Bilevel optimization approaches for learning the noise model in variational image processing

Juan Carlos De los Reyes

Centro de Modelización Matemática (MODEMAT)
EPN Quito, Ecuador

Joint with C.B. Schönlieb (Univ. of Cambridge), L. Calatroni (Univ. of Genova) and C. Cao (MODEMAT)

SVAN, March 2016

Outline

Outline

A generic inverse problem in imaging

The problem

Given data f, find the image information u which solves

$$f = T(u) + n$$

where T is a linear (or nonlinear) forward operator that models the relation between u and f and n is a noise component.

If T has an unbounded inverse, the problem is ill-posed. Causes: non-uniqueness, unstable inversion, noise, under sampling, ...

The problem has to be regularised by adding a-priori information on u...

The variational approach..

For given data f we seek a regularised image u by minimising

$$\mathcal{J}(u) = \underbrace{R(u)}_{\text{Prior}} + \lambda \underbrace{\phi(T(u), f)}_{\text{Data model}} \to \min_u,$$

where

- $R(u)$ is the prior (regularising) term: modelling a-priori information about the minimiser u in terms of regularity, e.g. $R(u) = \int u^2 \, dx$ which results in $u \in L^2$.
- $\phi(T(u), f)$ is a generic distance function, the data fidelity term of the functional which forces the minimiser u to obey (to a certain extent) the forward model.
- The parameter $\lambda > 0$ balances data model and prior.

Modelling

The result heavily depends on the correct modelling. There are two main degrees of freedom

- Image model: R, prior, regularity of the image, basis function representation, sparsity, ...
- Data model: $T, \ \phi, \ \lambda$, physical understanding, statistics, heuristics, ...

...and in both cases, we can try to extract this information directly from the data (experiments).

Modelling

The result heavily depends on the correct modelling. There are two main degrees of freedom

- Image model: R, prior, regularity of the image, basis function representation, sparsity, ...
- Data model: T, ϕ, λ, physical understanding, statistics, heuristics, ...

...and in both cases, we can try to extract this information directly from the data (experiments).

What difference does it make? A few examples . . .

H^1 versus TV regularisation

(a) original

(b) noisy

(c) $R(u) = \|\nabla u\|_2^2$

References: Rudin, Osher, Fatemi '92; Chambolle, Lions '97; Vese '01, . . .

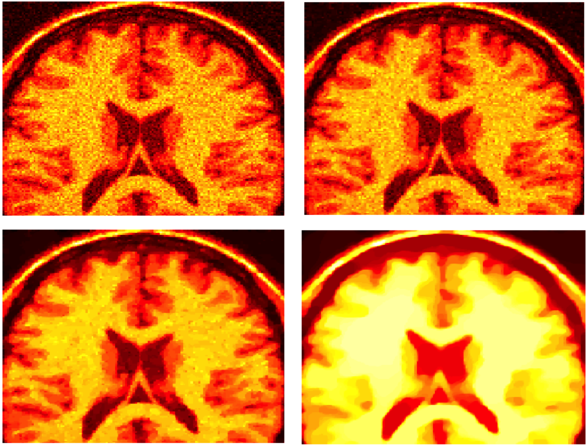
H^1 versus TV regularisation

(d) original (e) noisy (f) $R(u) = TV(u)$

References: Rudin, Osher, Fatemi '92; Chambolle, Lions '97; Vese '01, ...

Weight λ between image and data model

Total variation denoising for Gaussian noise



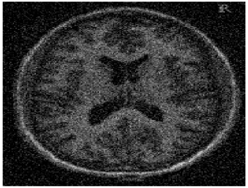
with increasing regularisation (from left to right).

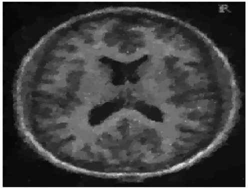
Effect of regulariser is complemented by effect of data term . . .

Choice of ϕ depends on data distribution

Gaussian

$$\phi(Tu, f) = \|Tu - f\|_2^2$$

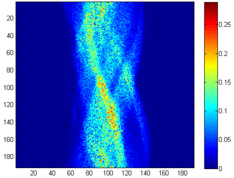


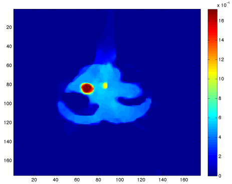


MRI

Poisson

$$\phi(Tu, f) = \int Tu - f \log(Tu)\, dx$$





PET*

Impulse

$$\phi(Tu, f) = \|Tu - f\|_1$$

sparse noise.

References: see recent works by Hohage and Werner ’12–

*Data courtesy of EIMI, Münster.

Optimise the choice of image and data model

State of the art in optimal model design

- Using a-priori information such as the noise level, e.g. Morozov '93; Engl, Hanke, Neubauer '96; ...; Baus, Nikolova, Steidl, JMIV '14;
- Adaptive parameter choice rules Hintermüller et al. '11; Frick, Marnitz, Munk '12–; Fornasier, Naumova, Pereverzyev '12–.
- Regulariser: Chung, O'Leary et al. '11– (optimal spectral filters); Sapiro et al. (dictionary learning); Peyré, Fadili '11; (learning sparsity priors).
- Bayesian statistics, e.g. Hero et al. Dobigeon, Hero, Tourneret '09; Park, Dobigeon, Hero '14.
- Statistical optimal design, e.g. Haber, Tenorio '03; Huang, Haber, Horesh, Seo '12; Ghattas et al. 08'–; Brune et al. '14.
- Examples of machine learning approaches: support vector machines, e.g. Tong, Chang '01, reproducible kernel Hilbert spaces Quang, Kang, Le '10, Gaussian mixture models Pedemonte, Bousse, Hutton, Arridge, Ourselin '11, learning by shape priors Cremers, Rousson '07, Schnoerr et al., Markov random fields Tappen '07; Domke '12–, non-smooth priors and noise models De Los Reyes, Schönlieb '13; Kunisch, Pock '13; Chung et al. '14.

State of the art in optimal model design

- Using a-priori information such as the noise level, e.g. Morozov '93; Engl, Hanke, Neubauer '96; . . . ; Baus, Nikolova, Steidl, JMIV '14;
- Adaptive parameter choice rules Hintermüller et al. '11; Frick, Marnitz, Munk '12–; Fornasier, Naumova, Pereverzyev '12–.
- Regulariser: Chung, O'Leary et al. '11– (optimal spectral filters); Sapiro et al. (dictionary learning); Peyré, Fadili '11; (learning sparsity priors).
- Bayesian statistics, e.g. Hero et al. Dobigeon, Hero, Tourneret '09; Park, Dobigeon, Hero '14.
- Statistical optimal design, e.g. Haber, Tenorio '03; Huang, Haber, Horesh, Seo '12; Ghattas et al. 08'–; Brune et al. '14.
- Examples of machine learning approaches: support vector machines, e.g. Tong, Chang '01, reproducible kernel Hilbert spaces Quang, Kang, Le '10, Gaussian mixture models Pedemonte, Bousse, Hutton, Arridge, Ourselin '11, learning by shape priors Cremers, Rousson '07, Schnoerr et al., Markov random fields Tappen '07; Domke '12–, non-smooth priors and noise models De Los Reyes, Schönlieb '13; Kunisch, Pock '13; Chung et al. '14.

Bilevel optimal reconstruction model

Assumptions

Training set of pairs $(f_k, u_k),\ k = 1, \dots, N$ with

- f_k imperfect data
- u_k represent the ground truth

Determine optimal regulariser R, data model ϕ, and λ in admissible set $\mathcal{A}$

$$\min_{(R,\phi,\lambda)\in\mathcal{A}} \sum_k \|\bar{u}_k - u_k\|^2_{L^2(\Omega)}$$

subject to

$$\bar{u}_k = \arg\min_u \left\{ R(u) + \int_\Omega \lambda\, \phi(Tu, f_k)\, dx \right\}$$

Bilevel optimal reconstruction model

Assumptions

Training set of pairs $(f_k, u_k),\ k = 1, \ldots, N$ with

- f_k imperfect data
- u_k represent the ground truth

Determine optimal regulariser R, data model ϕ, and λ in admissible set $\mathcal{A}$

$$\min_{(R,\phi,\lambda)\in\mathcal{A}} \sum_k \|\bar{u}_k - u_k\|^2_{L^2(\Omega)}$$

subject to

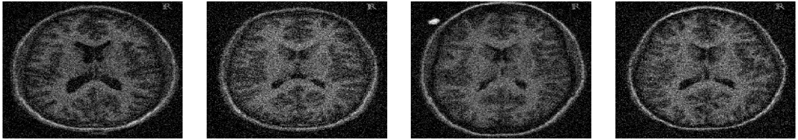
$$\bar{u}_k = \arg\min_u \left\{ R(u) + \int_\Omega \lambda\, \phi(Tu, f_k)\, dx \right\}$$

Learning from training sets

Image denoising training sets such as

Low resolution MRI scan

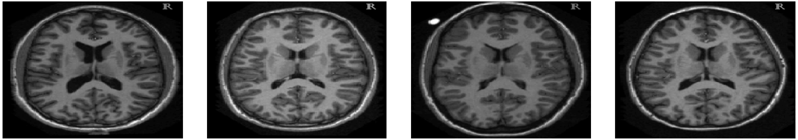
f_k



$\dots$

High resolution MRI scan

u_k

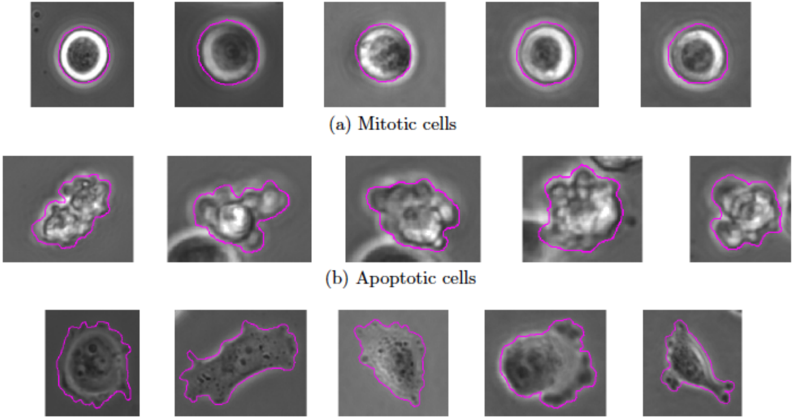


$\dots$

Simulated data from OASIS online database. Arridge, Kaipio, Kolehmainen, Schweiger, Somersalo, Tarvainen, Vauhkonen '06; Benning, Gladden, Holland, CBS,Valkonen '14

Learning from training sets

Image segmentation training sets such as



(a) Mitotic cells

(b) Apoptotic cells

(c) Flat cells

Figure 5.9.: Manually segmented test set of mitotic, apoptotic and flat cells
(Courtesy of Light Microscopy Core Facility, Cancer Research UK Cambridge Institute)

Learning by optimisation in imaging

Some related contributions

- Tappen et al. '07, '09; Domke '11–: Markov Random Field models; stochastic descent method
- Lui, Lin, Zhang and Su '09: optimal control approach, no analytical justification; promising numerical results.
- Horesh, Tenorio, Haber et al. '03–: optimal design (also for ℓ_1 minimisation).
- De los Reyes and Schönlieb '13: results in function space; derivative based optimization methods
- Kunisch and Pock '13: results for finite dimensional case; semismooth Newton method
- Chung et al. '14: finite dimensional; bounded operator T.

Learning in function space

Our goal: State and treat a nonsmooth optimization problem in function space (stick to the physical model).

- Infinite dimensional models more amenable to analysis of image features, e.g. edges.
- Lagrange multipliers and optimality condition.
- Compute optimal weights λ_i with a fast derivative-based and mesh independent optimisation method (second-order method); resolution independent imaging Viola, Fitzgibbon, Cipolla '12.
- Incorporate information of large image databases.
- Determination of the noise model present in the images.

Selection of the data model with a bilevel optimisation approach!

Outline

Total variation (TV) denoising

- Least squares minimization:

$$\min_u \int_\Omega |u - f|^2 \, dx, \qquad \text{(Gauss noise)}$$

where $f \in L^2(\Omega)$ is the noisy image

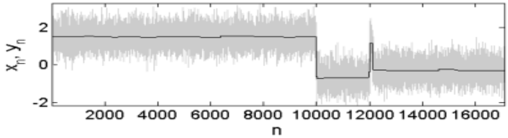
Total variation (TV) denoising

- Least squares minimization:

$$\min_u \int_\Omega |u - f|^2 \, dx, \qquad \text{(Gauss noise)}$$

  where $f \in L^2(\Omega)$ is the noisy image
- Include a total variation term in the minimization:

$$|Du|(\Omega)$$



Total variation (TV) denoising

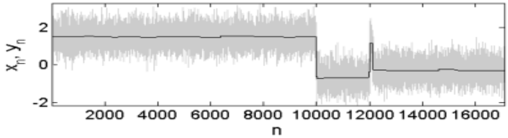
- Least squares minimization:

$$\min_u \int_\Omega |u - f|^2 \, dx, \qquad \text{(Gauss noise)}$$

where $f \in L^2(\Omega)$ is the noisy image

- Include a total variation term in the minimization:

$$|Du|(\Omega)$$



Minimization problem

$$\min_u \left(|Du|(\Omega) + \int_\Omega \lambda \, (u - f)^2 \, dx \right), \qquad \text{for } \lambda > 0.$$

TV denoising

$$\min_u \left(|Du|(\Omega) + \lambda \int_\Omega \phi(u,f)\, dx \right),$$

with

$$|Du|(\Omega) = TV(u) = \sup_{\mathbf{g}\in C_0^\infty(\Omega;\mathbb{R}^2), \|g\|_\infty \le 1} \int_\Omega u \,\nabla\cdot \mathbf{g}\, dx$$

$\longrightarrow$ the total variation of u in Ω

$\longrightarrow$ $\lambda > 0$ positive parameter

$\longrightarrow$ ϕ a suitable distance function called the data fidelity term.

TV denoising

$$\min_u \left(|Du|(\Omega) + \lambda \int_\Omega \phi(u, f)\, dx \right),$$

with

$$|Du|(\Omega) = TV(u) = \sup_{\mathbf{g} \in C_0^\infty(\Omega;\mathbb{R}^2), \|g\|_\infty \leq 1} \int_\Omega u \, \nabla \cdot \mathbf{g} \, dx$$

$\longrightarrow$ the total variation of u in Ω

$\longrightarrow$ $\lambda > 0$ positive parameter

$\longrightarrow$ ϕ a suitable distance function called the data fidelity term.

A generic TV denoising model

$$\min_u \left(|Du|(\Omega) + \sum_{i=1}^{d} \lambda_i \int_\Omega \phi_i(u, f)\, dx \right).$$

where

$\longrightarrow$ ϕ_i, $i = 1, \ldots, d$, convex & differentiable functions in u,

$\longrightarrow$ $\lambda_i \geq 0$

Choices for data fidelities ϕ_i's

ModeMat

- Gaussian noise: $\phi_1(u, f) = (u - f)^2$, ROF,Chambolle & Lions, Vese 1990's, ...
- Impulse noise: $\phi_2(u, f) = |u - f|$, Aujol, Gousseau, Nikolova, Osher, 2000's, ...
- Poisson noise: $\phi_3(u, f) = u - f \log u$, Burger et al. 2009-12
- Other possible choices, e.g. multiplicative noise, Rician noise Getreuer, Tong, Vese '11, ...

...weighted against each other with weights λ_i, which depend on the amount and strength of noise of different distributions in f.

Combinations of noise models also in Nikolova, Wen Chan '12

Learning TV denoising model

Assumptions

Training set of pairs $(f_k, u_k),\ k = 1, \dots, N$ with

- f_k noisy images
- u_k represent the ground truth

Learning TV denoising model

Assumptions

Training set of pairs $(f_k, u_k),\ k = 1, \ldots, N$ with

- f_k noisy images
- u_k represent the ground truth

Determine the optimal weights λ_i

$$\min_{\lambda_i \geq 0,\ i=1,\ldots,d} \sum_k \|\bar{u}_k - u_k\|^2_{L^2(\Omega)}$$

subject to: $$\bar{u}_k = \arg\min_u \left\{ |Du|(\Omega) + \sum_{i=1}^{d} \lambda_i \int_\Omega \phi_i(u, f_k) \right\}$$

Learning TV denoising model

Sobolev space setting

Determine the optimal weights λ_i

$$\min_{\lambda_i \geq 0,\ i=1,\dots,d} \sum_k \|\bar{u}_k - u_k\|^2_{L^2(\Omega)}$$

subject to: $$\bar{u}_k = \arg\min_u \left\{ \frac{\mu}{2} \int_\Omega |\nabla u|^2 + \int_\Omega |\nabla u| + \sum_{i=1}^{d} \lambda_i \int_\Omega \phi_i(u, f_k) \right\}$$

Learning TV denoising model

Sobolev space setting

Determine the optimal weights λ_i

$$\min_{\lambda_i \geq 0,\ i=1,\dots,d} \sum_k \|\bar{u}_k - u_k\|^2_{L^2(\Omega)}$$

subject to: $$\bar{u}_k = \arg\min_u \left\{ \frac{\mu}{2} \int_\Omega |\nabla u|^2 + \int_\Omega |\nabla u| + \sum_{i=1}^d \lambda_i \int_\Omega \phi_i(u, f_k) \right\}$$

Equivalently, due to optimality conditions:

$$\min_{\lambda_i \geq 0,\ i=1,\dots,d} \sum_k \|\bar{u}_k - u_k\|^2_{L^2(\Omega)}$$

subject to: $$-\mu \Delta \bar{u}_k + \sum_{i=1}^d \lambda_i\, \phi_i'(\bar{u}_k, f_k) + \partial(|\nabla \bar{u}_k|_{L^1}) \ni 0.$$

State of the art on optimality systems

Optimization of abstract variational inequalities

Barbu (1984, 1993), Tiba (1990), Bonnans-Tiba (1991), Wenbin-Rubio (1991), Bonnans-Casas (1995), Bergounioux (1998).

State of the art on optimality systems

Optimization of abstract variational inequalities

Barbu (1984, 1993), Tiba (1990), Bonnans-Tiba (1991), Wenbin-Rubio (1991), Bonnans-Casas (1995), Bergounioux (1998).

Not sharp enough!

State of the art on optimality systems

Optimization of abstract variational inequalities

Barbu (1984, 1993), Tiba (1990), Bonnans-Tiba (1991), Wenbin-Rubio (1991), Bonnans-Casas (1995), Bergounioux (1998).

Not sharp enough!

Renewed interest and improved results

Idea: exploit the special structure of TV term $\int_\Omega |\nabla u| \, dx$

DlRe (2011,2013), DlRe-Schönlieb (2013), Kunisch-Pock (2013), Hintermüller et al. (2015), DlRe-Meyer (2015).

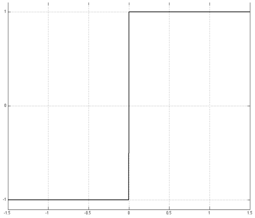
Learning TV denoising model

Tailored regularization

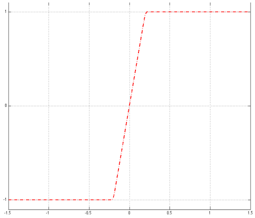
$$\min_{\lambda_i \geq 0,\ i=1,\ldots,d} \sum_k \|\bar{u}_k - u_k\|^2_{L^2(\Omega)}$$

$- \text{div } h_\gamma(\nabla u)$

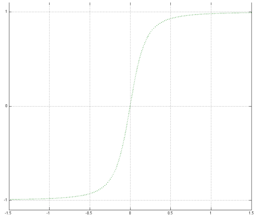
$$\text{subject to: } -\mu\Delta\bar{u}_k + \sum_{i=1}^{d} \lambda_i\ \phi_i'(\bar{u}_k, f_k) + \cancel{\partial(|\nabla\bar{u}_k|_{L^1})} \ni 0$$



Subdifferential of $|\cdot|$



Huber type function



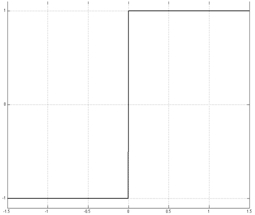
Learning TV denoising model

Tailored regularization

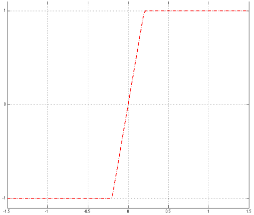
$$\min_{\lambda_i \geq 0,\ i=1,\ldots,d} \sum_k \|\bar{u}_k - u_k\|^2_{L^2(\Omega)}$$

$$- \operatorname{div} h_\gamma(\nabla u)$$

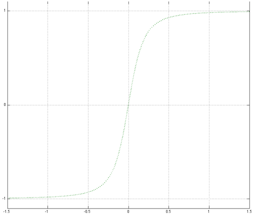
$$\text{subject to: } -\mu \Delta \bar{u}_k + \sum_{i=1}^{d} \lambda_i\ \phi_i'(\bar{u}_k, f_k) + \cancel{\partial(|\nabla \bar{u}_k|_{L^1})} \ni 0$$



Subdifferential of $|\cdot|$



Huber type function



$\frac{x}{\sqrt{x^2+\epsilon^2}}$

Breaktrough thanks to the type of regularization!

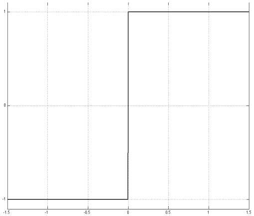
Learning TV denoising model

Tailored regularization

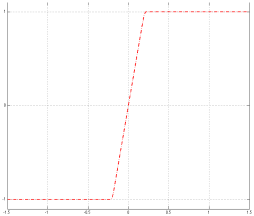
$$\min_{\lambda_i \geq 0,\ i=1,\dots,d} \sum_k \|\bar{u}_k - u_k\|^2_{L^2(\Omega)}$$

subject to: $$-\mu \Delta \bar{u}_k + \sum_{i=1}^{d} \lambda_i\, \phi_i'(\bar{u}_k, f_k) + \cancel{\partial(|\nabla \bar{u}_k|_{L^1})} \ni 0$$

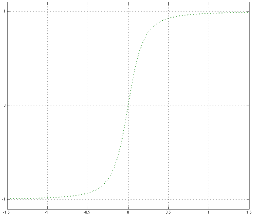
$- \text{div}\, h_\gamma(\nabla u)$



Subdifferential of $|\cdot|$



Huber type function ✓



$\cancel{\frac{x}{\sqrt{x^2+\epsilon^2}}}$

Breaktrough thanks to the type of regularization!

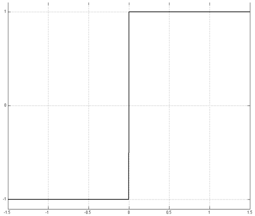
Learning TV denoising model

Tailored regularization

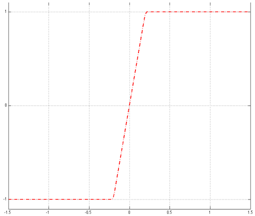
$$\min_{\lambda_i \geq 0,\ i=1,\dots,d} \sum_k \|\bar{u}_k - u_k\|^2_{L^2(\Omega)}$$

$$\text{subject to: } -\mu\Delta\bar{u}_k + \sum_{i=1}^{d} \lambda_i\ \phi_i'(\bar{u}_k, f_k) + \cancel{\partial(|\nabla\bar{u}_k|_{L^1})} \ni 0$$

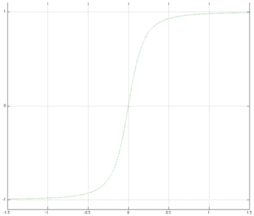
$$-\text{div } h_\gamma(\nabla u)$$



Subdifferential of $|\cdot|$



Huber type function ✓



$\cancel{\dfrac{x}{\sqrt{x^2+\epsilon^2}}}$

Breaktrough thanks to the type of regularization!

In this setting we can prove . . .

- existence of an optimal solution.
- differentiability of solution operator and derivation of sharp optimality system.
- convergence as Huber regularisation $\gamma \to \infty$ to sharp optimality system for non-smooth problem.
- Γ convergence of de-noising functional as ellipticity $\mu \to 0$.

In this setting we can prove . . .

- existence of an optimal solution.
- differentiability of solution operator and derivation of sharp optimality system.
- convergence as Huber regularisation $\gamma \to \infty$ to sharp optimality system for non-smooth problem.
- Γ convergence of de-noising functional as ellipticity $\mu \to 0$.

$\Rightarrow$ Consistency

In this setting we can prove . . .

- existence of an optimal solution.
- differentiability of solution operator and derivation of sharp optimality system.
- convergence as Huber regularisation $\gamma \to \infty$ to sharp optimality system for non-smooth problem.
- Γ convergence of de-noising functional as ellipticity $\mu \to 0$.

$\Rightarrow$ Consistency

. . . and in the numerics the parameters $0 < \mu \ll 1$ and $\gamma \gg 1$.

In this setting we can prove . . .

- existence of an optimal solution.
- differentiability of solution operator and derivation of sharp optimality system.
- convergence as Huber regularisation $\gamma \to \infty$ to sharp optimality system for non-smooth problem.
- Γ convergence of de-noising functional as ellipticity $\mu \to 0$.

$\Rightarrow$ Consistency

. . . and in the numerics the parameters $0 < \mu \ll 1$ and $\gamma \gg 1$.

For a direct (unregularized) approach: see the talk of David Villacís later this afternoon.

Optimality system for the regularized problems

There exist Lagrange multipliers $(p_\gamma, \varphi) \in H_0^1(\Omega) \times \mathbb{R}^d$ such that

$$-\mu\Delta u_\gamma - \text{div } h_\gamma(Du_\gamma) + \sum_{i=1}^{d} \int_\Omega \bar{\lambda}_i \, \phi_i'(u_\gamma, f) v \, dx = 0, \tag{1}$$

$$\begin{aligned} &-\mu\Delta p_\gamma - \text{div } (h_\gamma'(Du_\gamma)^* Dp_\gamma) \\ &\qquad + \sum_{i=1}^{d} \int_\Omega \lambda_i \, \phi_i''(u_\gamma, f) \, p_\gamma = -2(u_\gamma - u_k), \end{aligned} \tag{2}$$

$$\varphi_i = \int_\Omega p_\gamma \phi_i'(u_\gamma, f) \, dx, \qquad i = 1, ..., d, \tag{3}$$

$$\varphi_i \geq 0, \ \lambda_i \geq 0, \ \varphi_i \lambda_i = 0, \qquad i = 1, ..., d. \tag{4}$$

Optimality system for the regularized problems

There exist Lagrange multipliers $(p_\gamma, \varphi) \in H_0^1(\Omega) \times \mathbb{R}^d$ such that

$$-\mu\Delta u_\gamma - \text{div } h_\gamma(Du_\gamma) + \sum_{i=1}^{d} \int_\Omega \bar{\lambda}_i \, \phi_i'(u_\gamma, f) v \, dx = 0, \tag{1}$$

$$\begin{aligned} -\mu\Delta p_\gamma - \text{div } (h_\gamma'(Du_\gamma)^* Dp_\gamma) \\ + \sum_{i=1}^{d} \int_\Omega \lambda_i \, \phi_i''(u_\gamma, f) \, p_\gamma = -2(u_\gamma - u_k), \end{aligned} \tag{2}$$

$$\varphi_i = \int_\Omega p_\gamma \phi_i'(u_\gamma, f) \, dx, \qquad i = 1, ..., d, \tag{3}$$

$$\varphi_i \geq 0, \ \lambda_i \geq 0, \ \varphi_i \lambda_i = 0, \qquad i = 1, ..., d. \tag{4}$$

Characterization of the gradient

Optimality system for bilevel problem

Passing to the limit as $\gamma \to \infty$ we are able to derive a sharp OS:

$$-\mu\Delta\bar{u} - \text{div}\, q + \sum_{i=1}^{d} \int_\Omega \bar{\lambda}_i\, \phi_i'(\bar{u}) = 0, \tag{1}$$

$$\langle q, \nabla\bar{u}\rangle_{\mathbb{R}^2} = |\nabla\bar{u}| \qquad \text{a.e. in } \Omega, \tag{2}$$

$$\begin{aligned}\mu(\nabla p, \nabla v) &+ \langle \xi, \nabla v\rangle_{(\nabla H_0^1(\Omega))'} \\ &+ \sum_{i=1}^{d} \int_\Omega \bar{\lambda}_i\, \phi_i''(\bar{u}) p\, v\, dx = -2(\bar{u} - u_k, v), \forall v \in H_0^1(\Omega),\end{aligned} \tag{3}$$

$$\langle \xi, \nabla p\rangle_{(\nabla H_0^1(\Omega))'} \geq 0, \qquad \langle \xi, \nabla\bar{u}\rangle_{(\nabla H_0^1(\Omega))'} = 0, \tag{4}$$

$$\varphi_i = \int_\Omega p\phi_i'(\bar{u}, f)\, dx, \qquad i = 1, ..., d, \tag{5}$$

$$\varphi_i \geq 0,\ \lambda_i \geq 0,\ \varphi_i\lambda_i = 0, \qquad i = 1, ..., d. \tag{6}$$

Optimality system for bilevel problem

Passing to the limit as $\gamma \to \infty$ we are able to derive a sharp OS:

$$-\mu\Delta\bar{u} - \text{div}\, q + \sum_{i=1}^{d} \int_\Omega \bar{\lambda}_i\, \phi_i'(\bar{u}) = 0, \tag{1}$$

$$\langle q, \nabla\bar{u}\rangle_{\mathbb{R}^2} = |\nabla\bar{u}| \quad \text{a.e. in } \Omega, \tag{2}$$

$$\mu(\nabla p, \nabla v) + \langle \xi, \nabla v\rangle_{(\nabla H_0^1(\Omega))'} + \sum_{i=1}^{d} \int_\Omega \bar{\lambda}_i\, \phi_i''(\bar{u}) p\, v\, dx = -2(\bar{u} - u_k, v), \forall v \in H_0^1(\Omega), \tag{3}$$

$$\langle \xi, \nabla p\rangle_{(\nabla H_0^1(\Omega))'} \geq 0, \qquad \langle \xi, \nabla\bar{u}\rangle_{(\nabla H_0^1(\Omega))'} = 0, \tag{4}$$

$$\varphi_i = \int_\Omega p\phi_i'(\bar{u}, f)\, dx, \qquad i = 1, ..., d, \tag{5}$$

$$\varphi_i \geq 0,\ \lambda_i \geq 0,\ \varphi_i\lambda_i = 0, \qquad i = 1, ..., d. \tag{6}$$

C-stationarity

Numerical strategy

Solve

$$\min_{\lambda_i \geq 0,\ i=1,\ldots,d} \|\bar{u}_k - u_k\|^2_{L^2(\Omega)}$$

subject to

$$-\mu\Delta\bar{u} - \text{div}\,(h_\gamma(\nabla\bar{u})) + \sum_{i=1}^{d} \lambda_i\ \phi_i'(\bar{u}, f) = 0,$$

by quasi-Newton method (BFGS)

- state equation is solved by Newton type algorithm (varies with ϕ)
- evaluation of the gradient of the cost functional with adjoint information
- Armijo line search with curvature verification.

Optimal parameter for Gaussian model

$$\min_{\lambda \geq 0} \|u - u_k\|_{L^2}^2$$

subject to:

$$\min_{u \geq 0} \left\{ \frac{\mu}{2} \|Du\|_{L^2}^2 + \|Du\|_\gamma + \frac{\lambda}{2} \|u - f_k\|_{L^2}^2 \right\}$$

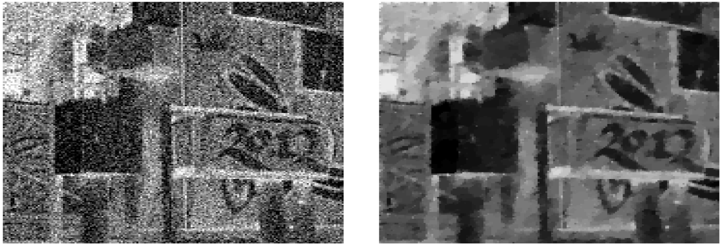
Noise $n \in N(0, 0.002)$ (optimal parameter $\lambda^* = 2980$)

Optimal parameter for Gaussian model

$$\min_{\lambda \geq 0} \|u - u_k\|_{L^2}^2$$

subject to:

$$\min_{u \geq 0} \left\{ \frac{\mu}{2}\|Du\|_{L^2}^2 + \|Du\|_\gamma + \frac{\lambda}{2}\|u - f_k\|_{L^2}^2 \right\}$$



Noise $n \in N(0, 0.02)$ (optimal parameter $\lambda^* = 1770.9$)

Mixed Gauss & Poisson noise

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

subject to:

$$\min_{u \geq 0} \left\{ \frac{\mu}{2} \|Du\|_{L^2}^2 + \|Du\|_\gamma + \frac{\lambda_1}{2} \|u - f_k\|_{L^2}^2 + \lambda_2 \int_\Omega (u - f_k \log u) \, dx \right\}.$$

Poisson noise and Gaussian noise $n \in N(0, 0.001)$. Optimal parameters $\lambda_1^* = 1847.75$ and $\lambda_2^* = 73.45$.

Impulse noise

$$\min \frac{1}{2}\|u-u_k\|_{L^2}^2$$

subject to:

$$\min_{u\geq 0}\left\{\frac{\mu}{2}\|Du\|_{L^2}^2+\|Du\|_\gamma+\lambda\|u-f_k\|_\gamma\right\}$$

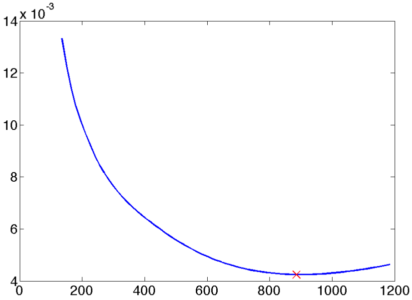
Impulse noise with 5% corrupted pixels; optimal parameter $\lambda^* = 45.88$

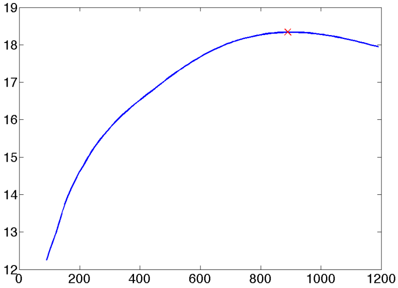
Optimality?

Quality measure

- Original cost functional (left figure) $\|u - u_k\|_{L^2}^2$
- Signal to noise ratio (right figure)

$$SNR = 20 \times \log_{10}\left(\frac{\|u_k\|_{L^2}}{\|u - u_k\|_{L^2}}\right),$$





Outline

Learning noise by means of a database

In applications the noise level can be tuned

In MRI or PET the accuracy of the measurements depends on the setup of the experiment. The training set can be provided by a series of measurements using (simulated) phantoms.

Learning noise by means of a database

In applications the noise level can be tuned

In MRI or PET the accuracy of the measurements depends on the setup of the experiment. The training set can be provided by a series of measurements using (simulated) phantoms.

Consider:

$$\min_{\lambda_i \geq 0,\ i=1,\dots,d} J(\lambda) := \frac{1}{2N} \sum_{k=1}^{\mathrm{N}} \left\| u_k^{TV} - \tilde{u}_k \right\|_{L^2(\Omega)}^2 + \frac{\beta}{2} \sum_{i=1}^{d} |\lambda_i|^2$$

subject to the set **of nonlinear constraints**:

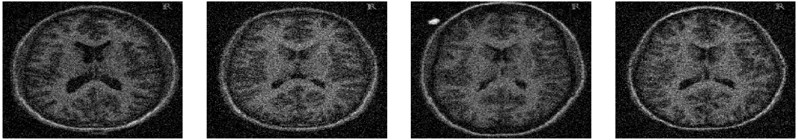
$$u_k^{TV} = \text{argmin}_{u \in BV(\Omega)} \left(|Du|(\Omega) + \sum_{i=1}^{d} \lambda_i \int_\Omega \phi_i(u, \tilde{f}_k)\, dx \right),\ k = 1, \dots, N$$

encoding the training set made up by the pairs $(\tilde{f}_k, \tilde{u}_k)$.

The database[†]

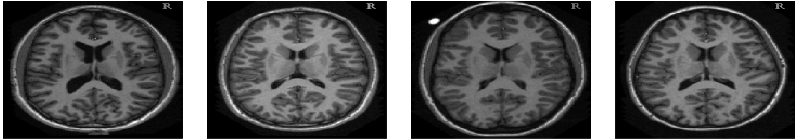
One parameter estimation for noisy images corrupted with Gaussian noise:

$\tilde{F} =$



$\dots$

$\tilde{U} =$

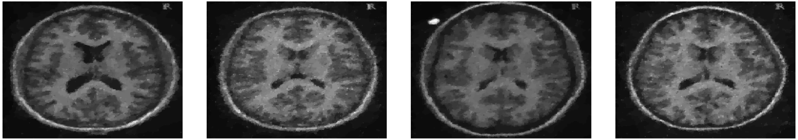


$\dots$

[†]OASIS online MRI database.

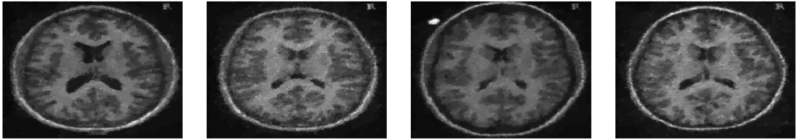
Optimal results

Denoised versions with optimal parameter $\hat{\lambda}$:

$\hat{U}^{TV} =$  $\cdots$

Optimal results

Denoised versions with optimal parameter $\hat{\lambda}$:

$\hat{U}^{TV} =$  $\dots$

Numerical difficulties

Numerically, the problem is hardly tractable due to the **large size** of the dictionary and the **nonsmooth nature** of the constraints which need to be solved in each iteration of the optimisation algorithm.

How we can compute it **efficiently** for **large** databases?

Batch sample methods

Ideally, we would like to **sample** randomly from the set of PDEs:

- **We want:** to reduce the number of PDEs that need to be solved.
- **We don't want:** to perform a “poor" approximation of the original problem.

Batch sample methods

Ideally, we would like to **sample** randomly from the set of PDEs:

- **We want:** to reduce the number of PDEs that need to be solved.
- **We don't want:** to perform a “poor" approximation of the original problem.

Questions:

* Batch approximation of which operators?
* Size of the sample? Update?
* How to check the quality of sampling approach?

BFGS with dynamic sampling

Modified version of algorithm by Byrd et al. (2012)

Algorithm 1 Dynamic sampling BFGS

1: Initialize: λ_0, sample $\mathcal{S}_0$ with $|S_0| \ll N$ and model parameter θ, $k = 0$.
2: **while** BFGS not converging, $k \geq 0$
3: sample $|S_k|$ PDE constraints to solve
4: update BFGS matrix
5: compute search direction d_k and steplength α_k (*Armijo*)
6: define new iterate: $\lambda_{k+1} = \lambda_k + \alpha_k d_k$
7: choose a sample S_{k+1} such that $|S_{k+1}| = |S_k|$
8: **if** appropriate condition on the quality of the approximation **then**
9: maintain the sample size $|S_{k+1}| = |S_k|$
10: **else** augment S_k such that 6: is verified.
11: **end**

Quality of the approximation $\rightarrow$ variance in replacing ∇J with

$$\nabla J_S = \frac{1}{2|S|} \sum_{k \in S} \left\| u_k^{TV} - \tilde{u}_k \right\|^2_{L^2(\Omega)} + \frac{\beta}{2} \sum_{i=1}^{d} |\lambda_i|^2$$

The condition on the batch gradient variance

For $\theta \in [0,1)$:

$\left\|\nabla J_S(\lambda) - \nabla J(\lambda)\right\|_2 \leq \theta \left\|\nabla J_S(\lambda)\right\|_2 \Rightarrow d = -\nabla J_S(\lambda)$ is a descent dir. $(*)$

The condition on the batch gradient variance

For $\theta \in [0, 1)$:

$\mathbb{E} \left\| \nabla J_S(\lambda) - \nabla J(\lambda) \right\|_2^2 \leq \theta^2 \left\| \nabla J_S(\lambda) \right\|_2^2 \Rightarrow d = -\nabla J_S(\lambda)$ is a descent dir. $(*)$

The condition on the batch gradient variance

For $\theta \in [0,1)$:

$\boxed{\|Var(\nabla J_S(\lambda))\|_1} \leq \theta^2 \|\nabla J_S(\lambda)\|_2^2 \Rightarrow d = -\nabla J_S(\lambda)$ is a descent dir. $(*)$

Approximating $Var(\nabla J_S(\lambda))$ using the **sample variance** provides a condition that needs to be checked for every random sample of size $S \ldots$

Is the condition $(*)$ satisfied?

- **Yes:** keep the size S fixed and pick another random sample of size S.
- **No:** augment S such that the condition is fulfilled.

Robustness and efficiency

- One parameter estimation: Gaussian noise $\mathcal{N}(0, 0.005)$.
- Database of variable size: 150×150 images.
- Not-sampling vs. sampling technique.

N	$\hat{\lambda}$	$\hat{\lambda}_S$	$\lvert S_0\rvert$	$\lvert S_{end}\rvert$	eff.	eff.$_S$	diff.
10	3334.5	3417.7	2	3	140	**84**	$2,4\%$
20	3437.0	3473.2	4	4	240	**120**	1.0%
30	3436.5	3471.6	6	6	420	**180**	1.0%
40	3431.5	3350.6	8	9	560	**272**	2.3%
50	3425.8	3280.5	10	10	700	**220**	4.2%
60	3426.0	3301.3	12	12	840	**264**	3.6%
70	3419.7	3417.8	14	14	980	**336**	$< 1\%$
80	3418.1	3283.5	16	16	1120	**480**	$3,9\%$
90	3416.6	3323.7	18	18	1260	**648**	2.7%
100	3413.6	3314.2	20	20	1400	**520**	2.9%

Parameters: $\beta = 10^{-10}$, $\lambda_0 = 1000$, $\theta = 0.5$, $|S_0| = 20\% N$.

Multiple noise case

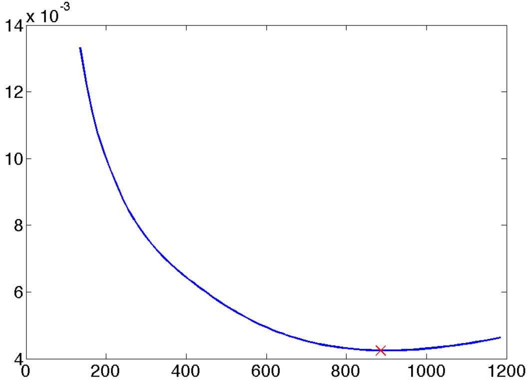
- **Two** parameters: Gaussian noise $\mathcal{N}(0, 0.005)$ + impulse noise.
- Database of variable size: 150×150 images.
- Not-sampling vs. sampling technique.

N	$\hat{\lambda}_{1S}$	$\hat{\lambda}_{2S}$	$\|S_0\|$	$\|S_{end}\|$	eff.	eff. Dyn.S.	diff.
10	86.31	28.43	2	7	180	**70**	5.2%
20	90.61	26.96	4	6	920	**180**	5.3%
30	94.36	29.04	6	7	2100	**314**	5.6%
40	88.88	31.56	8	8	880	**496**	1.2%
50	88.92	29.81	10	10	2200	**560**	$< 1\%$
60	89.64	28.36	12	12	1920	**336**	1.9%
70	86.09	28.09	14	14	2940	**532**	3.3%
80	87.68	29.97	16	16	3520	**448**	$< 1\%$

Parameters: $\beta = 0, \lambda_{1_0} = 10, \lambda_{2_0} = 10, \theta = 0.5, |S_0| = 20\% N$.

A new initialisation of BFGS

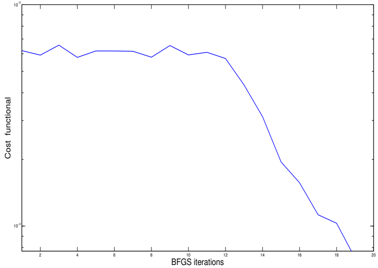
Looking at the cost functional:



$\Rightarrow$ very flat $\approx$ “small” Hessian..

A new initialisation of BFGS

T

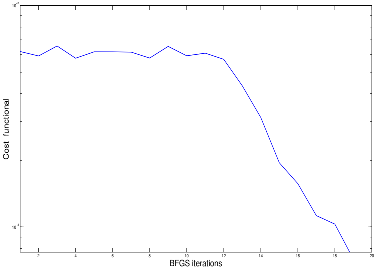


- Early oscillations;
- Late superlinear convergence (once $B \approx H$);

$\Longrightarrow$ (standard) initialisation $B_0 = Id$ is not ideal for our problem...

A new initialisation of BFGS

T



- Early oscillations;
- Late superlinear convergence (once $B \approx H$);

$\Longrightarrow$ (standard) initialisation $B_0 = Id$ is not ideal for our problem...

We approximate B_0 by considering $J(\lambda)$ as before, subject to:

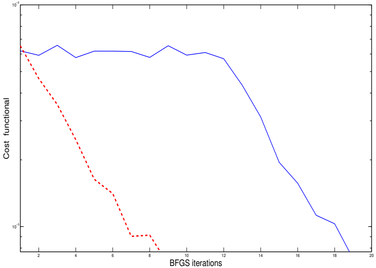
$$-\mu \Delta u_k + \sum_{i=1}^{d} \lambda_i \phi_i'(u_k, f) = 0 \quad \text{(linear constraints)} \quad \forall k = 1, \dots, N$$

and set:

$$\boxed{B_0 = J''(\lambda)}.$$

A new initialisation of BFGS

T



- Early oscillations;
- Late superlinear convergence (once $B \approx H$);

$\Longrightarrow$ (standard) initialisation $B_0 = Id$ is not ideal for our problem...

We approximate B_0 by considering $J(\lambda)$ as before, subject to:

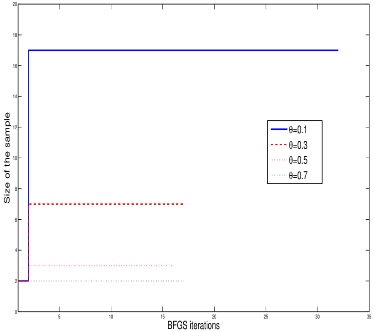
$$-\mu\Delta u_k + \sum_{i=1}^{d} \lambda_i \phi_i'(u_k, f) = 0 \quad \text{(linear constraints)} \quad \forall k = 1, \dots, N$$

and set:

$$\boxed{B_0 = J''(\lambda)}.$$

Accuracy and sample size selection

The parameter $\theta \in [0,1)$ affects *accuracy* and *efficiency*:



Sample size condition:

$$\boxed{\|Var(\nabla J_S(\lambda))\|_1} \leq \boldsymbol{\theta}^2 \,\|\nabla J_S(\lambda)\|_2^2$$

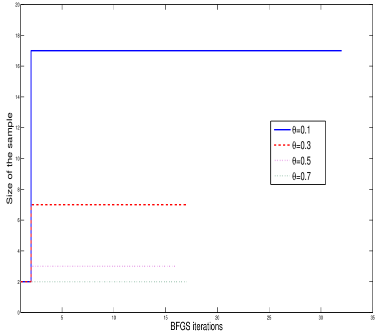
- $\boldsymbol{\theta} \nearrow$: larger variances allowed, smaller samples. Less accuracy, but gain in efficiency.
- $\boldsymbol{\theta} \searrow$: smaller variances are forced, bigger samples. More accuracy, but efficiency suffers.

Accuracy and sample size selection

The parameter $\theta \in [0, 1)$ affects *accuracy* and *efficiency*:

Sample size condition:

$$\|Var(\nabla J_S(\lambda))\|_1 \leq \boldsymbol{\theta}^2 \|\nabla J_S(\lambda)\|_2^2$$



- $\boldsymbol{\theta} \nearrow$: larger variances allowed, smaller samples. Less accuracy, but gain in efficiency.
- $\boldsymbol{\theta} \searrow$: smaller variances are forced, bigger samples. More accuracy, but efficiency suffers.

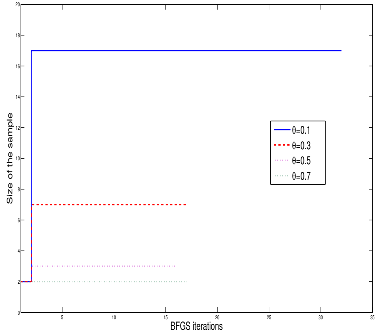
Furthermore, we would like to include the BFGS matrix B in the descent condition for faster convergence. . . Ongoing work.

Accuracy and sample size selection

The parameter $\theta \in [0,1)$ affects *accuracy* and *efficiency*:

Sample size condition:

$$\|Var(\nabla J_S(\lambda))\|_1 \leq \boldsymbol{\theta}^2 \left\|B^{-1}\nabla J_S(\lambda)\right\|_2^2$$



- $\boldsymbol{\theta} \nearrow$: larger variances allowed, smaller samples. Less accuracy, but gain in efficiency.
- $\boldsymbol{\theta} \searrow$: smaller variances are forced, bigger samples. More accuracy, but efficiency suffers.

Furthermore, we would like to include the BFGS matrix B in the descent condition for faster convergence... Ongoing work.

Outline

Bilevel optimization problem

Optimization problem in $H^1(\Omega)$

$$\min_{\lambda \geq 0} \frac{1}{2}\|\bar{u} - u_T\|^2_{L^2(\Omega)} + \frac{\beta}{2}\|\lambda\|^2_{H^1(\Omega)}$$

subject to:

$$\bar{u} = \operatorname*{arg\,min}_{u \in V \subset H^1(\Omega)} \left\{ \frac{\mu}{2}\int_\Omega |\nabla u|^2 + \int_\Omega |\nabla u| + \frac{1}{2}\int_\Omega \lambda(x)(u-f)^2 \right\}$$

Bilevel optimization problem

Second formulation

$$\min_{\lambda \in H^1(\Omega)} \frac{1}{2} \int_\Omega |u - u_T|^2 \, dx + \frac{\beta}{2} \|\lambda\|^2_{H^1(\Omega)}$$

subject to:

$$-\mu \Delta u - \text{div}\, (h_\gamma(\nabla u)) + \lambda(x)(u - f) = 0,$$

$$\lambda(x) \geq 0 \text{ in } \Omega.$$

PDE-constrained optimization problem with control constraints and control in the coefficients.

Ingredients for optimality conditions

- There exists an optimal weight $\lambda \in V \subset H^1(\Omega)$ solution of the bilevel problem.

Ingredients for optimality conditions

ModeMat

- There exists an optimal weight $\lambda \in V \subset H^1(\Omega)$ solution of the bilevel problem.
- The solution operator $\lambda \to u(\lambda)$ is Fréchet differentiable and its derivative corresponds to the unique solution of a linearized PDE. Moreover, if h_γ is C^2, then the solution operator is twice Fréchet differentiable.

Ingredients for optimality conditions

- There exists an optimal weight $\lambda \in V \subset H^1(\Omega)$ solution of the bilevel problem.
- The solution operator $\lambda \to u(\lambda)$ is Fréchet differentiable and its derivative corresponds to the unique solution of a linearized PDE. Moreover, if h_γ is C^2, then the solution operator is twice Fréchet differentiable.
- Consider the classical obstacle problem: Find $\lambda \geq 0$ such that

$$a(\lambda, v - \lambda) \geq (f, v - \lambda), \forall v \geq 0.$$

If the operator coefficients and the domain are regular enough, and $f \in L^2(\Omega)$, then $\lambda \in H^2(\Omega)$.

Optimality system

There exist Lagrange multipliers $(p, \varphi) \in H^1(\Omega) \times L^2(\Omega)$ such that

$$-\mu\Delta u - \text{div}\, q + \lambda(u - f) = 0,$$

$$q = h_\gamma(\nabla u)$$

$$-\mu\Delta p - \text{div}\, \zeta + \lambda\, p = -(u - u_T),$$

$$\zeta = h'_\gamma(\nabla u)^* \nabla p$$

$$-\beta\Delta\lambda + \beta\lambda + p(u - f) = \varphi$$

$$\varphi \geq 0,\ \lambda \geq 0,\ \varphi\lambda = 0 \text{ a.e. in } \Omega.$$

Optimality system

There exist Lagrange multipliers $(p, \varphi) \in H^1(\Omega) \times L^2(\Omega)$ such that

$$-\mu\Delta u - \text{div}\, q + \lambda(u - f) = 0,$$

$$q = h_\gamma(\nabla u)$$

$$-\mu\Delta p - \text{div}\, \zeta + \lambda\, p = -(u - u_T),$$

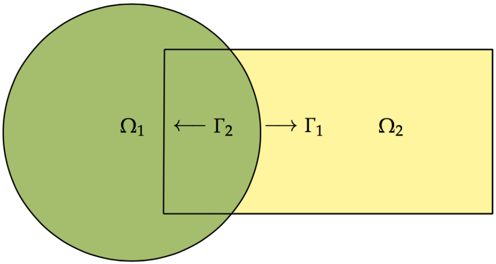
$$\zeta = h_\gamma'(\nabla u)^* \nabla p$$

$$-\beta\Delta\lambda + \beta\lambda + p(u - f) = \varphi$$

$$\varphi \geq 0,\ \lambda \geq 0,\ \varphi\lambda = 0 \text{ a.e. in } \Omega.$$

Need an efficient large scale nonlinear complementarity solver

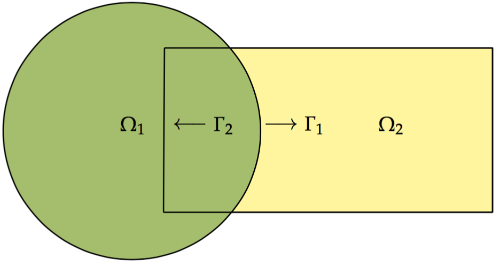
Schwarz domain decomposition



$$-\Delta u_1^{k+1} = f \text{ in } \Omega_1$$
$$u_1^{k+1} = u_2^k \text{ on } \Gamma_1$$
$$u_1^{k+1} = 0 \text{ on } \partial\Omega_1 \backslash \Gamma_1$$

$$-\Delta u_2^{k+1} = f \text{ in } \Omega_2$$
$$u_2^{k+1} = u_1^k \text{ on } \Gamma_1$$
$$u_2^{k+1} = 0 \text{ on } \partial\Omega_1 \backslash \Gamma_1$$

Schwarz domain decomposition



$$-\Delta u_1^{k+1} = f \text{ in } \Omega_1 \qquad -\Delta u_2^{k+1} = f \text{ in } \Omega_2$$
$$u_1^{k+1} = u_2^k \text{ on } \Gamma_1 \qquad u_2^{k+1} = u_1^k \text{ on } \Gamma_1$$
$$u_1^{k+1} = 0 \text{ on } \partial\Omega_1 \backslash \Gamma_1 \qquad u_2^{k+1} = 0 \text{ on } \partial\Omega_1 \backslash \Gamma_1$$

Direct application of Schwarz methods to TV denoising problems (see, e.g., Fornasier, Langer, Schönlieb (2009)).

Optimized Schwarz

- Modified transmission conditions:

$$
\begin{aligned}
-\Delta u_1^{k+1} &= f \text{ in } \Omega_1 & -\Delta u_2^{k+1} &= f \text{ in } \Omega_2 \\
(\partial_n + \sigma_1) u_1^{k+1} &= (\partial_n + \sigma_1) u_2^k \text{ on } \Gamma_1 & (\partial_n + \sigma_2) u_2^{k+1} &= (\partial_n + \sigma_2) u_1^k \text{ on } \Gamma_1 \\
u_1^{k+1} &= 0 \text{ on } \partial\Omega_1 \backslash \Gamma_1 & u_2^{k+1} &= 0 \text{ on } \partial\Omega_1 \backslash \Gamma_1
\end{aligned}
$$

Optimized Schwarz

- Modified transmission conditions:

$$
\begin{aligned}
&-\Delta u_1^{k+1} = f \text{ in } \Omega_1 && -\Delta u_2^{k+1} = f \text{ in } \Omega_2 \\
&(\partial_n + \sigma_1)u_1^{k+1} = (\partial_n + \sigma_1)u_2^k \text{ on } \Gamma_1 && (\partial_n + \sigma_2)u_2^{k+1} = (\partial_n + \sigma_2)u_1^k \text{ on } \Gamma_1 \\
&u_1^{k+1} = 0 \text{ on } \partial\Omega_1 \backslash \Gamma_1 && u_2^{k+1} = 0 \text{ on } \partial\Omega_1 \backslash \Gamma_1
\end{aligned}
$$

- Choice of weights according to high-low frequency Fourier analysis (Gander (2006))

Optimized Schwarz

- Modified transmission conditions:

$$-\Delta u_1^{k+1} = f \text{ in } \Omega_1 \qquad -\Delta u_2^{k+1} = f \text{ in } \Omega_2$$

$$(\partial_n + \sigma_1)u_1^{k+1} = (\partial_n + \sigma_1)u_2^k \text{ on } \Gamma_1 \qquad (\partial_n + \sigma_2)u_2^{k+1} = (\partial_n + \sigma_2)u_1^k \text{ on } \Gamma_1$$

$$u_1^{k+1} = 0 \text{ on } \partial\Omega_1 \backslash \Gamma_1 \qquad u_2^{k+1} = 0 \text{ on } \partial\Omega_1 \backslash \Gamma_1$$

- Choice of weights according to high-low frequency Fourier analysis (Gander (2006))
- Analysis of KKT matrix variants:

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

Semismooth Newton methods

- The last two equations can be reformulated as:

$$-\beta\Delta\lambda + \beta\lambda + (u - f)p - \max\big(0, -\beta\Delta\lambda + (u - f)p\big) = 0$$

Semismooth Newton methods

- The last two equations can be reformulated as:

$$-\beta\Delta\lambda + \beta\lambda + (u - f)p - \max\left(0, -\beta\Delta\lambda + (u - f)p\right) = 0$$

- Semismooth Newton methods may be considered in each subdomain.

Semismooth Newton methods

- The last two equations can be reformulated as:

$$-\beta\Delta\lambda + \beta\lambda + (u - f)p - \max\left(0, -\beta\Delta\lambda + (u - f)p\right) = 0$$

- Semismooth Newton methods may be considered in each subdomain.

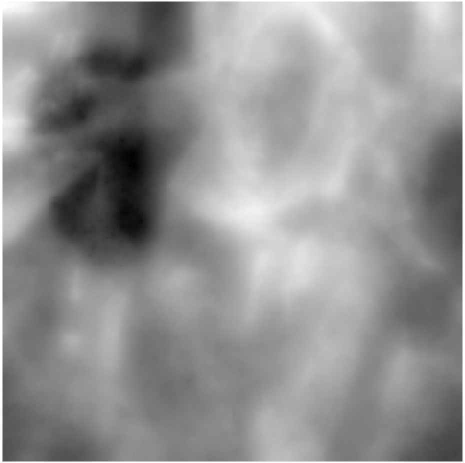
Theorem

The semi-smooth Newton method applied to the optimality system converges locally with superlinear convergence rate, provided that $\|y_0 - y^*\|$ *is sufficiently small.*

Experiments

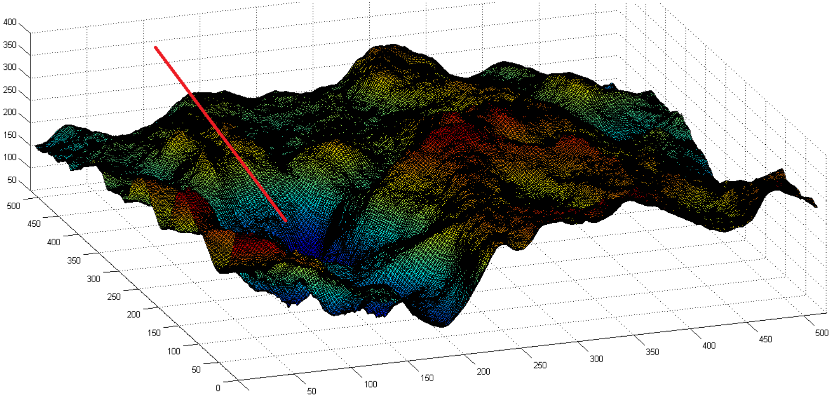
Original image and noisy image.

Experiments



Resulting image and optimal weight.

Experiments



Surface plot of optimal weight.

Outline

Conclusions and outlook

Conclusions:

- Optimise physical image and data model by bilevel optimisation.
- Optimise-then-discretise: model in the continuum (resolution independent);
- Setup of efficient numerics for Gaussian, Poisson and impulse noise, in case of small and large training sets;
- Spatial dependent $H^1(\Omega)$-weight functions results in a large OS, that can be efficiently solved by combining DD and SSN.

Outlook:

- Alternative cost functionals. How to measure optimality?
- More complex (realistic, mixed) noise models; sparse control on parameters.
- General linear operator T (inpainting, segmentation, ...)
- Optimising other elements in the model, e.g. regularisation procedure, acquisition (sampling), inpainting procedure ...

Thank you very much for your attention!

- J. C. De Los Reyes, and C.-B. Schönlieb, *Image denoising: Learning noise distribution via PDE-constrained optimisation*, Inverse Problems and Imaging, Vol. 7, 1183-1214, 2013.
- L. Calatroni, J. C. De Los Reyes, and C.-B. Schönlieb, *Dynamic sampling schemes for optimal noise learning under multiple nonsmooth constraints*, to appear in IFIP TC7-2013 proceedings.

More information see: `http://www.modemat.epn.edu.ec`