Can stable and accurate neural networks always be computed?

Matthew Colbrook (Cambridge and École normale supérieure)
m.colbrook@damtp.cam.ac.uk

Joint work with: **Vegard Antun** (Oslo), **Anders Hansen** (Cambridge)

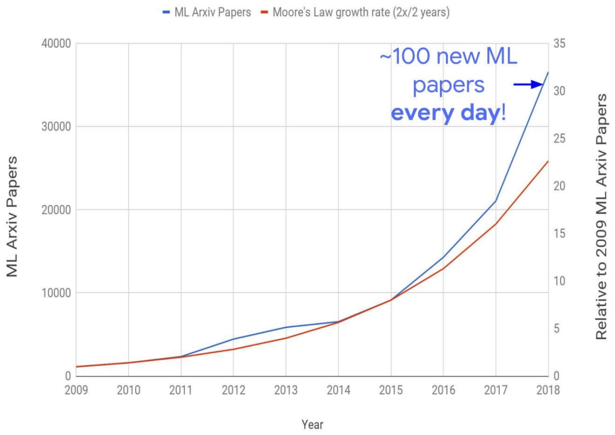
Based on: M. Colbrook, V. Antun, A. Hansen, "Can stable and accurate neural networks be computed? - On the barriers of deep learning and Smale's 18th problem" (to appear in PNAS)

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

Code: www.github.com/Comp-Foundations-and-Barriers-of-AI/firenet

Interest in deep learning unprecedented and exponentially growing

Machine learning papers on arXiv



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

The ubiquity of AI

AI techniques are replacing humans in problem solving:

- Self-driving vehicles
- Automated diagnosis in medicine and automated decision processes
- Automated weapon systems
- Music composition
- Call centres
- Any security system based on face or voice recognition
- Mathematical proofs

AI techniques are replacing established algorithms in science:

- Medical imaging (MRI, CT, etc.)
- Microscopy
- Imaging problems in general
- Radar, sonar, etc.
- Methods for solving PDEs

Augmenting the mathematician’s toolkit

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > articles > article

Article | Open Access | Published: 01 December 2021

Advancing mathematics by guiding human intuition with AI

Alex Davies, Petar Veličković, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomašev, Richard Tanburn, Peter Battaglia, Charles Blundell, András Juhász, Marc Lackenby, Geordie Williamson, Demis Hassabis & Pushmeet Kohli

Nature **600**, 70–74 (2021) | Cite this article

Abstract

The practice of mathematics involves discovering patterns and using these to formulate and prove conjectures, resulting in theorems. Since the 1960s, mathematicians have used computers to assist in the discovery of patterns and formulation of conjectures[1], most famously in the Birch and Swinnerton-Dyer conjecture[2], a Millennium Prize Problem[3]. Here we provide examples of new fundamental results in pure mathematics that have been discovered with the assistance of machine learning—demonstrating a method by which machine learning can aid mathematicians in discovering new conjectures and theorems. We propose a process of using machine learning to discover potential patterns and relations between mathematical objects, understanding them with attribution techniques and using these observations to guide intuition and propose conjectures. We outline this machine-learning-guided framework and demonstrate its successful application to current research questions in distinct areas of pure mathematics, in each case showing how it led to meaningful mathematical contributions on important open problems: a new connection between the algebraic and geometric structure of knots, and a candidate algorithm predicted by the combinatorial invariance conjecture for symmetric groups[4]. Our work may serve as a model for collaboration between the fields of mathematics and artificial intelligence (AI) that can achieve surprising results by leveraging the respective strengths of mathematicians and machine learning.

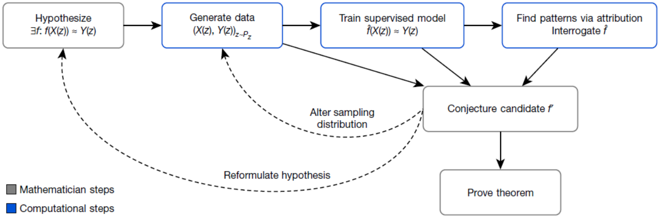


Fig. 1 | Flowchart of the framework. The process helps guide a mathematician’s intuition about a hypothesized function f, by training a machine learning model to estimate that function over a particular distribution of data P_Z. The insights from the accuracy of the learned function $\hat{f}$ and attribution techniques applied to it can aid in the understanding of the problem and the construction of a closed-form f'. The process is iterative and interactive, rather than a single series of steps.

Will AI replace standard algorithms in medical imaging?

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

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

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 tomography, ultrasound imaging and radio astronomy[1,2,3]. During image

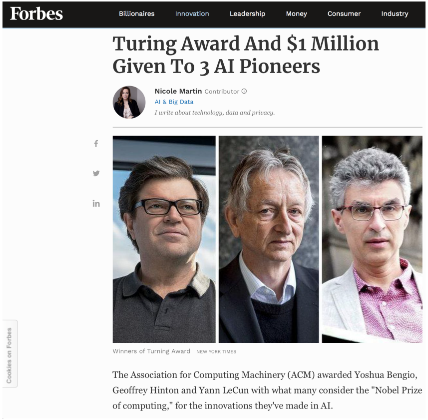
Editorial Summary

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

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.

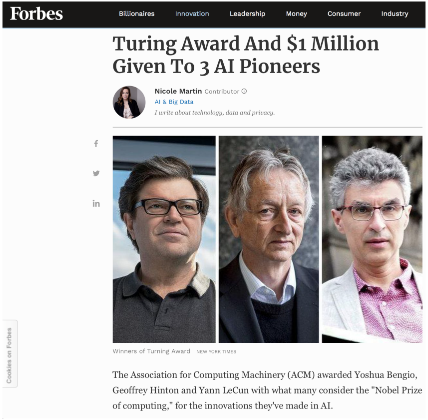
Winners of Turing Award NEW YORK TIMES

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

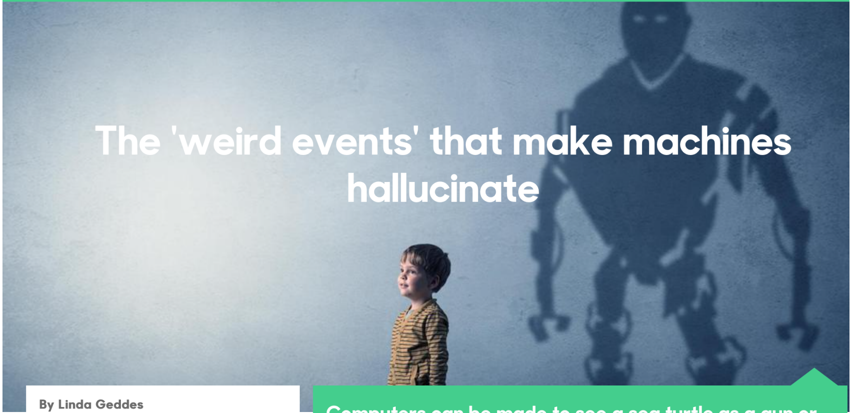
Geoffrey Hinton, The New Yorker, April 2017: “They should stop training radiologists now!”

BUT...

DANGER: AI generated hallucinations

BBC | Sign in | News | Sport | Weather | Shop | Reel | Travel | More | Search

machine minds | What is BBC Future? | Latest | Best of.. | Machine Minds | Future Now



By Linda Geddes
5 December 2018

Computers can be made to see a sea turtle as a gun or hear a concerto as someone's voice, which is raising concerns about using artificial intelligence in the real world.

The danger of false negatives

Uber's self-driving car saw the pedestrian but didn't swerve – report

Tuning of car's software to avoid false positives blamed, as US National Transportation Safety Board investigation continues

▲ Uber's modified Volvo XC90 SUV detected but did not react to the crossing pedestrian in first self-driving car fatality, report says. Photograph: Volvo

An Uber self-driving test car which killed a woman crossing the street detected her but decided not to react immediately, a report has said.

The car was travelling at 40mph (64km/h) in self-driving mode when it collided with 49-year-old Elaine Herzberg at about 10pm on 18 March. Herzberg was pushing a bicycle across the road outside of a crossing. She later died from her injuries.

Although the car's sensors detected Herzberg, its software which decides how it should react was tuned too far in favour of ignoring objects in its path which might be "false positives" (such as plastic bags), according to a report from the Information. This meant the modified Volvo XC90 did not react fast enough.

Deep Fool

FIGURE 4. Examples of natural images perturbed with the universal perturbation and their corresponding estimated labels with GoogLeNet. (a)–(h) Images belonging to the ILSVRC 2012 validation set. (i)–(l) Personal images captured by a mobile phone camera. (Figure used courtesy of [22].)

DL is also unstable in inverse problems!

PHYSICAL SCIENCES

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

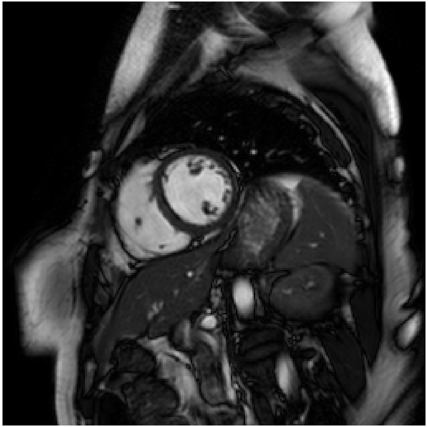
Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C. Hansen

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

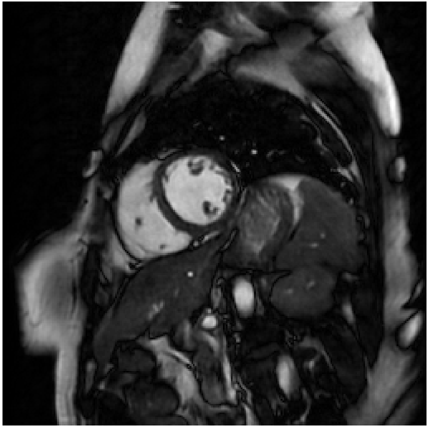
Edited by David L. Donoho, Stanford University, Stanford, CA, and approved March 12, 2020 (received for review June 4, 2019)

Example

$|x|$



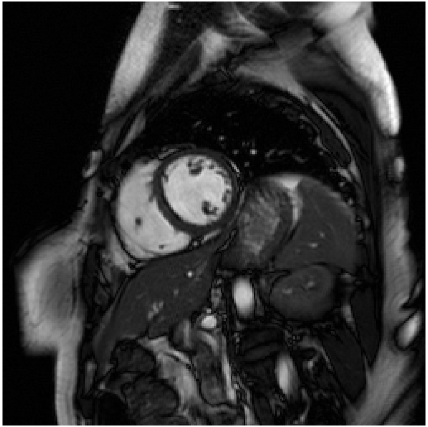
$|\Psi(Ax)|$



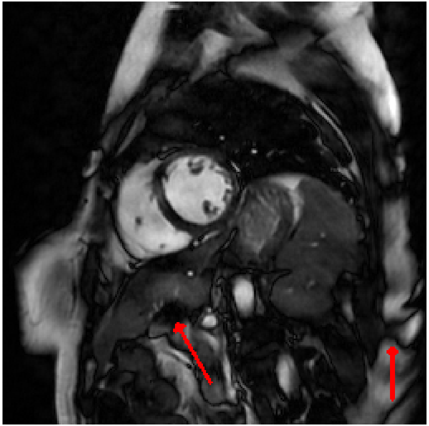
Network (33% subsampling) from: J. Schlemper, J. Caballero, J. V. Hajnal, A. Price and D. Rueckert, '*A deep cascade of convolutional neural networks for MR image reconstruction*', in International conference on information processing in medical imaging, Springer, 2017, pp. 647–658.
Figures from: Antun, 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

$|x + r_1|$



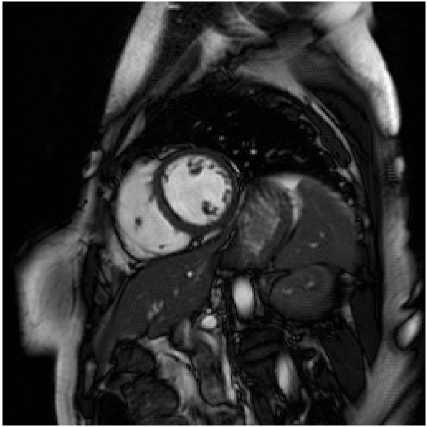
$|\Psi(A(x + r_1))|$



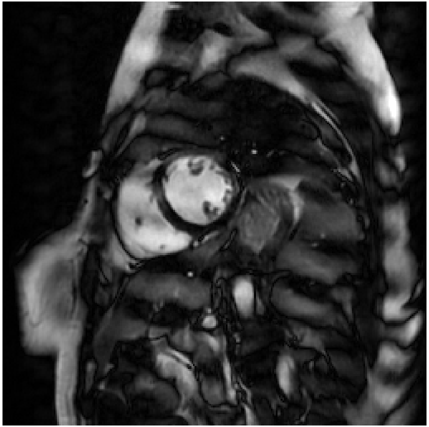
Network (33% subsampling) from: J. Schlemper, J. Caballero, J. V. Hajnal, A. Price and D. Rueckert, '*A deep cascade of convolutional neural networks for MR image reconstruction*', in International conference on information processing in medical imaging, Springer, 2017, pp. 647–658.
Figures from: Antun, 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

$|x + r_2|$



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

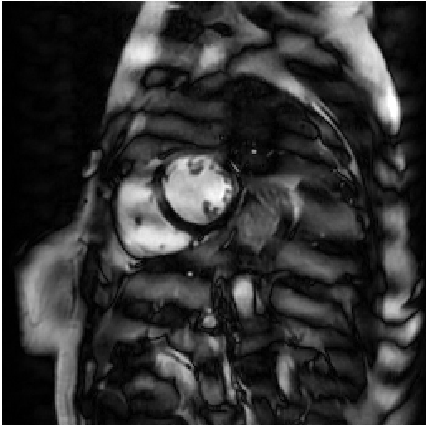


Network (33% subsampling) from: J. Schlemper, J. Caballero, J. V. Hajnal, A. Price and D. Rueckert, '*A deep cascade of convolutional neural networks for MR image reconstruction*', in International conference on information processing in medical imaging, Springer, 2017, pp. 647–658.
Figures from: Antun, 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

$|x + r_3|$

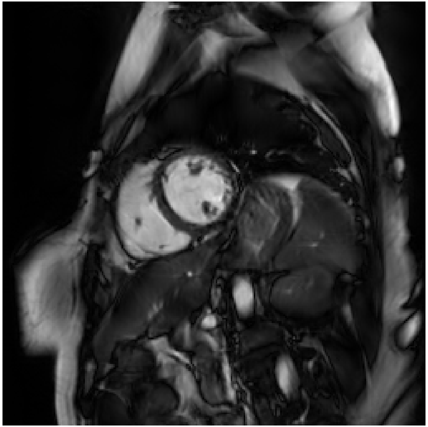
$|\Psi(A(x + r_3))|$



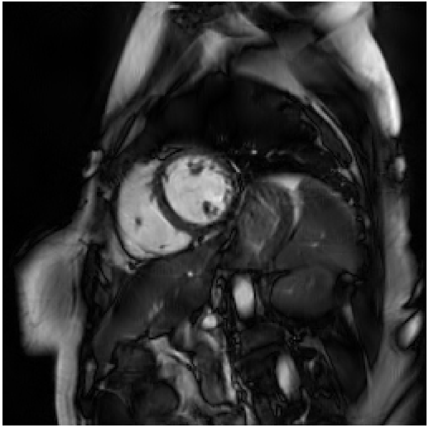
Network (33% subsampling) from: J. Schlemper, J. Caballero, J. V. Hajnal, A. Price and D. Rueckert, '*A deep cascade of convolutional neural networks for MR image reconstruction*', in International conference on information processing in medical imaging, Springer, 2017, pp. 647–658.
Figures from: Antun, 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

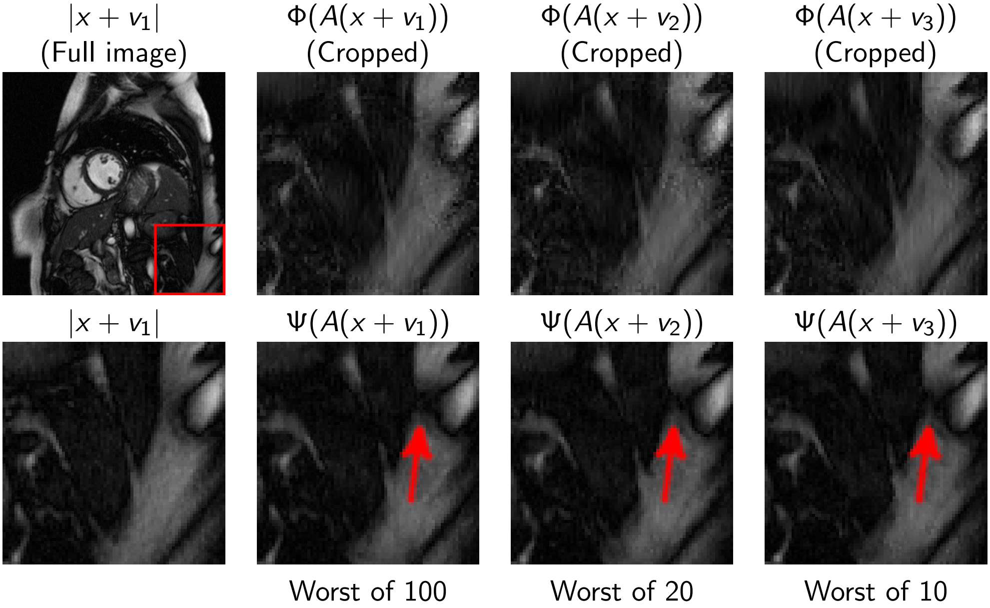
SoA from Ax



SoA from $A(x + r_3)$



AI generated hallucinations with random noise



Facebook and NYU's 2020 FastMRI challenge

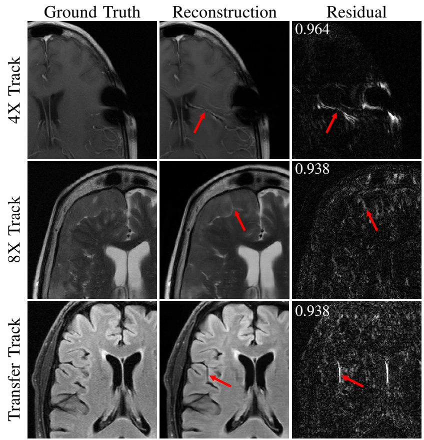


Fig. 6. Examples of reconstruction hallucinations among challenge submissions with SSIM scores over residual plots (residuals magnified by 5). (*top*) A 4X submission from Neurospin generated a false vessel, possibly related to susceptibilities introduced by surgical staples. (*middle*) An 8X submission 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.

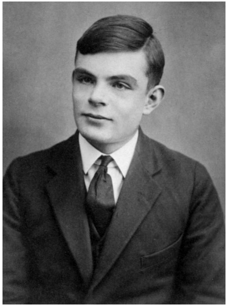
Optimism: Echoes of an old story

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

GOAL of 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)
- Balancing stability and accuracy

Can we compute neural networks that solve (P_j)?

Sparse regularisation (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 **minimising** 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 regularisation 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

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

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

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

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

Question: Given a collection Ω of $(A, \mathcal{S})$, does there <u>exist</u> a neural network approximating Ξ (solution map of (P_j)), and <u>can it be trained</u> 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* $K > 2$ *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 $K > 2$ 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* $K > 2$ *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* $y \in \mathcal{S}$,

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

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

Theorem

For (P_j), $N \geq 2$ and $m < N$. Let $K > 2$ 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 $y \in \mathcal{S}$,

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

(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 randomised can train (or compute) such a neural network accurate to K digits with probability greater than $1/2$.
- There exists a deterministic algorithm that computes a neural network with $K-1$ correct digits, but any such (even randomised) algorithm needs arbitrarily many training data.
- There exists a deterministic algorithm that computes 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/interpolation 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}	Ω_K
0.2999690	0.2597827	$n = 10$	10^{-1}	$K = 1$
0.3000000	0.2598050	$n = 20$	10^{-1}	$K = 1$
0.3000000	0.2598052	$n = 30$	10^{-1}	$K = 1$
0.0030000	0.0025980	$n = 10$	10^{-3}	$K = 3$
0.0030000	0.0025980	$n = 20$	10^{-3}	$K = 3$
0.0030000	0.0025980	$n = 30$	10^{-3}	$K = 3$
0.0000030	0.0000015	$n = 10$	10^{-6}	$K = 6$
0.0000030	0.0000015	$n = 20$	10^{-6}	$K = 6$
0.0000030	0.0000015	$n = 30$	10^{-6}	$K = 6$

Table: (**Impossibility of computing the existing neural network to arbitary 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: $f(x) < f^* + \epsilon$ does not in general imply $\min_{\hat{x}} \|x - \hat{x}\| \lesssim \epsilon$.

Question: Can we find 'good' input classes where

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

We shall see that the answer is yes!

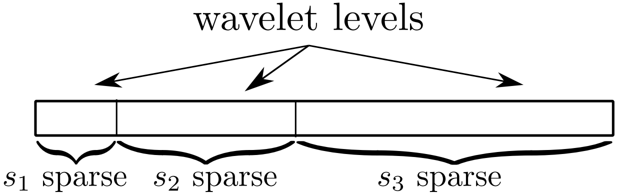
State-of-the-art model for sparse regularisation

Definition [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

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

The total sparsity is $s = s_1 + \ldots + s_r$. 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

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



The robust nullspace property

Definition [rNSP in levels]: Let $(\mathbf{s}, \mathbf{M})$ be local sparsities and sparsity levels respectively. $A \in \mathbb{C}^{m \times N}$ satisfies the robust null space property in levels (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,

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

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

Main result

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

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

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

More detailed comments

- Architecture inspired by restarted & reweighted unrolling of primal-dual algorithm.
- As well as stability, rNSPL allows exponential convergence.
- Naive unrolling gives slow convergence $\mathcal{O}(\delta + n^{-1})$.
- If we do not know constants for rNSPL, can perform log-scale grid search for suitable parameters (increase width of layers by a factor of $\log(n)$).
- Bound error of approximately applying hidden layers $\Rightarrow$ numerical stability.

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

Example in image recovery

Theorem

Consider recovering wavelet coeffs. $x = \Psi c$ *of* $c \in \mathbb{C}^{K^d}$ *from subsampled noisy Fourier or Walsh measurements* $y = DP_{\mathcal{I}}Vc + e$*. Let* $A = DP_{\mathcal{I}}V\Psi^*$, $m = |\mathcal{I}|$, $\epsilon_{\mathbb{P}} \in (0,1)$.

(i) *If* $\mathcal{I}$ *is a random sampling pattern drawn according to strategy in paper, and*

$$m \gtrsim (s_1 + \cdots + s_r) \cdot \mathcal{L}.$$

Then with prob. $1 - \epsilon_{\mathbb{P}}$, *A satisfies rNSPL of order* $(\mathbf{s}, \mathbf{M})$, $(\rho, \gamma) = (1/2, 2)$.

(ii) *For any* $\delta \in (0,1)$, *let* $\mathcal{J}(\delta, \mathbf{s}, \mathbf{M})$ *be the set of all* $y = Ax + e \in \mathbb{C}^m$ *where*

$$\|x\|_{\ell^2} \leq 1, \quad \max\{\sigma_{\mathbf{s},\mathbf{M}}(x)_{\ell^1}, \|e\|_{\ell^2}\} \leq \delta. \tag{2}$$

We provide an algorithm that constructs a neural network ϕ *with* $\mathcal{O}(\log(\delta^{-1}))$ *hidden layers (width* $2(N+m)$*) s.t. with probability at least* $1 - \epsilon_{\mathbb{P}}$,

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

Demonstration of convergence

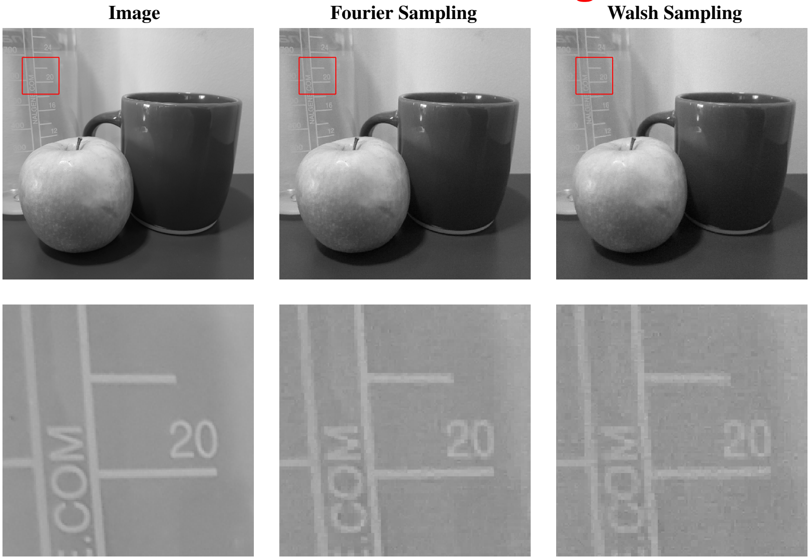
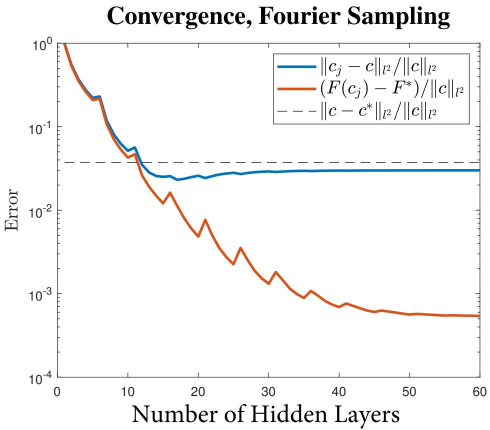
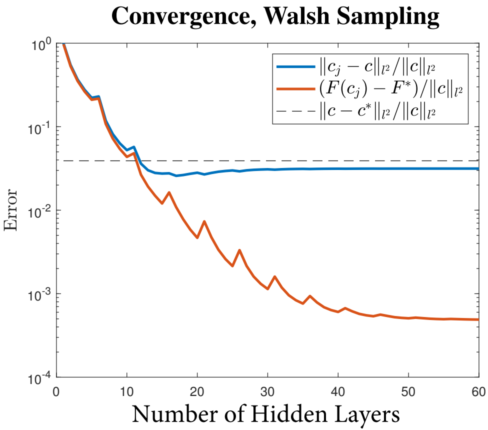


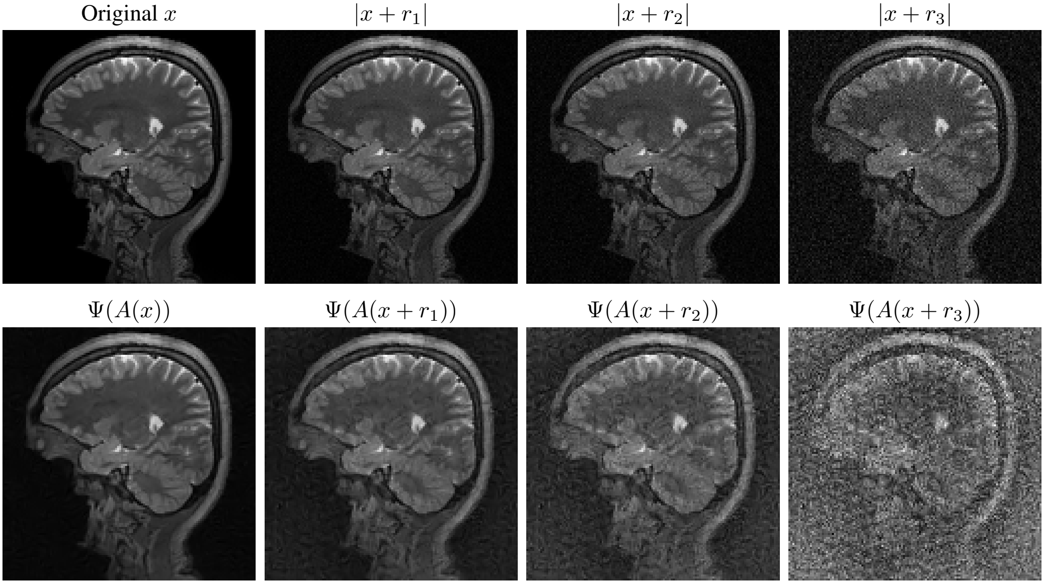
Figure: Images corrupted with 2% Gaussian noise and reconstructed using 15% sampling.

Demonstration of convergence

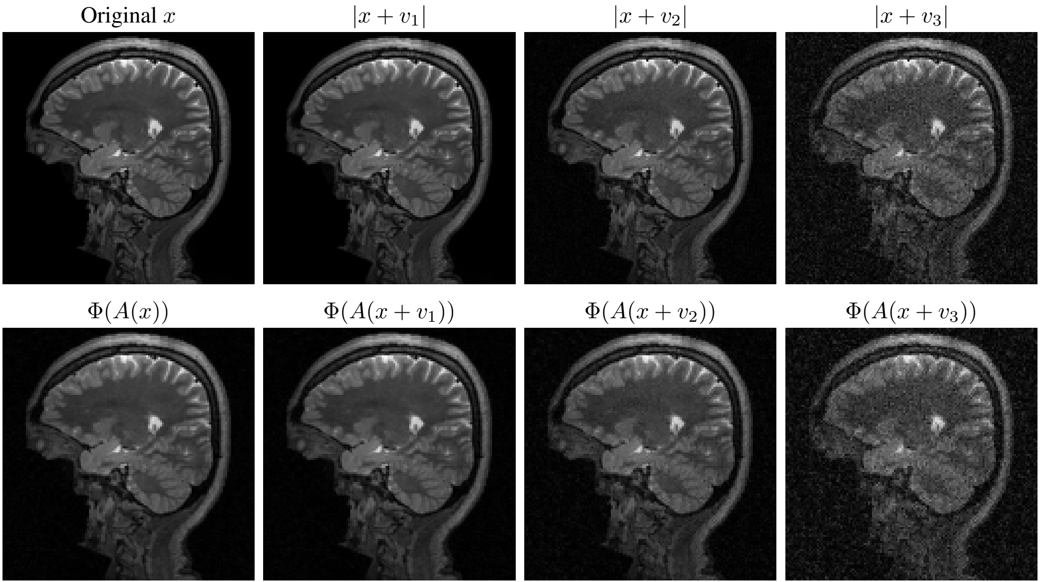




Stable? AUTOMAP ✗

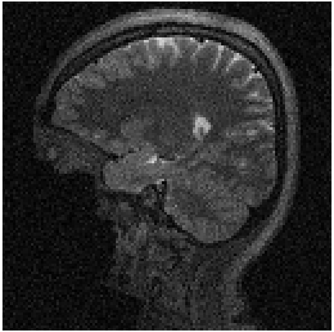


Stable? FIRENETs ✓



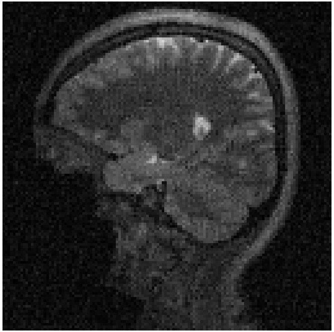
Adding FIRENET layers stabilises AUTOMAP

$|x + r_3|$



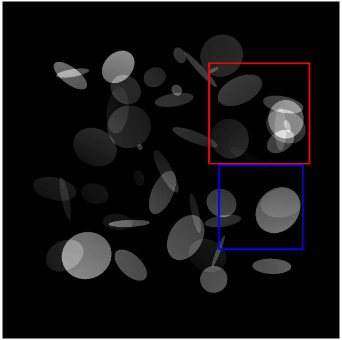
$\Psi(\tilde{y}),\ \tilde{y} = A(x + r_3)$

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

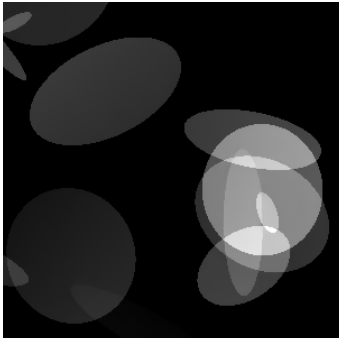


Stability and accuracy, and false negative

Original x
(full size)



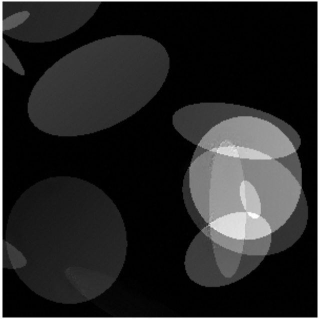
Original
(cropped, red frame)



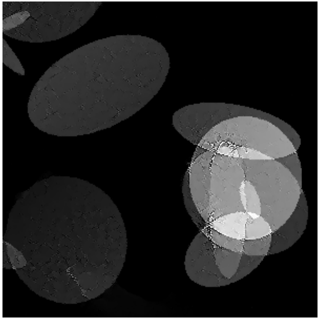
Original + detail ($x + h_1$)
(cropped, blue frame)

U-net trained without noise

Orig. + worst-case noise

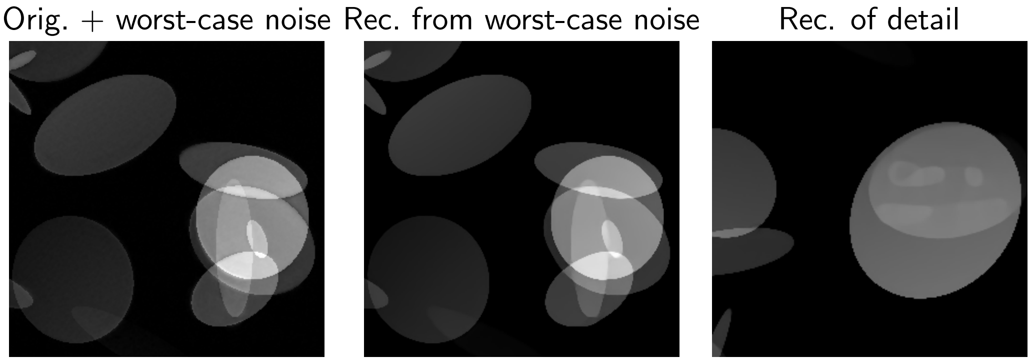


Rec. from worst-case noise



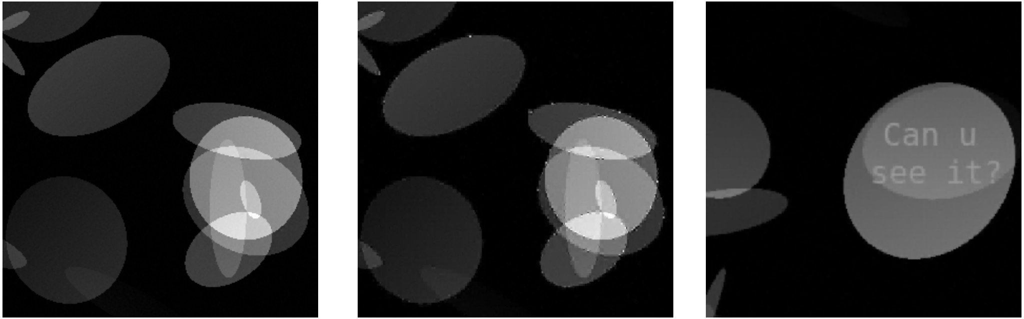
Rec. of detail

U-net trained with noise



FIRENET

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



Concluding remarks

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

- Well-conditioned problems where mappings from training data to suitable neural networks exist, but no training algorithm (even randomised) can approximate them.
- Existence of 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, there are algorithms that compute stable neural networks. E.g., Fast Iterative REstarted NETworks (FIRENETs) converge exponentially in the number of hidden layers. We prove FIRENETs withstand adversarial attacks.
- There is a trade-off between stability and accuracy in deep learning.

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

Different problems require different techniques for foundations!

Hopefully this talk has inspired you to build on these results and take up the challenge!