On the Barriers of Deep Learning, Approximate Sharpness, and Smale's 18th Problem

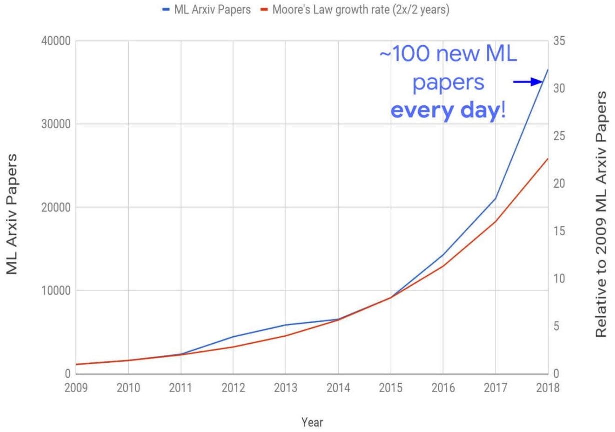
Matthew Colbrook
(University of Cambridge and École Normale Supérieure)

M. Colbrook, V. Antun, A. Hansen, "*The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smale's 18th problem*" (PNAS, to appear)
www.github.com/Comp-Foundations-and-Barriers-of-AI/firenet

M. Colbrook, "*WARPd: A linearly convergent first-order method for inverse problems with approximate sharpness conditions*" (SIIMS, under revision)
www.github.com/MColbrook/WARPd

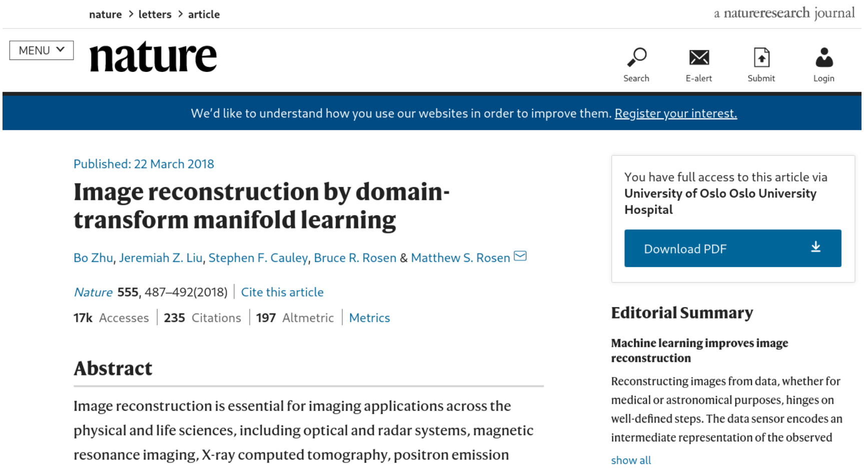
Interest in deep learning unprecedented and exponentially growing

Machine learning papers on arXiv



To keep up during first lockdown, would need to continually read a paper every 4 mins!

Will AI replace standard algorithms in medical imaging?



nature > letters > article

a natureresearch journal

MENU

nature

Search E-alert Submit Login

We'd like to understand how you use our websites in order to improve them. Register your interest.

Published: 22 March 2018

Image reconstruction by domain-transform manifold learning

Bo Zhu, Jeremiah Z. Liu, Stephen F. Cauley, Bruce R. Rosen & Matthew S. Rosen

Nature **555**, 487–492(2018) | Cite this article

17k Accesses | **235** Citations | **197** Altmetric | Metrics

Abstract

Image reconstruction is essential for imaging applications across the physical and life sciences, including optical and radar systems, magnetic resonance imaging, X-ray computed tomography, positron emission

You have full access to this article via **University of Oslo Oslo University Hospital**

Download PDF

Editorial Summary

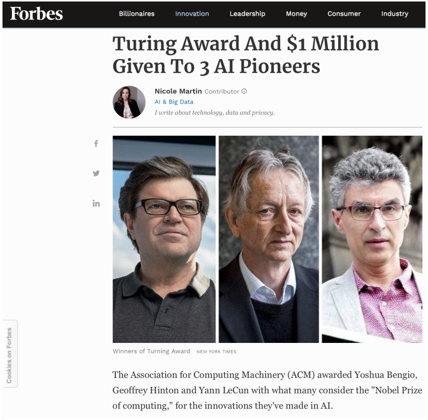
Machine learning improves image reconstruction

Reconstructing images from data, whether for medical or astronomical purposes, hinges on well-defined steps. The data sensor encodes an intermediate representation of the observed

show all

Claim: "superior immunity to noise and a reduction in reconstruction artefacts compared with conventional handcrafted reconstruction methods".

Very strong confidence in deep learning



Forbes Billionaires Innovation Leadership Money Consumer Industry

Turing Award And $1 Million Given To 3 AI Pioneers

Nicole Martin Contributor
AI & Big Data
I write about technology, data and privacy.

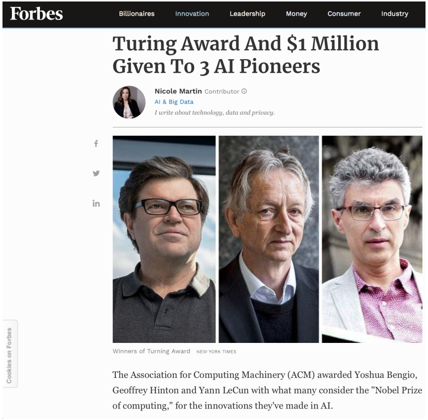
Winners of Turning Award NEW YORK TIMES

The Association for Computing Machinery (ACM) awarded Yoshua Bengio, Geoffrey Hinton and Yann LeCun with what many consider the "Nobel Prize of computing," for the innovations they've made in AI.

Cookies on Forbes

Geoffrey Hinton, The New Yorker, April 2017: "They should stop training radiologists now!"

Very strong confidence in deep learning


Forbes
Billionaires Innovation Leadership Money Consumer Industry

Turing Award And $1 Million Given To 3 AI Pioneers

Nicole Martin Contributor
AI & Big Data
I write about technology, data and privacy.

Winners of Turning Award NEW YORK TIMES

The Association for Computing Machinery (ACM) awarded Yoshua Bengio, Geoffrey Hinton and Yann LeCun with what many consider the "Nobel Prize of computing," for the innovations they've made in AI.

Cookies on Forbes

Geoffrey Hinton, The New Yorker, April 2017: “They should stop training radiologists now!”

BUT . . .

DANGER: AI generated hallucinations
Facebook and NYU's 2020 FastMRI challenge

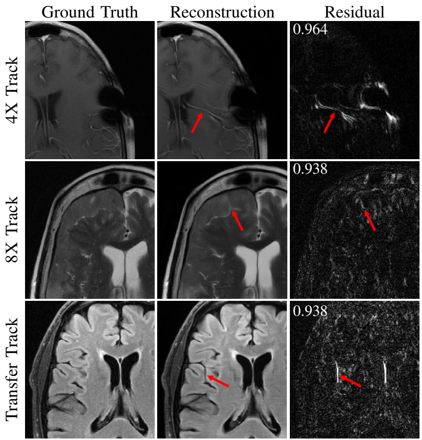


Fig. 6. Examples of reconstruction hallucinations among challenge submissions with SSIM scores over residual plots (residuals magnified by 5). (*top*) A 4X submission from Neurospin generated a false vessel, possibly related to susceptibilities introduced by surgical staples. (*middle*) An 8X submission from ATB introduced a linear bright signal mimicking a cleft of cerebrospinal fluid, as well as blurring of the boundaries of the extra-axial mass. (*bottom*) A submission from ResoNNance introduced a false sulcus or prominent vessel.

DL seems unstable in inverse problems!

Submit About Contact Journal Club Subscribe

Log in

PNAS Proceedings of the National Academy of Sciences of the United States of America

Keyword, Author, or DOI

Advanced Search

Home Articles Front Matter News Podcasts Authors

NEW RESEARCH IN Physical Sciences Social Sciences Biological Sciences

PHYSICAL SCIENCES

On instabilities of deep learning in image reconstruction and the potential costs of AI

Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C. Hansen

PNAS first published May 11, 2020 https://doi.org/10.1073/pnas.1907377117

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved March 12, 2020 (received for review June 4, 2019)

Article Figures & SI Info & Metrics PDF

Article Alerts Share
Email Article Tweet
Citation Tools Like 52
© Request Permissions Mendeley

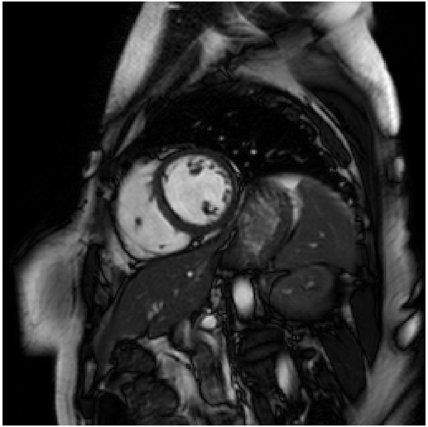
Submit

Sign up for Article Alerts

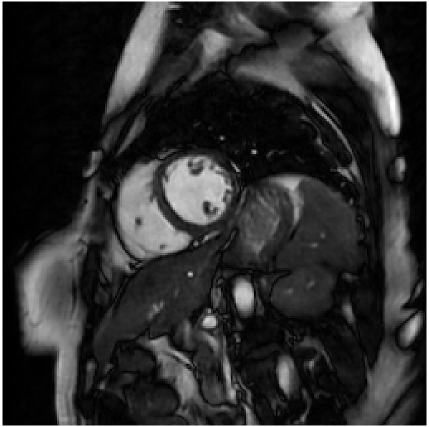
Enter Email Address Sign up

Example

$|x|$



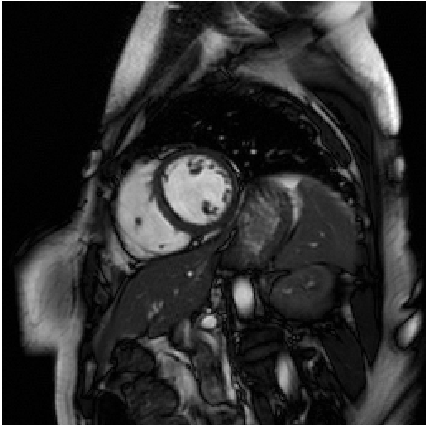
$|\Psi(Ax)|$



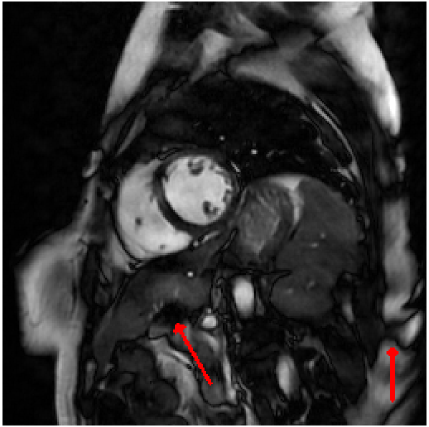
Network (33% subsampling) from: J. Schlemper, J. Caballero, J. V. Hajnal, A. Price and D. Rueckert, '*A deep cascade of convolutional neural networks for MR image reconstruction*', in International conference on information processing in medical imaging, Springer, 2017, pp. 647–658.
Figures from: Antun, V., Renna, F., Poon, C., Adcock, B., & Hansen, A. C., '*On instabilities of deep learning in image reconstruction and the potential costs of AI*'. Proc. Natl. Acad. Sci. USA, 2020..

Example

$|x + r_1|$



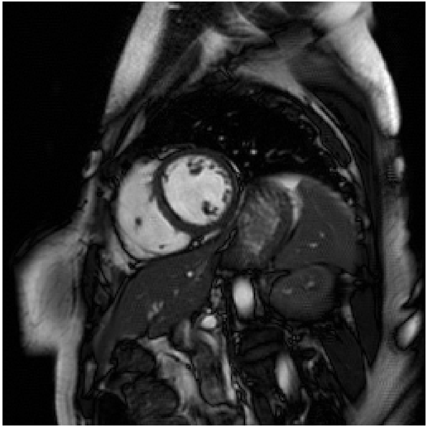
$|\Psi(A(x + r_1))|$



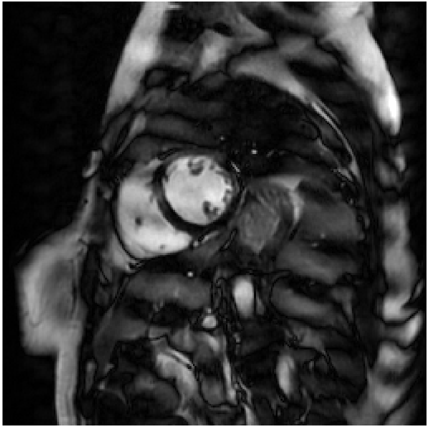
Network (33% subsampling) from: J. Schlemper, J. Caballero, J. V. Hajnal, A. Price and D. Rueckert, '*A deep cascade of convolutional neural networks for MR image reconstruction*', in International conference on information processing in medical imaging, Springer, 2017, pp. 647–658.
Figures from: Antun, V., Renna, F., Poon, C., Adcock, B., & Hansen, A. C., '*On instabilities of deep learning in image reconstruction and the potential costs of AI*'. Proc. Natl. Acad. Sci. USA, 2020..

Example

$|x + r_2|$



$|\Psi(A(x + r_2))|$

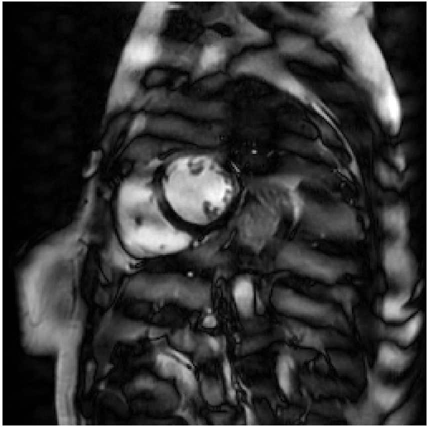


Network (33% subsampling) from: J. Schlemper, J. Caballero, J. V. Hajnal, A. Price and D. Rueckert, '*A deep cascade of convolutional neural networks for MR image reconstruction*', in International conference on information processing in medical imaging, Springer, 2017, pp. 647–658.
Figures from: Antun, V., Renna, F., Poon, C., Adcock, B., & Hansen, A. C., '*On instabilities of deep learning in image reconstruction and the potential costs of AI*'. Proc. Natl. Acad. Sci. USA, 2020..

Example

$|x + r_3|$

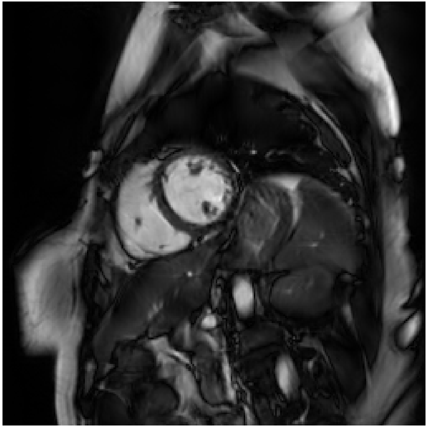
$|\Psi(A(x + r_3))|$



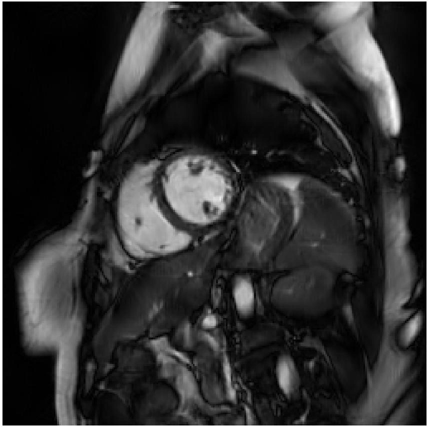
Network (33% subsampling) from: J. Schlemper, J. Caballero, J. V. Hajnal, A. Price and D. Rueckert, '*A deep cascade of convolutional neural networks for MR image reconstruction*', in International conference on information processing in medical imaging, Springer, 2017, pp. 647–658.
Figures from: Antun, V., Renna, F., Poon, C., Adcock, B., & Hansen, A. C., '*On instabilities of deep learning in image reconstruction and the potential costs of AI*'. Proc. Natl. Acad. Sci. USA, 2020..

Reconstruction using state-of-the-art standard methods

SoA from Ax



SoA from $A(x + r_3)$



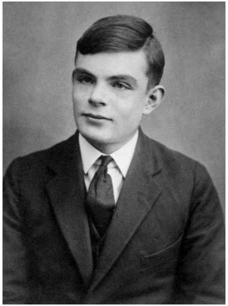
Optimism: Echoes of an old story

Hilbert's vision (start of 20th century): secure foundations for all mathematics.

- Mathematics should be written in a precise language, manipulated according to well defined rules.
- Completeness: a proof that all true mathematical statements can be proved in the formalism.
- Consistency: a proof that no contradiction can be obtained in the formalism of mathematics.
- Decidability: an algorithm for deciding the truth or falsity of any mathematical statement.

Hilbert's 10th problem: *Provide an algorithm which, for any given polynomial equation with integer coefficients, can decide whether there is an integer-valued solution.*

Foundations ⇒ better understanding, discover feasible directions for techniques, discover new methods, ...



Gödel (pioneer of **modern logic**) and Turing (pioneer of **modern computer science**) turned Hilbert's optimism upside down:

- There exist true statements in mathematics that cannot be proven!
- There exists problems that cannot be computed by an algorithm!

Hilbert's 10th problem: No such algorithm exists (1970, Matiyasevich).

A program for the foundations of DL and AI

Smale's 18th problem*: *What are the limits of artificial intelligence?*

A program determining the foundations/limitations of deep learning and AI is needed:

- Boundaries of methodologies.
- Universal/intrinsic boundaries (e.g., no algorithm can do it).

There is a key difference between existence and construction here.

Need to also incorporate two pillars of scientific computation:

- Stability
- Accuracy

A GOAL of this talk: Develop some results in this direction for inverse problems.

*Steve Smale's list of problems for the 21st century (requested by Vladimir Arnold), inspired by Hilbert's list.

Mathematical setup

Given measurements $\quad y = Ax + e \quad$ recover $\quad x \in \mathbb{C}^N$.

- $x \in \mathbb{C}^N$ be an unknown vector,
- $A \in \mathbb{C}^{m \times N}$ be a matrix $(m < N)$ describing modality (e.g., MRI), and
- $y = Ax + e$ the noisy measurements of x.

Outline:

- Fundamental barriers.
- Sufficient conditions and Fast Iterative REstarted NETworks (FIRENETs).
- Some numerical examples (e.g., stability and accuracy).
- Approximate sharpness conditions and Weighted, Accelerated and Restarted Primal-dual (WARPd).

Can we train neural networks that solve (P_j)?

Sparse regularization (benchmark problem):

$$\min_{x\in\mathbb{C}^N} \|x\|_{\ell^1} \quad \text{subject to} \quad \|Ax-y\|_{\ell^2} \leq \eta \tag{P_1}$$

$$\min_{x\in\mathbb{C}^N} \lambda\|x\|_{\ell^1} + \|Ax-y\|_{\ell^2}^2 \tag{P_2}$$

$$\min_{x\in\mathbb{C}^N} \lambda\|x\|_{\ell^1} + \|Ax-y\|_{\ell^2} \tag{P_3}$$

Denote the **minimizing** vectors by Ξ.

- Avoid bizarre, unnatural & pathological mappings: (P_j) well-understood & well-used!
- Simpler solution map than inverse problem $\Rightarrow$ stronger impossibility results.
- DL has also been used to speed up sparse regularization and tackle (P_j).

The set-up

$$A \in \mathbb{C}^{m \times N} \text{ (modality)}, \quad \mathcal{S} = \{y_k\}_{k=1}^{R} \subset \mathbb{C}^m \text{ (samples)}, \quad R < \infty$$

In practice, the matrix A is not known exactly or cannot be stored to infinite precision.

Assume access to: $\{y_{k,n}\}_{k=1}^{R}$ and A_n (rational approximations, e.g., floats) such that

$$\|y_{k,n} - y_k\| \leq 2^{-n}, \quad \|A_n - A\| \leq 2^{-n}, \quad \forall n \in \mathbb{N}.$$

Training set associated with $(A, \mathcal{S}) \in \Omega$ is

$$\iota_{A,\mathcal{S}} := \{(y_{k,n}, A_n) \mid k = 1, \ldots, R, \text{ and } n \in \mathbb{N}\}.$$

In a nutshell: allow access to arbitrary precision training data.

The set-up

$$A \in \mathbb{C}^{m\times N} \text{ (modality)}, \quad \mathcal{S} = \{y_k\}_{k=1}^{R} \subset \mathbb{C}^m \text{ (samples)}, \quad R < \infty$$

In practice, the matrix A is not known exactly or cannot be stored to infinite precision.

Assume access to: $\{y_{k,n}\}_{k=1}^{R}$ and A_n (rational approximations, e.g., floats) such that

$$\|y_{k,n} - y_k\| \leq 2^{-n}, \quad \|A_n - A\| \leq 2^{-n}, \quad \forall n \in \mathbb{N}.$$

Training set associated with $(A, \mathcal{S}) \in \Omega$ is

$$\iota_{A,\mathcal{S}} := \{(y_{k,n}, A_n) \mid k = 1, \dots, R, \text{ and } n \in \mathbb{N}\}.$$

In a nutshell: allow access to arbitrary precision training data.

Question: Given a collection Ω of $(A, \mathcal{S})$, does there exist a neural network approximating Ξ (solution map of (P_j)), and can it be trained by an algorithm?

What could go wrong?

$$\min_{x \in \mathbb{C}^N} \|x\|_{\ell^1} \quad \text{subject to} \quad \|Ax - y\|_{\ell^2} \leq \eta \tag{P_1}$$

$$\min_{x \in \mathbb{C}^N} \lambda \|x\|_{\ell^1} + \|Ax - y\|_{\ell^2}^2 \tag{P_2}$$

$$\min_{x \in \mathbb{C}^N} \lambda \|x\|_{\ell^1} + \|Ax - y\|_{\ell^2} \tag{P_3}$$

(i) **Non-existence:** There does not exist a neural network that approximates the function we are interested in.

(ii)

(iii)

What could go wrong?

$$\min_{x \in \mathbb{C}^N} \|x\|_{\ell^1} \quad \text{subject to} \quad \|Ax - y\|_{\ell^2} \leq \eta \tag{P_1}$$

$$\min_{x \in \mathbb{C}^N} \lambda \|x\|_{\ell^1} + \|Ax - y\|_{\ell^2}^2 \tag{P_2}$$

$$\min_{x \in \mathbb{C}^N} \lambda \|x\|_{\ell^1} + \|Ax - y\|_{\ell^2} \tag{P_3}$$

(i) ~~**Non-existence:** There does not exist a neural network that approximates the function we are interested in.~~

(ii)

(iii)

What could go wrong?

$$\min_{x \in \mathbb{C}^N} \|x\|_{\ell^1} \quad \text{subject to} \quad \|Ax - y\|_{\ell^2} \leq \eta \tag{P_1}$$

$$\min_{x \in \mathbb{C}^N} \lambda \|x\|_{\ell^1} + \|Ax - y\|_{\ell^2}^2 \tag{P_2}$$

$$\min_{x \in \mathbb{C}^N} \lambda \|x\|_{\ell^1} + \|Ax - y\|_{\ell^2} \tag{P_3}$$

(i) ~~**Non-existence:** There does not exist a neural network that approximates the function we are interested in.~~

(ii) **Non-trainable:** There exists a neural network that approximates the function. However, there does not exist an algorithm that can train the neural network.

(iii)

What could go wrong?

$$\min_{x \in \mathbb{C}^N} \|x\|_{\ell^1} \quad \text{subject to} \quad \|Ax - y\|_{\ell^2} \leq \eta \tag{P_1}$$

$$\min_{x \in \mathbb{C}^N} \lambda \|x\|_{\ell^1} + \|Ax - y\|_{\ell^2}^2 \tag{P_2}$$

$$\min_{x \in \mathbb{C}^N} \lambda \|x\|_{\ell^1} + \|Ax - y\|_{\ell^2} \tag{P_3}$$

(i) ~~**Non-existence:** There does not exist a neural network that approximates the function we are interested in.~~

(ii) **Non-trainable:** There exists a neural network that approximates the function. However, there does not exist an algorithm that can train the neural network.

(iii) **Not practical:** There exists a neural network that approximates the function, and an algorithm training it. However, the algorithm needs prohibitively many samples.

Bad news - can't necessarily approximate such a neural network

Theorem

For (P_j), $N \geq 2$ and $m < N$. Let $K \geq 3$ be a positive integer, $L \in \mathbb{N}$. Then there exists a **well-conditioned** *class (condition numbers ≤ 1) Ω of elements $(A, \mathcal{S})$ s.t. (Ω* **fixed** *in what follows):*

Bad news - can't necessarily approximate such a neural network

Theorem

For (P_j), $N \geq 2$ *and* $m < N$. *Let* $K \geq 3$ *be a positive integer,* $L \in \mathbb{N}$. *Then there exists a* **well-conditioned** *class (condition numbers* ≤ 1*)* Ω *of elements* $(A, \mathcal{S})$ *s.t. (*Ω **fixed** *in what follows):*

(i) *There* **does not exist any algorithm** *that, given a training set* $\iota_{A,\mathcal{S}}$, *produces a neural network* $\phi_{A,\mathcal{S}}$ *with*

$$\min_{y \in \mathcal{S}} \inf_{x^* \in \Xi(A,y)} \|\phi_{A,\mathcal{S}}(y) - x^*\|_{\ell^2} \leq 10^{-K}, \quad \forall (A, \mathcal{S}) \in \Omega. \tag{1}$$

Furthermore, for any $p > 1/2$, **no probabilistic algorithm** *can produce a neural network* $\phi_{A,\mathcal{S}}$ *such that (1) holds with probability at least* p.

Bad news - can't necessarily approximate such a neural network

Theorem

For (P_j), $N \geq 2$ *and* $m < N$. *Let* $K \geq 3$ *be a positive integer,* $L \in \mathbb{N}$. *Then there exists a* **well-conditioned** *class (condition numbers* ≤ 1*)* Ω *of elements* $(A, \mathcal{S})$ *s.t. (*Ω **fixed** *in what follows):*

(i) *There* **does not exist any algorithm** *that, given a training set* $\iota_{A,\mathcal{S}}$, *produces a neural network* $\phi_{A,\mathcal{S}}$ *with*

$$\min_{y \in \mathcal{S}} \inf_{x^* \in \Xi(A,y)} \|\phi_{A,\mathcal{S}}(y) - x^*\|_{\ell^2} \leq 10^{-K}, \quad \forall (A, \mathcal{S}) \in \Omega. \tag{1}$$

Furthermore, for any $p > 1/2$, **no probabilistic algorithm** *can produce a neural network* $\phi_{A,\mathcal{S}}$ *such that (1) holds with probability at least* p.

(ii) *There* **exists an algorithm** *that produces a neural network* $\phi_{A,\mathcal{S}}$ *such that*

$$\max_{y \in \mathcal{S}} \inf_{x^* \in \Xi(A,y)} \|\phi_{A,\mathcal{S}}(y) - x^*\|_{\ell^2} \leq 10^{-(K-1)}, \quad \forall (A, \mathcal{S}) \in \Omega.$$

However, for any such algorithm (even probabilistic), $M \in \mathbb{N}$ *and* $p \in \left[0, 1 - \frac{1}{N+1-m}\right)$, *there exists a training set* $\iota_{A,\mathcal{S}}$ *such that for all* $y \in \mathcal{S}$,

$$\mathbb{P}\left(\inf_{x^* \in \Xi(A,y)} \|\phi_{A,\mathcal{S}}(y) - x^*\|_{\ell^2} > 10^{-(K-1)} \text{ or size of training data needed} > M \right) > p.$$

Bad news - can't necessarily approximate such a neural network

Theorem

For (P_j), $N \geq 2$ *and* $m < N$. *Let* $K \geq 3$ *be a positive integer,* $L \in \mathbb{N}$. *Then there exists a* **well-conditioned** *class (condition numbers* ≤ 1*)* Ω *of elements* $(A, \mathcal{S})$ *s.t. (*Ω **fixed** *in what follows):*

(i) *There* **does not exist any algorithm** *that, given a training set* $\iota_{A,\mathcal{S}}$*, produces a neural network* $\phi_{A,\mathcal{S}}$ *with*

$$\min_{y \in \mathcal{S}} \inf_{x^* \in \Xi(A,y)} \|\phi_{A,\mathcal{S}}(y) - x^*\|_{\ell^2} \leq 10^{-K}, \quad \forall (A, \mathcal{S}) \in \Omega. \tag{1}$$

Furthermore, for any $p > 1/2$, **no probabilistic algorithm** *can produce a neural network* $\phi_{A,\mathcal{S}}$ *such that (1) holds with probability at least* p.

(ii) *There* **exists an algorithm** *that produces a neural network* $\phi_{A,\mathcal{S}}$ *such that*

$$\max_{y \in \mathcal{S}} \inf_{x^* \in \Xi(A,y)} \|\phi_{A,\mathcal{S}}(y) - x^*\|_{\ell^2} \leq 10^{-(K-1)}, \quad \forall (A, \mathcal{S}) \in \Omega.$$

However, for any such algorithm (even probabilistic), $M \in \mathbb{N}$ *and* $p \in \left[0, 1 - \frac{1}{N+1-m}\right)$*, there exists a training set* $\iota_{A,\mathcal{S}}$ *such that for all* $y \in \mathcal{S}$*,*

$$\mathbb{P}\left(\inf_{x^* \in \Xi(A,y)} \|\phi_{A,\mathcal{S}}(y) - x^*\|_{\ell^2} > 10^{-(K-1)} \text{ or size of training data needed} > M\right) > p.$$

(iii) *There* **exists an algorithm** *using only* L *training data from each* $\iota_{A,\mathcal{S}}$ *that produces a neural network* $\phi_{A,\mathcal{S}}(y)$ *such that*

$$\max_{y \in \mathcal{S}} \inf_{x^* \in \Xi(A,y)} \|\phi_{A,\mathcal{S}}(y) - x^*\|_{\ell^2} \leq 10^{-(K-2)}, \quad \forall (A, \mathcal{S}) \in \Omega.$$

In words . . .

Nice classes Ω where stable and accurate neural networks exist. But:

- No algorithm, even randomized can train such a neural network accurate to K digits with probability greater than $1/2$.
- There exists a deterministic algorithm that trains a neural network with $K-1$ correct digits, but any such (even randomized) algorithm needs arbitrarily many training data.
- There exists a deterministic algorithm that trains a neural network with $K-2$ correct digits using no more than L training samples.

Result **independent of neural network architecture** - a universal barrier.

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

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

Numerical example: fails with training methods

$\mathrm{dist}(\Psi_{A_n}(y_n), \Xi_3(A,y))$	$\mathrm{dist}(\Phi_{A_n}(y_n), \Xi_3(A,y))$	$\|A_n - A\| \leq 2^{-n}$ $\|y_n - y\|_{\ell^2} \leq 2^{-n}$	10^{-K}
0.2999690	0.2597827	$n = 10$	10^{-1}
0.3000000	0.2598050	$n = 20$	10^{-1}
0.3000000	0.2598052	$n = 30$	10^{-1}
0.0030000	0.0025980	$n = 10$	10^{-3}
0.0030000	0.0025980	$n = 20$	10^{-3}
0.0030000	0.0025980	$n = 30$	10^{-3}
0.0000030	0.0000015	$n = 10$	10^{-6}
0.0000030	0.0000015	$n = 20$	10^{-6}
0.0000030	0.0000015	$n = 30$	10^{-6}

Table: (**Impossibility of computing the existing neural network to arbitrary accuracy**). Matrix $A \in \mathbb{C}^{19\times 20}$ constructed from discrete cosine transform, $R = 8000$, solutions are 6-sparse. LISTA (learned iterative shrinkage thresholding algorithm) Ψ_{A_n}, and FIRENETs Φ_{A_n}. The table shows the shortest ℓ^2 distance between the output from the networks and the true minimizer of the problem $\min_{x\in\mathbb{C}^N} \|x\|_{\ell^1} + \|Ax - y\|_{\ell^2}$, for different values of n and K.

Can we avoid this?

$$\hat{x} \in \operatorname{argmin} f(x), \quad f^* = \min f(x)$$

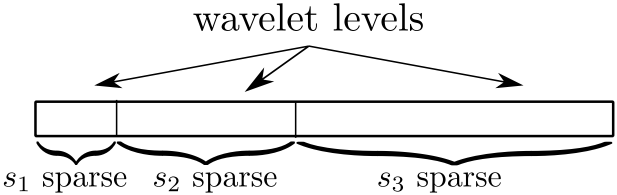
Problem: $f(x) \leq f^* + \epsilon$ does not in general imply x is close to set of minimizers.

Question: Can we find 'good' input classes where

$$f(x) \leq f^* + \epsilon \implies \inf_{\hat{x} \in \operatorname{argmin} f(x)} \|x - \hat{x}\| \lesssim \epsilon?$$

We shall see that the answer is yes!

State-of-the-art model for sparse regularisation



$\mathbf{M} = (M_1, \dots, M_r) \in \mathbb{N}^r$ and $\mathbf{s} = (s_1, \dots, s_r) \in \mathbb{Z}^r_{\geq 0}$. $x \in \mathbb{C}^N$ is $(\mathbf{s}, \mathbf{M})$-sparse in levels if

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

Denote set of $(\mathbf{s}, \mathbf{M})$-sparse vectors by $\Sigma_{\mathbf{s},\mathbf{M}}$, define

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

The robust nullspace property

Definition: $A \in \mathbb{C}^{m\times N}$ satisfies the **robust null space property in levels (rNSPL)** of order $(\mathbf{s},\mathbf{M})$ with constants $\rho \in (0,1)$ and $\gamma > 0$ if for any $(\mathbf{s},\mathbf{M})$ support set Δ,

$$\|x_\Delta\|_{\ell^2} \leq \frac{\rho\|x_{\Delta^c}\|_{\ell^1}}{\sqrt{r(s_1+\ldots+s_r)}} + \gamma\|Ax\|_{\ell^2}, \qquad \forall x \in \mathbb{C}^N.$$

Objective function: $f(x) = \lambda\|x\|_{\ell^1} + \|Ax-y\|_{\ell^2}$

$$\text{rNSPL} \Rightarrow \|z-x\|_{\ell^2} \lesssim \underbrace{\sigma_{\mathbf{s},\mathbf{M}}(x)_{\ell^1} + \|Ax-y\|_{\ell^2}}_{\text{“small”}} + \underbrace{\left(\lambda\|z\|_{\ell^1} + \|Az-y\|_{\ell^2} - \lambda\|x\|_{\ell^1} - \|Ax-y\|_{\ell^2}\right)}_{f(z)-f(x)\text{ objective function difference}},$$

In a nutshell: control $\|z-x\|_{\ell^2}$ by $f(z)-f(x)$, up to small approximation term.

Fast Iterative REstarted NETworks (FIRENETs)

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

Input: *Sparsity parameters* $(\mathbf{s}, \mathbf{M})$, $A \in \mathbb{C}^{m \times N}$ *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 width* $2(N+m)$ *such that:*

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

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

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

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

Demonstration of convergence

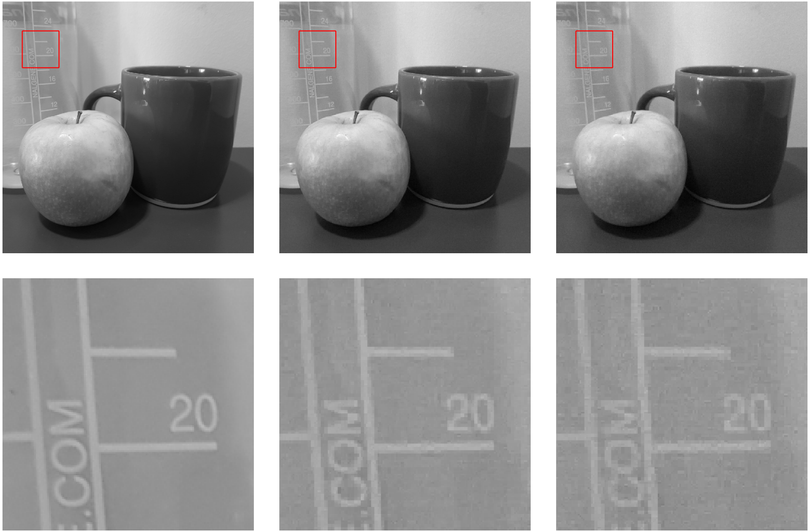
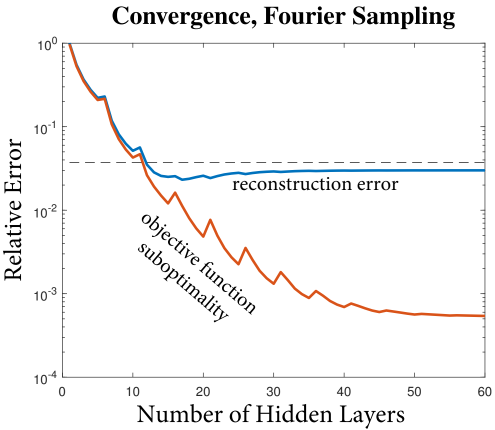
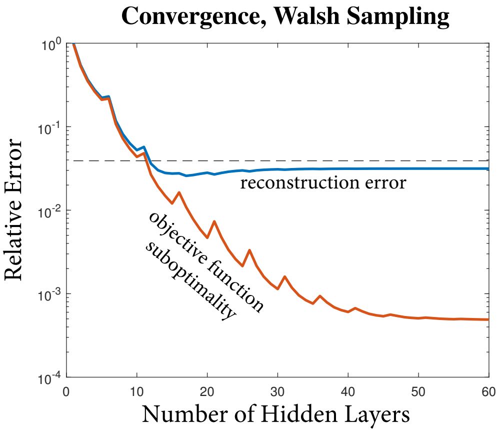


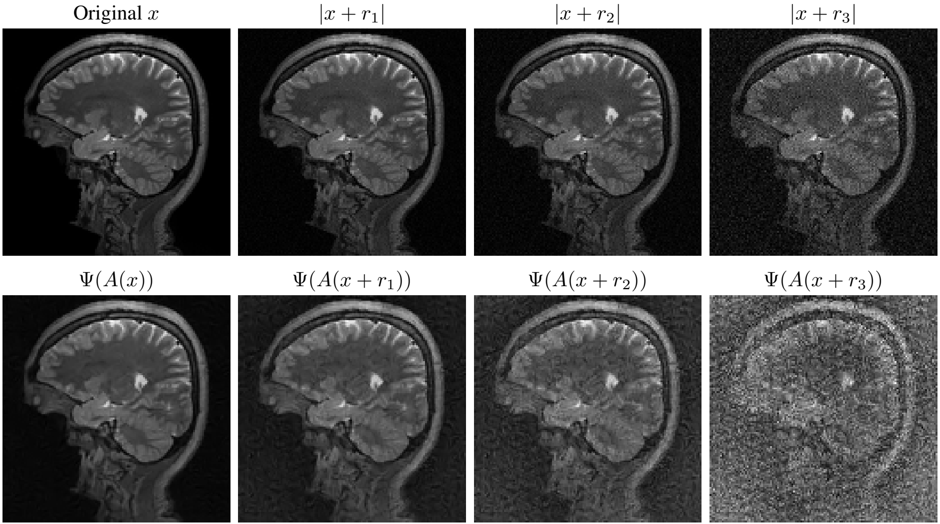
Figure: Images corrupted with 2% Gaussian noise and reconstructed using 15% sampling.

Demonstration of convergence



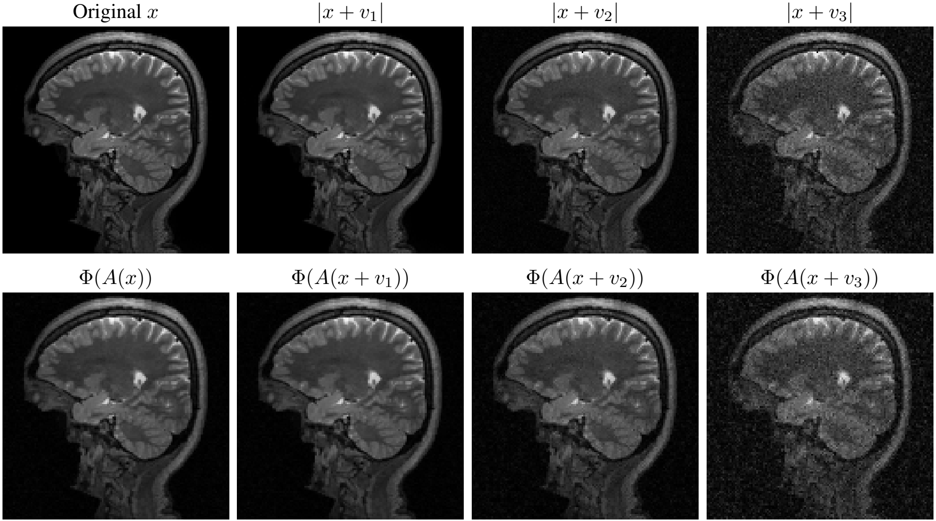


Stable? AUTOMAP ✗



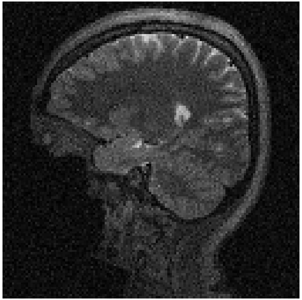
· V. Antun et al. "*On instabilities of deep learning in image reconstruction and the potential costs of AI*," PNAS, 2021.
· B. Zhu et al. "*Image reconstruction by domain-transform manifold learning*," Nature, 2018.

Stable? FIRENETs ✓



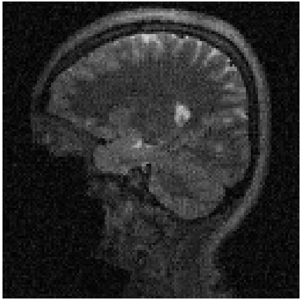
Adding FIRENET layers stabilizes AUTOMAP

$|x + \mathrm{e}_3|$



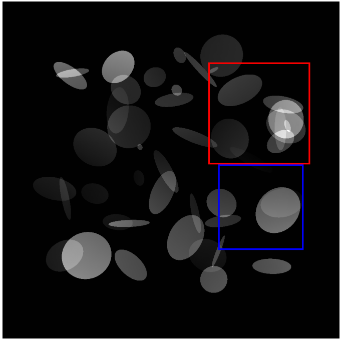
$\Psi(\tilde{y}),\ \tilde{y} = A(x + \mathrm{e}_3)$

$\Phi\left(\tilde{y}, \Psi(\tilde{y})\right)$

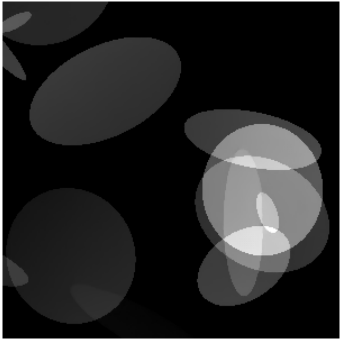


Stability vs. accuracy tradeoff

Original x
(full size)



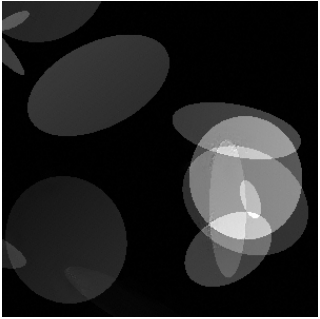
Original
(cropped, red frame)



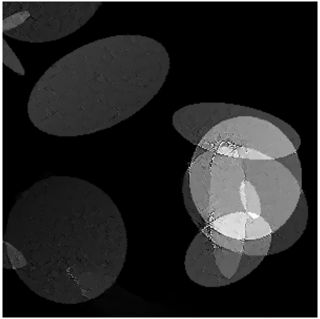
Original + detail $(x + h_1)$
(cropped, blue frame)

U-net trained without noise

Orig. + worst-case noise



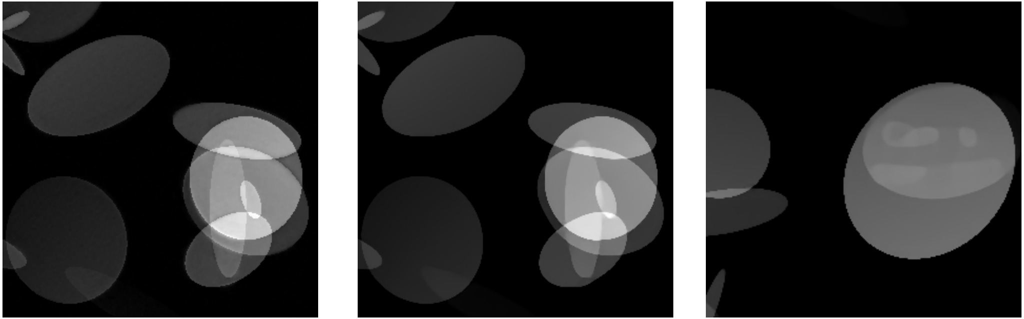
Rec. from worst-case noise



Rec. of detail

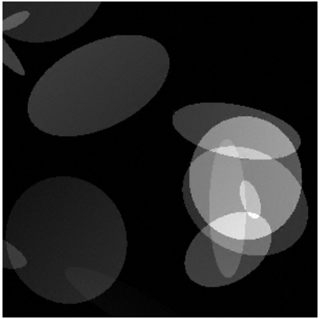
U-net trained with noise

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

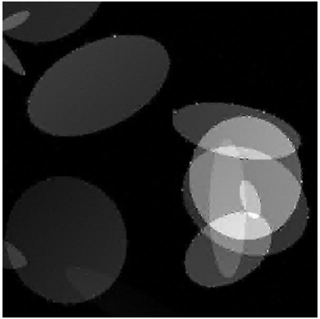


FIRENET

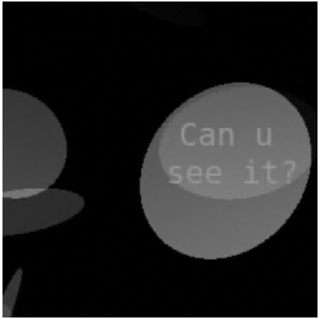
Orig. + worst-case noise



Rec. from worst-case noise



Rec. of detail



Broader framework: approximate sharpness conditions

Problem: Given $y = Ax + e \in \mathbb{C}^m$, recover $x \in \mathbb{C}^N$.

Optimization: $\min_{x\in\mathbb{C}^N} \mathcal{J}(x) + \|Bx\|_{\ell^1}$ s.t. $\|Ax - y\|_{\ell^2} \leq \epsilon$, seminorm $\mathcal{J}$, $B \in \mathbb{C}^{q\times N}$.

Assume: $\|\hat{x} - x\|_{\ell^2} \leq C_1\Big[\underbrace{\mathcal{J}(\hat{x}) + \|B\hat{x}\|_{\ell^1} - \mathcal{J}(x) - \|Bx\|_{\ell^1}}_{\text{objective function difference}} + C_2\underbrace{(\|A\hat{x} - y\|_{\ell^2} - \epsilon)}_{\text{feasibility gap}} + \underbrace{c(x,y)}_{\text{approx. term}}\Big].$

Examples: Sparse vector recovery, low-rank matrix recovery, matrix completion (local version holds), ℓ^1-analysis problems, TV minimization, mixed regularization problems, ...

Simplified version of Theorem: *Let $\delta > 0$. We provide a neural network ϕ of depth $\mathcal{O}(\log(\delta^{-1}))$ and width $\mathcal{O}(N+m+q)$ such that for all $(x,y) \in \mathbb{C}^N \times \mathbb{C}^m$*

$$\|Ax - y\|_{\ell^2} \leq \epsilon \text{ and } c(x,y) \leq \delta \quad \Rightarrow \quad \|\phi(y) - x\|_{\ell^2} \lesssim \delta.$$

· M. Colbrook "*WARPd: A linearly convergent first-order method for inverse problems with approximate sharpness conditions.*"

Weighted, Accelerated and Restarted Primal-dual (WARPd)

- Primal-dual iterations starting at x_0 ($X_k =$ ergodic average of first k iterates):

$$\underbrace{\mathcal{J}(X_k)+\|BX_k\|_{\ell^1}-\mathcal{J}(x)-\|Bx\|_{\ell^1}+C_2\left(\|AX_k-b\|_{\ell^2}-\epsilon\right)}_{=:G(X_k)} \leq \frac{1}{k}\left(\frac{\|x_0-x\|_{\ell^2}^2}{\tau_1}+\frac{C_2^2+q}{\tau_2}\right). \qquad (2)$$

- Assumption implies $\|X_k - x\|_{\ell^2} \leq C_1(G(X_k)+\delta)$, controls RHS of (2) upon restart.
- Reweighting trick and optimize parameters to form map H_k using k (constant) iterations s.t.

$$G(x_0) \leq \alpha_0 \Rightarrow G(H_k(x_0)) \leq \frac{C}{k}(\delta+\alpha_0)$$

- Restart iterations when $C/k \leq \nu \in (0,1)$ ($\nu = e^{-1}$ optimal). $\tilde{X}_p$ after p restarts:

$$G(\tilde{X}_p) \leq e^{-1}(\delta + e^{-1}(\delta + \ldots + e^{-1}(\delta+\alpha_0)) = (e^{-1}+e^{-2}+\ldots e^{-p})\delta + e^{-p}\alpha_0 \lesssim \delta + e^{-p}.$$

- Apply the assumption to get $\|\tilde{X}_p - x\|_{\ell^2} \lesssim \delta + e^{-p}$.

Remarks:

- Can be unrolled as a neural network (this inspired the architecture choice for FIRENETS).
  NB: Naive unrolling of PDHG gives slow $\mathcal{O}(\delta + p^{-1})$ convergence.
- Stability (w.r.t. input and execution) can be proven.
- If constants in assumption unknown, can perform grid search at extra logarithmic cost.

· A. Chambolle, T. Pock, '*A first-order primal-dual algorithm for convex problems with applications to imaging*,'' J Math Imaging Vis, 2011.

A final example with different regularizers

A is a DFT, 15% subsampled according to an inverse square law (optimal for TV).
Measurements are corrupted with 5% Gaussian noise.

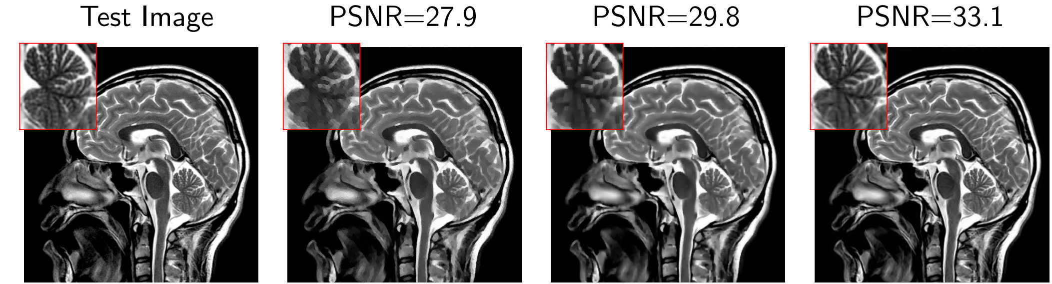


Figure: Middle-left: Converged reconstruction using TV. Middle-right: Converged reconstruction using TGV. Right: Reconstruction using (adaptively adjusted weighted) shearlets and TGV, after 25 iterations. All reconstructions were computed using WARPd.

WARPd can easily handle complicated mixed regularization problems.

$$\min_{x \in \mathbb{C}^N} \|WD^* x\|_{\ell^1} + \mathrm{TGV}_\alpha^2(x) \quad \text{s.t.} \quad \|Ax - b\|_{\ell^2} \leq \epsilon,$$

Concluding remarks

There is a **need for foundations** in AI/deep learning.

- Well-conditioned problems where mappings from training data to suitable neural networks exist, but no training algorithm (even randomized) can approximate them.
- Existence of training algorithms depends on desired accuracy. $\forall K \in \mathbb{Z}_{\geq 3}$, $\exists$ well-conditioned problems where simultaneously:
  - (i) Algorithms may compute neural networks to $K-1$ digits of accuracy, but not K.
  - (ii) Achieving $K-1$ digits of accuracy requires arbitrarily many training data.
  - (iii) Achieving $K-2$ correct digits requires only one training datum.
- Under *specific conditions*, algorithms can train stable and accurate neural networks. E.g., prove **FIRENETs** achieve exponential convergence & withstand adversarial attacks.
- There is a *trade-off* between stability and accuracy in deep learning.

- **WARPd** provides accelerated recovery under an *approximate sharpness condition*.
- Quantities controlling recovery also provide explicit approximate sharpness constants.
- WARPd $\underset{\text{unrolled}}{\Rightarrow}$ motivating architecture choices for FIRENETs.

Question: How do we *optimally* traverse the *stability* & *accuracy* trade-off?