Learning the Sampling for MRI

Matthias J. Ehrhardt

Institute for Mathematical Innovation, University of Bath, UK

June 24, 2020

Joint work with:
F. Sherry, M. Graves, G. Maierhofer, G. Williams, C.-B. Schönlieb (all Cambridge, UK), M. Benning (Queen Mary, UK), J.C. De los Reyes (EPN, Ecuador)

The Leverhulme Trust

UKRI Engineering and Physical Sciences Research Council

THE FARADAY INSTITUTION

Outline

1) Motivation

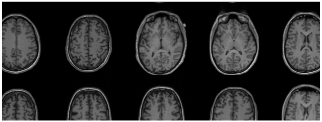
$\min_x \frac{1}{2}\|SFx - y\|_2^2 + \lambda \mathcal{R}(x)$

2) Bilevel Learning

$\min_{x,y} f(x,y)$

$x = \arg\min_z g(z,y)$

3) Learn sampling pattern in MRI



Inverse problems

$$Ax = y$$

x : desired solution

y : observed data

A : mathematical model

Goal: recover x given y

Inverse problems

$$Ax = y$$

x : desired solution

y : observed data

A : mathematical model

Goal: recover x given y

Hadamard (1902): We call an inverse problem $Ax = y$ **well-posed** if

(1) a solution x^* **exists**

(2) the solution x^* is **unique**

(3) x^* depends **continuously** on data y.

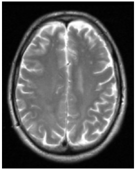
Otherwise, it is called **ill-posed**.

Jacques Hadamard

Most interesting problems are **ill-posed**.

Example: Magnetic Resonance Imaging (MRI)

MRI scanner

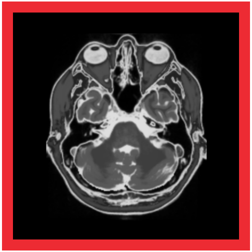


T_2^*

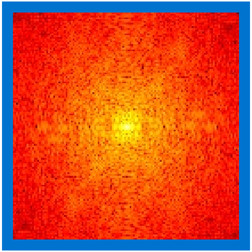
Continuous model: Fourier transform

$$Ax(s) = \int_{\mathbb{R}^2} x(s)\exp(-ist)dt$$

Dicrete model: $A = F \in \mathbb{C}^{N\times N}$

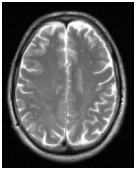


$\longrightarrow$



Example: Magnetic Resonance Imaging (MRI)

MRI scanner

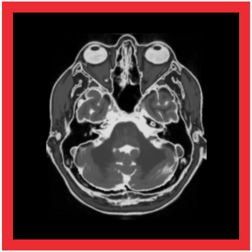


T_2^*

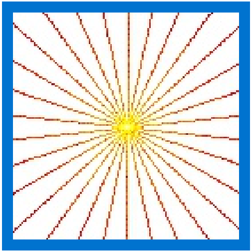
Continuous model: Fourier transform

$$Ax(s) = \int_{\mathbb{R}^2} x(s) \exp(-ist) dt$$

Dicrete model: $A = SF \in \mathbb{C}^{n \times N}$



$\longrightarrow$



Solution **not unique**.

How to solve inverse problems?

Variational regularization ($\sim$2000)
Approximate a solution x^* of $Ax = y$ via

$$\hat{x} \in \arg\min_{x} \left\{ \frac{1}{2} \|Ax - y\|_2^2 + \lambda \mathcal{R}(x) \right\}$$

$\mathcal{R}$ **regularizer**: penalizes unwanted features, ensures stability and uniqueness

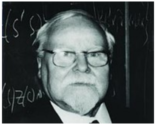
λ **regularization parameter**: $\lambda \geq 0$. If $\lambda = 0$, then an original solution is recovered. If $\lambda \to \infty$, more and more weight is given to the regularizer $\mathcal{R}$.

textbooks: Scherzer et al. 2008, Ito and Jin 2015, Benning and Burger 2018

Example: Regularizers

Tikhonov regularization ($\sim$1960):

$$\mathcal{R}(x) = \frac{1}{2}\|x\|_2^2$$

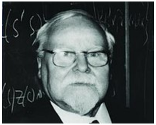


A. Tikhonov

Example: Regularizers

Tikhonov regularization ($\sim$1960):

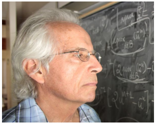
$$\mathcal{R}(x) = \frac{1}{2}\|x\|_2^2$$



A. Tikhonov

Total Variation Rudin, Osher, Fatemi 1992

$$\mathcal{R}(x) = \|\nabla x\|_1$$



S. Osher

H^1 ($\sim$1960-1990?)

$$\mathcal{R}(x) = \frac{1}{2}\|\nabla x\|_2^2$$

Wavelet sparsity ($\sim$1990)

$$\mathcal{R}(x) = \|Wx\|_1$$

Total Generalized Variation: Bredies, Kunisch, Pock 2010

$$\mathcal{R}(x) = \inf_v \|\nabla x - v\|_1 + \beta\|\nabla v\|_1$$

Example: MRI reconstruction

Compressed Sensing MRI:

$A = S \circ F$ Lustig, Donoho, Pauly 2007

Fourier transform F, sampling $Sw = (w_i)_{i \in \Omega}$

$$\hat{x} \in \arg\min_{x} \left\{ \frac{1}{2} \|SFx - y\|_2^2 + \lambda \|\nabla x\|_1 \right\}$$

Miki Lustig

Example: MRI reconstruction

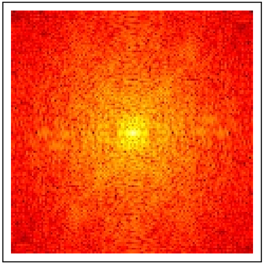
Compressed Sensing MRI:

$A = S \circ F$ Lustig, Donoho, Pauly 2007

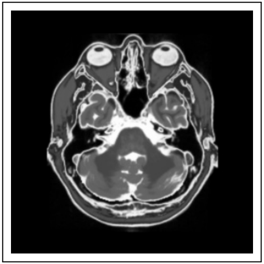
Fourier transform F, sampling $Sw = (w_i)_{i \in \Omega}$

$$\hat{x} \in \arg\min_x \left\{ \frac{1}{2} \|SFx - y\|_2^2 + \lambda \|\nabla x\|_1 \right\}$$

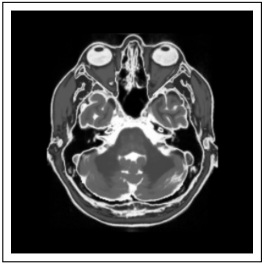
Miki Lustig



sampling $S^* y$



$\lambda = 0$



$\lambda = 1$

Example: MRI reconstruction

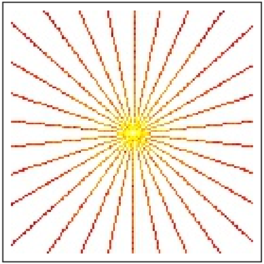
Compressed Sensing MRI:

$A = S \circ F$ Lustig, Donoho, Pauly 2007

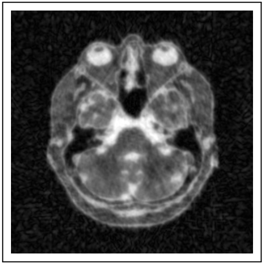
Fourier transform F, sampling $Sw = (w_i)_{i \in \Omega}$

$$\hat{x} \in \arg\min_{x} \left\{ \frac{1}{2} \|SFx - y\|_2^2 + \lambda \|\nabla x\|_1 \right\}$$

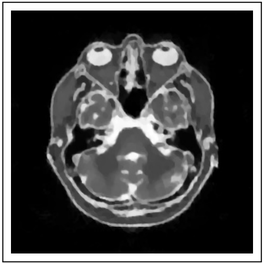
Miki Lustig



sampling $S^* y$



$\lambda = 0$



$\lambda = 10^{-4}$

Example: MRI reconstruction

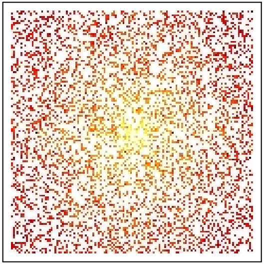
Compressed Sensing MRI:

$A = S \circ F$ Lustig, Donoho, Pauly 2007

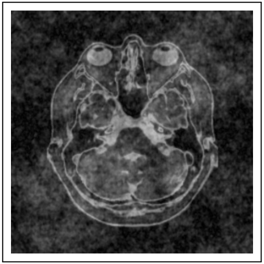
Fourier transform F, sampling $Sw = (w_i)_{i \in \Omega}$

$$\hat{x} \in \arg\min_x \left\{ \frac{1}{2} \|SFx - y\|_2^2 + \lambda \|\nabla x\|_1 \right\}$$

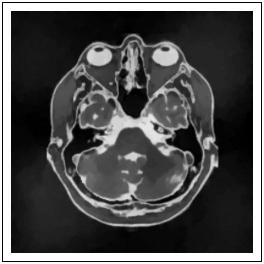
Miki Lustig



sampling $S^* y$



$\lambda = 0$



$\lambda = 10^{-4}$

Example: MRI reconstruction

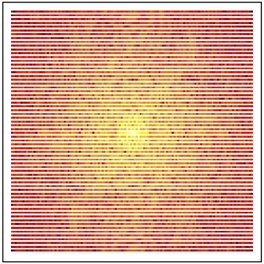
Compressed Sensing MRI:

$A = S \circ F$ Lustig, Donoho, Pauly 2007

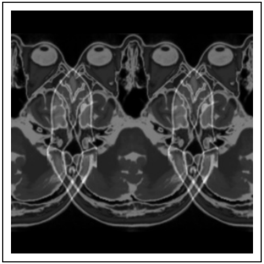
Fourier transform F, sampling $Sw = (w_i)_{i \in \Omega}$

$$\hat{x} \in \arg\min_x \left\{ \frac{1}{2} \|SFx - y\|_2^2 + \lambda \|\nabla x\|_1 \right\}$$

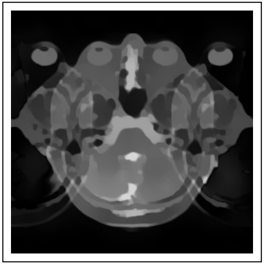
Miki Lustig



sampling $S^* y$



$\lambda = 0$



$\lambda = 10^{-3}$

Example: MRI reconstruction

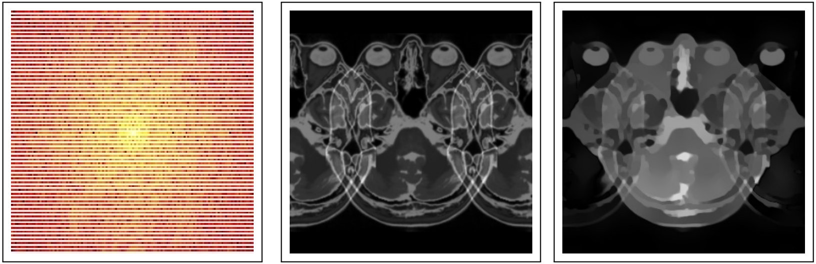
Compressed Sensing MRI:

$A = S \circ F$ Lustig, Donoho, Pauly 2007

Fourier transform F, sampling $Sw = (w_i)_{i \in \Omega}$

$$\hat{x} \in \arg\min_x \left\{ \frac{1}{2} \|SFx - y\|_2^2 + \lambda \|\nabla x\|_1 \right\}$$

Miki Lustig



sampling $S^* y$ $\lambda = 0$ $\lambda = 10^{-3}$

How to choose the sampling S? Is there an optimal sampling?

Example: MRI reconstruction

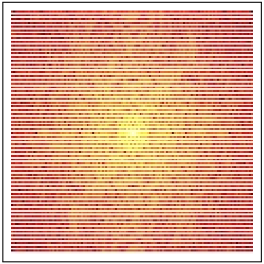
Compressed Sensing MRI:

$A = S \circ F$ Lustig, Donoho, Pauly 2007

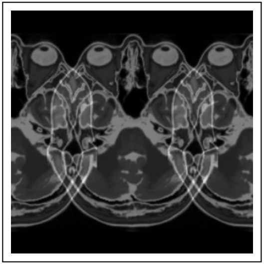
Fourier transform F, sampling $Sw = (w_i)_{i \in \Omega}$

$$\hat{x} \in \underset{x}{\arg\min} \left\{ \frac{1}{2} \| SFx - y \|_2^2 + \lambda \| \nabla x \|_1 \right\}$$

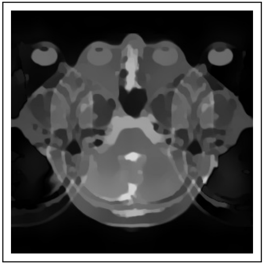
Miki Lustig



sampling $S^* y$



$\lambda = 0$



$\lambda = 10^{-3}$

How to choose the sampling S? Is there an optimal sampling?

Does the optimal sampling depend on $\mathcal{R}$ and λ?

Some important works on sampling for MRI

Uninformed

- Cartesian, radial, variable density ... e.g. Lustig et al. 2007
  - ✓ simple to implement
  - ✗ not tailored to application
  - ✗ not tailored to regularizer / reconstruction method
- compressed sensing **theory**: random sampling, mostly uniform
  e.g. Candes and Romberg 2007
  - ✓ mathematical guarantees
  - ✗ limited to few sparsity promoting regularizers: mostly ℓ^1 type
  - ✗ specific yet uninformed class of recoverable signals: sparse

Some important works on sampling for MRI

Uninformed

- Cartesian, radial, variable density ... e.g. Lustig et al. 2007
  - ✓ simple to implement
  - ✗ not tailored to application
  - ✗ not tailored to regularizer / reconstruction method
- compressed sensing **theory**: random sampling, mostly uniform
  e.g. Candes and Romberg 2007
  - ✓ mathematical guarantees
  - ✗ limited to few sparsity promoting regularizers: mostly ℓ^1 type
  - ✗ specific yet uninformed class of recoverable signals: sparse

Learned

- **Largest Fourier coefficients** of training set Knoll et al. 2011
  - ✓ simple to implement, computationally light
  - ✗ not tailored to regularizer / reconstruction method
- **greedy**: iteratively select "best" sample Gözcü et al. 2018
  - ✓ adaptive to dataset, regularizer / reconstruction method
  - ✗ only discrete values, e.g. can't learn regularization parameter
  - ✗ computationally heavy

Bilevel Learning

Bilevel learning for inverse problems

$$\hat{x} = \arg\min_{x} \left\{ \frac{1}{2}\|Ax - y\|_2^2 + \lambda \mathcal{R}(x) \right\}$$

$\mathcal{R}$ smooth and strongly convex

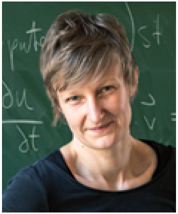
Bilevel learning for inverse problems

Upper level (learning):
Given $(x^\dagger, y), y = Ax^\dagger + \varepsilon$, solve

$$\min_{\lambda \geq 0, \hat{x}} \|\hat{x} - x^\dagger\|_2^2$$

Lower level (solve inverse problem):

$$\hat{x} = \arg\min_x \left\{ \frac{1}{2}\|Ax - y\|_2^2 + \lambda \mathcal{R}(x) \right\}$$



Carola Schönlieb

$\mathcal{R}$ smooth and strongly convex

von Stackelberg 1934, Kunisch and Pock 2013, De los Reyes and Schönlieb 2013

Bilevel learning for inverse problems

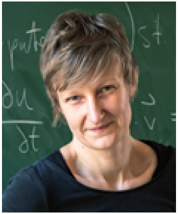
Upper level (learning):

Given $(x_i^\dagger, y_i)_{i=1}^n, y_i = Ax_i^\dagger + \varepsilon_i$, solve

$$\min_{\color{red}{\lambda \geq 0}, \color{blue}{\hat{x}_i}} \frac{1}{n} \sum_{i=1}^{n} \|\color{blue}{\hat{x}_i} - x_i^\dagger\|_2^2$$

Lower level (solve inverse problem):

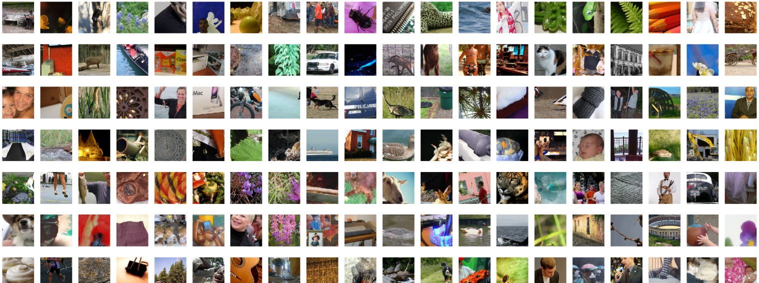
$$\color{blue}{\hat{x}_i} = \arg\min_x \left\{ \frac{1}{2}\|Ax - y_i\|_2^2 + \color{red}{\lambda}\mathcal{R}(x) \right\}$$



Carola Schönlieb

$\mathcal{R}$ smooth and strongly convex

von Stackelberg 1934, Kunisch and Pock 2013, De los Reyes and Schönlieb 2013



Bilevel learning: Reduced formulation

Upper level:

$$\min_{\lambda \geq 0, \hat{x}} \|\hat{x} - x^\dagger\|_2^2$$

Lower level:

$$\hat{x} = \arg\min_x \left\{ \frac{1}{2}\|Ax - y\|_2^2 + \lambda \mathcal{R}(x) \right\}$$

Bilevel learning: Reduced formulation

Upper level:

$$\min_{\lambda \geq 0, \hat{x}} U(\hat{x})$$

Lower level:

$$\hat{x} = \arg\min_{x} \left\{ \frac{1}{2} \|Ax - y\|_2^2 + \lambda \mathcal{R}(x) \right\}$$

Bilevel learning: Reduced formulation

Upper level:

$$\min_{\lambda \geq 0, \hat{x}} U(\hat{x})$$

Lower level:

$$\hat{x} = \arg\min_{x} L(x, \lambda)$$

Bilevel learning: Reduced formulation

Upper level:

$$\min_{\lambda \geq 0, \hat{x}} U(\hat{x})$$

Lower level:

$$x_\lambda := \hat{x} = \arg\min_x L(x, \lambda)$$

Reduced formulation: $\displaystyle\min_{\lambda \geq 0} U(x_\lambda) =: \tilde{U}(\lambda)$

Bilevel learning: Reduced formulation

Upper level: $$\min_{\lambda \geq 0, \hat{x}} U(\hat{x})$$

Lower level:
$$x_\lambda := \hat{x} = \arg\min_x L(x, \lambda) \quad \Leftrightarrow \quad \partial_x L(x_\lambda, \lambda) = 0$$

Reduced formulation: $$\min_{\lambda \geq 0} U(x_\lambda) =: \tilde{U}(\lambda)$$

Bilevel learning: Reduced formulation

Upper level: $\min_{\lambda \geq 0, \hat{x}} U(\hat{x})$

Lower level:

$$x_\lambda := \hat{x} = \arg\min_x L(x, \lambda) \quad \Leftrightarrow \quad \partial_x L(x_\lambda, \lambda) = 0$$

Reduced formulation: $\min_{\lambda \geq 0} U(x_\lambda) =: \tilde{U}(\lambda)$

$$0 = \partial_x^2 L(x_\lambda, \lambda)\partial_\lambda x_\lambda + \partial_\theta \partial_x L(x_\lambda, \lambda) \quad \Leftrightarrow \quad \partial_\lambda x_\lambda = -B^{-1}A$$

Bilevel learning: Reduced formulation

Upper level:
$$\min_{\lambda \geq 0, \hat{x}} U(\hat{x})$$

Lower level:
$$x_\lambda := \hat{x} = \arg\min_x L(x, \lambda) \quad \Leftrightarrow \quad \partial_x L(x_\lambda, \lambda) = 0$$

Reduced formulation: $\min_{\lambda \geq 0} U(x_\lambda) =: \tilde{U}(\lambda)$

$$0 = \partial_x^2 L(x_\lambda, \lambda) \partial_\lambda x_\lambda + \partial_\theta \partial_x L(x_\lambda, \lambda) \quad \Leftrightarrow \quad \partial_\lambda x_\lambda = -B^{-1} A$$

$$\nabla \tilde{U}(\lambda) = (\partial_\lambda x_\lambda)^* \nabla U(x_\lambda)$$

Bilevel learning: Reduced formulation

Upper level: $\min_{\lambda \geq 0, \hat{x}} U(\hat{x})$

Lower level:

$$x_\lambda := \hat{x} = \arg\min_x L(x, \lambda) \quad \Leftrightarrow \quad \partial_x L(x_\lambda, \lambda) = 0$$

Reduced formulation: $\min_{\lambda \geq 0} U(x_\lambda) =: \tilde{U}(\lambda)$

$$0 = \partial_x^2 L(x_\lambda, \lambda) \partial_\lambda x_\lambda + \partial_\theta \partial_x L(x_\lambda, \lambda) \quad \Leftrightarrow \quad \partial_\lambda x_\lambda = -B^{-1} A$$

$$\begin{aligned} \nabla \tilde{U}(\lambda) &= (\partial_\lambda x_\lambda)^* \nabla U(x_\lambda) \\ &= -A^* B^{-1} \nabla U(x_\lambda) = -A^* w \end{aligned}$$

where w solves $Bw = \nabla U(x_\lambda)$.

Algorithm for Bilevel learning

Upper level: $\min_{\lambda \geq 0, \hat{x}} U(\hat{x})$

Lower level: $x_\lambda := \arg\min_x L(x, \lambda)$

Reduced formulation: $\min_{\lambda \geq 0} U(x_\lambda) =: \tilde{U}(\lambda)$

- Solve reduced formulation via L-BFGS-B Nocedal and Wright 2000
- Compute gradients: Given λ
  (1) Compute x_λ, e.g. via PDHG Chambolle and Pock 2011
  (2) Solve $Bw = \nabla U(x_\lambda)$, $B := \partial_x^2 L(x_\lambda, \lambda)$ e.g. via CG
  (3) Compute $\nabla \tilde{U}(\lambda) = -A^* w$, $A := \partial_\theta \partial_x L(x_\lambda, \lambda)$

Learn sampling pattern in MRI

Learn sampling pattern in MRI

Lower level (MRI reconstruction):

$$R(\lambda, s, y) = \arg\min_x \left\{ \frac{1}{2} \| S(Fx - y) \|_2^2 + \lambda \mathcal{R}(x) \right\}$$

$$S = \mathrm{diag}(s), \quad s_i \in \{0, 1\}$$

Sherry et al. 2019, https://arxiv.org/pdf/1906.08754.pdf

Learn sampling pattern in MRI

Upper level (learning):
Given **training data** $(x_i^\dagger, y_i)_{i=1}^n$, solve

$$\min_{\lambda \geq 0, s \in \{0,1\}^m} \frac{1}{n} \sum_{i=1}^n \| R(\lambda, s, y_i) - x_i^\dagger \|_2^2$$

Lower level (MRI reconstruction):

$$R(\lambda, s, y) = \arg\min_x \left\{ \frac{1}{2} \| S(Fx - y) \|_2^2 + \lambda \mathcal{R}(x) \right\}$$

$$S = \mathrm{diag}(s), \quad s_i \in \{0, 1\}$$

Sherry et al. 2019, `https://arxiv.org/pdf/1906.08754.pdf`

Learn sampling pattern in MRI

Upper level (learning):
Given **training data** $(x_i^\dagger, y_i)_{i=1}^n$, solve

$$\min_{\lambda \geq 0, s \in [0,1]^m} \frac{1}{n} \sum_{i=1}^{n} \| R(\lambda, s, y_i) - x_i^\dagger \|_2^2$$

Lower level (MRI reconstruction):

$$R(\lambda, s, y) = \arg\min_x \left\{ \frac{1}{2} \| S(Fx - y) \|_2^2 + \lambda \mathcal{R}(x) \right\}$$

$$S = \mathrm{diag}(s), \quad s_i \in [0, 1]$$

Sherry et al. 2019, https://arxiv.org/pdf/1906.08754.pdf

Learn sampling pattern in MRI

Upper level (learning):

Given **training data** $(x_i^\dagger, y_i)_{i=1}^n$, solve

$$\min_{\lambda \geq 0, s \in [0,1]^m} \frac{1}{n} \sum_{i=1}^{n} \|R(\lambda, s, y_i) - x_i^\dagger\|_2^2 + \beta_1 \|s\|_1 + \beta_2 \|s(1-s)\|_1$$

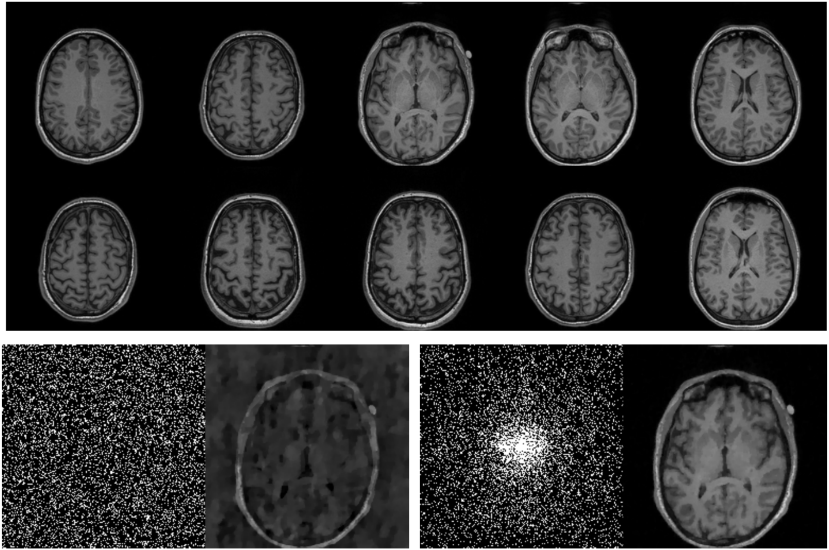
Lower level (MRI reconstruction):

$$R(\lambda, s, y) = \arg\min_x \left\{ \frac{1}{2} \|S(Fx - y)\|_2^2 + \lambda \mathcal{R}(x) \right\}$$

$$S = \operatorname{diag}(s), \quad s_i \in [0, 1]$$

Sherry et al. 2019, `https://arxiv.org/pdf/1906.08754.pdf`

Classical compressed sensing versus learned



Uniform random | Reconstruction | Learned | Reconstruction

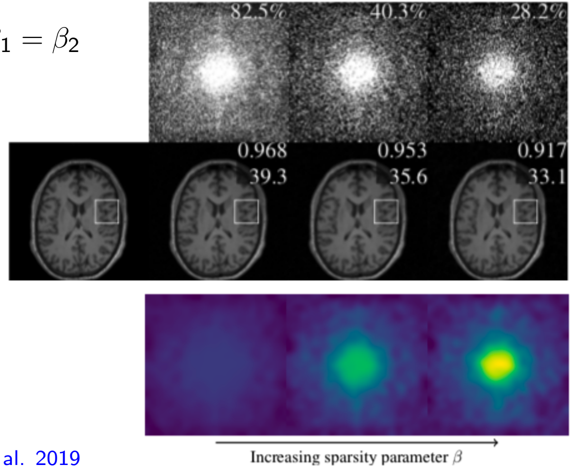
Sherry et al. 2019

Increasing sparsity

Reminder: **Upper level** (learning)

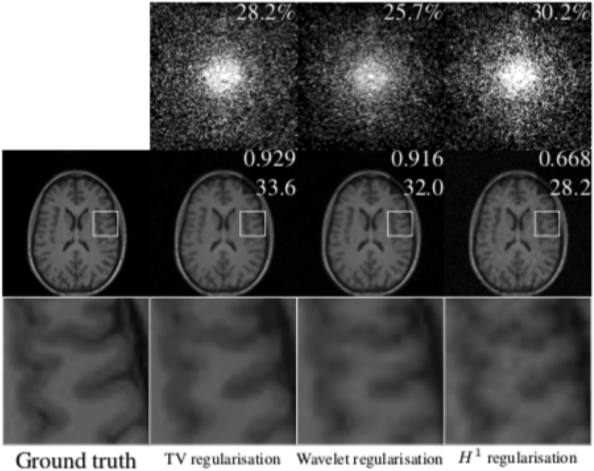
$$\min_{\lambda \geq 0, s \in [0,1]^m} \frac{1}{n} \sum_{i=1}^{n} \|R(\lambda, s, y_i) - x_i\|_2^2 + \beta_1 \|s\|_1 + \beta_2 \|s(1-s)\|_1$$

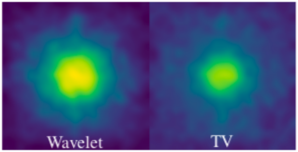
$$\beta = \beta_1 = \beta_2$$

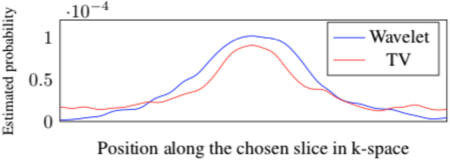


Sherry et al. 2019

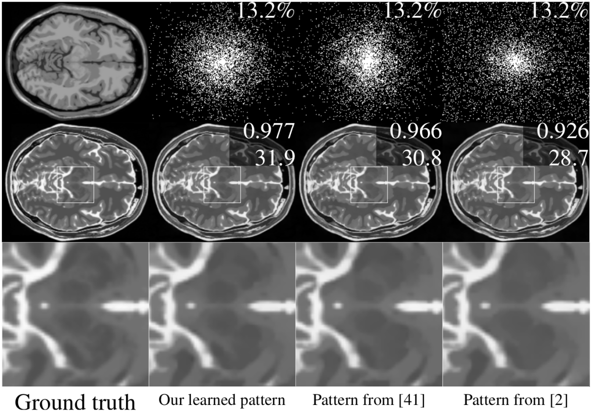
Compare regularizers







Compare "free" samplings



	Pattern type	SSIM	PSNR
Training	Our method	0.977 ± 0.002	32.5 ± 0.2
	Data-adapted [41]	0.968 ± 0.002	31.1 ± 0.1
	Uninformed VDS [2]	0.925 ± 0.005	28.9 ± 0.1
Testing	Our method	0.975 ± 0.003	32.1 ± 0.2
	Data-adapted [41]	0.967 ± 0.003	31.1 ± 0.2
	Uninformed VDS [2]	0.924 ± 0.003	28.8 ± 0.1

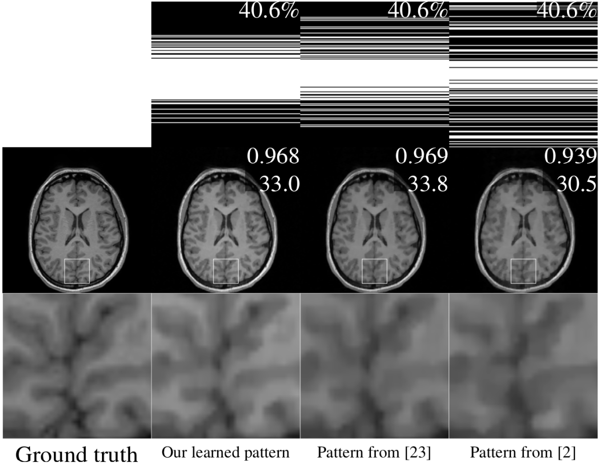
"ours" = Sherry et al. 2019

[41] = Knoll et al. 2011

[2] = Lustig et al. 2007

regularizer = dTV Ehrhardt and Betcke 2016

Compare Cartesian samplings



	Line sampling (40.6%)	Free pattern (34.7%)
Our method	4192	6494
The method from [23]	12087	$3.90 \cdot 10^8$

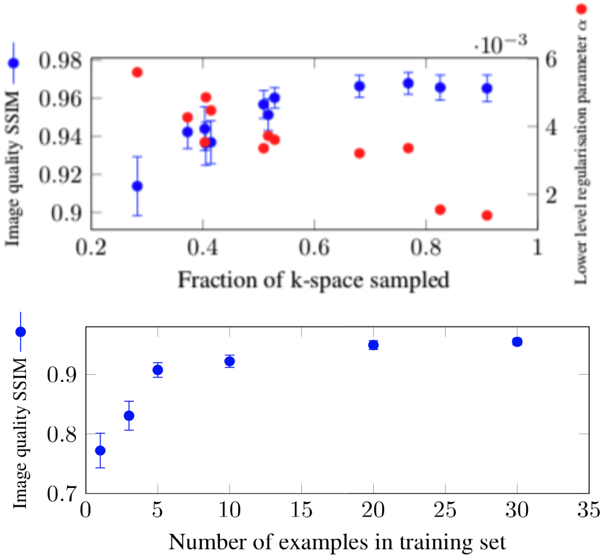
"ours" = Sherry et al. 2019

[23] = Gözcü et al. 2018

[2] = Lustig et al. 2007

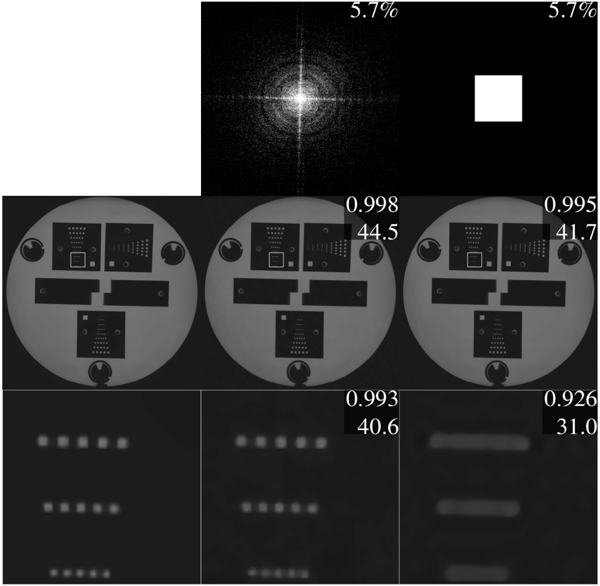
regularizer = TV

More insights: sampling and number of data



Sherry et al. 2019

High resolution imaging: 1024^2



Sherry et al. 2019

Conclusions and outlook

Conclusions

- Learn parameters via **Bilevel / machine learning**
- Learned sampling **better** than generic sampling
- "Optimal" sampling **depends on regularizer**
- **Very little data** needed

Outlook

- Investigate other **algorithms** tailored to problem
  - DFO with errors in objective (ongoing work with Lindon Roberts)
  - not based on reduced formulation, e.g. nonlinear ADMM
- **Unrolling**: replace lower level problem by algorithm
- **End-to-end learning**: learn reconstruction and sampling