On the existence of stable and accurate neural networks for image reconstruction

Matthew J. Colbrook
University of Cambridge, UK
Email: m.colbrook@damtp.cam.ac.uk

Vegard Antun
University of Oslo, Norway
Email: vegarant@math.uio.no

Anders C. Hansen
University of Cambridge, UK
Email: a.hansen@damtp.cam.ac.uk

***Abstract*—Deep learning has emerged as a competitive new tool in image reconstruction. However, recent results demonstrate such methods are typically highly unstable - tiny, almost undetectable perturbations cause severe artefacts in the reconstruction, a major concern in practice. This is paradoxical given the existence of stable state-of-the-art methods for these problems. Thus, approximation theoretical results non-constructively imply the existence of stable and accurate neural networks. Hence the fundamental question:** *Can we explicitly construct/train stable and accurate neural networks for image reconstruction?* **We prove the answer is yes and construct such networks. Numerical examples of the competitive performance are also provided.**

***Index Terms*—Deep learning, image reconstruction, neural networks, stability, compressed sensing**

I. Introduction

The existence of stable, accurate and fast methods for image reconstruction from incomplete noisy measurements is a crucial problem in both mathematics, the physical and the life sciences. Real world applications include Magnetic Resonance Imaging (MRI), Computerised Tomography (CT), fluorescence microscopy, electron tomography, Nuclear Magnetic Resonance (NMR), radio interferometry, lens-less cameras etc. Accordingly, there is a vast literature on different optimisation methods and reconstruction models, see for example [4] and the references therein. Over the last decade compressed sensing and sparse regularisation have become standard tools in imaging, providing reduced scanning time and enhanced image resolution [12], [28], [31], [34], [37], [47]. However, deep learning has provided a new view on inverse problems and image reconstruction [5], [32], [36], [39], [40], [48], [56] that may change the field.

Deep learning and neural networks have provided state-of-the-art methods for many problems such as image recognition [33], [53] and speech recognition [30]. Other examples even include the prediction of effects of drugs [38], language translation [49] (for instance in Google translate) and speech understanding [22]. However, there is also an increasing awareness that this may come at a high price - many of these methods are unstable. It is now well established that high performance deep learning methods for image classification are subject to failure given tiny, almost invisible perturbations of the image [41], [42], [50]. This is not just restricted to image classification but also applies to image reconstructions from (possibly noisy) partial measurements [6] (also demonstrated in this paper). The instability is a major concern in applications such as medical diagnosis. Given that compressed sensing provides stable methods tackling this problem, it is paradoxical that deep learning may lead to instabilities for an inherently stable problem. Indeed, we have the following paradox:

> *There may exist stable and accurate neural networks for image reconstruction, yet, current state-of-the-art neural networks provided by deep learning become unstable.*

It is important to remark that the image classification problem is inherently unstable (this can be proven). However, this is not the case for the problem considered in this paper. A natural question is therefore:

Question 1. *Is it possible to construct stable networks with recovery guarantees for the problem of image reconstruction?*

So far this question has been unanswered with no constructive proof of the existence of such neural networks. We provide the answer to this question in the affirmative in the cases of Fourier measurements (e.g. MRI) and binary measurements (e.g. digital signal processing) in arbitrary dimensions. In particular, our results in dimension two cover images, though the results extend to arbitrary dimensions. We demonstrate (via a constructive proof) the existence of *untrained* recursive[1] neural networks that are stable and achieve recovery rates at least as optimal as (and in some cases better than) the current state-of-the-art in compressed sensing. We prove constructively that a precision of order $\delta > 0$ can be achieved via a *stable* neural network with $O(\delta^{-1})$ layers. Our results:

- Provide insight into architectures of stable neural networks,
- Give a lower bound on the expected performance of stable neural networks,
- Show that the field of compressed sensing (in particular the notion of sparsity) has useful contributions to the emerging field of the theory of deep learning and neural networks.
- Numerically demonstrate the recovery guarantee and stability of the new network.

Both the aforementioned instability of a current state-of-the-art neural network and the stability of the new network are numerically demonstrated in Section IV.

II. Background

In simple mathematical terms, the problem of image reconstruction is described as follows. Given an image $x \in \mathbb{C}^N$ (interpreted as a vector for simplicity), we are given access to measurements of the form

$$y = Ax + e, \tag{1}$$

where $A \in \mathbb{C}^{m \times N}$ represents the sampling modality (we assume under-sampling with $m < N$), such as a discrete Fourier transform modelling MRI, and e represents the error in the measurement due to effects such as random noise. The task is to reconstruct x from the noisy measurements y. Note that without additional assumptions, such as sparsity of the vector x, this problem is highly ill-posed.

[1] See below for the relevant definition, this is not to be confused with the term recursive often used to describe a particular architecture of a neural network.

The compressed sensing literature typically studies the error of approximating x via a solution of

$$\min_{z\in\mathbb{C}^N} \|Wz\|_{l^1} \quad \text{s.t.} \quad \|Az-y\|_{l^2} \leq \epsilon, \tag{2}$$

or similar optimisation problems, where W is a sparsifying transform. This is known as basis pursuit denoising (see for example [11], [13], [23], [25]). Recently, neural networks have been applied to the recovery problem, yielding alternative non-linear reconstruction techniques.

A. Definition of neural networks

For introductions to the field of deep learning/neural networks we refer the reader to [35] and the references therein. In order to capture architectures such as skip connections or switches, we consider the following definition of a neural network. A neural network is a mapping $\phi : \mathbb{C}^m \to \mathbb{C}^N$ such that we can write

$$\phi(y) = W_L(\rho_{L-1}(...\rho_1(W_1(y)))), \tag{3}$$

where;

- Each W_j is an affine map $\mathbb{C}^{N_{j-1}} \to \mathbb{C}^{N_j}$ given by $W_j(x) = A_j x + b_j(y)$ where $A_j \in \mathbb{C}^{N_j \times N_{j-1}}$ and the $b_j(y) \in \mathbb{C}^{N_j}$ are affine functions of the input y.
- Each ρ_j is a non-linear function and is one of two forms:
  1) There exists an index set $I_j \subset \{1, ..., N_j\}$ such that ρ_j applies a non-linear function f_j element wise on the input vector's components with indices in I_j. (I_j is allowed to be a strict subset.)
  2) There exists a non-linear function f_j such that, after decomposing the input vector x as $(x_0, X, Y)^T$ for scalar x_0 and $X \in \mathbb{C}^{n_j}$, we have

$$\rho_j : \begin{pmatrix} x_0 \\ X \\ Y \end{pmatrix} \to \begin{pmatrix} 0 \\ f_j(x_0)X \\ Y \end{pmatrix}. \tag{4}$$

The affine dependence of the bias terms $b_j(y)$ on y simply allows skip connections from the input to the current level as in standard definitions of feed-forward neural networks (see for example [46] page 269). The type of architecture above has become standard [29], [32], [45], [55]. Note that we do not allow the matrices A_j to depend on y. We will denote the collection of all neural networks of the above form by $\mathcal{NN}_{D,L,q}$, where the vector $D = (N_0 = m, N_1, ..., N_L = N)$ denotes the dimensions in each layer, L denotes the number of layers and q denotes the number of different non-linear functions applied (including different I_j and n_j). In general, we will require that the N_j do not grow with j so that the size of each layer is of the same order as the sampling matrix A.

Deep learning for image reconstruction is based on the idea of constructing (recursively) a neural network from a set of training images $\{x_1, \ldots, x_n\}$ and the operator A from (1). The problem is that such an approach typically results in unstable methods as documented in [6], see also Figure 1. The networks we consider are untrained and recursive in the following sense.

Definition 2 (Recursive neural network)**.** *A class $\mathcal{C}$ of neural networks $\phi \in \mathcal{NN}_{D,L,q}$, as defined above, depending on inputs $\iota \in \Theta$, where $\Theta \subset \mathbb{C}^p$ is some infinite subset, is said to be recursive if the mapping $\Theta \ni \iota \mapsto \phi$ is recursive. Recursive here means in the sense of Turing [52] if one is given rational inputs and in the sense of Blum-Shub-Smale (BSS) [10] if one is given non-rational inputs. In layman terms this means that there exists an algorithm that can construct ϕ from the input.*

Remark 3 (Recursive networks depending on A)**.** Throughout, we will only consider recursive networks that depend on the measurement matrix $\iota = A \in \Theta \subset \mathbb{C}^{N\times N}$ and the number of layers. In fact, the affine functions in our neural networks can be constructed using arithmetic operations on the matrix A.

It is possible to extend this framework to trained neural networks and associate with the constructive algorithm a training cost [14]. Future work will extend these techniques to study the problem of training cost versus stable recovery guarantees.

B. Sparsity in levels and multilevel random subsampling

In order to state our main theorem we recall some basics from compressed sensing. Motivated from the observation that the sparsity structure of natural images is highly structured in bases such as standard wavelet bases, [4] introduced a new structured sparsity model, wherein a vector x is allowed to have different sparsities in separate levels. Since its introduction, this model has been used to explain the effectiveness of compressed sensing in real life applications [7], [8], [54]. Indeed, sparsity alone turns out to be an inaccurate model in many applications as shown by the so-called flip test [4], [8], [44]. Hence, the following will be crucial for proving our main result.

Definition 4 (Sparsity in levels)**.** *For $r \in \mathbb{N}$, let $\mathbf{M} = (M_1, ..., M_r)$, where $1 \leq M_1 < ... < M_r = N$, and $\mathbf{s} = (s_1, ..., s_r)$, where $s_k \leq M_k - M_{k-1}$ for $k = 1, ..., r$ and $M_0 = 0$. A vector $x \in \mathbb{C}^N$ is $(\mathbf{s}, \mathbf{M})$-sparse in levels if*

$$|\mathrm{supp}(x) \cap \{M_{k-1}+1, ..., M_k\}| \leq s_k, \quad k = 1, ..., r. \tag{5}$$

We denote the set of $(\mathbf{s}, \mathbf{M})$-sparse vectors by $\Sigma_{\mathbf{s},\mathbf{M}}$.

For simplicity, we will assume throughout that each $s_k > 0$ and will denote the sum $s_1 + ... + s_k$ by s. We also need to describe a multilevel random sub-sampling model.

Definition 5 (Multilevel random sampling)**.** *Let $l \in \mathbb{N}, \mathbf{N} = (N_1, \ldots, N_l) \in \mathbb{N}^l$ with $1 \leq N_1 < \ldots < N_l$, $\mathbf{m} = (m_1, \ldots, m_l) \in \mathbb{N}^l$, with $m_k \leq N_k - N_{k-1}$, $k = 1, \ldots, l$, and suppose that*

$$\Omega_k \subset \{N_{k-1}+1, \ldots, N_k\}, \ |\Omega_k| = m_k, \quad k = 1, \ldots, l,$$

are chosen uniformly at random, where $N_0 = 0$. We refer to the set $\Omega = \Omega_{\mathbf{N},\mathbf{m}} = \Omega_1 \cup \ldots \cup \Omega_l$ as an $(\mathbf{N}, \mathbf{m})$- multilevel sampling scheme.

For a multilevel random sampling scheme $\Omega = \Omega_{\mathbf{N},\mathbf{m}}$, with $|\Omega| = m$, we let $P_\Omega \colon \mathbb{C}^N \to \mathbb{C}^m$ be the projection onto the canonical basis e_j indexed by Ω.

C. Fourier, Walsh function and wavelet bases

Our theorems consider the case where $A = P_\Omega U$ corresponds to either discrete Fourier measurements of the image or binary measurements (Walsh-Hadamard measurements). We work in d dimensions and assume that $N = 2^{r\cdot d}$, though the results trivially extend to rectangular images.

When using Fourier measurements, the matrix U corresponds to the d-dimensional discrete Fourier transform. For example, in the one dimensional case, let $x = \{x(t)\}_{t=0}^{N-1} \in \mathbb{C}^N$ be a signal. We denote its Fourier transform by

$$\mathcal{F}x(\omega) = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} x(t) \exp\big(\frac{2\pi i \omega t}{N}\big), \quad \omega \in \mathbb{R}$$

and let $U = U_{\mathrm{df1}} \in \mathbb{C}^{N\times N}$ denote the corresponding matrix such that

$$U_{\mathrm{df1}} x = \{\mathcal{F}x(\omega)\}_{\omega=-N/2+1}^{N/2}.$$

We divide the different frequencies into dyadic bands B_k, where $B_1 = \{0, 1\}$ and for $k = 2, ..., r$

$$B_k = \{-2^{k-1} + 1, ..., -2^{k-2}\} \cup \{2^{k-2} + 1, ..., 2^{k-1}\}.$$

In the general d-dimensional case we set

$$B_{\mathbf{k}}^{(d)} = B_{k_1} \times ... \times B_{k_d}, \quad \mathbf{k} = (k_1, ..., k_d) \in \mathbb{N}^d.$$

Given $(m_{\mathbf{k}=(k_1,...,k_d)})_{k_1,...,k_d=1}^r$ with $|m_{\mathbf{k}}| \leq \left|B_{\mathbf{k}}^{(d)}\right|$, we will use a multilevel random sampling such that $m_{\mathbf{k}}$ measurements are chosen uniformly and independently from $B_{\mathbf{k}}^{(d)}$. In Definition 5, this corresponds to $l = r^d$ and the N_i's can be chosen given a suitable ordering of the Fourier basis.

When using binary measurements, the matrix U corresponds to Walsh-Hadamard transform. The (Paley-ordered) Walsh functions are defined by

$$V_m(x) = (-1)^{\sum_{j=1}^{\infty} m_j x_j}, \quad x \in [0, 1), \quad m \in \mathbb{Z}_{\geq 0}, \tag{6}$$

where $(x_i)_{i\in\mathbb{N}}$ denotes the binary expansion of x (terminating if x is a dyadic rational) and $m = \sum_{j=1}^{\infty} m_j 2^{j-1}$. For properties we refer the reader to [27]. The vector form of these functions are

$$v_{j,p}(i) = 2^{-r/2} V_{2^j+p}(i2^{-r}), \tag{7}$$

for $i = 0, ..., 2^r - 1, j = 0, ..., r - 1, p = 0, ..., 2^j - 1$. For the general $d-$dimensional case we will take tensor products and define for $\mathbf{k} = (k_1, ..., k_d) \in \{1, ..., r\}^d$ the levels

$$W_{\mathbf{k}}^{(d)} = \{v_{k_1-1,p_1} \otimes ... \otimes v_{k_d-1,p_d} : 0 \leq p_j \leq 2^{k_j-1} - 1\}. \tag{8}$$

Again, this corresponds to $l = r^d$ and the N_i's can be chosen given a suitable ordering of the basis.

For the sparsifying basis we use the standard multidimensional Haar wavelets, formed via tensorizing the relevant multiresolution analysis. Sparsity in levels as per Definition 4 then corresponds to the r wavelet scales. We let $W \in \mathbb{C}^{N\times N}$ be the matrix corresponding to this basis so that Wx form the wavelet coefficients, i.e. we expect Wx to be sparse in the sense of Definition 4. Our results also carry over to the infinite-dimensional setting with the use of higher order Daubechies wavelets (though the results are more complicated to write down).

III. Main Result

To state our results in their most general form, let $\mathbf{w} = (w_i)_{i=1}^N$ be a vector of strictly positive weights and define the weighted l_w^1 norm via

$$\|x\|_{l_w^1} = \sum_{i=1}^{N} w_i |x_i|. \tag{9}$$

Given a sparsity pattern $(\mathbf{s}, \mathbf{M})$, define the best weighted $(\mathbf{s}, \mathbf{M})$-term approximation error as

$$\sigma_{\mathbf{s},\mathbf{M}}(x)_{l_w^1} = \inf\{\|x - z\|_{l_w^1} : z \in \Sigma_{\mathbf{s},\mathbf{M}}\}. \tag{10}$$

For an image which is compressible in the wavelet basis, $\sigma_{\mathbf{s},\mathbf{M}}(Wx)_{l_w^1}$ is expected to be small. In general, the weights are a prior on the anticipated support of the vector [26], [43], [51]. From now on, we will assume that if $M_{j-1}+1 \leq i \leq M_j$ then $w_i = w_{(j)}$ (i.e. constant in each level). We also define

$$\lambda(\mathbf{w}, \mathbf{s}) = \frac{\sum_{j=1}^r s_j w_{(j)}^2}{\min_{j=1,...,r} s_j w_{(j)}^2}, \quad \eta(\mathbf{w}, \mathbf{s}) = \sum_{j=1}^{r} s_j w_{(j)}^2, \tag{11}$$

and the quantities

$$\mathcal{M}_{\mathcal{F}}(\mathbf{s}, \mathbf{k}) = \sum_{l=1}^{\|\mathbf{k}\|_\infty} s_l \prod_{i=1}^{d} 2^{-|k_i - l|} + \sum_{l=\|\mathbf{k}\|_\infty+1}^{r} s_l 2^{-2(l-\|\mathbf{k}\|_\infty)} \prod_{i=1}^{d} 2^{-|k_i-l|}, \tag{12}$$

$$\mathcal{M}_{\mathcal{B}}(\mathbf{s}, \mathbf{k}) = s_{\|\mathbf{k}\|_\infty} \prod_{i=1}^{d} 2^{-|k_i - \|\mathbf{k}\|_\infty|}. \tag{13}$$

These will be used to state the sample size needed for the recovery guarantee.

Our main results state that stable neural networks exist with recovery guarantees for the problem of image reconstruction given noisy Fourier measurements or noisy binary measurements. The notation $a \lesssim b$ means that there is some constant (in our case of order 1 and explicit from the proof), independent of all the parameters, such that $a \leq Cb$.

Theorem 6 (Stable Neural Networks Exist)**.** *Let $\epsilon_{\mathbb{P}} \in (0, 1)$, $r, d \in \mathbb{N}$, $N = 2^{r\cdot d}$ and $\mathbf{M} = (M_1, ..., M_r)$, $\mathbf{s} = (s_1, ..., s_r)$ describe $(\mathbf{s}, \mathbf{M})$-sparse vectors corresponding to the scales in a d-dimensional wavelet basis. Let $\Omega = \Omega_{\mathbf{N},\mathbf{m}}$ be the multilevel sampling scheme discussed in §II-C where the $m_{\mathbf{k}}$ satisfy*

$$m_{\mathbf{k}} \gtrsim \lambda(\mathbf{w}, \mathbf{s}) \cdot \mathcal{M}_{\mathcal{F}}(\mathbf{s}, \mathbf{k}) \cdot L, \tag{14}$$

in the case of Fourier measurements and

$$m_{\mathbf{k}} \gtrsim \lambda(\mathbf{w}, \mathbf{s}) \cdot \mathcal{M}_{\mathcal{B}}(\mathbf{s}, \mathbf{k}) \cdot L, \tag{15}$$

in the case of binary measurements. Here L is a logarithmic factor given by

$$L = d \cdot r^2 \cdot \log(m) \cdot \log^2(s\lambda(\mathbf{w}, \mathbf{s})) + \log(\epsilon_{\mathbb{P}}^{-1}). \tag{16}$$

Then, for each $n \in \mathbb{N}$, there exists a recursive neural network ϕ_n^A, i.e. the map

$$(n, A) \to \phi_n^A \tag{17}$$

is recursive, with the following properties:

1) *(Stability and accuracy) With probability at least $1 - \epsilon_{\mathbb{P}}$, the following uniform recovery guarantee holds. For any $x \in \mathbb{C}^N$ with $\|x\|_{l^2} \lesssim 1$ and $y = P_\Omega U x + e$,*

$$\|\phi_n^A(y) - x\|_{l^2} \lesssim \frac{\lambda(\mathbf{w}, \mathbf{s})^{\frac{1}{4}}}{\sqrt{\eta(\mathbf{w}, \mathbf{s})}} \sigma_{\mathbf{s},\mathbf{M}}(Wx)_{l_w^1} + \lambda(\mathbf{w}, \mathbf{s})^{\frac{1}{4}} \|e\|_{l^2} + \frac{\lambda(\mathbf{w}, \mathbf{s})^{\frac{1}{4}}}{n} \sup_{\boldsymbol{k}\in\{1,...,r\}^d} \frac{\left|B_{\boldsymbol{k}}^{(d)}\right|}{\sqrt{m_{\boldsymbol{k}}}}. \tag{18}$$

2) *($3n$ layers of bounded size) $\phi_n^A \in \mathcal{NN}_{D,3n,3}$ with*

$$D = (m, m, \underbrace{2N + m, 2(N + m), 2N + m + 1}_{\text{repeated } n-1 \text{ times}}, 2N, N). \tag{19}$$

Remark 7. The word recursive in the theorem means in the BSS sense and Θ as in Definition 2 is the set of all $N \times N$ matrices. However, the theorem, with a slightly different wording, also holds in the Turing sense with Θ being the set of all $N \times N$ matrices with computable (Turing sense) entries.

Note that we can choose the weights $w_{(j)} = \sqrt{s/s_j}$ to minimise (14) so that $\lambda(\mathbf{w}, \mathbf{s}) = r$ and $\eta(\mathbf{w}, \mathbf{s}) = rs$. Up to log-factors, this measurement condition then becomes equivalent to that for the oracle

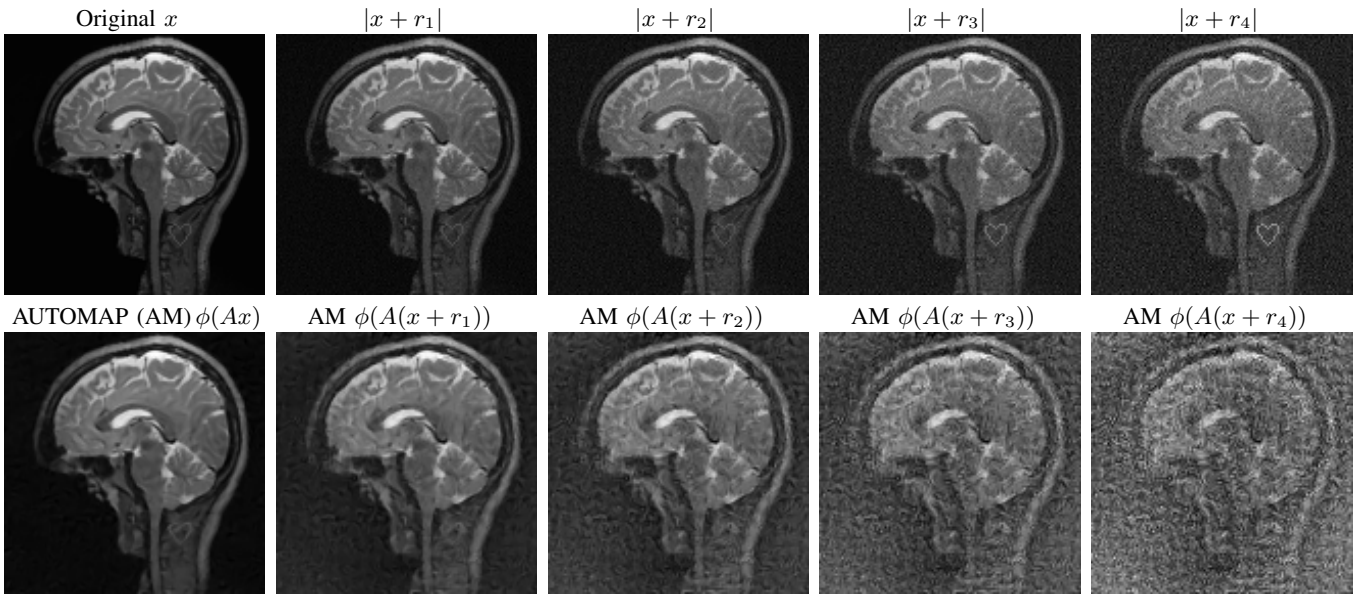

Fig. 1. Results of stability test for AUTOMAP taken from [6], and where $A = P_\Omega U$ is a subsampled Fourier transform. To visualise we show $|x + r_j|$. Top row: original image with perturbations r_j. Bottom row: reconstructions from $A(x + r_j)$ by the deep learning AUTOMAP (AM) network ϕ [57]. A detail in form of a heart (with increasing intensity) is added to visualise the loss in quality.

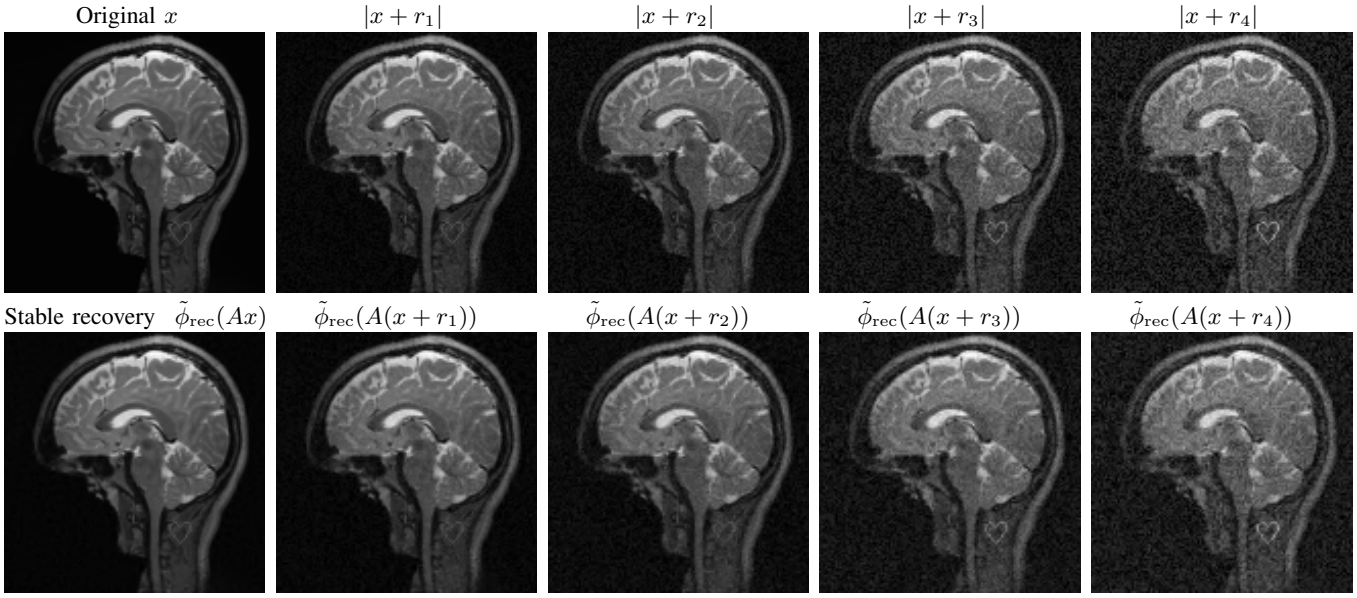

Fig. 2. Results of stability test for new stable networks, where $A = P_\Omega U$ is a sub-sampled Fourier transform (same measurement matrix as Figure 1). These perturbations are of the same size (measured in l^2 norm) as in Figure 1 but are found by searching for instabilities of the new neural network $\tilde{\phi}_{\text{rec}}$ (and hence the perturbations differ from those in Figure 1). Again, to visualise we show $|x + r_j|$. Top row: original image with perturbations r_j. Bottom row: reconstructions from $A(x + r_j)$ by the stable neural network $\tilde{\phi}_{\text{rec}}$.

estimator (where one assumes a-priori knowledge of the support of the vector) [1] and we also have

$$L \approx \log(N)^2 \log(m) d^{-1}. \tag{20}$$

Further interpretations of the measurement conditions (14) and (15) are as follows. In the Fourier case, if $d = 1$, this estimate yields the multi-level sampling estimates

$$m_k \gtrsim \left(s_k + \sum_{l=1}^{k-1} s_l 2^{-(k-l)} + \sum_{l=k+1}^{r} s_l 2^{-3(l-k)} \right) rL. \tag{21}$$

In other words, up to logarithmic factors s_l measurements are needed in each level. Furthermore, if $s_1 = ... = s_r = s_*$ and $d = 2$ then (14) is implied if

$$m_{(k_1,k_2)} \gtrsim s_* 2^{-|k_1-k_2|} rL. \tag{22}$$

Another interpretation is gained by considering

$$m_k = \sum_{\|\mathbf{k}\|=k} m_{\mathbf{k}}, \quad k = 1, ..., r, \tag{23}$$

the number of samples per annular region. We then have

$$m_k \gtrsim 3^d d \left(s_k + \sum_{l=1}^{k-1} s_l 2^{-(k-l)} + \sum_{l=k+1}^{r} s_l 2^{-3(l-k)} \right) rL, \tag{24}$$

which is the same estimate as the one dimensional case for bounded d. Note that the number of samples required in each annular region is (up to the usual logarithmic factors) proportional to the corresponding sparsity s_k with additional exponentially decaying terms dependent on $s_l, l \neq k$. In the binary measurement case, (22) remains the same whereas (24) becomes

$$m_k \gtrsim 2^d d s_k r L, \tag{25}$$

and there are no terms from the sparsity levels $s_l, l \neq k$.

Finally we remark that the result can be extended to infinite dimensional models of spaces such as $L^2([0,1]^d)$ [14]. We refer the reader to [2], [3] for compressed sensing in infinite dimensions.

IV. Numerical Example

To demonstrate the instabilities of deep learning, and the stability of the new neural networks, we perform an instability test [6] for the case of Fourier measurements. The test includes an algorithm that does the following. Given an image $x \in \mathbb{C}^N$ and measurement matrix $A = P_\Omega U \in \mathbb{C}^{m \times N}$ as above, consider a neural network $\phi : \mathbb{C}^m \to \mathbb{C}^N$ which aims to reconstruct the image $\phi(y) \approx x$ from the (noisy) measurements $y = Ax + e$. The algorithm seeks a vector $r \in \mathbb{R}^N$ such that

$$\|\phi(y + Ar) - \phi(y)\|_{l^2} \text{ is large, while } \|r\|_{l^2} \text{ is small.} \tag{26}$$

In other words, the algorithm searches for a perturbation of the image that makes the most severe change in the output of the network while still keeping the perturbation small. Specifically, consider the optimisation problem

$$r^*(y) \in \operatorname*{argmax}_r \frac{1}{2}\|\phi(y + Ar) - x\|_2^2 - \frac{\lambda}{2}\|r\|_2^2. \tag{27}$$

The problem (27) seeks perturbations in the image domain since this provides an easy way to compare the original image with a perturbed image and deduce whether the reconstruction of the perturbed image is acceptable/unacceptable. Of course, we could have just as easily considered perturbations in the sampling domain instead.

Due to the non-linearity of ϕ, finding a global maximiser of (27) is very difficult (if not impossible), even for small values of m and N. The test aims to locate local maxima of (27) by using a gradient search method. Let

$$Q_y^\phi(r) = \frac{1}{2}\|\phi(y + Ar) - x\|_2^2 - \frac{\lambda}{2}\|r\|_2^2 \tag{28}$$

be the objective function. A natural method to solve (27) is *gradient ascent with momentum*. This uses the gradient of Q_y^ϕ (which can easily be written down) along with two parameters $\gamma > 0$ (the momentum) and $\eta > 0$ (the learning rate) in each step towards a local maximum. Namely, $r(0)$ is initialised randomly and then we update the perturbation at the ith step via $v(i+1) = \gamma v(i) + \eta \nabla_r Q_y^\phi(r(i))$ and $r(i+1) = r(i) + v(i+1)$. The final perturbation is taken after M steps, where typically we run a few hundred steps, seeking the perturbation which causes the worst reconstructed image. Just as in the case when training neural networks using stochastic gradient descent with momentum, choosing the parameters γ and η is an art of engineering, and the optimal choices of γ, η are based on empirical testing.

First we report the results of [6] on the AUTOMAP [57] network used for MRI reconstruction with 60% subsampling (this is considered the current state-of-the-art). The network weights are provided by the authors of [57] and had been trained on de-identified brain images from the MGH–USC HCP dataset [24] where the image measurements $y = Ax + e$ were contaminated with small additive white noise e. In this test an image x from the mentioned dataset is picked with an added detail in form of a heart to easier see the loss of quality (see Figure 1). The mentioned algorithm is run on the AUTOMAP network to find a sequence of perturbations $|r_1| < |r_2| < |r_3| < |r_4|$. In order to illustrate how small the perturbations are we have visualised $|x + r_j|$ in the first row of Figure 1. As can be seen from the second row in the figure, the network reconstruction completely deforms the image and the added detail gradually disappears completely in the network reconstruction (similar results hold without this structural change).

In contrast, we have performed the instability test, but now for the new neural networks reported in this paper. Figure 2 shows the instability test applied to the constructed stable neural networks described by Theorem 6. We now see that despite the search for adversarial perturbations, the reconstruction remains stable. The error in the reconstruction was also found to be of the same order of the perturbation (as expected from the stability Theorem 6).

In applying the test to the new stable neural networks, we tested/tuned the parameters in the gradient ascent algorithm considerably (much more so than was needed for applying the test to AUTOMAP where finding instabilities was straightforward), yet we could not produce instabilities (as expected from Theorem 6). However, it should be mentioned that this search algorithm is just one form of test. It is likely that there are many other tests for creating instabilities for neural networks for inverse problems. Future work will consider these and also apply them to the new class of constructed neural networks.

Trained neural networks do not come with any understanding nor guarantee of their accuracy or stability. This poses a problem as real world measurements always come with noise, both structural and random. It is, however, not clear how to protect against the potential bad tiny noise. Indeed, a detail may be washed out, as shown in the experiment, but the similarity between the standard artefact may make it difficult to judge that this is an untrustworthy image. Thus stable and accurate neural networks are needed. Here we have constructed and demonstrated the first such class of neural networks for image reconstruction. Further results can be proven using the setup of the SCI hierarchy [9], [15]–[21].

V. Acknowledgments

MJC acknowledges support from EPSRC grant EP/L016516/1. ACH thanks NVIDIA for a GPU grant in form of a Titan X Pascal and acknowledges support from a Royal Society University Research Fellowship and EPSRC grant EP/L003457/1.

References

[1] B. Adcock, C. Boyer, and S. Brugiapaglia. On oracle-type local recovery guarantees in compressed sensing. *arXiv preprint arXiv:1806.03789*, 2018.

[2] B. Adcock and A. C. Hansen. A generalized sampling theorem for stable reconstructions in arbitrary bases. *Journal of Fourier Analysis and Applications*, 18(4):685–716, 2012.

[3] B. Adcock and A. C. Hansen. Generalized sampling and infinite-dimensional compressed sensing. *Foundations of Computational Mathematics*, 16(5):1263–1323, 2016.

[4] B. Adcock, A. C. Hansen, C. Poon, and B. Roman. Breaking the coherence barrier: A new theory for compressed sensing. In *Forum of Mathematics, Sigma*, volume 5. Cambridge University Press, 2017.

[5] J. Adler and O. Öktem. Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Problems*, 33(12):124007, 2017.

[6] V. Antun, F. Renna, C. Poon, B. Adcock, and A. C. Hansen. On instabilities of deep learning in image reconstruction - Does AI come at a cost? *Submitted*, 2019.

[7] A. Bastounis, B. Adcock, and A. C. Hansen. From global to local: Getting more from compressed sensing. *SIAM News*, Oct., 2017.

[8] A. Bastounis and A. C. Hansen. 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 on Imaging Sciences*, 10(1):335–371, 2017.

[9] J. Ben-Artzi, M. J. Colbrook, A. C. Hansen, O. Nevanlinna, and M. Seidel. Computing Spectra – On the Solvability Complexity Index hierarchy and towers of algorithms. *arXiv:1508.03280v5*, 2020.

[10] L. Blum, F. Cucker, M. Shub, and S. Smale. *Complexity and Real Computation*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1998.

[11] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[12] K. Choi, J. Wang, L. Zhu, T.-S. Suh, S. Boyd, and L. Xing. Compressed sensing based cone-beam computed tomography reconstruction with a first-order method a. *Medical physics*, 37(9):5113–5125, 2010.

[13] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *Journal of the American mathematical society*, 22(1):211–231, 2009.

[14] M. J. Colbrook. Stable and accurate neural networks for image reconstruction.

[15] M. J. Colbrook. Computing spectral measures and spectral types. *arXiv:1908.06721v2*, 2019.

[16] M. J. Colbrook. The foundations of spectral computations via the solvability complexity index hierarchy: Part II. *arXiv:1908.09598*, 2019.

[17] M. J. Colbrook. *The Foundations of Infinite-Dimensional Spectral Computations*. PhD thesis, University of Cambridge, 2020.

[18] M. J. Colbrook. Pseudoergodic operators and periodic boundary conditions. *Mathematics of Computation*, 89(322):737–766, 2020.

[19] M. J. Colbrook and A. C. Hansen. The foundations of spectral computations via the solvability complexity index hierarchy: Part I. *arXiv:1908.09592*, 2019.

[20] M. J. Colbrook and A. C. Hansen. On the infinite-dimensional QR algorithm. *Numerische Mathematik*, 143(1):17–83, 2019.

[21] M. J. Colbrook, B. Roman, and A. C. Hansen. How to compute spectra with error control. *Physical Review Letters*, 122(25):250201, 2019.

[22] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

[23] D. L. Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[24] Q. Fan, T. Witzel, A. Nummenmaa, K. R. Van Dijk, J. D. Van Horn, M. K. Drews, L. H. Somerville, M. A. Sheridan, R. M. Santillana, J. Snyder, et al. MGH–USC human connectome project datasets with ultra-high b-value diffusion mri. *Neuroimage*, 124:1108–1114, 2016.

[25] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*, volume 1. Birkhäuser Basel, 2013.

[26] M. P. Friedlander, H. Mansour, R. Saab, and Ö. Yilmaz. Recovering compressively sampled signals using partial support information. *IEEE Transactions on Information Theory*, 58(2):1122–1134, 2012.

[27] B. Golubov, A. Efimov, and V. Skvortsov. *Walsh series and transforms: theory and applications*, volume 64. Springer Science & Business Media, 2012.

[28] M. Guerquin-Kern, M. Haberlin, K. P. Pruessmann, and M. Unser. A fast wavelet-based reconstruction method for magnetic resonance imaging. *IEEE transactions on medical imaging*, 30(9):1649–1660, 2011.

[29] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll. Learning a variational network for reconstruction of accelerated mri data. *Magnetic resonance in medicine*, 79(6):3055–3071, 2018.

[30] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

[31] D. J. Holland, M. J. Bostock, L. F. Gladden, and D. Nietlispach. Fast multidimensional nmr spectroscopy using compressed sensing. *Angewandte Chemie*, 123(29):6678–6681, 2011.

[32] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[34] R. Leary, Z. Saghi, P. A. Midgley, and D. J. Holland. Compressed sensing electron tomography. *Ultramicroscopy*, 131:70–91, 2013.

[35] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436 EP –, 05 2015.

[36] A. Lucas, M. Iliadis, R. Molina, and A. K. Katsaggelos. Using deep neural networks for inverse problems in imaging: beyond analytical methods. *IEEE Signal Processing Magazine*, 35(1):20–36, 2018.

[37] M. Lustig, D. Donoho, and J. M. Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

[38] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik. Deep neural nets as a method for quantitative structure–activity relationships. *Journal of chemical information and modeling*, 55(2):263–274, 2015.

[39] M. Mardani, E. Gong, J. Y. Cheng, S. Vasanawala, G. Zaharchuk, M. Alley, N. Thakur, S. Han, W. Dally, J. M. Pauly, et al. Deep generative adversarial networks for compressed sensing automates mri. *arXiv preprint arXiv:1706.00051*, 2017.

[40] M. T. McCann, K. H. Jin, and M. Unser. Convolutional neural networks for inverse problems in imaging: A review. *IEEE Signal Processing Magazine*, 34(6):85–95, Nov 2017.

[41] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. *arXiv preprint*, 2017.

[42] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016.

[43] H. Rauhut and R. Ward. Interpolation via weighted l1 minimization. *Applied and Computational Harmonic Analysis*, 40(2):321–351, 2016.

[44] B. Roman, A. Hansen, and B. Adcock. On asymptotic structure in compressed sensing. *arXiv preprint arXiv:1406.4178*, 2014.

[45] 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*, pages 647–658. Springer, 2017.

[46] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[47] V. Studer, J. Bobin, M. Chahid, H. S. Mousavi, E. Candes, and M. Dahan. Compressive fluorescence microscopy for biological and hyperspectral imaging. *Proceedings of the National Academy of Sciences*, 109(26):E1679–E1687, 2012.

[48] J. Sun, H. Li, Z. Xu, et al. Deep admm-net for compressive sensing mri. In *Advances in Neural Information Processing Systems*, pages 10–18, 2016.

[49] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[50] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[51] Y. Traonmilin and R. Gribonval. Stable recovery of low-dimensional cones in hilbert spaces: One rip to rule them all. *Applied and Computational Harmonic Analysis*, 45(1):170–205, 2018.

[52] A. M. Turing. On Computable Numbers, with an Application to the Entscheidungsproblem. *Proc. London Math. Soc.*, S2-42(1):230, 1936.

[53] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun, and R. Fergus. Regularization of neural networks using dropconnect. In *International Conference on Machine Learning*, pages 1058–1066, 2013.

[54] Q. Wang, M. Zenge, H. E. Cetingul, E. Mueller, and M. S. Nadar. Novel sampling strategies for sparse mr image reconstruction. *Proc. Int. Soc. Mag. Res. in Med.,(22)*, 2014.

[55] G. Yang, S. Yu, H. Dong, G. Slabaugh, P. L. Dragotti, X. Ye, F. Liu, S. Arridge, J. Keegan, Y. Guo, et al. Dagan: Deep de-aliasing generative adversarial networks for fast compressed sensing mri reconstruction. *IEEE transactions on medical imaging*, 37(6):1310–1321, 2018.

[56] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555:487 EP –, 03 2018.

[57] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487, 2018.