Bilevel optimisation for hyperparameter estimation in imaging inverse problems

Luca Calatroni, CR CNRS, Laboratoire I3S
CNRS, UCA, Inria SAM, France

(Virtual) Peyresq school
June 2021

Structure of the course

Lecture I:

- Need for hyperparameter estimation in imaging
- Review of *a posteriori*/*a priori* approaches
- Bilevel modelling: general viewpoint, specific instances
- Theoretical guarantees and algorithmic insights $\rightarrow$ technical

Lecture II:

- Learning the noise model
- Learning the regularisation models
- Extensions: learning spatially-dependent regularisation weights, non-convex modelling
- Learning problem operators, relations with other (deep) learning approaches

Introduction

Hyperparameter setting in variational inverse problems: general framework

Problem: for $A \in \mathbb{R}^{m \times n}$ and $f \in \mathbb{R}^m$, seek $x \in \mathbb{R}^n$ such that:

$$\boxed{f = \mathcal{T}(Ax)}$$

. . . Pierre's course: due to ill-posedness, regularization is needed!

Following a Bayesian/MAP approach consider:

$$P(f|Ax, \theta_l) \quad \text{(likelihood/fidelity)}, \qquad P(x; \theta_p) \quad \text{(prior/regularisation)}$$

with θ_l and θ_p hyperparameters of the distributions.

Hyperparameter setting in variational inverse problems: general framework

Problem: for $A \in \mathbb{R}^{m\times n}$ and $f \in \mathbb{R}^m$, seek $x \in \mathbb{R}^n$ such that:

$$\boxed{f = Ax + b}$$

...Pierre's course: due to ill-posedness, regularization is needed!

Following a Bayesian/MAP approach consider:

$$P(f|Ax, \theta_l) \quad \text{(likelihood/fidelity)}, \qquad P(x; \theta_p) \quad \text{(prior/regularisation)}$$

with θ_l and θ_p hyperparameters of the distributions.

Example: quadratic case

Assume noise is additive, white, Gaussian (AWGN) + Gaussian prior:

$$b \sim \mathcal{N}(0, \sigma_b^2 Id) \qquad x \sim \mathcal{N}(0, \sigma_x^2 Id), \qquad \sigma_b, \sigma_x > 0$$

MAP estimation reduces to the following problem:

$$\text{find } x^* = \arg\min_x \frac{1}{2\sigma_b^2}\|f - Ax\|_2^2 + \frac{1}{2\sigma_x^2}\|x\|^2$$

Hyperparameter setting in variational inverse problems: general framework

Problem: for $A \in \mathbb{R}^{m\times n}$ and $f \in \mathbb{R}^m$, seek $x \in \mathbb{R}^n$ such that:

$$\boxed{f = Ax + b}$$

. . . Pierre's course: due to ill-posedness, regularization is needed!

Following a Bayesian/MAP approach consider:

$$P(f|Ax, \theta_l) \quad \text{(likelihood/fidelity)}, \qquad P(x; \theta_p) \quad \text{(prior/regularisation)}$$

with θ_l and θ_p hyperparameters of the distributions.

Example: quadratic case

Assume noise is additive, white, Gaussian (AWGN) + Gaussian prior:

$$b \sim \mathcal{N}(0, \sigma_b^2 Id) \qquad x \sim \mathcal{N}(0, \sigma_x^2 Id), \qquad \sigma_b, \sigma_x > 0$$

MAP estimation reduces to the following problem:

$$\text{find } x^* = \arg\min_x \frac{\sigma_x^2}{2\sigma_b^2}\|f - Ax\|_2^2 + \frac{1}{2}\|x\|^2$$

Probabilistic interpretation: balance between regularisation/fidelity = ratio between underlying probabilistic hyperparameters.

Hyperparameter setting in variational inverse problems: general framework

Problem: for $A \in \mathbb{R}^{m\times n}$ and $f \in \mathbb{R}^m$, seek $x \in \mathbb{R}^n$ such that:

$$\boxed{f = Ax + b}$$

... Pierre's course: due to ill-posedness, regularization is needed!

Following a Bayesian/MAP approach consider:

$$P(f|Ax, \theta_l) \quad \text{(likelihood/fidelity)}, \qquad P(x; \theta_p) \quad \text{(prior/regularisation)}$$

with θ_l and θ_p hyperparameters of the distributions.

Example: quadratic case

Assume noise is additive, white, Gaussian (AWGN) + Gaussian prior:

$$b \sim \mathcal{N}(0, \sigma_b^2 Id) \qquad x \sim \mathcal{N}(0, \sigma_x^2 Id), \qquad \sigma_b, \sigma_x > 0$$

MAP estimation reduces to the following problem:

$$\text{find } x^* = \arg\min_x \frac{\mu}{2}\|f - Ax\|_2^2 + \frac{1}{2}\|x\|^2$$

Probabilistic interpretation: balance between regularisation/fidelity = ratio between underlying probabilistic hyperparameters.

Hyperparameter setting in variational inverse problems: general framework

Problem: for $A \in \mathbb{R}^{m \times n}$ and $f \in \mathbb{R}^m$, seek $x \in \mathbb{R}^n$ such that:

$$\boxed{f = Ax + b}$$

. . . Pierre's course: due to ill-posedness, regularization is needed!

Following a Bayesian/MAP approach consider:

$$P(f|Ax, \theta_l) \quad \text{(likelihood/fidelity)}, \qquad P(x; \theta_p) \quad \text{(prior/regularisation)}$$

with θ_l and θ_p hyperparameters of the distributions.

Example: quadratic case

Assume noise is additive, white, Gaussian (AWGN) + Gaussian prior:

$$b \sim \mathcal{N}(0, \sigma_b^2 Id) \qquad x \sim \mathcal{N}(0, \sigma_x^2 Id), \qquad \sigma_b, \sigma_x > 0$$

MAP estimation reduces to the following problem:

$$\text{find } x^* = \arg\min_x \frac{1}{2}\|f - Ax\|_2^2 + \frac{\alpha}{2}\|x\|^2$$

with $\alpha = 1/\mu$.

Probabilistic interpretation: balance between regularisation/fidelity = ratio between underlying probabilistic hyperparameters.

Hyperparameter setting: example I (TV restoration)

AWGN + Total Variation regularisation (Rudin, Osher, Fatemi,'92):

TV regularisation

AWGN noise + Laplace distribution on discrete image gradient magnitudes

$$b \sim \mathcal{N}(0, \sigma_b^2 Id) \qquad |(Dx)_i|_2 \sim \mathcal{L}(0, \tau), \ i = 1, \ldots, n \qquad \sigma_b, \tau > 0$$

MAP estimation:

$$\arg\min_x \frac{1}{2}\|f - Ax\|_2^2 + \alpha\|Dx\|_{2,1}$$

where $\alpha = \alpha(\sigma_b^2, \tau)$ and $\|Dx\|_{2,1} = \sum_{i=1}^{n} \sqrt{(D_h x)_i^2 + (D_v x)_i^2}$, ("1-norm" of Dx)

Hyperparameter setting: example I (TV restoration)

AWGN + Total Variation regularisation (Rudin, Osher, Fatemi,'92):

TV regularisation

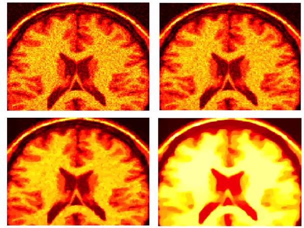
AWGN noise + Laplace distribution on discrete image gradient magnitudes

$$b \sim \mathcal{N}(0, \sigma_b^2 Id) \qquad |(Dx)_i|_2 \sim \mathcal{L}(0, \tau),\ i = 1, \dots, n \qquad \sigma_b, \tau > 0$$

MAP estimation:

$$\arg\min_x \frac{1}{2}\|f - Ax\|_2^2 + \alpha \|Dx\|_{2,1}$$

where $\alpha = \alpha(\sigma_b^2, \tau)$ and $\|Dx\|_{2,1} = \sum_{i=1}^{n} \sqrt{(D_h x)_i^2 + (D_v x)_i^2}$, ("1-norm" of Dx)



Importance of parameter selection in TV restoration

Hyperparameter setting: example II (noise modelling)

Non-Gaussian noise scenarios. Popular noise models:

- AWLN/impulsive noise: $b \sim \mathcal{L}(0,\tau) \to \frac{1}{\tau}\|f - Ax\|_1$
- Poisson noise (non-additive) [1]: $f = \mathcal{P}(Ax)$ with

  $$f_j \sim \mathcal{P}((Ax)_j), j = 1, \dots, n \to KL(f, Ax) = \mu \sum_{j=1}^{n} \left((Ax)_j - f_j \log(Ax)_j\right)$$

[1] Review for astronomical/biological imaging: Bertero, Boccacci, Ruggiero, '18

Hyperparameter setting: example II (noise modelling)

Non-Gaussian noise scenarios. Popular noise models:

- AWLN/impulsive noise: $b \sim \mathcal{L}(0,\tau) \to \frac{1}{\tau}\|f - Ax\|_1$
- Poisson noise (non-additive) [1]: $f = \mathcal{P}(Ax)$ with

$$f_j \sim \mathcal{P}((Ax)_j), j = 1,\dots,n \to KL(f, Ax) = \mu \sum_{j=1}^{n} \left((Ax)_j - f_j \log(Ax)_j\right)$$

Mixed noise models

- linear combination of data fidelities (De Los Reyes, Schoenlieb, '13...):

$$\arg\min_x \sum_{i=1}^{d} \mu_i \Phi_i(Ax; f) + R(x), \qquad \mu_j \geq 0$$

- non-linear combinations (exact log-likelihood Chouzenoux, Jezierska, Pesquet, Talbot, '15,... infimal-convolution Calatroni, De Los Reyes, Schoenlieb, '17...):

$$\arg\min_x \mathcal{G}(\Phi_1(Ax; f), \dots, \Phi_d(Ax; f); \mu_1, \dots, \mu_d) + R(x), \qquad \mu_j \geq 0$$

[1] Review for astronomical/biological imaging: Bertero, Boccacci, Ruggiero, '18

Hyperparameter setting: example II (noise modelling)

Non-Gaussian noise scenarios. Popular noise models:

- AWLN/impulsive noise: $b \sim \mathcal{L}(0,\tau) \to \frac{1}{\tau}\|f - Ax\|_1$
- Poisson noise (non-additive) [1]: $f = \mathcal{P}(Ax)$ with

$$f_j \sim \mathcal{P}((Ax)_j), j = 1, \dots, n \to KL(f, Ax) = \mu \sum_{j=1}^{n} \left((Ax)_j - f_j \log(Ax)_j\right)$$

Mixed noise models

- linear combination of data fidelities (De Los Reyes, Schoenlieb, '13...):

$$\arg\min_x \sum_{i=1}^{d} \mu_i \Phi_i(Ax; f) + R(x), \qquad \mu_j \geq 0$$

- non-linear combinations (exact log-likelihood Chouzenoux, Jezierska, Pesquet, Talbot, '15,... infimal-convolution Calatroni, De Los Reyes, Schoenlieb, '17...):

$$\arg\min_x \mathcal{G}(\Phi_1(Ax; f), \dots, \Phi_d(Ax; f); \mu_1, \dots, \mu_d) + R(x), \qquad \mu_j \geq 0$$

Here, hyperparameters control fidelities VS. regularisation, but also balance each other

[1] Review for astronomical/biological imaging: Bertero, Boccacci, Ruggiero, '18

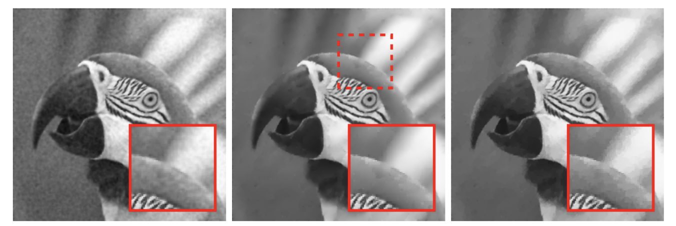
Hyperparameter setting: example II (higher-order & analysis-type reg.)

Higher-order regularisation: combine gradient with higher-order information (ICTV Chambolle, Lions, '97, TGV Bredies,Kunisch, Pock, '10):

$$TGV^2_{(\alpha,\beta)}(x) := \min_w \ \alpha \int_\Omega |\nabla x - w| + \beta \int_\Omega |Ew|$$

where, roughly, $Ew = \frac{1}{2}(\nabla w + \nabla w^T)$. Here, $\theta = (\alpha, \beta) > 0$ control the amount of TV-type regularisation against higher-order smoothing

$$\arg\min_x TGV^2_{(\alpha,\beta)}(x) + \Phi(Ax; f).$$



Too low, optimal and too high β ($\sim$ TV)

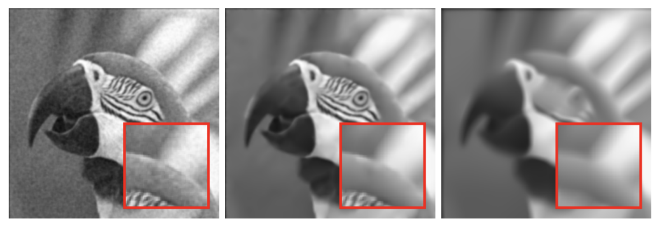
Hyperparameter setting: example II (higher-order & analysis-type reg.)

Higher-order regularisation: combine gradient with higher-order information (ICTV Chambolle, Lions, '97, TGV Bredies,Kunisch, Pock, '10):

$$TGV^2_{(\alpha,\beta)}(x) := \min_w \ \alpha \int_\Omega |\nabla x - w| + \beta \int_\Omega |Ew|$$

where, roughly, $Ew = \frac{1}{2}(\nabla w + \nabla w^T)$. Here, $\theta = (\alpha, \beta) > 0$ control the amount of TV-type regularisation against higher-order smoothing

$$\arg\min_x TGV^2_{(\alpha,\beta)}(x) + \Phi(Ax; f).$$



Too low, too high α ($\sim$ TV^2)

Hyperparameter setting: example II (higher-order & analysis-type reg.)

Higher-order regularisation: combine gradient with higher-order information (ICTV Chambolle, Lions, '97, TGV Bredies,Kunisch, Pock, '10):

$$TGV^2_{(\alpha,\beta)}(x) := \min_w \ \alpha \int_\Omega |\nabla x - w| + \beta \int_\Omega |Ew|$$

where, roughly, $Ew = \frac{1}{2}(\nabla w + \nabla w^T)$. Here, $\theta = (\alpha, \beta) > 0$ control the amount of TV-type regularisation against higher-order smoothing

$$\arg\min_x TGV^2_{(\alpha,\beta)}(x) + \Phi(Ax; f).$$

Analysis approach (Elad, Milanfar, Rubinstein, '07): express image prior in terms of linear combination of linear (convolutional) operators $K_k \in \mathbb{R}^{s\times n}, k = 1, \ldots, q$ (Kunisch, Pock, '13)

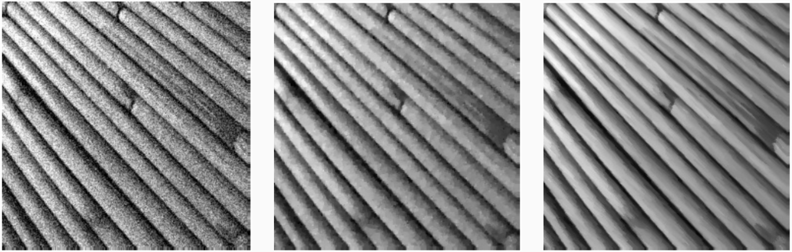
$$\arg\min_x \frac{1}{2} \sum_{k=1}^{q} \theta_k \|K_k x\|_2^2 + \Phi(Ax; f)$$

$$\|K_k x\|_2^2 = \sum_{i=1}^{n} |(K_k x)_i|^2$$

(see later. . .)

Hyperparameter setting: example II (space-variant regularisation)[2]

At each pixel, local image scale $\alpha_i > 0$, directionality $\gamma_i \in [-\pi/2, \pi/2)$ and anisotropy $a_i \in (0, 1]$ is encoded as hyperparameter.



Noisy, TV, DTV results

Hyperparameter setting: example II (space-variant regularisation)[2]

At each pixel, local image scale $\alpha_i > 0$, directionality $\gamma_i \in [-\pi/2, \pi/2)$ and anisotropy $a_i \in (0, 1]$ is encoded as hyperparameter.

* Weighted TV (Hintermueller et al., '17-...)

$$\boxed{\mathrm{WTV}_{(\alpha_1,\dots,\alpha_n)}(x) := \sum_{i=1}^{n} \alpha_i |(Dx)_i|}$$

[2] review: Pragliola, Calatroni, Lanza, Sgallari, '21

Hyperparameter setting: example II (space-variant regularisation)[2]

At each pixel, local image scale $\alpha_i > 0$, directionality $\gamma_i \in [-\pi/2, \pi/2)$ and anisotropy $a_i \in (0, 1]$ is encoded as hyperparameter.

* Weighted TV (Hintermueller et al., '17-...)

$$\boxed{\mathrm{WTV}_{(\alpha_1,\ldots,\alpha_n)}(x) := \sum_{i=1}^{n} \alpha_i |(Dx)_i|}$$

* Directional weighted TV (Bayram, Kamasak, '12, Kongskov, Dong, Knudsen, '19)

$$\boxed{\mathrm{DTV}_{\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{a}}(u) := \sum_{i=1}^{n} \alpha_i |\mathrm{diag}(1, a_i) R_{-\gamma_i} (Dx)_i|}$$

[2] review: Pragliola, Calatroni, Lanza, Sgallari, '21

Hyperparameter setting: example II (space-variant regularisation)[2]

At each pixel, local image scale $\alpha_i > 0$, directionality $\gamma_i \in [-\pi/2, \pi/2)$ and anisotropy $a_i \in (0, 1]$ is encoded as hyperparameter.

* Weighted TV (Hintermueller et al., '17-...)

$$\mathrm{WTV}_{(\alpha_1,\ldots,\alpha_n)}(x) := \sum_{i=1}^{n} \alpha_i |(Dx)_i|$$

* Directional weighted TV (Bayram, Kamasak, '12, Kongskov, Dong, Knudsen, '19)

$$\mathrm{DTV}_{\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{a}}(u) := \sum_{i=1}^{n} \alpha_i |\mathrm{diag}(1, a_i) R_{-\gamma_i} (Dx)_i|$$

...possibly **many** parameters to estimate!

---

[2]review: Pragliola, Calatroni, Lanza, Sgallari, '21

Classical approaches

A posteriori, a priori, error-free parameter choice

- **A posteriori** (Morozov, '66, Miller '70): available estimate on the data discrepancy and/or value of the regulariser at the ideal solution :

$$\Phi(A\tilde{f}; f) \leq \epsilon \qquad \text{and/or} \qquad R(\tilde{f}) \leq S, \qquad \epsilon, S > 0$$

  Morozov's discrepancy principle: choose θ s.t. $\Phi(Ax(\theta); f) \leq \epsilon(\sigma^2)$, where σ^2 relates to noise intensity, e.g., Gaussian noise variance.

A posteriori, a priori, error-free parameter choice

- **A posteriori** (Morozov, '66, Miller '70): available estimate on the data discrepancy and/or value of the regulariser at the ideal solution :

$$\Phi(A\tilde{f}; f) \leq \epsilon \qquad \text{and/or} \qquad R(\tilde{f}) \leq S, \qquad \epsilon, S > 0$$

  Morozov's discrepancy principle: choose θ s.t. $\Phi(Ax(\theta); f) \leq \epsilon(\sigma^2)$, where σ^2 relates to noise intensity, e.g., Gaussian noise variance.

- **A priori** (Engl, Neubauer,'00): estimate on noise level + *prior* smoothness assumption on solution. No need to compute $x(\theta)$! Typically find optimal θ by 'measuring' optimality (convergence rates. . .)

A posteriori, a priori, error-free parameter choice

- **A posteriori** (Morozov, '66, Miller '70): available estimate on the data discrepancy and/or value of the regulariser at the ideal solution :

$$\Phi(A\tilde{f}; f) \leq \epsilon \qquad \text{and/or} \qquad R(\tilde{f}) \leq S, \qquad \epsilon, S > 0$$

  Morozov's discrepancy principle: choose θ s.t. $\Phi(Ax(\theta); f) \leq \epsilon(\sigma^2)$, where σ^2 relates to noise intensity, e.g., Gaussian noise variance.

- **A priori** (Engl, Neubauer,'00): estimate on noise level + *prior* smoothness assumption on solution. No need to compute $x(\theta)$! Typically find optimal θ by 'measuring' optimality (convergence rates...)

- **Error-free**: "early" (heuristic) attempts of learning-from-data strategies.

  Generalised cross-validation (Golub et al. '79): let $x(\theta)^{[k]} \in \mathbb{R}^n$ be obtained from measurements $f^{[k]} = (f_1, f_2, \ldots, f_{k-1}, f_{k+1}, \ldots, f_m)$.

  Choose θ s.t. it minimizes ('leave one out')

$$\theta \mapsto \sum_{k=1}^{m} |(Ax(\theta)^{[k]})_k - f_k|^2 \quad \text{s.t.} \ (Ax(\theta)^{[k]})_k \approx f_k$$

Other approaches: L-curve (Hansen, '92)...

Pros VS. cons

- **A posteriori**/Morozov's discrepancy principle: choose θ s.t. $\Phi(Ax(\theta); f) \leq \epsilon(\sigma^2)$, where $\epsilon(\sigma^2)$ depends on noise level (unknown in many applications!)
- **A priori** (Engl, Neubauer,'00): estimate on noise level + *prior* smoothness assumption on solution: very limiting in practice!
- **Error-free**: generalised cross-validation (Golub et al. '79) requires computation of large matrix traces: intractable for large-scale problems.

What is still done in practice

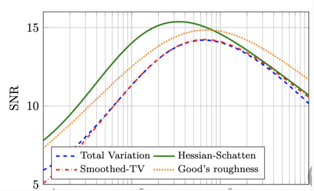
Brute force approach: scalar problem, $\theta > 0$

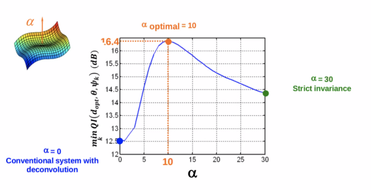
What is still done in practice

Brute force approach: scalar problem, $\theta > 0$

1. Choose $\theta \in \Theta := \{\theta_1, \theta_2, \ldots, \theta_K\} \to$ (discretised parameter range)
2. Solve:
$$x(\theta) \in \arg\min_x \; \mathcal{F}(x; f, \theta)$$
3. Do the same for all $\theta_j, j = 1, \ldots, K$
4. Optimise θ w.r.t. to ground truth $\tilde{f}$ and in terms of **task-dependent quality measures** (RMSE, PSNR, SSIM. . .):

$$\arg\min_{\theta\in\Theta} \; RMSE(x(\theta); \tilde{f}), \quad \arg\min_{\theta\in\Theta} \; -PSNR(x(\theta); \tilde{f}), \quad \arg\min_{\theta\in\Theta} \; -SSIM(x(\theta); \tilde{f}) \ldots$$





Exemple of parameter optimisation

What is still done in practice

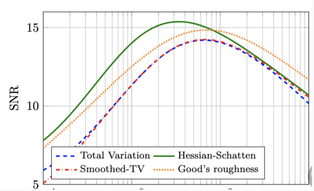
Brute force approach: scalar problem, $\theta > 0$

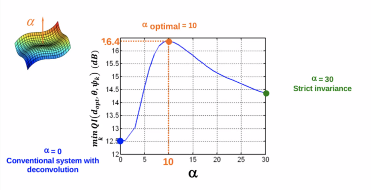
1. Choose $\theta \in \Theta := \{\theta_1, \theta_2, \ldots, \theta_K\} \to$ (discretised parameter range)
2. Solve:

$$x(\theta) \in \arg\min_x \; \mathcal{F}(x; f, \theta)$$

3. Do the same for all $\theta_j, j = 1, \ldots, K$
4. Optimise θ w.r.t. to ground truth $\tilde{f}$ and in terms of **task-dependent quality measures** (RMSE, PSNR, SSIM...):

$$\arg\min_{\theta\in\Theta} \; RMSE(x(\theta); \tilde{f}), \quad \arg\min_{\theta\in\Theta} \; -PSNR(x(\theta); \tilde{f}), \quad \arg\min_{\theta\in\Theta} \; -SSIM(x(\theta); \tilde{f}) \ldots$$





Exemple of parameter optimisation

Motivation for bilevel approaches: can we formalise this idea?

Related and previous approaches

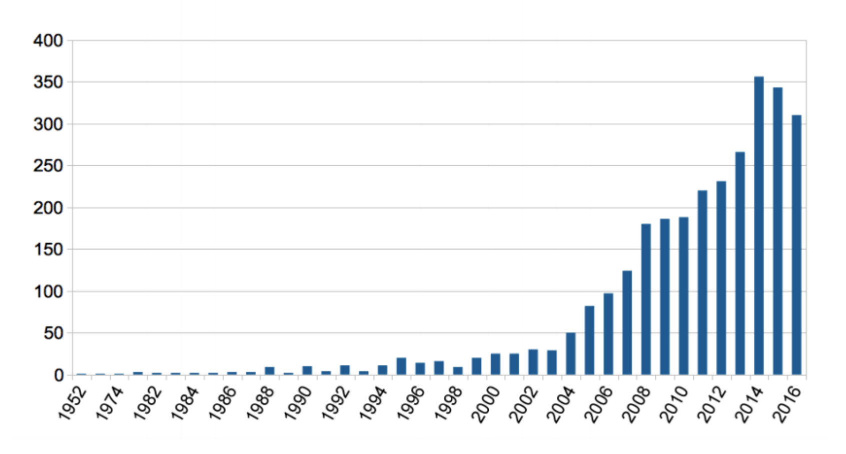
- Optimal model design (Haber, Tenorio, '03, no proofs, many examples):

$$R(x;\theta) = \sum_{i=1}^{n} |\theta_i (Dx)_i|^2, \quad R(x;\theta) = \sum_{i=1}^{n} \theta_i |(Dx)_i|$$

  $\rightarrow$ supervised learning technique effective for learning (parametrised) regularisation models.

- Optimal parameter estimation in the context of Markov Random Field modelling (Samuel, Tappen, '09): lower-level problem defined in terms of log-posterior. . .

Interest in bilevel optimization



Number of works on bilevel optimization over time

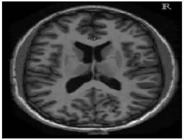
Bilevel modelling

Problem formulation

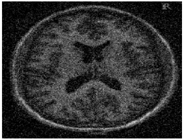
Given **one exemplar** training pair $(\tilde{f}, f) \in \mathbb{R}^n \times \mathbb{R}^m$ with:

- $\tilde{f}$: "ground truth" data, degradation-free example used for training;
- f: corresponding blurred, undersampled, noisy version (in the setting considered)

$$\boxed{f = \mathcal{T}(A\tilde{f})}$$



$\tilde{f}$



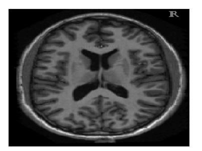
f

Problem formulation

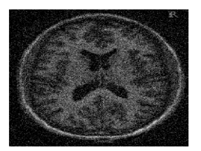
Given **one exemplar** training pair $(\tilde{f}, f) \in \mathbb{R}^n \times \mathbb{R}^m$ with:

- $\tilde{f}$: "ground truth" data, degradation-free example used for training;
- f: corresponding blurred, undersampled, noisy version (in the setting considered)

$$\boxed{f = \mathcal{T}(A\tilde{f})}$$



$\tilde{f}$



f

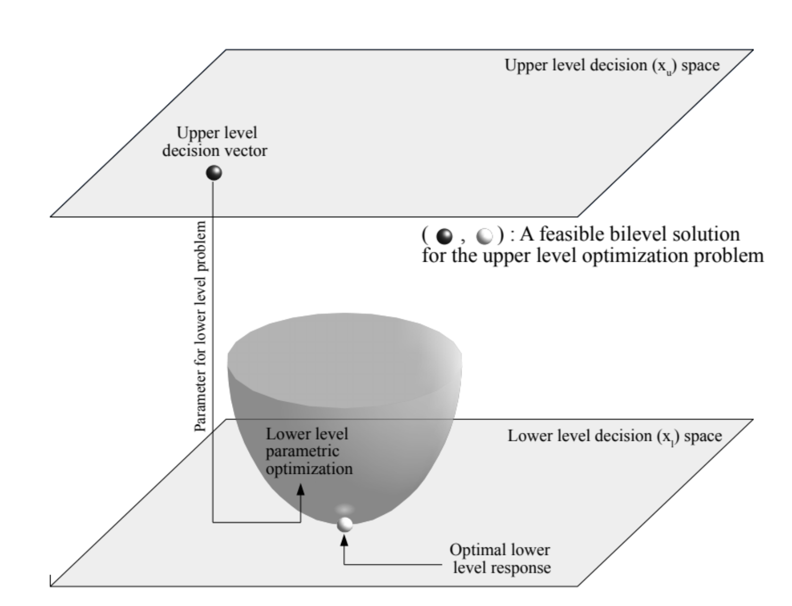
For $q \geq 1$ parameters

$$\boxed{\begin{cases} \min_{\theta \in \mathbb{R}^q_{\geq 0}} \quad \mathcal{E}(x(\theta); \tilde{f}) \\ \text{s.t.} \quad x(\theta) \in \arg\min_{x \in \mathbb{R}^n} \quad \mathcal{F}(x, \theta; f) \end{cases}}$$

- **Upper level functional**: $\mathcal{E}(\cdot; \tilde{f}) : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ (task-dependent)
- **Lower level functional**: $\mathcal{F}(\cdot, \cdot; f) : \mathbb{R}^n \times \mathbb{R}^q_{\geq 0} \to \mathbb{R}_{\geq 0}$ (reconstruction model)

Reference book: J. Bard, *Practical Bilevel Optimization*, '98

Idea



Upper level decision (x_u) space
Upper level decision vector
(,) : A feasible bilevel solution for the upper level optimization problem
Parameter for lower level problem
Lower level parametric optimization
Lower level decision (x_l) space
Optimal lower level response

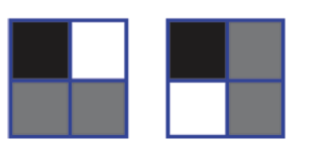
Problem formulation, example I: TV image denoising

TV denoising (Rudin, Osher, Fatemi, '92):

- $A = Id$, $\mathcal{T}(\cdot) = \cdot + n$ with $n \sim \mathcal{N}(0, \sigma^2 Id)$, $q = 1$
- Note: noise level σ^2 **does not need to be known!**
- SNR-like upper level functional $\mathcal{E}(x(\theta); \tilde{f}) = \frac{1}{2}\|x(\theta) - \tilde{f}\|_2^2$ for assessing **reconstruction** ($\sim -SNR$)

$$\begin{cases} \min_{\theta \geq 0} \quad \left\{ \mathcal{E}(x(\theta); \tilde{f}) := \frac{1}{2}\|x(\theta) - \tilde{f}\|_2^2 \right\} \\ \text{s.t.} \quad x(\theta) = \arg\min_{x \in \mathbb{R}^n} \quad \left\{ \mathcal{F}(x, \theta; f) := \theta \|Dx\|_{2,1} + \frac{1}{2}\|x - f\|_2^2 \right\} \end{cases}$$

D is finite difference gradient $(Dx)_{i,j} = (u_{i+1,j} - u_{i,j}, u_{i,j+1} - u_{i,j})$.



First-order derivative filter

Problem formulation, example II: analysis-based priors

Regularisation is chosen as sum of sparsity-promoting terms defined in terms of filters (Elad, Milanfar, Rubinstein, '07), for **fixed** $p \in \{1, 2\}$ (convexity)

$$R(x) = \frac{1}{p} \sum_{k=1}^{q} \theta_k \|K_k x\|_p^p = \frac{1}{p} \sum_{k=1}^{q} \theta_k \sum_{i=1}^{n} |(K_k x)_i|^p$$

- Filter operators $K_k \in \mathbb{R}^{s \times n}$ (generalisation of TV)
- $\theta \in \mathbb{R}^q_{\geq 0}$
- $A = Id$, $\mathcal{T}(\cdot) = \cdot + n$ with $n \sim \mathcal{N}(0, \sigma^2 Id)$.
- Note: noise level σ^2 **does not need to be known!**
- SNR-like $\mathcal{E}(x(\theta); \tilde{f}) = \frac{1}{2}\|x(\theta) - \tilde{f}\|_2^2$, assessing **reconstruction**

$$\begin{cases} \min_{\theta \in \mathbb{R}^q_{\geq 0}} \ \frac{1}{2}\|x(\theta) - \tilde{f}\|_2^2 \\ \text{s.t.} \quad x(\theta) = \arg\min_{x \in \mathbb{R}^n} \ \left\{ \mathcal{F}(x, \theta; f) := \frac{1}{p}\sum_{k=1}^{q} \theta_k \|K_k x\|_p^p + \frac{1}{2}\|x - f\|_2^2 \right\} \end{cases}$$

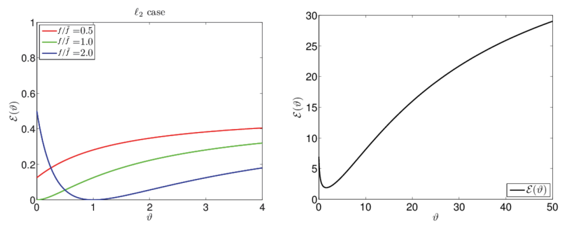
Understanding (non-)convexity of the bilevel problem: scalar example

Set: $x, f, \tilde{f} \in \mathbb{R}$, $q = 1$, $K_1 = 1$, $p = 2$. Explicit solution for lower-level problem!

$$x(\theta) = \arg\min \frac{1}{2}\theta|x|^2 + \frac{1}{2}|x - f|^2 \Rightarrow (\theta + 1)x(\theta) = f$$

Plug into upper level functional:

$$\min_{\theta \geq 0} \left\{ \mathcal{E}(x(\theta)) = \frac{1}{2}|x(\theta) - \tilde{f}|^2 \right\} = \min_{\theta \geq 0} \left\{ \frac{1}{2}\left| \frac{f}{1+\theta} - \tilde{f} \right|^2 = \mathcal{E}(\theta) \right\}$$



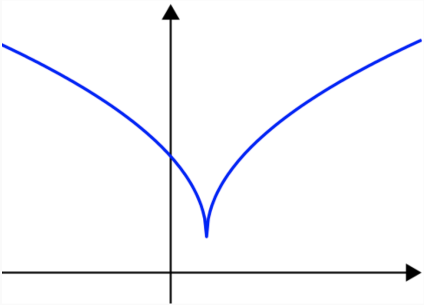
Shape of upper cost functional

Quasi-convexity

For all $a \geq 0$, the set $S_a = \{\theta \geq 0 : \mathcal{E}(\theta) \leq a\}$ is convex $\Leftrightarrow$ $\mathcal{E}(\theta)$ is **quasi-convex**.

Quasi-convexity

- **Convexity$\rightarrow$ quasi-convexity**
- The viceversa is not true:



Quasiconvex but non convex function

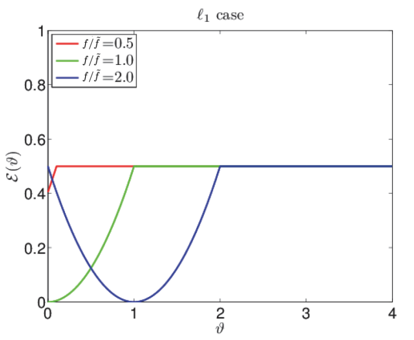
- Better property than concavity: it improves the chance of computing the optimal parameters of the model.

Understanding convexity of the bilevel problem: scalar non-smooth example

Set: $x, f, \tilde{f} \in \mathbb{R}$, $q = 1$, $K_1 = 1$, $p = 1$.

Few calculations show: $x(\theta) = \max(0, |f| - \theta)\text{sign}(f)$. Upper-level functional becomes:

$$\min_{\theta \geq 0} \frac{1}{2} \Big| \max(0, |f| - \theta)\text{sign}(f) - \tilde{f} \Big|^2$$



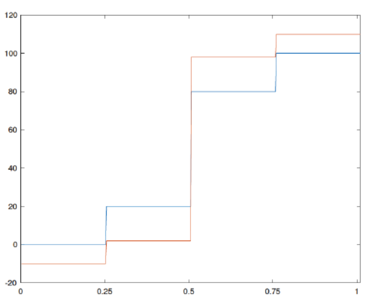
Shape of upper cost functional for 1D quadratic problems

Still quasi-convex.

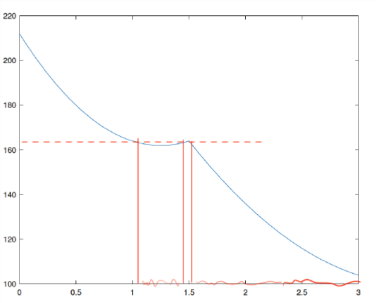
Understanding convexity of the bilevel problem: scalar TV example

Set: $x, f, \tilde{f} \in \mathbb{R}^n$, $q = 1$, $(K_1 x)_i = x_{i+1} - x_i$, $p = 1$.

Computations here are not trivial (exact TV solution in 1D Strong, '96), but one can compute reduced upper level-functional...



$\tilde{f}$ (red), f (blue)



$\mathcal{E}(\theta)$

The functional here is not quasi-convex (Arridge, Maas, Oktem, Schoenlieb, '19)

We will focus on denoising problem
(see later for the reasons why)

Theoretical guarantees

Existence of solution: quadratic case, $p = 2$

For $\theta = (\theta_1, \ldots, \theta_q) \in \mathbb{R}^q_{\geq 0}$, the lower-level problem

$$x(\theta) = \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \; \frac{1}{2} \sum_{k=1}^{q} \theta_k \|K_k x\|_2^2 + \frac{1}{2} \|x - f\|_2^2 \qquad (*)$$

Existence of solution: quadratic case, $p = 2$

For $\theta = (\theta_1, \ldots, \theta_q) \in \mathbb{R}^q_{\geq 0}$, the lower-level problem

$$x(\theta) = \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \ \frac{1}{2} \sum_{k=1}^{q} \theta_k \|K_k x\|_2^2 + \frac{1}{2} \|x - f\|_2^2 \qquad (*)$$

Set $\mathcal{K}_k := K_k^T K_k$, the optimality condition reads:

$$x(\theta) + \sum_{k=1}^{q} \theta_k \mathcal{K}_k x(\theta) = f \quad \Leftrightarrow \quad x(\theta) = \left(Id + \sum_{k=1}^{q} \theta_k \mathcal{K}_k \right)^{-1} f$$

Existence of solution: quadratic case, $p = 2$

For $\theta = (\theta_1, \ldots, \theta_q) \in \mathbb{R}^q_{\geq 0}$, the lower-level problem

$$x(\theta) = \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \; \frac{1}{2} \sum_{k=1}^{q} \theta_k \|K_k x\|_2^2 + \frac{1}{2}\|x - f\|_2^2 \qquad (*)$$

Set $\mathcal{K}_k := K_k^T K_k$, the optimality condition reads:

$$x(\theta) + \sum_{k=1}^{q} \theta_k \mathcal{K}_k x(\theta) = f \quad \Leftrightarrow \quad x(\theta) = \left(Id + \sum_{k=1}^{q} \theta_k \mathcal{K}_k \right)^{-1} f$$

Hence, the reduced upper-level functional is:

$$\min_{\theta \in \mathbb{R}^q_{\geq 0}} \; \frac{1}{2}\|x(\theta) - \tilde{f}\|_2^2 = \frac{1}{2} \left\| \underbrace{\left(Id + \sum_{k=1}^{q} \theta_k \mathcal{K}_k \right)^{-1}}_{=:\mathcal{R}} f - \tilde{f} \right\|_2^2 =: \mathcal{E}(\theta; \tilde{f}) \quad \text{(reduced cost)}$$

Existence + local optimality (Kunisch, Pock, '13)

If $\inf \left\{ \|\tilde{x} - \tilde{f}\| : \tilde{x} \in \ker(K_k) \text{ for some } k \right\} > \|f - \tilde{f}\|_2$, then (*) has solution. Let $\{\mathcal{K}_k \mathcal{R} f\}_{k=1}^q$ be **linearly independent** and let $\theta^* \in \mathbb{R}^q_{\geq 0}$. Then, if

$$\left\| \left(Id + \sum_{k=1}^{q} \theta_k^* \mathcal{K}_k \right)^{-1} f - \tilde{f} \right\|_2$$

is small enough, then, $\nabla^2 \mathcal{E}(\theta^*) > 0$, hence θ^* is a locally unique minimum.

$p = 1$

ℓ_1 type priors have great impact in signal processing/compressed sensing.

$$\begin{cases} \min_{\theta \in \mathbb{R}^q_{\geq 0}} \ \frac{1}{2}\|x(\theta) - \tilde{f}\|_2^2 \\ \text{s.t.} \quad x(\theta) = \arg\min_{x \in \mathbb{R}^n} \ \sum_{k=1}^q \theta_k \|K_k x\|_1 + \frac{1}{2}\|x - f\|_2^2 \end{cases}$$

$p = 1$

ℓ_1 type priors have great impact in signal processing/compressed sensing.

$$\begin{cases} \min_{\theta \in \mathbb{R}^q_{\geq 0}} \ \frac{1}{2}\|x(\theta) - \tilde{f}\|_2^2 \\ \text{s.t.} \quad x(\theta) = \arg\min_{x \in \mathbb{R}^n} \ \sum_{k=1}^q \theta_k \|K_k x\|_1 + \frac{1}{2}\|x - f\|_2^2 \end{cases}$$

Optimality conditions can still be found:

$$x(\theta) \quad \text{s.t.} \quad \begin{cases} \sum_{k=1}^q \theta_k K_k^T \lambda_k + x(\theta) = f, \\ (\lambda_k)_i \in \begin{cases} \operatorname{sgn}(K_k x(\theta))_i & \text{if} \quad (K_k x(\theta))_i \neq 0 \\ [-1, 1] & \text{if} \quad (K_k x(\theta))_i = 0. \end{cases} \end{cases}$$

Existence (Kunisch, Pock, '13)

If $\inf \left\{ \|\tilde{x} - \tilde{f}\| : \tilde{x} \in \ker(K_k) \text{ for some } k \right\} > \|f - \tilde{f}\|_2$, then the lower-level problem has a solution $\theta^* \in \mathbb{R}^q_{\geq 0}$.

$p = 1$

ℓ_1 type priors have great impact in signal processing/compressed sensing.

$$\begin{cases} \min_{\theta \in \mathbb{R}^q_{\geq 0}} \ \frac{1}{2}\|x(\theta) - \tilde{f}\|_2^2 \\ \text{s.t.} \quad x(\theta) = \arg\min_{x \in \mathbb{R}^n} \ \sum_{k=1}^q \theta_k \|K_k x\|_1 + \frac{1}{2}\|x - f\|_2^2 \end{cases}$$

Optimality conditions can still be found:

$$x(\theta) \quad \text{s.t.} \quad \begin{cases} \sum_{k=1}^q \theta_k K_k^T \lambda_k + x(\theta) = f, \\ (\lambda_k)_i \in \begin{cases} \operatorname{sgn}(K_k x(\theta))_i & \text{if} \quad (K_k x(\theta))_i \neq 0 \\ [-1, 1] & \text{if} \quad (K_k x(\theta))_i = 0. \end{cases} \end{cases}$$

Existence (Kunisch, Pock, '13)

If $\inf \left\{ \|\tilde{x} - \tilde{f}\| : \tilde{x} \in \ker(K_k) \text{ for some } k \right\} > \|f - \tilde{f}\|_2$, then the lower-level problem has a solution $\theta^* \in \mathbb{R}^q_{\geq 0}$.

Plugging in upper-level?!

Smoothing: motivation and IFT

$$\begin{cases} \min_{\theta \in \mathbb{R}^q_{\geq 0}} \ \mathcal{E}(x(\theta)) \\ \text{s.t.} \quad x(\theta) \in \arg\min_{x \in \mathbb{R}^n} \ \mathcal{F}(x, \theta) \end{cases}$$

Look for optimality of the bilevel problem by **chain-rule**:

$$\frac{\partial}{\partial \theta} \mathcal{E}(x(\theta)) = \frac{\partial \mathcal{E}}{\partial x}(x(\theta)) \frac{\partial X}{\partial \theta}(\theta)$$

where

$$X : \theta \mapsto x(\theta) \quad \text{is the } \textbf{solution map}$$

Smoothing: motivation and IFT

$$\begin{cases} \min_{\theta \in \mathbb{R}^q_{\geq 0}} \ \mathcal{E}(x(\theta)) \\ \text{s.t.} \quad x(\theta) \in \arg\min_{x \in \mathbb{R}^n} \ \mathcal{F}(x, \theta) \end{cases}$$

Look for optimality of the bilevel problem by **chain-rule**:

$$\frac{\partial}{\partial \theta} \mathcal{E}(x(\theta)) = \frac{\partial \mathcal{E}}{\partial x}(x(\theta)) \frac{\partial X}{\partial \theta}(\theta)$$

where

$$X : \theta \mapsto x(\theta) \quad \text{is the } \textbf{solution map}$$

- **Optimality**: if $\hat{x} = x(\hat{\theta})$ exists, then

$$F(\hat{x}, \hat{\theta}) := \frac{\partial \mathcal{F}}{\partial x}(\hat{x}, \hat{\theta}) = 0$$

Smoothing: motivation and IFT

$$\begin{cases} \min_{\theta \in \mathbb{R}^q_{\geq 0}} \ \mathcal{E}(x(\theta)) \\ \text{s.t.} \quad x(\theta) \in \arg\min_{x \in \mathbb{R}^n} \ \mathcal{F}(x, \theta) \end{cases}$$

Look for optimality of the bilevel problem by **chain-rule**:

$$\frac{\partial}{\partial \theta} \mathcal{E}(x(\theta)) = \frac{\partial \mathcal{E}}{\partial x}(x(\theta)) \frac{\partial X}{\partial \theta}(\theta)$$

where

$$X : \theta \mapsto x(\theta) \quad \text{is the } \textbf{solution map}$$

- **Optimality**: if $\hat{x} = x(\hat{\theta})$ exists, then

$$F(\hat{x}, \hat{\theta}) := \frac{\partial \mathcal{F}}{\partial x}(\hat{x}, \hat{\theta}) = 0$$

- **Implicit function theorem**: if F is C^1 and $\frac{\partial F}{\partial x}(\hat{x}, \hat{\theta})$ is invertible, then there exists $X : \theta \mapsto x(\theta)$ in a neighbourhood of $(\hat{x}, \hat{\theta})$, X is C^1 there and

$$\frac{\partial X}{\partial \theta}(\theta) = \left(-\frac{\partial F}{\partial x}(X(\theta), \theta)\right)^{-1} \frac{\partial F}{\partial \theta}(X(\theta), \theta) = -\left(H_{\mathcal{F}}(X(\theta), \theta)\right)^{-1} \frac{\partial^2 \mathcal{F}}{\partial \theta \partial x}(X(\theta), \theta)$$

where $H_{\mathcal{F}}(X(\theta)) = \partial \mathcal{F} / \partial x^2$ is the Hessian of $\mathcal{F}$.

Smoothing: motivation and IFT

$$\begin{cases} \min_{\theta \in \mathbb{R}^q_{\geq 0}} \ \mathcal{E}(x(\theta)) \\ \text{s.t.} \quad x(\theta) \in \arg\min_{x \in \mathbb{R}^n} \ \mathcal{F}(x, \theta) \end{cases}$$

Look for optimality of the bilevel problem by **chain-rule**:

$$\frac{\partial}{\partial \theta} \mathcal{E}(x(\theta)) = \frac{\partial \mathcal{E}}{\partial x}(x(\theta)) \frac{\partial X}{\partial \theta}(\theta)$$

where

$$X : \theta \mapsto x(\theta) \quad \text{is the } \textbf{solution map}$$

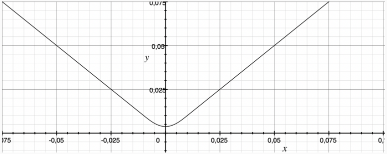
- **Optimality**: if $\hat{x} = x(\hat{\theta})$ exists, then
$$F(\hat{x}, \hat{\theta}) := \frac{\partial \mathcal{F}}{\partial x}(\hat{x}, \hat{\theta}) = 0$$

- **Implicit function theorem**: if F is C^1 and $\frac{\partial F}{\partial x}(\hat{x}, \hat{\theta})$ is invertible, then there exists $X : \theta \mapsto x(\theta)$ in a neighbourhood of $(\hat{x}, \hat{\theta})$, X is C^1 there and
$$\frac{\partial X}{\partial \theta}(\theta) = \left(-\frac{\partial F}{\partial x}(X(\theta), \theta)\right)^{-1} \frac{\partial F}{\partial \theta}(X(\theta), \theta) = -\left(H_{\mathcal{F}}(X(\theta), \theta)\right)^{-1} \frac{\partial^2 \mathcal{F}}{\partial \theta \partial x}(X(\theta), \theta)$$
where $H_{\mathcal{F}}(X(\theta)) = \partial \mathcal{F} / \partial x^2$ is the Hessian of $\mathcal{F}$.

Smoothness of $\mathcal{F}(\cdot, \theta)$ ($\sim C^2$) typically not encountered in applications. . .

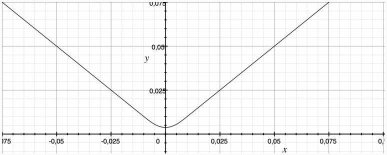
$$\begin{cases} \min_{\theta \in \mathbb{R}^q_{\geq 0}} \ \frac{1}{2}\|x_\varepsilon(\theta) - \tilde{f}\|_2^2 \\ \text{s.t.} \quad x_\varepsilon(\theta) = \arg\min_{x \in \mathbb{R}^n} \ \sum_{k=1}^q \theta_k \sum_{j=1}^n \eta_\varepsilon((K_k x)_j) + \frac{1}{2}\|x - f\|_2^2 \end{cases}$$

where $\varepsilon \ll 1$ and $\eta_\varepsilon(\cdot)$ is C^2, $\eta''_\epsilon \geq 0$.



$$\begin{cases} \min_{\theta\in\mathbb{R}^q_{\geq 0}} & \frac{1}{2}\|x_\varepsilon(\theta) - \tilde{f}\|_2^2 \\ \text{s.t.} & x_\varepsilon(\theta) = \arg\min_{x\in\mathbb{R}^n} \; \sum_{k=1}^q \theta_k \sum_{j=1}^n \eta_\varepsilon((K_k x)_j) + \frac{1}{2}\|x - f\|_2^2 \end{cases}$$

where $\varepsilon \ll 1$ and $\eta_\varepsilon(\cdot)$ is C^2, $\eta_\epsilon'' \geq 0$.



Proposition

There exists a unique solution $x_\varepsilon(\theta)$ of the lower-level problem and the solution map $X_\varepsilon : \theta \mapsto x_\varepsilon(\theta)$ is differentiable for all $\varepsilon > 0$.

...Solutions of bilevel problem?

Adjoint states and optimality system

$$\begin{cases} \min_{\theta \in \mathbb{R}^q_{\geq 0}} \ \frac{1}{2}\|x_\varepsilon(\theta) - \tilde{f}\|_2^2 = \mathcal{E}(x_\varepsilon(\theta)) \\ \text{s.t.} \quad x_\varepsilon(\theta) = \arg\min_{x \in \mathbb{R}^n} \ \sum_{k=1}^q \theta_k \sum_{j=1}^n \eta_\varepsilon((K_k x)_j) + \frac{1}{2}\|x - f\|_2^2 \end{cases}$$

If $\theta_\varepsilon \in \mathbb{R}^q_{\geq 0}$ solves the bilevel problem, then (optimality condition):

$$\frac{\partial}{\partial \theta}\mathcal{E}(x_\varepsilon(\theta_\varepsilon))(\theta - \theta_\varepsilon) = \langle x_\varepsilon(\theta) - \tilde{f}, \frac{\partial X_\varepsilon}{\partial \theta}(\theta_\varepsilon)(\theta - \theta_\varepsilon)\rangle \geq 0 \quad \forall \theta \in \mathbb{R}^q_{\geq 0} \qquad (*)$$

Adjoint states and optimality system

$$\begin{cases} \min_{\theta \in \mathbb{R}^q_{\geq 0}} \ \frac{1}{2}\|x_\varepsilon(\theta) - \tilde{f}\|_2^2 = \mathcal{E}(x_\varepsilon(\theta)) \\ \text{s.t.} \quad x_\varepsilon(\theta) = \arg\min_{x \in \mathbb{R}^n} \ \sum_{k=1}^q \theta_k \sum_{j=1}^n \eta_\varepsilon((K_k x)_j) + \frac{1}{2}\|x - f\|_2^2 \end{cases}$$

If $\theta_\varepsilon \in \mathbb{R}^q_{\geq 0}$ solves the bilevel problem, then (optimality condition):

$$\frac{\partial}{\partial \theta}\mathcal{E}(x_\varepsilon(\theta_\varepsilon))(\theta - \theta_\varepsilon) = \langle x_\varepsilon(\theta) - \tilde{f}, \frac{\partial X_\varepsilon}{\partial \theta}(\theta_\varepsilon)(\theta - \theta_\varepsilon)\rangle \geq 0 \quad \forall \theta \in \mathbb{R}^q_{\geq 0} \qquad (*)$$

Eliminate dependence on $\frac{\partial X_\varepsilon}{\partial \theta}$ by introducing **adjoint state** $p \in \mathbb{R}^n$ which solves:

$$p + \sum_{k=1}^{q} \theta_k K_k^T N_\varepsilon''(K_k x) K_k p = -(x_\varepsilon(\theta_\varepsilon) - \tilde{f}) \qquad (**)$$

Adjoint states and optimality system

$$\begin{cases} \min_{\theta\in\mathbb{R}^q_{\geq 0}} \ \frac{1}{2}\|x_\varepsilon(\theta)-\tilde{f}\|_2^2 = \mathcal{E}(x_\varepsilon(\theta)) \\ \text{s.t.} \quad x_\varepsilon(\theta) = \arg\min_{x\in\mathbb{R}^n} \ \sum_{k=1}^q \theta_k \sum_{j=1}^n \eta_\varepsilon((K_k x)_j) + \frac{1}{2}\|x-f\|_2^2 \end{cases}$$

If $\theta_\varepsilon \in \mathbb{R}^q_{\geq 0}$ solves the bilevel problem, then (optimality condition):

$$\frac{\partial}{\partial\theta}\mathcal{E}(x_\varepsilon(\theta_\varepsilon))(\theta-\theta_\varepsilon) = \langle x_\varepsilon(\theta)-\tilde{f}, \frac{\partial X_\varepsilon}{\partial\theta}(\theta_\varepsilon)(\theta-\theta_\varepsilon)\rangle \geq 0 \quad \forall\theta\in\mathbb{R}^q_{\geq 0} \qquad (*)$$

Eliminate dependence on $\frac{\partial X_\varepsilon}{\partial\theta}$ by introducing **adjoint state** $p\in\mathbb{R}^n$ which solves:

$$p + \sum_{k=1}^q \theta_k K_k^T N''_\varepsilon(K_k x) K_k p = -(x_\varepsilon(\theta_\varepsilon)-\tilde{f}) \qquad (**)$$

Smoothed optimality system (necessary condition)

For $N'_\varepsilon(z) = (\eta'_\varepsilon(z_j))_j^T \in \mathbb{R}^s$ and $N''_\varepsilon(z) = \operatorname{diag}\left((\eta''_\varepsilon(z_j))_j\right) \in \mathbb{R}^{s\times s}$

$$\begin{cases} x_\varepsilon + \sum_{k=1}^q \theta_{\varepsilon,k} K_k^T N'_\varepsilon(K_k x_\varepsilon) = f & \text{(optimality l.l.)} \\ p_\varepsilon + \sum_{k=1}^q \theta_{\varepsilon,k} K_k^T N''_\varepsilon(K_k x_\varepsilon) K_k p_\varepsilon = -(x_\varepsilon - \tilde{f}) & \text{(adjoint eq. } \to \text{ linear)} \\ \langle N'_\varepsilon(K_k x_\varepsilon), K_k p_\varepsilon\rangle(\theta_k - \theta_{\varepsilon,k}) \geq 0, \quad \forall\theta_k\geq 0, \quad k=1,\dots,q & \text{(optimality bilevel)} \end{cases} \qquad \text{(OC)}$$

Adjoint states and optimality system

$$\begin{cases} \min_{\theta\in\mathbb{R}^q_{\geq 0}} \ \frac{1}{2}\|x_\varepsilon(\theta)-\tilde{f}\|_2^2 = \mathcal{E}(x_\varepsilon(\theta)) \\ \text{s.t.} \quad x_\varepsilon(\theta) = \arg\min_{x\in\mathbb{R}^n} \ \sum_{k=1}^q \theta_k \sum_{j=1}^n \eta_\varepsilon((K_k x)_j) + \frac{1}{2}\|x-f\|_2^2 \end{cases}$$

If $\theta_\varepsilon \in \mathbb{R}^q_{\geq 0}$ solves the bilevel problem, then (optimality condition):

$$\frac{\partial}{\partial\theta}\mathcal{E}(x_\varepsilon(\theta_\varepsilon))(\theta-\theta_\varepsilon) = \langle x_\varepsilon(\theta)-\tilde{f}, \frac{\partial X_\varepsilon}{\partial\theta}(\theta_\varepsilon)(\theta-\theta_\varepsilon)\rangle \geq 0 \quad \forall\theta\in\mathbb{R}^q_{\geq 0} \qquad (*)$$

Eliminate dependence on $\frac{\partial X_\varepsilon}{\partial\theta}$ by introducing **adjoint state** $p\in\mathbb{R}^n$ which solves:

$$p + \sum_{k=1}^q \theta_k K_k^T N_\varepsilon''(K_k x) K_k p = -(x_\varepsilon(\theta_\varepsilon)-\tilde{f}) \qquad (**)$$

Smoothed optimality system (necessary condition)

For $N_\varepsilon'(z) = (\eta_\varepsilon'(z_j))_j^T \in \mathbb{R}^s$ and $N_\varepsilon''(z) = \operatorname{diag}\left((\eta_\varepsilon''(z_j))_j\right) \in \mathbb{R}^{s\times s}$

$$\begin{cases} x_\varepsilon + \sum_{k=1}^q \theta_{\varepsilon,k} K_k^T N_\varepsilon'(K_k x_\varepsilon) = f & \text{(optimality l.l.)} \\ p_\varepsilon + \sum_{k=1}^q \theta_{\varepsilon,k} K_k^T N_\varepsilon''(K_k x_\varepsilon) K_k p_\varepsilon = -(x_\varepsilon - \tilde{f}) & \text{(adjoint eq. } \to \text{linear)} \\ \langle N_\varepsilon'(K_k x_\varepsilon), K_k p_\varepsilon\rangle(\theta_k - \theta_{\varepsilon,k}) \geq 0, \quad \forall\theta_k\geq 0, \quad k=1,\dots,q & \text{(optimality bilevel)} \end{cases} \qquad \text{(OC)}$$

Limit and convergence results for $\varepsilon\to 0$: very technical (Kunisch, Pock, '13).

Algorithmic approaches (nods)

Newton-type algorithms for solving smoothed optimality system

Choose $\varepsilon \ll 1$ to well-approximate the non-smooth behaviour.

Idea: use fast optimisation approaches ($\sim$ Newton's method) to solve the bilevel problem **by solving** the optimality system.

Newton-type algorithms for solving smoothed optimality system

Choose $\varepsilon \ll 1$ to well-approximate the non-smooth behaviour.

Idea: use fast optimisation approaches ($\sim$ Newton's method) to solve the bilevel problem **by solving** the optimality system.

Complementarity condition for getting 2 equations (depending on μ) from inequality

$$\mu - \max(0, \mu - \theta) = 0, \qquad \left(\left(\langle N_\varepsilon'(K_1 x), K_1 p\rangle, \ldots, \langle N_\varepsilon'(K_q x), K_q p\rangle\right)\right)^T - \mu = 0,$$

Newton-type algorithms for solving smoothed optimality system

Choose $\varepsilon \ll 1$ to well-approximate the non-smooth behaviour.

Idea: use fast optimisation approaches ($\sim$ Newton's method) to solve the bilevel problem **by solving** the optimality system.

Complementarity condition for getting 2 equations (depending on μ) from inequality

Finally, end up with a problem expressed in the form of linear system:

$$\text{find } x_\varepsilon \in \mathbb{R}^n, \theta_\varepsilon \in \mathbb{R}^q_{\geq 0}, p_\varepsilon \in \mathbb{R}^n, \mu_\varepsilon \in \mathbb{R}^q : G(x_\varepsilon, \theta_\varepsilon, p_\varepsilon, \mu_\varepsilon) = 0.$$

Algorithm: Newton-type bilevel learning

inputs: $(x^0, \theta^0, p^0, \mu^0)$

1. Solve: $J(x^n, \theta^n, p^n, \mu^n) \begin{bmatrix} \delta x \\ \delta\theta \\ \delta p \\ \delta\mu \end{bmatrix} = -G(x^n, \theta^n, p^n, \mu^n)$
2. Update $(x^{n+1}, \theta^{n+1}, p^{n+1}, \mu^{n+1}) = (x^n, \theta^n, p^n, \mu^n) + (\delta x, \delta\theta, \delta p, \delta\mu)$ (+ linesearch)
3. Iterate

Newton-type algorithms for solving smoothed optimality system

Choose $\varepsilon \ll 1$ to well-approximate the non-smooth behaviour.

Idea: use fast optimisation approaches ($\sim$ Newton's method) to solve the bilevel problem **by solving** the optimality system.

Complementarity condition for getting 2 equations (depending on μ) from inequality

Finally, end up with a problem expressed in the form of linear system:

$$\text{find } x_\varepsilon \in \mathbb{R}^n, \theta_\varepsilon \in \mathbb{R}^q_{\geq 0}, p_\varepsilon \in \mathbb{R}^n, \mu_\varepsilon \in \mathbb{R}^q : G(x_\varepsilon, \theta_\varepsilon, p_\varepsilon, \mu_\varepsilon) = 0.$$

Algorithm: Newton-type bilevel learning

inputs: $(x^0, \theta^0, p^0, \mu^0)$

1. Solve: $J(x^n, \theta^n, p^n, \mu^n) \begin{bmatrix} \delta x \\ \delta\theta \\ \delta p \\ \delta\mu \end{bmatrix} = -G(x^n, \theta^n, p^n, \mu^n)$
2. Update $(x^{n+1}, \theta^{n+1}, p^{n+1}, \mu^{n+1}) = (x^n, \theta^n, p^n, \mu^n) + (\delta x, \delta\theta, \delta p, \delta\mu)$ (+ linesearch)
3. Iterate

Proposition

Upon suitable regularity conditions & initialisation (θ_0, x_0) around a stationary point $G(x_\varepsilon, \theta_\varepsilon, p_\varepsilon, \mu_\varepsilon) = 0$, then the algorithm converges locally superlinearly .

Taking a step back: optimisation VS. discretisation

We started from a finite-dimensional formulation of the problem, then we designed suitable algorithms. Could we do the reverse (to make Pierre happy)?

Taking a step back: optimisation VS. discretisation

We started from a finite-dimensional formulation of the problem, then we designed suitable algorithms. Could we do the reverse (to make Pierre happy)?

First discretise, then optimise

- Replace function spaces with finite-dimensional subsets
- Constraints are discretised too
- Integrals become sums, duality is defined in terms of discrete surrogates…

Pro: easy to explain/design. **Con**: mesh-dependency due to choice of discretisation.

Taking a step back: optimisation VS. discretisation

We started from a finite-dimensional formulation of the problem, then we designed suitable algorithms. Could we do the reverse (to make Pierre happy)?

First discretise, then optimise

- Replace function spaces with finite-dimensional subsets
- Constraints are discretised too
- Integrals become sums, duality is defined in terms of discrete surrogates...

Pro: easy to explain/design. **Con**: mesh-dependency due to choice of discretisation.

First optimise, then discretise

- Carry out analysis using variational calculus/PDE tools
- Related to PDE-constrained optimisation

Pro: better understanding of regularity, mesh independence. **Con**: analysis requires technical tools

Hinze, Rosch, '11

Bilevel problem in infinite-dimensional setting

For $\mathcal{X}, \mathcal{H}$ suitable (reflexive Banach) function spaces consider:

$$\begin{cases} \min_{\theta \in \mathcal{X}} \ \mathcal{E}(x(\theta); \tilde{f}) \\ \text{s.t.} \quad x(\theta) \in \arg\min_{x \in \mathcal{H}} \ \mathcal{F}(x, \theta) \end{cases}$$

Examples:

-
-
-

Bilevel problem in infinite-dimensional setting

For $\mathcal{X}, \mathcal{H}$ suitable (reflexive Banach) function spaces consider:

$$\begin{cases} \min_{\theta \geq 0} \ \frac{1}{2}\|x(\theta) - \tilde{f}\|_{L^2}^2 \\ \text{s.t.} \quad x(\theta) \in \arg\min_{x \in BV(\Omega)} \ TV(x) + \frac{\theta}{2}\|x - f\|_{L^2}^2 \end{cases}$$

Examples:

- TV denoising (Schoenlieb, De Los Reyes, '13): bad regularity properties of the solution map $X : \theta \mapsto x(\theta)$ due to BV topology...
-
-

Bilevel problem in infinite-dimensional setting

For $\mathcal{X}, \mathcal{H}$ suitable (reflexive Banach) function spaces consider:

$$\begin{cases} \min_{\theta \in \mathbb{R}^d_{\geq 0}} \ \frac{1}{2}\|x(\theta) - \tilde{f}\|^2_{L^2} \\ \text{s.t.} \quad x(\theta) \in \arg\min_{x \in H^1(\Omega)} \ TV_\varepsilon(x) + \frac{\epsilon}{2}\|\nabla x\|^2_{L^2} + \sum_{i=1}^d \theta_i \Phi_i(x; f) \end{cases}$$

Examples:

- TV denoising (Schoenlieb, De Los Reyes, '13): bad regularity properties of the solution map $X : \theta \mapsto x(\theta)$ due to BV topology...
- Smoothed TV denoising (with general fidelities) (Calatroni, Chung, De Los Reyes, Schoenlieb, Valkonen, '15) in Hilbert scenarios
-

Bilevel problem in infinite-dimensional setting

For $\mathcal{X}, \mathcal{H}$ suitable (reflexive Banach) function spaces consider:

$$\begin{cases} \min_{\theta \in \mathbb{R}^d_{\geq 0}} \; \frac{1}{2}\|x(\theta) - \tilde{f}\|^2_{L^2} \\ \text{s.t.} \quad x(\theta) \in \arg\min_{x \in H^1(\Omega)} \; TGV^{\theta}_{\varepsilon}(x) + \frac{\epsilon}{2}\|\nabla x\|^2_{L^2} + \frac{1}{2}\|x - f\|^2_{L^2} \end{cases}$$

Examples:

- TV denoising (Schoenlieb, De Los Reyes, '13): bad regularity properties of the solution map $X : \theta \mapsto x(\theta)$ due to BV topology...
- Smoothed TV denoising (with general fidelities) (Calatroni, Chung, De Los Reyes, Schoenlieb, Valkonen, '15) in Hilbert scenarios
- Higher-order regularisation $\theta = (\alpha, \beta)$ (Valkonen, De Los Reyes, Schoenlieb, '17, Hintermueller et al. '17-...)

Bilevel problem in infinite-dimensional setting

For $\mathcal{X}, \mathcal{H}$ suitable (reflexive Banach) function spaces consider:

$$\begin{cases} \min_{\theta \in \mathbb{R}^d_{\geq 0}} \ \frac{1}{2}\|x(\theta) - \tilde{f}\|^2_{L^2} \\ \text{s.t.} \quad x(\theta) \in \arg\min_{x \in H^1(\Omega)} \ TGV^{\theta}_{\varepsilon}(x) + \frac{\epsilon}{2}\|\nabla x\|^2_{L^2} + \frac{1}{2}\|x - f\|^2_{L^2} \end{cases}$$

Examples:

- TV denoising (Schoenlieb, De Los Reyes, '13): bad regularity properties of the solution map $X : \theta \mapsto x(\theta)$ due to BV topology...
- Smoothed TV denoising (with general fidelities) (Calatroni, Chung, De Los Reyes, Schoenlieb, Valkonen, '15) in Hilbert scenarios
- Higher-order regularisation $\theta = (\alpha, \beta)$ (Valkonen, De Los Reyes, Schoenlieb, '17, Hintermueller et al. '17-...)

General idea: use optimality condition (elliptic PDE, thanks to Hilbert) of lower-level

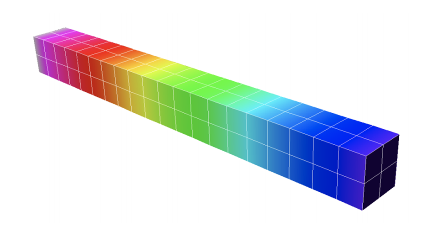
$$e(x, \theta) = 0$$

Then, use variational calculus to derive (semismooth, quasi-)Newton-type schemes.

AkA: PDE-constrained optimisation (theory Hinze, Pinnau, Ulbrich, Ulbrich, '10, numerical + codes: De Los Reyes, '15)

Optimal control/design

Problem: Given a metal bar, generate temperature distribution x depending on forcing term/boundary conditions θ closest to a given reference $\tilde{f}$.



$$\begin{cases} \min_\theta & \frac{1}{2}\|x(\theta) - \tilde{f}\|^2 + \frac{\beta}{2}\|\theta\|^2 \\ \text{s.t.} & x(\theta) \quad \text{solves} \quad \Delta x = \theta \\ (\text{or } x = \theta \text{ on } \partial\Omega) & \end{cases}$$

PDE-constrained optimisation/optimal control problems:

- Regularity/topological assumptions
- Refined notions of differentiability (B-ouligand, Mordukhovich,. . .) for dealing with non-smoothness
- Similar ideas as in discrete case (smoothing, adjoint systems, Newton-type approaches. . .)

Bilevel learning of noise modelling

Optimal bilevel Gaussian image denoising

$$f = \tilde{f} + n, \qquad n \sim \mathcal{N}(0, \sigma^2 Id)$$

Scalar bilevel learning from one image pair[3], $\varepsilon \ll 1$:

$$\begin{cases} \min_{\theta \geq 0} \ \frac{1}{2}\|x_\varepsilon(\theta) - \tilde{f}\|_2^2 \\ \text{s.t.} \quad x_\varepsilon(\theta) = \arg\min_{x \in \mathbb{R}^n} \ \|Dx\|_{2,1,\varepsilon} + \frac{\theta}{2}\|x - f\|_2^2 \end{cases}$$

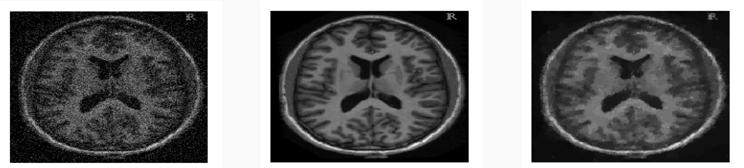
[3] De Los Reyes, Schoenlieb, '13

Optimal bilevel Gaussian image denoising

$$f = \tilde{f} + n, \qquad n \sim \mathcal{N}(0, \sigma^2 Id)$$

Scalar bilevel learning from one image pair[3], $\varepsilon \ll 1$:

$$\begin{cases} \min_{\theta \geq 0} \ \frac{1}{2}\|x_\varepsilon(\theta) - \tilde{f}\|_2^2 \\ \text{s.t.} \quad x_\varepsilon(\theta) = \arg\min_{x \in \mathbb{R}^n} \ \|Dx\|_{2,1,\varepsilon} + \frac{\theta}{2}\|x - f\|_2^2 \end{cases}$$



f $\qquad$ $\tilde{f}$ $\qquad$ $x(\hat{\theta})$

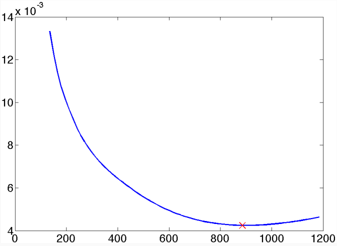
[3] De Los Reyes, Schoenlieb, '13

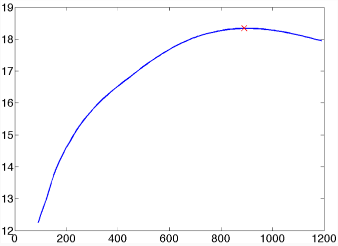
Optimal bilevel Gaussian image denoising

$$f = \tilde{f} + n, \qquad n \sim \mathcal{N}(0, \sigma^2 Id)$$

Scalar bilevel learning from one image pair[3], $\varepsilon \ll 1$:

$$\begin{cases} \min_{\theta \geq 0} \ \frac{1}{2}\|x_\varepsilon(\theta) - \tilde{f}\|_2^2 \\ \text{s.t.} \quad x_\varepsilon(\theta) = \arg\min_{x \in \mathbb{R}^n} \ \|Dx\|_{2,1,\varepsilon} + \frac{\theta}{2}\|x - f\|_2^2 \end{cases}$$





Cost VS. SNR

[3] De Los Reyes, Schoenlieb, '13

Optimal bilevel Poisson image denoising

$$f = \mathcal{P}(\tilde{f}) \sim \mathrm{Poiss}(\tilde{f})$$

Scalar bilevel learning from one image pair[4], $\varepsilon, \delta \ll 1$ and $\rho \gg 1$ (positivity penalty)

$$\begin{cases} \min_{\theta \geq 0} \ \frac{1}{2}\|x(\theta) - \tilde{f}\|_2^2 \\ \text{s.t.} \quad x_\varepsilon(\theta) = \arg\min_{x \in \mathbb{R}^n} \ \|Dx\|_{2,1,\varepsilon} + \frac{\theta}{2}\sum_{i=1}^n (x_i - f_i \log x_i) + \frac{\rho}{2}\|\min(\delta, x)\|_2^2 \end{cases}$$

[4] Calatroni, Chung, De Los Reyes, Schoenlieb, Valkonen, '16

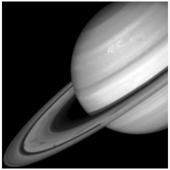
Optimal bilevel Poisson image denoising

$$\boxed{f = \mathcal{P}(\tilde{f}) \sim \text{Poiss}(\tilde{f})}$$

Scalar bilevel learning from one image pair[4], $\varepsilon, \delta \ll 1$ and $\rho \gg 1$ (positivity penalty)

$$\begin{cases} \min_{\theta \geq 0} \ \frac{1}{2} \| x(\theta) - \tilde{f} \|_2^2 \\ \text{s.t.} \quad x_\varepsilon(\theta) = \arg\min_{x \in \mathbb{R}^n} \ \| Dx \|_{2,1,\varepsilon} + \frac{\theta}{2} \sum_{i=1}^n (x_i - f_i \log x_i) + \frac{\rho}{2} \| \min(\delta, x) \|_2^2 \end{cases}$$

f



$\tilde{f}$

$x_\varepsilon(\hat{\theta})$

[4] Calatroni, Chung, De Los Reyes, Schoenlieb, Valkonen, '16

Large training sets for robustness

Heuristic

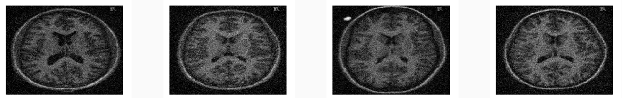
The quality of imaging measurements depends on the experimental setup. A training set $(f_k, \tilde{f}_k),\ k = 1, \ldots, K$ can be provided using (simulated) **phantoms**. Then, the estimated optimal $\theta \in \mathbb{R}^q_{\geq 0}$ can be applied to restore unseen data acquired within the same standard setup.

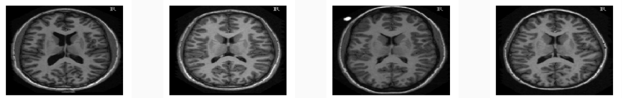
Large training sets for robustness

Heuristic

The quality of imaging measurements depends on the experimental setup. A training set $(f_k, \tilde{f}_k)$, $k = 1, \dots, K$ can be provided using (simulated) **phantoms**. Then, the estimated optimal $\theta \in \mathbb{R}^q_{\geq 0}$ can be applied to restore unseen data acquired within the same standard setup.

- f_k: imperfect noisy data (standard clinical acquisition time)
- $\tilde{f}_k$: (approximation of) ground truth (longer acquisition time)

$f_k =$  $\dots$

$\tilde{f}_k =$  $\dots$

Simulated data from OASIS online database.

Bilevel problem: multiple constraints case

$$\begin{cases} \min_{\theta \geq 0} \; \frac{1}{2K} \sum_{i=1}^{K} \|x_\varepsilon^i(\theta) - \tilde{f}_i\|_2^2 \\ \text{s.t.} \quad x_\varepsilon^i(\theta) = \arg\min_{x \in \mathbb{R}^n} \; \|Dx\|_{2,1,\varepsilon} + \frac{\theta}{2}\|x - f_i\|_2^2, \quad i = 1, \ldots, K \end{cases}$$

Large-scale opt. problem: need to solve $K \gg 1$ (parallel) lower-level problems...

[5] Byrd, Nocedal et al. '13

Bilevel problem: multiple constraints case

$$\begin{cases} \min_{\theta\geq 0} \ \frac{1}{2K}\sum_{i=1}^{K}\|x_\varepsilon^i(\theta)-\tilde{f}_i\|_2^2 \\ \text{s.t.} \quad x_\varepsilon^i(\theta) = \arg\min_{x\in\mathbb{R}^n} \ \|Dx\|_{2,1,\varepsilon} + \frac{\theta}{2}\|x-f_i\|_2^2, \quad i=1,\dots,K \end{cases}$$

Large-scale opt. problem: need to solve $K \gg 1$ (parallel) lower-level problems...

- **Stochastic optimisation approaches**: adapt the size $|S|$ of the training set *dynamically*, balancing convergence VS. accuracy[5]
- **Parallelisation**
- **Batch gradients** evaluated on $S \subset \{1,\dots,K\}$ for computation of $\hat{\theta}_S \geq 0$

K	$\hat{\theta}_S$	$\lvert S_0\rvert$	$\lvert S_{end}\rvert$	eff.	eff. Dyn. S.	diff.
10	86.31	2	7	180	**70**	5.2%
20	90.61	4	6	920	**180**	5.3%
30	94.36	6	7	2100	**314**	5.6%
40	88.88	8	8	880	**496**	1.2%
50	88.92	10	10	2200	**560**	$<1\%$
60	89.64	12	12	1920	**336**	1.9%
70	86.09	14	14	2940	**532**	3.3%
80	87.68	16	16	3520	**448**	$<1\%$

[5] Byrd, Nocedal et al. '13

Learning the noise model: multiple fidelities

Problem: often times, the noise distribution is unknown.

Idea: feed the model with possibly $d > 1$ data terms (associated each to single noise models) and perform bilevel estimation.

Due to the fact that we may have $\theta_i = 0$ for some $i = 1, \ldots, d$, only "active" noise data terms will be selected.

Learning the noise model: multiple fidelities

Problem: often times, the noise distribution is unknown.

Idea: feed the model with possibly $d > 1$ data terms (associated each to single noise models) and perform bilevel estimation.

Due to the fact that we may have $\theta_i = 0$ for some $i = 1, \ldots, d$, only "active" noise data terms will be selected.

$$\begin{cases} \min_{\theta \in \mathbb{R}^d_{\geq 0}} \ \frac{1}{2}\|x_\varepsilon(\theta) - \tilde{f}\|_2^2 \\ \text{s.t.} \quad x_\varepsilon(\theta) = \arg\min_{x \in \mathbb{R}^n} \ \|Dx\|_{2,1,\varepsilon} + \sum_{i=1}^d \theta_i \Phi_i(x; f) \end{cases}$$

$$\Phi_i(x; f) = \begin{cases} \frac{1}{2}\|x - f\|_2^2 & \text{Gaussian}, \\ \|x - f\|_1 & \text{Laplace/S.P} \to \text{smoothed}, \\ \sum_{j=1}^n (x_j - f_j \log x_j) + \iota_{\geq 0}(x) & \text{Poisson}, \\ \ldots & \text{other}, \end{cases}$$

Learning the noise model: multiple fidelities

Problem: often times, the noise distribution is unknown.

Idea: feed the model with possibly $d > 1$ data terms (associated each to single noise models) and perform bilevel estimation.

Due to the fact that we may have $\theta_i = 0$ for some $i = 1, \dots, d$, only "active" noise data terms will be selected.

$$\begin{cases} \min_{\theta \in \mathbb{R}^d_{\geq 0}} \ \frac{1}{2}\|x_\varepsilon(\theta) - \tilde{f}\|_2^2 \\ \text{s.t.} \quad x_\varepsilon(\theta) = \arg\min_{x \in \mathbb{R}^n} \ \|Dx\|_{2,1,\varepsilon} + \sum_{i=1}^d \theta_i \Phi_i(x; f) \end{cases}$$

$$\Phi_i(x; f) = \begin{cases} \frac{1}{2}\|x - f\|_2^2 & \text{Gaussian,} \\ \|x - f\|_1 & \text{Laplace/S.P} \to \text{smoothed,} \\ \sum_{j=1}^n (x_j - f_j \log x_j) + \iota_{\geq 0}(x) & \text{Poisson,} \\ \dots & \text{other,} \end{cases}$$

Does the estimation respect the noise mixture/predominance?

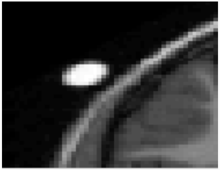
Learning the noise model: Gaussian + impulsive noise

$$\begin{cases} \min_{(\theta_1,\theta_2)\in\mathbb{R}^2_{\geq 0}} \quad \frac{1}{2K}\sum_{i=1}^{K} \|x^i_\varepsilon(\theta) - \tilde{f}^i\|_2^2 \\ \text{s.t.} \quad x^i_\varepsilon(\theta) = \arg\min_{x\in\mathbb{R}^n} \quad \|Dx\|_{2,1,\varepsilon} + \theta_1 \sum_{j=1}^{n} \eta_\varepsilon(x_j - f^i_j) + \frac{\theta_2}{2}\|x - f^i\|_2^2 \end{cases}$$

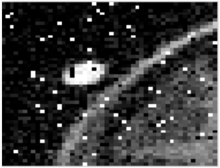
Learning the noise model: Gaussian + impulsive noise

$$
\begin{cases}
\min_{(\theta_1,\theta_2)\in\mathbb{R}^2_{\geq 0}} \ \frac{1}{2K}\sum_{i=1}^{K} \|x_\varepsilon^i(\theta) - \tilde{f}^i\|_2^2 \\
\text{s.t.} \quad x_\varepsilon^i(\theta) = \arg\min_{x\in\mathbb{R}^n} \ \|Dx\|_{2,1,\varepsilon} + \theta_1 \sum_{j=1}^{n} \eta_\varepsilon(x_j - f_j^i) + \frac{\theta_2}{2}\|x - f^i\|_2^2
\end{cases}
$$

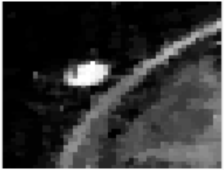
Actual mixed noise scenario:



f



$\tilde{f}$

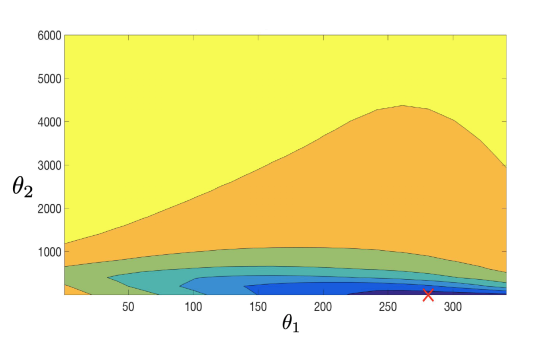


$x_\varepsilon(\hat{\theta})$

Learning the noise model: Gaussian + impulsive noise

$$\begin{cases} \min_{(\theta_1,\theta_2)\in\mathbb{R}^2_{\geq 0}} & \frac{1}{2K}\sum_{i=1}^K \|x_\varepsilon^i(\theta)-\tilde{f}^i\|_{1,\delta} \\ \text{s.t.} & x_\varepsilon^i(\theta) = \arg\min_{x\in\mathbb{R}^n} \|Dx\|_{2,1,\varepsilon} + \theta_1 \sum_{j=1}^n \eta_\varepsilon(x_j - f_j^i) + \frac{\theta_2}{2}\|x-f^i\|_2^2 \end{cases}$$

"Blind" denoising test: only impulsive noise in the data. Can we discriminate it?



$\hat{\theta}_1 \neq 0$, $\hat{\theta}_2 \approx 0$. Smoothed ℓ_1 cost.

Bilevel learning of regularisation models

Analysis approach

Task: learn optimal regularisation parameters $\theta = (\theta_1, \dots, \theta_q) \in \mathbb{R}^q_{\geq 0}$ weighting (many) regularisation filters $K_k \in \mathbb{R}^{s \times n}$

$$\begin{cases} \min_{\theta \in \mathbb{R}^q_{\geq 0}} \ \frac{1}{2K} \sum_{i=1}^K \|x_\varepsilon^i(\theta) - \tilde{f}^i\|_2^2 \\ \text{s.t.} \quad x_\varepsilon^i(\theta) = \arg\min_{x \in \mathbb{R}^n} \ \displaystyle\sum_{k=1}^q \theta_k \sum_{j=1}^n \eta_\varepsilon((K_k x)_j) + \frac{1}{2}\|x - f^i\|_2^2, \quad i = 1, \dots, K \end{cases}$$

Analysis approach

Task: learn optimal regularisation parameters $\theta = (\theta_1, \ldots, \theta_q) \in \mathbb{R}^q_{\geq 0}$ weighting (many) regularisation filters $K_k \in \mathbb{R}^{s \times n}$

$$\begin{cases} \min_{\theta \in \mathbb{R}^q_{\geq 0}} \ \frac{1}{2K} \sum_{i=1}^K \|x_\varepsilon^i(\theta) - \tilde{f}^i\|_2^2 \\ \text{s.t.} \quad x_\varepsilon^i(\theta) = \arg\min_{x \in \mathbb{R}^n} \ \displaystyle\sum_{k=1}^q \theta_k \sum_{j=1}^n \eta_\varepsilon((K_k x)_j) + \frac{1}{2}\|x - f^i\|_2^2, \quad i = 1, \ldots, K \end{cases}$$

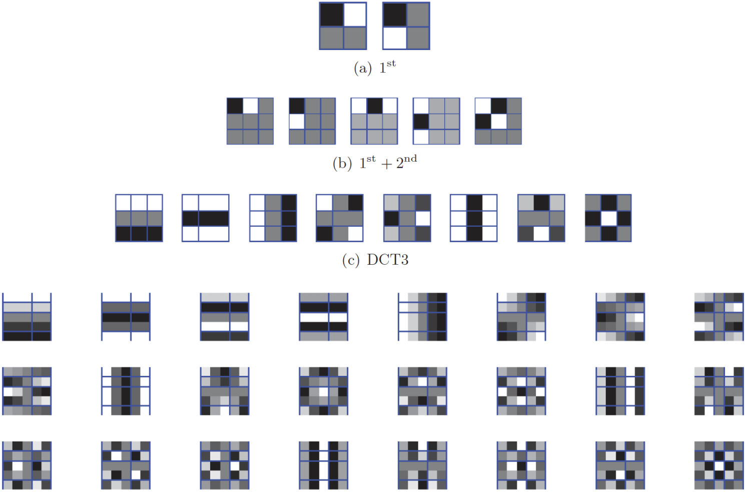
Convolutional structure of filters employed:

$$K_k \mathrm{vec}(x) = \kappa_k * \mathrm{mat}(x)$$

Examples:

- Standard discretisation of 1st and 2nd order derivatives via finite differences
- Higher-order linear operators obtained from 2D DCT basis (from JPEG compression problems, it is known that they provide a sparse representation of the image)
- Can be seen as a generalisation of TV

Filters


(a) 1^{st}

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

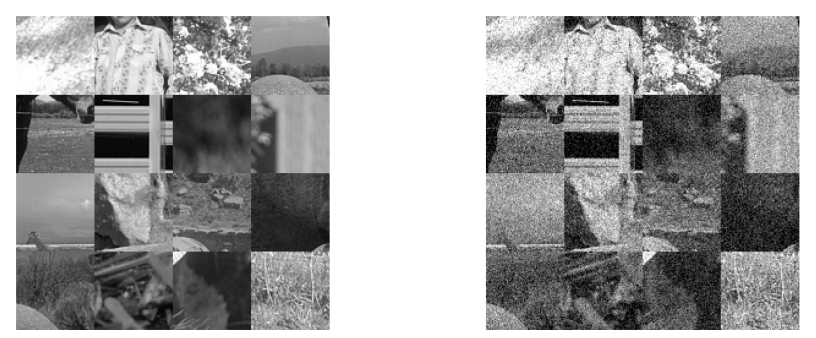
(c) DCT3

(d) DCT5

Filters κ_k

Dataset and results

Training Dataset: Sample of 64 patches of size 64×64 from Berkley dataset[6] + AWGN noise of different intensities $\sigma \in \{15, 25, 50\}$.



Images $(\tilde{f}, f)$ obtained by adding AWGN with $\sigma = 25$.

[6] https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/

Dataset and results

	$\sigma = 15$		$\sigma = 25$		$\sigma = 50$	
Filters	k	$\mathcal{E}(\vartheta)$	k	$\mathcal{E}(\vartheta)$	k	$\mathcal{E}(\vartheta)$
1^{st}	8	162.87	24	302.69	16	601,88
$1^{st} + 2^{nd}$	18	152.45	33	282.02	43	562.44
DCT3	12	147.55	20	270.62	37	542.90
DCT5	16	$\boxed{144.69}$	44	$\boxed{265.41}$	100	$\boxed{525.97}$

Number of Newton steps and value of $\mathcal{E}(\theta)$.

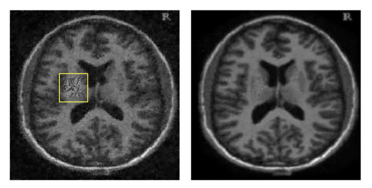
- Lowest energy achieved for very diverse filter banks (DCT5)
- Adding 2nd order information improves significantly TV-type results (known idea of higher-order regularisations)
- Different test on piecewise constant dataset: simple (1st) filters preferred over more complex ones: TV is a good choice for these data.

Learning better models

1) Learning spatially-varying regularisation parameters for WTV[6]

$$\begin{cases} \min_{\theta \in \mathbb{R}^n_{\geq 0}} \ \frac{1}{2}\|x(\theta) - \tilde{f}\|_2^2 + \beta\|D\theta\|_2^2, \qquad \beta \ll 1 \\ \text{s.t.} \quad x(\theta) = \arg\min_{x\in\mathbb{R}^n} \ \left\{ \sum_{j=1}^n \theta_j |(Dx)_j|_\varepsilon + \frac{1}{2}\|x - f\|_2^2 \right\} \end{cases}$$

- The noise may be not homogeneous in the image (due to device faults): adjust regularisation strength locally → requires parameter smoothness



Reconstruction by constant $\hat{\theta} \geq 0$ VS. spatially-varying $\hat{\theta} \in \mathbb{R}^n_{\geq 0}$.

- Local parameters better adapt to capture local image scales
- Treated via Schwarz domain decomposition methods/duality theory
- No clear extension to unseen data!!

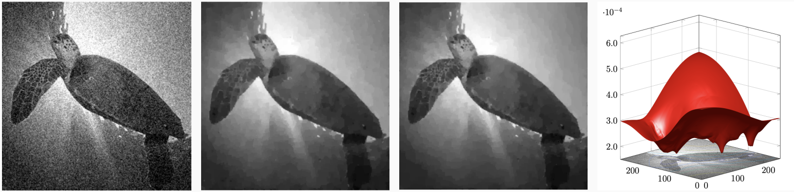
[6] Chung, De Los Reyes, Schoenlieb, '17, Hintermueller, Papafitsoros, '19

Learning better models

1) Learning spatially-varying regularisation parameters for WTV[6]

$$\begin{cases} \min_{\theta\in\mathbb{R}^n_{\geq 0}} \; \frac{1}{2}\|x(\theta)-\tilde{f}\|_2^2 + {\color{red}\beta\|D\theta\|_2^2}, \qquad \beta \ll 1 \\ \text{s.t.} \quad x(\theta) = \arg\min_{x\in\mathbb{R}^n} \; \left\{ \sum_{j=1}^n \theta_j |(Dx)_j|_\varepsilon + \frac{1}{2}\|x-f\|_2^2 \right\} \end{cases}$$

- The noise may be not homogeneous in the image (due to device faults): adjust regularisation strength locally → requires parameter smoothness
- Local parameters better adapt to capture local image scales
- Treated via Schwarz domain decomposition methods/duality theory



Noisy image, scalar/weighted TV reconstruction and plot of optimal θ^*.

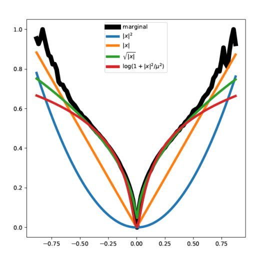
- No clear extension to unseen data!!

[6] Chung, De Los Reyes, Schoenlieb, '17, Hintermueller, Papafitsoros, '19

Learning better models

2) Learning with non-convex priors, e.g. “Field Of Experts” Roth, Black, '09, Samuel, Tappen, '09

Motivation: ℓ_1-sparsity does not match very well with the actual filter distribution



- Better match achieved by $\log(1+t^2/\mu^2) \sim \sqrt{|t|}$
- Consider

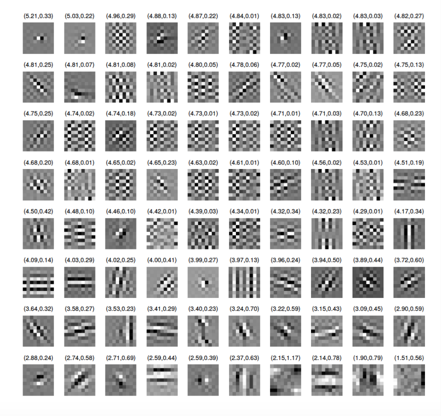
$$R(x) = \sum_{k=1}^{q} \theta_k \sum_{i=1}^{n} \rho((K_k x)_i),$$

$\rho(t) = \log(1+t^2)$ and $K_k = \sum_j \beta_{kj} B_j$, $\beta_{kj} \in \mathbb{R}$ and $\{B_j\}$ is DCT basis.

- Learn (θ, β) (Chen, Ranftl, Pock, '14)

$$\begin{cases} \min_{\theta \geq 0, \beta} \ \frac{1}{2K} \sum_{i=1}^{K} \|x_\varepsilon^i(\theta,\beta) - \tilde{f}^i\|_2^2 \\ \text{s.t.} \quad x_\varepsilon^i(\theta,\beta) = \arg\min_{x \in \mathbb{R}^n} \ \sum_{k=1}^{q} \theta_k \sum_{i=1}^{n} \rho\left(\left(\sum_j \beta_{kj} B_j x\right)_i\right) + \frac{1}{2}\|x - f^i\|_2^2, \end{cases}$$

2) Learning with non-convex priors, e.g. “Field Of Experts” Roth, Black, ’09, Samuel, Tappen, ’09

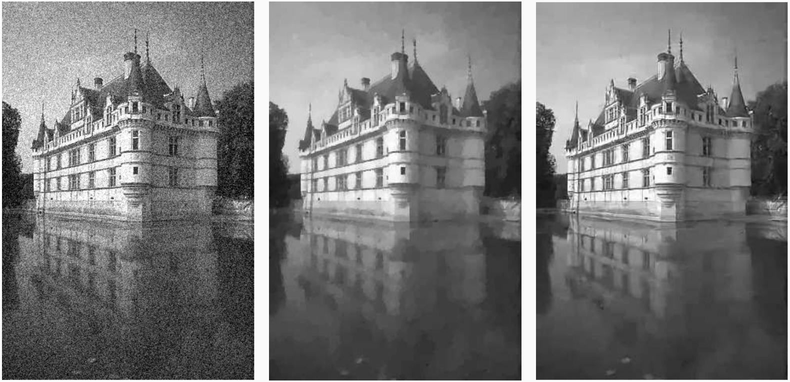


- Used on unseen data they perform as well as state-of-the art (BM3D, FoE. . .)
- These filters can be used on **different tasks** (deblurring, inpainting. . .) **outperforming** classical methods

DCT-7 atoms, $q = 80$, random initialisation + normalisation

Learning better models

2) Learning with non-convex priors, e.g. "Field Of Experts" Roth, Black, '09, Samuel, Tappen, '09



Noisy, TV reconstructed and optimal FoE reconstruction

Learning better models

2) Learning with non-convex priors, e.g. "Field Of Experts" Roth, Black, '09, Samuel, Tappen, '09

Deblurring test (20 pix. motion blur + noise), comparison with deblurring-tuned models. GMM-EPLL (27.46 dB), GOAL (27.97 dB), learned FoE (28.26 dB).

Analogous work in learning reaction-diffusion models (Chen, Yu, Pock, '15)...

Extensions

Blind bilevel learning

Motivation: very often (e.g. in biological imaging) the theoretical form of the convolution operator (PSF) is not known/does not correspond with the actual one...

Problem: estimate both solution and (convolutional) model operator

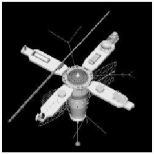
$$\boxed{\text{find} \quad x, h \quad \text{s.t.} \quad f = h * x + b}$$

Motivation: very often (e.g. in biological imaging) the theoretical form of the convolution operator (PSF) is not known/does not correspond with the actual one...

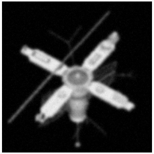
Problem: estimate both solution and (convolutional) model operator

$$\text{find} \quad x, h \quad \text{s.t.} \quad f = h * x + b$$

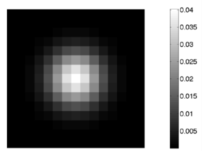
Training data (no training PSF!):



$\tilde{f}$



f, AWGN



$\tilde{h}$ (unknown)

$$\begin{cases} \min_{\theta \geq 0, h \in Q_h} \|x_\varepsilon(\theta) - \tilde{f}\|_2^2 + \frac{\beta}{2}\|Dh\|_2^2 \\ \text{s.t.} \quad x_\varepsilon(\theta) = \arg\min_{x \in \mathbb{R}^n} \ \frac{\epsilon}{2}\|Dx\|_2^2 + \|Dx\|_{2,1,\varepsilon} + \frac{\theta}{2}\|h * x - f\|_2^2 \end{cases}$$

$Q_h := \left\{ h \in \mathbb{R}^{|\Omega_h|} : \sum h_j = 1, h_j \geq 0 \right\}$. See Hintermueller, Wu, '15

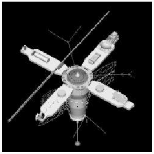
Blind bilevel learning

Motivation: very often (e.g. in biological imaging) the theoretical form of the convolution operator (PSF) is not known/does not correspond with the actual one...

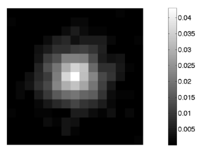
Problem: estimate both solution and (convolutional) model operator

$$\boxed{\text{find} \quad x, h \quad \text{s.t.} \quad f = h * x + b}$$

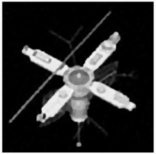
Training data (no training PSF!):



$\tilde{f}$



$\hat{h}$



$x(\hat{\theta})$

$$\begin{cases} \min\limits_{\theta \geq 0, h \in Q_h} \|x_\varepsilon(\theta) - \tilde{f}\|_2^2 + \frac{\beta}{2}\|Dh\|_2^2 \\ \text{s.t.} \quad x_\varepsilon(\theta) = \arg\min_{x \in \mathbb{R}^n} \ \frac{\epsilon}{2}\|Dx\|_2^2 + \|Dx\|_{2,1,\varepsilon} + \frac{\theta}{2}\|h * x - f\|_2^2 \end{cases}$$

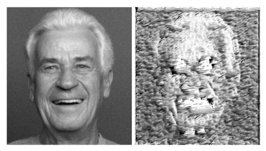
$Q_h := \left\{ h \in \mathbb{R}^{|\Omega_h|} : \sum h_j = 1, h_j \geq 0 \right\}$. See Hintermueller, Wu, '15

Towards deep learning approaches

Instead of solving lower-level problem exactly, **unroll** the iterative algorithm used to solve the lower level.

$$\begin{cases} \min_{\theta \geq 0, T} \; \frac{1}{2K} \sum_{i=1}^{K} \|x^i(\theta) - \tilde{f}_i\|_2^2 \\ \text{s.t.} \quad x_{t+1}^i = x_t^i(\theta) - \tau_t \left(\sum_{k=1}^{q} \theta_k \sum_{j=1}^{n} K_k^T \rho'((K_k x_t^i)_j) + (x_t^i - f^i) \right), \quad i = 1, \ldots, K, \quad t = 1, \ldots, T \\ x^i(\theta) = x_T^i \end{cases}$$

- **Smooth lower-level problems**: choosing the right T (early stopping) $\rightarrow$ optimal control problem



T is too large

Towards deep learning approaches

Instead of solving lower-level problem exactly, **unroll** the iterative algorithm used to solve the lower level.

$$\begin{cases} \min_{\theta \geq 0} \ \frac{1}{2K} \sum_{i=1}^{K} \|x^i(\theta) - \tilde{f}_i\|_2^2 \\ \text{s.t.} \quad x_{t+1}^i = \text{Algo}\left(x_t^i; f^i, \mathcal{G}(\theta)\right), \quad i = 1, \ldots, K, \quad t \geq 1 \end{cases}$$

- **Smooth lower-level problems**: choosing the right T (early stopping) $\rightarrow$ optimal control problem
- **Non-smooth lower-level problems**: proximal gradient algorithms (Variational networks Kobler, Klazer et al. '17, Plug & Play [6] with learned denoiser $\mathcal{G}$ Meinhardt, Moeller, Hazirbas, Cremers, '17, primal-dual algorithms (Ochs, Rantfl, Brox, Pock, '14)

  Computation of derivatives is still possible via backpropagation

[6] Journée ISIS, September 10 2021:
http://www.gdr-isis.fr/index.php?page=reunion&idreunion=454

Comparison with deep learning approaches

Deep learning: analogies

- Use many $(\tilde{f}_i, f_i)$, $i = 1, \dots K$ as training examples
- Choose a parametric function $\mathcal{G}$ (network) s.t. $\mathcal{G}(f_i; \theta) \approx \tilde{f}_i$
- For training: compute optimal $\theta \in X$ s.t.

$$\min_{\theta} \ \frac{1}{K} \sum_{i=1}^{K} \mathcal{L}(\mathcal{G}(f_i; \theta); \tilde{f}_i)$$

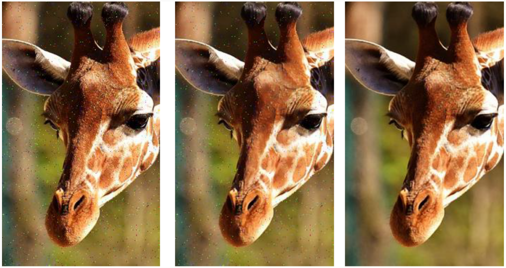
Comparison with deep learning approaches

Deep learning: analogies

- Use many $(\tilde{f}_i, f_i)$, $i = 1, \dots K$ as training examples
- Choose a parametric function $\mathcal{G}$ (network) s.t. $\mathcal{G}(f_i; \theta) \approx \tilde{f}_i$
- For training: compute optimal $\theta \in X$ s.t.

$$\min_{\theta} \frac{1}{K} \sum_{i=1}^{K} \mathcal{L}(\mathcal{G}(f_i; \theta); \tilde{f}_i)$$

Bilevel learning allows for more versatility (no black box): hybrid approach.



Stick with ℓ_1 fidelity, but use $\mathcal{G}$ as denoiser: $x_{t+1} = \mathcal{G}(x_t - \tau_t \nabla \Phi(x_t; f))$.

Take-home messages

Bilevel learning

Mathematically grounded idea for learning within an interpretable (variational) framework.

Pro's

- Interpretability ($\neq$ many deep learning approaches)
- Adaptivity to different parameter estimation problems
- Trained on **denoising** models, can be applied/tuned also on **more complex** tasks

Take-home messages

Bilevel learning

Mathematically grounded idea for learning within an interpretable (variational) framework.

Pro's

- Interpretability ($\neq$ many deep learning approaches)
- Adaptivity to different parameter estimation problems
- Trained on **denoising** models, can be applied/tuned also on **more complex** tasks

Con's

- Smoothness of lower level problem (for computing/inverting Hessians) and exact minimisation or early stopping
- Computationally heavy (despite stochastic optimisation ideas...): does not scale well for large parameter spaces...
- Non-convexity

Material and codes

- K. Kunisch and T. Pock, *A bilevel optimization approach for parameter learning in variational models*, SIAM J. Imaging Sci., 6(2):938-983, 2013.
- J. C. De los Reyes and C.-B. Schönlieb,*Image denoising: Learning the noise model via nonsmooth PDE-constrained optimization*, Inverse Probl. Imaging, 7(4), 2013.
- L. Calatroni, C. Cao, J. C. De Los Reyes, C.-B. Schönlieb, T. Valkonen, *Bilevel approaches for learning of variational imaging models*, RADON book series, vol. 18 on Variational Methods, (2016).
- Y. Chen, T. Pock, R. Ranftl, H. Bischof, *Revisiting loss-specific training of filter-based MRFs for image restoration*, GCPR, 2014.

- **Codes**: `https://github.com/VLOGroup/pgmo-lecture` (T. Pock's lectures on optimisation and learning + Python notebooks)
`https://github.com/VLOGroup/denoising-variationalnetwork` +
`https://github.com/dvillacis/bilevel_toolbox` (MATLAB)

Thanks!

Questions?

calatroni@i3s.unice.fr