On the barriers of AI and the trade-off between stability and accuracy in deep learning

Vegard Antun (Oslo, vegarant@math.uio.no)
Matthew J. Colbrook (Cambridge, m.colbrook@damtp.cam.ac.uk)

Joint work with:

Ben Adcock (SFU), Nina Gottschling (Cambridge), Anders Hansen (Cambridge), Clarice Poon (Bath), Francesco Renna (Porto)

Geilo Winter School, January 2021

MAIN GOAL

Determine the barriers of computations in deep learning
(i.e. what is and what is not possible)

$$\Downarrow$$

Stability and Accuracy in AI

Outline of lectures

DAY I	DAY II	Day III
Gravity of AI	Inverse Problems	Achieving Kernel Awareness
Image Classification	Instabilities & Kernel Awareness	FIRENETs
Need for Foundations	Intriguing Barriers	Imaging Applications
AI for Image Reconstruction	Algorithm Unrolling	Numerical Examples

Slides will be hosted at http://www.damtp.cam.ac.uk/user/mjc249/Talks.html.

Useful references for further reading in grey boxes.

Comments and suggestions welcome! (vegarant@math.uio.no, m.colbrook@damtp.cam.ac.uk)

Recap: Problem

Given measurements $y = Ax + e$, of $x \in \mathcal{M}_1 \subset \mathbb{C}^N$, recover x.

- In imaging $A \in \mathbb{C}^{m \times N}$ is a model of the sampling modality with $m < N$.
- x is the **unknown** signal of interest,
- and e is noise or perturbations.

Recap: How do we find sparse solutions?

Solve one of the problems:

Quadratically constrained basis pursuit (QCBP):

$$\min_{z\in\mathbb{C}^N} \|z\|_{l^1} \quad \text{subject to} \quad \|Az-y\|_{l^2} \leq \eta \qquad (P_1)$$

Unconstrained LASSO (U-LASSO):

$$\min_{z\in\mathbb{C}^N} \|Az-y\|_{l^2}^2 + \lambda\|z\|_{l^1} \qquad (P_2)$$

Square-root LASSO (SR-LASSO):

$$\min_{z\in\mathbb{C}^N} \|Az-y\|_{l^2} + \lambda\|z\|_{l^1} \qquad (P_3)$$

We let $\Xi_j(y, A)$ denote the set of minimizers for (P_j), given input $A \in \mathbb{C}^{m\times N}, y \in \mathbb{C}^m$.

Recap: Computational barriers

Nice classes $\Omega \subset \{(y, A) : y \in \mathbb{C}^m, A \in \mathbb{C}^{m\times N}\}$ where one can prove NNs with great approximation qualities exist. But:

- No algorithm, even randomised can train (or compute) such a NN accurate to K digits with probability greater than $1/2$.

Existence vs computation (universal approximation/interpolation theorems **not** enough).

Conclusion: Theorems on existence of neural networks may have little to do with the neural networks produced in practice.

Recap: Very crude reason why...

Let $f\colon \mathbb{R}^N \to \mathbb{R}$ be the function we want to minimize. Set

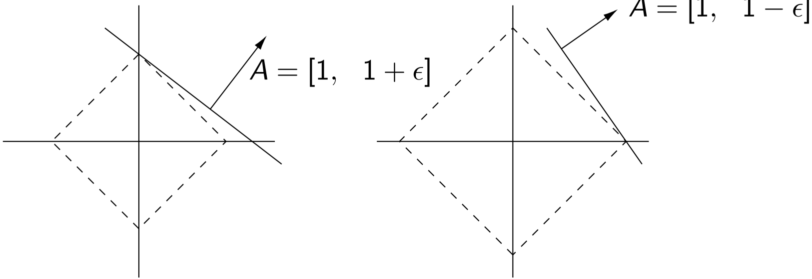
$$f^* = \min_{z \in \mathbb{R}^N} f(z).$$

Let $\hat{x}$ be a minimizer of f. Suppose $x \in \mathbb{R}^N$ satisfy

$$f(x) < f^* + \epsilon.$$

This does **not imply** that $\|x - \hat{x}\| \lesssim \epsilon$.

Recap: Very crude reason why...



Question: Can we find ‘good’ input classes where

$$f(x) < f^* + \epsilon \implies \|x - \hat{x}\| \lesssim \epsilon$$

We shall see that the answer is yes!

Robust null space property

Notation: Let $\Omega \subset \{1, \dots, N\}$ and let $P_\Omega \in \mathbb{R}^{N \times N}$ be the projection

$$P_\Omega x = \begin{cases} x_i & i \in \Omega \\ 0 & \text{otherwise} \end{cases}.$$

Definition (Robust Null Space Property)

A matrix $A \in \mathbb{C}^{m \times N}$ satisfies the robust Null Space Property (rNSP) of order $1 \leq s \leq N$ with constants $0 < \rho < 1$ and $\gamma > 0$ if

$$\|P_\Omega x\|_{l^2} \leq \frac{\rho}{\sqrt{s}} \|P_\Omega^\perp x\|_{l^1} + \gamma \|Ax\|_{l^2},$$

for all $x \in \mathbb{C}^N$ and any $\Omega \subseteq \{1, \dots, N\}$ with $|\Omega| \leq s$.

μ-suboptimality for SR-LASSO

Definition 1 (μ-suboptimality for SR-LASSO)

A vector $\tilde{x} \in \mathbb{C}^N$ is *μ-suboptimal* for the problem (P_3) if

$$\lambda \|\tilde{x}\|_{l^1} + \|A\tilde{x} - y\|_{l^2} \leq \mu + \min_{z \in \mathbb{C}^N} \left\{ \lambda \|z\|_{l^1} + \|Az - y\|_{l^2} \right\}.$$

μ-suboptimality + rNSP implies closeness to minimizer

Theorem 2

Suppose that $A \in \mathbb{C}^{m \times N}$ has the rNSP of order s with constants $0 < \rho < 1$ and $\gamma > 0$. Let $x \in \mathbb{C}^N$ and $y = Ax + e \in \mathbb{C}^m$ and

$$\lambda \leq \frac{C_1}{C_2\sqrt{s}},$$

where $C_1, C_2 > 0$ are constant depending only on ρ and γ. Then, every vector $\tilde{x} \in \mathbb{C}^N$ that is μ-suboptimal for $\min_{z \in \mathbb{C}^N} \lambda\|z\|_{l^1} + \|Az - y\|_{l^2}$ satisfies

$$\|\tilde{x} - x\|_{l^2} \leq 2C_1 \frac{\sigma_s(x)_{l^1}}{\sqrt{s}} + \frac{C_1}{\sqrt{s}\lambda}\mu + \left(\frac{C_1}{\sqrt{s}\lambda} + C_2\right)\|e\|_{l^2}.$$

See:
Adcock, B., & Hansen, A. C., '*Compressive Imaging: Structure, Sampling, Learning*', Cambridge University Press, 2021 (to appear). https://www.compressiveimagingbook.com

Theorem 3 (Universal Instability Theorem)

Let $A \in \mathbb{C}^{m \times N}$, *where* $m < N$, *and let* $\Psi : \mathbb{C}^m \to \mathbb{C}^N$ *be a continuous map. Suppose there are* $x, x' \in \mathbb{C}^N$ *and* $\eta > 0$ *such that*

$$\|\Psi(Ax) - x\| < \eta, \quad \text{and} \quad \|\Psi(Ax') - x'\| < \eta, \tag{1}$$

and

$$\|Ax - Ax'\| < \eta. \tag{2}$$

We then have the following:

(i) **(Instability with respect to worst-case perturbations)** *Then the local* ε*-Lipschitz constant at* $y = Ax$ *satisfies*

$$L^\varepsilon(\Psi, y) := \sup_{0<\|z-y\|\leq\varepsilon} \frac{\|\Psi(z) - \Psi(y)\|}{\|z - y\|} \geq \frac{1}{\varepsilon}\left(\|x - x'\| - 2\eta\right), \qquad \forall \varepsilon \geq \eta. \tag{3}$$

See: Gottschling, Antun, Adcock, and Hansen, 2020. *The troublesome kernel: why deep learning for inverse problems is typically unstable.* arXiv:2001.01258.

rNSP $\implies$ kernel awareness for sparse vectors

Theorem 4

Suppose the matrix $A \in \mathbb{C}^{m \times N}$ satisfies the robust null space property (rNSP) or order s, with constants $0 < \rho < 1$ and $\gamma > 0$. Then for all s-sparse vectors $x, z \in \mathbb{C}^N$,

$$\|z - x\|_{l^2} \leq \frac{C_2}{2} \|A(z - x)\|_{l^2}$$

where

$$C_2 = \frac{(3\rho + 5)\gamma}{1 - \rho}. \tag{4}$$

See:
Foucart, S., & Rauhut, H., '*A Mathematical Introduction to Compressive Sensing*', birkhäuser, 2013.

Typical compressive sensing theorem

Theorem 5

Let $A \in \mathbb{C}^{m\times N}$ *with* $m < N$ *and let* $W \in \mathbb{C}^{N\times N}$ *be unitary. Suppose that* AW^{-1} *has the rNSP of order* s *with constants* $0 < \rho < 1$ *and* $\gamma > 0$*. Let* $y = Ax + e$ *and let* $0 < \lambda \leq C_1/(\sqrt{s}C_2)$*. Then every minimizer* $\hat{x} \in \mathbb{C}^N$ *of the problem*

$$\min_{z\in\mathbb{C}^N} \lambda\|Wz\|_{l^1} + \|Az - y\|_{l^2} \tag{P$_3$}$$

satisfies

$$\|\hat{x} - x\|_{l^2} \leq 2C_1 \frac{\sigma_s(Wx)_{l^1}}{\sqrt{s}} + \left(\frac{C_1}{\sqrt{s}\lambda} + C_2\right)\|e\|_{l^2},$$

where C_1 *and* C_2 *are the constants in* (4)*, and*

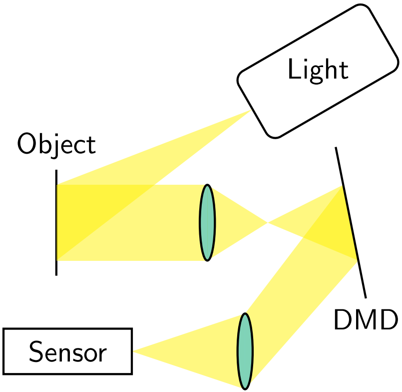
$$\sigma_s(z)_{l^1} := \inf\{\|z - t\|_{l^1} : t \textit{ is a s-sparse vector}\}$$

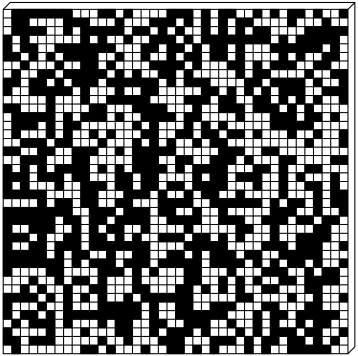
denotes the distance to a s-sparse vector.

Do the matrices that we use in imaging have the robust null space property?

Example 1: Binary imaging

Examples: Fluorescence microscopy and single-pixel imaging

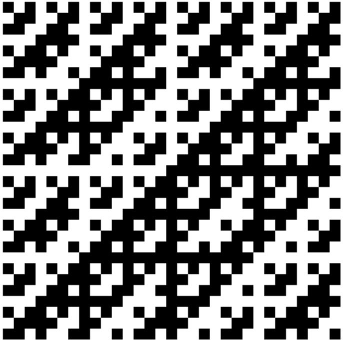




DMD

Example 1: Binary imaging – Walsh-Hadamard sampling

Three different ordering of the Hadamard matrix $U_{\mathrm{had}} \in \mathbb{R}^{N\times N}$.



We select a subset $\Omega \subset \{1,\ldots,N\}$, $|\Omega| = m$, of the rows $P_\Omega U_{\mathrm{had}}$.

Example 2: Fourier Sampling – MRI

Many sampling modalities can be modeled by the Fourier transform

$$\mathcal{F}f(\omega) = \int_{[0,1]^2} f(t) e^{-2\pi i \omega \cdot t} \, dt,$$

We discretize this integral to get a linear system

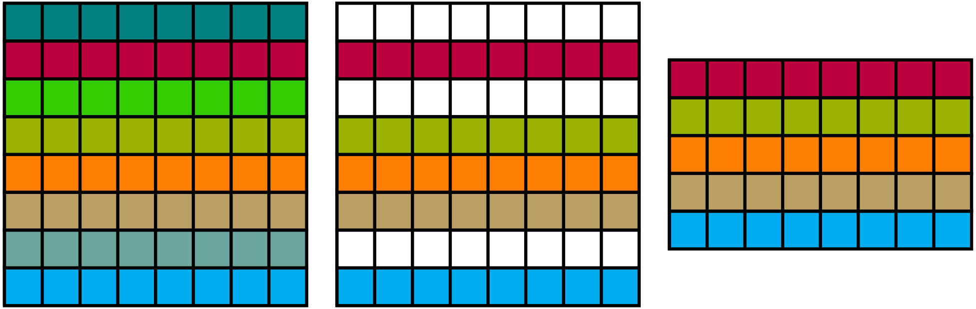
$$\mathcal{F}f(\omega_1, \omega_2) \approx \sum_{k=0}^{N-1} \sum_{j=0}^{N-1} x_{j,k} \frac{1}{N} e^{2\pi i(\omega_1 j + \omega_2 k)/N}$$

where $x_{j,k} = f(k/N, l/N)$ and $\omega = (\omega_1, \omega_2) \in \{-N/2 + 1, \ldots, N/2\}^2$. We write this system as

$$y = U_{\mathrm{dft}} x$$

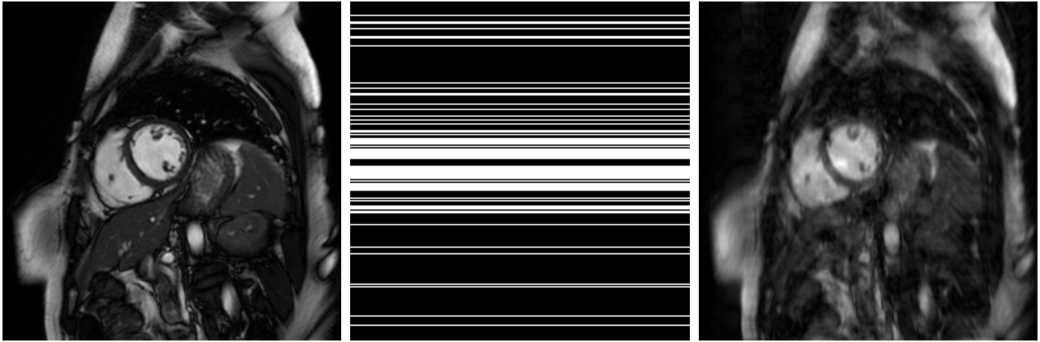
where $U_{\mathrm{dft}} \in \mathbb{C}^{N^2 \times N^2}$ is the Fourier matrix. This matrix is unitary.

The matrix $P_\Omega U$ with $\Omega = \{2, 4, 5, 6, 8\}$



Example 2: Fourier Sampling – MRI

Let $A = P_\Omega F$ and $y = Ax$.



Original x — Sampling pattern Ω — Adjoint: $A^* y$

Sparse regularization in imaging

- Given the linear system

$$Ux_0 = y.$$

- Solve

$$\min_{z \in \mathbb{C}^N} \lambda \|z\|_{l^1} + \|P_\Omega U z - P_\Omega y\|_{l^2} \tag{P$_3$}$$

- In imaging we use for example $U = U_{\mathrm{dft}} U_{\mathrm{dwt}}^{-1}$

Original image

$d = U_{\mathrm{dwt}}^{-1} x_0$

5% of the w. coeff.

$P_{\widetilde{\Omega}} x_0$

Compressed image

$\tilde{d} = U_{\mathrm{dwt}}^{-1} P_{\widetilde{\Omega}} x_0$

Sparse regularization in imaging

- Given the linear system
$$Ux_0 = y.$$
- Solve
$$\min_{z\in\mathbb{C}^N} \lambda\|z\|_{l^1} + \|P_\Omega Uz - P_\Omega y\|_{l^2}$$
where P_Ω is a projection and $\Omega \subset \{1,\dots,N\}$ is subsampled with $|\Omega| = m$.

Traditional idea: If U is unitary, Ω is chosen uniformly at random and

$$m \gtrsim N \cdot \mu(U) \cdot s \cdot L(\epsilon^{-1}, s, N)$$

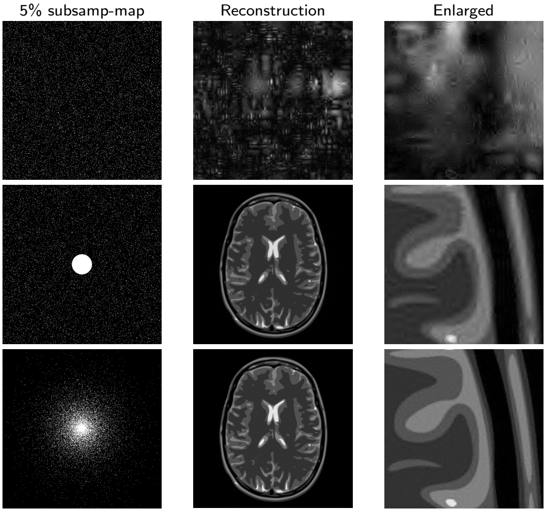
then with probability $1-\epsilon$, $P_\Omega U$ has the robust null space property (rNSP) of order s (with certain constants). Here

$$\mu(U) := \max_{i,j} |U_{i,j}|^2 \in [1/N, 1]$$

is referred to as the incoherence parameter and $L(\epsilon^{-1}, s, N)$ is a polylogarithmic factor.

Uniform Random Subsampling

$$U = U_{\mathrm{dft}} V_{\mathrm{dwt}}^{-1}.$$



Sparsity

- The classical idea of sparsity in sparse regularization is that there are s important coefficients in the vector x_0 that we want to recover.
- The location of these coefficients is arbitrary.

The Flip Test and the rNSP

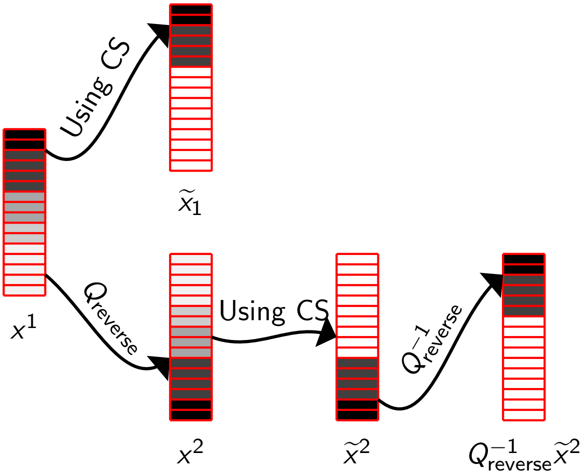


Figure from: Bastounis, A. & H. C, Anders Christian (2017). *On the absence of uniform recovery in many real-world applications of compressed sensing and the restricted isometry property and nullspace property in levels*. SIAM Journal of Imaging Sciences.

Sparsity - The Flip Test

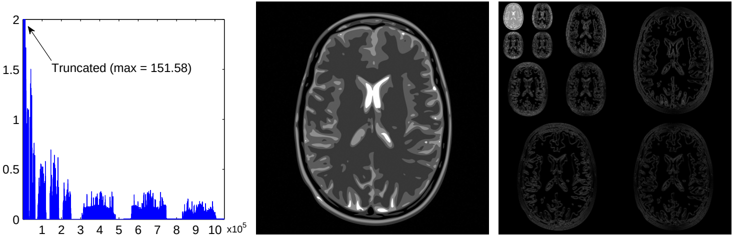
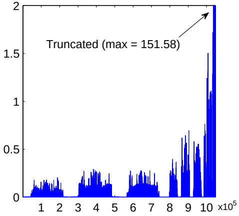


Figure: Wavelet coefficients and subsampling reconstructions from 10% of Fourier coefficients with distributions $(1 + \omega_1^2 + \omega_2^2)^{-1}$ and $(1 + \omega_1^2 + \omega_2^2)^{-3/2}$.

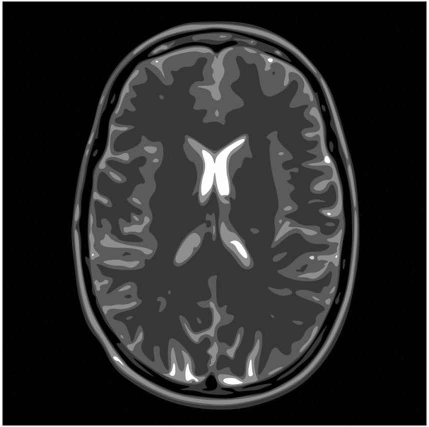
If sparsity is the right model we should be able to flip the coefficients. Let

$$z_f =$$

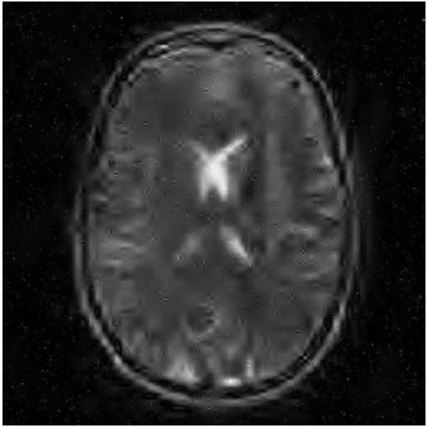


Sparsity- The Flip Test: Results

Rec. not flipped coeff.



Rec. flipped coeff.



Conclusion: The ordering of the coefficients did matter. Moreover, this phenomenon happens with all wavelets, curvelets, contourlets and shearlets and any reasonable subsampling scheme.

Question: Is sparsity really the right model?

The Flip Test and the rNSP

CS reconstr. | CS reconstr, w/ flip coeffs. | Subsampling pattern

2048, 12%

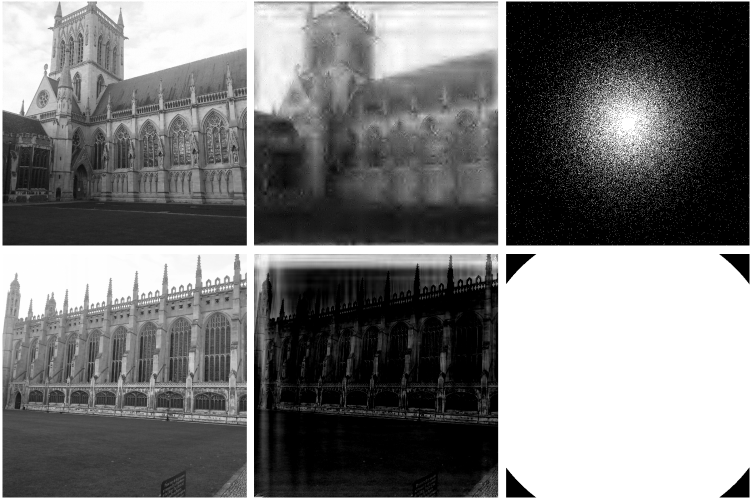
$U_{\mathrm{dft}} V_{\mathrm{dwt}}^{-1}$

Magnetic Resonance Imaging

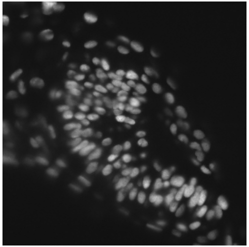
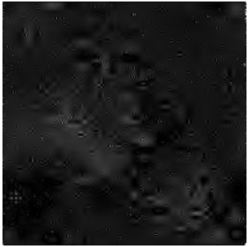
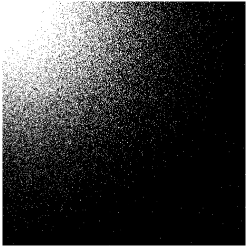
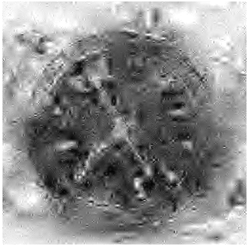
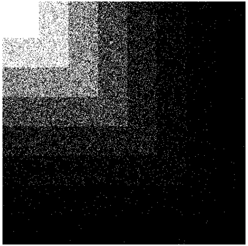
2048, 97%

$U_{\mathrm{dft}} V_{\mathrm{dwt}}^{-1}$

Magnetic Resonance Imaging

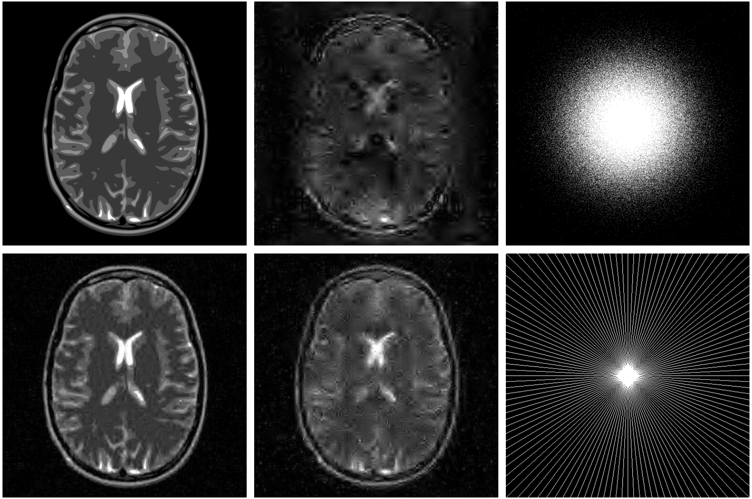


Sparsity - The Flip Test

	CS reconstr.	CS reconstr, w/ flip coeffs.	Subsampling pattern
512, 20% $U_{\mathrm{Had}} V_{\mathrm{dwt}}^{-1}$ Fluorescence Microscopy			
1024, 12% $U_{\mathrm{Had}} V_{\mathrm{dwt}}^{-1}$ Compressive Imaging, Hadamard Spectroscopy			

Sparsity - The Flip Test (contd.)

CS reconstr. | CS reconstr, w/ flip coeffs. | Subsampling pattern



1024, 20%
$U_{\mathrm{dft}} V_{\mathrm{dwt}}^{-1}$
Magnetic Resonance Imaging

512, 12%
$U_{\mathrm{dft}} V_{\mathrm{dwt}}^{-1}$
Tomography, Electron Microscopy

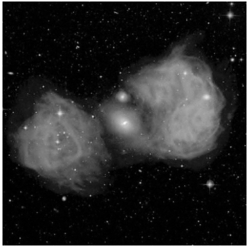
Sparsity - The Flip Test (contd.)

1024, 10%

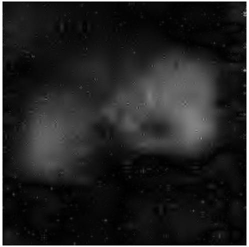
$U_{\mathrm{dft}} V_{\mathrm{dwt}}^{-1}$

Radio interferometry

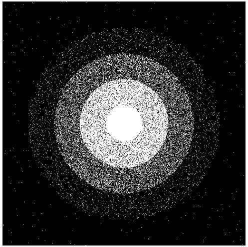
CS reconstr.



CS reconstr, w/ flip coeffs.



Subsampling pattern



The Flip Test and the rNSP

		Matrix method		rNSP
		$\mathrm{DFT} \cdot \mathrm{DWT}^{-1}$	$\mathrm{HAD} \cdot \mathrm{DWT}^{-1}$	
Problem	MRI	✓	✗	✗
	Tomography	✓	✗	✗
	Spectroscopy	✓	✗	✗
	Electron microscopy	✓	✗	✗
	Radio interferometry	✓	✗	✗
	Fluorescence microscopy	✗	✓	✗
	Lensless camera	✗	✓	✗
	Single pixel camera	✗	✓	✗
	Hadamard spectroscopy	✗	✓	✗

Table: A table displaying various applications of compressive sensing. For each application, a suitable matrix is suggested along with information on whether or not that matrix has the rNSP of a sufficiently large order s.

Sparse regularization in imaging

- Given the linear system
$$Ux_0 = y.$$
- Solve
$$\min_{z\in\mathbb{C}^N} \lambda\|z\|_{l^1} + \|P_\Omega Uz - P_\Omega y\|_{l^2}$$
where P_Ω is a projection and $\Omega \subset \{1,\ldots,N\}$ is subsampled with $|\Omega| = m$.

Traditional idea: If U is unitary, Ω is chosen uniformly at random and

$$m \gtrsim N \cdot \mu(U) \cdot s \cdot L(\epsilon^{-1}, s, N)$$

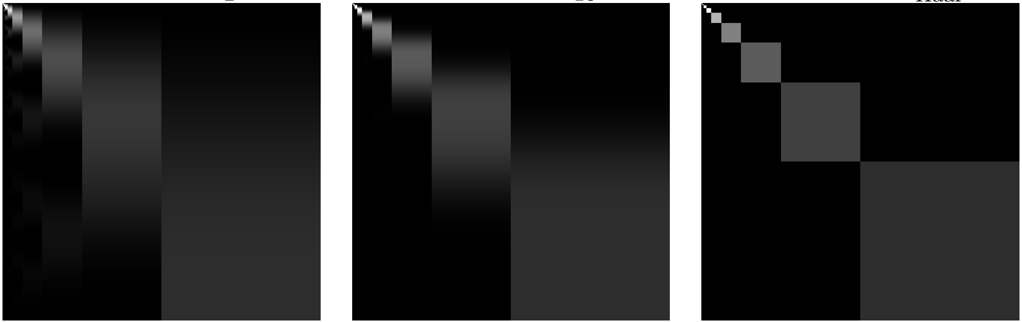
then with probability $1-\epsilon$, $P_\Omega U$ has the robust null space property (rNSP) of order s (with certain constants). Here

$$\mu(U) := \max_{i,j} |U_{i,j}|^2 \in [1/N, 1]$$

is referred to as the incoherence parameter and $L(\epsilon^{-1}, s, N)$ is a polylogarithmic factor.

What kind of structure do we have?

$\mathrm{DFT} \cdot \mathrm{DWT}_{2}^{-1}$ $\mathrm{DFT} \cdot \mathrm{DWT}_{10}^{-1}$ $\mathrm{HAD} \cdot \mathrm{DWT}_{\mathrm{Haar}}^{-1}$



The three images display the absolute values of various sensing matrices. A lighter colour represents larger absolute values. Here DFT is the Discrete Fourier Transform, HAD the Hadamard transform and $\mathrm{DWT}_{\mathrm{N}}^{-1}$ the Inverse Wavelet Transform corresponding to Daubechies wavelets with N vanishing moments.

Reading material

- Adcock, B., & Hansen, A. C., '*Compressive Imaging: Structure, Sampling, Learning*', Cambridge University Press, 2021 (to appear). `https://www.compressiveimagingbook.com`
- Bastounis, A., Adcock, B., & Hansen, A. C. (2017). '*From global to local: Getting more from compressed sensing*'. SIAM News, Oct.
- Adcock, B., Hansen, A. C., Poon, C., & Roman, B. (2017). '*Breaking the coherence barrier: A new theory for compressed sensing*'. In Forum of Mathematics, Sigma (Vol. 5). Cambridge University Press.
- Adcock, B., Antun, V., & Hansen, A. C. (2019). '*Uniform recovery in infinite-dimensional compressed sensing and applications to structured binary sampling*'. arXiv:1905.00126.
- Roman, B., Hansen, A., & Adcock, B. (2014). '*On asymptotic structure in compressed sensing*'.arXiv:1406.4178.

Sparsity in levels

Definition 6 (Sparsity in levels)

Let $\mathbf{M} = (M_1, \ldots, M_r) \in \mathbb{N}^r$, where $1 \leq M_1 < \cdots < M_r = N$, and $\mathbf{s} = (s_1, \ldots, s_r) \in \mathbb{N}_0^r$, where $s_k \leq M_k - M_{k-1}$ for $k = 1, \ldots, r$ and $M_0 = 0$. A vector $x \in \mathbb{C}^N$ is $(\mathbf{s}, \mathbf{M})$-sparse in levels if

$$|\mathrm{supp}(x) \cap \{M_{k-1} + 1, ..., M_k\}| \leq s_k, \quad k = 1, ..., r.$$

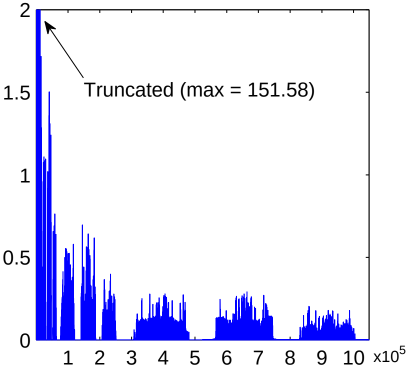
The total sparsity is $s = s_1 + ... + s_r$. We denote the set of $(\mathbf{s}, \mathbf{M})$-sparse vectors by $\Sigma_{\mathbf{s},\mathbf{M}}$. We also define the following measure of distance of a vector x to $\Sigma_{\mathbf{s},\mathbf{M}}$ by

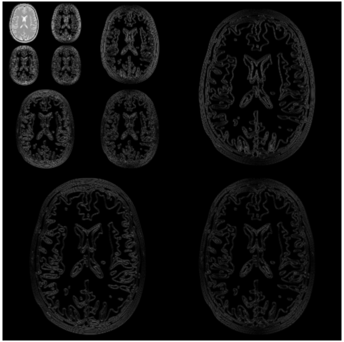
$$\sigma_{\mathbf{s},\mathbf{M}}(x)_{l^1_w} = \inf\{\|x - z\|_{l^1_w} : z \in \Sigma_{\mathbf{s},\mathbf{M}}\}.$$

Here $\|z\|_{l^1_w} := \sum_{j=1}^{N} w_j |z_j|$, is the weighted l^1-norm for positive weights $\{w_j\}$.

Sparsity - The Flip Test in Levels

Let





denote the vector of the wavelet coefficients. Let z_f^L denote the flipped version of z where the flipping of coefficients only happens within the levels.

Sparsity - The Flip Test in Levels

- Let
$$\tilde{y} = U_{\text{dft}} U_{\text{dwt}}^{-1} z_f^L$$
- Solve
$$\min_{z \in \mathbb{C}^N} \lambda \|z\|_{l^1} + \|P_\Omega U_{\text{dft}} U_{\text{dwt}}^{-1} z - P_\Omega \tilde{y}\|_{l^2} \qquad (\mathrm{P}_3)$$
to get $\hat{z}_f^L$.
- Flip the coefficients of $\hat{z}_f^L$ back to get $\hat{z}$, and let $\hat{x} = U_{\text{dwt}}^{-1} \hat{z}$.

The Flip Test in levels

CS reconstr.

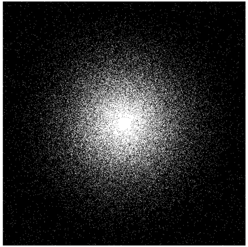
CS rec., w/ flip (levels) coeffs.

Subsampling pattern

2048, 12%

$U_{\mathrm{dft}} V_{\mathrm{dwt}}^{-1}$

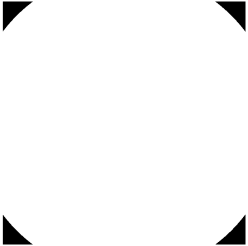
Magnetic Resonance Imaging



2048, 97%

$U_{\mathrm{dft}} V_{\mathrm{dwt}}^{-1}$

Magnetic Resonance Imaging

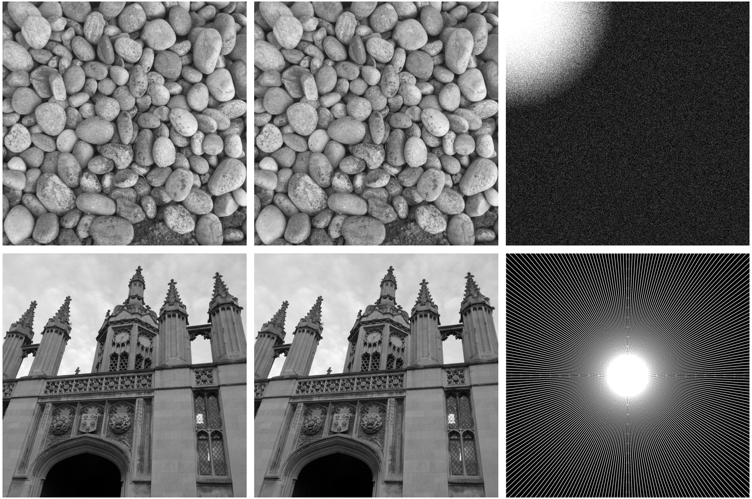


The Flip Test in levels

CS reconstr. | CS rec., w/ flip (levels) coeffs. | Subsampling pattern

2048, 12%
$U_{\mathrm{Had}} V_{\mathrm{dwt}}^{-1}$
Fluorescence microscopy

2048, 27.5%
$U_{\mathrm{dft}} V_{\mathrm{dwt}}^{-1}$
Tomography



The weighted Robust Nullspace Property in Levels (wrNSPL)

Definition 7 (wrNSP in levels)

Let $(\mathbf{s}, \mathbf{M})$ be local sparsities and sparsity levels respectively. For weights $\{w_j\}_{j=1}^N$ ($w_j > 0$), we say that $A \in \mathbb{C}^{m \times N}$ satisfies the weighted robust null space property in levels (wrNSPL) of order $(\mathbf{s}, \mathbf{M})$ with constants $0 < \rho < 1$ and $\gamma > 0$ if for any $(\mathbf{s}, \mathbf{M})$ support set Ω,

$$\|P_\Omega x\|_{l^2} \leq \frac{\rho \|P_{\Omega^c} x\|_{l^1_w}}{\sqrt{\xi}} + \gamma \|Ax\|_{l^2}, \qquad \text{for all } x \in \mathbb{C}^N.$$

Some key points so far . . .

- In general no NN can solve the problems (P_j), $j = 1, 2, 3$ for arbitrary input, but if A has the rNSP or wrNSPL we can.
- The assumption of sparsity and uniformly random subsampling is too general to explain the success of sparse regularization in imaging. Additional structure is needed!
- The wrNSPL provide sufficient conditions for kernel awareness for images which are sparse in wavelets.
- By sampling in a structured way we can achieve the wrNSPL.

Fast Iterative REstarted NETworks
(FIRENETs)

The model

Definition [Sparsity in levels]: Let $\mathbf{M} = (M_1, \dots, M_r) \in \mathbb{N}^r$, where $1 \le M_1 < \cdots < M_r = N$, and $\mathbf{s} = (s_1, \dots, s_r) \in \mathbb{N}_0^r$, where $s_k \le M_k - M_{k-1}$ for $k = 1, \dots, r$ and $M_0 = 0$. A vector $x \in \mathbb{C}^N$ is $(\mathbf{s}, \mathbf{M})$-sparse in levels if

$$|\mathrm{supp}(x) \cap \{M_{k-1} + 1, ..., M_k\}| \le s_k, \quad k = 1, ..., r.$$

The total sparsity is $s = s_1 + ... + s_r$. We denote the set of $(\mathbf{s}, \mathbf{M})$-sparse vectors by $\Sigma_{\mathbf{s},\mathbf{M}}$. We also define the following measure of distance of a vector x to $\Sigma_{\mathbf{s},\mathbf{M}}$ by

$$\sigma_{\mathbf{s},\mathbf{M}}(x)_{l^1_w} = \inf\{\|x - z\|_{l^1_w} : z \in \Sigma_{\mathbf{s},\mathbf{M}}\}.$$

For simplicity, assume $s_k > 0$ and l^1_w weights constant in each level:

$$w_i = w_{(j)}, \quad \text{if } M_{j-1} + 1 \le i \le M_j.$$

Kernel awareness: the robust nullspace property

Definition [weighted rNSP in levels]: Let $(\mathbf{s}, \mathbf{M})$ be local sparsities and sparsity levels respectively. For weights $\{w_i\}_{i=1}^N$ $(w_i > 0)$, we say that $A \in \mathbb{C}^{m \times N}$ satisfies the weighted robust null space property in levels (weighted rNSPL) of order $(\mathbf{s}, \mathbf{M})$ with constants $0 < \rho < 1$ and $\gamma > 0$ if for any $(\mathbf{s}, \mathbf{M})$ support set Δ,

$$\|x_\Delta\|_{l^2} \leq \frac{\rho \|x_{\Delta^c}\|_{l^1_w}}{\sqrt{\xi}} + \gamma \|Ax\|_{l^2}, \qquad \text{for all } x \in \mathbb{C}^N.$$

The goal of this section

Simplified version of Theorem: *We provide an algorithm such that:*

Input: *Sparsity parameters* $(\mathbf{s}, \mathbf{M})$, *weights* $\{w_i\}_{i=1}^N$, $A \in \mathbb{C}^{m \times N}$ *(with the input* A *given by* $\{A_l\}$*) satisfying the rNSPL with constants* $0 < \rho < 1$ *and* $\gamma > 0$, $n \in \mathbb{N}$ *and positive* $\{\delta, b_1, b_2\}$.

Output: *A neural network* ϕ_n *with* $\mathcal{O}(n)$ *layers and the following property.*

For any $x \in \mathbb{C}^N$ *and* $y \in \mathbb{C}^m$ *with*

$$\underbrace{\sigma_{\mathbf{s},\mathbf{M}}(x)_{l_w^1}}_{\text{distance to sparse in levels vectors}} + \underbrace{\|Ax - y\|_{l^2}}_{\text{noise of measurements}} \lesssim \delta, \quad \|x\|_{l^2} \lesssim b_1, \quad \|y\|_{l^2} \lesssim b_2,$$

we have the following **stable** *and* **exponential convergence** *guarantee in* n

$$\|\phi_n(y) - x\|_{l^2} \lesssim \delta + e^{-n}.$$

Comments

- Strategy: restarted & reweighted unrolling of primal-dual algorithm applied to:

$$(P_3) \quad \mathrm{argmin}_{x\in\mathbb{C}^N} F_3^A(x,y,\lambda) := \lambda\|x\|_{l^1_w} + \|Ax-y\|_{l^2}.$$

- As well as stability, rNSPL allows exponential convergence.
- Even ignoring stability, naive unrolling of iterative methods only gives slow convergence $\mathcal{O}(\delta + n^{-1})$ (and in certain regimes $\mathcal{O}(\delta + n^{-2})$).
- If we do not know ρ or γ (constants for rNSPL), can perform log-scale grid search for suitable parameters (increase number of layers by a factor of $\log(n)$). Sometimes (see below) we know ρ and γ with probabilistic bounds.

Precise definition of neural network

$\phi\colon \mathbb{C}^m \to \mathbb{C}^N$ s.t. $\phi(y) = V_T(\rho_{T-1}(...\rho_1(V_1(y))))$, and

- Each V_j is an affine map $\mathbb{C}^{N_{j-1}} \to \mathbb{C}^{N_j}$ given by $V_j(x) = W_j x + b_j(y)$ where $W_j \in \mathbb{C}^{N_j \times N_{j-1}}$ and the $b_j(y) = R_j y + c_j \in \mathbb{C}^{N_j}$ are affine functions of the input y.
- Each $\rho_j\colon \mathbb{C}^{N_j} \to \mathbb{C}^{N_j}$ is one of two forms:
  (i) $I_j \subset \{1, ..., N_j\}$ s.t. ρ_j applies $f_j : \mathbb{C} \to \mathbb{C}$ element-wise on components with indices in I_j:

$$\rho_j(x)_k = \begin{cases} f_j(x_k), & \text{if } k \in I_j \\ x_k, & \text{otherwise.} \end{cases}$$

  (ii) $f_j : \mathbb{C} \to \mathbb{C}$ s.t. after decomposing the input vector x as $(x_0, X^\top, Y^\top)^\top$ for scalar x_0, $X \in \mathbb{C}^{m_j}$, $Y \in \mathbb{C}^{N_j - 1 - m_j}$,

$$\rho_j : \begin{pmatrix} x_0 \\ X \\ Y \end{pmatrix} \to \begin{pmatrix} 0 \\ f_j(x_0)X \\ Y \end{pmatrix}.$$

Precise definition of neural network

$$\begin{pmatrix} x_0 \\ X \\ Y \end{pmatrix} \to \begin{pmatrix} f_j(x_0) \\ X \\ Y \end{pmatrix} \to \begin{pmatrix} f_j(x_0)\mathbf{1} \\ X \\ f_j(x_0)\mathbf{1} + X \\ Y \end{pmatrix}$$

$$\to \begin{pmatrix} f_j(x_0)^2\mathbf{1} \\ X^2 \\ [f_j(x_0)\mathbf{1} + X]^2 \\ Y \end{pmatrix}$$

$$\to \begin{pmatrix} 0 \\ \frac{1}{2}\left[[f_j(x_0)\mathbf{1} + X]^2 - f_j(x_0)^2\mathbf{1} - X^2\right] = f_j(x_0)X \\ Y \end{pmatrix}.$$

Precise definition of neural network

- Recall that we assume knowledge $A_l \in \mathbb{Q}[i]^{m\times N}$ such that
$$\|A_l - A\| \leq 2^{-l}, \quad \forall l \in \mathbb{N}.$$
- Our nonlinear activation functions will be built using square roots. We assume that we have access to a routine "sqrt_θ" such that $|\mathrm{sqrt}_\theta(x) - \sqrt{x}| \leq \theta$.
- An interpretation of θ: numerical stability, or accumulation of errors, of the forward pass of the NN. A key point is that θ doesn't need to be small.

For brevity, will ignore these points in presentation below.

Step 1: Preliminary constructions

$$\psi^0_\beta(x) = \max\left\{0, 1 - \frac{\beta}{\|x\|_{l^2}}\right\} x, \quad \psi^1(x) = \min\left\{1, \frac{1}{\|x\|_{l^2}}\right\} x.$$

Lemma: Let $M \in \mathbb{N}$, $\beta \in \mathbb{Q}_{>0}$ and $\theta \in \mathbb{Q}_{>0}$. Then there exists NNs $\phi^0_{\beta,\theta}, \phi^1_\theta$ with $T = 3$ s.t.

$$\left\|\phi^0_{\beta,\theta}(x) - \psi^0_\beta(x)\right\|_{l^2} \le \theta, \quad \left\|\phi^1_\theta(x) - \psi^1(x)\right\|_{l^2} \le \theta.$$

E.g. $\phi^0_{\beta,\theta} : x \xrightarrow{\text{L}} \begin{pmatrix} x \\ x \end{pmatrix} \xrightarrow{\text{NL}} \begin{pmatrix} |x_1|^2 \\ |x_2|^2 \\ \vdots \\ |x_M|^2 \\ x \end{pmatrix} \xrightarrow{\text{L}} \begin{pmatrix} \sum_{j=1}^M |x_j|^2 \\ x \end{pmatrix} \xrightarrow{\text{NL}} \begin{pmatrix} 0 \\ \max\left\{0, 1 - \frac{\beta}{\mathrm{sqrt}_\theta(\|x\|^2_{l^2})}\right\} x \end{pmatrix}$

$$\xrightarrow{\text{L}} \max\left\{0, 1 - \frac{\beta}{\mathrm{sqrt}_\theta(\|x\|^2_{l^2})}\right\} x.$$

Step 1: Preliminary constructions

Lemma: Let $s, \theta \in \mathbb{Q}_{>0}$, $w \in \mathbb{Q}^N_{>0}$ and for $\hat{x} \in \mathbb{C}^N$ consider the minimisation problem

$$\text{argmin}_{x \in \mathbb{C}^N} \|x\|_{l^1_w} + s\|x - \hat{x}\|^2_{l^2}. \tag{5}$$

Let $\tilde{x}_s(\hat{x})$ be the solution of (5). Then, there exists NNs $\phi_{s,\theta}$ ($T = 2$) s.t.

$$\|\phi_{s,\theta}(\hat{x}) - \tilde{x}_s(\hat{x})\|_{l^2} \leq \theta \|w\|_{l^2}.$$

Proof.

Fun exercise in algorithm unrolling! □

Step 2: Unrolling primal-dual iterations

X, Y finite-dimensional real vectors spaces, $K : X \to Y$ linear

$$\min_{x \in X} \max_{y \in Y} \langle Kx, y \rangle + G(x) - F^*(y)$$

For convex $H : Z \to [0, \infty]$, define

$$(I + \tau \partial H)^{-1}(w) = \mathrm{argmin}_z H(z) + \frac{\|z - w\|_{l^2}^2}{2\tau}$$

If easy to compute for $H = G, F$, then iterate updates of primal and dual variables.

Chambolle, A. and Pock, T., 2011. *A first-order primal-dual algorithm for convex problems with applications to imaging.* Journal of mathematical imaging and vision, 40(1), pp.120-145.

Step 2: Unrolling primal-dual iterations

122 J Math Imaging Vis (2011) 40: 120–145

Algorithm 1

- Initialization: Choose $\tau, \sigma > 0$, $\theta \in [0, 1]$, $(x^0, y^0) \in X \times Y$ and set $\bar{x}^0 = x^0$.
- Iterations ($n \geq 0$): Update $x^n, y^n, \bar{x}^n$ as follows:

$$\begin{cases} y^{n+1} = (I + \sigma \partial F^*)^{-1}(y^n + \sigma K \bar{x}^n) \\ x^{n+1} = (I + \tau \partial G)^{-1}(x^n - \tau K^* y^{n+1}) \\ \bar{x}^{n+1} = x^{n+1} + \theta(x^{n+1} - x^n) \end{cases} \tag{7}$$

We can use previous constructions for the proximal maps!

$\Rightarrow$ unrolled primal-dual iterations

Chambolle, A. and Pock, T., 2016. *On the ergodic convergence rates of a first-order primal–dual algorithm.* Mathematical Programming, 159(1-2), pp.253-287.

Step 2: Unrolling primal-dual iterations

Theorem: Suppose $L_A \geq 1$ is an upper bound for $\|A\|$, and that $\tau, \sigma > 0$ are such that $\tau\sigma L_A^2 < 1$. Let $p \in \mathbb{N}$, then there exists an algorithm that constructs a sequence of neural networks $\phi_{p,\lambda}^A$ (each with $T = \mathcal{O}(p)$) such that:

(i) $\phi_{p,\lambda}^A : \mathbb{C}^{m+N} \to \mathbb{C}^N$ takes an input $y \in \mathbb{C}^m$ and an initial guess $x_0 \in \mathbb{C}^N$.

(ii) For any inputs $y \in \mathbb{C}^m$ and $x_0 \in \mathbb{C}^N$, and for any $x \in \mathbb{C}^N$,

$$\underbrace{\lambda\|\phi_{p,\lambda}^A(y,x_0)\|_{l_w^1} + \|A\phi_{p,\lambda}^A(y,x_0) - y\|_{l^2}}_{F_3^A(\phi_{p,\lambda}^A(y,x_0),y,\lambda)} \underbrace{-\lambda\|x\|_{l_w^1} - \|Ax - y\|_{l^2}}_{-F_3^A(x,y,\lambda)} \leq \frac{1}{p}\left(\frac{\|x - x_0\|_{l^2}^2}{\tau} + \frac{1}{\sigma}\right).$$

$$(P_3) \quad \operatorname{argmin}_{x\in\mathbb{C}^N} F_3^A(x, y, \lambda) := \lambda\|x\|_{l_w^1} + \|Ax - y\|_{l^2}.$$

Step 3: "Recalling" some compressed sensing results

$$\xi := \sum_{k=1}^{r} w_{(k)}^2 s_k, \quad \zeta := \min_{k=1,\dots,r} w_{(k)}^2 s_k, \quad \kappa := \frac{\xi}{\zeta}.$$

$$\begin{aligned} \text{rNSPL} \Rightarrow \|z_1 - z_2\|_{l^2} \le & \frac{2C_1}{\sqrt{\xi}} \sigma_{\mathbf{s},\mathbf{M}}(z_2)_{l^1_w} + 2C_2\|Az_2 - y\|_{l^2} \\ & + \frac{C_1}{\lambda\sqrt{\xi}} \left(\lambda\|z_1\|_{l^1_w} + \|Az_1 - y\|_{l^2} - \lambda\|z_2\|_{l^1_w} - \|Az_2 - y\|_{l^2}\right), \end{aligned} \tag{6}$$

$$\begin{aligned} \text{Set} \quad G(z_1, z_2, y) &:= \lambda\|z_1\|_{l^1_w} + \|Az_1 - y\|_{l^2} - \lambda\|z_2\|_{l^1_w} - \|Az_2 - y\|_{l^2}, \\ &= F_3^A(z_1, y, \lambda) - F_3^A(z_2, y, \lambda) \\ c(z, y) &:= \frac{2C_1}{C_2\sqrt{\xi}} \cdot \sigma_{\mathbf{s},\mathbf{M}}(z)_{l^1_w} + 2\|Az - y\|_{l^2}. \end{aligned}$$

Choosing $\lambda \le C_1/(C_2\sqrt{\xi})$,

$$\|z_1 - z_2\|_{l^2} \le \frac{C_1}{\lambda\sqrt{\xi}} \left(c(z_2, y) + G(z_1, z_2, y)\right), \tag{7}$$

which holds for completely general z_1, z_2 and y.

Step 4: Combine with constructed neural networks

Define the following map from unrolled primal-dual iterations

$$H_p^\beta : \mathbb{C}^m \times \mathbb{C}^N \to \mathbb{C}^N, \quad H_p^\beta(y, x_0) = p\beta\phi_{p,\lambda}^A\left(\frac{y}{p\beta}, \frac{x_0}{p\beta}\right).$$

Use previous theorem ($\tau, \sigma \sim \|A\|^{-1}$) to get

$$G\left(H_p^\beta(y, x_0), x, y\right) \le C_3\left(\frac{\|A\|}{p^2\beta}\|x - x_0\|_{l^2}^2 + \|A_I\|\beta\right).$$

Combine with (7) to get

$$G\left(H_p^\beta(y, x_0), x, y\right) \le \frac{C_4}{p^2\beta}\left[c(x, y) + G(x_0, x, y)\right]^2 + C_5\|A_I\|\beta. \tag{8}$$

Step 5: Perform a reweight and restart

Idea: Balance the two terms in (8) so that every p iterations we have errors decreasing by a constant factor (up to δ). Optimal parameters give

$$\epsilon_0 \approx b_2, \quad \epsilon_n = e^{-1}(\delta + \epsilon_{n-1}), \quad \beta_n = \frac{\epsilon_n}{2\|A\|}.$$

$$\phi_n(y, x_0) = H_p^{\beta_n}(y, \phi_{n-1}(y, x_0))$$

$$\Rightarrow G\left(\phi_n(y, x_0), x, y\right) \leq \epsilon_n \lesssim \delta + e^{-n}$$

Combining this with (6), we obtain (for $x_0 = 0$)

$$\|\phi_n(y) - x\|_{l^2} \lesssim \underbrace{\sigma_{\mathbf{s},\mathbf{M}}(x)_{l^1_w}}_{\text{distance to sparse in levels vectors}} + \underbrace{\|Ax - y\|_{l^2}}_{\text{noise of measurements}} + \delta + \underbrace{e^{-n}}_{\text{"convergence" error}}.$$

□

Algorithm 1: `FIRENETcomp` constructs a FIRENET which corresponds to n iterations of `InnerIt` with a rescaling scheme. We write the output as the map ϕ_n to emphasise that `FIRENETcomp` defines a NN. `InnerIt` performs p iterations of Chambolle and Pock's primal-dual algorithm for square-root LASSO (the order of updates is swapped compared to [37]). The functions φ_s and ψ^1 are proximal maps:

$$[\varphi_s(x)]_j = \max\left\{0, 1 - \frac{s}{|x_j|}\right\} x_j, \quad \psi^1(y) = \min\left\{1, \frac{1}{\|y\|_{l^2}}\right\} y.$$

Both of these are approximated by NNs in our proof.

Function `FIRENETcomp` $(A, p, \tau, \sigma, \lambda, \{w_j\}_{j=1}^N, \epsilon_0, \delta, n)$

  Initiate with $\phi_0 \equiv 0$ (other initial vectors can also be chosen).
  (NB: ϵ_0 should be of the same order as $\|y\|_{l^2}$ for inputs $y \in \mathbb{C}^m$.)
  for $k = 1, ..., n$ **do**
    $\epsilon_k = e^{-1}(\delta + \epsilon_{k-1})$,
    $\beta_k = \frac{\epsilon_k}{2\|A\|}$
    $\phi_k(\cdot) = p\beta_k \cdot \texttt{InnerIt}\left(\frac{\cdot}{p\beta_k}, \frac{\phi_{k-1}(\cdot)}{p\beta_k}, A, p, \sigma, \tau, \lambda, \{w_j\}_{j=1}^N\right)$
  end
  return: FIRENET $\phi_n : \mathbb{C}^m \to \mathbb{C}^N$
end

Function `InnerIt` $(y, x_0, A, p, \tau, \sigma, \lambda, \{w_j\}_{j=1}^N)$

  Set $B = \mathrm{diag}(w_1, ..., w_N) \in \mathbb{C}^{N \times N}$.
  Initiate with $x^0 = x_0$, $y^0 = 0 \in \mathbb{C}^m$ (the superscripts denote indices not powers).
  for $k = 0, ..., p-1$ **do**
    $x^{k+1} = B\varphi_{\tau\lambda}(B^{-1}(x^k - \tau A^* y^k))$
    $y^{k+1} = \psi^1(y^k + \sigma A(2x^{k+1} - x^k) - \sigma y)$
  end
  $X = \sum_{k=1}^p \frac{x^k}{p}$
  return: $X \in \mathbb{C}^N$ (ergodic average of p iterates)
end

Applications in compressive imaging.

Demonstration of convergence

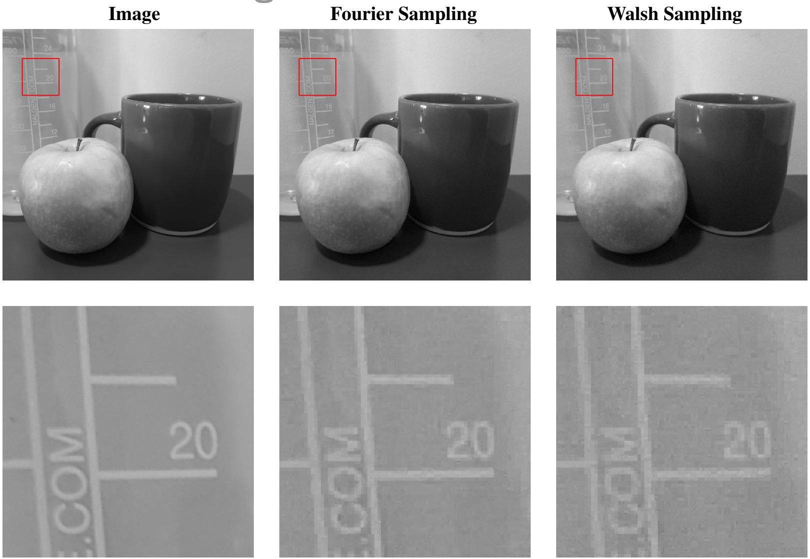
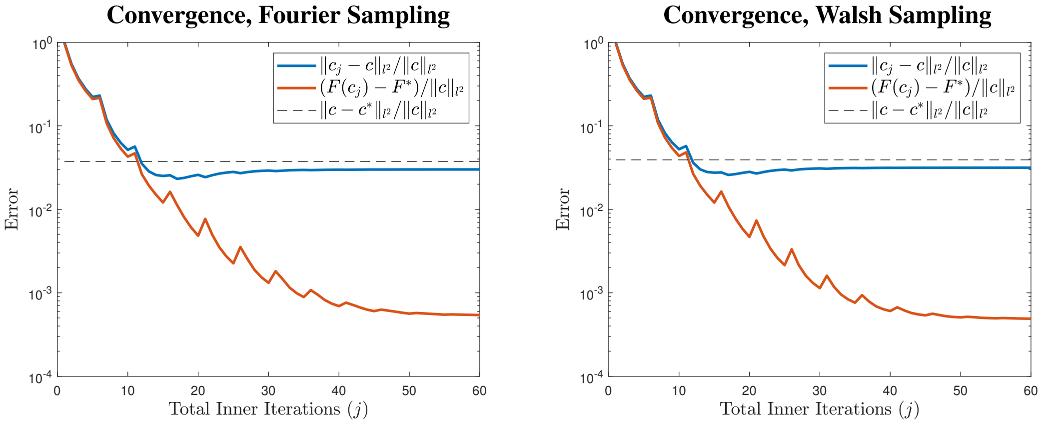


Figure: Images corrupted with 2% Gaussian noise and reconstructed using only 15% sampling with $n = p = 5$.

Demonstration of convergence



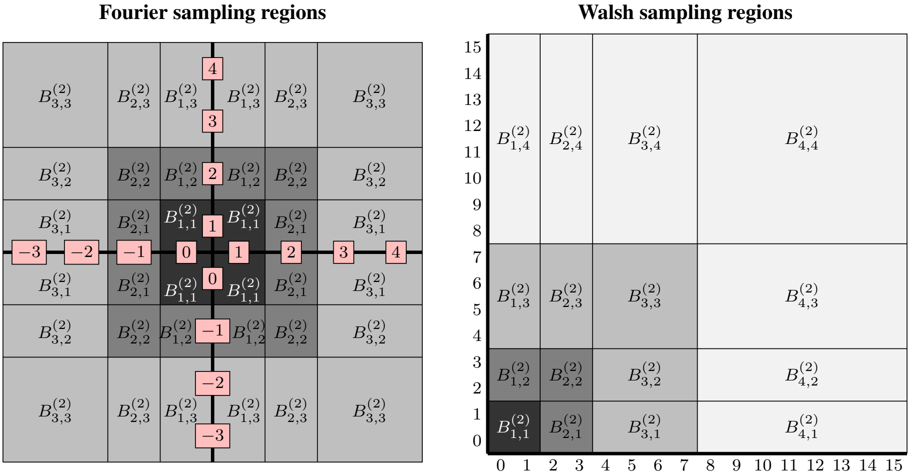
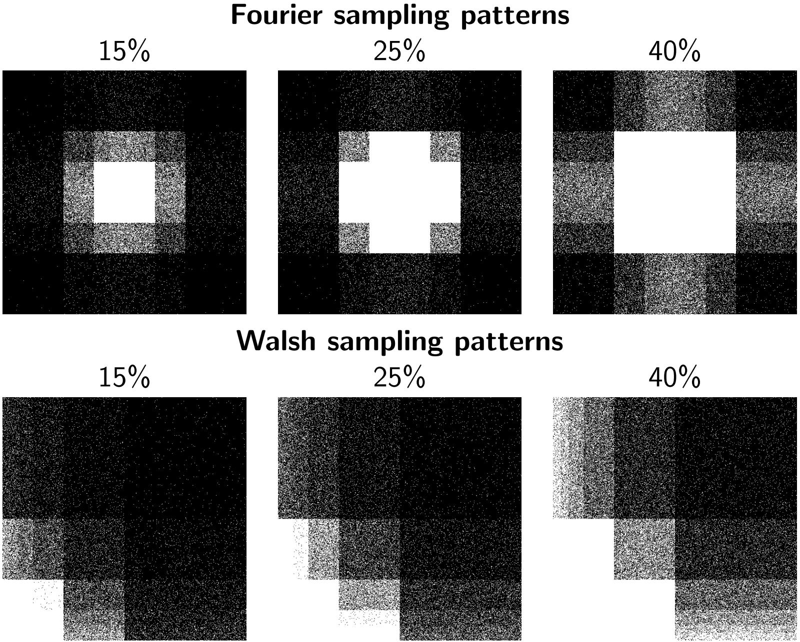


Figure: The different sampling regions used for the sampling patterns for Fourier (left, $r = 3$) and Walsh (right, $r = 4$). The axis labels correspond to the frequencies in each band and the annular regions are shown as the shaded greyscale regions.


Fourier sampling patterns
15%
25%
40%
Walsh sampling patterns
15%
25%
40%

The main result of this section

Theorem

Let $\epsilon_{\mathbb{P}} \in (0,1)$ *and* $\mathcal{L} = \log^3(N) \cdot \log(m) \cdot \log^2(s \cdot \log(N)) + \log(\epsilon_{\mathbb{P}}^{-1})$. *Suppose:*

- *(a) In the Fourier case:* $m_{\mathbf{k}} \gtrsim \mathcal{M}_{\mathcal{F}}(\mathbf{s}, \mathbf{k}) \cdot \mathcal{L}$.
- *(b) In the Walsh case:* $m_{\mathbf{k}} \gtrsim \mathcal{M}_{\mathcal{W}}(\mathbf{s}, \mathbf{k}) \cdot \mathcal{L}$.

For $\delta \in (0,1)$, *let* $\mathcal{J}(\delta, \mathrm{s}, \mathrm{M}, w)$ *be collection of all* $y \in \mathbb{C}^m$ *with* $y = Ac + e$ *where*

$$\|c\|_{l^2} \leq 1, \quad \max\left\{\frac{\sigma_{\mathbf{s},\mathbf{M}}(\Psi c)_{l^1_w}}{\sqrt{\xi}}, \|e\|_{l^2}\right\} \leq \delta.$$

We provide an algorithm that computes a neural network ϕ *with* $\mathcal{O}(\log(\delta^{-1}))$ *layers s.t. with probability at least* $1 - \epsilon_{\mathbb{P}}$,

$$\|\phi(y) - c\|_{l^2} \lesssim \delta, \quad \forall y = Ac + e \in \mathcal{J}(\delta, \mathrm{s}, \mathrm{M}, w).$$

$$
\begin{aligned}
\mathcal{M}_{\mathcal{F}}(\mathbf{s},\mathbf{k}) &:= \sum_{j=1}^{\|\mathbf{k}\|_{l^\infty}} s_j \prod_{i=1}^{d} 2^{-|k_i-j|} + \sum_{j=\|\mathbf{k}\|_{l^\infty}+1}^{r} s_j 2^{-2(j-\|\mathbf{k}\|_{l^\infty})} \prod_{i=1}^{d} 2^{-|k_i-j|} \\
\mathcal{M}_{\mathcal{W}}(\mathbf{s},\mathbf{k}) &:= s_{\|\mathbf{k}\|_{l^\infty}} \prod_{i=1}^{d} 2^{-|k_i-\|\mathbf{k}\|_{l^\infty}|}.
\end{aligned}
$$

Remarks

- Up to log-factors, measurement condition equivalent to the currently best-known oracle estimator (where one assumes apriori knowledge of the support of the vector).
- Consider number of samples per annular region

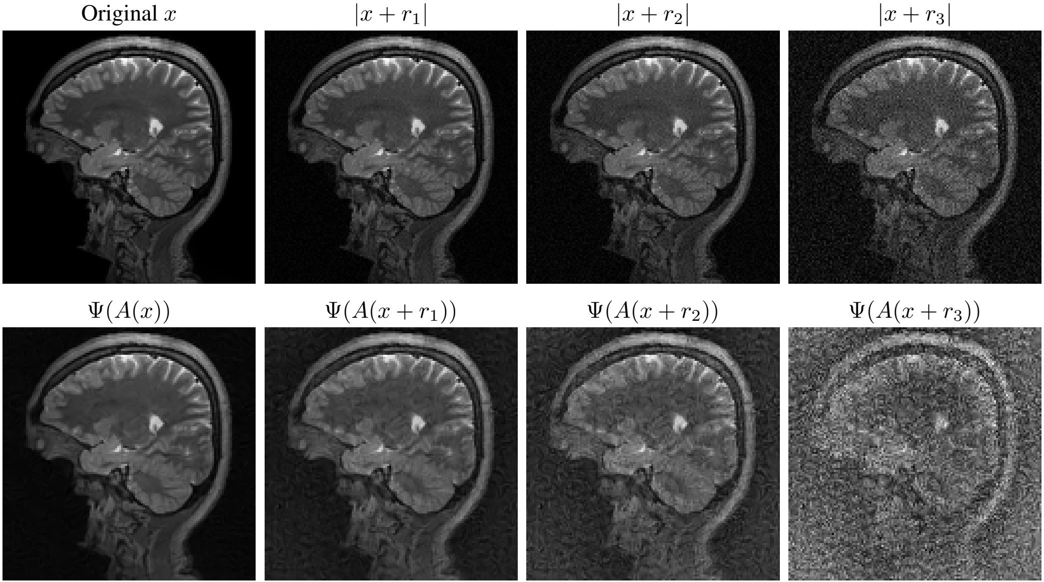
$$m_k = \sum_{\|\mathbf{k}\|_{l^\infty}=k} m_{\mathbf{k}}, \quad k = 1, \dots, r,$$

then up to logarithmic factors and exponentially small terms, s_k measurements are needed in each region.

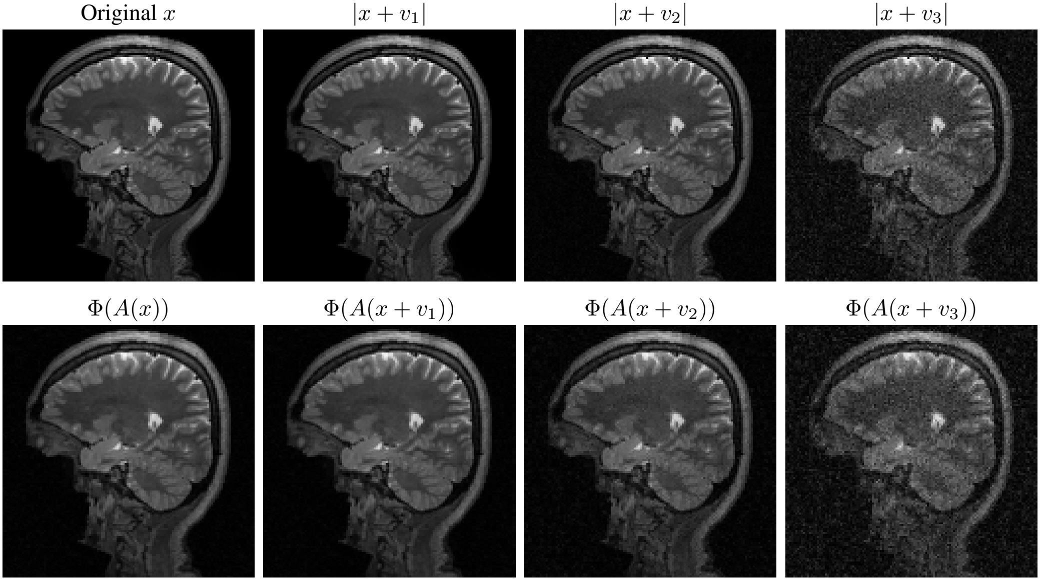
Take home message: Using the above machinery, we get optimal recovery in terms of the number of samples needed and only need $\mathcal{O}(\log(\delta^{-1}))$ many layers!!

Numerical experiments.

Stable? AUTOMAP ✗



Stable? FIRENETs ✓



Adding FIRENET layers stabilises AUTOMAP

$|x + r_3|$ $\qquad$ $\Psi(\tilde{y}),\ \tilde{y} = A(x + r_3)$ $\qquad$ $\Phi\left(\tilde{y}, \Psi(\tilde{y})\right)$

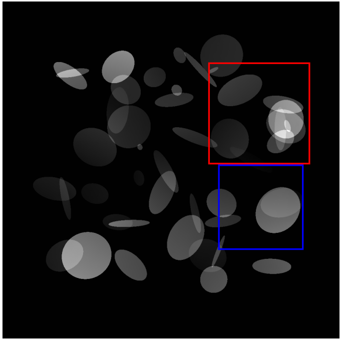
FIRENET withstand worst-case perturbations and generalises well

$|x_1 + v_1|$ $|x_2 + v_2|$ $|x_3 + v_3|$

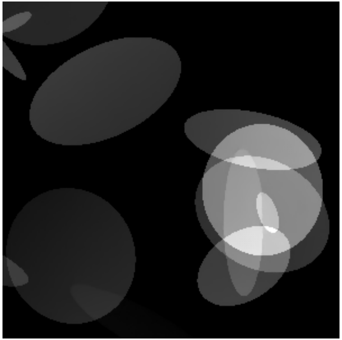
$\Phi(A(x_1 + v_1))$ $\Phi(A(x_2 + v_2))$ $\Phi(A(x_3 + v_3))$

Stability and accuracy, and false negative

Original x (full size)



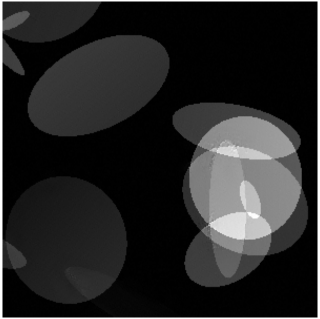
Original (cropped, red frame)



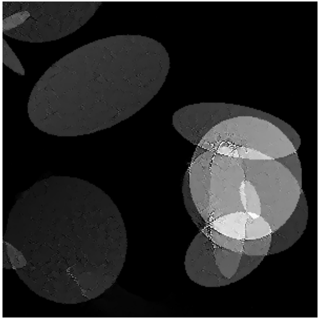
Original + detail $(x + h_1)$ (cropped, blue frame)

U-net trained without noise

Orig. + worst-case noise



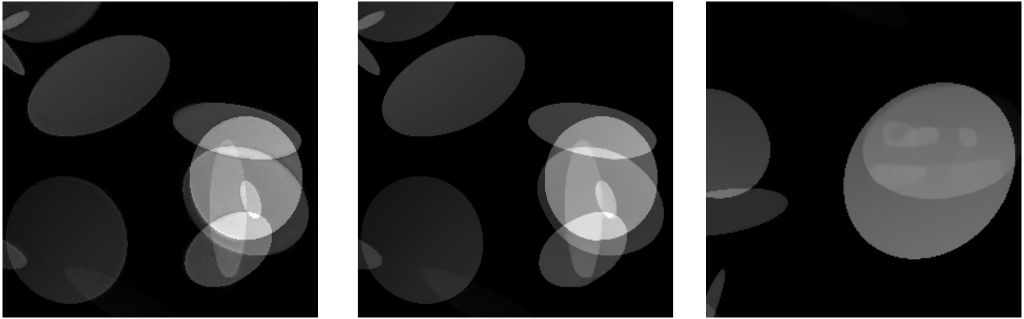
Rec. from worst-case noise



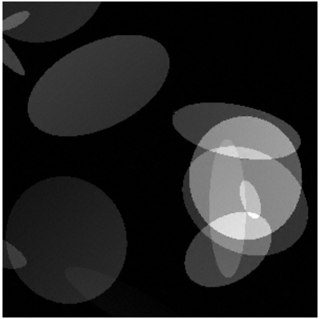
Rec. of detail

U-net trained with noise

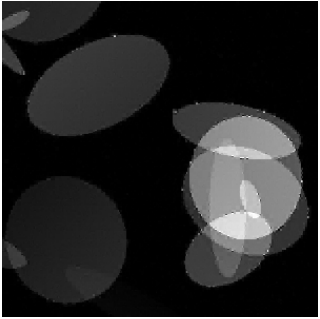
Orig. + worst-case noise | Rec. from worst-case noise | Rec. of detail



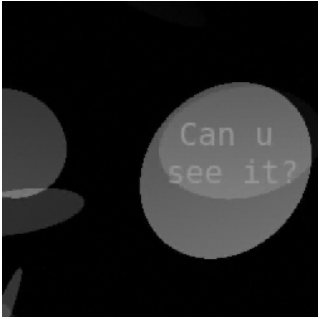
Orig. + worst-case noise



Rec. from worst-case noise



Rec. of detail



Final question: How do we optimally traverse the stability & accuracy trade-off?

FIRENETs provide a balance but are likely not the end of the story.

Answering this question will require a foundations framework for AI.

Hopefully we've inspired you to build on these results and take up the challenge!

Extra slides.

Multilevel random subsampling

Definition [Multilevel random subsampling]: Let $\mathrm{N} = (N_1, \dots, N_l) \in \mathbb{N}^l$, where $1 \leq N_1 < \cdots < N_l = N$ and $\mathrm{m} = (m_1, \dots, m_l) \in \mathbb{N}^l$ with $m_k \leq N_k - N_{k-1}$ for $k = 1, \dots, l$, and $N_0 = 0$. For each $k = 1, \dots, l$, let $\mathcal{I}_k = \{N_{k-1} + 1, \dots, N_k\}$ if $m_k = N_k - N_{k-1}$ and if not, let $t_{k,1}, \dots, t_{k,m_k}$ be chosen uniformly and independently from the set $\{N_{k-1} + 1, \dots, N_k\}$ (with possible repeats), and set $\mathcal{I}_k = \{t_{k,1}, \dots, t_{k,m_k}\}$. If $\mathcal{I} = \mathcal{I}_{\mathrm{N},\mathrm{m}} = \mathcal{I}_1 \cup \cdots \cup \mathcal{I}_l$ we refer to $\mathcal{I}$ as an (N, m)-multilevel subsampling scheme.

Definition [Multilevel subsampled unitary matrix]: A matrix $A \in \mathbb{C}^{m \times N}$ is an (N, m)-multilevel subsampled unitary matrix if $A = P_{\mathcal{I}} D U$ for a unitary matrix $U \in \mathbb{C}^{N \times N}$ and (N, m)-multilevel subsampling scheme $\mathcal{I}$. D is a diagonal scaling matrix:

$$D_{ii} = \sqrt{\frac{N_k - N_{k-1}}{m_k}}, \quad i = N_{k-1} + 1, ..., N_k, \quad k = 1, ..., l$$

and $P_{\mathcal{I}}$ is the projection onto the linear span of the subset of the canonical basis indexed by $\mathcal{I}$.

The orthonormal bases

$K = 2^r$ for $r \in \mathbb{N}$, and consider d-dimensional tensors in $\mathbb{C}^{K\times\cdots\times K} = \mathbb{C}^{K^d}$, $N = K^d$.

$V \in \mathbb{C}^{N\times N}$: corresponds to d-dimensional discrete Fourier or Walsh transform.

Fourier case: divide frequencies $\{-K/2+1,\ldots,K/2\}^d$ into dyadic bands. For $d=1$, $B_1 = \{0,1\}$ and $B_k = \{-2^{k-1}+1,\ldots,-2^{k-2}\} \cup \{2^{k-2}+1,\ldots,2^{k-1}\}$ for $k = 2,\ldots,r$.

Walsh case: $B_1 = \{0,1\}$ and $B_k = \{2^{k-1},\ldots,2^k-1\}$ for $k=2,\ldots,r$.

d-**dimensional case:** $B_{\mathbf{k}}^{(d)} = B_{k_1} \times \ldots \times B_{k_d}, \quad \mathbf{k} = (k_1,\ldots,k_d) \in \mathbb{N}^d$.

Observe: subsampled measurements of $V(c)$.

Sparse rep: Haar wavelet coefficients Ψc, $U = V\Psi^*$.

Sampling: Given $\{m_{\mathbf{k}=(k_1,\ldots,k_d)}\}_{k_1,\ldots,k_d=1}^r$, use a multilevel random sampling such that $m_{\mathbf{k}}$ measurements are chosen from $B_{\mathbf{k}}^{(d)}$.

Reduction to previous theorem

$U = \left[U^{(\mathbf{k},j)}\right]_{\mathbf{k}=1,j=1}^{\|\mathbf{k}\|_{l^\infty}\leq r,r}$ be defined as above. Then the $(\mathbf{k},j)$th local coherence of U is

$$\mu(U^{\mathbf{k},j}) = \left|B_{\mathbf{k}}^{(d)}\right| \max_{p,q} \ |(U^{\mathbf{k},j})_{pq}|^2, \quad \text{where } \left|B_{\mathbf{k}}^{(d)}\right| \text{ is the cardinality of } B_{\mathbf{k}}^{(d)}.$$

Proposition: Let $\epsilon_{\mathbb{P}} \in (0,1)$, $(\mathbf{s},\mathbf{M})$ be local sparsities and sparsity levels with $2 \leq s \leq N$, and consider the (N,m)-multilevel subsampling scheme to form A. Let

$$t_j = \min\left\{\left\lceil \frac{\xi(\mathbf{s},\mathbf{M},w)}{w_{(j)}^2} \right\rceil, M_j - M_{j-1}\right\}, \quad j = 1, ..., r,$$

and suppose that

$$m_k \gtrsim \mathcal{L}' \cdot \sum_{j=1}^{r} t_j \mu(U^{k,j}), \quad k = 1, ..., l$$

where $\mathcal{L}' = r \cdot \log(2m) \cdot \log^2(t) \cdot \log(N) + \log(\epsilon_{\mathbb{P}}^{-1})$. Then with probability at least $1-\epsilon_{\mathbb{P}}$, A satisfies the weighted rNSPL of order $(\mathbf{s},\mathbf{M})$ with constants $\rho = 1/2, \gamma = \sqrt{2}$.

Coherence bound for Fourier case

Lemma: Consider the d-dimensional Fourier–Haar–wavelet matrix with blocks $U^{\mathbf{k},j}$, then the local coherences satisfy

$$\mu(U^{\mathbf{k},j}) \lesssim 2^{-2(j-\|\mathbf{k}\|_{l^\infty})_+} \prod_{i=1}^{d} 2^{-|k_i-j|},$$

where for $t \in \mathbb{R}$, $t_+ = \max\{0, t\}$. It follows that

$$\sum_{j=1}^{r} s_j \mu(U^{\mathbf{k},j}) \lesssim \sum_{j=1}^{\|\mathbf{k}\|_{l^\infty}} s_j \prod_{i=1}^{d} 2^{-|k_i-j|} + \sum_{j=\|\mathbf{k}\|_{l^\infty}+1}^{r} s_j 2^{-2(j-\|\mathbf{k}\|_{l^\infty})} \prod_{i=1}^{d} 2^{-|k_i-j|} = \mathcal{M}_{\mathcal{F}}(\mathbf{s}, \mathbf{k}).$$

Proof.

Exercise in using the discrete Fourier transform and some trigonometric identities. □

Coherence bound for Walsh case

Lemma: Consider the d-dimensional Walsh–Haar–wavelet matrix with blocks $U^{(\mathbf{k},j)}$, then the local coherences satisfy

$$\mu(U^{(\mathbf{k},j)}) \lesssim \begin{cases} \prod_{i=1}^{d} 2^{-|k_i - j|} & \text{if } k_i \leq j \text{ for } i = 1, \ldots, d \text{ with at least one equality,} \\ 0 & \text{otherwise} \end{cases}.$$

It follows that

$$\sum_{j=1}^{r} s_j \mu(U^{(\mathbf{k},j)}) \lesssim s_{\|\mathbf{k}\|_{l^\infty}} \prod_{i=1}^{d} 2^{-|k_i - \|\mathbf{k}\|_{l^\infty}|} = \mathcal{M}_{\mathcal{W}}(\mathbf{s}, \mathbf{k}).$$

Proof.

Exercise in keeping track of supports of Haar wavelet basis. □