Inexact Derivative-Free Optimization for Bilevel Learning

Joint work with Matthias Ehrhardt (Bath)

Lindon Roberts, ANU (`lindon.roberts@anu.edu.au`)

SIAM Computational Science & Engineering (online)
4 March 2021

Outline

Reference

M. J. Ehrhardt and R. Inexact derivative free optimization for bilevel learning. *Journal of Mathematical Imaging and Vision*, 2021.

https://doi.org/10.1007/s10851-021-01020-8

Outline

1. Bilevel Learning for Variational Regularization
2. Inexact Derivative-Free Optimization
3. Numerical Results

Variational Regularization

Many inverse problems can be posed in the form

$$\min_x \mathcal{D}(Ax, y) + \alpha\mathcal{R}(x),$$

where

- x is the quantity we wish to find;
- y is some observed data: $y \approx Ax$ (usually with noise);
- $\mathcal{D}(\cdot,\cdot)$ measures data fidelity
- $\mathcal{R}(\cdot)$ is a regularizer (what types of solutions x do we prefer?);
- $\alpha > 0$ is a parameter.

Without a regularizer, inverse problems are typically ill-posed.

Image Denoising

Given a noisy image y, find a denoised image x by solving:

$$\min_{x} \underbrace{\frac{1}{2}\|x-y\|_2^2}_{\mathcal{D}(x,y)} + \alpha \underbrace{\sum_{j}\sqrt{\|\nabla x_j\|_2^2 + \nu^2}}_{\approx \mathrm{TV}(x)} + \frac{\xi}{2}\|x\|_2^2$$

- Smooth and strongly convex optimization problem
  - Iterative methods converge linearly (e.g. gradient descent, FISTA)
- Solution depends on choices of α, ν and ξ:

Example

$(\alpha = 1,\ \nu = \xi = 10^{-3})$

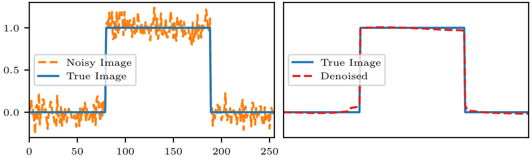


Image Denoising

Given a noisy image y, find a denoised image x by solving:

$$\min_x \underbrace{\frac{1}{2}\|x-y\|_2^2}_{\mathcal{D}(x,y)} + \alpha \underbrace{\sum_j \sqrt{\|\nabla x_j\|_2^2 + \nu^2}}_{\approx \mathrm{TV}(x)} + \frac{\xi}{2}\|x\|_2^2$$

- Smooth and strongly convex optimization problem
  - Iterative methods converge linearly (e.g. gradient descent, FISTA)
- Solution depends on choices of α, ν and ξ:

Vary α
($\nu = 10^{-3}$, $\xi = 10^{-3}$)

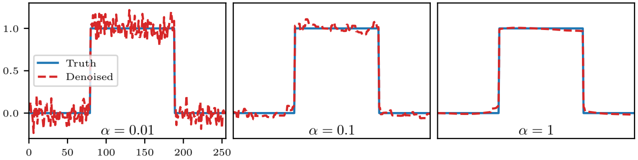


Image Denoising

Given a noisy image y, find a denoised image x by solving:

$$\min_{x} \underbrace{\frac{1}{2}\|x-y\|_2^2}_{\mathcal{D}(x,y)} + \alpha \underbrace{\sum_j \sqrt{\|\nabla x_j\|_2^2 + \nu^2}}_{\approx \mathrm{TV}(x)} + \frac{\xi}{2}\|x\|_2^2$$

- Smooth and strongly convex optimization problem
  - Iterative methods converge linearly (e.g. gradient descent, FISTA)
- Solution depends on choices of α, ν and ξ:

Vary ν

$(\alpha = 1,\ \xi = 10^{-3})$

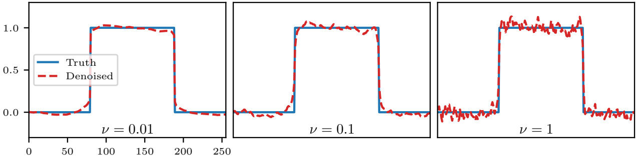


Image Denoising

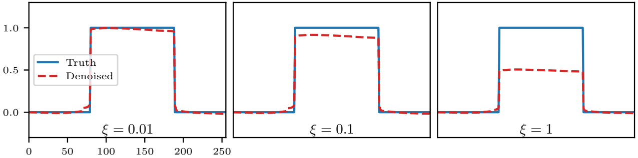
Given a noisy image y, find a denoised image x by solving:

$$\min_x \underbrace{\frac{1}{2}\|x-y\|_2^2}_{\mathcal{D}(x,y)} + \alpha \underbrace{\sum_j \sqrt{\|\nabla x_j\|_2^2 + \nu^2}}_{\approx \mathrm{TV}(x)} + \frac{\xi}{2}\|x\|_2^2$$

- Smooth and strongly convex optimization problem
  - Iterative methods converge linearly (e.g. gradient descent, FISTA)
- Solution depends on choices of α, ν and ξ:

Vary ξ

($\alpha = 1$, $\nu = 10^{-3}$)



Choosing Parameters

Solution depends on problem parameters (e.g. α, ν and ξ)

Question

How to choose good problem parameters?

Choosing Parameters

Solution depends on problem parameters (e.g. α, ν and ξ)

Question

How to choose good problem parameters?

- Trial & error
- L-curve criterion
- **Bilevel Learning** — learn from data

Bilevel Learning

Suppose we have training data $(x_1, y_1), \ldots, (x_n, y_n)$ — ground truth <u>and</u> noisy observations.

Attempt to recover x_i from y_i by solving inverse problem with parameters $\theta \in \mathbb{R}^m$:

$$\hat{x}_i(\theta) := \arg\min_x \Phi_i(x, \theta), \qquad \text{e.g. } \Phi_i(x, \theta) = \mathcal{D}(Ax, y_i) + \theta \mathcal{R}(x).$$

Try to find θ by making $\hat{x}_i(\theta)$ close to x_i

$$\min_\theta \frac{1}{n} \sum_{i=1}^{n} \|\hat{x}_i(\theta) - x_i\|^2 + \mathcal{J}(\theta),$$

with optional (smooth) term $\mathcal{J}(\theta)$ to encourage particular θ (e.g. sparsity).

Bilevel Optimization

The bilevel learning problem is:

$$\begin{aligned} \min_{\theta} \quad & f(\theta) := \frac{1}{n}\sum_{i=1}^{n} \|\hat{x}_i(\theta) - x_i\|^2 + \mathcal{J}(\theta), \\ \text{s.t.} \quad & \hat{x}_i(\theta) := \arg\min_{x} \Phi_i(x, \theta), \quad \forall i = 1, \ldots, n. \end{aligned}$$

- If Φ_i are strongly convex in x and sufficiently smooth in x and θ, then $\hat{x}_i(\theta)$ is well-defined and continuously differentiable.
- Upper-level problem ($\min_\theta f(\theta)$) is a smooth nonconvex optimization problem

Problem

Convergent algorithms require exact derivatives of $f(\theta)$, but not available (cannot even compute $\hat{x}_i(\theta)$ exactly)! [e.g. Kunisch & Pock (2013), Sherry et al. (2019)]

Bilevel Optimization with DFO

Problem

Convergent algorithms require exact derivatives of $f(\theta)$, but not available (cannot even compute $\hat{x}_i(\theta)$ exactly)!

In practice, calculate $\hat{x}_i(\theta)$ and derivatives by running N iterations of strongly convex solver (but how to choose N?).

Bilevel Optimization with DFO

Problem

Convergent algorithms require exact derivatives of $f(\theta)$, but not available (cannot even compute $\hat{x}_i(\theta)$ exactly)!

In practice, calculate $\hat{x}_i(\theta)$ and derivatives by running N iterations of strongly convex solver (but how to choose N?).

Solution:

- Use algorithms which assume $f(\theta)$ is smooth, but do not require exact evaluations of $f(\theta)$
- Don't compute (approximate) gradients of f at all: slow in practice

Bilevel Optimization with DFO

Problem

Convergent algorithms require exact derivatives of $f(\theta)$, but not available (cannot even compute $\hat{x}_i(\theta)$ exactly)!

In practice, calculate $\hat{x}_i(\theta)$ and derivatives by running N iterations of strongly convex solver (but how to choose N?).

Solution:

- Use algorithms which assume $f(\theta)$ is smooth, but do not require exact evaluations of $f(\theta)$
- Don't compute (approximate) gradients of f at all: slow in practice
- Use derivative-free optimization (DFO)

Model-Based DFO

Several types of DFO, focus on model-based DFO (mimics classical methods):

$$\min_{\theta} f(\theta)$$

For $k = 0, 1, 2, \ldots$

1. Sample f in a neighbourhood of θ_k — reuse existing evaluations where possible
2. Build an interpolating function (local model) $m_k(\theta) \approx f(\theta)$, accurate for $\theta \approx \theta_k$
3. Minimize m_k in a neighbourhood of θ_k to get θ_{k+1}

(commonly based on trust-region methods)

Theorem (Conn, Scheinberg & Vicente)

If interpolation points are close to θ_k and Λ-poised ("well-spaced"), then interpolating model is a fully linear approximation of f (accuracy $\approx$ Taylor error).

Inexact DFO for Bilevel Optimization

How to adapt to bilevel learning?

Inexact DFO for Bilevel Optimization

How to adapt to bilevel learning?

Theorem (Ehrhardt & R., extension of Conn & Vicente (2012))

If interpolation points are close to θ_k and Λ-poised, and computed minimizers of $\Phi_i(x_i, \theta)$ are sufficiently close to $\hat{x}_i(\theta)$, then interpolating model is a fully linear approximation of f.

- Allow inexact minimization of Φ_i early, only ask for high accuracy when needed
- Exploit sum-of-squares structure of f to improve performance [Cartis & R. (2019)]

Theoretical Guarantees

Algorithm converges with inexact evaluations of $\hat{x}_i(\theta)$:

Theorem (Ehrhardt & R.)

If f is sufficiently smooth and bounded below, then:

- *The inexact bilevel DFO algorithm produces a sequence θ_k such that $\|\nabla f(\theta_k)\| < \epsilon$ after at most $k = \mathcal{O}(\epsilon^{-2})$ iterations. That is, $\liminf_{k\to\infty} \|\nabla f(\theta_k)\| = 0$.*
- *All evaluations of $\hat{x}_i(\theta)$ together require at most $\mathcal{O}(\epsilon^{-2}|\log \epsilon|)$ iterations (of gradient descent, FISTA, etc.)*

- $\mathcal{O}(\epsilon^{-2})$ bound matches known results for standard DFO and trust-region methods.
- Convergence without exact function values or gradients (more robust in practice), independent of lower-level algorithm.

Numerical Results

- Implement inexact algorithm in DFO-LS (state-of-the-art DFO software)
  - Github: `numerical algorithms group/dfols`
- Use gradient descent & FISTA to calculate $\hat{x}_i(\theta) = \min_x \Phi_i(x, \theta)$
  - Using known Lipschitz and strong convexity constants (depending on θ)
  - Allow arbitrary accuracy in $\hat{x}_i(\theta)$: terminate when $\|\nabla_x \Phi\|$ sufficiently small
  - A priori linear convergence bounds too conservative in practice
- Compare to regular DFO-LS with "fixed accuracy" lower-level solutions (constant $\#$ iterations of GD/FISTA)
  - In practice, have to guess appropriate $\#$ iterations
- Measure decrease in $f(\theta)$ as function of total GD/FISTA iterations

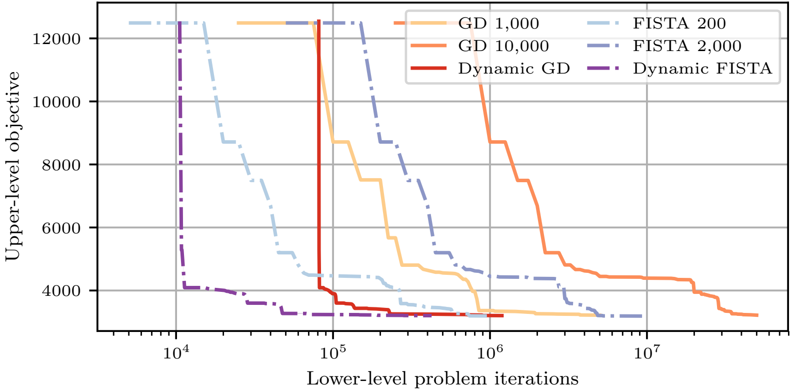
2D Denoising Problem (learn α, ν and ξ)

2D denoising — final learned parameters give good reconstructions

Final reconstruction of $x_1, \ldots, x_6$ after 100 evaluations of $f(\theta)$

2D Denoising Problem (learn α, ν and ξ)

Dynamic accuracy is faster than "fixed accuracy" (at least 10x speedup):



Objective value $f(\theta)$ vs. computational effort

Learning MRI Sampling Patterns

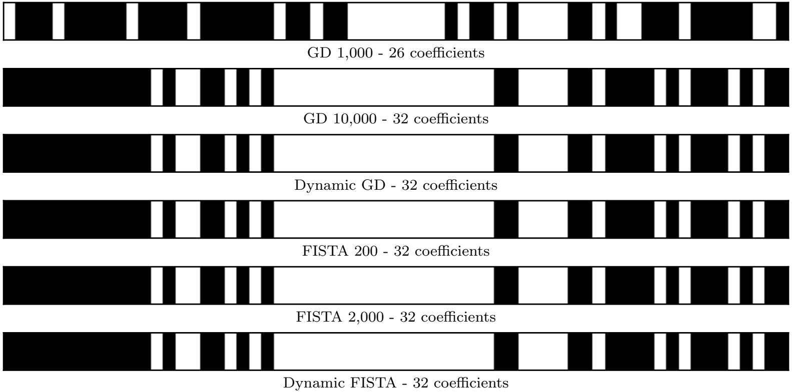
MRIs measure a subset of Fourier coefficients of an image: reconstruct using

$$\min_x \frac{1}{2}\|\mathcal{F}(x) - y\|_S^2 + \mathcal{R}(x)$$

where $\|v\|_S^2 := v^T S v$ and sampling pattern $S = \mathrm{diag}(s_1, \dots, s_d)$ for $s_j \in [0, 1]$.

- Use same smoothed TV regularizer $\mathcal{R}(x)$ (with fixed α, ν and ξ)
- Learn $s_1, \dots, s_d$, with parametrization $s_j(\theta) := \theta_j/(1-\theta_j)$ [Chen et al. (2014)]
- Measuring each coefficient takes time, so target sparsity: use $\mathcal{J}(\theta) = \|\theta\|_1$.

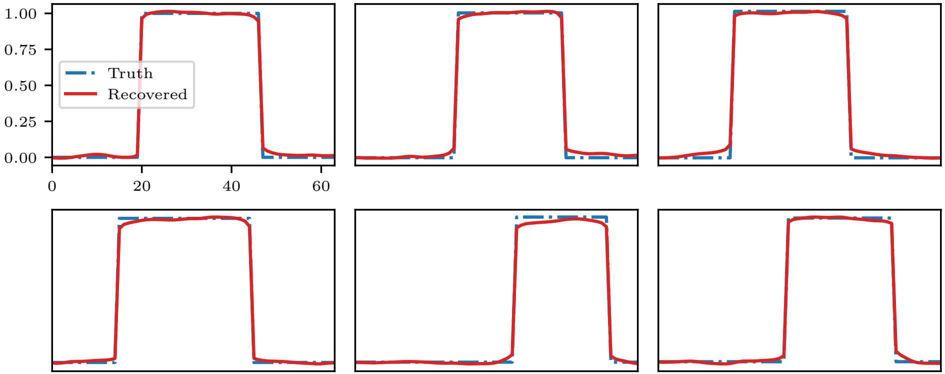
All variants learn 50% sparse sampling patterns:



Learned sampling patterns (white = active)

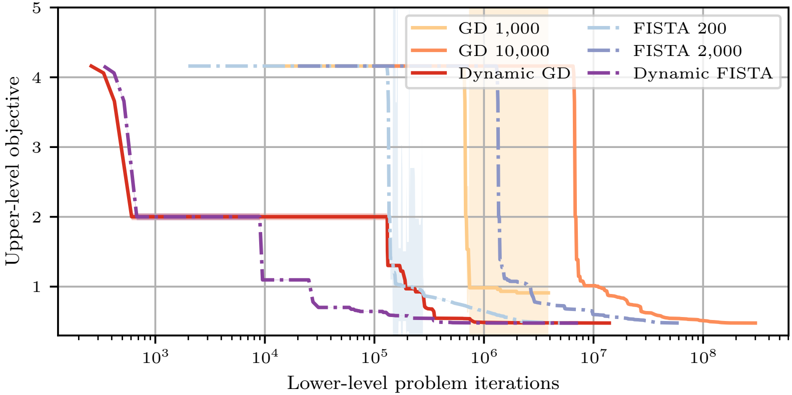
Learning MRI Sampling Patterns

Learned sampling patterns give good reconstructions:



Final reconstruction of $x_1, \ldots, x_6$ after 3000 **evaluations of $f(\theta)$**

... and dynamic accuracy is still substantially faster than fixed accuracy:



Objective value $f(\theta)$ vs. computational effort

Conclusion & Future Work

Conclusions

- Bilevel learning can be used to determine good parameters for inverse problems
- Inexact DFO method gives convergence guarantees with inexact evaluations
  - Practical & theoretical algorithms match, don't guess fixed # GD/FISTA iterations
  - Our results independent of lower-level solver choice
- Tested on 1D and 2D denoising, learning MRI sampling patterns
- Using dynamic accuracy dramatically reduces computational requirements
- Robust to choice of starting point (results in paper)

Future work:

- Subsampling algorithms (à la stochastic gradient descent)
- Learn 2D MRI sampling patterns

Coralia Cartis, Jan Fiala, Benjamin Marteau, and Lindon Roberts.
Improving the flexibility and robustness of model-based derivative-free optimization solvers.
ACM Transactions on Mathematical Software, 45(3):32:1–32:41, 2019.

Coralia Cartis and Lindon Roberts.
A derivative-free Gauss-Newton method.
Mathematical Programming Computation, 11(4):631–674, 2019.

Yunjin Chen, René Ranftl, Thomas Brox, and Thomas Pock.
A bi-level view of inpainting-based image compression.
In *19th Computer Vision Winter Workshop*, 2014.

- Andrew R. Conn, Katya Scheinberg, and Luís N. Vicente.
  Introduction to Derivative-Free Optimization, volume 8 of MPS-SIAM Series on Optimization.
  MPS/SIAM, Philadelphia, 2009.
- Andrew R. Conn and Luís N. Vicente.
  Bilevel derivative-free optimization and its application to robust optimization.
  Optimization Methods and Software, 27(3):561–577, 2012.
- Matthias J. Ehrhardt and Lindon Roberts.
  Inexact derivative free optimization for bilevel learning.
  Journal of Mathematical Imaging and Vision, 2021.

Karl Kunisch and Thomas Pock.
A Bilevel Optimization Approach for Parameter Learning in Variational Models.
SIAM Journal on Imaging Sciences, 6(2):938–983, 2013.

Ferdia Sherry, Martin Benning, Juan Carlos De los Reyes, Martin J. Graves, Georg Maierhofer, Guy Williams, Carola-Bibiane Schönlieb, and Matthias J. Ehrhardt.
Learning the Sampling Pattern for MRI.
IEEE Transactions on Medical Imaging, 39(12):4310–4321, 2020.