Compressive sensing with un-trained neural networks: Gradient descent finds the smoothest approximation

Reinhard Heckel [1] **Mahdi Soltanolkotabi** [2]

Abstract

Un-trained convolutional neural networks have emerged as highly successful tools for image recovery and restoration. They are capable of solving standard inverse problems such as denoising and compressive sensing with excellent results by simply fitting a neural network model to measurements from a single image or signal without the need for any additional training data. For some applications, this critically requires additional regularization in the form of early stopping the optimization. For signal recovery from a few measurements, however, un-trained convolutional networks have an intriguing self-regularizing property: Even though the network can perfectly fit any image, the network recovers a natural image from few measurements when trained with gradient descent until convergence. In this paper, we provide numerical evidence for this property and study it theoretically. We show that—without any further regularization—an un-trained convolutional neural network can approximately reconstruct signals and images that are sufficiently structured, from a near minimal number of random measurements.

1. Introduction

Un-trained convolutional neural networks have emerged as highly successful tools for image recovery and restoration, for a variety of problems including denoising, compressive sensing, and inpainting (Ulyanov et al., 2018; Jin et al., 2019; Veen et al., 2018; Jagatap & Hegde, 2019; Heckel, 2019; Heckel & Hand, 2019; Bostan et al., 2020; Wang et al., 2020; Hyder & Asif, 2020; Arora et al., 2020). As opposed to trained convolutional neural networks, that learn an image prior from training data, un-trained convolutional networks act as an image prior without any training and solely based on the architecture of the network and the optimization procedure used to fit them.

The benefit of untrained networks was first observed in the Deep Image Prior (DIP) paper (Ulyanov et al., 2018). The key observation of Ulyanov et al. (2018) is that fitting a standard over-parameterized convolutional autoencoder (specifically, the U-net (Ronneberger et al., 2015) or variations thereoff) to a single noisy/corrupted image, when combined with early stopping, yields excellent denoising, inpainting, and super-resolution performance. Subsequent literature has demonstrated that many elements of the architecture of a convolutional autoencoder—such as the encoder part—are irrelevant for this behavior to emerge. In particular the papers (Heckel & Hand, 2019; Heckel & Soltanolkotabi, 2020) highlight the critical role of convolutions with fixed convolutional kernels.

Un-trained convolutional networks are empirically most effective when the network is over-parametrized, meaning that is has more parameters than image pixels. This holds even though in this regime the neural network can in principle fit any image perfectly, including random noise. Therefore, further regularization is critical to performance in many applications. For instance denoising (Ulyanov et al., 2018; Heckel & Soltanolkotabi, 2020) critically requires early stopping, as without early stopping the noisy image is fitted perfectly and no noise is removed. However, perhaps surprisingly, for some inverse problems including inpainting (Ulyanov et al., 2018) and compressive sensing, no further regularization is necessary! That is, a convolutional neural network, when fitted to compressive measurements from a single image (no other training data) can estimate the original image well, as illustrated in Figure 1. This phenomenon demonstrates an intriguing self-regularization capability in the context of compressive sensing.

The overarching goal of this paper is to study compressive sensing with un-trained convolutional generators theoretically in order to explain the above phenomenon. In particular, our goal is to understand (i) why for compressive sensing problems gradient descent can reconstruct a good

[1]Dept. of Electrical and Computer Engineering, Technical University of Munich [2]Dept. of Electrical and Computer Engineering, University of Southern California. Correspondence to: Reinhard Heckel <reinhard.heckel@tum.de>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

signal estimate without any further regularization or additional training data and to (ii) prove that this is possible with a minimal number of measurements that is proportional to an appropriately defined notion of signal dimensionality.

1.1. Compressive sensing with un-trained neural networks

We consider the problem of recovering an unknown signal $\mathbf{x}^* \in \mathbb{R}^n$ from $m \ll n$ linear measurements of the form

$$\mathbf{y} = \mathbf{A}\mathbf{x}^* \in \mathbb{R}^m, \tag{1}$$

with $\mathbf{A} \in \mathbb{R}^{m \times n}$ representing the measurement matrix. This problem formulation includes the compressive sensing problem relevant for computational imaging as well as inpainting. To understand how un-trained networks can be utilized to recover the unknown signal, consider an over-parameterized, un-trained convolutional image prior $G\colon \mathbb{R}^N \to \mathbb{R}^n$ mapping an $N \gg n$ dimensional parameter vector $\mathbf{C}$ to an n dimensional signal. We take G to be the deep decoder, a simple *un-trained* convolutional network, defined formally in Section 2. We emphasize that G is an un-trained neural networks that is randomly initialized and has never seen any training data. To reconstruct the signal from its measurements we fit a compressed version of the generator output to these measurements via randomly initialized gradient descent on the loss

$$\mathcal{L}(\mathbf{C}) = \frac{1}{2}\|\mathbf{A}G(\mathbf{C}) - \mathbf{y}\|_2^2. \tag{2}$$

Let $\hat{\mathbf{C}}$ denote the solution found by gradient descent. The signal estimate can then be calculated as $\hat{\mathbf{x}} = G(\hat{\mathbf{C}})$.

A number of recent papers have shown that with the deep image prior (a convolutional autoencoder) or the deep decoder (a convolutional generator) as a prior G, this approach is rather effective (Veen et al., 2018; Jagatap & Hegde, 2019; Heckel, 2019). Most recently Arora et al. (Arora et al., 2020) have shown that this approach significantly improves upon classical compressive sensing methods (ℓ_1-regularization and total-variation norm minimization) for accelerating multi-coil magnetic resonance imaging, which is arguably one of the most prominent real-world application of compressive sensing.

The generator G is over-parameterized and can express any image $\mathbf{x}^*$, including unstructured noise. Nevertheless, typically no further regularization in the form of early stopping the optimization is necessary. We demonstrate this phenomenon in Figure 1. This figure shows that running gradient descent on the loss $\mathcal{L}(\mathbf{C})$ eventually yields an estimate that is very close to the original image. This is surprising because i) there is no additional training data and ii) even though the generator G can fit any image, including noise, gradient descent still finds an image close to the original one.

1.2. Contributions

The main contribution of this paper is to show that un-trained convolutional image priors provably enable recovery of natural images from a few random linear measurements. This holds by simply running gradient descent until convergence—without any further regularization. More specifically, we show that fitting an over-parameterized convolutional network with fixed convolutions (via gradient descent) to random measurements of a smooth signal essentially recovers that signal. Furthermore, the required number of measurements is commensurate to how smooth the signal is with more measurements required when the signal has "high-frequency" components. In more detail:

- Suppose we have m-linear measurements $\mathbf{y} = \mathbf{A}\mathbf{x}^*, \mathbf{A} \in \mathbb{R}^{m \times n}$ of an unknown signal $\mathbf{x}^*$ with $\mathbf{A}$ a Gaussian measurement matrix. Furthermore, assume that the signal $\mathbf{x}^*$ is p-smooth, in the sense that it can be represented as a linear combination of the p lowest frequency orthonormal trigonometric basis functions $\mathbf{w}_1, \dots, \mathbf{w}_n \in \mathbb{R}^n$ as

  $$\mathbf{x}^* = \sum_{i=1}^{p} \mathbf{w}_i \langle \mathbf{w}_i, \mathbf{x}^* \rangle .$$

  We plot these trigonometric basis functions in Figure 2 and formally define them later on in Section 4. Note that the smaller p, the smoother the signal $\mathbf{x}^*$ is, thus p is a measure of smoothness.

  Our main result shows that the estimate $\mathbf{C}_\infty$, obtained by running gradient descent on the loss (2) until convergence, yields an output $G(\mathbf{C}_\infty)$ which is very close to $\mathbf{x}^*$, i.e., $G(\mathbf{C}_\infty) \approx \mathbf{x}^*$. This holds as soon as the number of measurements exceeds the degrees of smoothness present in the signal (p). Since natural images are approximately smooth, this results provides a theoretical explanation why compressive sensing on natural images with over-parameterized convolutional generators works so well (see (Veen et al., 2018; Jagatap & Hegde, 2019; Heckel, 2019; Arora et al., 2020) for corresponding empirical results).

- In a nutshell, our main insight is that the behavior of large over-parameterized neural networks is dictated by the spectral properties of their Jacobian mapping. For the convolutional generators considered in this paper, the associated Jacobian matrix has singular vectors that can be well approximated by the orthonormal trigonometric basis function and singular values that decay very quickly from the low-frequency to the high-frequency trigonometric basis functions. Specifically, the associated singular values decay approximately geometrically.

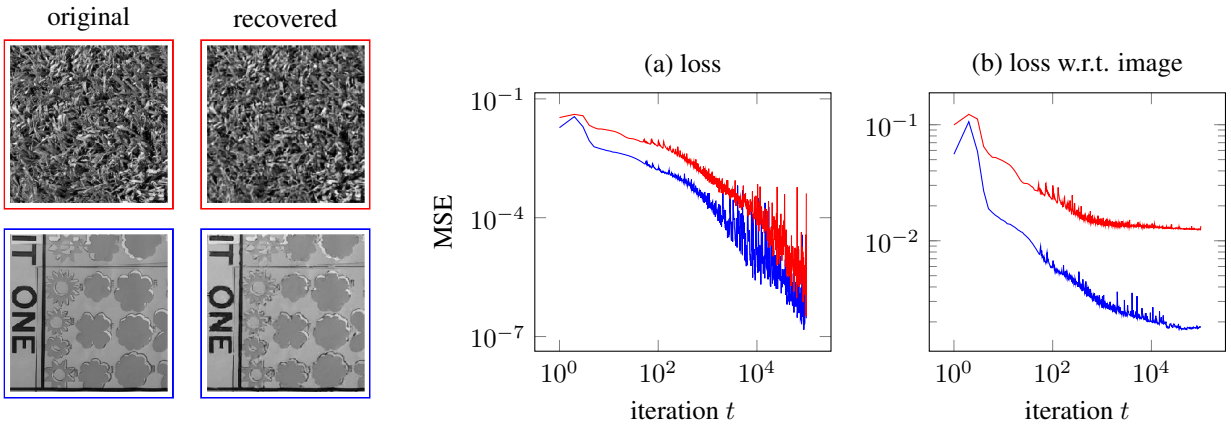


Figure 1. Compressive sensing of two different images $\mathbf{x}^*$ displayed on the right with a random matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, $m = n/4$, from the measurement $\mathbf{y} = \mathbf{A}\mathbf{x}^*$. Panel (a) shows the loss at iteration t, i.e., $\frac{1}{2}\|\mathbf{A}G(\mathbf{C}_t) - \mathbf{y}\|_2^2$, and panel (b) is the loss with respect to the original image, i.e., $\|G(\mathbf{C}_t) - \mathbf{x}^*\|_2^2$. Here, G is a 5-layer deep decoder (Heckel & Hand, 2019); a convolutional network with fixed convolutional filters. The figure looks qualitatively the same if we take G as the deep image prior (Ulyanov et al., 2018), a U-net like convolutional autoencoder. It can be seen that early stopping is not required: gradient descent converges to a good solution, and early stopping does not improve performance for this example. Moreover, the simple and smooth image (blue) achieves a smaller loss with the same number of measurements than the non-smooth grass texture (red). Both features are captured by our theory.

To prove our result, we first characterize the least-squares solution of a randomly sketched least-squares problem with a design matrix with a decaying spectrum. To prove the result for convolutional generators we show that this non-linear learning problem behaves like an associated linear model with the above spectral characteristics. We then conclude the proof for the corresponding convolutional generator, by showing that the solutions obtained by running gradient descent on the non-linear problem is close to that obtained by running gradient descent on the linear problem.

- In order to develop a better understanding of compressive sensing with untrained priors, we also carry out compressive sensing experiments for accelerating magnetic resonance imaging (MRI). Our experiments corroborate our theoretical finding that simply iterating until convergence is effective. This also suggests that there is little or no benefit to additional regularization.

Our paper is organized as follows: We start by stating the convolutional architecture considered in this paper in Section 2. In Section 3 we study the reconstruction of a signal from few a measurements with a *linear* over-parameterized generator to form intuition. In Section 4 we state our main results for signal recovery with convolutional generators. Section 5 contains our numerical result for MRI imaging. We conclude the paper with related work and a brief proof sketch, all formal proofs are deferred to the Appendix.

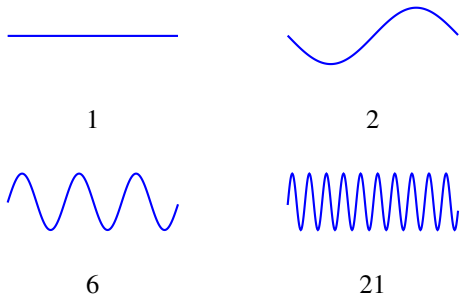


Figure 2. The 1st, 2nd, 6th, and 21st trigonometric basis functions in dimension $n = 300$.

2. Convolutional generators

A convolutional generator generates an image through convolutional operations and applications of non-linearities. In this paper, we study a two-layer convolutional generator $G\colon \mathbb{R}^{nk \times n} \to \mathbb{R}^n$ theoretically. The generator has the form

$$G(\mathbf{C}) = \mathrm{ReLU}(\mathbf{U}\mathbf{C})\mathbf{v}. \tag{3}$$

Here, $\mathbf{v} = [1, \ldots, 1, -1, \ldots, -1]/\sqrt{k}$ are the fixed weights of the output layer, of which half are positive and the other half are negative, and $\mathbf{C} \in \mathbb{R}^{n \times k}$ is the coefficient matrix of the generator, corresponding to the weights in the first layer of the network. Critical for the performance of the generator is the convolutional operation with a fixed kernel $\mathbf{u}$, implemented through multiplication with the circulant matrix $\mathbf{U} \in \mathbb{R}^{n \times n}$.

This architecture is a two-dimensional version of the deep decoder (Heckel & Hand, 2019). The deep decoder in turn

is a sub-set of the deep image prior (Ulyanov et al., 2018) and the U-net (Ronneberger et al., 2015), as commented on below.

The deep decoder with d layers (typically, $d = 4, 5, 6$) is defined as

$$G(\mathbf{C}) = \mathrm{ReLU}(\mathbf{U}\mathbf{B}_d\mathbf{C}_d)\mathbf{v}, \tag{4}$$

where

$$\mathbf{B}_{i+1} = \mathrm{cn}(\mathrm{ReLU}(\mathbf{U}_i\mathbf{B}_i\mathbf{C}_i)), i = 0, \ldots, d-1.$$

Here $\mathrm{cn}(\cdot)$ is a channel normalization operation, which normalizes each channel/column of the volume/matrix $\mathrm{ReLU}(\mathbf{U}_i\mathbf{B}_i\mathbf{C}_i) \in \mathbb{R}^{n_i \times k}$ individually and can be viewed as a special case of the batch normalization operation. Note that if the signal to be generated is an image and thus two-dimensional ($n_i \in \mathbb{Z}^2$), then $\mathbf{B}_i$ is a three-dimensional tensor consisting of k many channels, and if the signal is one-dimensional ($n_i \in \mathbb{Z}$), those tensors are two-dimensional and can be viewed as matrices consisting of k many columns (or channels). Moreover, $\mathbf{B}_0$ is a fixed input tensor, which we assume to have full row rank. The parameters of the deep decoder are the weight matrices $\mathbf{C}_1, \ldots, \mathbf{C}_d \in \mathbb{R}^{k \times k}$. Multiplication with those weight matrices is performing linear combinations of the channels, which in turn is equivalent to performing 1x1-convolutions.

For $d = 2$, the deep decoder reduces to the two-dimensional version in (3). To see this, note that for $d = 2$, because $\mathbf{B}_0$ has full column rank, optimizing over $\mathbf{B}_0\mathbf{C}_0 \in \mathbb{R}^{n \times k}$ is equivalent to optimizing over $\mathbf{C} \in \mathbb{R}^{n \times k}$ instead.

Finally, as mentioned before, the deep decoder can be viewed as the relevant part of a convolutional generator to function as an image prior. It can be deduced from a convolutional autoencoder (such as the deep image prior (Ulyanov et al., 2018) and the U-net (Ronneberger et al., 2015)) by removing the encoder part, any skip connections, and most surprisingly, the trainable convolutional filters of spatial extent larger than one. As demonstrated in (Heckel & Soltanolkotabi, 2020), the critical aspect for an un-trained deep image prior are the convolutions with fixed convolutional kernels, implemented here by the operator $\mathbf{U}$.

3. Signal recovery with over-parameterized linear generators

Consider an over-parameterized linear generator $\tilde{G}(\mathbf{c}) = \mathbf{J}\mathbf{c}$ defined by a wide, full-rank, generator matrix $\mathbf{J} \in \mathbb{R}^{n \times N}, N \geq n$, and an arbitrary and unknown signal $\mathbf{x}^* \in \mathbb{R}^n$. Because $\mathbf{J}$ has full rank, the signal can be expressed as $\mathbf{x}^* = \mathbf{J}\mathbf{c}^*$. However, the coefficient vector $\mathbf{c}^*$ in this representation is non-unique, as $\mathbf{J}$ is a wide matrix containing more columns than rows. We observe m linear measurements of the unknown signal of the form

$$\mathbf{y} = \mathbf{A}\mathbf{x}^*,$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a wide ($m < n$) Gaussian measurement matrix, with iid $\mathcal{N}(0, 1/m)$ entries. We note that with this variance, norms are approximately preserved (i.e., for a fixed $\mathbf{z}$, with high probability $\|\mathbf{z}\|_2 \approx \|\mathbf{A}\mathbf{z}\|_2$).

Our goal is to estimate the signal $\mathbf{x}^*$ based on the measurement $\mathbf{y}$. We estimate the signal $\mathbf{x}^*$ by first computing a coefficient estimate $\hat{\mathbf{c}}$ by minimizing the loss

$$\mathcal{L}(\mathbf{c}) = \frac{1}{2}\|\mathbf{A}\mathbf{J}\mathbf{c} - \mathbf{y}\|_2^2,$$

via running gradient descent with sufficiently small step size until convergence. We then estimate the signal via $\hat{\mathbf{x}} = \mathbf{J}\hat{\mathbf{c}}$. Since gradient descent applied on a least-squares problem yields the minimum-norm solution, the estimate $\hat{\mathbf{c}}$ can equivalently be expressed as

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \|\mathbf{c}\|_2^2 \text{ subject to } \mathbf{A}\mathbf{J}\mathbf{c} = \mathbf{y}. \tag{5}$$

In closed form, $\hat{\mathbf{c}}$ is given as

$$\hat{\mathbf{c}} = (\mathbf{A}\mathbf{J})^\dagger \mathbf{A}\mathbf{J}\mathbf{c}^* = \mathbf{P}_{\mathbf{J}^T\mathbf{A}^T}\mathbf{c}^*,$$

where $(\mathbf{A}\mathbf{J})^\dagger$ is the pseudo-inverse of $\mathbf{A}\mathbf{J}$, and $\mathbf{P}_{\mathbf{J}^T\mathbf{A}^T}$ is a orthogonal projection operator onto the range of $(\mathbf{A}\mathbf{J})^T$. Thus, the signal estimation error is

$$\hat{\mathbf{x}} - \mathbf{x}^* = \mathbf{J}(\hat{\mathbf{c}} - \mathbf{c}^*) = \mathbf{J}(\mathbf{I} - \mathbf{P}_{\mathbf{J}^T\mathbf{A}^T})\mathbf{c}^*. \tag{6}$$

The following theorem characterizes this signal estimation error.

Theorem 1. *Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be a random Gaussian matrix with $m \geq 12$, and let $\mathbf{w}_1, \ldots, \mathbf{w}_n$ be the left singular vectors of $\mathbf{J}$ with associated singular values $\sigma_1 \geq \ldots \geq \sigma_n$. Then, for any $\mathbf{x}^* \in \mathbb{R}^n$, with probability at least $1 - 3e^{-1/2m}$, the signal estimate $\hat{\mathbf{x}} = \mathbf{J}\hat{\mathbf{c}}$ based on the measurement $\mathbf{y} = \mathbf{A}\mathbf{x}^*$, with the coefficient estimate $\hat{\mathbf{c}}(\mathbf{y})$ defined in (5), obeys*

$$\|\hat{\mathbf{x}} - \mathbf{x}^*\|_2^2 \leq C\left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\left\langle \mathbf{w}_i, \mathbf{x}^* \right\rangle^2\right) \sum_{i > 2m/3} \sigma_i^2. \tag{7}$$

Here, C is a fixed numerical constant.

The proof, given in the appendix, relies on arguments from (Halko et al., 2011, Sec. 8 and Sec. 9) developed for approximating low-rank matrices through random sampling.

The theorem guarantees that the error in estimating the signal $\mathbf{x}^*$ from compressive measurements $\mathbf{y} = \mathbf{A}\mathbf{x}^*$ is small provided that two conditions are satisfied:

(i) The signal $\mathbf{x}^*$ lies (approximately) in the span of the leading $O(m)$ singular vectors of $\mathbf{J}$, where m is the number of linear measurements.

(ii) The singular values of the generator matrix $\mathbf{J}$ decay sufficiently fast (for example geometrically).

To see this, let us consider a concrete example. Suppose the singular values decay geometrically, i.e., $\sigma_i^2 = \gamma^i$ for some $\gamma \in (0,1)$. Moreover, suppose that the signal $\mathbf{x}^*$ lies in the span of the leading $m/3$ singular values of $\mathbf{J}$, i.e., $\mathbf{x}^* \in \text{span}(\mathbf{w}_1, \dots, \mathbf{w}_{m/3})$. Then, Theorem 1 guarantees that the estimate $\hat{\mathbf{x}}$ based on m random linear measurements obeys

$$\|\hat{\mathbf{x}} - \mathbf{x}^*\|_2^2 \leq C \frac{\gamma^{m/3}}{1-\gamma} \|\mathbf{x}^*\|_2^2. \tag{8}$$

Here, we used that the first term in the right-hand-side of (1) is bounded by $1/\sigma_{m/3}^2 \|\mathbf{x}^*\|_2^2$, using that $\mathbf{x}^*$ is in the span of the leading singular vectors, and that $\sum_{i>2m/3} \sigma_i^2 \leq \frac{\gamma^{2m/3}}{1-\gamma}$, by the formula for a geometric series. The bound (8) is very small provided that γ is slightly below one (since $\gamma^{m/3}$ decays exponentially)—thus guaranteeing almost perfect recovery of a signal that is aligned with the leading singular vectors of $\mathbf{J}$.

4. Main results for compressive sensing with convolutional generators

We are now ready to state our main results for compressive sensing with convolutional generators. We consider the non-linear least-squares objective

$$\mathcal{L}(\mathbf{C}) = \frac{1}{2} \|\mathbf{A}G(\mathbf{C}) - \mathbf{y}\|_2^2,$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}, m \leq n$, is a Gaussian random matrix with iid $\mathcal{N}(0, 1/m)$ entries and $G(\mathbf{C})$ is the two-layer decoder network defined in section 2. We minimize this objective by running gradient descent with a constant step-size η, starting from a random initialization $\mathbf{C}_0$, with entries drawn iid from a Gaussian distribution $\mathcal{N}(0, \omega^2)$, and with variance ω^2 specified later. The coefficients at iterations $t = 1, 2, \dots$ are given by

$$\mathbf{C}_{t+1} = \mathbf{C}_t - \eta \nabla \mathcal{L}(\mathbf{C}_t). \tag{9}$$

In the previous section we studied a linear generator with generator matrix $\mathbf{J}$ with quickly decaying spectrum. In this section we extend the insights from the previous section to the non-linear case by replacing the role of the generator matrix $\mathbf{J}$ with the Jacobian of the non-linear generator G, defined as $[\mathcal{J}(\mathbf{C})]_{ij} = \frac{\partial}{\partial c_i}[G(\mathbf{C})]_j$. In contrast to the linear case, however, the Jacobian changes across iterations of gradient descent. Nevertheless, we can account for these changes in the Jacobian in our analysis.

As found in (Heckel & Soltanolkotabi, 2020), for the two-layer deep decoder that we consider, the left singular vectors of the Jacobian can be well approximated by the trigonometric basis function $\mathbf{w}_1, \dots, \mathbf{w}_n \in \mathbb{R}^n$ plotted in Figure 2, and defined as

$$[\mathbf{w}_i]_j = \frac{1}{\sqrt{n}} \begin{cases} 1 & i = 0 \\ \sqrt{2}\cos(2\pi ji/n) & i = 1, \dots, n/2-1 \\ (-1)^j & i = n/2 \\ \sqrt{2}\sin(2\pi ji/n) & i = n/2+1, \dots, n-1 \end{cases}. \tag{10}$$

Moreover, the singular values of the Jacobian throughout the iterates can be well approximated by associated values that only depend on the convolution kernel $\mathbf{u}$ associated with the convolution operator $\mathbf{U}$. Those values $\boldsymbol{\sigma} \in \mathbb{R}^n$ are given by

$$\boldsymbol{\sigma} = \|\mathbf{u}\|_2 \sqrt{\left| \mathbf{F} g\left(\frac{\mathbf{u} \circledast \mathbf{u}}{\|\mathbf{u}\|_2^2} \right) \right|} \tag{11}$$

with

$$g(z) = \frac{1}{2}\left(1 - \frac{\cos^{-1}(z)}{\pi}\right) z.$$

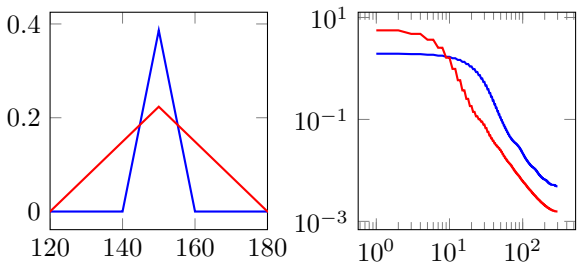
Here, for two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, $\mathbf{u} \circledast \mathbf{v}$ denotes their circular convolution, $\mathbf{F}$ is the discrete Fourier transform matrix, and the scalar non-linearity g is applied entrywise. As a concrete relevant example, in Figure 3 we depict the triangular kernel that is used in the original deep decoder network. The most important observation from this plot is that the associated weights $\boldsymbol{\sigma} = [\sigma_1, \dots, \sigma_n]$ decay very fast, namely geometrically.

With those definition, we are now ready to state our main result.

Theorem 2. *Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be a random Gaussian matrix with $m \geq 12$ and suppose we are given a linear measurement $\mathbf{y} = \mathbf{A}\mathbf{x}^*$ of an arbitrary signal $\mathbf{x}^* \in \mathbb{R}^n$. Consider a two layer generator network $G(\mathbf{C}) = \text{ReLU}(\mathbf{UC})\mathbf{v}$, $\mathbf{C} \in \mathbb{R}^{n \times k}$, with*

$$k \geq C_{\mathbf{u}} \frac{m}{\xi^8}, \tag{12}$$

channels and with convolutional kernel $\mathbf{u}$ of the convolutional operator $\mathbf{U}$ and associated weights $\boldsymbol{\sigma} = [\sigma_1, \dots, \sigma_n]$. Here, $\xi \leq 1$ is arbitrary and $C_{\mathbf{u}}$ is a constant that only depends on the convolutional kernel $\mathbf{u}$. In order to estimate the signal, we fit the convolutional generator to the signal by running gradient descent starting from a random initialization $\mathbf{C}_0$ with i.i.d. $\mathcal{N}(0, \omega^2)$, entries, $\omega \propto \frac{\|\mathbf{y}\|_2}{\sqrt{n}}$, and



(a) triangular kernel (b) associated weights

Figure 3. Triangular kernels and the weights associated to low-frequency trigonometric functions they induce, for a generator network of output dimension $n = 300$. The wider the kernel is, the more the weights are concentrated towards the low-frequency components of the signal. Note that the lower singular values decay geometrically (as evident from the straight line in the log-log plot)—as the singular values in our example in Section 3.

sufficiently small stepsize to the loss $\frac{1}{2}\|\mathbf{A}G(\mathbf{C}) - \mathbf{y}\|_2^2$ *until convergence. Then, with high probability, the reconstruction error with parameters* $\mathbf{C}_\infty$ *at convergence obeys*

$$\|G(\mathbf{C}_\infty) - \mathbf{x}^*\|_2^2 \leq C\left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\left\langle \mathbf{w}_i, \mathbf{x}^*\right\rangle^2\right) \sum_{i>2m/3} \sigma_i^2 + \xi^2\|\mathbf{x}^*\|_2^2. \tag{13}$$

Here, C is a fixed numerical constant.

Theorem 2 establishes that a convolutional generator enables the reconstruction of a natural signal from a few linear measurements. To see this, note that a good model for a natural image is a smooth signal, i.e., a signal that can be well-approximated by few leading trigonometric basis functions. More concretely, Figure 4 in (Simoncelli & Olshausen, 2001) shows that the power spectrum of a natural image (i.e., the energy distribution by frequency) decays rapidly from low frequencies to high frequencies.

Thus it is reasonably to assume that the signal $\mathbf{x}^*$ can be represented with few of the trigonometric basis function; for concreteness say that $\mathbf{x}^*$ lies in the span of $\mathbf{w}_1, \ldots, \mathbf{w}_{m/3}$. Next, recall from Figure 3 that the weights associated with a triangular kernel decay geometrically (i.e., $\sigma_i^2 = \gamma^i$ for some $\gamma \in (0, 1)$). Thus, from the same argument as used for (8), the bound (13) established by the theorem yields that the reconstruction error is bounded by

$$\|G(\mathbf{C}_\infty) - \mathbf{x}^*\|_2^2 \leq C\frac{\gamma^{m/3}}{1-\gamma}\|\mathbf{x}^*\|_2^2 + \xi^2\|\mathbf{x}^*\|_2^2.$$

Thus our theorem guarantees the recovery of a sufficiently smooth signal by optimizing over the range of the generator. In particular if the signal is p-smooth, i.e., lies in the span of $\mathbf{w}_1, \ldots, \mathbf{w}_p$, then $O(p)$ measurements are sufficient to provide an accurate estimate.

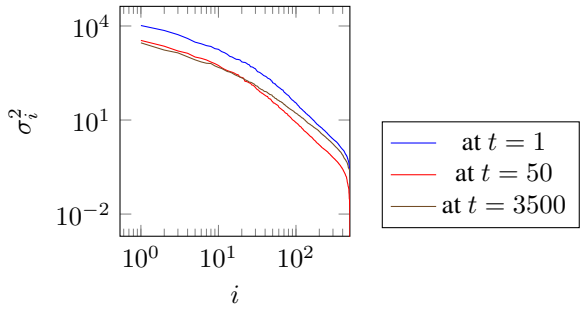


Figure 4. The singular value distribution of the Jacobian of a four-layer deep decoder at different iterations of gradient descent; the spectrum changes only slightly, and the singular values decay slightly faster than geometrically.

4.1. Beyond two layer networks

Our main theorem from the previous section relies on two critical ingredients:

(i) The finding from (Heckel & Soltanolkotabi, 2020) that the leading singular vectors of the Jacobian of a two-layer deep decoder are approximately the trigonometric basis function throughout all iterations of gradient descent.

(ii) The weights $\sigma_1, \ldots, \sigma_n$ associated with the trigonometric basis functions decaying sufficiently fast, specifically approximately geometric. That is required for gradient descent applied to fitting m compressive measurements until convergence to (approximately) only fit the signal to the leading $O(m)$ trigonometric basis functions.

Those results extend to deeper networks as follows. First, as shown numerically in (Heckel & Soltanolkotabi, 2020), the leading singular vectors of the Jacobian of a four-layer deep decoder are also close to the trigonometric basis functions, and change only little across iterations. Second, as shown in Figure 4, the singular values of a four-layer deep decoder also decay (at least) geometrically, and the spectrum changes only little across iterations. Thus, the implications of our theory continue to apply for deeper deep decoders.

5. Numerical experiments for magnetic resonance imaging

In the final part of our paper we consider accelerating magnetic resonance imaging (MRI), one of the major application of compressive sensing. MRI is a medical imaging technique where measurements of an object can only be taken in the Fourier domain, referred to as k-space. If the full k-space measurement is collected, an image of the object can be computed almost perfectly (up the noise inherent in the measurement process). In order to accelerate the imag-

ing process, it is common to only collect a small part of the k-space, which corresponds to taking few linear Fourier measurements; or in the notation of our paper, a measurement matrix $\mathbf{A}$ with subsampled rows of the Fourier matrix.

In order to understand whether our main finding—that signal reconstruction from compressive measurements without further regularization is possible—applies in practice, we consider the problem of reconstructing an image from few k-space measurements. We consider reconstruction of an image from 8-fold undersampled k-space measurements from the fastMRI dataset, recently released by facebook and NYU (Zbontar et al., 2018). We reconstruct with a $d = 5$ layer and highly over-parameterized deep decoder. Figure 5 shows the corresponding loss curves. It can be seen that early stopping at the optimal early stopping point gives only marginally better performance than when optimizing until convergence, and in addition the optimal early stopping point is unknown in practice (because we do not have access to a reconstruction from a full measurement).

6. Related literature

In this paper we focus on un-trained neural network for solving inverse problems. In contrast a large body of recent result concentrates on using trained deep convolutional neural networks for image recovery and reconstruction. Training based deep learning methods for solving inverse problems are either trained end-to-end for tasks like denoising (Burger et al., 2012; Zhang et al., 2017), or are based on learning a generative image model (by training an autoencoder or GAN (Hinton & Salakhutdinov, 2006; Goodfellow et al., 2014)) and then using the resulting image models to regularize problems such as compressed sensing (Bora et al., 2017; Hand & Voroninski, 2018; Huang et al., 2018), denoising (Heckel et al., 2020), or phase retrieval (Hand et al., 2018; Shamshad & Ahmed, 2018). In contrast to un-trained network, where optimization is over the weights of the untrained generator, in the aformentioned papers it is over the input of the (trained) network.

Our proof relies on relating the dynamics of gradient descent on an over-parameterized network to that of gradient descent on an associated linear network. This proof technique has been used in a variety of recent publication (Soltanolkotabi et al., 2018; Venturi et al., 2019; Du et al., 2018; Oymak & Soltanolkotabi, 2019a;b; Arora et al., 2019; Oymak et al., 2019; Basri et al., 2019; Li et al., 2019). Most related to our work is the recent paper (Heckel & Soltanolkotabi, 2020) that shows that the deep decoder enables denoising. Neither of the publications, however, addresses compressive sensing or reconstruction from randomly sketched data, and most of our technical results are specific to this setup.

Finally note that regularizing *linear* models with gradient descent via early stopping has a rich history in the signal processing community. In the 50s, Landweber proposed to recover a signal from linear measurements via gradient descent (Landweber, 1951) which became known as the Landweber algorithm in the inverse problems community. Subsequent work in this literature proposed to early-stop the Landweber iterations (i.e., gradient descent) in order to regularize ill-posed inverse problems (Trussell & Civanlar, 1985).

7. Proof sketch

In this section we provide a sketch of our argument. Our statement and formal proof pertains to the two-layer case, in this section we provide the sketch for the general case where $G(\boldsymbol{\theta})$ is a generic network with a N-dimensional parameter vector $\boldsymbol{\theta}$, and then comment on how this general proof strategy is particularized to the two layer case.

Given a measurement $\mathbf{y}$, we characterize the solution of running gradient descent with fixed step size η on the nonlinear least-squares objective

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\|f(\boldsymbol{\theta}) - \mathbf{y}\|_2^2, \quad f(\boldsymbol{\theta}) = \mathbf{A}G(\boldsymbol{\theta}),$$

starting from an initial point $\boldsymbol{\theta}_0$. The updates take the form

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\nabla\mathcal{L}(\boldsymbol{\theta}_t), \quad \nabla\mathcal{L}(\boldsymbol{\theta}) = \mathcal{J}^T(\boldsymbol{\theta})(f(\boldsymbol{\theta}) - \mathbf{y}), \tag{14}$$

where $\mathcal{J}(\boldsymbol{\theta})$ is the Jacobian of f at $\boldsymbol{\theta}$. We start gradient descent from a random initialization $\boldsymbol{\theta}_0$ with iid $\mathcal{N}(0, \omega)$ entries. Central to our analysis are the following objects. Let $\mathcal{J}_G(\boldsymbol{\theta}) \in \mathbb{R}^{n\times N}$ be the Jacobian of $G(\boldsymbol{\theta})$ and define $\mathbf{J}_G$ as a reference generator Jacobian that we set to a matrix that is very close to the generator Jacobian at initialization, i.e., $\mathbf{J}_G \approx \mathcal{J}_G(\boldsymbol{\theta}_0)$. For the two-layer network for which we state a precise result, this matrix only depends on the convolutional operator $\mathbf{U}$.

Relevant for the dynamics of gradient descent, however, are the corresponding sketched original and reference Jacobians, defined as

$$\mathcal{J}(\boldsymbol{\theta}) = \mathbf{A}\mathcal{J}_G(\boldsymbol{\theta}) \in \mathbb{R}^{m\times N} \quad \text{and} \quad \mathbf{J} = \mathbf{A}\mathbf{J}_G \in \mathbb{R}^{m\times N}.$$

Since we chose $\mathbf{J}_G \approx \mathcal{J}_G(\boldsymbol{\theta}_0)$, we also have $\mathbf{J} \approx \mathcal{J}(\boldsymbol{\theta}_0)$.

7.1. Closeness to an associated linear problem

To characterized the behavior of the gradient descent updates in (22), we relate the non-linear least squares problem to a linearized one in a ball around the initialization $\boldsymbol{\theta}_0$. This general strategy has been utilized in a number of recent publications (Soltanolkotabi et al., 2018; Du et al., 2018; Heckel & Soltanolkotabi, 2020; Arora et al., 2019; Oymak & Soltanolkotabi, 2019b; Oymak et al., 2019).

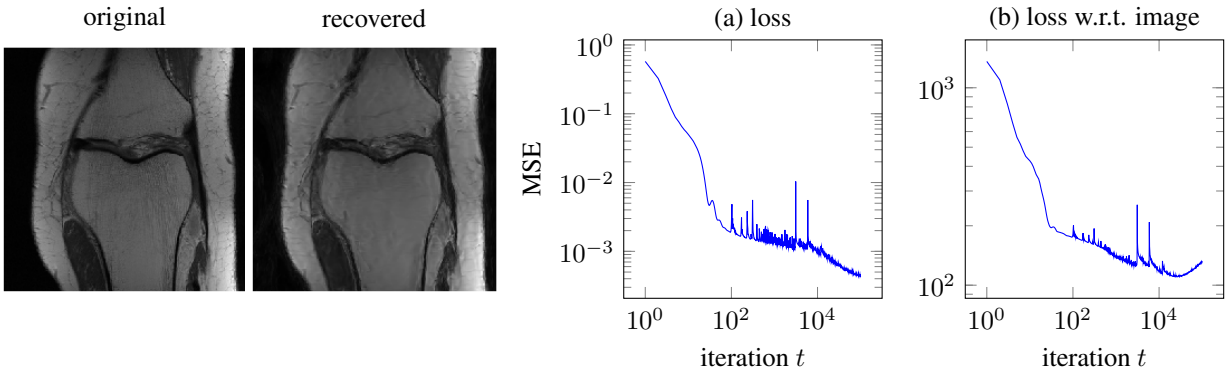


Figure 5. Compressive sensing MRI: MSE of reconstructing an image from 8-fold undersampled k-space MRI measurements. While early stopping is not absolutely necessary, stopping at about 2000 iterations slightly improves performance relative to optimizing until convergence.

We define the associated linearized least-squares problem as

$$\mathcal{L}_{\text{lin}}(\boldsymbol{\theta}) = \frac{1}{2}\|f(\boldsymbol{\theta}_0) + \mathbf{J}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \mathbf{y}\|_2^2. \qquad (15)$$

Code

Code to reproduce the experiments is available at `https://github.com/MLI-lab/cs_deep_decoder`.

Acknowledgements

R. Heckel is partially supported by NSF award IIS-1816986, acknowledges support of the NVIDIA Corporation in form of a GPU, and would like to thank Tobit Klug for proofreading a previous version of this manuscript. M. Soltanolkotabi is supported by the Packard Fellowship in Science and Engineering, a Sloan Research Fellowship in Mathematics, an NSF-CAREER under award #1846369, the Air Force Office of Scientific Research Young Investigator Program (AFOSR-YIP) under award #FA9550-18-1-0078, an NSF-CIF award #1813877, DARPA under the Learning with Less Labels (LwLL) and Fast Network Interface Cards (FastNICs) program, and a Google faculty research award.

References

Arora, S., Du, S. S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, 2019.

Arora, S., Roeloffs, V., and Lustig, M. Untrained modified deep decoder for joint denoising parallel imaging reconstruction. In *International Society for Magnetic Resonance in Medicine Annual Meeting*, 2020.

Basri, R., Jacobs, D., Kasten, Y., and Kritchman, S. The convergence rate of neural networks for learned functions of different frequencies. In *Advances in Neural Information Processing Systems*, 2019.

Bora, A., Jalal, A., Price, E., and Dimakis, A. G. Compressed sensing using generative models. In *International Conference on Machine Learning*, 2017.

Bostan, E., Heckel, R., Chen, M., Kellman, M., and Waller, L. Deep phase decoder: Self-calibrating phase microscopy with an untrained deep neural network. *Optica*, 2020.

Burger, H. C., Schuler, C. J., and Harmeling, S. Image denoising: Can plain neural networks compete with BM3d? In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2392–2399, 2012.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680. 2014.

Halko, N., Martinsson, P. G., and Tropp, J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, Jan 2011.

Hand, P. and Voroninski, V. Global guarantees for enforcing deep generative priors by empirical risk. In *Conference on Learning Theory*, 2018.

Hand, P., Leong, O., and Voroninski, V. Phase retrieval under a generative prior. In *Advances in Neural Information Processing Systems*, 2018.

Heckel, R. Regularizing linear inverse problems with convolutional neural networks. *arXiv:1907.03100*, 2019.

Heckel, R. and Hand, P. Deep decoder: Concise image representations from untrained non-convolutional networks. In *International Conference on Learning Representations*, 2019.

Heckel, R. and Soltanolkotabi, M. Denoising and regularization via exploiting the structural bias of convolutional generators. In *International Conference on Learning Representations*, 2020.

Heckel, R., Huang, W., Hand, P., and Voroninski, V. Deep denoising: Rate-optimal recovery of structured signals with a deep prior. *Information and Inference: A Journal of the IMA*, 2020.

Hinton, G. E. and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science*, 313 (5786):504–507, 2006.

Huang, W., Hand, P., Heckel, R., and Voroninski, V. A provably convergent scheme for compressive sensing under random generative priors. *arXiv:1812.04176 [math]*, 2018.

Hyder, R. and Asif, M. S. Generative models for low-dimensional video representation and reconstruction. *IEEE Transactions on Signal Processing*, 68:1688–1701, 2020.

Jagatap, G. and Hegde, C. Algorithmic guarantees for inverse imaging with untrained network priors. In *Advances in Neural Information Processing Systems*, 2019.

Jin, K. H., Gupta, H., Yerly, J., Stuber, M., and Unser, M. Time-dependent deep image prior for dynamic mri. *arXiv:1910.01684 [cs, eess]*, 2019.

Landweber, L. An iteration formula for fredholm integral equations of the first kind. *American Journal of Mathematics*, 73(3):615–624, 1951.

Li, M., Soltanolkotabi, M., and Oymak, S. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. *arXiv:1903.11680*, 2019.

Oymak, S. and Soltanolkotabi, M. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, 2019a.

Oymak, S. and Soltanolkotabi, M. Towards moderate overparameterization: Global convergence guarantees for training shallow neural networks. *arXiv:1902.04674*, 2019b.

Oymak, S., Fabian, Z., Li, M., and Soltanolkotabi, M. Generalization guarantees for neural networks via harnessing the low-rank structure of the jacobian. *arXiv:1906.05392*, 2019.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. Lecture Notes in Computer Science, 2015.

Shamshad, F. and Ahmed, A. Robust compressive phase retrieval via deep generative priors. *arXiv preprint arXiv:1808.05854*, 2018.

Simoncelli, E. P. and Olshausen, B. A. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216, 2001.

Soltanolkotabi, M., Javanmard, A., and Lee, J. D. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 2018.

Trussell, H. and Civanlar, M. The Landweber iteration and projection onto convex sets. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(6):1632–1634, 1985.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. Deep image prior. In *Conference on Computer Vision and Pattern Recognition*, 2018.

Veen, D. V., Jalal, A., Soltanolkotabi, M., Price, E., Vishwanath, S., and Dimakis, A. G. Compressed sensing with deep image prior and learned regularization. *arXiv:1806.06438*, 2018.

Venturi, L., Bandeira, A., and Bruna, J. Spurious valleys in two-layer neural network optimization landscapes. *Journal on Machine Learning Research*, 2019.

Vershynin, R. *Introduction to the non-asymptotic analysis of random matrices*, pp. 210–268. Cambridge University Press, 2012.

Wang, F., Bian, Y., Wang, H., Lyu, M., Pedrini, G., Osten, W., Barbastathis, G., and Situ, G. Phase imaging with an untrained neural network. *Light: Science & Applications*, 9(1):1–7, 2020.

Zbontar, J., Knoll, F., Sriram, A., Muckley, M. J., Bruno, M., Defazio, A., Parente, M., Geras, K. J., Katsnelson, J., Chandarana, H., Zhang, Z., Drozdzal, M., Romero, A., Rabbat, M., Vincent, P., Pinkerton, J., Wang, D., Yakubova, N., Owens, E., Zitnick, C. L., Recht, M. P., Sodickson, D. K., and Lui, Y. W. fastMRI: An open dataset and benchmarks for accelerated MRI. *arXiv:1811.08839*, 2018.

Zhang, K., Zuo, W., Chen, Y., Meng, D., and Zhang, L. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.