# Inexact Derivative-Free Optimization for Bilevel Learning

Joint work with Matthias Ehrhardt (Bath)

*J. Math. Imag. Vision (2021) & OPT2020 at NeurIPS (2020)*

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

Machine Intelligence and Learning Systems Seminar, Université Paris Dauphine-PSL
14 October 2021

# Outline

1. Bilevel Learning for Variational Regularization
2. Inexact Derivative-Free Optimization
   - Practical algorithm for bilevel learning with convergence guarantees
3. Numerical Results:
   - Image denoising
   - MRI sampling patterns
   - Logistic regression

# 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

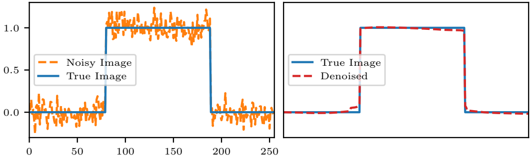
Given a noisy image $y$, find a denoised image $x$ by solving:

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

- Smooth and strongly convex optimization problem
  - Iterative methods converge linearly (e.g. gradient descent, FISTA)
- Solution depends on choices of $\alpha$, $\nu$ and $\xi$:

**Example**

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



# Image Denoising

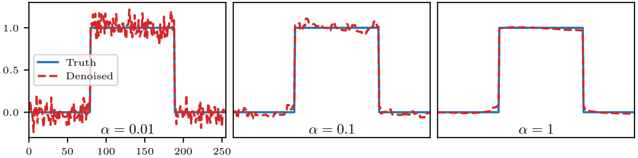
Given a noisy image $y$, find a denoised image $x$ by solving:

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

- Smooth and strongly convex optimization problem
  - Iterative methods converge linearly (e.g. gradient descent, FISTA)
- Solution depends on choices of $\alpha$, $\nu$ and $\xi$:

**Vary $\alpha$**

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



# Image Denoising

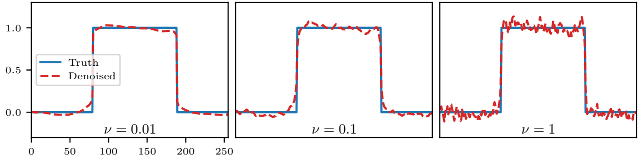
Given a noisy image $y$, find a denoised image $x$ by solving:

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

- Smooth and strongly convex optimization problem
  - Iterative methods converge linearly (e.g. gradient descent, FISTA)
- Solution depends on choices of $\alpha$, $\nu$ and $\xi$:

**Vary $\nu$**

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

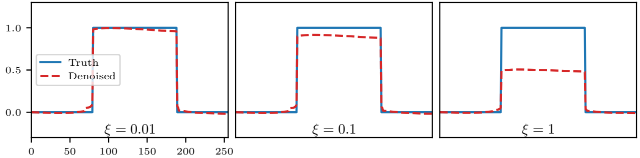


Given a noisy image $y$, find a denoised image $x$ by solving:

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

- Smooth and strongly convex optimization problem
  - Iterative methods converge linearly (e.g. gradient descent, FISTA)
- Solution depends on choices of $\alpha$, $\nu$ and $\xi$:

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



# Choosing Parameters

Recovered solution depends strongly on problem parameters (e.g. $\alpha$, $\nu$ and $\xi$)

**Question**

How to choose good problem parameters?

# Choosing Parameters

Recovered solution depends strongly on problem parameters (e.g. $\alpha$, $\nu$ and $\xi$)

**Question**

How to choose good problem parameters?

- Trial & error
- L-curve criterion
- **Bilevel learning** — data-driven approach

# 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 $\theta$ 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 choices of $\theta$.

# Bilevel Optimization

The bilevel learning problem is:

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

$$\text{s.t.} \quad \hat{x}_i(\theta) := \arg\min_{x} \Phi_i(x, \theta), \quad \forall i = 1, \ldots, n.$$

- If $\Phi_i$ are strongly convex in $x$ and sufficiently smooth in $x$ and $\theta$, 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 do not require exact evaluations of $f(\theta)$
- Don't compute gradients of $f$ at all, since slow in practice $\Rightarrow$ derivative-free optimization (DFO)

# Model-Based DFO

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

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

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

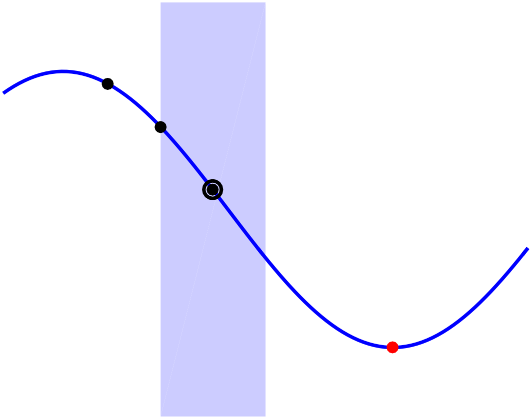
1. Sample $f$ in a neighborhood of $\theta_k$ — reuse existing evaluations where possible
2. Build an interpolating function (local model) $m_k(\theta) \approx f(\theta)$, accurate for $\theta \approx \theta_k$
3. Calculate tentative new point by minimizing model in a neighborhood

$$\theta_k^+ = \arg\min_{\theta} m_k(\theta), \qquad \text{subject to } \|\theta - \theta_k\| \leq \Delta_k.$$

4. Accept/reject step and adjust $\Delta_k$ based on quality of new point $f(\theta_k^+)$

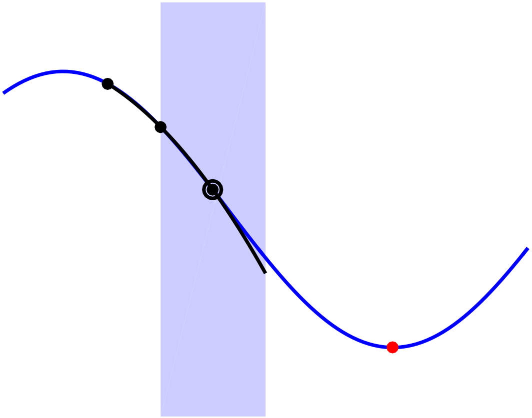
$$\theta_{k+1} = \begin{cases} \theta_k^+, & \text{if sufficient decrease,} \qquad \longleftarrow \text{(maybe increase } \Delta_k\text{)} \\ \theta_k, & \text{otherwise.} \qquad \longleftarrow \text{(decrease } \Delta_k\text{)} \end{cases}$$

# Example: Model-Based DFO



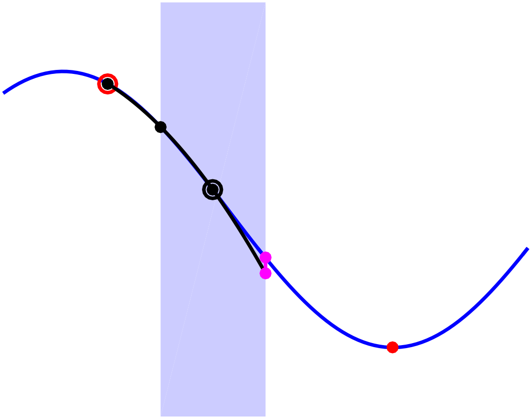
**1. Choose interpolation set**

# Example: Model-Based DFO



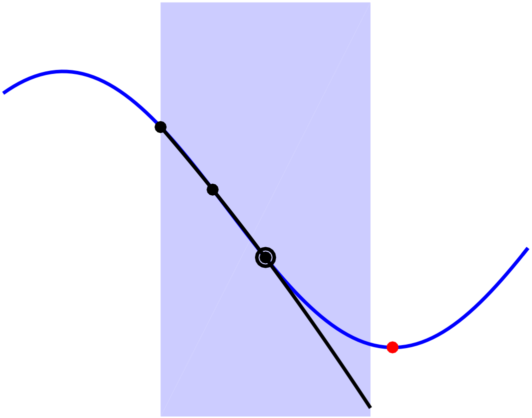
**2. Interpolate & minimize...**

# Example: Model-Based DFO



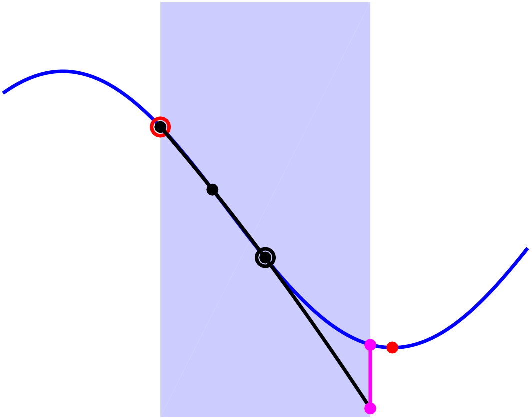
**3. Add new point to interpolation set (replace a bad point)**

# Example: Model-Based DFO



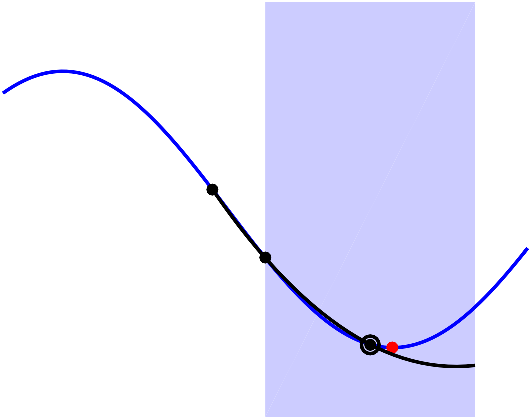
**4. Repeat with new interpolation set & model**

# Example: Model-Based DFO



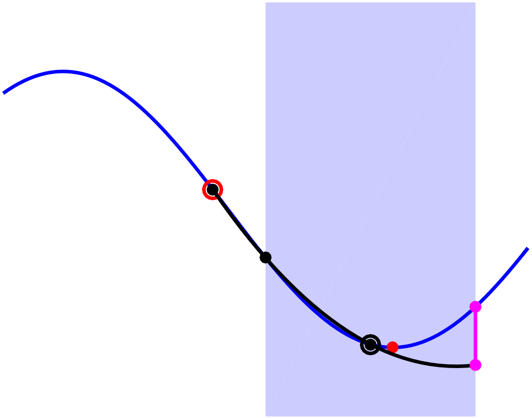
**4. Repeat with new interpolation set & model**

# Example: Model-Based DFO



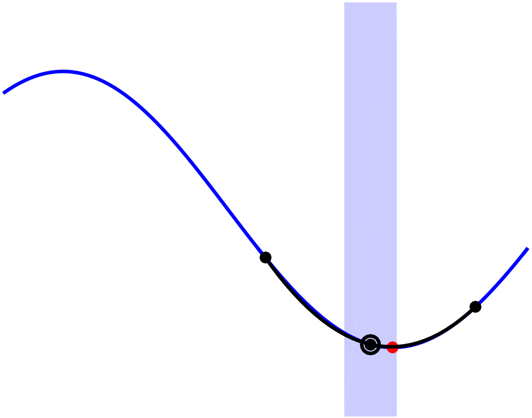
**4. Repeat with new interpolation set & model**

# Example: Model-Based DFO



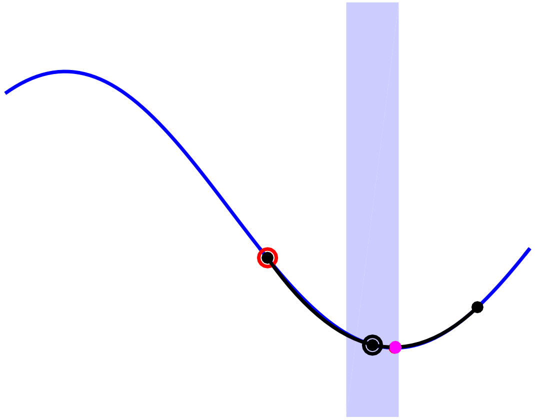
**4. Repeat with new interpolation set & model**

# Example: Model-Based DFO



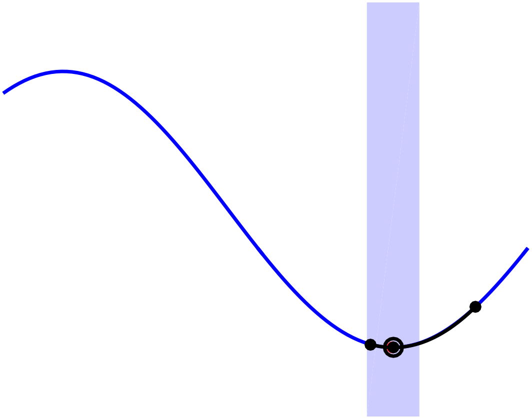
**4. Repeat with new interpolation set & model**

# Example: Model-Based DFO



**4. Repeat with new interpolation set & model**

# Example: Model-Based DFO



**4. Repeat with new interpolation set & model**

# Inexact DFO for Bilevel Optimization

Key to DFO convergence theory is the following approximation result:

**Theorem (Conn, Scheinberg & Vicente)**

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

How to adapt to bilevel learning?

# Inexact DFO for Bilevel Optimization

Key to DFO convergence theory is the following approximation result:

**Theorem (Conn, Scheinberg & Vicente)**

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

How to adapt to bilevel learning?

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

*If interpolation points are close to $\theta_k$ and "well-spaced", 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 $\Phi_i$ 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 $\theta_k$ 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.)*

# 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 $\theta_k$ 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.)*

**Key Benefit**

Using inexact information in a structured way gives a faster learning algorithm plus guaranteed convergence (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 $\theta$)
  - 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

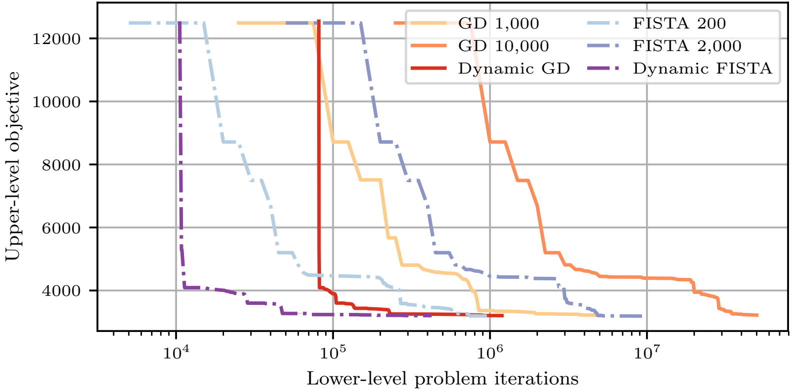
# 2D Denoising Problem (learn $\alpha$, $\nu$ and $\xi$)

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 $\alpha$, $\nu$ and $\xi$)

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

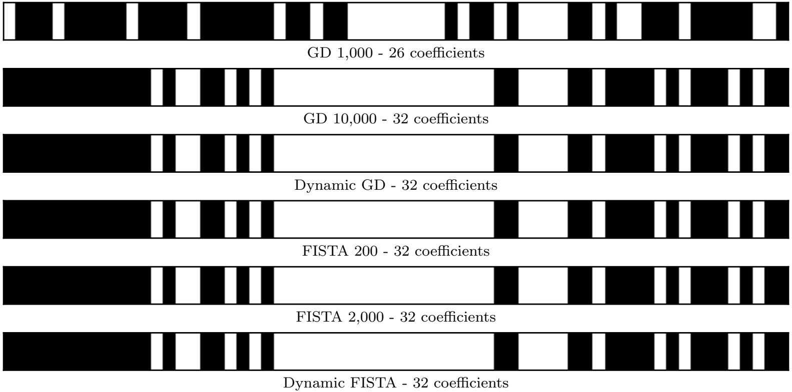
$$\min_x \frac{1}{2}\|\mathcal{F}(x) - y\|_S^2 + \mathcal{R}(x)$$

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

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

# Learning MRI Sampling Patterns

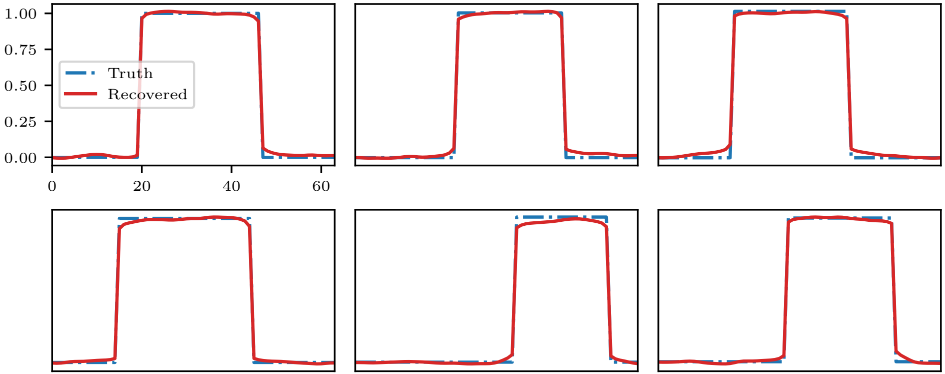
All variants learn 50% sparse sampling patterns:




**Learned sampling patterns (white = active)**

# Learning MRI Sampling Patterns

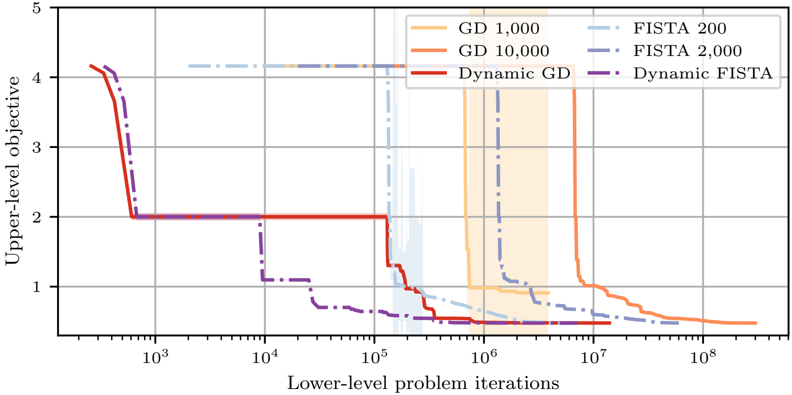
Learned sampling patterns give good reconstructions:




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

# Learning MRI Sampling Patterns

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

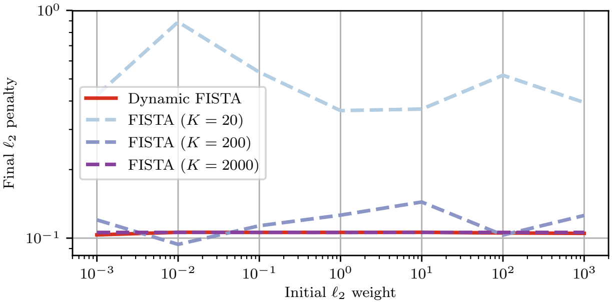



**Objective value $f(\theta)$ vs. computational effort**

# Robustness

We also gain robustness to starting point (because of relevant convergence guarantees).

Example: learning regularizer for logistic regression (on MNIST dataset)




**Final parameter $\theta^*$ vs. starting point $\theta^0$**

# 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: no guesswork required!
  - Our results independent of lower-level solver choice
- Order-of-magnitude speedup and improved robustness on several problem categories

## Future work

- Incorporate inexact gradient information (without losing convergence guarantees)
- Subsampling algorithms (à la stochastic gradient descent)
- Large-scale applications: learning 2D MRI sampling patterns, convex neural net regularizers

# References i


C. Cartis, J. Fiala, B. Marteau, and L. Roberts, *Improving the flexibility and robustness of model-based derivative-free optimization solvers*, ACM Transactions on Mathematical Software, 45 (2019), pp. 32:1–32:41.

C. Cartis and L. Roberts, *A derivative-free Gauss-Newton method*, Mathematical Programming Computation, 11 (2019), pp. 631–674.

Y. Chen, R. Ranftl, T. Brox, and T. Pock, *A bi-level view of inpainting-based image compression*, in 19th Computer Vision Winter Workshop, 2014.

A. R. Conn, K. Scheinberg, and L. N. Vicente, *Introduction to Derivative-Free Optimization*, vol. 8 of MPS-SIAM Series on Optimization, MPS/SIAM, Philadelphia, 2009.

A. R. Conn and L. N. Vicente, *Bilevel derivative-free optimization and its application to robust optimization*, Optimization Methods and Software, 27 (2012), pp. 561–577.

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

K. Kunisch and T. Pock, *A Bilevel Optimization Approach for Parameter Learning in Variational Models*, SIAM Journal on Imaging Sciences, 6 (2013), pp. 938–983.

F. Sherry, M. Benning, J. C. D. los Reyes, M. J. Graves, G. Maierhofer, G. Williams, C.-B. Schönlieb, and M. J. Ehrhardt, *Learning the Sampling Pattern for MRI*, IEEE Transactions on Medical Imaging, 39 (2020), pp. 4310–4321.