of optimal parameters in $(0, +\infty]^{2N}$ is proven. For convenience, we provide a simplified proof for the case where box constraints on the parameters are imposed. We start with an auxiliary lower semicontinuity result for the Huber regularised functionals.

Lemma 2.1. *Let $u, v \in L^p(\Omega)$, $1 \leq p < \infty$. Then, the functional $u \mapsto \int_{\Omega} |u - v|_{\gamma} dx$, where $|\cdot|_{\gamma}$ is the Huber regularisation in Definition 2.1, is lower semicontinuous with respect to weak* convergence in $\mathcal{M}(\Omega; \mathbb{R}^d)$*

Proof. Recall that for $g \in \mathbb{R}^m$, the Huber-regularised norm may be written in dual form as

$$|g|_{\gamma} = \sup \left\{ \langle q, g \rangle - \frac{\gamma}{2} \|q\|_2^2 : \|q\|_2 \leq 1 \right\}.$$

Therefore, we find that

$$\begin{aligned} G(u) := \int_{\Omega} |u - v|_{\gamma} dx = & \sup \left\{ \int_{\Omega} u(x) \cdot \varphi(x) dx - \int_{\Omega} \frac{\gamma}{2} \|\varphi(x)\|_2^2 dx : \right. \\ & \left. \varphi \in C_c^{\infty}(\Omega), \|\varphi(x)\|_2 \leq \alpha \text{ for every } x \in \Omega \right\}. \end{aligned}$$

The functional G is of the form $G(u) = \sup \{ \langle u, \varphi \rangle - G^*(\varphi) \}$, where G^* is the convex conjugate of G . Now, let $\{u^i\}_{i=1}^{\infty}$ converge to u weakly* in $\mathcal{M}(\Omega; \mathbb{R}^d)$. Taking a supremising sequence $\{\varphi^j\}_{j=1}^{\infty}$ for this functional at any point u , we easily see lower semicontinuity by considering the sequences $\{ \langle u^i, \varphi^j \rangle - G^*(\varphi^j) \}_{i=1}^{\infty}$ for each j . \square

Our main existence result is the following.

Theorem 2.1. *We consider the learning problem (2.3) for TGV² and ICTV regularisation, optimising over parameters (α, β) such that $0 \leq \alpha \leq \bar{\alpha}, 0 \leq \beta \leq \bar{\beta}$. Here $(\bar{\alpha}, \bar{\beta}) < \infty$ is an arbitrary but fixed vector in \mathbb{R}^2 that defines a box constraint on the parameter space. Then, there exists an optimal solution $(\hat{\alpha}, \hat{\beta}) \in \mathbb{R}^2$ for this problem for both choices of cost functionals, $F = L_2^2$ and $F = F_{L_1^\nabla}$.*

Proof. Let $(\alpha_n, \beta_n) \subset \mathbb{R}^2$ be a minimising sequence. Due to the box constraints we have that the sequence (α_n, β_n) is bounded in \mathbb{R}^2 . Moreover, we get for the corresponding sequences of states $u_n := u_{(\alpha_n, \beta_n)}$ that

$$J^{\gamma, \mu}(u_n; \alpha_n, \beta_n) \leq J^{\gamma, \mu}(u; \alpha_n, \beta_n), \quad \forall u \in H^1(\Omega),$$

in particular this holds for $u = 0$. Hence,

$$(2.4) \quad \frac{1}{2} \|u_n - f\|_{L^2(\Omega)}^2 + R_{\alpha_n, \beta_n}^{\gamma, \mu}(u_n) \leq \frac{1}{2} \|f\|_{L^2(\Omega)}^2.$$

Exemplarily, we consider here the case for the TGV regulariser, that is $R_{\alpha_n, \beta_n}^{\gamma, \mu} = \text{TGV}_{\alpha, \beta}^{2, \gamma, \mu}$. The proof for the ICTV regulariser can be done in a similar fashion. Inequality (2.4) in particular gives

$$\|u_n\|_{H^1(\Omega)}^2 + \|w_n\|_{\mathbb{H}^1(\Omega)}^2 \leq \frac{1}{\mu} \|f\|_{L^2(\Omega)},$$

where w_n is the optimal w for u_n . This gives that (u_n, w_n) is uniformly bounded in $H^1(\Omega) \times \mathbb{H}^1(\Omega)$ and that there exists a subsequence $\{(\alpha_n, \beta_n, u_n, w_n)\}$ which converges weakly in $\mathbb{R}^2 \times H^1(\Omega) \times \mathbb{H}^1(\Omega)$ to a limit point $(\hat{\alpha}, \hat{\beta}, \hat{u}, \hat{w})$. Moreover, $u_n \rightarrow \hat{u}$ strongly in $L^p(\Omega)$ and $w_n \rightarrow \hat{w}$ in $L^p(\Omega; \mathbb{R}^n)$. Using the continuity of the L^2 fidelity term with respect to strong convergence in L^2 , and the weak lower semicontinuity of the H^1 term with respect to weak convergence in H^1 and of the Huber regularised functional even with respect to weak* convergence in \mathcal{M} (cf. Lemma 2.1) we get

$$\begin{aligned} & \frac{1}{2} \|\hat{u} - f\|_{L^2(\Omega)}^2 + \int_{\Omega} \hat{\alpha} |D\hat{u} - \hat{w}|_{\gamma} dx + \int_{\Omega} \hat{\beta} |Ew|_{\gamma} dx \\ & + \frac{\mu}{2} \left(\|\hat{u}\|_{H^1(\Omega)}^2 + \|\hat{w}\|_{\mathbb{H}^1(\Omega)}^2 \right) \\ \leq & \liminf_n \frac{1}{2} \|u_n - f\|_{L^2(\Omega)}^2 + \int_{\Omega} \hat{\alpha} |Du_n - w_n|_{\gamma} dx + \int_{\Omega} \hat{\beta} |Ew_n|_{\gamma} dx \\ & + \frac{\mu}{2} \left(\|u_n\|_{H^1(\Omega)}^2 + \|w_n\|_{\mathbb{H}^1(\Omega)}^2 \right) \\ \leq & \liminf_n \frac{1}{2} \|u_n - f\|_{L^2(\Omega)}^2 + \int_{\Omega} \alpha_n |Du_n - w_n|_{\gamma} dx + \int_{\Omega} \beta_n |Ew_n|_{\gamma} dx \\ & + \frac{\mu}{2} \left(\|u_n\|_{H^1(\Omega)}^2 + \|w_n\|_{\mathbb{H}^1(\Omega)}^2 \right), \end{aligned}$$

where in the last step we have used the boundedness of the sequence $R_{\alpha_n, \beta_n}^{\gamma, \mu}(u_n)$ from (2.4) and the convergence of (α_n, β_n) in \mathbb{R}^2 . This shows that the limit point \hat{u} is an optimal solution for $(\hat{\alpha}, \hat{\beta})$. Moreover, due to the weak lower semicontinuity of the cost functional F and the fact that the set $\{(\alpha, \beta) : 0 \leq \alpha \leq \bar{\alpha}, 0 \leq \beta \leq \bar{\beta}\}$ is closed, we have that $(\hat{\alpha}, \hat{\beta}, \hat{u})$ is optimal for (2.3). \square

Remark 2.1.

- Using the existence result in [17], in principle we could allow infinite values for α and β . This would include both TV² and TV as possible optimal regularisers in our learning problem.

- In [17], in the case of the L^2 cost and assuming that

$$R_{\alpha,\beta}^\gamma(f) > R_{\alpha,\beta}^\gamma(f_0),$$

we moreover show that the parameters (α, β) are strictly larger than 0. In the case of the Huberised TV cost this can only be proven in a discretised setting. Please see [17] for details.

- The existence of solutions with $\mu = 0$, that is without elliptic regularisation, is also proven in [17]. Note that here, we focus on the $\mu > 0$ case since the elliptic regularity is required for proving the existence of Lagrange multipliers in the next section.

3. LAGRANGE MULTIPLIERS

In this section we prove the existence of Lagrange multipliers for the learning problem (2.3) and derive an optimality system that characterizes its solution. Moreover, a gradient formula for the reduced cost functional is obtained, which plays an important role in the development of fast solution algorithms for the learning problems (see Section 5.1).

In what follows all proofs are presented for the TGV² regularisation case, that is $R_{\alpha,\beta}^\gamma = \text{TGV}_{\alpha,\beta}^{2,\gamma}$. However, possible modifications to cope with the ICTV model will also be commented.

We start by investigating the differentiability of the solution operator.

3.1. Differentiability of the solution operator. We recall that the TGV² denoising problem is given by

$$u = (v, w) = \arg \min_{BV \times BD} \left\{ \frac{1}{2} \int_{\Omega} |v - f|^2 + \int_{\Omega} \alpha |Dv - w|_{\gamma} + \int_{\Omega} \beta |Ew|_{\gamma} \right\}.$$

Using an elliptic regularization we then get

$$u = \arg \min_{H^1(\Omega) \times \mathbb{H}^1(\Omega)} \left\{ a(u, u) + \frac{1}{2} \int_{\Omega} |v - f|^2 + \int_{\Omega} \alpha |Dv - w|_{\gamma} + \int_{\Omega} \beta |Ew|_{\gamma} \right\},$$

where $a(u, u) = \mu (\|v\|_{H^1}^2 + \|w\|_{\mathbb{H}^1}^2)$. A necessary and sufficient optimality condition for the latter is then given by the following variational equation

$$(3.1) \quad a(u, \Psi) + \int_{\Omega} \alpha h_{\gamma}(Dv - w)(D\phi - \varphi) dx \\ + \int_{\Omega} \beta h_{\gamma}(Ew)E\varphi dx + \int_{\Omega} (v - f)\phi dx = 0, \text{ for all } \Psi \in U,$$

where $\Psi = (\phi, \varphi)$ and $U = H^1(\Omega) \times \mathbb{H}^1(\Omega)$.

Theorem 3.1. *The solution operator $S : \mathbb{R}^2 \mapsto U$, which assigns to each pair $(\alpha, \beta) \in \mathbb{R}^2$ the corresponding solution to the denoising problem (3.1), is Fréchet differentiable and its derivative is characterized by the unique solution $z = S'(\alpha, \beta)[\theta_1, \theta_2] \in U$ of the following linearized equation:*

$$(3.2) \quad a(z, \Psi) + \int_{\Omega} \theta_1 h_{\gamma}(Dv - w)(D\phi - \varphi) dx \\ + \int_{\Omega} \alpha h'_{\gamma}(Dv - w)(Dz_1 - z_2)(D\phi - \varphi) dx + \int_{\Omega} \theta_2 h_{\gamma}(Ew)E\varphi dx \\ + \int_{\Omega} \beta h'_{\gamma}(Ew)Ez_2E\varphi dx + \int_{\Omega} z_1\phi dx = 0, \text{ for all } \Psi \in U.$$

Proof. Thanks to the ellipticity of $a(\cdot, \cdot)$ and the monotonicity of h_γ , existence of a unique solution to the linearized equation follows from the Lax-Milgram theorem.

Let $\xi := u^+ - u - z$, where $u = S(\alpha, \beta)$ and $u^+ = S(\alpha + \theta_1, \beta + \theta_2)$. Our aim is to prove that $\|\xi\|_U = o(|\theta|)$. Combining the equations for u^+ , u and z we get that

$$\begin{aligned} a(\xi, \Psi) &+ \int_{\Omega} (\alpha + \theta_1) h_\gamma(Dv^+ - w^+)(D\phi - \varphi) dx - \int_{\Omega} \alpha h_\gamma(Dv - w)(D\phi - \varphi) dx \\ &- \int_{\Omega} \theta_1 h_\gamma(Dv - w)(D\phi - \varphi) dx - \int_{\Omega} \alpha h'_\gamma(Dv - w)(Dz_1 - z_2)(D\phi - \varphi) dx \\ &\quad + \int_{\Omega} (\beta + \theta_2) h_\gamma(Ew^+) E\varphi dx - \int_{\Omega} \beta h_\gamma(Ew) E\varphi dx \\ &- \int_{\Omega} \theta_2 h_\gamma(Ew) E\varphi dx - \int_{\Omega} \beta h'_\gamma(Ew) Ez_2 E\varphi dx + 2 \int_{\Omega} \xi_1 \phi dx = 0, \text{ for all } \Psi \in U, \end{aligned}$$

where $\xi := (\xi_1, \xi_2) \in H^1(\Omega) \times \mathbb{H}^1(\Omega)$. Adding and subtracting the terms

$$\int_{\Omega} \alpha h'_\gamma(Dv - w)(D\delta_v - \delta_w)(D\phi - \varphi) dx$$

and

$$\int_{\Omega} \beta h'_\gamma(Ew) E\delta_w : E\varphi dx,$$

where $\delta_v := v_{\alpha+\theta} - v$ and $\delta_w := w_{\alpha+\theta} - w$, we obtain that

$$\begin{aligned} a(\xi, \Psi) &+ \int_{\Omega} \alpha h'_\gamma(Dv - w)(D\xi_1 - \xi_2)(D\phi - \varphi) \\ &\quad + \int_{\Omega} \beta h'_\gamma(Ew) E\xi_2 : E\varphi dx + 2 \int_{\Omega} \xi_1 \phi dx \\ &= - \int_{\Omega} \alpha [h_\gamma(Dv^+ - w^+) - h_\gamma(Dv - w) - h'_\gamma(Dv - w)(D\delta_v - \delta_w)] (D\phi - \varphi) \\ &\quad - \int_{\Omega} \theta_1 [h_\gamma(Dv^+ - w^+) - h_\gamma(Dv - w)] (D\phi - \varphi) dx \\ &\quad - \int_{\Omega} \beta [h_\gamma(Ew^+) - h_\gamma(Ew) - h'_\gamma(Ew) E\delta_w] : E\varphi dx \\ &\quad - \int_{\Omega} \theta_2 [h_\gamma(Ew_{\alpha+\theta}) - h_\gamma(Ew)] : E\varphi dx, \text{ for all } \Psi \in U. \end{aligned}$$

Testing with $\Psi = \xi$ and using the monotonicity of $h'_\gamma(\cdot)$ we get that

$$\begin{aligned} \|\xi\|_U \leq C \Big\{ &|\alpha| \|h_\gamma(Dv^+ - w^+) - h_\gamma(Dv - w) - h'_\gamma(Dv - w)(D\delta_v - \delta_w)\|_{L^2} \\ &+ |\theta_1| \|h_\gamma(Dv^+ - w^+) - h_\gamma(Dv - w)\|_{L^2} \\ &+ |\beta| \|h_\gamma(Ew^+) - h_\gamma(Ew) - h'_\gamma(Ew) E\delta_w\|_{L^2} \\ &\quad + |\theta_2| \|h_\gamma(Ew_{\alpha+\theta}) - h_\gamma(Ew)\|_{L^2} \Big\}, \end{aligned}$$

for some generic constant $C > 0$. Considering the differentiability and Lipschitz continuity of $h'_\gamma(\cdot)$, it then follows that

$$(3.3) \quad \|\xi\|_U \leq C \left(|\alpha| o(\|u^+ - u\|_{1,p}) + |\theta_1| \|u_{\alpha+\theta} - u\|_U + |\beta| o(\|w^+ - w\|_{1,p}) + |\theta_2| \|w_{\alpha+\theta} - w\|_{\mathbb{H}^1(\Omega)} \right),$$

where $\|\cdot\|_{1,p}$ stands for the norm in the space $\mathbb{W}^{1,p}(\Omega)$. From regularity results for second order systems (see [19, Thm. 1, Rem. 14]), it follows that

$$\begin{aligned} \|u^+ - u\|_{1,p} &\leq L|\theta| (\|\operatorname{Div} h_\gamma(Dv - w)\|_{-1,p} + \|h_\gamma(Dv - w)\|_{-1,p} + \|\operatorname{Div} h_\gamma(Ew)\|_{-1,p}) \\ &\leq L|\theta| (2\|h_\gamma(Dv - w)\|_{L^\infty} + \|h_\gamma(Ew)\|_{L^\infty}) \\ &\leq \tilde{L}|\theta|, \end{aligned}$$

since $|h_\gamma(\cdot)| \leq 1$. Inserting the latter in estimate (3.3), we finally get that

$$\|\xi\|_U = o(|\theta|).$$

□

Remark 3.1. The Fréchet differentiability proof makes use of the quasilinear structure of the TGV² variational form, making it difficult to extend to the ICTV model without further regularisation terms. For the latter, however, a Gateaux differentiability result may be obtained using the same proof technique as in [16].

3.2. The adjoint equation. Next, we use the Lagrangian formalism for deriving the adjoint equations for both the TGV² and ICTV learning problems. Existence of a solution to the adjoint equation then follows from the well-posedness of the linearized equation.

Defining the Lagrangian associated to TGV² learning problem by:

$$\begin{aligned} \mathcal{L}(v, w, \alpha, \beta, p_1, p_2) &= F(u) + \mu(v, p_1)_{H^1} + \mu(w, p_2)_{\mathbb{H}^1} \\ &\quad + \int_{\Omega} \alpha h_\gamma(Dv - w)(Dp_1 - p_2) + \int_{\Omega} \beta h_\gamma(Ew)Ep_2 + \int_{\Omega} (v - f)p_1, \end{aligned}$$

and taking the derivative with respect to the state variable (v, w) , we get the necessary optimality condition

$$\begin{aligned} \mathcal{L}'_{(u,v)}(u, v, \alpha, \beta, p_1, p_2)[(\delta_v, \delta_w)] &= F'(u)\delta_u + \mu(p_1, \delta_v)_{H^1} + \mu(p_2, \delta_w)_{\mathbb{H}^1} \\ &\quad + \int_{\Omega} \alpha h'_\gamma(Dv - w)(D\delta_v - \delta_w)(Dp_1 - p_2) \\ &\quad + \int_{\Omega} \beta h'_\gamma(Ew)E\delta_w Ep_2 + \int_{\Omega} p_1\delta_v = 0. \end{aligned}$$

If $\delta_w = 0$, then

$$\mu(p_1, \delta_v)_{H^1} + \int_{\Omega} \alpha h'_\gamma(Dv - w)(Dp_1 - p_2)D\delta_v + \int_{\Omega} p_1\delta_v = -\nabla_v F(u)\delta_v,$$

whereas if $\delta_v = 0$, then

$$\begin{aligned} \mu(p_2, \delta_w)_{\mathbb{H}^1} - \int_{\Omega} \alpha h'_\gamma(Dv - w)(Dp_1 - p_2)\delta_w \\ + \int_{\Omega} \beta h'_\gamma(Ew) Ep_2 E\delta_w = -\nabla_w F(u)\delta_w. \end{aligned}$$

Existence of a unique solution then follows from the transposition method, since the linearised equation is well-posed.

Remark 3.2. For the ICTV model it is possible to proceed formally with the Lagrangian approach. We recall that a necessary and sufficient optimality condition for

the ICTV functional is given by

$$(3.4) \quad \mu(u, \phi)_{H^1} + \mu(\nabla v, \nabla \varphi)_{\mathbb{H}^1} + \int_{\Omega} \alpha h_{\gamma}(Du - \nabla v)(D\phi - \nabla \varphi) \\ + \int_{\Omega} \beta h_{\gamma}(D\nabla v)D\nabla \varphi + \int_{\Omega} (u - f)\phi = 0, \text{ for all } (\phi, \varphi) \in H^1(\Omega) \times \mathbb{H}^1(\Omega)$$

and the correspondent Lagrangian functional \mathcal{L} is given by

$$\mathcal{L}(u, v, \alpha, \beta, p_1, p_2) = F(u) + \mu(u, p_1)_{H^1} + \mu(\nabla v, \nabla p_2)_{\mathbb{H}^1} \\ + \int_{\Omega} \alpha h_{\gamma}(Du - \nabla v)(Dp_1 - \nabla p_2) + \int_{\Omega} \beta h_{\gamma}(D\nabla v)D\nabla p_2 + \int_{\Omega} (u - f)p_1.$$

Deriving the Lagrangian with respect to the state variable (u, v) and setting it equal to zero yields

$$\mathcal{L}'_{(u,v)}(u, v, \alpha, \beta, p_1, p_2)[(\delta_u, \delta_v)] = F'(u)\delta_u + \mu(p_1, \delta_u)_{H^1} + \mu(\nabla p_2, \nabla \delta_v)_{\mathbb{H}^1} \\ + \int_{\Omega} \alpha h'_{\gamma}(Du - \nabla v)(D\delta_u - \nabla \delta_v)(Dp_1 - \nabla p_2) \\ + \int_{\Omega} \beta h'_{\gamma}(D\nabla v)D\nabla \delta_v D\nabla p_2 + \int_{\Omega} p_1 \delta_u = 0.$$

By taking succesively $\delta_v = 0$ and $\delta_u = 0$, the following system is obtained

$$(3.5a) \quad \mu(p_1, \delta_u)_{H^1} + \int_{\Omega} \alpha h'_{\gamma}(Du - \nabla v)(Dp_1 - \nabla p_2)D\delta_u + \int_{\Omega} p_1 \delta_u = -F'(u)\delta_u.$$

$$(3.5b) \quad \mu(\nabla p_2, \nabla \delta_v)_{\mathbb{H}^1} + \int_{\Omega} \alpha h'_{\gamma}(Du - \nabla v)(Dp_1 - \nabla p_2)\nabla \delta_v \\ + \int_{\Omega} \beta h'_{\gamma}(D\nabla v)D\nabla p_2 D\nabla \delta_v = 0.$$

3.3. Optimality condition. Using the differentiability of the solution operator and the well-posedness of the adjoint equation, we derive next an optimality system for the characterization of local minima of the bilevel learning problem. Besides the optimality condition itself, a gradient formula arises as byproduct, which is of importance in the design of solution algorithms for the learning problems.

Theorem 3.2. *Let $(\bar{\alpha}, \bar{\beta}) \in \mathbb{R}_+^2$ be a local optimal solution for problem (2.3). Then there exist Lagrange multipliers $\Pi \in U$ and $\lambda_1, \lambda_2 \in L^2(\Omega)$ such that the following system holds:*

$$(3.6a) \quad a(u, \Psi) + \alpha \int_{\Omega} h_{\gamma}(Dv - w)(D\phi - \varphi) dx \\ + \beta \int_{\Omega} h_{\gamma}(Ew)E\varphi dx + 2 \int_{\Omega} (v - f)\phi dx = 0, \text{ for all } \Psi \in H^1(\Omega) \times \mathbb{H}^1(\Omega),$$

$$(3.6b) \quad a(\Pi, \Psi) + \alpha \int_{\Omega} h'_{\gamma}(Dv - w)(Dp_1 - p_2)(D\phi - \varphi) dx \\ + \beta \int_{\Omega} h'_{\gamma}(Ew) Ep_2 E\varphi dx + 2 \int_{\Omega} p_1 \phi dx = -F_u(u)[\Psi], \text{ for all } \Psi \in H^1(\Omega) \times \mathbb{H}^1(\Omega),$$

$$(3.6c) \quad \lambda_1 = \int_{\Omega} h_{\gamma}(Dv - w)(Dp_1 - p_2)$$

$$(3.6d) \quad \lambda_2 = \int_{\Omega} h_{\gamma}(Ew) Ep_2$$

$$(3.6e) \quad \lambda_1 \geq 0, \quad \lambda_2 \geq 0$$

$$(3.6f) \quad \lambda_1 \cdot \bar{\alpha} = \lambda_2 \cdot \bar{\beta} = 0.$$

Proof. Consider the reduced cost functional $\mathcal{F}(\alpha, \beta) = F(u(\alpha, \beta))$. The bilevel optimization problem can then be formulated as

$$\min_{(\alpha, \beta) \in C} \mathcal{F}(\alpha, \beta),$$

where $\mathcal{F} : \mathbb{R}^2 \rightarrow \mathbb{R}$ and C corresponds to the positive orthant in \mathbb{R}^2 . From [38, Thm. 3.1], there exist multipliers λ_1, λ_2 such that

$$\begin{aligned} \lambda_1 &= \nabla_{\alpha} \mathcal{F}(\bar{\alpha}, \bar{\beta}) \\ \lambda_2 &= \nabla_{\beta} \mathcal{F}(\bar{\alpha}, \bar{\beta}) \\ \lambda_1 &\geq 0, \quad \lambda_2 \geq 0 \\ \lambda_1 \cdot \bar{\alpha} &= \lambda_2 \cdot \bar{\beta} = 0, \end{aligned}$$

By taking the derivative with respect to (α, β) and denoting by u' the solution to the linearized equation (3.2) we get, together with the adjoint equation (3.6b), that

$$\begin{aligned} \mathcal{F}'(\alpha, \beta)[\phi] &= F_u(u)u'(\alpha, \beta)[\phi] \\ &= -a(\Pi, u') - \alpha \int_{\Omega} h'_{\gamma}(Dv - w)(Dp_1 - p_2)(Dv' - w') \\ &\quad - \beta \int_{\Omega} h'_{\gamma}(Ew)Ep_2 Ew' - 2 \int_{\Omega} p_1 v' \\ &= -a(u', \Pi) - \alpha \int_{\Omega} h'_{\gamma}(Dv - w)(Dv' - w')(Dp_1 - p_2) \\ &\quad - \beta \int_{\Omega} h'_{\gamma}(Ew)Ew' Ep_2 - 2 \int_{\Omega} v' p_1 \end{aligned}$$

which, taking into account the linearized equation, yields

$$(3.7) \quad \mathcal{F}'(\alpha, \beta)[\phi] = \phi_1 \int_{\Omega} h_{\gamma}(Dv - w)(Dp_1 - p_2) + \phi_2 \int_{\Omega} h_{\gamma}(Ew)Ep_2.$$

Altogether we proved the result. \square

Remark 3.3. From the existence result (see Remark 2.1), we actually know that, under some assumptions, $\bar{\alpha}$ and $\bar{\beta}$ are strictly greater than zero. This implies that the multipliers $\lambda_1 = \lambda_2 = 0$ and the problem is of unconstrained nature. This plays an important role in the design of solution algorithms, since only a mild treatment of the constraints has to be taken into account, as will be showed in Section 6.

4. NUMERICAL ALGORITHMS

In this section we propose a second order quasi-Newton method for the solution of the learning problem with scalar regularisation parameters. The algorithm is based on a BFGS update, preserving the positivity of the iterates through the line search strategy and updating the matrix cyclically depending on the satisfaction of the curvature condition. For the solution of the lower level problem, a semismooth Newton method with a properly modified Jacobi matrix is considered. Moreover, warm initialisation strategies have to be taken into account in order to get convergence for the TGV² problem. The developed algorithm is also extended to a simple linear polynomial case.

4.1. BFGS algorithm. Thanks to the gradient characterization obtained in Theorem 3.2, we next devise a BFGS algorithm to solve the bilevel learning problems. We employ a few technical tricks to ensure convergence of the classical method. In particular, for numerical stability we need to avoid the boundary of the constraint set on the parameters, so we pick $0 < \theta < \Theta$, considered numerically almost zero or infinity, respectively, and require the box constraints

$$(4.1) \quad \theta \leq \alpha, \beta \leq \Theta.$$

We also limit the step length to get at most a fraction closer to the boundary. As we show in [17] the solution is in the interior for the regularisation and cost functionals we are interested in. Below this limit, we use Armijo line search.

Moreover, the good behaviour of the BFGS method depends upon the BFGS matrix staying positive definite. This would be ensured by the Wolfe conditions, but because of our step length limitation, the curvature condition is not necessarily satisfied. (The Wolfe conditions are guaranteed to be satisfied for some step length σ , if our domain is unbounded, but the range where the step satisfies the criterion, may be beyond our maximum step length, and is not necessarily satisfied closer to the current point.) Instead we skip the BFGS update if the curvature is negative.

Overall our learning algorithm may be written as follows.

Algorithm 4.1 (BFGS for denoising parameter learning). Pick Armijo line search constant c , and target residual ρ . Pick initial iterate (α^0, β^0) . Solve the denoising problem (2.3b) for $(\alpha, \beta) = (\alpha^0, \beta^0)$, yielding u^0 . Initialise $B^1 = I$. Set $i := 0$, and iterate the following steps:

- (1) Solve the adjoint equation (3.6b) for Π^i , and calculate $\nabla \mathcal{F}(\alpha^i, \beta^i)$ from (3.7).
- (2) If $i \geq 2$ do the following:
 - (a) Set $s := (\alpha^i, \beta^i) - (\alpha^{i-1}, \beta^{i-1})$, and $r := \nabla \mathcal{F}(\alpha^i, \beta^i) - \nabla \mathcal{F}(\alpha^{i-1}, \beta^{i-1})$.
 - (b) Perform the BFGS update

$$B^i := \begin{cases} B^{i-1}, & s^T r < 0, \\ B^{i-1} - \frac{B^{i-1} s \otimes B^{i-1} s}{t^T B^{i-1} s} + \frac{r \otimes r}{s^T r}, & s^T r \geq 0. \end{cases}$$

- (3) Compute $\delta_{\alpha, \beta}$ from

$$B^i \delta_{\alpha, \beta} = g^i.$$

- (4) Initialise $\sigma := \min\{1, \sigma_{\max}/2\}$, where

$$\sigma_{\max} := \max\{\sigma \geq 0 \mid (\alpha^i, \beta^i) + \sigma \delta_{\alpha, \beta} \text{ satisfies (4.1)}\}.$$

Repeat the following:

- (a) Let $(\alpha_\sigma, \beta_\sigma) := (\alpha^i, \beta^i) + \sigma \delta_{\alpha, \beta}$, and solve the denoising problem (2.3b) for $(\alpha, \beta) = (\alpha_\sigma, \beta_\sigma)$, yielding u_σ .
 - (b) If the residual $\|(\alpha_\sigma, \beta_\sigma) - (\alpha^i, \beta^i)\| / \|(\alpha_\sigma, \beta_\sigma)\| < \rho$ do the following:
 - (i) If $\min_\sigma \mathcal{F}(\alpha_\sigma, \beta_\sigma) < \mathcal{F}(\alpha^i, \beta^i)$ over all σ tried, choose σ^* the minimiser, set $(\alpha^{i+1}, \beta^{i+1}) := (\alpha_{\sigma^*}, \beta_{\sigma^*})$, $u^{i+1} := u_{\sigma^*}$, and continue from Step 5
 - (ii) Otherwise end the algorithm with solution $(\alpha^*, \beta^*) := (\alpha^i, \beta^i)$.
 - (c) Otherwise, if Armijo condition $\mathcal{F}(\alpha_\sigma, \beta_\sigma) \leq \mathcal{F}(\alpha^i, \beta^i) + \sigma c \nabla \mathcal{F}(\alpha^i, \beta^i)^T \delta_{\alpha, \beta}$ holds, set $(\alpha^{i+1}, \beta^{i+1}) := (\alpha_\sigma, \beta_\sigma)$, $u^{i+1} := u_\sigma$, and continue from Step 5.
 - (d) In all other cases, set $\sigma := \sigma/2$ and continue from Step 4a.
- (5) If the residual $\|(\alpha^{i+1}, \beta^{i+1}) - (\alpha^i, \beta^i)\| / \|(\alpha^{i+1}, \beta^{i+1})\| < \rho$, end the algorithm with $(\alpha^*, \beta^*) := (\alpha^{i+1}, \beta^{i+1})$. Otherwise continue from Step 1 with $i := i + 1$.

Step (4) ensures that the iterates remain feasible, without making use of a projection step. This is justified since it's been analytically proved that the optimal parameters are greater than zero (see [17]).

4.2. **An infeasible semi-smooth Newton method.** In variational form, the TGV² denoising problem can be written as

$$\begin{aligned} \mu \int_{\Omega} (Dv \cdot D\phi + v\phi) + \int_{\Omega} \alpha h_{\gamma}(Dv - w)D\phi + \int_{\Omega} (v - f)\phi &= 0, \quad \forall \phi \in H^1(\Omega) \\ \mu \int_{\Omega} (Ew : E\varphi + w\varphi) - \int_{\Omega} \alpha h_{\gamma}(Dv - w)D\varphi \\ &+ \int_{\Omega} \beta h_{\gamma}(Ew) E\varphi = 0, \quad \forall \varphi \in \mathbb{H}^1(\Omega) \end{aligned}$$

or, in general abstract primal-dual form, as

$$(4.2a) \quad Lu + \sum_{i=1}^N A_i^* q_i = f \quad \text{in } \Omega$$

$$(4.2b) \quad \max\{1/\gamma, |[A_j u](x)|_2\} q_j(x) - \alpha_j [A_j u](x) = 0 \text{ a.e. in } \Omega, \quad j = 1, \dots, N.$$

where $L \in \mathcal{L}(H^1(\Omega; \mathbb{R}^m), H^1(\Omega; \mathbb{R}^m)')$ is a second order linear elliptic operator, A_j , $j = 1, \dots, N$, are linear operators acting on u and $q_j(x)$, $j = 1, \dots, N$, correspond to the dual multipliers.

Let us set

$$\mathbf{m}_j(u) := \max\{1/\gamma, |[A_j u](x)|_2\}.$$

Let us also define the diagonal application $\mathfrak{D}(u) : L^2(\Omega; \mathbb{R}^m) \rightarrow L^2(\Omega; \mathbb{R}^m)$ by

$$[\mathfrak{D}(u)q](x) = u(x)q(x), \quad (x \in \Omega)$$

We may derive $\nabla_u[\mathfrak{D}(\mathbf{m}_j(u))q_j]$ being defined by

$$\nabla_u[\mathfrak{D}(\mathbf{m}_j(u))p_j] = A_j^* \mathfrak{D}(q_j) \mathfrak{N}(A_j u) \quad \text{where} \quad \mathfrak{N}(z) := \begin{cases} 0, & |z(x)|_2 < 1/\gamma \\ \frac{z(x)}{|z(x)|_2}, & |z(x)|_2 \geq 1/\gamma. \end{cases}$$

Then (4.2a), (4.2b) may be written as

$$\begin{aligned} Lu + \sum_{i=1}^N A_i^* q_i &= f \quad \text{in } \Omega \\ \mathfrak{D}(\mathbf{m}_j(u))q_j - \alpha_j A_j u &= 0, \quad \text{a.e. in } \Omega, \quad (j = 1, \dots, N). \end{aligned}$$

Linearising, we obtain the system

(SSN-1)

$$\begin{pmatrix} L & A_1^* & \dots & A_N^* \\ -\alpha_1 A_1 + \mathfrak{N}(A_1 u) \mathfrak{D}(q_1) A_1 & \mathfrak{D}(\mathbf{m}_1(u)) & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ -\alpha_N A_N + \mathfrak{N}(A_N u) \mathfrak{D}(q_N) A_N & 0 & 0 & \mathfrak{D}(\mathbf{m}_N(u)) \end{pmatrix} \begin{pmatrix} \delta u \\ \delta q_1 \\ \vdots \\ \delta q_N \end{pmatrix} = R$$

where

$$R := \begin{pmatrix} -Lu - \sum_{i=1}^N A_i^* q_i + f \\ \alpha_1 A_1 u - \mathfrak{D}(\mathbf{m}_1(u)) q_1 \\ \vdots \\ \alpha_N A_N u - \mathfrak{D}(\mathbf{m}_N(u)) q_N \end{pmatrix}.$$

The semi-smooth Newton method solves (SSN-1) at a current iterate $(u^i, q_1^i, \dots, q_N^i)$. It then updates

$$(SSN-2) \quad (u^{i+1}, \tilde{q}_1^{i+1}, \dots, \tilde{q}_N^{i+1}) := (u^i + \tau \delta u, q_1^i + \tau \delta q_1, q_N^i + \tau \delta q_N),$$

for a suitable step length τ , allowing \tilde{q}^{i+1} to become infeasible in the process. That is, it may hold that $|\tilde{q}_j^{i+1}(x)|_2 > \alpha_j$, which may lead to non-descent directions. In order to globalize the method, one projects

$$(SSN-3) \quad q_j^{i+1} := \mathfrak{P}(\tilde{q}_j^{i+1}; \alpha_j), \quad \text{where} \quad \mathfrak{P}(q, \alpha)(x) := \text{sgn}(q(x)) \min\{\alpha, |q(x)|\},$$

in the building of the Jacobi matrix. Following [23, 33], it can be shown that a discrete version of the method (SSN-1)–(SSN-3) converges globally and locally superlinearly near a point where the subdifferentials of the operator on (u, q_1, \dots, q_N) corresponding (4.2) are non-singular. Further dampening as in [23] guarantees local superlinear convergence at any point. We do not represent the proof, as going into the discretisation and dampening details would expand this work considerably.

Remark 4.1. The system (SSN-1) can be simplified, which is crucial to obtain acceptable performance with TGV². Indeed observe that B is invertible, so we may solve δu from

$$(4.3) \quad B\delta u = R_1 - \sum_{j=1}^N A_j^* \delta q_j.$$

Thus we may simplify δu out of (SSN-1), and only solve for $\delta q_1, \dots, \delta q_N$ using a reduced system matrix. Finally we calculate δu from (4.3).

For the denoising sub-problem (2.3b) we use the method (SSN-1)–(SSN-3) with the reduced system matrix of Remark 4.1. Here, we denote by z in the case of TGV² the parameters

$$z = (v, w),$$

and in the case of ICTV

$$z = (u, v).$$

For the calculation of the step length τ , we use Armijo line search with parameter $c = 1E^{-4}$. We end the SSN iterations when

$$\tau \frac{\|\delta y^i\|}{\max\{1, \|y^i\|\}} \leq 1E^{-5},$$

where $\delta y^i = (\delta z^i, \delta q_1^i, \dots, \delta q_N^i)$, and $y^i = (z^i, q_1^i, \dots, q_N^i)$.

4.3. Warm initialisation. In our numerical experimentation we generally found Algorithm 4.1 to perform well for learning the regularisation parameter for TV denoising as was done in [16]. For learning the two (or even more) regularisation parameters for TGV² denoising, we found that a warm initialisation is needed to obtain convergence. More specifically, we use TV as an aid for discovering both the initial iterate (α^0, β^0) as well as the initial BFGS matrix B^1 . This is outlined in the following algorithm.

Algorithm 4.2 (BFGS initialisation for TGV² parameter learning). Pick a heuristic factor $\delta_0 > 0$. Then do the following:

- (1) Solve the corresponding problem for TV using Algorithm 4.1. This yields optimal TV denoising parameter α_{TV}^* , as well as the BFGS estimate B_{TV} for $\nabla^2 \mathcal{F}(\alpha_{\text{TV}}^*)$.
- (2) Run Algorithm 4.1 for TGV² with initialisation $(\alpha^0, \beta^0) := (\alpha_{\text{TV}}^* \delta_0, \alpha_{\text{TV}}^*)$, and initial BFGS matrix $B^1 := \text{diag}(B_{\text{TV}} \delta_0, B_{\text{TV}})$.

With $\Omega = (0, 1)^2$, we pick $\delta_0 = 1/\ell$, where the original discrete image has $\ell \times \ell$ pixels. This corresponds to the heuristic [35, 2] that if $\ell \approx 128$ or 256 and the discrete image is mapped into the corresponding domain $\Omega = (0, \ell)^2$ directly (corresponding to spatial step size of one in the discrete gradient operator), then $\beta \in (\alpha, 1.5\alpha)$ tends to be a good choice. We will later verify this through the use of our algorithms.

Now, if $f \in \text{BV}((0, \ell)^2)$ is rescaled to $\text{BV}((0, 1)^2)$, i.e. $\tilde{f}(x) := f(x/\ell)$, then with $\tilde{u}(x) := u(x/\ell)$ and $\tilde{w}(x) := w(x/\ell)/\ell$, we have

$$(4.4) \quad \frac{1}{2} \|f - u\|_{L^2((0, \ell)^2)}^2 + \alpha \|Du - w\|_{\mathcal{M}((0, \ell)^2; \mathbb{R}^2)} + \beta \|Ew\|_{\mathcal{M}((0, \ell)^2; \mathbb{R}^{2 \times 2})} \\ = n^2 \left(\frac{1}{2} \|\tilde{f} - \tilde{u}\|_{L^2((0, 1)^2)}^2 + n\alpha \|D\tilde{u} - \tilde{w}\|_{\mathcal{M}((0, 1)^2; \mathbb{R}^2)} + n^2 \beta \|E\tilde{w}\|_{\mathcal{M}((0, 1)^2; \mathbb{R}^{2 \times 2})} \right).$$

This introduces the factor $1/\ell = |\Omega|^{-1/2}$ between rescaled α, β .

5. EXPERIMENTS

In this section we present some numerical experiments to verify the theoretical properties of the bilevel learning problems and the efficiency of the proposed solution algorithms. In particular, we exhaustively compare the performance of the new proposed cost functional with respect to well-known quality measures, showing a better behaviour of the new cost for the chosen tested images. The performance of the proposed BFGS algorithm, combined with the semismooth Newton method for the lower level problem, is also examined.

5.1. Gaussian denoising. We tested Algorithm 4.1 for TV and Algorithm 4.2 for TGV^2 Gaussian denoising parameter learning on various images. Here we report the results for two images, the parrot image in Figure 4a, and the geometric image in Figure 5. We applied synthetic noise to the original images, such that the PSNR of the parrot image is 24.7, and the PSNR of the geometric image is 24.8.

In order to learn the regularisation parameter α for TV, we picked initial $\alpha^0 = 0.1/\ell$. For TGV^2 initialisation by TV was used as in Algorithm 4.1. We chose the other parameters of Algorithm 4.1 as $c = 1\text{E}-4$, $\rho = 1\text{E}-5$, $\theta = 1\text{E}-8$, and $\Theta = 10$. For the SSN denoising method the parameters $\gamma = 100$ and $\mu = 1\text{E}-10$ were chosen.

We have included results for both the L^2 -squared cost functional L_2^2 and the Huberised total variation cost functional $L_\eta^1 \nabla$. The learning results are reported in Table 1 for the parrot images, and Table 2 for the geometric image. The denoising results with the discovered parameters can be found in the aforementioned Figure 4 and Figure 5. We report the resulting optimal parameter values, the cost functional value, PSNR, SSIM [37], as well as the number of iterations taken by the outer BFGS method.

Our first observation is that all approaches successfully learn a denoising parameter that gives a good-quality denoised image. Secondly, we observe that the gradient cost functional $L_\eta^1 \nabla$ performs visually and in terms of SSIM significantly better for TGV^2 parameter learning than the cost functional L_2^2 . In terms of PSNR the roles are reversed, as should be, since the L_2^2 is equivalent to PSNR. This again confirms that PSNR is a poor quality measure for images. For TV there is no significant difference between different cost functionals in terms of visual quality, although the PSNR and SSIM differ.

We also observe that the optimal TGV^2 parameters (α^*, β^*) generally satisfy $\beta^*/\alpha^* \in (0.75, 1.5)/\ell$. This confirms the earlier observed heuristic that if $\ell \approx 128, 256$ then $\beta \in (1, 1.5)\alpha$ tends to be a good choice. As we can observe from Figure 4 and Figure 5, this optimal TGV^2 parameter choice also avoids the stair-casing effect that can be observed with TV in the results.

In Figure 3, we have plotted by the red star the discovered regularisation parameter (α^*, β^*) reported in Figure 4. Studying the location of the red star, we may conclude that Algorithm 4.1 and Algorithm 4.2 manage to find a nearly optimal parameter in very few BFGS iterations.

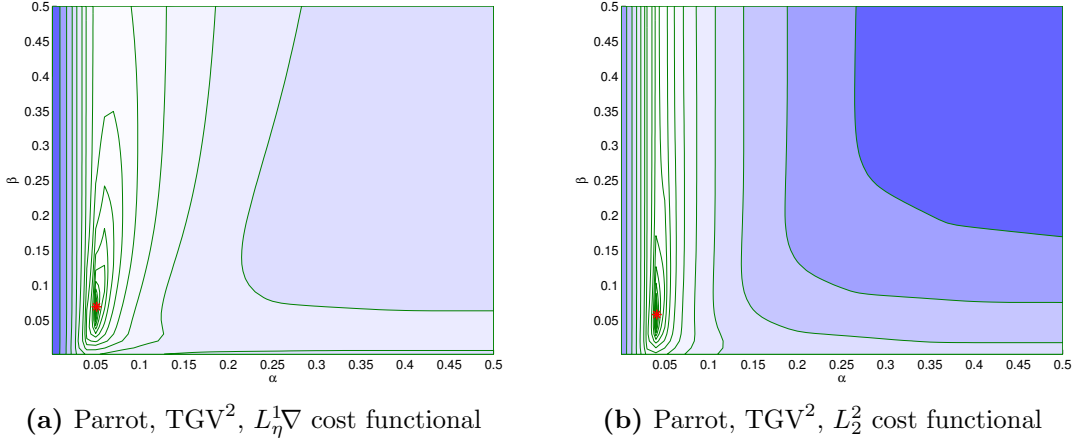


Figure 3. Cost functional value versus (α, β) for TGV^2 denoising, for the parrot test images, for both L_2^2 and $L_\eta^1 \nabla$ cost functionals. The illustrations are contour plots of function value versus (α, β) .

Table 1. Quantified results for the parrot image ($\ell = 256 =$ image width/height in pixels)

Denoise	Cost	Initial (α, β)	Result (α^*, β^*)	Cost	SSIM	PSNR	Its.	Fig.
TGV^2	$L_\eta^1 \nabla$	$(\alpha_{TV}^*/\ell, \alpha_{TV}^*)$	$(0.069/\ell^2, 0.051/\ell)$	6.615	0.897	31.720	12	4(c)
TGV^2	L_2^2	$(\alpha_{TV}^*/\ell, \alpha_{TV}^*)$	$(0.058/\ell^2, 0.041/\ell)$	6.412	0.890	31.992	11	4(d)
ICTV	$L_\eta^1 \nabla$	$(\alpha_{TV}^*/\ell, \alpha_{TV}^*)$	$(0.068/\ell^2, 0.051/\ell)$	6.656	0.895	31.667	16	4(e)
ICTV	L_2^2	$(\alpha_{TV}^*/\ell, \alpha_{TV}^*)$	$(0.051/\ell^2, 0.041/\ell)$	6.439	0.887	31.954	7	4(f)
TV	$L_\eta^1 \nabla$	$0.1/\ell$	$0.057/\ell$	6.944	0.887	31.298	10	4(g)
TV	L_2^2	$0.1/\ell$	$0.042/\ell$	6.623	0.879	31.710	12	4(h)

Table 2. Quantified results for the synthetic image ($\ell = 256 =$ image width/height in pixels)

Denoise	Cost	Initial $\vec{\alpha}$	Result $\vec{\alpha}^*$	Value	SSIM	PSNR	Its.	Fig.
TGV^2	$L_\eta^1 \nabla$	$(\alpha_{TV}^*/\ell, \alpha_{TV}^*)$	$(0.453/\ell^2, 0.071/\ell)$	3.769	0.989	36.606	17	5(c)
TGV^2	L_2^2	$(\alpha_{TV}^*/\ell, \alpha_{TV}^*)$	$(0.307/\ell^2, 0.055/\ell)$	3.603	0.986	36.997	19	5(d)
ICTV	$L_\eta^1 \nabla$	$(\alpha_{TV}^*/\ell, \alpha_{TV}^*)$	$(0.505/\ell^2, 0.103/\ell)$	4.971	0.970	34.201	23	5(e)
ICTV	L_2^2	$(\alpha_{TV}^*/\ell, \alpha_{TV}^*)$	$(0.056/\ell^2, 0.049/\ell)$	3.947	0.965	36.206	7	5(f)
TV	$L_\eta^1 \nabla$	$0.1/\ell$	$0.136/\ell$	5.521	0.966	33.291	6	5(g)
TV	L_2^2	$0.1/\ell$	$0.052/\ell$	4.157	0.948	35.756	7	5(h)

5.2. Statistical testing. To obtain a statistically significant outlook to the performance of different regularisers and cost functionals, we made use of the Berkeley segmentation dataset BSDS300 [28], displayed in Figure 6. We resized each image to 128 pixels on its shortest edge, and take the 128×128 top-left square of the image. To this data set, we applied pixelwise Gaussian noise of variance $\sigma^2 = 2, 10$, and 20. We tested the performance of both cost functionals, $L_\eta^1 \nabla$ and L_2^2 , as well as the TGV^2 , ICTV, and TV regularisers on this dataset, for all noise levels. In the first instance, reported in Figures 7–10 (noise levels $\sigma^2 = 2, 20$ only), and Tables 3–5, we applied the proposed bi-level learning model on each image individually, to learn the optimal parameters specifically for that image, and a corresponding noisy image for

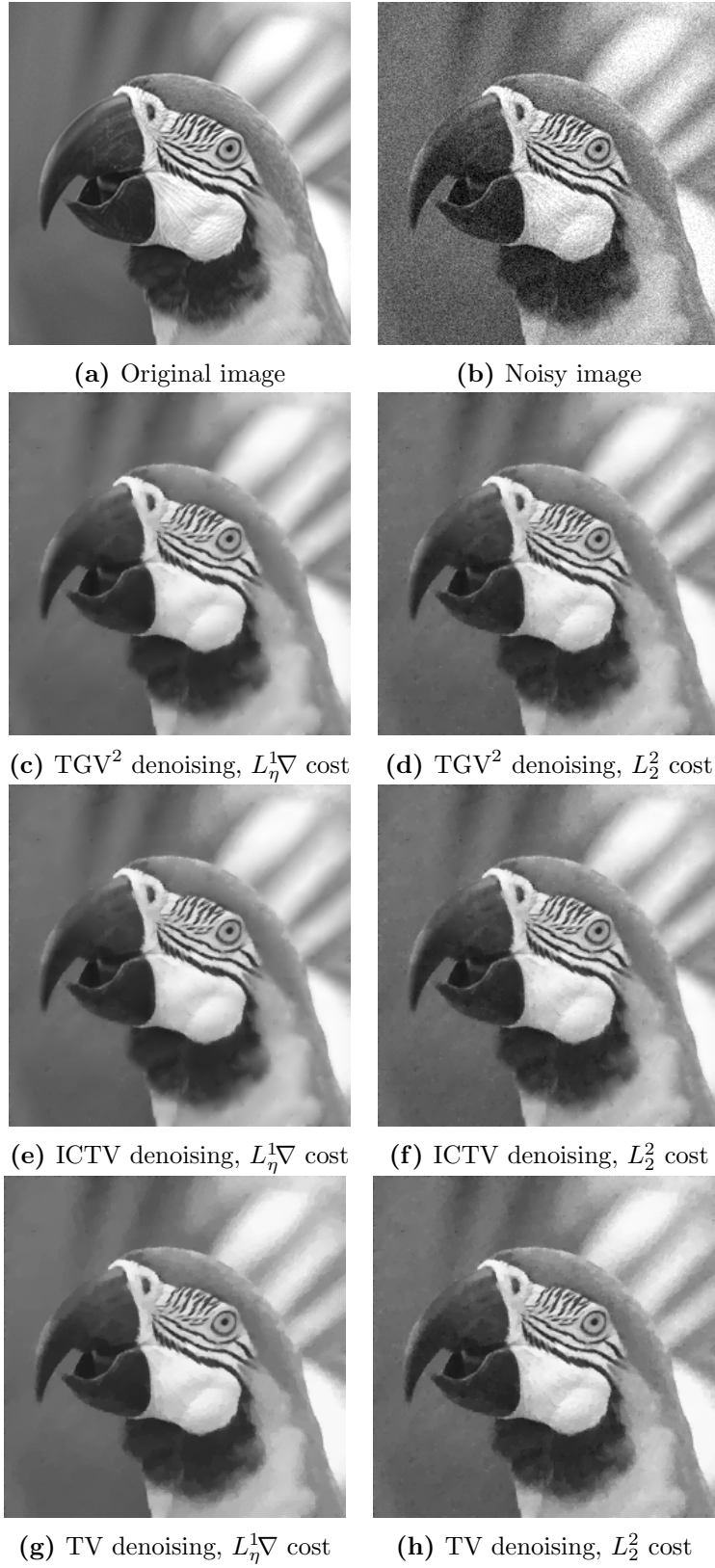


Figure 4. Optimal denoising results for initial guess $\vec{\alpha} = (\alpha_{TV}^*/\ell, \alpha_{TV}^*)$ for TGV^2 and $\vec{\alpha} = 0.1/\ell$ for TV

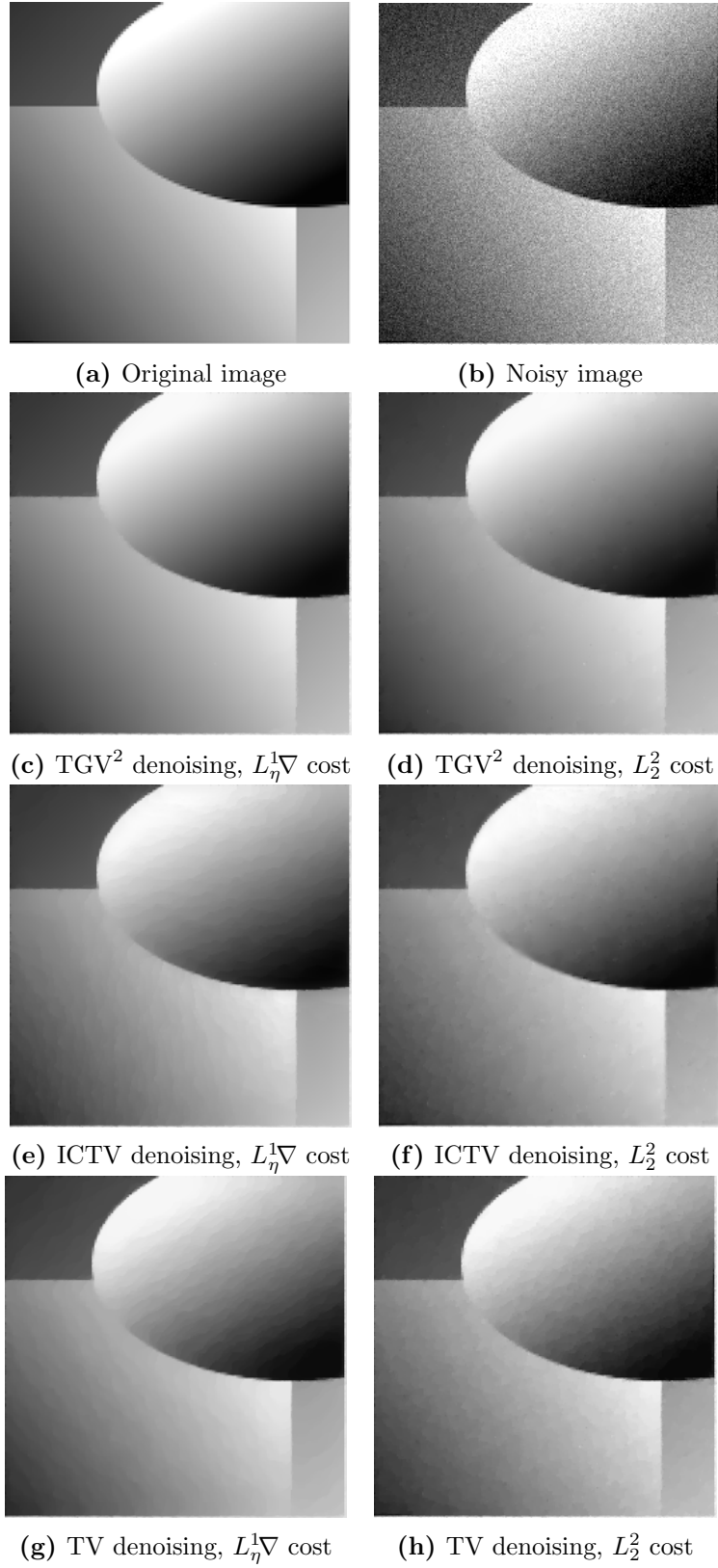


Figure 5. Optimal denoising results for initial guess $\vec{\alpha} = (\alpha_{TV}^*/\ell, \alpha_{TV}^*)$ for TGV^2 and $\vec{\alpha} = 0.2/\ell$ for TV



Figure 6. The 200 images of the Berkeley segmentation dataset BSDS300 [28], cropped to be rectangular, keeping top left corner, and resized to 128×128 .

all of the noise levels separately. For the algorithm, we use the same parametrisation as in Section 5.1.

The figures display the noisy images, and indicate by colour coding the best result as judged by the structural similarity measure SSIM [37], PSNR, and the objective function value ($L_\eta^1 \nabla$ or L_2^2 cost). These criteria are, respectively, the top, middle, and bottom rows of colour-coding squares. Red square indicates that TV performed the best, green square indicates that ICTV performed the best, and blue square indicates that TGV^2 performed the best—this is naturally for the optimal parameters for the corresponding regulariser and cost functional discovered by our algorithms.

In the tables, we report the information in a more concise numerical fashion, indicating the mean, standard deviation, and median for all the different criteria (SSIM, PSNR, and cost functional value), as well as the number of images for which each regulariser performed the best. We recall that SSIM is normalised to $[0, 1]$, with higher value better. Moreover, we perform a statistical 95% one-tailed paired t-test on each of the criteria, and a pair of regularisers, to see whether any pair of regularisers can be ordered. If so, this is indicated in the last row of each of the tables.

Overall, studying the t-test and other data, the ordering of the regularisers appears to be

$$ICTV > TGV^2 > TV.$$

This is rather surprising, as in many specific examples, TGV^2 has been observed to perform better than ICTV, see our Figures 4 and 5, as well as [4, 1]. Only when the noise is high, appears TGV^2 to come on par with ICTV with the $L_\eta^1 \nabla$ cost functional in Figure 9 and Table 5.

A more detailed study of the results in Figures 7–10 seems to indicate that TGV^2 performs better than ICTV when the image contains large smooth areas, but ICTV generally performs better for more chaotic images. This observation agrees with the results in Figures 4 and 5, as well as [4, 1], where the images are of the former type.

One possible reason for the better performance of ICTV could be that TGV^2 has more degrees of freedom—in ICTV we essentially constrain $w = \nabla v$ for some function v —and therefore overfits to the noisy data, until the noise level becomes so high that overfitting would become too high for any parameter. To see whether this is true, we



Figure 7. Ordering of regularisers with individual learning, $L_\eta^1 \nabla$ cost, and noise variance $\sigma^2 = 2$, on the 200 images of the BSDS300 dataset, resized. Best regulariser: red=TV, green=ICTV, blue=TGV²; top=SSIM, middle=PSNR, bottom=objective value.



Figure 8. Ordering of regularisers with individual learning, L_2^2 cost, and noise variance $\sigma^2 = 2$, on the 200 images of the BSDS300 dataset, resized. Best regulariser: red=TV, green=ICTV, blue=TGV²; top=SSIM, middle=PSNR, bottom=objective value.

also performed batch learning, learning a single set of parameters for all images with the same noise level. That is, we studied the model

$$\min_{\bar{\alpha}} \sum_{i=1}^N F_i(u_{i,\bar{\alpha}}) \quad \text{s.t.} \quad u_{i,\bar{\alpha}} \in \arg \min_{u \in H^1(\Omega)} \frac{1}{2} \|f_i - u\|_{L^2(\Omega)}^2 + R_{\bar{\alpha}}^{\gamma,\mu}(u),$$

with

$$F_i(u) = \frac{1}{2} \|f_{0,i} - u\|_{L^2(\Omega)}^2, \quad \text{or} \quad F_i(u) = \int_{\Omega} |\nabla(f_{0,i} - u)|_{\gamma} dx,$$

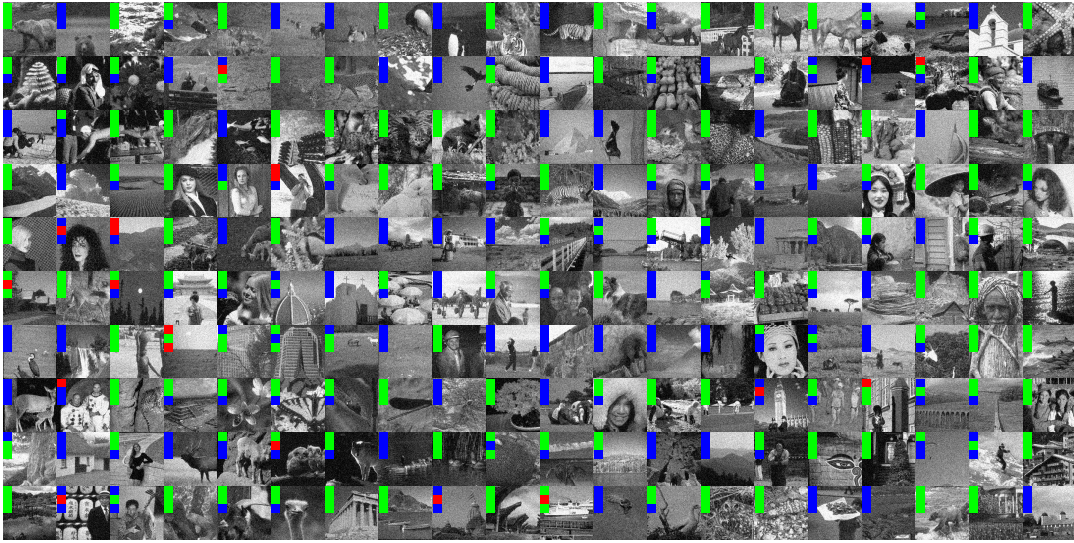


Figure 9. Ordering of regularisers with individual learning, $L_\eta^1 \nabla$ cost, and noise variance $\sigma^2 = 20$, on the 200 images of the BSDS300 dataset, resized. Best regulariser: red=TV, green=ICTV, blue=TGV²; top=SSIM, middle=PSNR, bottom=objective value.



Figure 10. Ordering of regularisers with individual learning, L_2^2 cost, and noise variance $\sigma^2 = 20$, on the 200 images of the BSDS300 dataset, resized. Best regulariser: red=TV, green=ICTV, blue=TGV²; top=SSIM, middle=PSNR, bottom=objective value.

where $\vec{\alpha} = (\alpha, \beta)$, f_1, \dots, f_N are the $N = 200$ noisy images with the same noise level, and $f_{0,1}, \dots, f_{0,N}$ the original noise free images.

The results are in Figures 11–14 (noise levels $\sigma^2 = 2, 20$ only), and Tables 6–8. The results are still roughly the same as with individual learning. Again, only with high noise in Table 8, does TGV² not lose to ICTV. Another interesting observation is that TV starts to be frequently the best regulariser for individual images, although still statistically does worse than either ICTV or TGV².

	SSIM				PSNR				value			
	mean	std	med	best	mean	std	med	best	mean	std	med	best
Noisy data	0.978	0.015	0.981	0	41.56	0.86	41.95	0	2.9E ⁴	3.1E ²	2.9E ⁴	0
$L_\eta^1 \nabla$ -TV	0.988	0.005	0.989	1	42.57	1.10	42.46	5	2.4E ⁴	3.7E ³	2.5E ⁴	1
$L_\eta^1 \nabla$ -ICTV	0.989	0.005	0.990	141	42.74	1.16	42.62	143	2.3E ⁴	3.9E ³	2.4E ⁴	137
$L_\eta^1 \nabla$ -TGV ²	0.989	0.005	0.989	58	42.70	1.17	42.55	52	2.4E ⁴	4.0E ³	2.5E ⁴	62
95% t-test	ICTV > TGV ² > TV				ICTV > TGV ² > TV				ICTV > TGV ² > TV			
L_2^2 -TV	0.988	0.005	0.988	2	42.64	1.14	42.50	2	0.41	0.08	0.43	2
L_2^2 -ICTV	0.988	0.005	0.989	142	42.79	1.18	42.64	148	0.39	0.08	0.41	148
L_2^2 -TGV ²	0.988	0.005	0.989	56	42.76	1.19	42.58	50	0.40	0.08	0.42	50
95% t-test	ICTV > TGV ² > TV				ICTV > TGV ² > TV				ICTV > TGV ² > TV			

Table 3. Regulariser performance with individual learning, L_2^2 and $L_\eta^1 \nabla$ costs and noise variance $\sigma^2 = 2$; BSDS300 dataset, resized.

	SSIM				PSNR				value			
	mean	std	med	best	mean	std	med	best	mean	std	med	best
Noisy data	0.731	0.120	0.744	0	27.72	0.88	28.09	0	1.4E ⁵	2.5E ³	1.4E ⁵	0
$L_\eta^1 \nabla$ -TV	0.898	0.036	0.900	4	31.28	1.63	30.97	8	7.3E ⁴	2.2E ⁴	7.3E ⁴	1
$L_\eta^1 \nabla$ -ICTV	0.906	0.034	0.909	139	31.54	1.68	31.21	142	7.1E ⁴	2.2E ⁴	7.1E ⁴	121
$L_\eta^1 \nabla$ -TGV ²	0.905	0.035	0.907	57	31.47	1.72	31.10	50	7.1E ⁴	2.2E ⁴	7.1E ⁴	78
95% t-test	ICTV > TGV ² > TV				ICTV > TGV ² > TV				ICTV > TGV ² > TV			
L_2^2 -TV	0.897	0.033	0.898	9	31.54	1.76	31.15	2	5.52	1.89	5.51	2
L_2^2 -ICTV	0.903	0.032	0.903	131	31.72	1.76	31.33	148	5.30	1.81	5.35	148
L_2^2 -TGV ²	0.902	0.033	0.903	60	31.67	1.80	31.28	50	5.38	1.87	5.39	50
95% t-test	ICTV > TGV ² > TV				ICTV > TGV ² > TV				ICTV > TGV ² > TV			

Table 4. Regulariser performance with individual learning, L_2^2 and $L_\eta^1 \nabla$ costs and noise variance $\sigma^2 = 10$; BSDS300 dataset, resized.

	SSIM				PSNR				value			
	mean	std	med	best	mean	std	med	best	mean	std	med	best
Noisy data	0.505	0.143	0.516	0	21.80	0.92	22.14	0	2.8E ⁵	7.9E ³	2.8E ⁵	0
$L_\eta^1 \nabla$ -TV	0.795	0.063	0.799	7	27.27	1.64	27.02	11	1.0E ⁵	3.5E ⁴	9.7E ⁴	1
$L_\eta^1 \nabla$ -ICTV	0.810	0.061	0.814	120	27.52	1.66	27.24	125	9.7E ⁴	3.4E ⁴	9.6E ⁴	79
$L_\eta^1 \nabla$ -TGV ²	0.808	0.062	0.814	73	27.50	1.74	27.15	64	9.8E ⁴	3.5E ⁴	9.5E ⁴	120
95% t-test	ICTV > TGV ² > TV				ICTV, TGV ² > TV				ICTV, TGV ² > TV			
L_2^2 -TV	0.802	0.056	0.804	8	27.70	1.93	27.28	0	13.65	5.53	13.14	0
L_2^2 -ICTV	0.811	0.056	0.816	126	27.86	1.91	27.45	138	13.14	5.22	12.62	138
L_2^2 -TGV ²	0.810	0.057	0.814	66	27.83	1.94	27.41	62	13.28	5.38	12.77	62
95% t-test	ICTV > TGV ² > TV				ICTV > TGV ² > TV				ICTV > TGV ² > TV			

Table 5. Regulariser performance with individual learning, L_2^2 and $L_\eta^1 \nabla$ costs and noise variance $\sigma^2 = 20$; BSDS300 dataset, resized.

For the first image of the data set, ICTV does in all of the Figures 7–14 better than TGV², while for the second image, the situation is reversed. We have highlighted these two images for the $L_\eta^1 \nabla$ cost in Figures 15–18, for both noise levels $\sigma = 2$ and $\sigma = 20$. In the case where ICTV does better, hardly any difference can be observed by the eye, while for second image TGV² clearly has less stair-casing in the smooth areas of the image, especially with the noise level $\sigma = 20$.

Based on this study, it therefore seems that ICTV is the most reliable regulariser of the ones tested, when the type of image being processed is unknown, and low SSIM, PSNR or $L_\eta^1 \nabla$ cost functional value is desired. But as can be observed for individual images, it can within large smooth areas exhibit artefacts that are avoided by the use of TGV².

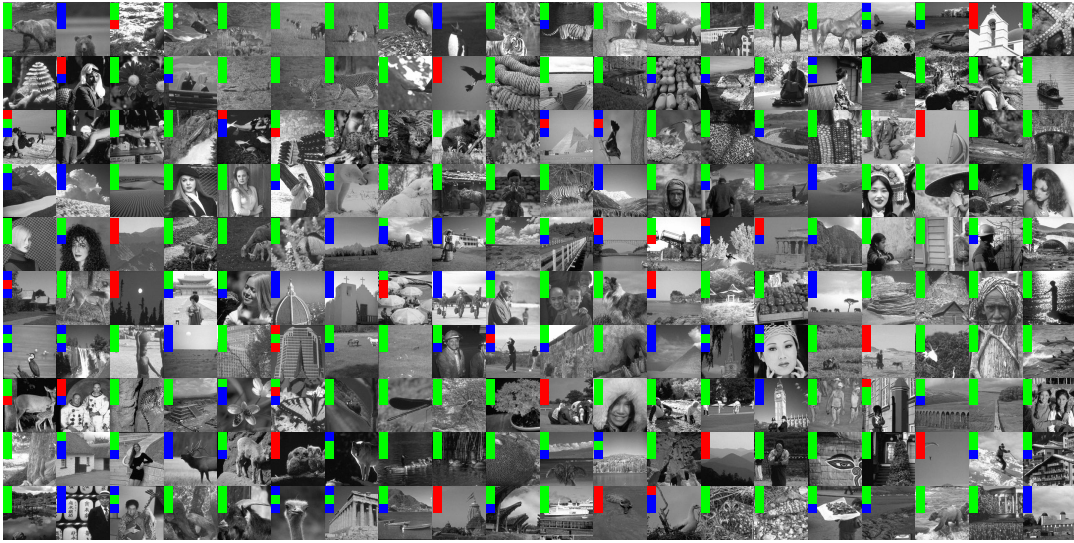


Figure 11. Ordering of regularisers with batch learning, $L_\eta^1 \nabla$ cost, and noise variance $\sigma^2 = 2$, on the 200 images of the BSDS300 dataset, resized. Best regulariser: red=TV, green=ICTV, blue=TGV²; top=SSIM, middle=PSNR, bottom=objective value.



Figure 12. Ordering of regularisers with batch learning, L_2^2 cost, and noise variance $\sigma^2 = 2$, on the 200 images of the BSDS300 dataset, resized. Best regulariser: red=TV, green=ICTV, blue=TGV²; top=SSIM, middle=PSNR, bottom=objective value.

5.3. The choice of cost functional. The L_2^2 cost functional naturally obtains better PSNR than $L_\eta^1 \nabla$, as the two former are equivalent. Comparing the results for the two cost functionals in Tables 3–5, we may however observe that for low noise levels $\sigma^2 = 2, 10$, and generally for batch learning, $L_\eta^1 \nabla$ attains better (higher) SSIM. Since SSIM better captures [37] the visual quality of images than PSNR, this recommends the use of our novel total variation cost functional $L_\eta^1 \nabla$. Of course, one might attempt to optimise the SSIM. This is however a non-convex functional, which will pose additional numerical challenges avoided by the convex total variation cost.

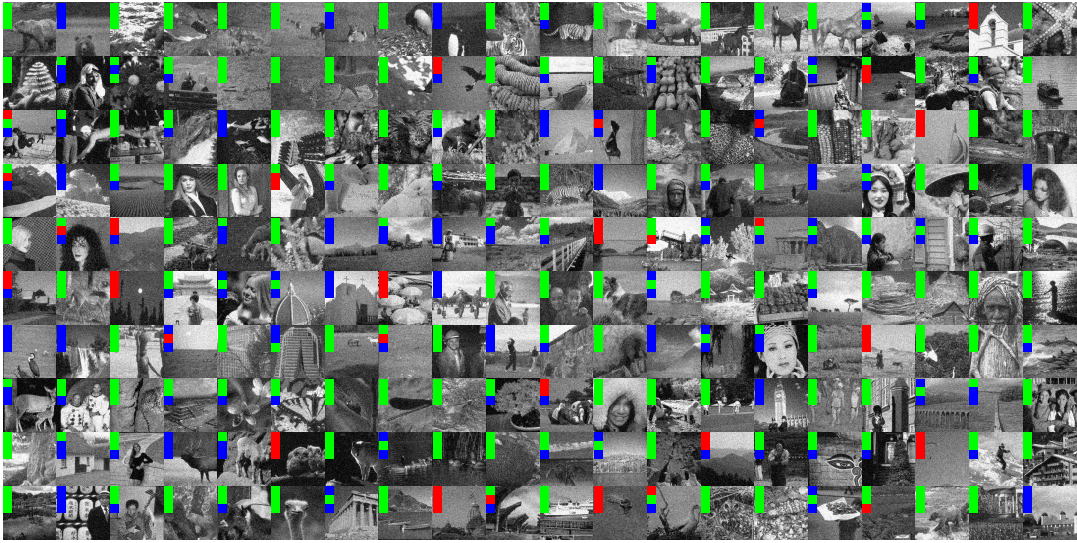


Figure 13. Ordering of regularisers with batch learning, $L^1_\eta \nabla$ cost, and noise variance $\sigma^2 = 20$, on the 200 images of the BSDS300 dataset, resized. Best regulariser: red=TV, green=ICTV, blue=TGV²; top=SSIM, middle=PSNR, bottom=objective value.

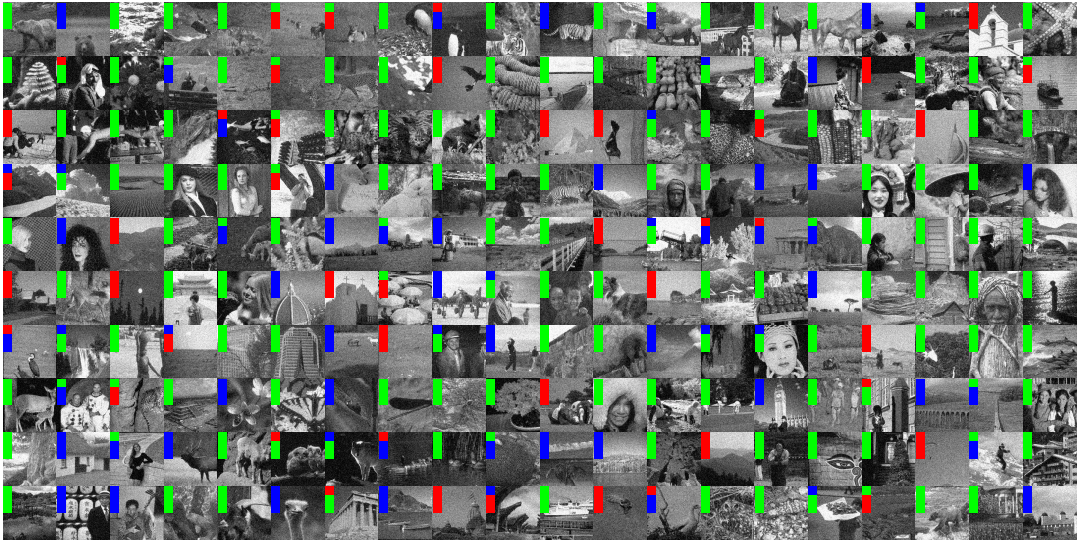


Figure 14. Ordering of regularisers with batch learning, L^2_2 cost, and noise variance $\sigma^2 = 20$, on the 200 images of the BSDS300 dataset, resized. Best regulariser: red=TV, green=ICTV, blue=TGV²; top=SSIM, middle=PSNR, bottom=objective value.

CONCLUSION AND OUTLOOK

In this paper we propose a bilevel optimisation method in function space for learning the optimal choice of parameters in higher-order total variation regularisation. We present a rigorous analysis of this optimisation problem as well as a numerical discussion in the context of image denoising. In particular, we make use of the bilevel learning approach to compare the performance – in terms of returned image quality – of TV, ICTV and TGV regularisation. A statistical analysis, carried out on a dataset of 200 images, suggest that ICTV performs slightly better than TGV, and

	SSIM				PSNR				value			
	mean	std	med	best	mean	std	med	best	mean	std	med	best
Noisy data	0.978	0.015	0.981	16	41.56	0.86	41.95	24	2.9E ⁴	3.1E ²	2.9E ⁴	16
$L_\eta^1 \nabla$ -TV	0.987	0.006	0.988	23	42.43	1.07	42.37	21	2.5E ⁴	3.4E ³	2.5E ⁴	20
$L_\eta^1 \nabla$ -ICTV	0.988	0.006	0.989	119	42.56	1.06	42.51	135	2.4E ⁴	3.5E ³	2.5E ⁴	113
$L_\eta^1 \nabla$ -TGV ²	0.987	0.006	0.989	42	42.51	1.09	42.44	20	2.4E ⁴	3.6E ³	2.5E ⁴	51
95% t-test	ICTV > TGV ² > TV				ICTV > TGV ² > TV				ICTV > TGV ² > TV			
L_2^2 -TV	0.986	0.007	0.987	13	42.46	0.95	42.43	17	0.42	0.07	0.43	17
L_2^2 -ICTV	0.987	0.007	0.988	139	42.57	0.95	42.56	128	0.41	0.07	0.42	128
L_2^2 -TGV ²	0.987	0.007	0.988	38	42.53	0.97	42.51	40	0.41	0.07	0.42	40
95% t-test	ICTV > TGV ² > TV				ICTV > TGV ² > TV				ICTV > TGV ² > TV			

Table 6. Regulariser performance with batch learning, $L_\eta^1 \nabla$ and L_2^2 costs, noise variance $\sigma^2 = 2$; BSDS300 dataset, resized.

	SSIM				PSNR				value			
	mean	std	med	best	mean	std	med	best	mean	std	med	best
Noisy data	0.731	0.120	0.744	8	27.72	0.88	28.09	2	1.4E ⁵	2.5E ³	1.4E ⁵	0
$L_\eta^1 \nabla$ -TV	0.893	0.035	0.897	23	31.24	1.87	30.94	23	7.5E ⁴	2.2E ⁴	7.3E ⁴	18
$L_\eta^1 \nabla$ -ICTV	0.897	0.034	0.902	134	31.36	1.81	31.11	150	7.4E ⁴	2.2E ⁴	7.2E ⁴	107
$L_\eta^1 \nabla$ -TGV ²	0.896	0.035	0.901	35	31.31	1.88	31.01	25	7.4E ⁴	2.3E ⁴	7.2E ⁴	75
95% t-test	ICTV > TGV ² > TV				ICTV > TGV ² > TV				ICTV, TGV ² > TV			
L_2^2 -TV	0.887	0.035	0.889	29	31.31	1.50	31.15	25	5.72	1.91	5.51	25
L_2^2 -ICTV	0.889	0.036	0.893	127	31.41	1.44	31.28	131	5.57	1.83	5.37	131
L_2^2 -TGV ²	0.888	0.035	0.891	44	31.38	1.50	31.20	44	5.64	1.90	5.44	44
95% t-test	ICTV > TGV ² > TV				ICTV > TGV ² > TV				ICTV > TGV ² > TV			

Table 7. Regulariser performance with batch learning, $L_\eta^1 \nabla$ and L_2^2 costs, noise variance $\sigma^2 = 10$; BSDS300 dataset, resized.

	SSIM				PSNR				value			
	mean	std	med	best	mean	std	med	best	mean	std	med	best
Noisy data	0.505	0.143	0.516	4	21.80	0.92	22.14	1	2.8E ⁵	7.9E ³	2.8E ⁵	0
$L_\eta^1 \nabla$ -TV	0.789	0.067	0.798	18	27.37	2.13	26.98	24	1.0E ⁵	3.7E ⁴	9.8E ⁴	14
$L_\eta^1 \nabla$ -ICTV	0.795	0.065	0.804	139	27.46	2.10	27.05	141	1.0E ⁵	3.6E ⁴	9.6E ⁴	91
$L_\eta^1 \nabla$ -TGV ²	0.794	0.066	0.804	39	27.44	2.12	27.04	34	1.0E ⁵	3.7E ⁴	9.6E ⁴	95
95% t-test	ICTV > TGV ² > TV				ICTV > TGV ² > TV				TGV ² > ICTV > TV			
L_2^2 -TV	0.786	0.053	0.790	31	27.50	1.71	27.27	33	14.11	5.78	13.16	33
L_2^2 -ICTV	0.790	0.054	0.790	123	27.56	1.64	27.37	119	13.84	5.54	12.75	119
L_2^2 -TGV ²	0.789	0.053	0.793	46	27.55	1.70	27.33	48	13.93	5.73	12.95	48
95% t-test	ICTV, TGV ² > TV				ICTV, TGV ² > TV				ICTV > TGV ² > TV			

Table 8. Regulariser performance with batch learning, $L_\eta^1 \nabla$ and L_2^2 costs, noise variance $\sigma^2 = 20$; BSDS300 dataset, resized.

both perform better than TV, in average. For denoising of images with a high noise level ICTV and TGV score comparably well. For images with large smooth areas TGV performs better than ICTV.

Moreover, we propose a new cost functional for the bilevel learning problem, which exhibits interesting theoretical properties and has a better behaviour with respect to the PSNR related L^2 cost used previously in the literature. This study raises the question of other, alternative cost functionals. For instance, one could be tempted to use the SSIM as cost, but its non-convexity might present several analytical and numerical difficulties. The new cost functional, proposed in this paper, turns out to be a good compromise between image quality measure and analytically tractable cost term.

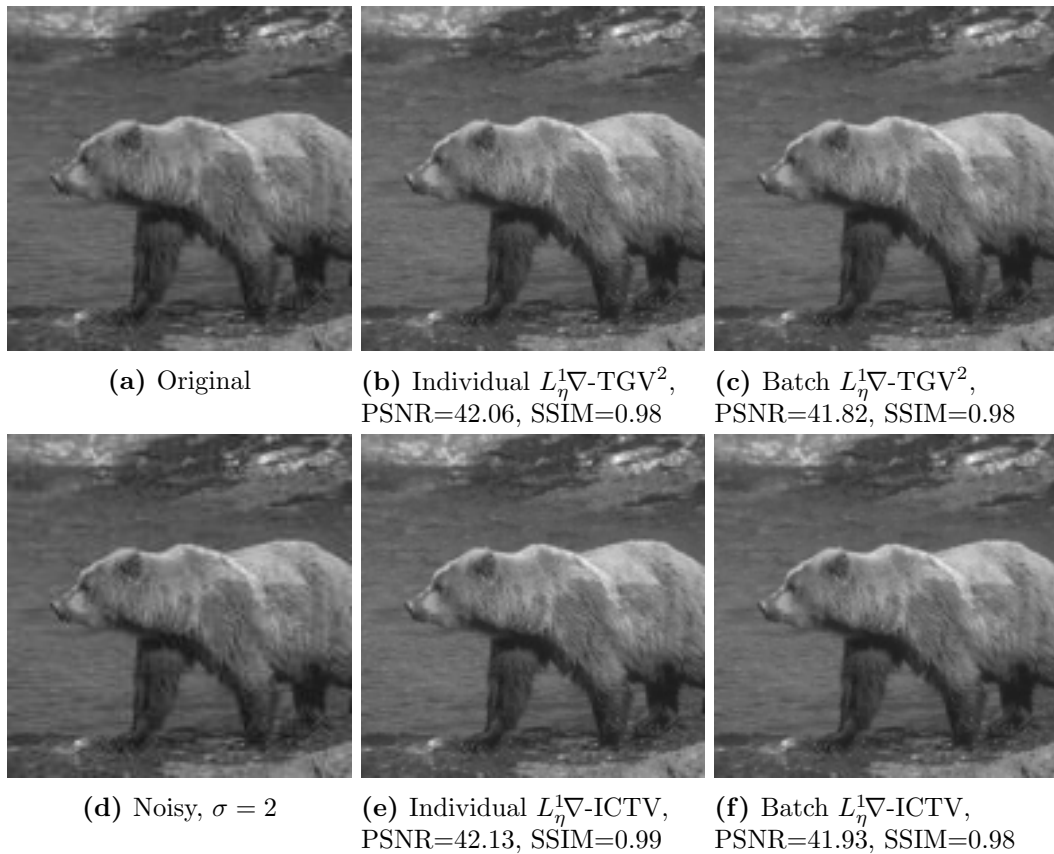


Figure 15. Image for which ICTV performs better than TGV², $\sigma = 2$

ACKNOWLEDGEMENTS

This project has been supported by King Abdullah University of Science and Technology (KAUST) Award No. KUK-I1-007-43, EPSRC grants Nr. EP/J009539/1 and Nr. EP/M00483X/1, the Escuela Politécnica Nacional de Quito under award PIS 12-14 and the MATHAmSud project SOCDE ‘Sparse Optimal Control of Differential Equations’. While in Quito, T. Valkonen has moreover been supported by a Prometeo scholarship of the Senescyt (Ecuadorian Ministry of Science, Technology, Education, and Innovation).

REFERENCES

- [1] Martin Benning, Christoph Brune, Martin Burger, and Jahn Müller. Higher-order TV methods—enhancement via Bregman iteration. *Journal of Scientific Computing*, 54(2-3):269–310, 2013.
- [2] Martin Benning, Lynn Gladden, Daniel Holland, Carola-Bibiane Schönlieb, and Tuomo Valkonen. Phase reconstruction from velocity-encoded MRI measurements – A survey of sparsity-promoting variational approaches. *Journal of Magnetic Resonance*, 238:26–43, 2014.
- [3] Lorenz Biegler, George Biros, Omar Ghattas, Matthias Heinkenschloss, David Keyes, Bani Mallick, Luis Tenorio, Bart van Bloemen Waanders, Karen Willcox, and Youssef Marzouk. *Large-scale inverse problems and quantification of uncertainty*, volume 712. John Wiley & Sons, 2011.
- [4] K. Bredies, K. Kunisch, and T. Pock. Total generalized variation. *SIAM Journal on Imaging Sciences*, 3:492–526, 2011.

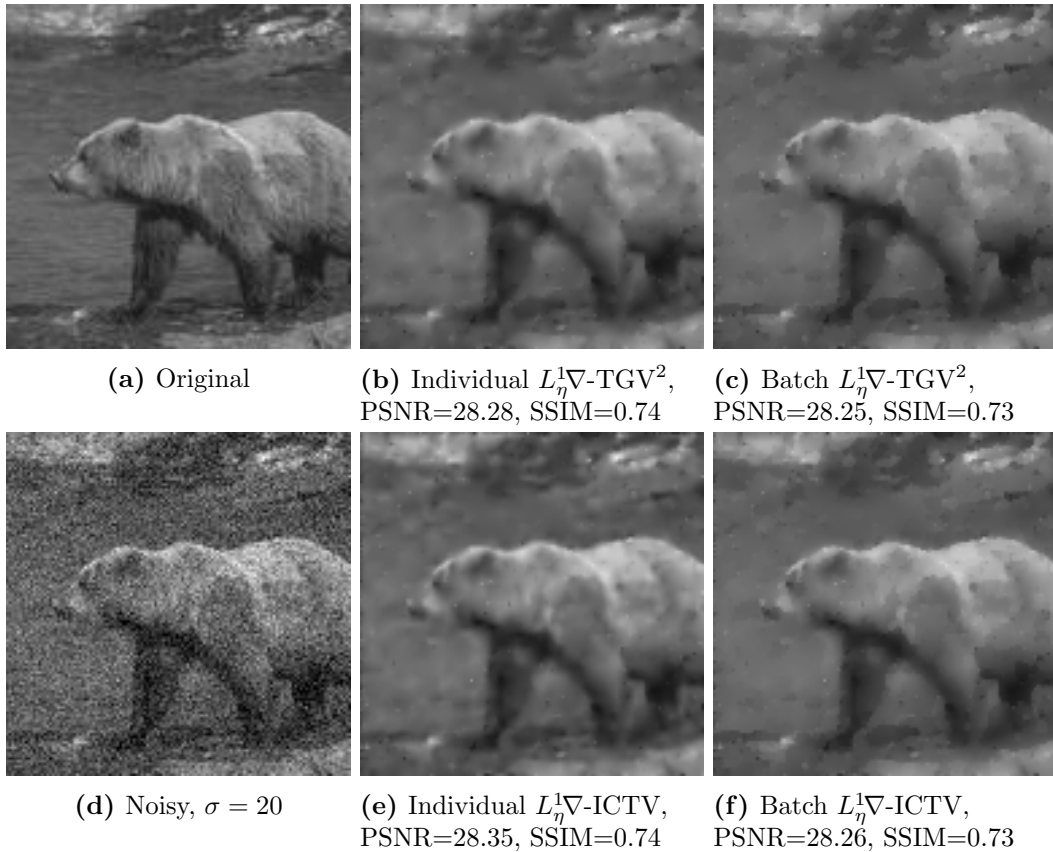


Figure 16. Image for which ICTV performs better than TGV², $\sigma = 20$

- [5] Kristian Bredies and Martin Holler. A total variation-based jpeg decompression model. *SIAM Journal on Imaging Sciences*, 5(1):366–393, 2012.
- [6] Kristian Bredies, Karl Kunisch, and Tuomo Valkonen. Properties of L^1 -TGV²: The one-dimensional case. *Journal of Mathematical Analysis and Applications*, 398:438–454, 2013.
- [7] Kristian Bredies and Tuomo Valkonen. Inverse problems with second-order total generalized variation constraints. In *Proceedings of the 9th International Conference on Sampling Theory and Applications (SampTA) 2011, Singapore*, 2011.
- [8] Tan Bui-Thanh, Karen Willcox, and Omar Ghattas. Model reduction for large-scale systems with high-dimensional parametric input space. *SIAM Journal on Scientific Computation*, 30(6):3270–3288, 2008.
- [9] Luca Calatroni, Juan Carlos De los Reyes, and Carola-Bibiane Schönlieb. Dynamic sampling schemes for optimal noise learning under multiple nonsmooth constraints. In *System Modeling and Optimization*, pages 85–95. Springer Verlag, 2014.
- [10] Antonin Chambolle and Pierre-Louis Lions. Image recovery via total variation minimization and related problems. *Numerische Mathematik*, 76:167–188, 1997.
- [11] T. Chan, A. Marquina, and P. Mulet. High-order total variation-based image restoration. *SIAM Journal on Scientific Computation*, 22(2):503–516, 2000.
- [12] Tony F Chan, Sung Ha Kang, and Jianhong Shen. Euler’s elastica and curvature-based inpainting. *SIAM Journal on Applied Mathematics*, pages 564–592, 2002.
- [13] Yunjin Chen, Thomas Pock, and Horst Bischof. Learning ℓ_1 -based analysis and synthesis sparsity priors using bi-level optimization. In *Workshop on Analysis Operator Learning vs. Dictionary Learning, NIPS 2012*, 2012.
- [14] Yunjin Chen, René Ranftl, and Thomas Pock. Insights into analysis operator learning: From patch-based sparse models to higher-order mrfs. *Image Processing, IEEE Transactions on*, 2014. to appear.

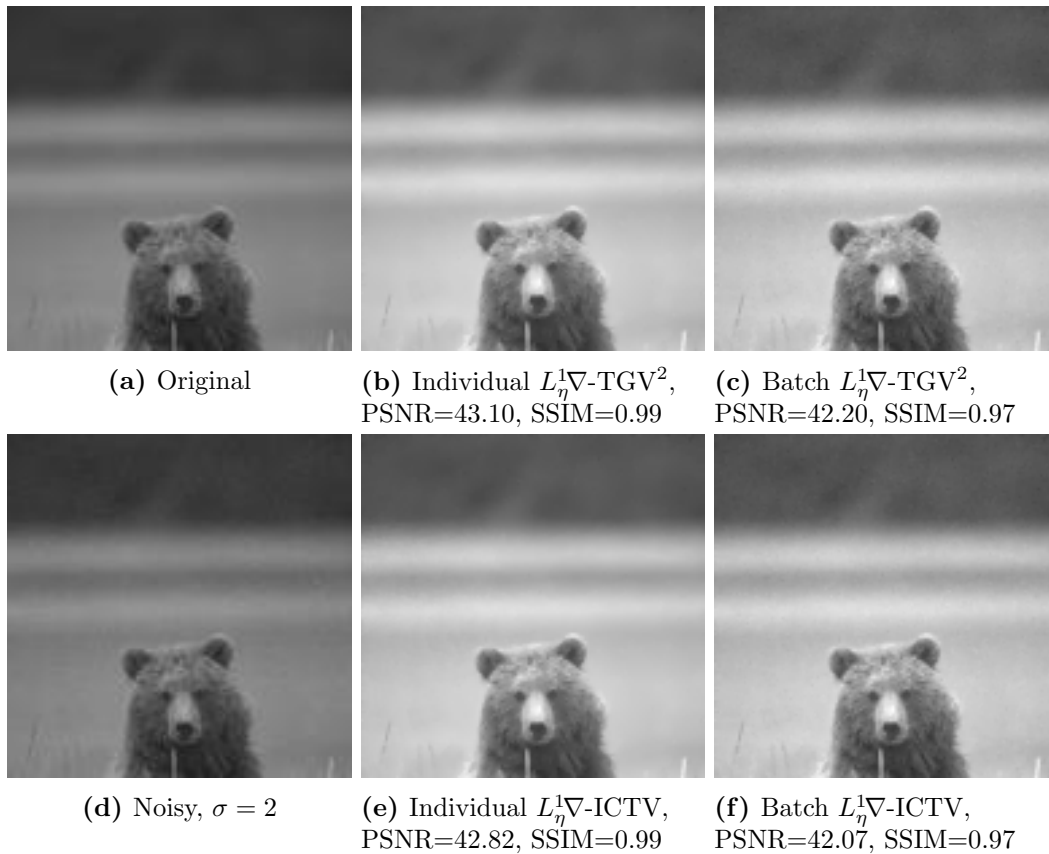


Figure 17. Image for which TGV² performs better than ICTV, $\sigma = 2$

- [15] Julianne Chung, Malena I Español, and Tuan Nguyen. Optimal regularization parameters for general-form tikhonov regularization. *arXiv preprint arXiv:1407.1911*, 2014.
- [16] Juan Carlos De los Reyes and Carola-Bibiane Schönlieb. Image denoising: Learning the noise model via nonsmooth PDE-constrained optimization. *Inverse Problems & Imaging*, 7(4), 2013.
- [17] Juan Carlos de Los Reyes, Carola-Bibiane Schönlieb, and Tuomo Valkonen. The structure of optimal parameters for image restoration problems. Submitted, 2015.
- [18] Justin Domke. Generic methods for optimization-based modeling. In *International Conference on Artificial Intelligence and Statistics*, pages 318–326, 2012.
- [19] K. Gröger. A $W^{1,p}$ -estimate for solutions to mixed boundary value problems for second order elliptic differential equations. *Math. Ann.*, 283(4):679–687, 1989.
- [20] E Haber and L Tenorio. Learning regularization functionals—a supervised training approach. *Inverse Problems*, 19(3):611, 2003.
- [21] Eldad Haber, Lior Horesh, and Luis Tenorio. Numerical methods for the design of large-scale nonlinear discrete ill-posed inverse problems. *Inverse Problems*, 26(2):025002, 2010.
- [22] Walter Hinterberger and Otmar Scherzer. Variational methods on the space of functions of bounded hessian for convexification and denoising. *Computing*, 76(1):109–133, 2006.
- [23] M. Hintermüller and G. Stadler. An infeasible primal-dual algorithm for total bounded variation-based inf-convolution-type image restoration. *SIAM Journal on Scientific Computation*, 28(1):1–23, 2006.
- [24] Michael Hintermüller and Tao Wu. Bilevel optimization for calibrating point spread functions in blind deconvolution. Preprint, 2014.
- [25] Florian Knoll, Kristian Bredies, Thomas Pock, and Rudolf Stollberger. Second order total generalized variation (TGV) for MRI. *Magnetic Resonance in Medicine*, 65(2):480–491, Feb 2011.

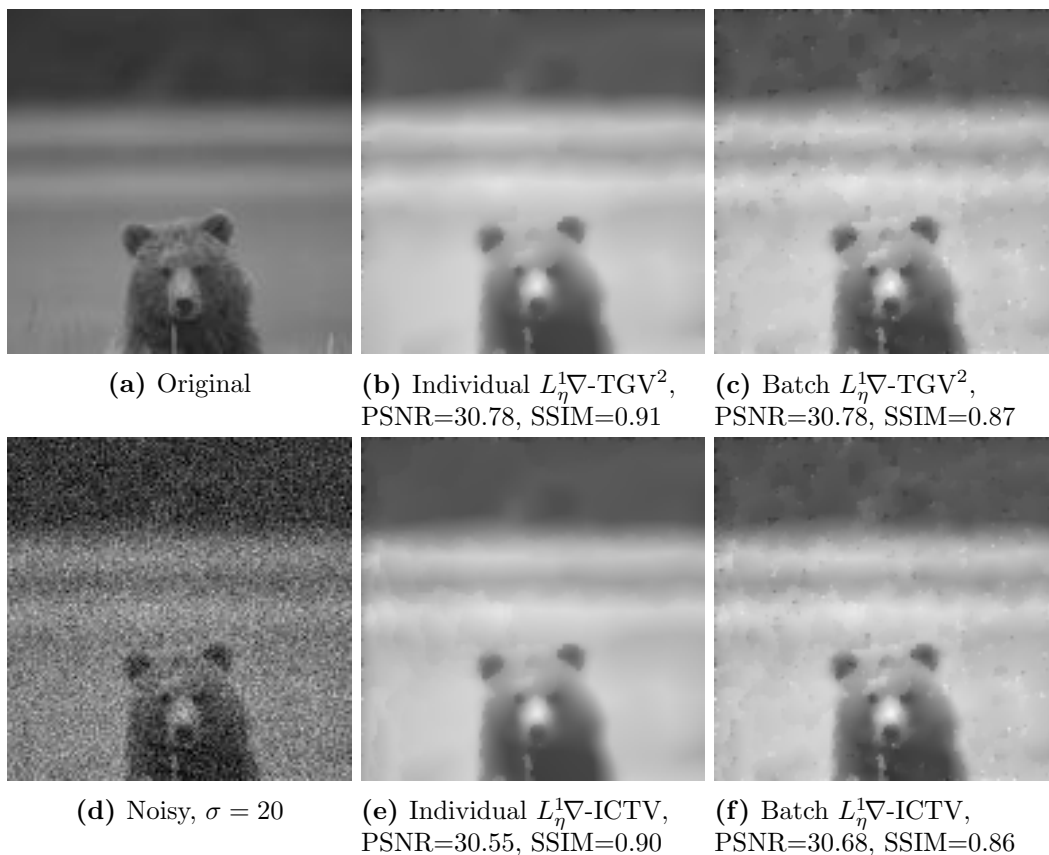


Figure 18. Image for which TGV² performs better than ICTV, $\sigma = 20$

- [26] K. Kunisch and T. Pock. A bilevel optimization approach for parameter learning in variational models. *SIAM Journal on Imaging Sciences*, 6(2):938–983, 2013.
- [27] Marius Lysaker and Xue-Cheng Tai. Iterative image restoration combining total variation minimization and a second-order functional. *International Journal of Computer Vision*, 66(1):5–18, 2006.
- [28] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001.
- [29] Simon Masnou and J-M Morel. Level lines based disocclusion. In *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, pages 259–263. IEEE, 1998.
- [30] K. Papafitsoros and C.-B. Schönlieb. A combined first and second order variational approach for image reconstruction. *Journal of Mathematical Imaging and Vision*, 48(2):308–338, 2014.
- [31] Wolfgang Ring. Structural properties of solutions to total variation regularization problems. *ESAIM: Mathematical Modelling and Numerical Analysis*, 34:799–810, 7 2000.
- [32] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.
- [33] D. Sun and J. Han. Newton and Quasi-Newton methods for a class of nonsmooth equations and related problems. *SIAM Journal on Optimization*, 7(2):463–480, 1997.
- [34] Marshall F Tappen. Utilizing variational optimization to learn Markov random fields. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.
- [35] Tuomo Valkonen, Kristian Bredies, and Florian Knoll. Total generalised variation in diffusion tensor imaging. *SIAM Journal on Imaging Sciences*, 6(1):487–525, 2013.

- [36] Fabio Viola, Andrew Fitzgibbon, and Roberto Cipolla. A unifying resolution-independent formulation for early vision. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 494–501. IEEE, 2012.
- [37] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [38] J. Zowe and S. Kurcyusz. Regularity and stability for the mathematical programming problem in Banach spaces. *Appl. Math. Optim.*, 5(1):49–62, 1979.