# Bilevel approaches for learning of variational imaging models

Luca Calatroni
Cambridge Centre for Analysis, University of Cambridge
Wilberforce Road, CB3 0WA, Cambridge, UK
(lc524@cam.ac.uk)

Cao Chung
Research Center on Mathematical Modelling (MODEMAT),
Escuela Politécnica Nacional, Quito, Ecuador
(cao.vanchung@epn.edu.ec)

Juan Carlos De Los Reyes
Research Center on Mathematical Modelling (MODEMAT),
Escuela Politécnica Nacional, Quito, Ecuador
(juan.delosreyes@epn.edu.ec)

Carola-Bibiane Schönlieb
Department of Applied Mathematics and Theoretical Physics (DAMTP)
University of Cambridge, Wilberforce Road, CB3 0WA, Cambridge, UK
(cbs31@cam.ac.uk)

Tuomo Valkonen
Department of Applied Mathematics and Theoretical Physics (DAMTP)
University of Cambridge, Wilberforce Road, CB3 0WA, Cambridge, UK
(tuomo.valkonen@iki.fi)



**Abstract**

We review some recent learning approaches in variational imaging, based on bilevel optimisation, and emphasize the importance of their treatment in function space. The paper covers both analytical and numerical techniques. Analytically, we include results on the existence and structure of minimisers, as well as optimality conditions for their characterisation. Based on this information, Newton type methods are studied for the solution of the problems at hand, combining them with sampling techniques in case of large databases. The computational verification of the developed techniques is extensively documented, covering instances with different type of regularisers, several noise models, spatially dependent weights and large image databases.

**Keywords:** Image denoising, variational methods, bilevel optimisation, supervised learning.

**Classification:** 49J40, 49J21, 49K20, 68U10, 68T05, 90C53, 65K10

## 1 Overview of learning in variational imaging

A myriad of different imaging models and reconstruction methods exist in the literature and their analysis and application is mostly being developed in parallel in different disciplines. The task of image reconstruction from noisy and under-sampled measurements, for instance, has been attempted in engineering and statistics (in particular signal processing) using filters [60, 73, 28] and multi scale analysis [79, 50, 80], in statistical inverse problems using Bayesian inversion and machine learning [37] and in mathematical analysis using variational calculus, PDEs and numerical optimisation [71]. Each one of these methodologies has its advantages and disadvantages, as well as multiple different levels of interpretation and formalism. In this paper we focus on the formalism of variational reconstruction approaches.

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

$$\alpha\, R(u) + d(K(u), f), \tag{1}$$

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 against the need of regularisation. The parameter $\alpha$ 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 $\alpha$, 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.

Several strategies for conceiving optimization problems have been considered. One approach is the a-priori modelling of image priors, forward operator $K$ and distance function $d$. Total variation regularisation, for instance, has been introduced as an image prior in [71] due to its edge-preserving properties. Its reconstruction qualities have subsequently been thoroughly analysed in works of the variational calculus and partial differential equations community, e.g. [2, 20, 1, 7, 11, 6, 65, 81] only to name a few. The forward operator in magnetic resonance imaging (MRI), for instance, can be derived by formalising the physics behind MRI which roughly results in $K = \mathcal{S}\mathcal{F}$ a sampling operator applied to the Fourier transform. Appropriate data fidelity distances $d$ are mostly driven by statistical considerations that model our knowledge of the data distribution [47, 49]. Poisson distributed data, as it appears in photography [29] and emission tomography applications [82], is modelled by the Kullback-Leibler divergence [72], while a normal data distribution, as for Gaussian noise, results in a least squares fit model. In the context of data driven learning approaches we mention statistically grounded methods for optimal model design [42] and marginalization [13, 53], adaptive and multiscale regularization [76, 35, 39], learning in the context of sparse coding and dictionary learning [64, 57, 56, 85, 66], learning image priors using Markov Random fields [70, 77, 34], deriving optimal priors and regularised inversion matrices by analysing their spectrum [26, 40], and many recent approaches that – based on a more or less generic model setup such as (1) – aim to optimise operators (i.e., matrices and expansion) and functions (i.e. distance functions $d$) in a functional variational regularisation approach by bilevel learning from 'examples' [46, 31, 54, 4, 74, 24, 33, 32], among others. All these approaches vary in their philosophy and mathematics. The main dividing line is between model-based derivation of (1) that uses a-priori knowledge on the data and the image, and data-based derivation of (1) that learns the setup of the model from the data itself.

While functional modelling 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) 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). That is, given a set of training images we find a setup of (1) which minimises an a-priori determined cost functional $F$ measuring the performance of (1) with respect to the training set, compare Section 2 for details. The setup of (1) can be optimised for the choice of regularisation $R$ as will be discussed in Section 4, for the data fitting distance $d$ as in Section 5, or for an appropriate forward operator $K$ as in blind image deconvolution [45] for example. In the present article, rather than working on the discrete problem, as is done in standard parameter learning and model optimisation methods, we discuss the optimisation of variational 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 [70, 77, 34, 23], 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 [42, 41], as well as Ghattas et al. [13, 8]. Recently parameter learning in the context of functional variational regularisation models (1) also entered the image processing community with works by the authors [31, 18, 33, 32, 19, 25], Kunisch, Pock and co-workers [54, 22, 24], Chung et al. [27], and others [4, 74]. Interesting recent works also include bilevel learning approaches for image segmentation [67] and learning of support vector machines [51].

Apart from the work of the authors [31, 18, 33, 32, 25, 19], all approaches so far are formulated and optimised in the discrete setting. In what follows, we review modelling, analysis and optimisation of bilevel learning approaches in function space rather than on a discretisation of (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. HD photography produces larger and larger images with a frequently increasing number of megapixels, compare Figure 1. 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 [83]. Moreover, optimisation methods conceived in function space potentially

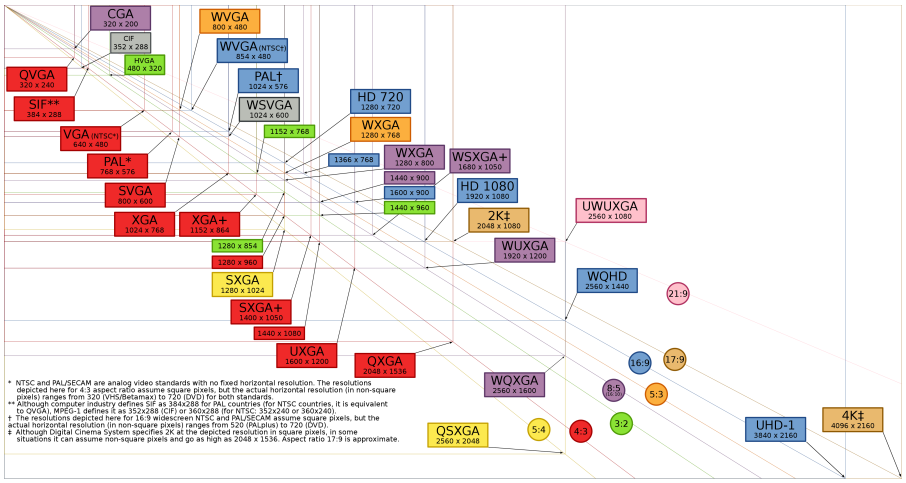



Figure 1: Camera technology tending towards continuum images? Most image processing and analysis algorithms are designed for a finite number of pixels. But camera technology allows to capture images of higher and higher resolution and therefore the number of pixels in images changes constantly. Functional analysis, partial differential equations and continuous optimisation allow us to design image processing models in the continuum.

result in numerical iterative schemes which are resolution and mesh-independent upon discretisation [44].

**Outline of the paper** In what follows we focus on bilevel learning of an optimal variational regularisation model in function space. We give an account on the analysis for a generic learning approach in infinite dimensional function space presented in [33] in Section 2. In particular, we will discuss under which conditions on the learning approach, in particular the regularity of the variational model and the cost functional, we can indeed prove existence of optimal parameters in the interior of the domain (guaranteeing compactness), and derive an optimal system exemplarily for parameter learning for total variation denoising. Section 3 discusses the numerical solution of bilevel learning approaches. Here, we focus on the second-order iterative optimisation methods such as quasi and semismooth Newton approaches [30], which are combined with stochastic (dynamic) sampling strategies for efficiently solving the learning problem even in presence of a large training set [18]. In Sections 4 and 5 we discuss the application of the generic learning model from Section 2 to conceiving optimal regularisation functionals (in the simplest setting this means computing optimal regularisation parameters; in the most complex setting this means computing spatially dependent and vector valued regularisation parameters) [32, 25] and optimal data fidelity functions in presence of different noise distributions [31, 19].

# 2 The learning model and its analysis in function space

## 2.1 The abstract model

Our image domain will be an open bounded set $\Omega \subset \mathbb{R}^n$ with Lipschitz boundary. Our data $f$ lies in $Y = L^2(\Omega; \mathbb{R}^m)$. We look for positive parameters $\lambda = (\lambda_1, \dots, \lambda_M)$ and $\alpha = (\alpha_1, \dots, \alpha_N)$ in abstract parameters sets $\mathcal{P}_\lambda^+$ and $\mathcal{P}_\alpha^+$ They are intended to solve for some convex, proper, weak* lower semicontinuous cost functional $F : X \to \mathbb{R}$ the problem

$$\min_{\alpha \in \mathcal{P}_\alpha^+, \lambda \in \mathcal{P}_\lambda^+} F(u_{\alpha,\lambda}) \quad \text{s.t.} \quad u_{\alpha,\lambda} \in \operatorname*{arg\,min}_{u \in X} J(u; \lambda, \alpha), \tag{P}$$

for

$$J(u; \lambda, \alpha) := \sum_{i=1}^{M} \int_\Omega \lambda_i(x) \phi_i(x, [Ku](x))\, dx + \sum_{j=1}^{N} \int_\Omega \alpha_j(x)\, d|A_j u|(x).$$

Our solution $u$ lies in an abstract space $X$, mapped by the linear operator $K$ to $Y$. Several further technical assumptions discussed in detail in [33] cover $A$, $K$, and the $\phi_i$. In Section 2.2 of this review we concentrate on specific examples of the framework.

For the approximation of problem (P) we consider various smoothing steps. For one, we require Huber regularisation of the Radon norms. Secondly, we take a convex, proper, and weak* lower-semicontinous smoothing functional $H : X \to [0, \infty]$. The typical choice that we concentrate on is $H(u) = \frac{1}{2}\|\nabla u\|^2$.

For parameters $\mu \geq 0$ and $\gamma \in (0, \infty]$, we then consider the problem

$$\min_{\alpha \in \mathcal{P}_\alpha^+,\, \lambda \in \mathcal{P}_\lambda^+} F(u_{\alpha,\lambda,\gamma,\mu}) \quad \text{s.t.} \quad u_{\alpha,\lambda,\gamma,\mu} \in \operatorname*{arg\,min}_{u \in X \cap \operatorname{dom} \mu H} J^{\gamma,\mu}(u; \lambda, \alpha) \tag{$\mathrm{P}^{\gamma,\mu}$}$$

for

$$J^{\gamma,\mu}(u; \lambda, \alpha) := \mu H(u) + \sum_{i=1}^{M} \int_\Omega \lambda_i(x) \phi_i(x, [Ku](x))\, dx + \sum_{j=1}^{N} \int_\Omega \alpha_j(x)\, d|A_j u|_\gamma(x).$$

Here we denote by $|A_j u|_\gamma$ the Huberised total variation measure per the following definition.

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

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

Then if $\nu = f\mathcal{L}^n + \nu^s$ is the Lebesgue decomposition of $\nu \in \mathcal{M}(\Omega; \mathbb{R}^n)$ into the absolutely continuous part $f\mathcal{L}^n$ and the singular part $\nu^s$, we set

$$|\nu|_\gamma(V) := \int_V |f(x)|_\gamma\, dx + |\nu^s|(V), \quad (V \in \mathcal{B}(\Omega)).$$

The measure $|\nu|_\gamma$ is the Huber-regularisation of the total variation measure $|\nu|$.

In all of these, we interpret the choice $\gamma = \infty$ to give back the standard unregularised total variation measure or norm.

## 2.2 Existence and structure: $L^2$-squared cost and fidelity

We now choose

$$F(u) = \frac{1}{2}\|Ku - f_0\|_Y^2, \quad \text{and} \quad \phi_1(x, v) = \frac{1}{2}|f(x) - v|^2, \tag{2}$$

with $M = 1$. We also take $\mathcal{P}_\lambda^+ = \{1\}$, i.e., we do not look for the fidelity weights. Our next results state for specific regularisers with discrete parameters $\alpha = (\alpha_1, \ldots, \alpha_N) \in \mathcal{P}_\alpha^+ = [0, \infty]^N$, conditions for the optimal parameters to satisfy $\alpha > 0$. Observe how we allow infinite parameters, which can in some cases distinguish between different regularisers.

We note that these results are not a mere existence results; they are structural results as well. If we had an additional lower bound $0 < c \leq \alpha$ in (P), we could without the conditions (3) for TV and (4) for $\mathrm{TGV}^2$ [10] denoising, show the existence of an optimal parameter $\alpha$. Also with fixed numerical regularisation ($\gamma < \infty$ and $\mu > 0$), it is not difficult to show the existence of an optimal parameter without the lower bound. What our very natural conditions provide is existence of optimal interior solution $\alpha > 0$ to (P) without any additional box constraints or the numerical regularisation. Moreover, the conditions (3) and (4) guarantee convergence of optimal parameters of the numerically regularised $H^1$ problems ($\mathrm{P}^{\gamma,\mu}$) to a solution of the original $\mathrm{BV}(\Omega)$ problem (P).

**Theorem 2.2** (Total variation Gaussian denoising [33])**.** Suppose $f, f_0 \in \mathrm{BV}(\Omega) \cap L^2(\Omega)$, and

$$\mathrm{TV}(f) > \mathrm{TV}(f_0). \tag{3}$$

Then there exist $\bar{\mu}, \bar{\gamma} > 0$ such that any optimal solution $\alpha_{\gamma,\mu} \in [0, \infty]$ to the problem

$$\min_{\alpha \in [0,\infty]} \frac{1}{2}\|f_0 - u_\alpha\|_{L^2(\Omega)}^2$$

with

$$u_\alpha \in \operatorname*{arg\,min}_{u \in \mathrm{BV}(\Omega)} \Big( \frac{1}{2}\|f - u\|_{L^2(\Omega)}^2 + \alpha |Du|_\gamma(\Omega) + \frac{\mu}{2}\|\nabla v\|_{L^2(\Omega; \mathbb{R}^n)}^2 \Big)$$

satisfies $\alpha_{\gamma,\mu} > 0$ whenever $\mu \in [0, \bar{\mu}]$, $\gamma \in [\bar{\gamma}, \infty]$.

This says that for the optimal parameter to be strictly positive, the noisy image $f$ should, in terms of the total variation, oscillate more than the noise-free image $f_0$ – exactly what we would naturally expect!

*First steps of proof: modelling in the abstract framework.* The modelling of total variation is is based on the choice of $K$ as the embedding of $X = \mathrm{BV}(\Omega) \cap L^2(\Omega)$ into $Y = L^2(\Omega)$, and $A_1 = D$. For the smoothing term we take $H(u) = \frac{1}{2}\|\nabla v\|_{L^2(\Omega; \mathbb{R}^n)}^2$. For the rest of the proof we refer to [33]. □

**Theorem 2.3** (Second-order total generalised variation Gaussian denoising [33])**.** Suppose that the data $f, f_0 \in L^2(\Omega) \cap \mathrm{BV}(\Omega)$ satisfies for some $\alpha_2 > 0$ the condition

$$\mathrm{TGV}^2_{(\alpha_2,1)}(f) > \mathrm{TGV}^2_{(\alpha_2,1)}(f_0). \tag{4}$$

Then there exists $\bar{\mu}, \bar{\gamma} > 0$ such any optimal solution $\alpha_{\gamma,\mu} = ((\alpha_{\gamma,\mu})_1, (\alpha_{\gamma,\mu})_2)$ to the problem

$$\min_{\alpha \in [0,\infty]^2} \frac{1}{2}\|f_0 - v_\alpha\|^2_{L^2(\Omega)}$$

with

$$(v_\alpha, w_\alpha) \in \operatorname*{arg\,min}_{\substack{v \in \mathrm{BV}(\Omega) \\ w \in \mathrm{BD}(\Omega)}} \Big( \frac{1}{2}\|f - v\|^2_{L^2(\Omega)} + \alpha_1 |Dv - w|_\gamma(\Omega) + \alpha_2 |Ew|_\gamma(\Omega) + \frac{\mu}{2}\|(\nabla v, \nabla w)\|^2_{L^2(\Omega;\mathbb{R}^n \times \mathbb{R}^{n\times n})} \Big)$$

satisfies $(\alpha_{\gamma,\mu})_1, (\alpha_{\gamma,\mu})_2 > 0$ whenever $\mu \in [0, \bar{\mu}]$, $\gamma \in [\bar{\gamma}, \infty]$.

Here we recall that $\mathrm{BD}(\Omega)$ is the space of vector fields of bounded deformation [78]. Again, the noisy data has to oscillate more in terms of $\mathrm{TGV}^2$ than the ground-truth does, for the existence of an interior optimal solution to (P). This of course allows us to avoid constraints on $\alpha$.

Observe that we allow for infinite parameters $\alpha$. We do not seek to restrict them to be finite, as this will allow us to decide between $\mathrm{TGV}^2$, TV, and $\mathrm{TV}^2$ regularisation.

*First steps of proof: modelling in the abstract framework.* To present $\mathrm{TGV}^2$ in the abstract framework, we take take $X = (\mathrm{BV}(\Omega) \cap L^2(\Omega)) \times \mathrm{BD}(\Omega)$, and $Y = L^2(\Omega)$. We denote $u = (v, w)$, and set

$$K(v, w) = v, \quad A_1 u = Dv - w, \quad \text{and} \quad A_2 u = Ew$$

for $E$ the symmetrised differential. For the smoothing term we take

$$H(u) = \frac{1}{2}\|(\nabla v, \nabla w)\|^2_{L^2(\Omega;\mathbb{R}^n \times \mathbb{R}^{n\times n})}.$$

For all the gory details we again point the reader to [33]. □

We also have a result on the approximation properties of the numerical models as $\gamma \nearrow \infty$ and $\mu \searrow 0$. Roughly, the the outer semicontinuity [69] of the solution map $\mathcal{S}$ in the next theorem means that as the numerical regularisation vanishes, any optimal parameters for the regularised models ($\mathrm{P}^{\gamma,\mu}$) tend to some optimal parameters of the original model (P).

**Theorem 2.4** ([33])**.** In the setting of Theorem 2.2 and Theorem 2.3, there exist $\bar{\gamma} \in (0, \infty)$ and $\bar{\mu} \in (0, \infty)$ such that the solution map

$$(\gamma, \mu) \mapsto \alpha_{\gamma,\mu}$$

is outer semicontinuous within $[\bar{\gamma}, \infty] \times [0, \bar{\mu}]$.

We refer to [33] for further, more general results of the type in this section. These include analogous of the above ones for a novel Huberised total variation cost functional.

## 2.3 Optimality conditions

In order to compute optimal solutions to the learning problems, a proper characterization of them is required. Since ($\mathrm{P}^{\gamma,\mu}$) constitute PDE-constrained optimisation problems, suitable techniques from this field may be utilized. For the limit cases, an additional asymptotic analysis needs to be performed in order to get a sharp characterization of the solutions as $\gamma \to \infty$ or $\mu \to 0$, or both.

Several instances of the abstract problem ($\mathrm{P}^{\gamma,\mu}$) have been considered in previous contributions. The case with Total Variation regularization was considered in [31] in presence of several noise models. There the Gâteaux differentiability of the solution operator was proved, which lead to the derivation of an optimality system. Thereafter an asymptotic analysis with respect to $\gamma \to \infty$ was carried out (with $\mu > 0$), obtaining an optimality system for the corresponding problem. In that case the optimisation problem corresponds to one with variational inequality constraints and the characterization concerns C-stationary points.

Differentiability properties of higher order regularisation solution operators were also investigated in [32]. A stronger Fréchet differentiability result was proved for the $\mathrm{TGV}^2$ case, which also holds for TV. These stronger results open the door, in particular, to further necessary and sufficient optimality conditions.

For the general problem ($P^{\gamma,\mu}$), using the Lagrangian formalism the following optimality system is obtained:

$$\mu \int_\Omega \langle \nabla u, \nabla v\rangle \, dx + \sum_{i=1}^{M} \int_\Omega \lambda_i \, \phi_i'(Ku) Kv \, dx + \sum_{j=1}^{N} \int_\Omega \alpha_j \langle h_\gamma(A_j u), A_j v\rangle \, dx = 0, \quad \forall v \in V, \quad (5)$$

$$\mu \int_\Omega \langle \nabla p, \nabla v\rangle \, dx + \sum_{i=1}^{M} \int_\Omega \langle \lambda_i \phi_i''(Ku) Kp, Kv\rangle \, dx + \sum_{j=1}^{N} \int_\Omega \alpha_j \langle h_\gamma'^{*}(A_j u) A_j p, A_j v\rangle \, dx = -F'(u)v, \quad \forall v \in V, \quad (6)$$

$$\int_\Omega \phi_i(Ku) Kp(\zeta - \lambda_i) \, dx \geq 0, \quad \forall \zeta \geq 0, \ i = 1, \ldots, M, \quad (7)$$

$$\int_\Omega h_\gamma(A_j u) A_j p(\eta - \alpha_j) \, dx \geq 0, \quad \forall \eta \geq 0, \ j = 1, \ldots, N, \quad (8)$$

where $V$ stands for the Sobolev space where the regularised image lives (typically a subspace of $H^1(\Omega; \mathbb{R}^m)$ with suitable homogeneous boundary conditions), $p \in V$ stands for the adjoint state and $h_\gamma$ is a regularized version of the TV subdifferential, for instance,

$$h_\gamma(z) := \begin{cases} \frac{z}{|z|} & \text{if } \ \gamma|z| - 1 \geq \frac{1}{2\gamma} \\ \frac{z}{|z|}(1 - \frac{\gamma}{2}(1 - \gamma|z| + \frac{1}{2\gamma})^2) & \text{if } \ \gamma|z| - 1 \in (-\frac{1}{2\gamma}, \frac{1}{2\gamma}) \\ \gamma z & \text{if } \ \gamma|z| - 1 \leq -\frac{1}{2\gamma}. \end{cases} \quad (9)$$

This optimality system is stated here formally. Its rigorous derivation has to be justified for each specific combination of spaces, regularisers, noise models and cost functionals.

With help of the adjoint equation (6) also gradient formulas for the reduced cost functional $\mathcal{F}(\lambda, \alpha) := F(u_{\alpha,\lambda}, \lambda, \alpha)$ are derived:

$$(\nabla_\lambda \mathcal{F})_i = \int_\Omega \phi_i(Ku) Kp \, dx, \qquad (\nabla_\alpha \mathcal{F})_j = \int_\Omega h_\gamma(A_j u) A_j p \, dx, \quad (10)$$

for $i = 1, \ldots, M$ and $j = 1, \ldots, N$, respectively. The gradient information is of numerical importance in the design of solution algorithms. In the case of finite dimensional parameters, thanks to the structure of the minimisers reviewed in Section 2, the corresponding variational inequalities (7) and (8) turn into equalities. This has important numerical consequences, since in such cases the gradient formulas (10) may be used without additional projection steps. This will be commented in detail in the next section.

# 3 Numerical optimisation of the learning problem

## 3.1 Adjoint based methods

The derivative information provided through the adjoint equation (6) may be used in the design of efficient second-order algorithms for solving the bilevel problems under consideration. Two main directions may be considered in this context: Solving the original problem via optimisation methods [18, 32, 63], and solving the full optimality system of equations [54, 25]. The main advantage of the first one consists in the reduction of the computational cost when a large image database is considered (this issue will be treated in detail below). When that occurs, the optimality system becomes extremely large, making it difficult to solve it in a manageable amount of time. The advantage of the second approach, on the other hand, consists in the possibility of using efficient (possibly generalized) Newton solvers, which have been intensively developed in the last years.

Let us first describe the quasi-Newton methodology considered in [18, 32] and further developed in [32]. For the design of a quasi-Newton algorithm for the bilevel problem with, e.g., one noise model ($\lambda_1 = 1$), the cost functional has to be considered in reduced form as $\mathcal{F}(\alpha) := F(u_\alpha, \alpha)$, where $u_\alpha$ is implicitly determined by solving the denoising problem

$$u_\alpha = \arg\min_{u \in V} \ \frac{\mu}{2} \int_\Omega \|\nabla u\|^2 \, dx + \sum_{j=1}^{N} \int_\Omega \alpha_j \, d|A_j u|_\gamma + \int_\Omega \phi(u) \, dx, \quad \mu > 0. \quad (11)$$

Using the gradient formula for $\mathcal{F}$,

$$(\nabla \mathcal{F}(\alpha^{(k)}))_j = \int_\Omega h_\gamma(A_j u) A_j p \, dx, \quad j = 1, \dots, N, \tag{12}$$

the BFGS matrix may be updated with the classical scheme

$$B_{k+1} = B_k - \frac{B_k s_k \otimes B_k s_k}{(B_k s_k, s_k)} + \frac{z_k \otimes z_k}{(z_k, s_k)}, \tag{13}$$

where $s_k = \alpha^{(k+1)} - \alpha^{(k)}$, $z_k = \nabla \mathcal{F}(\alpha^{(k+1)}) - \nabla \mathcal{F}(\alpha^{(k)})$ and $(w \otimes v)\varphi := (v, \varphi)w$. For the line search strategy, a backtracking rule may be considered, with the classical Armijo criteria

$$\mathcal{F}(\alpha^{(k)} + t_k d^{(k)}) - \mathcal{F}(\alpha^{(k)}) \leq t_k \beta \nabla \mathcal{F}(\alpha^{(k)})^T d^{(k)}, \quad \beta \in (0, 1], \tag{14}$$

where $d^{(k)}$ stands for the quasi-Newton descent direction and $t_k$ the length of the quasi-Newton step. We consider, in addition, a cyclic update based on curvature verification, i.e., we update the quasi-Newton matrix only if the curvature condition $(z_k, s_k) > 0$ is satisfied. The positivity of the parameter values is usually preserved along the iterations, making a projection step superfluous in practice. In more involved problems, like the ones with $\mathrm{TGV}^2$ or ICTV denoising, an extra criteria may be added to the Armijo rule, guaranteeing the positivity of the parameters in each iteration. Experiments with other line search rules (like Wolfe) have also been performed. Although these line search strategies automatically guarantee the satisfaction of the curvature condition (see, e.g., [62]), the interval where the parameter $t_k$ has to be chosen appears to be quite small and is typically missing.

The denoising problems (11) may be solved either by efficient first- or second-order methods. In previous works we considered primal-dual Newton type algorithms (either classical or semismooth) for this purpose. Specifically, by introducing the dual variables $q_i$, $i = 1, \dots, N$, a necessary and sufficient condition for the lower level is given by

$$\mu \int_\Omega \langle \nabla u, \nabla v \rangle \, dx + \sum_{i=1}^N \int_\Omega \langle q_i, A_i v \rangle \, dx + \int_\Omega \langle \phi'(u), v \rangle \, dx = 0, \qquad \forall v \in V, \tag{15}$$

$$q_i = \alpha_i \, h_\gamma(A_i u) \quad \text{a.e. in } \Omega, \ i = 1, \dots, N, \tag{16}$$

where $h_\gamma(z) := \frac{z}{\max(1/\gamma, |z|)}$ is a regularized version of the TV subdifferential, and the generalized Newton step has the following Jacobi matrix

$$\begin{pmatrix} L + \phi''(u) & A_1^* & \dots & A_N^* \\ -\alpha_1 \left[ \mathfrak{N}(A_1 u) - \chi_1 \frac{A_1 u \otimes A_1 u}{|A_1 u|^3} \right] A_1 & I & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ -\alpha_N \left[ \mathfrak{N}(A_N u) - \chi_N \frac{A_N u \otimes A_N u}{|A_N u|^3} \right] A_N & 0 & 0 & I \end{pmatrix} \tag{17}$$

where $L$ is an elliptic operator, $\chi_i(x)$ is the indicator function of the set $\{x : \gamma |A_i u| > 1\}$ and $\mathfrak{N}(A_i u) := \frac{\min(1, \gamma |A_i u|)}{|A_i u|}$, for $i = 1, \dots, N$. In practice, the convergence neighbourhood of the classical method is too small and some sort of globalization is required. Following [44] a modification of the matrix was systematically considered, where the term $\frac{A_i u \otimes A_i u}{|A_i u|^3}$ is replaced by $\frac{q_i}{\max(|q_i|, \alpha_i)} \otimes \frac{A_i u}{|A_i u|^2}$. The resulting algorithm exhibits both a global and a local superlinear convergent behaviour.

For the coupled BFGS algorithm a warm start of the denoising Newton methods was considered, using the image computed in the previous quasi-Newton iteration as initialization for the lower level problem algorithm. The adjoint equations, used for the evaluation of the gradient of the reduced cost functional, are solved by means of sparse linear solvers.

Alternatively, as mentioned previously, the optimality system may be solved at once using nonlinear solvers. In this case the solution is only a stationary point, which has to be verified a-posteriori to be a minimum of the cost functional. This approach has been considered in [54] and [25] for the finite- and infinite-dimensional cases, respectively. The solution of the optimality system also presents some challenges due to the nonsmoothness of the regularisers and the positivity constraints.

For simplicity, consider the bilevel learning problem with the TV-seminorm, a single Gaussian noise model

and a scalar weight $\alpha$. The optimality system for the problems reads as follows

$$\mu \int_\Omega \langle \nabla u, \nabla v\rangle \, dx + \int_\Omega \alpha h_\gamma(\nabla u)\nabla v \, dx + \int_\Omega (u-f)v \, dx = 0, \forall v \in V, \tag{18a}$$

$$\mu \int_\Omega \langle \nabla p, \nabla v\rangle \, dx + \int_\Omega \alpha \langle h_\gamma'^*(\nabla u)\nabla p, \nabla v\rangle \, dx + \int_\Omega p\, v \, dx = -F'(u)v, \quad \forall v \in V, \tag{18b}$$

$$\sigma = \int_\Omega \langle h_\gamma(\nabla u), \nabla p\rangle \, dx. \tag{18c}$$

$$\sigma \geq 0, \ \alpha \geq 0, \ \sigma \cdot \alpha = 0. \tag{18d}$$

where $h_\gamma$ is given by, e.g., equation (9). The Newton iteration matrix for this coupled system has the following form:

$$\begin{pmatrix} L + \nabla^* \alpha^{(k)} h_\gamma'(\nabla u^k)\nabla & 0 & \nabla^* h_\gamma(\nabla u^k) \\ \nabla^* \alpha^{(k)} h_\gamma''(\nabla u^k)\nabla p \nabla + F''(u^k) & L + \nabla^* \alpha^{(k)} h_\gamma'(\nabla u^k)\nabla & \nabla^* h_\gamma'(\nabla u^k)\nabla p \\ h_\gamma'(\nabla u^k)\nabla p \nabla & h_\gamma(\nabla u^k)\nabla & 0 \end{pmatrix}.$$

The structure of this matrix leads to similar difficulties as for the denoising Newton iterations described above. To fix this and get good convergence properties, Kunisch and Pock [54] proposed an additional feasibility step, where the iterates are projected on the nonlinear constraining manifold. In [25], similarly as for the lower level problem treatment, modified Jacobi matrices are built by replacing the terms $h_\gamma'(u_k)$ in the diagonal, using projections of the dual multipliers. Both approaches lead to globally convergent algorithm with locally superlinear convergence rates. Also domain decomposition techniques were tested in [25] for the efficient numerical solution of the problem.

By using this optimize-then-discretize framework, resolution independent solution algorithms may be obtained. Once the iteration steps are well specified, both strategies outlined above use a suitable discretization of the image. Typically a finite differences scheme with mesh size step $h > 0$ is used for this purpose. The minimum possible value of $h$ is related to the resolution of the image. For the discretization of the Laplace operator the usual five point stencil is used, while forward and backward finite differences are considered for the discretization of the divergence and gradient operators, respectively. Alternative discretization methods (finite elements, finite volumes, etc) may be considered as well, with the corresponding operators.

## 3.2 Dynamic sampling

For a robust and realistic learning of the optimal parameters, ideally, a rich database of $K$ images, $K \gg 1$ should be considered (like, for instance, MRI applications, compare Section 5.1). Numerically, this consists in solving a large set of nonsmooth PDE-constraints of the form (15)- (16) in each iteration of the BFGS optimisation algorithm (13).

In [18] we extended to our imaging framework a *dynamic sample size* stochastic approximation method proposed by Byrd et al. [16]. The algorithm starts by selecting from the whole dataset a sample $S$ whose size $|S|$ is small compared to the original size $K$. In the following iterations, if the approximation of the optimal parameters computed produces an improvement in the cost functional, then the sample size is kept unchanged and the optimisation process continues selecting in the next iteration a new sample of the same size. Otherwise, if the approximation computed is not a good one, a new, larger, sample size is selected and a new sample $S$ of this new size is used to compute the new step. The key point in this procedure is clearly the rule that checks throughout the progression of the algorithm, whether the approximation we are performing is good enough, i.e. the sample size is big enough, or has to be increased. Because of this systematic check, such sampling strategy is called *dynamic*. Denoting by $u_\alpha^k$ the solution of (15)-(16) and by $f_0^k$ the ground-truth images for every $k = 1, \ldots, K$, we consider now the reduced cost functional $\mathcal{F}(\alpha)$ in correspondence of the whole database

$$\mathcal{F}(\alpha) = \frac{1}{2K} \sum_{k=1}^{K} \|u_\alpha^k - f_0^k\|_{L^2}^2,$$

we consider, for every sample $S \subset \{1, \ldots, K\}$, the batch objective function:

$$\mathcal{F}_S(\alpha) := \frac{1}{2|S|} \sum_{k \in S} \|u_\alpha^k - f_0^k\|_{L^2}^2.$$

As in [16], we formulate in [18] a condition on the batch gradient $\nabla \mathcal{F}_S$ which imposes in every stage of the optimisation that the direction $-\nabla \mathcal{F}_S$ is a descent direction for $\mathcal{F}$ at $\alpha$ if the following condition holds:

$$\|\nabla \mathcal{F}_S(\alpha) - \nabla \mathcal{F}(\alpha)\|_{L^2} \leq \theta \|\nabla \mathcal{F}_S(\alpha)\|_{L^2}, \quad \theta \in [0, 1). \tag{19}$$

**Algorithm 1** Dynamic Sampling BFGS

Initialize: $\alpha_0$, sample $\mathcal{S}_0$ with $|S_0| \ll K$ and model parameter $\theta$, $k = 0$.
**while** BFGS not converging, $k \geq 0$
  sample $|S_k|$ PDE constraints to solve;
  update the BFGS matrix;
  compute direction $d_k$ by BFGS and steplength $t_k$ by Armijo cond. (14);
  define new iterate: $\alpha_{k+1} = \alpha_k + t_k d_k$;
  **if variance condition** is satisfied then
    maintain the sample size: $|S_{k+1}| = |S_k|$;
  **else** augment $S_k$ such that condition **variance condition** is verified.
**end**

(a) original (b) noisy (c) $R(u) = \|\nabla u\|_2^2$ (d) $R(u) = |Du|(\Omega)$

Figure 2: The effect of the choice of regularisation in (1): Choosing the $L^2$ norm squared of the gradient of $u$ as a regulariser imposes isotropic smoothing on the image and smoothes the noise equally as blurring the edges. Choosing the total variation (TV) as a regulariser we are able to eliminate the noise while preserving the main edges in the image.

The computation of $\nabla\mathcal{F}$ may be very expensive for applications involving large databases and nonlinear constraints, so we rewrite (19) as an estimate of the variance of the random vector $\nabla\mathcal{F}_S(\alpha)$. We do not report here the details of the derivation of such estimate, but we refer the interested reader to [18, Section 2]. Here, we just underline that through such a condition on the variance one can control in each iteration of the BFGS optimisation whether the sampling approximation is accurate enough and, if this is not the case, a new larger sample size may be determined in order to reach the desired level of accuracy, depending on the parameter $\theta$ in (19).

In order to improve upon the traditional slow convergence drawback of steepest descent methods, we combined the Dynamic Sampling strategy described above with BFGS method (13), as described in Algorithm 1.

# 4 Learning the image model

One of the main aspects of discussion in the modelling of variational image reconstruction is the type and strength of regularisation that should be imposed on the image. That is, what is the correct choice of regularity that should be imposed on an image and how much smoothing is needed in order to counteract imperfections in the data such as noise, blur or undersampling. In our variational reconstruction approach (1) this boils down to the question of choosing the regulariser $R(u)$ for the image function $u$ and the regularisation parameter $\alpha$. In this section we will demonstrate how functional modelling and data learning can be combined to derive optimal regularisation models. To do so, we focus on **T**otal **V**ariation (TV) type regularisation approaches and their optimal setup. The following discussion constitutes the essence of our derivations in [32], including an extended numerical discussion with an interesting application of our approach to cartoon-texture decomposition.

## 4.1 Total variation type regularisation

The TV is the total variation measure of the distributional derivative of $u$ [3], that is for $u$ defined on $\Omega$

$$TV(u) = |Du|(\Omega) = \int_\Omega d|Du|. \tag{20}$$

As the seminal work of Rudin, Osher and Fatemi [71] and many more contributions in the image processing community have proven, a non-smooth first-order regularisation procedure as TV results in a nonlinear

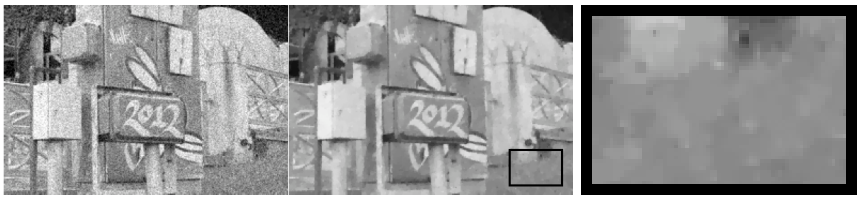


Figure 3: TV image denoising and the staircasing effect: (l.) noisy image, (m.) denoised image, (r.) detail of the bottom right hand corner of the denoised image to visualise the staircasing effect (the creation of blocky-like patterns due to the first-order regulariser).

smoothing of the image, smoothing more in homogeneous areas of the image domain and preserving characteristic structures such as edges, compare Figure **??**. More precisely, when TV is chosen as a regulariser in (1) the reconstructed image is a function in $BV$ the space of functions of bounded variation, allowing the image to be discontinuous as its derivative is defined in the distributional sense only. Since edges are discontinuities in the image function they can be represented by a BV regular image. In particular, the TV regulariser is tuned towards the preservation of edges and performs very well if the reconstructed image is piecewise constant.

Because one of the main characteristics of images are edges as they define divisions between objects in a scene, the preservation of edges seems like a very good idea and a favourable feature of TV regularisation. 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 [68], compare Figure 3.

One possibility to counteract such artefacts is the introduction of higher-order derivatives in the image regularisation. Here, we mainly concentrate on two second-order total variation models: the recently proposed **T**otal **G**eneralized **V**ariation (TGV) [10] and the **I**nfimal-**C**onvolution **T**otal **V**ariation (ICTV) model of Chambolle and Lions [21]. We focus on second-order TV regularisation only since this is the one which seems to be most relevant in imaging applications [52, 9]. For $\Omega \subset \mathbb{R}^2$ open and bounded, the ICTV regulariser reads

$$\mathrm{ICTV}_{\alpha,\beta}(u) := \min_{v \in W^{1,1}(\Omega),\ \nabla v \in BV(\Omega)} \alpha \|Du - \nabla v\|_{\mathcal{M}(\Omega;\mathbb{R}^2)} + \beta \|D\nabla v\|_{\mathcal{M}(\Omega;\mathbb{R}^{2\times 2})}. \tag{21}$$

On the other hand, second-order TGV [12, 11] reads

$$\mathrm{TGV}^2_{\alpha,\beta}(u) := \min_{w \in BD(\Omega)} \alpha \|Du - w\|_{\mathcal{M}(\Omega;\mathbb{R}^2)} + \beta \|Ew\|_{\mathcal{M}(\Omega;\mathrm{Sym}^2(\mathbb{R}^2))}. \tag{22}$$

Here $\mathrm{BD}(\Omega) := \{w \in L^1(\Omega;\mathbb{R}^n) \mid \|Ew\|_{\mathcal{M}(\Omega;\mathbb{R}^{n\times n})} < \infty\}$ is the space of vector fields of bounded deformation on $\Omega$, $E$ denotes the *symmetrised gradient* and $\mathrm{Sym}^2(\mathbb{R}^2)$ the space of symmetric tensors of order 2 with arguments in $\mathbb{R}^2$. The parameters $\alpha, \beta$ are fixed positive parameters. The main difference between (21) and (22) 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 [6]. Substituting $\alpha R(u)$ in (1) by $\alpha TV(u)$, $\mathrm{TGV}^2_{\alpha,\beta}(u)$ or $\mathrm{ICTV}_{\alpha,\beta}(u)$ gives the TV image reconstruction model, TGV image reconstruction model and the ICTV image reconstruction model, respectively.

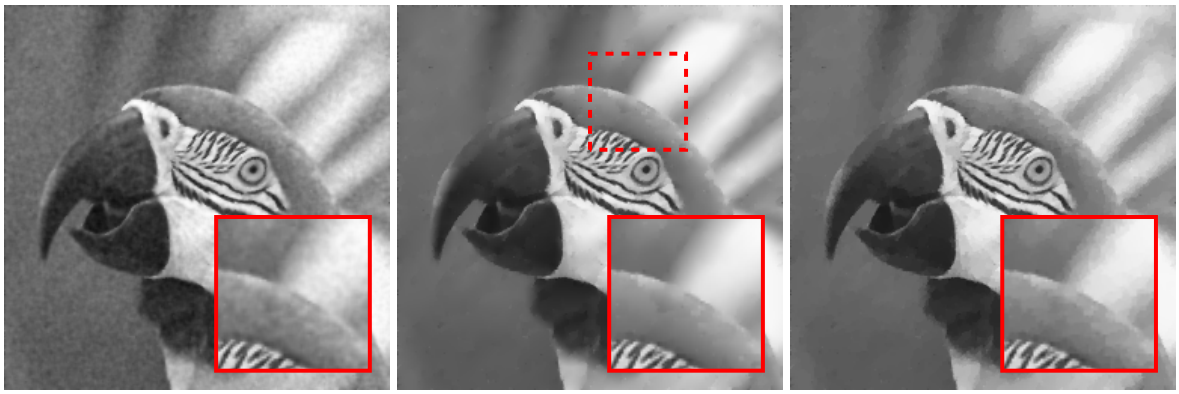
## 4.2 Optimal parameter choice for TV type regularisation

The regularisation effect of TV and second-order TV approaches as discussed above heavily depends on the choice of the regularisation parameters $\alpha$ (i.e. $(\alpha, \beta)$ for second-order TV approaches). In Figures 4 and 5 we show the effect of different choices of $\alpha$ and $\beta$ in $\mathrm{TGV}^2$ denoising. In what follows we show some results from [32] applying the learning approach ($\mathrm{P}^{\gamma,\mu}$) to find optimal parameters in TV type reconstruction models, as well as a new application of bilevel learning to optimal cartoon-texture decomposition.

**Optimal TV, $\mathrm{TGV}^2$ and $ICTV$ denoising** We focus on the special case of $K = Id$ and $L^2$-squared cost $F$ and fidelity term $\Phi$ as introduced in Section 2.2. In [33, 32] we also discuss the analysis and the effect of Huber regularised $L^1$ costs, but this is beyond the scope of this paper and we refer the reader to the respective papers. We consider the problem for finding optimal parameters $(\alpha, \beta)$ for the variational regularisation model

$$u_{(\alpha,\beta)} \in \arg\min_{u \in X} R_{(\alpha,\beta)}(u) + \|u - f\|^2_{L^2(\Omega)},$$

where $f$ is the noisy image, $R_{(\alpha,\beta)}$ is either TV in (20) multiplied by $\alpha$ (then $\beta$ is obsolete), $\mathrm{TGV}^2_{(\alpha,\beta)}$ in (22) or $ICTV_{(\alpha,\beta)}$ in (21). We employ the framework of ($\mathrm{P}^{\gamma,\mu}$) with a training pair $(f_0, f)$ of original image $f_0$ and



(a) Too low $\beta$ / High oscillation (b) Optimal $\beta$ (c) Too high $\beta$ / almost TV

Figure 4: Effect of $\beta$ on $\mathrm{TGV}^2$ denoising with optimal $\alpha$

(a) Too low $\alpha$, low $\beta$.
Good match to noisy data

(b) Too low $\alpha$, optimal $\beta$.
optimal $TV^2$-like behaviour

(c) Too high $\alpha$, high $\beta$.
Bad $\mathrm{TV}^2$-like behaviour

Figure 5: Effect of choosing $\alpha$ too large in $\mathrm{TGV}^2$ denoising

noisy image $f$, using $L^2$-squared cost $F_{L^2_2}(v) := \frac{1}{2}\|f_0 - v\|^2_{L^2(\Omega;\mathbb{R}^d)}$. As a first example we consider a photograph of a parrot to which we add Gaussian noise such that the PSNR of the parrot image is 24.72. In Figure 6, we plot by the red star the discovered regularisation parameter $(\alpha^*, \beta^*)$ reported in Figure 7. Studying the location of the red star, we may conclude that the algorithm managed to find a nearly optimal parameter in very few BFGS iterations, compare Table 1.

**Optimizing cartoon-texture decomposition using a sketch** It is not possible to distinguish noise from texture by the $G$-norm and related approaches [58]. Therefore, learning an optimal cartoon-texture decomposition based on a noise image and a ground-truth image is not feasible. What we did instead, is to make a hand-drawn sketch as our expected "cartoon" $f_0$, and then use the bi-level framework to find the true "cartoon" and "texture" as split by the model

$$J(u, v; \alpha) = \frac{1}{2}\|f - u - v\|^2 + \alpha_1\|v\|_{\mathrm{KR}} + \alpha_2 \mathrm{TV}(u)$$

for the Kantorovich-Rubinstein norm of [55]. For comparison we also include basic TV regularisation results, where we define $v = f - u$. The results for two different iages are in Figure 8 and Table 2, and Figure 9 and Table 3, respectively.

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

| Denoise | Cost | Initial $(\alpha, \beta)$ | Result $(\alpha^*, \beta^*)$ | Cost | SSIM | PSNR | Its. | Fig. |
|---|---|---|---|---|---|---|---|---|
| $\mathrm{TGV}^2$ | $L^2_2$ | $(\alpha^*_{\mathrm{TV}}/\ell, \alpha^*_{\mathrm{TV}})$ | $(0.058/\ell^2, 0.041/\ell)$ | 6.412 | 0.890 | 31.992 | 11 | 7(b) |
| ICTV | $L^2_2$ | $(\alpha^*_{\mathrm{TV}}/\ell, \alpha^*_{\mathrm{TV}})$ | $(0.051/\ell^2, 0.041/\ell)$ | 6.439 | 0.887 | 31.954 | 7 | 7(c) |
| TV | $L^2_2$ | $0.1/\ell$ | $0.042/\ell$ | 6.623 | 0.879 | 31.710 | 12 | 7(a) |

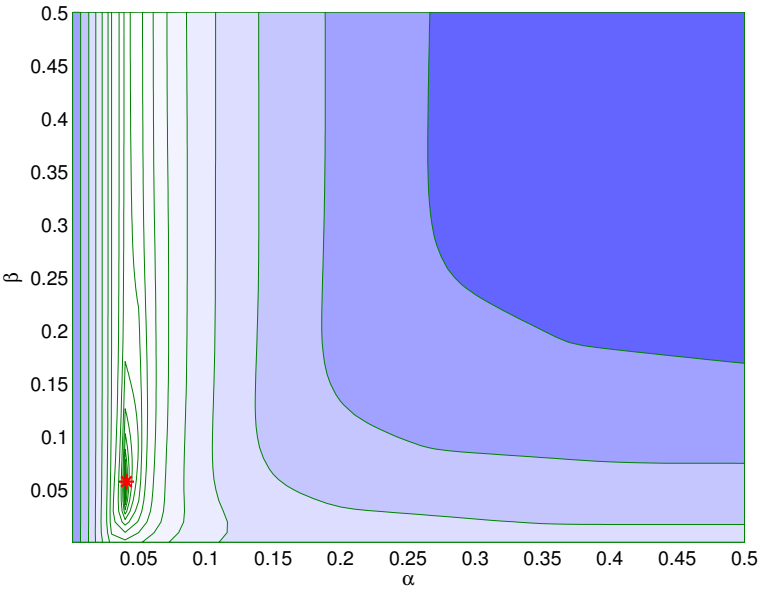



Figure 6: Cost functional value for the $L^2_2$ cost functional plotted versus $(\alpha, \beta)$ for $\mathrm{TGV}^2$ denoising. The illustration is a contour plot of function value versus $(\alpha, \beta)$.

Table 2: Quantified results for cartoon-texture decomposition of the parrot image ($\ell = 256 =$ image width/height in pixels)

| Denoise | Cost | Initial $\vec{\alpha}$ | Result $\vec{\alpha}^*$ | Value | SSIM | PSNR | Its. | Fig. |
|---|---|---|---|---|---|---|---|---|
| KRTV | $L^2_2$ | $(\alpha^*_{\mathrm{TV}}/\ell^{1.5}, \alpha^*_{\mathrm{TV}})$ | $0.006/\ell$ | 81.245 | 0.565 | 9.935 | 11 | 8(f) |
| TV | $L^2_2$ | $0.1/\ell$ | $0.311/\ell$ | 81.794 | 0.546 | 9.876 | 7 | 8(g) |

# 5 Learning the data model

The correct mathematical modelling of the data fidelity terms $\phi_i, i = 1, \ldots, M$ in (P) is crucial for the design of a realisitc denoising model. Their choice corresponds to physical and statistical properties of the noise distribution corrupting the ground-truth $f_0$ and varies significantly depending on applications. Typically, the noise is assumed to be additive, Gaussian-distributed with 0 mean and variance $\sigma^2$ determining the noise intensity. This assumption is reasonable in most of the applications because of the Central Limit Theorem. However, there are cases where this modelling assumption does not correspond to the actual statistical properties characterising the physics of the application considered. For instance, when considering astronomical images, different physical properties corresponding to the quantised (discrete) nature of light and to the independence of photons detection lead to consider a *Poisson* noise distribution, which is signal dependent. Impulse noise seems to be more appropriate for modelling transmission errors affecting only some of the pixels in the image. For those pixels, the intensity value of the signal is switched to either the maximum/minimum value of the dynamic range of the image intensity or to a random value, with positive probability.

For what follows, we will focus on these three noise distributions and on their possible combination. Other distributions can be considered as well: in general, they suit specific applications (like radar or medical ultrasound images) where intrinsically the noise corrupting the image cannot be considered signal-independent.

From a mathematical point of view, variational models reflecting the statistical properties of the noise have been derived for the design of consistent denoising models. Starting from the pioneering work of Rudin, Osher and Fatemi [71], in the case of Gaussian noise a $L^2$-type data fidelity $\phi$ is typically considered. In the case of impulse noise, a variational model based on the use of the $L^1$ norm has been considered in [61]: statistically, this corresponds to consider a Laplace distribution. Poisson noise-based models have been considered in several papers by approximating such distribution with a weighted-Gaussian distribution through variance-stabilising techniques [75, 14]. In [72] a statistically-consistent analytical modelling for Poisson noise distributions has been derived: this results in a Kullback-Leibler-type fidelity.

As a result of different physical factors, very often in applications the presence of different noise distributions

Table 3: Quantified results for cartoon-texture decomposition of the Barbara image ($\ell = 256 =$ image width/height in pixels)

| Denoise | Cost | Initial $\vec{\alpha}$ | Result $\vec{\alpha}^*$ | Value | SSIM | PSNR | Its. | Fig. |
|---|---|---|---|---|---|---|---|---|
| KRTV | $L^2_2$ | $(\alpha^*_{\mathrm{TV}}/\ell, \alpha^*_{\mathrm{TV}})$ | $0.423/\ell$ | 97.291 | 0.551 | 8.369 | 6 | 9(e) |
| TV | $L^2_2$ | $0.1/\ell$ | $0.563/\ell$ | 97.205 | 0.552 | 8.377 | 7 | 9(f) |

(a) Noisy image

(b) $\mathrm{TGV}^2$ denoising, $L^2_2$ cost

(c) ICTV denoising, $L^2_2$ cost

(a) TV denoising, $L^2_2$ cost

Figure 7: Optimal denoising results for initial guess $\vec{\alpha} = (\alpha^*_{\mathrm{TV}}/\ell, \alpha^*_{\mathrm{TV}})$ for $\mathrm{TGV}^2$ and $ICTV$, and $\vec{\alpha} = 0.1/\ell$ for TV

has to be considered as well. In [43] a combined $L^1$-$L^2$ TV-based model is considered for impulse and Gaussian noise removal. A two-phase approach is considered in [17] where the selection of the $L^1/L^2$ term is performed depending on the intensity of the noise. In general, though, the literature on these combined noise models is rather scarse. Gaussian-Poisson noise mixture has been considered in several papers from different point of views: in [48] the exact log-likelihood estimator of the model is derived and then computed via a primal-dual splitting, while in other works (see, e.g., [38]) the discrete-continuous nature of the model (due to the Poisson-Gaussian component, respectively) is approximated by neglecting or modifying one of the two noise models, typically by means of variance-stabilising techniques or a weighted-$L^2$ approximation.

We now proceed differently from Section 2.2 and focus on the modelling of the optimal fidelity terms $\phi_i$ best fitting the acquired data, providing some examples for the single and multiple noise estimation case. In particular, we focus on the estimation of the optimal fidelity weights $\lambda_i, i = 1, \dots, M$ appearing in (P) and ($\mathrm{P}^{\gamma,\mu}$), focusing on the Total-Variation regularisation (20) only applied to denoising problems. Compared to Section 2.1, this corresponds to fix $\mathcal{P}^+_\alpha = \{1\}$ and $K = Id$. We base our presentation on [31, 18], where a careful analysis in term of well-posedness of the problem and derivation of the optimality system in this framework is carried out.

**Shorthand notation** In order not to make the notation too heavy, we warn the reader that we will use a shorthand notation for the quantities appearing in the regularised problem ($\mathrm{P}^{\gamma,\mu}$), that is we will write $\Phi_i(v)$ for the data fidelities $\phi_i(x, v), i = 1 \dots, M$ and $u$ for $u_{\lambda,\gamma,\mu}$, the minimiser of $J^{\gamma,\mu}(\cdot;\lambda)$.

## 5.1 Single noise estimation

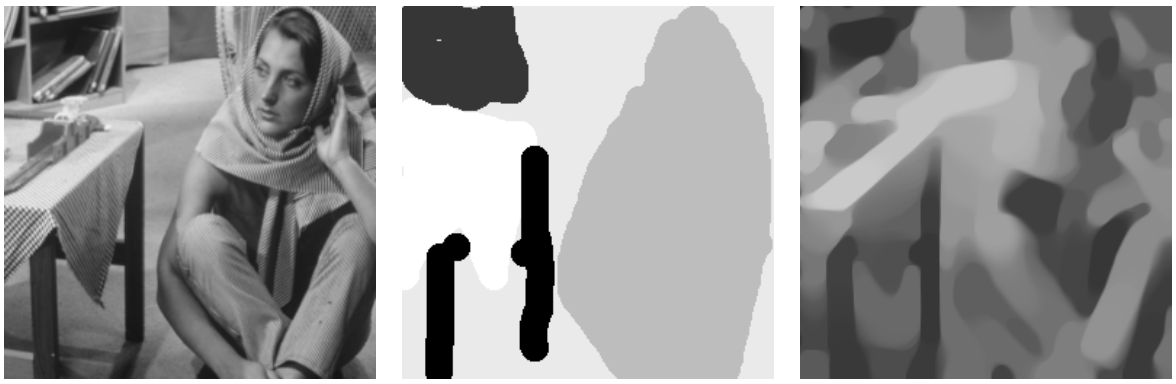
In this section we consider the one-noise distribution case ($M = 1$) where we aim to determine the constant optimal fidelity weight $\lambda$ by solving the following optimisation problem:

$$\min_{\lambda \geq 0} \ \frac{1}{2}\|f_0 - u\|^2_{L^2} \tag{23a}$$

(b) Original image (c) Cartoon sketch (d) KRTV denoising, $L_2^2$ cost

(e) TV denoising, $L_2^2$ cost (f) Texture component for KRTV (g) Texture component for TV

Figure 8: Optimal sketch-based cartoonification for initial guess $\vec{\alpha} = (\alpha_{\mathrm{TV}}^*/\ell^{1.5}, \alpha_{\mathrm{TV}}^*)$ for KRTV and $\vec{\alpha} = 0.1/\ell$ for TV



(a) Original image (b) Cartoon sketch (c) KRTV denoising, $L_2^2$ cost

(d) TV denoising, $L_2^2$ cost (e) Texture component for KRTV (f) Texture component for TV

Figure 9: Optimal sketch-based cartoonification for initial guess $\vec{\alpha} = (\alpha_{\mathrm{TV}}^*/\ell, \alpha_{\mathrm{TV}}^*)$ for KRTV and $\vec{\alpha} = 0.1/\ell$ for TV

subject to (compare (11))

$$\mu\langle\nabla u, \nabla(v-u)\rangle_{L^2} + \lambda \int_\Omega \Phi'(u)(v-u)\,dx + \int_\Omega \|\nabla v\|\,dx - \int_\Omega \|\nabla u\|\,dx \geq 0 \quad \text{for all } v \in H_0^1(\Omega), \quad (23b)$$

where the fidelity term $\Phi$ will change according to the different noise distributions considered and the pair $(f_0, f)$ is the training pair composed by a noise-free and noisy version of the same image, respectively.

Note that in the case the noise level is known there are classical techniques in inverse problems for choosing an optimal parameter $\lambda^*$ in a variational regularisation approach, e.g. the discrepancy principle or the L-curve approach [36]. In our discussion we do not use any knowledge of the noise level but rather extract this information indirectly from our training set and translate it to the optimal choice of $\lambda$. As we will see later such an approach is also naturally extendable to multiple noise models as well as inhomogeneous noise.

**Gaussian noise** We start by considering (23) for determining the regularisation parameter in the standard TV denoising model assuming that the noise in the image is normally distributed. In this case the fidelity term reads $\Phi(u) = |u-f|^2$. The optimisation problem 23 takes the following form:

$$\min_{\lambda \geq 0} \ \frac{1}{2}\|f_0 - u\|_{L^2}^2 \quad (24a)$$

subject to:

$$\mu\langle\nabla u, \nabla(v-u)\rangle_{L^2} + \int_\Omega \lambda(u-f)(v-u)\,dx + \int_\Omega \|\nabla v\|\,dx - \int_\Omega \|\nabla u\|\,dx \geq 0, \forall v \in H_0^1(\Omega). \quad (24b)$$

For the numerical solution of the regularised variational inequality we use a primal-dual algorithm presented in [44].

As an example, we compute the optimal parameter $\lambda^*$ in (24) for a noisy image distorted by Gaussian noise with zero mean and variance 0.02 . Results are reported in Figure 10. The optimisation result has been obtained for the parameter values $\mu = 1e-12, \gamma = 100$ and $h = 1/177$.

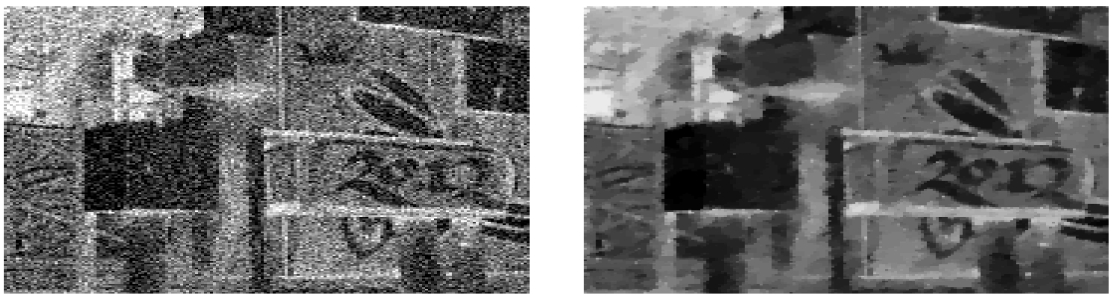
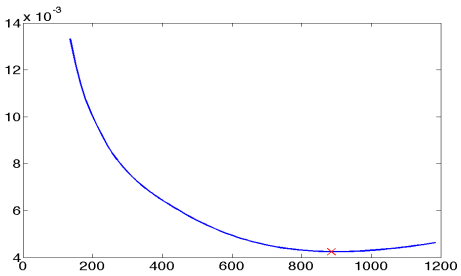


Figure 10: Noisy (left) and optimal denoised (right) image. Noise variance: 0.02. Optimal parameter $\lambda^* = 1770.9$.

In order to check the optimality of the computed regularisation parameter $\lambda^*$, we consider the $80 \times 80$ pixel bottom left corner of the noisy image in Figure 10. In Figure 11 the values of the cost functional and of the **S**ignal to **N**oise **R**atio $SNR = 20 \times \log_{10}\left(\frac{\|f_0\|_{L^2}}{\|u-f_0\|_{L^2}}\right)$, for parameter values between 150 and 1200, are plotted. Also the cost functional value corresponding to the computed optimal parameter $\lambda^* = 885.5$ is shown with a cross. It can be observed that the computed weight actually corresponds to an optimal solution of the bilevel problem. Here we have used $h = 1/80$ and the other parameters as above.

The problem presented consists in the optimal choice of the TV regularisation parameter, if the original image is known in advance. This is a toy example for proof of concept only. In applications, this image would be replaced by a training set of images.



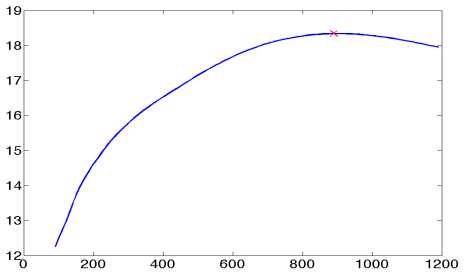



Figure 11: Plot of the cost functional value (left) and the SNR (right) vs. the parameter $\lambda$. Parameters: the input is the $80 \times 80$ pixel crop of the bottom left corner of the noisy image in Figure 10, $h = 1/80$, $\gamma = 100$, $\mu = 1e-12$. The red cross in the plot corresponds to the optimal $\lambda^* = 885.5$.

| $K$ | $\lambda^*$ | $\lambda_S^*$ | $\lvert S_0 \rvert$ | $\lvert S_{end} \rvert$ | eff. | eff. Dyn.S. | BFGS its. | BFGS its. Dyn.S. | diff. |
|---|---|---|---|---|---|---|---|---|---|
| 10 | 3334.5 | 3427.7 | 2 | 3 | 140 | **84** | 7 | 21 | 2.7% |
| 20 | 3437.0 | 3475.1 | 4 | 4 | 240 | **120** | 7 | 15 | 1.1% |
| 30 | 3436.5 | 3478.2 | 6 | 6 | 420 | **180** | 7 | 15 | 1.2% |
| 40 | 3431.5 | 3358.3 | 8 | 9 | 560 | **272** | 7 | 16 | 2.1% |
| 50 | 3425.8 | 3306.4 | 10 | 10 | 700 | **220** | 7 | 11 | 3.5% |
| 60 | 3426.0 | 3543.4 | 12 | 12 | 840 | **264** | 7 | 11 | 3.3% |
| 70 | 3419.7 | 3457.7 | 14 | 14 | 980 | **336** | 7 | 12 | 1.1% |
| 80 | 3418.1 | 3379.3 | 16 | 16 | 1120 | **480** | 7 | 15 | $< 1\%$ |
| 90 | 3416.6 | 3353.5 | 18 | 18 | 1260 | **648** | 7 | 18 | 2.3% |
| 100 | 3413.6 | 3479.0 | 20 | 20 | 1400 | **520** | 7 | 13 | 1.9% |

Table 4: Optimal $\lambda^*$ estimation for large training sets: computational costs are reduced via Dynamic Sampling Algorithm 1.

**Robust estimation with training sets** Gaussian noise images typically arise within the framework of **M**agnetic **R**esonance **I**maging (MRI). The challenge in this case consists in training the TV denoising method such that with one fixed optimally computed $\lambda^*$ clearer images are obtained from noisy acquisitions taken on a single MR tomograph with fixed settings. MR images seem to be a natural choice for our methodology, since a training set of images is often at hand. Let us consider a training database $\left\{(f_0^k, f_k)\right\}_{k=1,\dots,K}, K \gg 1$ of clean and noisy images. We modify (24) as:

$$\min_{\lambda \geq 0} \frac{1}{2K} \sum_{k=1}^{K} \|f_0^k - u_k\|_{L^2}^2 \tag{25}$$

subject to the set of regularised versions of (24b), for $k = 1, \dots, K$.

As explained in [18], dealing with large training sets of images and non-smooth PDE constraints of the form (24b) may result is very high computational costs as, in principle, each constraint needs to be solved in each iteration of the optimisation loop. On the other hand, in MRI applications, a large database of images is desirable in order to make the optimal noise estimation robust. In order to overcome the computational efforts, we estimate $\lambda^*$ using the Dynamic Sampling Algorithm 1.

For the following numerical tests, the parameters are chosen as follows: $\mu = 1e-12$, $\gamma = 100$ and $h = 1/150$. The noise in the images has distribution $\mathcal{N}(0, 0.005)$ and the accuracy parameter $\theta$ of the Algorithm 1, is chosen to be $\theta = 0.5$.

Table 4 shows the numerical values of the optimal parameter $\lambda^*$ and $\lambda_S^*$ computed varying $N$ after solving all the PDE constraints and using Dynamic Sampling algorithm, respectively. We measure the efficiency of the algorithms in terms of the number of the PDEs solved during the whole optimisation and we compare the efficiency of solving (25) subject to the whole set of constraints (24b) with the one where solution is computed by means of the Dynamic Sampling strategy, observing a clear improvement. Computing also the relative error $\|\hat{\lambda}_S - \hat{\lambda}\|_1 / \|\lambda_S|\|_1$ we note a good level of accuracy: the error remains always below 5%.

Figure 12 shows an example of database of brain images[1] together with the optimal denoised version obtained by Algorithm 1 for Gaussian noise estimation.

In order to test the adaptability of our method to images which are very diverse between each other, we test our model for a very diversified database [2], see Fig. 13. From Table 5 we can observe that increasing the size of the database, the estimation of the optimal parameter $\lambda^*$ may vary significantly, due to the diversity of images considered. This reflects the property of our approach to estimate the parameter $\lambda^*$ which is optimal with respect to the *entire* database, cf. cost functional (25).

[1] OASIS online database, `http://www.oasis-brains.org/`.

[2] Berkeley database, available online at: `http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/BSDS300/html/dataset/images.`

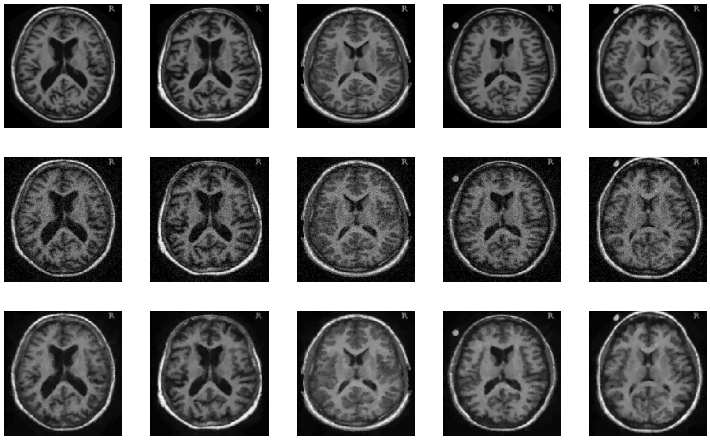


Figure 12: Sample of 5 images of OASIS MRI brain database: original images (upper row), noisy images (middle row) and optimal denoised images (bottom row), $\hat{\lambda_S} = 3280.5$.

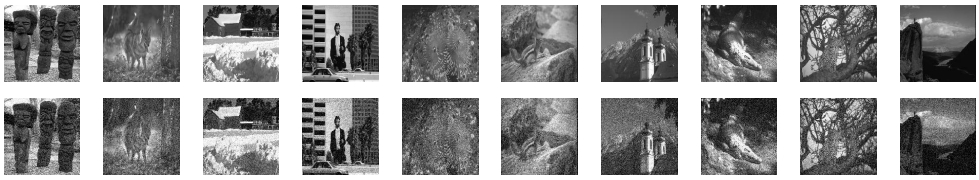


Figure 13: Noise-free and noisy versions of images from the Berkeley database. The Gaussian noise distribution is 0 mean and variance $\sigma^2 = 0.01$.

| $\boldsymbol{K}$ | 10 | 20 | 30 | 40 |
|---|---|---|---|---|
| $\boldsymbol{\lambda^*}$ | 2732.15 | 2766.32 | 2170.23 | 2292.51 |

Table 5: Optimal $\lambda^*$ estimation for heterogeneous database, see Fig. 13. The numerical value adapts to the diversity of the images considered.

**Poisson noise** As a second example, we consider the case of images corrupted by Poisson noise. The corresponding data fidelity in this case has been shown in [72] to be a KL-type fidelity defined as $\Phi(u) = u - f \log u$, which requires the additional condition for $u$ to be strictly positive. We enforce this constraint by using a standard penalty method and solve:

$$\min_{\lambda \geq 0} \ \frac{1}{2}\|f_0 - u\|^2_{L^2}$$

where $u$ is the solution of the minimisation problem:

$$\min_{v>0} \left\{ \frac{\mu}{2}\|\nabla v\|^2_{L^2} + |Dv|(\Omega) + \lambda \int_\Omega (v - f \ \log v) \ dx + \frac{\eta}{2}\|\min(v, \delta)\|^2_{L^2} \right\}, \tag{26}$$

where $\eta \gg 1$ is a penalty parameter enforcing the positivity constraint and $\delta \ll 1$ ensures strict positivity throughout the optimisation. After Huber-regularising the TV term using (2.1), we write the primal-dual form of the corresponding optimality condition for the optimisation problem (26) similarly as in (15)-(16) :

$$-\mu\Delta u - \operatorname{div} q + \lambda \ (1 - \frac{f}{u}) + \eta\chi_{\mathcal{T}_\gamma} \ u = 0, \quad q = \frac{\gamma\nabla u}{\max(\gamma|\nabla u|, 1)}, \tag{27}$$

where $\mathcal{T}_\delta$ is the active set $\mathcal{T}_\delta := \{x \in \Omega : u(x) < \delta\}$. We then design a modified SSN iteration solving (27) similarly as described in Section 3.1, see [31, Section 4] for more details. Figure 14 shows the optimal denoising result for the Poisson noise case in correspondence of the value $\lambda^* = 1013.76$.

**Spatially dependent weight** We continue with an example where $\lambda$ is spatially-dependent. Specifically, we choose as parameter space $V = \{v \in H^1(\Omega) : \partial_n u = 0 \text{ on } \Gamma\}$ in combination with a TV regulariser and a single Gaussian noise model. For this example the noisy image is distorted non-uniformly: A Gaussian noise with

Figure 14: Poisson denoising: Original (left), noisy (center) and optimal denoised (right) images. Parameters: $\gamma = 1e3, \mu = 1e-10, h = 1/128, \eta = 1e4$. Optimal weight: $\lambda^* = 1013.76$.

zero mean and variance 0.04 is present on the whole image and an additional noise with variance 0.06 is added on the area marked by red line.

Since the spatially dependent parameter does not allow to get rid of the positivity constraints in an automatic way, we solved the whole optimality system by means of the semismooth Newton method described in Section 3, combined with a Schwarz domain decomposition method. Specifically, we decomposed the domain first and apply the globalized Newton algorithm in each subdomain afterwards. The detailed numerical performance of this approach is reported in [25].

The results are shown in Figure 15 for the parameters $\mu = 1e-16$, $\gamma = 25$ and $h = 1/500$. The values of $\lambda$ on whole domain are between 100.0 to 400.0. From the right image in Figure 15 we can see the dependence of the optimal parameter $\lambda^*$ on the distribution of noise. As expected, at the high-level noise area in the input image, values of $\lambda^*$ are lower (darker area) than in the rest of the image.

Figure 15: Noisy image (left), denoised image (center) and intensity of $\lambda^*$ (right).

## 5.2 Multiple noise estimation

In many applications, the acquired image may be possibly corrupted by different types of noise, each one corresponding to a different data fidelity term $\Phi_i$ weighted by a non-negative weighting $\lambda_i$. In this multiple noise case, we consider the following optimisation lower level problem:

$$\min_u \left\{ \frac{\mu}{2}\|\nabla u\|_{L^2}^2 + |Du|(\Omega) + \int_\Omega \Psi(\lambda_1, \dots, \lambda_M, \Phi_1(u), \dots, \Phi_M(u))\, dx \right\},$$

where the modelling function $\Psi$ combines the different fidelity terms $\Phi_i$ and weights $\lambda_i$ in order to deal with the multiple noise case. The case when $\Psi$ is a linear a linear combination of fidelities $\Phi_i$ with coefficients $\lambda_i$ is the one presented in the general model (P) and ($\mathrm{P}^{\gamma,\mu}$) and has been considered in [31]. In the following, we present also the case when $\Psi$ is an infimal-convolution operation of fidelities, as considered in [19].

**Impulse and Gaussian noise** Motivated by some previous work in literature on the use of the infimal-convolution operation [5, Chapter 16] for image decomposition, cf. [21, 15], we consider in [19] the modelling of mixed noise distribution through such operation with the intent of obtaining an optimal denoised image thanks to the decomposition of the noise into its different components. In the case of combined Gaussian and impulse noise, the optimisation model reads:

$$\min_{\lambda_1,\lambda_2 \geq 0} \ \frac{1}{2}\|f_0 - u\|_{L^2}^2$$

where $u$ is the solution of the optimisation problem:

$$\min_{\substack{v\in BV\\ n\in L^2}} \left\{ \frac{\mu}{2}\|\nabla v\|_{L^2}^2 + |Dv|(\Omega) + \lambda_1\|n\|_{L^1} + \lambda_2\|f - v - n\|_{L^2}^2 \right\}, \tag{28}$$

where $n$ represents the impulse noise component (and, as such, is treated using the $L^1$ norm) and the optimisation runs over $v$ and $n$. We use once again a single training pair $(f_0, f)$ and consider a Huber-regularisation depending on a parameter $\gamma$ for both the TV term and the $L^1$ norm in (28). The corresponding Euler-Lagrange equations are:

$$\begin{aligned} -\mu\Delta u - \operatorname{div}\left(\frac{\gamma\nabla u}{\max(\gamma|\nabla u|,1)}\right) - \lambda_2(f-u-n) = 0,\\ \lambda_1\frac{\gamma\, n}{\max(\gamma|n|,1)} - \lambda_2(f-u-n) = 0. \end{aligned}$$

Again, writing the equations above in a primal-dual form, we can write the modified SSN iteration and solve the optimisation problem with BFGS as described in Section 3.1.

In Figure 16 we present the results of the model considered. The original image $f_0$ has been corrupted with Gaussian noise of zero mean and variance 0.005 and then a percentage of 5% of pixels has been corrupted with impulse noise. The parameters have been chosen to be $\gamma = 1e3$, $\mu = 1e-15$ and the mesh step size $h = 1/120$. The computed optimal weights are $\lambda_1^* = 351.23$ and $\lambda_2^* = 5200.1$. The results show the actual decomposition of the noise into its sparse and Gaussian components.

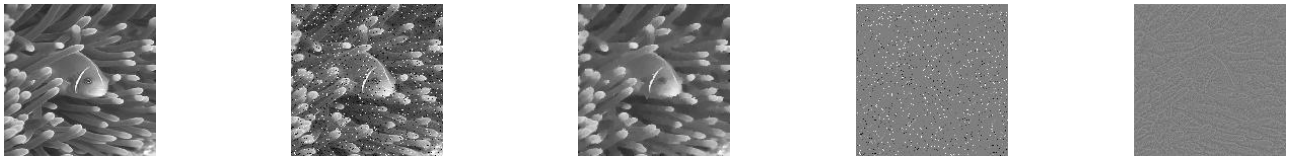


Figure 16: Impulse-Gaussian denoising. From left to right: Original image, noisy image corrupted by impulse noise and Gaussian noise with mean zero and variance 0.005, denoised image, impulse noise residuum and Gaussian noise residuum. Optimal parameters: $\lambda_1^* = 351.23$ and $\lambda_2^* = 5200.1$.

**Gaussian and Poisson noise** We consider now the optimisation problem with $\Phi_1(u) = |u - f|^2$ for the Gaussian noise component and $\Phi_2(u) = (u - f\log u)$ for the Poisson distributed one. We aim to determine the optimal weighting $(\lambda_1, \lambda_2)$ as follows:

$$\min_{\lambda_1,\lambda_2\geq 0} \ \frac{1}{2}\|f_0 - u\|_{L^2}^2$$

subject to $u$ be the solution of:

$$\min_{v>0}\left\{ \frac{\mu}{2}\|\nabla v\|_{L^2}^2 + |Dv|(\Omega) + \frac{\lambda_1}{2}\|v-f\|_{L^2}^2 + \lambda_2\int_\Omega (v - f\log v)\, dx \right\}, \tag{29}$$

for one training pair $(f_0, f)$, where $f$ corrupted by Gaussian and Poisson noise. After Huber-regularising the Total Variation term in (29), we derive (formally) the following Euler-Lagrange equation

$$\begin{aligned} &-\mu\Delta u - \operatorname{div}\left(\frac{\gamma\nabla u}{\max(\gamma|\nabla u|,1)}\right) + \lambda_1(u-f) + \lambda_2(1-\frac{f}{u}) - \alpha = 0\\ &\alpha\cdot u = 0, \end{aligned}$$

with non-negative Lagrange multiplier $\alpha \in L^2(\Omega)$. As in [72] we multiply the first equation by $u$ and obtain

$$u\cdot\left(-\mu\Delta u - \operatorname{div}\left(\frac{\gamma\nabla u}{\max(\gamma|\nabla u|,1)}\right) + \lambda_1(u-f)\right) + \lambda_2(u-f) = 0,$$

where we have used the complementarity condition $\alpha\cdot u = 0$. Next, the solution $u$ is computed iteratively by using a semismooth Newton type method combined with the outer BFGS iteration as above.

In Figure 17 we show the optimisation result. The original image $f_0$ has been first corrupted by Poisson noise and then Gaussian noise was added, with zero mean and variance 0.001. Choosing the parameter values to be $\gamma = 100$ and $\mu = 1e-15$, the optimal weights $\lambda_1^* = 1847.75$ and $\lambda_2^* = 73.45$ were computed on a grid with mesh size step $h = 1/200$.

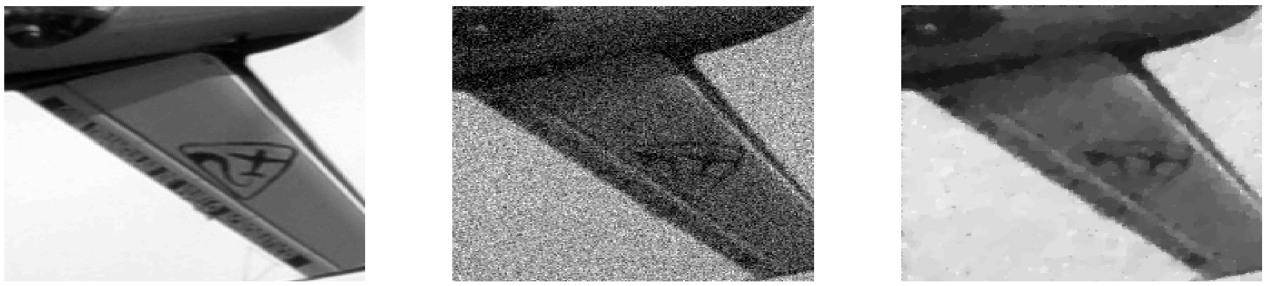


Figure 17: Poisson-Gaussian denoising: Original image (left), noisy image corrupted by Poisson noise and Gaussian noise with mean zero and variance 0.001 (center) and denoised image (right). Optimal parameters $\lambda_1^* = 1847.75$ and $\lambda_2^* = 73.45$.

# 6 Conclusion and outlook

Machine learning approaches in image processing and computer vision have mostly developed in parallel to their mathematical analysis counterparts, which have variational regularisation models at their core. Variational regularisation techniques offer rigorous and intelligible image analysis – which gives reliable and stable answers that provide us with insight in the constituents of the process and error estimates. This guarantee of giving a meaningful and stable result is crucial in most image processing applications, in biomedical and seismic imaging, in remote sensing and astronomy: provably giving an answer which is correct up to some error bounds is important when diagnosing patients, deciding upon a surgery or when predicting earthquakes. Machine learning methods, on the other hand, are extremely powerful as they learn from examples and are hence able to adapt to a specific task. The recent rise of deep learning gives us a glimpse on what is possible when intelligently using data to learn from. Todays (29 April 2015) search on a Google on 'deep learning image' just gave 59,800,000 hits. Deep learning is employed for all kinds of image processing and computer vision tasks, with impressive results! The weak point of machine learning approaches, however, is that they generally cannot offer stability or error bounds, neither provide most of them understanding about the driving factors (e.g. the important features in images) that led to their answer.

In this paper we wanted to give an account to a recent realisation in mathematical image processing that a marriage between machine learning and variational regularisation might be interesting – an attempt to bring together the Good from both worlds. In particular, we have discussed bilevel optimisation approaches in which optimal image regularisers and data fidelity terms are learned making use of a training set. We discussed the analysis of such a bilevel strategy in the continuum as well as their efficient numerical solution by quasi-Newton methods, and presented numerical examples for computing optimal regularisation parameters for TV, $\text{TGV}^2$ and *ICTV* denoising, as well as for deriving optimal data fidelity terms for TV image denoising for data corrupted with pure or mixed noise distributions.

Although the techniques presented in this article are mainly focused on denoising problems, the perspectives of using similar approaches in other image reconstruction problems (inpainting, segmentation, etc.) appear to be promising. Also the extension to color images deserves to be further studied.

Finally, there are still several open questions which deserve to be investigated in the future. Here a short list:

- Is it possible to obtain an optimality system for (P) by performing an asymptotic analysis when $\mu \to 0$?
- How to measure optimality? Are quality measures such as the signal-to-noise ratio and generalisations thereof [84] enough? Should one try to match characteristic expansions of the image such as Fourier or Wavelet expansions? [59].
- How to decide about the presence of a specific noise model? Is it possible to use sparse optimization for the automatic identification of one specific model? Can it be used to identify mixed models?

# 7 Acknowledgments

The original research behind this review has been supported by the King Abdullah University of Science and Technology (KAUST) Award No. KUK-I1-007-43, the EPSRC grants Nr. EP/J009539/1 "Sparse & Higher-order Image Restoration", and Nr. EP/M00483X/1 "Efficient computational tools for inverse imaging problems", the Escuela Politécnica Nacional de Quito under award PIS 12-14 and the MATHAmSud project SOCDE "Sparse

Optimal Control of Differential Equations". C. Cao and T. Valkonen have also been supported by Prometeo scholarships of SENESCYT (Ecuadorian Ministry of Higher Education, Science, Technology and Innovation).

**A data statement for the EPSRC** The data leading to this *review* publication will be made available, as appropriate, as part of the original publications that this work summarises.

# References

[1] W. Allard, Total variation regularization for image denoising, I. Geometric theory. *SIAM J. Math. Anal.* 39 (2008) 1150–1190.

[2] L. Ambrosio, A. Coscia and G. Dal Maso, Fine Properties of Functions with Bounded Deformation. *Arch. Ration. Mech. Anal.* 139 (1997) 201–238.

[3] L. Ambrosio, N. Fusco and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford University Press (2000).

[4] F. Baus, M. Nikolova and G. Steidl, Fully smoothed L1-TV models: Bounds for the minimizers and parameter choice. *J. Math. Imaging Vision* 48 (2014) 295–307.

[5] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books in Mathematics, Springer (2011).

[6] M. Benning, C. Brune, M. Burger and J. Müller, Higher-Order TV Methods—Enhancement via Bregman Iteration. *J. Sci. Comput.* 54 (2013) 269–310.

[7] M. Benning and M. Burger, Ground states and singular vectors of convex variational regularization methods. *Methods and Applications of Analysis* 20 (2013) 295–334, arXiv:1211.2057.

[8] L. Biegler, G. Biros, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Mallick, L. Tenorio, B. van Bloemen Waanders, K. Willcox and Y. Marzouk, *Large-scale inverse problems and quantification of uncertainty*, volume 712, John Wiley & Sons (2011).

[9] K. Bredies and M. Holler, A total variation-based JPEG decompression model. *SIAM J. Imaging Sci.* 5 (2012) 366–393.

[10] K. Bredies, K. Kunisch and T. Pock, Total Generalized Variation. *SIAM J. Imaging Sci.* 3 (2011) 492–526.

[11] K. Bredies, K. Kunisch and T. Valkonen, Properties of $L^1$-TGV$^2$: The one-dimensional case. *J. Math. Anal Appl.* 398 (2013) 438–454.

[12] K. Bredies and T. Valkonen, Inverse problems with second-order total generalized variation constraints, in: *Proc. SampTA 2011* (2011).

[13] T. Bui-Thanh, K. Willcox and O. Ghattas, Model reduction for large-scale systems with high-dimensional parametric input space. *SIAM J. Sci. Comput.* 30 (2008) 3270–3288.

[14] M. Burger, J. Müller, E. Papoutsellis and C.-B. Schönlieb, Total Variation Regularisation in Measurement and Image space for PET reconstruction. *Inverse Problems* 10 (2014).

[15] M. Burger, K. Papafitsoros, E. Papoutsellis and C.-B. Schönlieb, Infimal convolution regularisation functionals on BV and $L^p$ spaces. Part I: The finite p case (2015), submitted.

[16] R. H. Byrd, G. M. Chin, J. Nocedal and Y. Wu, Sample size selection in optimization methods for machine learning. *Math. Program.* 134 (2012) 127–155.

[17] J.-F. Cai, R. H. Chan and M. Nikolova, Two-phase approach for deblurring images corrupted by impulse plus gaussian noise. *Inverse Probl. Imaging* 2 (2008) 187–204.

[18] L. Calatroni, J. C. De los Reyes and C.-B. Schönlieb, Dynamic sampling schemes for optimal noise learning under multiple nonsmooth constraints, in: *System Modeling and Optimization*, 85–95, Springer Verlag (2014).

[19] L. Calatroni, J. C. D. los Reyes and C.-B. Schönlieb, Learning the optimal Total Variation denoising model for multiple noise distributions, in preparation.

[20] V. Caselles, A. Chambolle and M. Novaga, The discontinuity set of solutions of the TV denoising problem and some extensions. *Multiscale Model. Simul.* 6 (2007) 879–894.

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

[22] Y. Chen, T. Pock and H. Bischof, Learning $\ell_1$-based analysis and synthesis sparsity priors using bi-level optimization, in: *Workshop on Analysis Operator Learning vs. Dictionary Learning, NIPS 2012* (2012).

[23] Y. Chen, R. Ranftl and T. Pock, Insights into analysis operator learning: From patch-based sparse models to higher-order MRFs. *Image Processing, IEEE Transactions on* (2014), to appear.

[24] Y. Chen, W. Yu and T. Pock, On learning optimized reaction diffusion processes for effective image restoration, in: *IEEE Conference on Computer Vision and Pattern Recognition* (2015), to appear.

[25] C. V. Chung and J. C. De los Reyes, Learning optimal spatially-dependent regularization parameters in total variation image restoration, in preparation.

[26] J. Chung, M. Chung and D. P. O'Leary, Designing optimal spectral filters for inverse problems. *SIAM J. Sci. Comput.* 33 (2011) 3132–3152.

[27] J. Chung, M. I. Español and T. Nguyen, Optimal Regularization Parameters for General-Form Tikhonov Regularization. *arXiv preprint arXiv:1407.1911* (2014).

[28] A. Cichocki, S.-i. Amari et al., *Adaptive Blind Signal and Image Processing*, John Wiley Chichester (2002).

[29] R. Costantini and S. Susstrunk, Virtual sensor design, in: *Electronic Imaging 2004*, 408–419, International Society for Optics and Photonics (2004).

[30] J. C. De los Reyes, *Numerical PDE-Constrained Optimization*, Springer (2015).

[31] J. C. De los Reyes and C.-B. Schönlieb, Image denoising: Learning the noise model via Nonsmooth PDE-constrained optimization. *Inverse Probl. Imaging* 7 (2013).

[32] J. C. De los Reyes, C.-B. Schönlieb and T. Valkonen, Optimal parameter learning for higher-order total variation regularisation models, in preparation.

[33] J. C. De los Reyes, C.-B. Schönlieb and T. Valkonen, The structure of optimal parameters for image restoration problems, in preparation.

[34] J. Domke, Generic methods for optimization-based modeling, in: *International Conference on Artificial Intelligence and Statistics*, 318–326 (2012).

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

[36] H. W. Engl, M. Hanke and A. Neubauer, *Regularization of Inverse Problems*, volume 375, Springer (1996).

[37] S. N. Evans and P. B. Stark, Inverse problems as statistics. *Inverse Problems* 18 (2002) R55.

[38] A. Foi, Clipped noisy images: Heteroskedastic modeling and practical denoising. *Signal Processing* 89 (2009) 2609 – 2629, special Section: Visual Information Analysis for Security.

[39] K. Frick, P. Marnitz, A. Munk et al., Statistical multiresolution Dantzig estimation in imaging: Fundamental concepts and algorithmic framework. *Electronic Journal of Statistics* 6 (2012) 231–268.

[40] G. Gilboa, A total variation spectral framework for scale and texture analysis. *SIAM J. Imaging Sci.* 7 (2014) 1937–1961.

[41] E. Haber, L. Horesh and L. Tenorio, Numerical methods for the design of large-scale nonlinear discrete ill-posed inverse problems. *Inverse Problems* 26 (2010) 025002.

[42] E. Haber and L. Tenorio, Learning regularization functionals – a supervised training approach. *Inverse Problems* 19 (2003) 611.

[43] M. Hintermüller and A. Langer, Subspace Correction Methods for a Class of Nonsmooth and Nonadditive Convex Variational Problems with Mixed $L^1/L^2$ Data-Fidelity in Image Processing. *SIAM J. Imaging Sci.* 6 (2013) 2134–2173.

[44] 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) 1–23.

[45] M. Hintermüller and T. Wu, Bilevel Optimization for Calibrating Point Spread Functions in Blind Deconvolution (2014), preprint.

[46] H. Huang, E. Haber, L. Horesh and J. K. Seo, Optimal Estimation Of L1-regularization Prior From A Regularized Empirical Bayesian Risk Standpoint. *Inverse Probl. Imaging* 6 (2012).

[47] J. Idier, *Bayesian approach to inverse problems*, John Wiley & Sons (2013).

[48] A. Jezierska, E. Chouzenoux, J.-C. Pesquet and H. Talbot, A Convex Approach for Image Restoration with Exact Poisson-Gaussian Likelihood, Technical report (2013).

[49] J. Kaipio and E. Somersalo, *Statistical and computational inverse problems*, volume 160, Springer Science & Business Media (2006).

[50] N. Kingsbury, Complex wavelets for shift invariant analysis and filtering of signals. *Applied and Computational Harmonic Analysis* 10 (2001) 234–253.

[51] T. Klatzer and T. Pock, Continuous Hyper-parameter Learning for Support Vector Machines, in: *Computer Vision Winter Workshop (CVWW)* (2015).

[52] F. Knoll, K. Bredies, T. Pock and R. Stollberger, Second order total generalized variation (TGV) for MRI. *Magnetic Resonance in Medicine* 65 (2011) 480–491.

[53] V. Kolehmainen, T. Tarvainen, S. R. Arridge and J. P. Kaipio, Marginalization of uninteresting distributed parameters in inverse problems—application to diffuse optical tomography. *International Journal for Uncertainty Quantification* 1 (2011).

[54] K. Kunisch and T. Pock, A bilevel optimization approach for parameter learning in variational models. *SIAM J. Imaging Sci.* 6 (2013) 938–983.

[55] J. Lellmann, D. Lorenz, C.-B. Schönlieb and T. Valkonen, Imaging with Kantorovich-Rubinstein discrepancy. *SIAM J. Imaging Sci.* 7 (2014) 2833–2859, arXiv:1407.0221.

[56] J. Mairal, F. Bach, J. Ponce and G. Sapiro, Online dictionary learning for sparse coding, in: *Proceedings of the 26th Annual International Conference on Machine Learning*, 689–696, ACM (2009).

[57] J. Mairal, B. F, J. Ponce, G. Sapiro and A. Zisserman, Discriminative learned dictionaries for local image analysis. *CVPR* (2008).

[58] Y. Meyer, *Oscillating patterns in image processing and nonlinear evolution equations*, AMS (2001).

[59] D. Mumford and B. Gidas, Stochastic models for generic images. *Quarterly of Applied Mathematics* 59 (2001) 85–112.

[60] F. Natterer and F. Wübbeling, *Mathematical Methods in Image Reconstruction*, Monographs on Mathematical Modeling and Computation Vol 5, Philadelphia, PA: SIAM) (2001).

[61] M. Nikolova, A variational approach to remove outliers and impulse noise. *J. Math. Imaging Vision* 20 (2004) 99–120.

[62] J. Nocedal and S. Wright, *Numerical Optimization*, Springer Series in Operations Research and Financial Engineering, Springer (2006).

[63] P. Ochs, R. Ranftl, T. Brox and T. Pock, Bilevel Optimization with Nonsmooth Lower Level Problems, in: *International Conference on Scale Space and Variational Methods in Computer Vision (SSVM)* (2015), to appear.

[64] B. Olshausen and D. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381 (1996) 607–609.

[65] K. Papafitsoros and K. Bredies, A study of the one dimensional total generalised variation regularisation problem. *arXiv preprint arXiv:1309.5900* (2013).

[66] G. Peyré and J. M. Fadili, Learning analysis sparsity priors. *Sampta'11* (2011).

[67] R. Ranftl and T. Pock, A Deep Variational Model for Image Segmentation, in: *36th German Conference on Pattern Recognition (GCPR)* (2014).

[68] W. Ring, Structural Properties of Solutions to Total Variation Regularization Problems. *ESAIM: Math. Model. Numer. Anal.* 34 (2000) 799–810.

[69] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer (1998).

[70] S. Roth and M. J. Black, Fields of experts: A framework for learning image priors, in: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, 860–867, IEEE (2005).

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

[72] A. Sawatzky, C. Brune, J. Müller and M. Burger, Total Variation Processing of Images with Poisson Statistics, in: *Computer Analysis of Images and Patterns*, *Lecture Notes in Computer Science*, volume 5702, Edited by X. Jiang and N. Petkov, 533–540, Springer Berlin Heidelberg (2009).

[73] L. L. Scharf, *Statistical Signal Processing*, volume 98, Addison-Wesley Reading, MA (1991).

[74] U. Schmidt and S. Roth, Shrinkage fields for effective image restoration, in: *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2774–2781, IEEE (2014).

[75] J.-L. Starck, F. D. Murtagh and A. Bijaoui, Image restoration with noise suppression using a wavelet transform and a multiresolution support constraint (1994).

[76] E. Tadmor, S. Nezzar and L. Vese, A multiscale image representation using hierarchical (BV, L 2) decompositions. *Multiscale Model. Simul.* 2 (2004) 554–579.

[77] M. F. Tappen, Utilizing variational optimization to learn Markov random fields, in: *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 1–8, IEEE (2007).

[78] R. Temam, *Mathematical problems in plasticity*, Gauthier-Villars (1985).

[79] M. Unser, Texture classification and segmentation using wavelet frames. *Image Processing, IEEE Transactions on* 4 (1995) 1549–1560.

[80] M. Unser and N. Chenouard, A unifying parametric framework for 2D steerable wavelet transforms. *SIAM J. Imaging Sci.* 6 (2013) 102–135.

[81] T. Valkonen, The jump set under geometric regularisation. Part 1: Basic technique and first-order denoising. *SIAM J. Math. Anal.* (2015), accepted, arXiv:1407.1531.

[82] Y. Vardi, L. Shepp and L. Kaufman, A statistical model for positron emission tomography. *Journal of the American Statistical Association* 80 (1985) 8–20.

[83] F. Viola, A. Fitzgibbon and R. Cipolla, A unifying resolution-independent formulation for early vision, in: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 494–501, IEEE (2012).

[84] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Processing* 13 (2004) 600–612.

[85] G. Yu, G. Sapiro and S. Mallat, Image modeling and enhancement via structured sparse model selection. *Proc. IEEE Int. Conf. Image Processing* (2010).