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

Reinhard Heckel* and Mahdi Soltanolkotabi†

*Dept. of Electrical and Computer Engineering, Technical University of Munich
†Dept. of Electrical and Computer Engineering, University of Southern California

May 7, 2020

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 [Uly+18; Jin+19; Vee+18; JH19; Hec19; HH19; Bos+20; Wan+20; HA20; Aro+20]. 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 [Uly+18]. The key observation of Ulyanov et al. [Uly+18] is that fitting a standard over-parameterized convolutional autoencoder (specifically, the U-net [Ron+15] 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 [HH19; HS20] 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 [Uly+18; HS20] 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 [Uly+18] and compressive sensing, no further regularization is necessary!

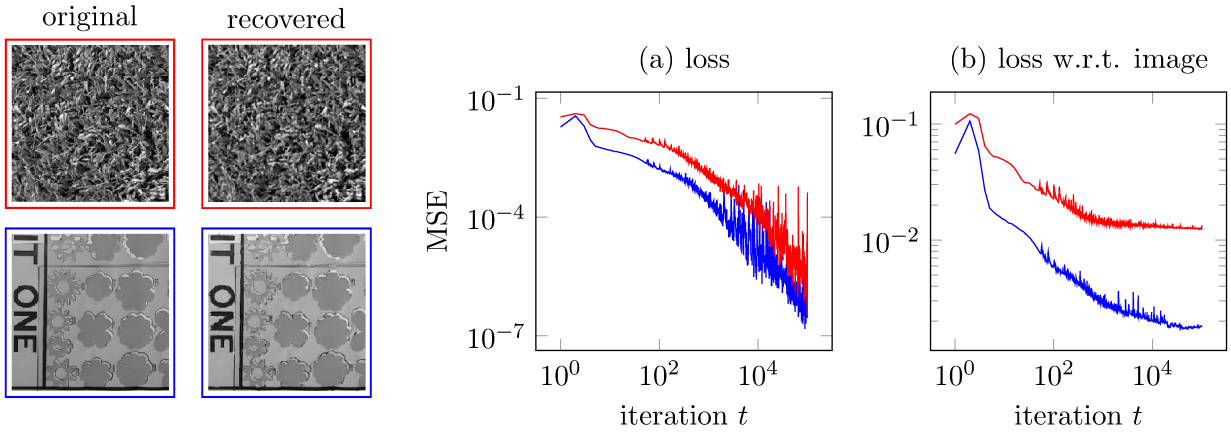


Figure 1: Compressive sensing of two different images $\mathbf{x}^*$ displayed on the right with a random matrix $\mathbf{A} \in \mathbb{R}^{m \times n}, m = n/4$, from the measurement $\mathbf{y} = \mathbf{A}\mathbf{x}^*$. Panel (a) shows the loss at iteration t, i.e., $\frac{1}{2}\|\mathbf{A}G(\mathbf{C}_t) - \mathbf{y}\|_2^2$, 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 [HH19]; a convolutional network with fixed convolutional filters. The figure looks qualitatively the same if we take G as the deep image prior [Uly+18], 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.

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 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 [Vee+18; JH19; Hec19]. Most recently Arora et al. [Aro+20] 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

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

  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.

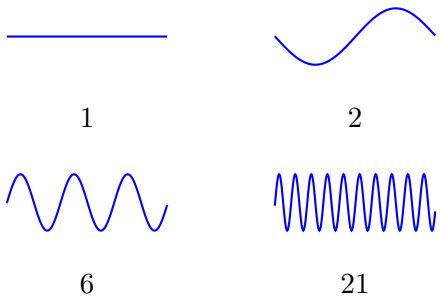


Figure 2: The 1st, 2nd, 6th, and 21st trigonometric basis functions in dimension $n = 300$.

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 [Vee+18; JH19; Hec19; Aro+20] 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.

  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.

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

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

Here, $\mathbf{v} = [1,\dots,1,-1,\dots,-1]/\sqrt{k}$ 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 [HH19]. The deep decoder in turn is a sub-set of the deep image prior [Uly+18] and the U-net [Ron+15], as commented on below.

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

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

where

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

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 $\mathbf{C}_1,\dots,\mathbf{C}_d \in \mathbb{R}^{k\times k}$. 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 [Uly+18] and the U-net [Ron+15]) by removing the encoder part, any skip connections, and most surprisingly, the trainable convolutional filters of spatial extent larger than one. As demonstrated in [HS20], 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 $\tilde{G}(\mathbf{c}) = \mathbf{Jc}$ 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*

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

Here, C is a fixed numerical constant.

The proof, given in the appendix, relies on arguments from [Hal+11, 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

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

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 $\sum_{i>2m/3} \sigma_i^2 \le \frac{\gamma^{2m/3}}{1-\gamma}$, 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

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

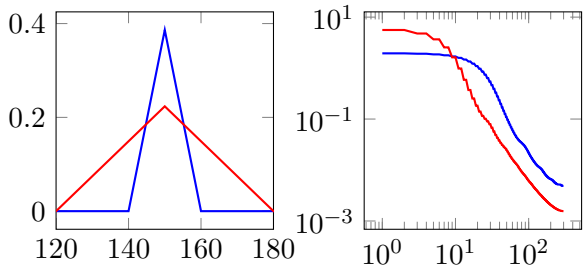
where $\mathbf{A} \in \mathbb{R}^{m \times n}, m \le n$, is a Gaussian random matrix with iid $\mathcal{N}(0, 1/m)$ entries and $G(\mathbf{C})$ is the two-layer decoder network defined in section 2. We minimize this objective by running gradient descent with a constant stepsize η, starting from a random 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 [HS20], 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

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



(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.

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 $\boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_n]$ 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 $G(\mathbf{C}) = \mathrm{ReLU}(\mathbf{UC})\mathbf{v}$, $\mathbf{C} \in \mathbb{R}^{n \times k}$, 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 $\boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_n]$. 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 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*

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

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 [SO01] 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.

4.1 Beyond two layer networks

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

(i) The finding from [HS20] 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 [HS20], 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.

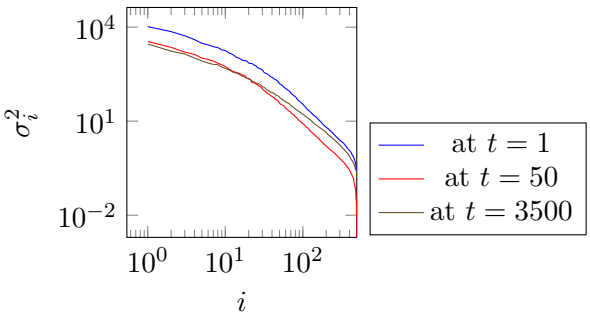


Figure 4: The singular value distribution of the Jacobian of a four-layer deep decoder at different iterations of gradient descent; the spectrum changes only slightly, and the singular values decay slightly faster than geometrically.

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 imaging 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 [Zbo+18]. 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 [Bur+12; Zha+17], or are based on learning a generative image model (by training an autoencoder or GAN [HS06; Goo+14]) and then using the resulting image models to regularize problems such as compressed sensing [Bor+17; HV18; Hua+18], denoising [Hec+20], or phase retrieval [Han+18; SA18]. In contrast to un-trained network, where optimization is over the weights of the un-trained generator, in the aformentioned papers it is over the input of the (trained) network.

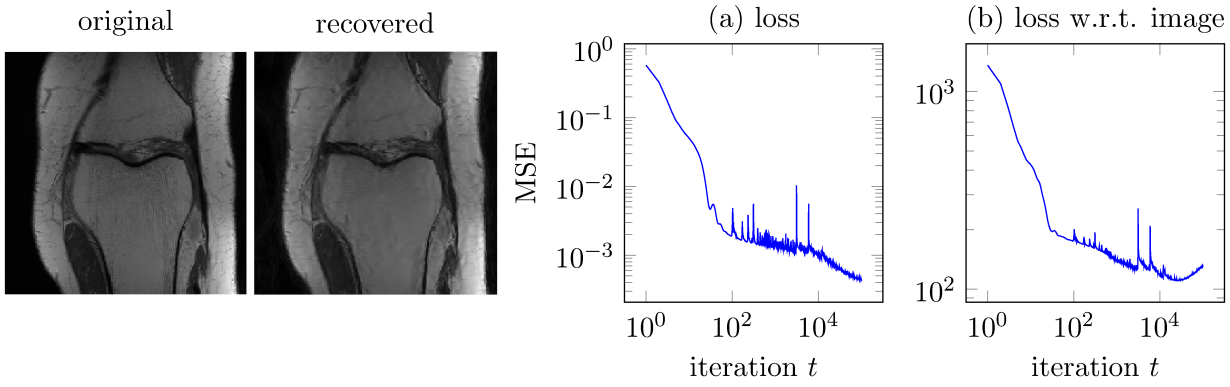


Figure 5: Compressive sensing MRI: MSE of reconstructing an image from 8-fold undersampled k-space MRI measurements. While early stopping is not absolutely necessary, stopping at about 2000 iterations slightly improves performance relative to optimizing until convergence.

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 [Sol+18; Ven+19; Du+18; OS19a; OS19b; Aro+19; Oym+19; Bas+19; Li+19]. Most related to our work is the recent paper [HS20] 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 [Lan51] 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 [TC85].

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 (24), 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 [Sol+18; Du+18; Aro+19; OS19b; Oym+19; HS20]. We define the associated linearized least-squares problem as

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

Starting from the same initial point $\boldsymbol{\theta}_0$, the gradient descent updates of the linearized problem are

$$\widetilde{\boldsymbol{\theta}}_{t+1} = \widetilde{\boldsymbol{\theta}}_t - \eta \mathbf{J}^T \left(f(\boldsymbol{\theta}_0) + \mathbf{J}(\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0) - \mathbf{y} \right). \tag{16}$$

The iterates and residuals of the non-linear and linear updates are close throughout the entire run of gradient descent provided the following assumptions are satisfied:

(i) The smallest and largest singular values of the generator reference Jacobian are lower and upper bounded by constants α and β, respectively.

(ii) The reference Jacobian approximates the Jacobian at initialization, i.e., for $\epsilon_0 > 0$,

$$\|\mathbf{J} - \mathcal{J}(\boldsymbol{\theta}_0)\| \leq \epsilon_0,$$

where $\|\cdot\|$ is the standard operator (matrix) norm.

(iii) Within a radius R around the initialization, the Jacobian varies by no more than ϵ in the sense that

$$\|\mathcal{J}(\boldsymbol{\theta}) - \mathcal{J}(\boldsymbol{\theta}_0)\| \leq \frac{\epsilon}{2}, \quad \text{for all} \quad \boldsymbol{\theta} \in \mathcal{B}_R(\boldsymbol{\theta}_0). \tag{17}$$

Here, $\mathcal{B}_R(\boldsymbol{\theta}_0) \coloneqq \{\boldsymbol{\theta} \colon \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \leq R\}$ is the ball with radius R around $\boldsymbol{\theta}_0$.

Under these assumptions, we establish that the residuals of the linear problem,

$$\widetilde{\mathbf{r}}_t \coloneqq f(\boldsymbol{\theta}_0) + \mathbf{A}\mathbf{J}_G(\tilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0) - \mathbf{y}$$

and that of the non-linear problem,

$$\mathbf{r}_t \coloneqq \mathbf{A}G(\boldsymbol{\theta}_t) - \mathbf{y},$$

are close during the entire run of gradient descent, and most importantly for proving our result, that the iterates of the linear and non-linear problem are close, again during the entire run of gradient descent:

$$\left\|\boldsymbol{\theta}_t - \widetilde{\boldsymbol{\theta}}_t\right\|_2 \leq O(\epsilon_0 + \epsilon)\|\mathbf{r}_0\|_2.$$

7.2 Inheriting the properties of the linear problem

Recall that our goal is to characterize the signal estimate $G(\boldsymbol{\theta}_\infty)$ at convergence. We characterize this estimate by

i) characterizing the estimate $\hat{\mathbf{x}} = \mathbf{J}_G\boldsymbol{\theta}_\infty$ obtained by running the linear problem until convergence and

ii) showing that this estimate is close to the original estimate, i.e., $\mathbf{J}_G\boldsymbol{\theta}_\infty \approx G(\boldsymbol{\theta}_\infty)$.

In more detail, suppose that the assumption i-iii are satisfied for sufficiently small closeness parameters ϵ_0 and ϵ. Then, as discussed above, the iterates of the non-linear problem and the linear problem are close at any iteration, in particular at convergence. Since the Jacobians are also close, we can establish that $\hat{\mathbf{x}} = \mathbf{J}_G\boldsymbol{\theta}_\infty \approx G(\boldsymbol{\theta}_\infty)$.

In more detail, we can bound the signal estimation error at convergence as

$$\begin{aligned}\|G(\boldsymbol{\theta}_\infty) - \mathbf{x}^*\|_2 &\leq \|\hat{\mathbf{x}} - \mathbf{x}^*\|_2 + \|G(\boldsymbol{\theta}_\infty) - \hat{\mathbf{x}}\|_2 \\ &\leq \|\hat{\mathbf{x}} - \mathbf{x}^*\|_2 + O(\epsilon_0 + \epsilon).\end{aligned}$$

The first term is controlled by analyzing the linear case with Theorem 1 from Section 3. To control the second term we need a simple definition

$$\mathcal{J}_G(\boldsymbol{\theta}_\infty, 0) = \int_0^1 \mathcal{J}_G(t\boldsymbol{\theta}_\infty)dt.$$

With this definition in place we can proceed to bound the second term as follows

$$\begin{aligned}&\|G(\boldsymbol{\theta}_\infty) - \hat{\mathbf{x}}\|_2 \\ &= \left\|\mathcal{J}_G(\boldsymbol{\theta}_\infty, 0)\boldsymbol{\theta}_\infty - \mathbf{J}_G\tilde{\boldsymbol{\theta}}_\infty\right\|_2 \\ &= \left\|\mathcal{J}_G(\boldsymbol{\theta}_\infty, 0)\boldsymbol{\theta}_\infty - \mathcal{J}_G(\boldsymbol{\theta}_\infty, 0)\tilde{\boldsymbol{\theta}}_\infty + \mathcal{J}_G(\boldsymbol{\theta}_\infty, 0)\tilde{\boldsymbol{\theta}}_\infty - \mathbf{J}_G\tilde{\boldsymbol{\theta}}_\infty\right\|_2 \\ &\leq \|\mathcal{J}_G(\boldsymbol{\theta}_\infty, 0)\|\left\|\boldsymbol{\theta}_\infty - \tilde{\boldsymbol{\theta}}_\infty\right\|_2 + \|\mathcal{J}_G(\boldsymbol{\theta}_\infty, 0) - \mathbf{J}_G\|\left\|\tilde{\boldsymbol{\theta}}_\infty\right\|_2 \\ &\leq O(\epsilon_0 + \epsilon).\end{aligned}$$

For the last bound we used that by our discussion above, the iterates of the non-linear problem are close at any iteration, in particular at convergence, so that $\left\|\boldsymbol{\theta}_\infty - \tilde{\boldsymbol{\theta}}_\infty\right\|_2 \leq O(\epsilon_0 + \epsilon)$.

7.3 Concluding the proof sketch

The proof for the two-layer case is then concluded by analyzing the associated linear problem. In particular, we use that the matrix $\mathbf{J}_G$ has as its left-singular vectors the trigonometric basis function, and its spectrum are the associated weights $\sigma_1, \ldots, \sigma_n$ specified in Section 4.

In order to extend this proof to a multi-layer deep decoder $G(\boldsymbol{\theta})$, all we need to do is to characterize the associated matrix $\mathbf{J}_G$, in particular its left-singular vectors and corresponding singular values.

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. 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

[Aro+19] S. Arora, S. S. Du, W. Hu, Z. Li, and R. Wang. "Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks". In: *International Conference on Machine Learning*. 2019.

[Aro+20] S. Arora, V. Roeloffs, and M. Lustig. "Untrained modified deep decoder for joint denoising parallel imaging reconstruction". In: *International Society for Magnetic Resonance in Medicine Annual Meeting*. 2020.

[Bas+19] R. Basri, D. Jacobs, Y. Kasten, and S. Kritchman. "The convergence rate of neural networks for learned functions of different frequencies". In: *Advances in Neural Information Processing Systems*. 2019.

[Bor+17] A. Bora, A. Jalal, E. Price, and A. G. Dimakis. "Compressed sensing using generative models". In: *International Conference on Machine Learning*. 2017.

[Bos+20] E. Bostan, R. Heckel, M. Chen, M. Kellman, and L. Waller. "Deep Phase Decoder: Self-calibrating phase microscopy with an untrained deep neural network". In: *Optica* (2020).

[Bur+12] H. C. Burger, C. J. Schuler, and S. Harmeling. "Image denoising: Can plain neural networks compete with BM3D?" In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2012, pp. 2392–2399.

[Du+18] S. S. Du, X. Zhai, B. Poczos, and A. Singh. "Gradient Descent Provably Optimizes Over-parameterized Neural Networks". In: *International Conference on Learning Representations*. 2018.

[Goo+14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A Courville, and Y. Bengio. "Generative adversarial nets". In: *Advances in Neural Information Processing Systems*. 2014, pp. 2672–2680.

[Hal+11] N. Halko, P. G. Martinsson, and J. A. Tropp. "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions". In: *SIAM Review* 53.2 (2011), pp. 217–288.

[HV18] P. Hand and V. Voroninski. "Global guarantees for enforcing deep generative priors by empirical risk". In: *Conference on Learning Theory*. 2018.

[Han+18] P. Hand, O. Leong, and V. Voroninski. "Phase Retrieval Under a Generative Prior". In: *Advances in Neural Information Processing Systems*. 2018.

[Hec19] R. Heckel. "Regularizing linear inverse problems with convolutional neural networks". In: *arXiv:1907.03100* (2019).

[HH19] R. Heckel and P. Hand. "Deep Decoder: Concise image representations from untrained non-convolutional networks". In: *International Conference on Learning Representations*. 2019.

[HS20] R. Heckel and M. Soltanolkotabi. "Denoising and regularization via exploiting the structural bias of convolutional generators". In: *International Conference on Learning Representations*. 2020.

[Hec+20] R. Heckel, W. Huang, P. Hand, and V. Voroninski. "Deep denoising: Rate-optimal recovery of structured signals with a deep prior". In: *Information and Inference: A Journal of the IMA* (2020).

[HS06] G. E. Hinton and R. R. Salakhutdinov. "Reducing the dimensionality of data with neural networks". In: *Science* 313.5786 (2006), pp. 504–507.

[Hua+18] W. Huang, P. Hand, R. Heckel, and V. Voroninski. "A Provably Convergent Scheme for Compressive Sensing under Random Generative Priors". In: *arXiv:1812.04176 [math]* (2018).

[HA20] R. Hyder and M. S. Asif. "Generative Models for Low-Dimensional Video Representation and Reconstruction". In: *IEEE Transactions on Signal Processing* 68 (2020), pp. 1688–1701.

[JH19] G. Jagatap and C. Hegde. "Algorithmic guarantees for inverse imaging with untrained network priors". In: *Advances in Neural Information Processing Systems*. 2019.

[Jin+19] K. H. Jin, H. Gupta, J. Yerly, M. Stuber, and M. Unser. "Time-Dependent Deep Image Prior for Dynamic MRI". In: *arXiv:1910.01684 [cs, eess]* (2019).

[Lan51] L. Landweber. "An iteration formula for fredholm integral equations of the first kind". In: *American Journal of Mathematics* 73.3 (1951), pp. 615–624.

[Li+19] M. Li, M. Soltanolkotabi, and S. Oymak. "Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks". In: *arXiv:1903.11680* (2019).

[OS19a] S. Oymak and M. Soltanolkotabi. "Overparameterized nonlinear learning: Gradient descent takes the shortest path?" In: *International Conference on Machine Learning*. 2019.

[OS19b] S. Oymak and M. Soltanolkotabi. "Towards moderate overparameterization: Global convergence guarantees for training shallow neural networks". In: *arXiv:1902.04674* (2019).

[Oym+19] S. Oymak, Z. Fabian, M. Li, and M. Soltanolkotabi. "Generalization guarantees for neural networks via harnessing the low-rank structure of the Jacobian". In: *arXiv:1906.05392* (2019).

[Ron+15] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional networks for biomedical image segmentation". In: Lecture Notes in Computer Science. 2015.

[SA18] F. Shamshad and A. Ahmed. "Robust compressive phase retrieval via deep generative priors". In: *arXiv preprint arXiv:1808.05854* (2018).

[SO01] E. P. Simoncelli and B. A. Olshausen. "Natural image statistics and neural representation". In: *Annual Review of Neuroscience* 24.1 (2001), pp. 1193–1216.

[Sol+18] M. Soltanolkotabi, A. Javanmard, and J. D. Lee. "Theoretical insights into the optimization landscape of over-parameterized shallow neural networks". In: *IEEE Transactions on Information Theory* (2018).

[TC85] H. Trussell and M. Civanlar. "The Landweber iteration and projection onto convex sets". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 33.6 (1985), pp. 1632–1634.

[Uly+18] D. Ulyanov, A. Vedaldi, and V. Lempitsky. "Deep image prior". In: *Conference on Computer Vision and Pattern Recognition*. 2018.

[Vee+18] D. V. Veen, A. Jalal, M. Soltanolkotabi, E. Price, S. Vishwanath, and A. G. Dimakis. "Compressed sensing with Deep Image Prior and learned regularization". In: *arXiv:1806.06438* (2018).

[Ven+19] L. Venturi, A. Bandeira, and J. Bruna. "Spurious valleys in two-layer neural network optimization landscapes". In: *Journal on Machine Learning Research* (2019).

[Ver12] R. Vershynin. "Introduction to the non-asymptotic analysis of random matrices". In: *Compressed sensing theory and applications*. Cambridge University Press, 2012, pp. 210–268.

[Wan+20] F. Wang, Y. Bian, H. Wang, M. Lyu, G. Pedrini, W. Osten, G. Barbastathis, and G. Situ. "Phase Imaging with an Untrained Neural Network". en. In: *Light: Science & Applications* 9.1 (2020), pp. 1–7.

[Zbo+18] J. Zbontar et al. "fastMRI: An Open Dataset and Benchmarks for Accelerated MRI". In: *arXiv:1811.08839* (2018).

[Zha+17] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising". In: *IEEE Transactions on Image Processing* 26.7 (2017), pp. 3142–3155.

A Proof of Theorem 1

The statement follows from the following more general result.

Proposition 1. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be a Gaussian random matrix with $m = k + p$, and $p \geq 4$, and let $\mathbf{J}^T = \mathbf{U}_n \mathbf{\Sigma} \mathbf{V}^T$ with $\mathbf{U}_n \in \mathbb{R}^{d\times n}$ and $\mathbf{\Sigma}, \mathbf{V} \in \mathbb{R}^{n\times n}$, be the singular value decomposition of $\mathbf{J}^T$ with singular values $\sigma_1 \geq \ldots \geq \sigma_n$. Then, for any $\mathbf{c}^* \in \mathbb{R}^d$, with probability at least $1 - 2e^{-p} - e^{-u^2/2}$, the estimate $\hat{\mathbf{c}} = \mathbf{P}_{\mathbf{A}^T\mathbf{J}^T}\mathbf{c}$ obeys*

$$\|\mathbf{J}\hat{\mathbf{c}} - \mathbf{J}\mathbf{c}\|_2^2 \leq \left\|\mathbf{U}_n^T\mathbf{c}^*\right\|_2^2 \left(\left(\sigma_{k+1} e\left(\sqrt{\frac{3k}{p+1}} + \frac{e\sqrt{k+p}}{p+1}u\right) + \sqrt{\sum_{i>k}\sigma_i^2}\frac{e\sqrt{k+p}}{p+1}u\right)^2 + \sum_{j>k}\sigma_j^2\right).$$

To see this, note that with $p = k/2$ and $u = \sqrt{p}$, the proposition guarantees that with probability at least $1 - 3e^{-p}$,

$$\begin{aligned}\|\mathbf{J}\hat{\mathbf{c}} - \mathbf{J}\mathbf{c}^*\|_2^2 &\leq \left\|\mathbf{U}_n^T\mathbf{c}^*\right\|_2^2 \left(\left(25\sigma_{k+1} + 7\sqrt{\sum_{i>k}\sigma_i^2}\right)^2 + \sum_{i>k}\sigma_i^2\right)\\ &\leq \left\|\mathbf{U}_n^T\mathbf{c}^*\right\|_2^2 32^2 \sum_{i>k}\sigma_i^2.\end{aligned}$$

Noting that $m = 3/2k$, $\hat{\mathbf{x}} = \mathbf{J}\hat{\mathbf{c}}$ and $\mathbf{x}^* = \mathbf{J}\mathbf{c}^*$ concludes the proof.

Proof of Proposition 1: By the characterization (6), our goal is to upper bound

$$\|\mathbf{J}\hat{\mathbf{c}} - \mathbf{J}\mathbf{c}^*\|_2^2 = \left\|\mathbf{c}^{*T}(\mathbf{I} - \mathbf{P}_{\mathbf{J}^T\mathbf{A}^T})\mathbf{J}^T\right\|_2^2. \tag{18}$$

Our proof relies on arguments from [Hal+11, Sec. 8 and Sec. 9] developed for approximating low-rank matrices through random sampling.

We start by partitioning the right-singular vectors of $\mathbf{J}^T$ into two blocks $\mathbf{V}_1$ and $\mathbf{V}_2$ containing k and $n - k$ columns, respectively.

$$\mathbf{J}^T = \mathbf{U}_n \begin{bmatrix}\mathbf{\Sigma}_1 & 0\\ 0 & \mathbf{\Sigma}_2\end{bmatrix}\begin{bmatrix}\mathbf{V}_1^T\\ \mathbf{V}_2^T\end{bmatrix}.$$

Define the random matrices

$$\Omega_1 = \mathbf{V}_1^T\mathbf{A}^T \in \mathbb{R}^{k\times m}, \quad \Omega_2 = \mathbf{V}_2^T\mathbf{A}^T \in \mathbb{R}^{n-k\times m}.$$

Note that both matrices are standard Gaussian, and, because they are non-overlapping sub-matrices of $\mathbf{VA}$, they are also stochastically independent. Moreover, Ω_1 has full row-rank with probability one.

For convenience, define

$$\tilde{\mathbf{J}}^T = \mathbf{\Sigma}\mathbf{V}.$$

Next, we record a useful property from [Hal+11, Prop. 8.4]: For a unitary matrix $\mathbf{U}$ any matrix $\mathbf{M}$,

$$\mathbf{P}_{\mathbf{M}} = \mathbf{U}\mathbf{P}_{\mathbf{U}^T\mathbf{M}}\mathbf{U}^T. \tag{19}$$

To see that the identity (19) holds, first note that the matrix $\mathbf{P} = \mathbf{U}^T\mathbf{P}_\mathbf{M}\mathbf{U}$ is an orthogonal projection operator because it is Hermitian an $\mathbf{P}^2 = \mathbf{P}$. Moreover,

$$\text{range}(\mathbf{P}) = \mathbf{U}^T\text{range}(\mathbf{M}) = \text{range}(\mathbf{U}^T\mathbf{M}).$$

Since the range determines the orthogonal projector onto its range, we have that $\mathbf{P} = \mathbf{U}^T\mathbf{P}_\mathbf{M}\mathbf{U} = \mathbf{P}_{\mathbf{U}^T\mathbf{M}}$, concluding the proof of (19). Next, let

$$\mathbf{J}^T = \underbrace{[\mathbf{U}_n\mathbf{U}_{d-n}]}_{\mathbf{U}} \begin{bmatrix} \mathbf{\Sigma}_1 & 0 \\ 0 & \mathbf{\Sigma}2 \\ 0 & 0 \end{bmatrix} \mathbf{V}^T$$

be the full singular value decomposition of $\mathbf{J}^T$, including the singular vectors $\mathbf{U}_{d-n}$ multiplying with zero singular values. Applying the identity (19) and that $\mathbf{U}^T\mathbf{U}$ we proceed as

$$\begin{aligned}
\left\|\mathbf{c}^T(\mathbf{I} - \mathbf{P}_{\mathbf{J}^T\mathbf{A}^T})\mathbf{J}^T\right\|_2^2 &= \left\|\mathbf{c}^T\mathbf{U}(\mathbf{I} - \mathbf{P}_{\mathbf{U}^T\mathbf{J}^T\mathbf{A}^T})\mathbf{U}^T\mathbf{J}^T\right\|_2^2 \\
&= \left\|\mathbf{c}^T[\mathbf{U}_n\mathbf{U}_{d-n}](\mathbf{I} - \mathbf{P}_{\begin{bmatrix}\tilde{\mathbf{J}}^T\mathbf{A}^T \\ 0\end{bmatrix}})\begin{bmatrix}\mathbf{\Sigma} \\ 0\end{bmatrix}\right\|_2^2 \\
&= \left\|\mathbf{c}^T[\mathbf{U}_n\mathbf{U}_{d-n}]\begin{bmatrix}(\mathbf{I} - \mathbf{P}_{\tilde{\mathbf{J}}^T\mathbf{A}^T})\mathbf{\Sigma} \\ 0\end{bmatrix})\right\|_2^2 \\
&= \left\|\mathbf{c}^T\mathbf{U}_n(\mathbf{I} - \mathbf{P}_{\tilde{\mathbf{J}}^T\mathbf{A}^T})\mathbf{\Sigma}\right\|_2^2.
\end{aligned}$$

Moreover,

$$\begin{aligned}
\left\|\mathbf{c}^T\mathbf{U}_n(\mathbf{I} - \mathbf{P}_{\tilde{\mathbf{J}}^T\mathbf{A}^T})\mathbf{\Sigma}\right\|_2^2 &\leq \left\|\mathbf{c}^T\mathbf{U}_n\right\|_2^2\left\|(\mathbf{I} - \mathbf{P}_{\tilde{\mathbf{J}}^T\mathbf{A}^T})\mathbf{\Sigma}\right\|^2 \\
&= \left\|\mathbf{c}^T\mathbf{U}_n\right\|_2^2\left\|\mathbf{\Sigma}^T(\mathbf{I} - \mathbf{P}_{\tilde{\mathbf{J}}^T\mathbf{A}^T})\mathbf{\Sigma}\right\| \\
&\leq \left\|\mathbf{\Sigma}_2\Omega_2\Omega_1^\dagger\right\|^2 + \|\mathbf{\Sigma}_2\|^2 \\
&\leq \left\|\mathbf{c}^T\mathbf{U}_n\right\|_2^2\left(\left(\|\mathbf{\Sigma}_2\|e\left(\sqrt{\frac{3k}{p+1}} + \frac{e\sqrt{k+p}}{p+1}u\right) + \|\mathbf{\Sigma}_2\|_F\frac{e\sqrt{k+p}}{p+1}ut\right)^2 + \|\mathbf{\Sigma}_2\|^2\right),
\end{aligned}$$

where the second-to-last inequality follows from [Hal+11, Last ineq in Sec. 9.2]. Finally, the last inequality holds with the probability specified in the proposition because by [Hal+11, Last inequality in Sec. 10.3], for $p \geq 4$ and $u > 0$,

$$\mathrm{P}\left[\left\|\mathbf{\Sigma}_2\Omega_2\Omega_1^\dagger\right\| \geq \|\mathbf{\Sigma}_2\|e\left(\sqrt{\frac{3k}{p+1}} + \frac{e\sqrt{k+p}}{p+1}u\right) + \|\mathbf{\Sigma}_2\|_F\frac{e\sqrt{k+p}}{p+1}ut\right] \leq 2e^{-p} + e^{-u^2/2}.$$

This concludes the proof of the proposition.

B Proof of Theorem 2

The result stated in the main text (Theorem 2) is obtained from a slightly more general result which applies beyond convolutional networks. Specifically, we consider neural network generators of the form

$$G(\mathbf{C}) = \mathrm{ReLU}(\mathbf{UC})\mathbf{v},$$

with $\mathbf{C} \in \mathbb{R}^{n\times k}$, and $\mathbf{U} \in \mathbb{R}^{n\times n}$ an arbitrary fixed matrix, and $\mathbf{v} \in \mathbb{R}^k$, with half of the entries of $\mathbf{v}$ equal to $+1/\sqrt{k}$ and the other half equal to $-1/\sqrt{k}$.

The (transposed) Jacobian $\mathrm{ReLU}(\mathbf{Uc})$ is $\mathbf{U}^T\mathrm{diag}(\mathrm{ReLU}'(\mathbf{Uc}))$. Thus the Jacobian of $G(\mathbf{C})$ is given by

$$\mathcal{J}_G^T(\mathbf{C}) = \begin{bmatrix} v_1\mathbf{U}^T\mathrm{diag}(\mathrm{ReLU}'(\mathbf{Uc}_1)) \\ \vdots \\ v_k\mathbf{U}^T\mathrm{diag}(\mathrm{ReLU}'(\mathbf{Uc}_k)) \end{bmatrix} \in \mathbb{R}^{nk\times n}, \tag{20}$$

where ReLU' is the derivative of the activation function. Next we define a notion of expected Jacobian. Towards this goal, we first define the matrix

$$\mathbf{\Sigma}(\mathbf{U}) \coloneqq \mathbb{E}\left[\mathcal{J}_G(\mathbf{C})\mathcal{J}_G^T(\mathbf{C})\right],$$

associated with the function $G(\mathbf{C}) = \mathrm{ReLU}(\mathbf{UC})\mathbf{v}$. Here, expectation is over $\mathbf{C}$ with iid $\mathcal{N}(0,\omega)$ entries. Consider the eigenvalue decomposition of $\mathbf{\Sigma}(\mathbf{U})$ given by

$$\mathbf{\Sigma}(\mathbf{U}) = \sum_{i=1}^{n} \sigma_i^2 \mathbf{w}_i\mathbf{w}_i^T.$$

Our results depend on the largest and smallest eigenvalue of $\mathbf{\Sigma}(\mathbf{U})$ denoted by σ_n^2 and $\|\mathbf{U}\|^2$ and in particular a condition number denoted by κ formally defined as

$$\kappa_{\mathbf{u}} := \frac{\|\mathbf{U}\|^2}{\sigma_n^2}.$$

With these definitions in place we are now ready to state our result about neural generators.

Theorem 3. *Consider a compressive observation $\mathbf{y} \in \mathbb{R}^m$ given by*

$$\mathbf{y} = \mathbf{Ax},$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ with $m \le \frac{n}{9}$ is a Gaussian random matrix with iid $\mathcal{N}(0, 1/m)$ entries. Suppose that the number of channels obeys

$$k \ge C\frac{\kappa_{\mathbf{u}}^{26}}{\xi^8}m \tag{21}$$

for an error tolerance parameter $0 < \xi \le \frac{1}{\sqrt{2\log\left(\frac{2n}{\delta}\right)}}$. We fit the neural generator $G(\mathbf{C})$ to the signal $\mathbf{y} \in \mathbb{R}^n$ by minimizing a loss of the form

$$\mathcal{L}(\mathbf{C}) = \frac{1}{2}\|\mathbf{A}G(\mathbf{C}) - \mathbf{y}\|_2^2 \tag{22}$$

via running gradient descent with iterations $\mathbf{C}_{t+1} = \mathbf{C}_t - \eta \nabla \mathcal{L}(\mathbf{C}_t)$*, starting from* $\mathbf{C}_0$ *with i.i.d.* $\mathcal{N}(0, \omega^2)$ *entries,* $\omega = \frac{\xi \|\mathbf{y}\|_2}{2\sqrt{n}\|\mathbf{U}\|}$*, and step size obeying* $\eta \leq \frac{m}{4n\|\mathbf{U}\|^2}$*. Then, with probability at least* $1 - ne^{-k^2} - 2e^{-\frac{m}{2}} - \delta$*, for all iterations* t*,*

$$\|\mathbf{x} - G(\mathbf{C}_t)\|_2 \leq \xi \|\mathbf{x}^*\|_2 + C \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2} \langle \mathbf{w}_i, \mathbf{x}^* \rangle^2 \right) \sum_{i > 2m/3} \sigma_i^2. \tag{23}$$

Theorem 2 follows directly from Theorem 3 by noting that for $\mathbf{U}$ a circulant matrix (implementing a convolution), as found in [HS20], the left singular vectors of $\mathbf{\Sigma}(\mathbf{U})$ are given by the trigonometric basis functions in (10) and the singular values are given by (11).

C The dynamics of linear and nonlinear least-squares

Theorem 3, proven below, builds on a result on the dynamics of a general non-linear least squares problem that is stated and discussed in this section. Consider a nonlinear least-squares fitting problem of the form

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

Here, $f \colon \mathbb{R}^N \to \mathbb{R}^n$ is a non-linear model with parameters $\boldsymbol{\theta} \in \mathbb{R}^N$.

To solve this problem, we run gradient descent with a fixed stepsize η, starting from an initial point $\boldsymbol{\theta}_0$, with updates of the form

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla \mathcal{L}(\boldsymbol{\theta}_t) \quad \text{where} \quad \nabla \mathcal{L}(\boldsymbol{\theta}) = \mathcal{J}^T(\boldsymbol{\theta})(f(\boldsymbol{\theta}) - \mathbf{y}). \tag{24}$$

Here, $\mathcal{J}(\boldsymbol{\theta}) \in \mathbb{R}^{n \times N}$ is the Jacobian associated with the nonlinear map f with entries given by $[\mathcal{J}(\boldsymbol{\theta})]_{i,j} = \frac{\partial f_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}$. In order to study the properties of the gradient descent iterates in (24), 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 variety of recent publications [Sol+18; Du+18; Aro+19; OS19b; Oym+19], our specific argument is most similar to [HS20]. Contrary to the result in [HS20], which holds for a certain number of initial iterations, our statement applied to all iterations.

The associated linearized least-squares problem is defined as

$$\mathcal{L}_{\text{lin}}(\boldsymbol{\theta}) = \frac{1}{2} \|f(\boldsymbol{\theta}_0) + \mathbf{J}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \mathbf{y}\|_2^2. \tag{25}$$

Here, $\mathbf{J} \in \mathbb{R}^{n \times N}$, refered to as the reference Jacobian, is a fixed matrix independent of the parameter $\boldsymbol{\theta}$ that approximates the Jacobian mapping at initialization, $\mathcal{J}(\boldsymbol{\theta}_0)$. Starting from the same initial point $\boldsymbol{\theta}_0$, the gradient descent updates of the linearized problem are

$$\widetilde{\boldsymbol{\theta}}_{t+1} = \widetilde{\boldsymbol{\theta}}_t - \eta \mathbf{J}^T \left(f(\boldsymbol{\theta}_0) + \mathbf{J}(\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0) - \mathbf{y} \right). \tag{26}$$

To show that the non-linear updates (24) are close to the linearized iterates (26), we make the following assumptions:

Assumption 1 (Bounded spectrum)**.** *We assume the singular values of the reference Jacobian obey for some* α, β

$$\sqrt{2}\alpha \leq \sigma_n \leq \sigma_1 \leq \beta. \tag{27}$$

Furthermore, we assume that the Jacobian mapping associated with the nonlinear model f *obeys*

$$\|\mathcal{J}(\boldsymbol{\theta})\| \leq \beta \quad \text{for all} \quad \boldsymbol{\theta} \in \mathbb{R}^N. \tag{28}$$

Assumption 2 (Closeness of the reference and initialization Jacobians)**.** *We assume the reference Jacobian and the Jacobian of the nonlinearity at initialization* $\mathcal{J}(\boldsymbol{\theta}_0)$ *are* ϵ_0*-close in the sense that*

$$\|\mathcal{J}(\boldsymbol{\theta}_0) - \mathbf{J}\| \leq \epsilon_0. \tag{29}$$

Assumption 3 (Bounded variation of Jacobian around initialization)**.** *We assume that within a radius* R *around the initialization, the Jacobian varies by no more than* ϵ *in the sense that*

$$\|\mathcal{J}(\boldsymbol{\theta}) - \mathcal{J}(\boldsymbol{\theta}_0)\| \leq \frac{\epsilon}{2}, \quad \text{for all} \quad \boldsymbol{\theta} \in \mathcal{B}_R(\boldsymbol{\theta}_0), \tag{30}$$

where $\mathcal{B}_R(\boldsymbol{\theta}_0) \coloneqq \{\boldsymbol{\theta} \colon \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq R\}$ *is the ball with radius* R *around* $\boldsymbol{\theta}_0$*.*

Under these assumptions i) the difference of the nonlinear iterative updates (24) and the linear iterative updates (26) is bounded, and ii) the difference of the linear and non-linear residuals, defined as

$$\text{nonlinear residual:} \quad \mathbf{r}_t \coloneqq f(\boldsymbol{\theta}_t) - \mathbf{y} \tag{31}$$

$$\text{linear residual:} \quad \widetilde{\mathbf{r}}_t \coloneqq f(\boldsymbol{\theta}_0) + \mathbf{J}(\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0) - \mathbf{y} \tag{32}$$

are close throughout the entire run of gradient descent; both in the proximity of the initialization.

Theorem 4 (Closeness of linear and nonlinear least-squares problems)**.** *Assume the Jacobian mapping* $\mathcal{J}(\boldsymbol{\theta}) \in \mathbb{R}^{n \times N}$ *associated with the function* $f(\boldsymbol{\theta})$ *obeys Assumptions 1, 2, and 3 around an initial point* $\boldsymbol{\theta}_0 \in \mathbb{R}^N$ *with respect to a reference Jacobian* $\mathbf{J} \in \mathbb{R}^{n \times N}$ *and with parameters* $\alpha, \beta, \epsilon_0, \epsilon$*, obeying* $2\beta(\epsilon_0 + \epsilon) \leq \alpha^2$*, and* R*. Furthermore, assume the radius* R *is given by*

$$\frac{R}{2} \coloneqq \left\|\mathbf{J}^\dagger \mathbf{r}_0\right\|_2 + 2.5\frac{\beta^2}{\alpha^4}(\epsilon_0 + \epsilon)\|\mathbf{r}_0\|_2. \tag{33}$$

Here, $\mathbf{J}^\dagger$ *is the pseudo-inverse of* $\mathbf{J}$*. We run gradient descent with stepsize* $\eta \leq \frac{1}{\beta^2}$ *on the linear and non-linear least squares problem, starting from the same initialization* $\boldsymbol{\theta}_0$*. Then, for all iterations* t*,*

i) the non-linear residual converges geometrically

$$\|\mathbf{r}_t\|_2 \leq \left(1 - \eta\alpha^2\right)^t \|\mathbf{r}_0\|_2, \tag{34}$$

ii) the residuals of the original and the linearized problems are close

$$\|\mathbf{r}_t - \widetilde{\mathbf{r}}_t\|_2 \leq 2\beta\eta(\epsilon_0 + \epsilon)(1 - \eta\alpha^2)^{t-1} t \|\mathbf{r}_0\|_2 \tag{35}$$

$$\leq \frac{2\beta(\epsilon_0 + \epsilon)}{e(\ln 2)\alpha^2}\|\mathbf{r}_0\|_2, \tag{36}$$

iii) the parameters of the original and the linearized problems are close

$$\left\|\boldsymbol{\theta}_t - \widetilde{\boldsymbol{\theta}}_t\right\|_2 \leq 2.5\frac{\beta^2}{\alpha^4}(\epsilon_0 + \epsilon)\|\mathbf{r}_0\|_2, \tag{37}$$

iv) and finally, the parameters are not far from the initialization

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2 \leq \frac{R}{2}. \tag{38}$$

The above theorem formalizes that in a (small) radius around the initialization, the non-linear problem behaves similarly as its linearization. Thus to characterize the dynamics of the nonlinear problem, it suffices to characterize the dynamics of the linearized problem. This is the subject of our next theorem, which is a standard results on the iterates of least squares, see [HS20, Thm. 5] for the proof.

Proposition 2 (Theorem 5 in [HS20])**.** *Consider a linear least squares problem (25) and let* $\mathbf{J} = \mathbf{W}\boldsymbol{\Sigma}\mathbf{V}^T \in \mathbb{R}^{n\times p} = \sum_{i=1}^n \sigma_i \mathbf{w}_i \mathbf{v}_i^T$ *be the singular value decomposition of the matrix* $\mathbf{J}$*. Then the residual* $\widetilde{\mathbf{r}}_t$ *after* t *iterations of gradient descent with updates (26) is*

$$\widetilde{\mathbf{r}}_t = \sum_{i=1}^n \left(1 - \eta\sigma_i^2\right)^t \mathbf{w}_i \left\langle \mathbf{w}_i, \mathbf{r}_0 \right\rangle. \tag{39}$$

Moreover, using a step size satisfying $\eta \leq \frac{1}{\sigma_1^2}$*, the linearized iterates (26) obey*

$$\left\|\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0\right\|_2^2 = \sum_{i=1}^n \left(\left\langle \mathbf{w}_i, \mathbf{r}_0 \right\rangle \frac{1 - (1 - \eta\sigma_i^2)^t}{\sigma_i}\right)^2. \tag{40}$$

In the next section we show we can combine these two general theorems to provide guarantees for compressed sensing using general neural networks.

C.1 Proof of Theorem 4 (closeness of linear and non-linear least-squares)

The proof is by induction. We note that the base case $t = 0$ is trivially true. We suppose the statement, in particular the bounds (34), (35), (36), (37), and (38) hold for all iterations $\tau \leq t-1$. We then show that those relations continue to hold for iteration t in five steps: In Step I, we show that a weaker version of (38) holds, specifically that $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2 \leq R$. This guarantees that we can work with our assumptions; those require the iterates to be sufficiently close to the initial values. In Step II we show that the nonlinear residual decreases at a geometric rate proving (34). In Steps III and IV we show that the residuals and the coefficients of the linear and non-linear problem are close, respectively. Finally, in Step V we utilize Steps I-IV to complete the proof by showing that the iterates of the non-linear problem are close to its initialization (i.e., equation (38)).

Linear convergence of linear residual: Before we start, we note that under our assumption, the residual of the linear problem converges linearly. Specifically, by the updates of the linear problem (26), we have that

$$\widetilde{\mathbf{r}}_{t+1} = (\mathbf{I} - \eta\mathbf{J}\mathbf{J}^T)\widetilde{\mathbf{r}}_t. \tag{41}$$

Using that the smallest singular values of $\mathbf{J}\mathbf{J}^T$ is lower bounded by $2\alpha^2$, this guarantees that

$$\|\widetilde{\mathbf{r}}_t\|_2 \leq (1-2\eta\alpha^2)^t \|\widetilde{\mathbf{r}}_0\|_2,$$

establishing linear convergence of the linear problem.

Step I: Next iterate obeys $\boldsymbol{\theta}_t \in \mathcal{B}_R(\boldsymbol{\theta}_0)$. We start by using a coarse argument that establishes $\boldsymbol{\theta}_t \in \mathcal{B}_R(\boldsymbol{\theta}_0)$. First note that by the triangle inequality and the induction assumption (38) we have

$$\begin{aligned}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2 \leq& \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}\|_2 + \|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}_0\|_2, \\ \leq& \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}\|_2 + \frac{R}{2}.\end{aligned}$$

So to prove $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2 \leq R$ it suffices to show that $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}\|_2 \leq R/2$. To this aim note that

$$\begin{aligned}\frac{1}{\eta}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}\|_2 &= \|\nabla \mathcal{L}(\boldsymbol{\theta}_{t-1})\|_2 \\ &= \left\|\mathcal{J}^T(\boldsymbol{\theta}_{t-1})\mathbf{r}_{t-1}\right\|_2 \\ &\leq \left\|\mathcal{J}^T(\boldsymbol{\theta}_{t-1})\widetilde{\mathbf{r}}_{t-1}\right\|_2 + \|\mathcal{J}(\boldsymbol{\theta}_{t-1})\|\|\mathbf{r}_{t-1} - \widetilde{\mathbf{r}}_{t-1}\|_2 \\ &\leq \left\|\mathbf{J}^T\widetilde{\mathbf{r}}_{t-1}\right\|_2 + \|\mathcal{J}(\boldsymbol{\theta}_{t-1}) - \mathbf{J}\|\|\widetilde{\mathbf{r}}_{t-1}\|_2 + \|\mathcal{J}(\boldsymbol{\theta}_{t-1})\|\|\mathbf{r}_{t-1} - \widetilde{\mathbf{r}}_{t-1}\|_2 \\ &\overset{\text{(i)}}{\leq} \beta^2\left\|\mathbf{J}^\dagger\mathbf{r}_0\right\|_2 + (\epsilon+\epsilon_0)\|\mathbf{r}_0\|_2 + \frac{2\beta^2(\epsilon_0+\epsilon)}{e(\ln 2)\alpha^2}\|\mathbf{r}_0\|_2 \\ &\overset{\text{(ii)}}{\leq} \beta^2\left\|\mathbf{J}^\dagger\mathbf{r}_0\right\|_2 + \frac{2\beta^2}{\alpha^2}(\epsilon_0+\epsilon)\|\mathbf{r}_0\|_2. \end{aligned} \tag{42}$$

Here, (ii) follows from the fact that $\frac{1}{2} \leq \frac{\beta^2}{\alpha^2}$ and inequality (i) follows from Assumptions 1-3, the induction hypothesis (36), $\|\widetilde{\mathbf{r}}_{\tau-1}\| \leq \|\mathbf{r}_0\|$, and the bound

$$\begin{aligned}\left\|\mathbf{J}^T\widetilde{\mathbf{r}}_{t-1}\right\|_2 &= \left\|\mathbf{J}^T(\mathbf{I}-\eta\mathbf{J}\mathbf{J}^T)^{t-1}\mathbf{r}_0\right\|_2 \\ &= \left\|\boldsymbol{\Sigma}(\mathbf{I}-\eta\boldsymbol{\Sigma}^2)^{t-1}\mathbf{W}^T\mathbf{r}_0\right\|_2 \\ &\leq \sqrt{\sum_{j=1}^n \sigma_j^2\langle \mathbf{w}_j, \mathbf{r}_0\rangle^2} \\ &\leq \beta^2\sqrt{\sum_{j=1}^n \frac{1}{\sigma_j^2}\langle \mathbf{w}_j, \mathbf{r}_0\rangle^2} \\ &= \beta^2\left\|\mathbf{J}^\dagger\mathbf{r}_0\right\|_2.\end{aligned}$$

To continue we use the fact that $\eta \leq \frac{1}{\beta^2}$ in (42) to conclude that

$$\begin{aligned}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}\|_2 &\leq \eta\beta^2\left\|\mathbf{J}^\dagger\mathbf{r}_0\right\|_2 + \eta\frac{2\beta^2(\epsilon_0+\epsilon)}{\alpha^2}\|\mathbf{r}_0\|_2. \\ &\leq \left\|\mathbf{J}^\dagger\mathbf{r}_0\right\|_2 + \frac{2(\epsilon_0+\epsilon)}{\alpha^2}\|\mathbf{r}_0\|_2. \\ &\leq \frac{R}{2}.\end{aligned}$$

The last inequality follows by definition of R in (33), and concludes the proof of Step I.

Step II: Geometric decay of non-linear iterate. Since the linear residuals converge linearly and the Jacobian of the non-linear problem is close the Jacobian of the linear problem, $\mathbf{J}$, the non-linear problem also converges linearly. To see this, with $\mathcal{J}(\mathbf{a}, \mathbf{b}) = \int_0^1 \mathcal{J}(s\mathbf{b} - (1-s)\mathbf{a})ds$, we have that, by the mean value theorem

$$\begin{aligned} f(\boldsymbol{\theta}_t) &= f(\boldsymbol{\theta}_{t-1} - \eta \nabla \mathcal{L}(\boldsymbol{\theta}_{t-1})) \\ &= f(\boldsymbol{\theta}_{t-1}) - \eta \mathcal{J}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}) \nabla \mathcal{L}(\boldsymbol{\theta}_{t-1}) \\ &= f(\boldsymbol{\theta}_{t-1}) - \eta \mathcal{J}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}) \mathcal{J}^T(\boldsymbol{\theta}_{t-1})(f(\boldsymbol{\theta}_{t-1}) - \mathbf{y}) \\ &= f(\boldsymbol{\theta}_{t-1}) - \eta \mathbf{B}_1 \mathbf{B}_2 (f(\boldsymbol{\theta}_{t-1}) - \mathbf{y}). \end{aligned}$$

where in the last equality we defined the matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ accordingly for notational convenience. This implies that

$$\begin{aligned} \mathbf{r}_t &= f(\boldsymbol{\theta}_t) - \mathbf{y} \\ &= (\mathbf{I} - \eta \mathbf{B}_1 \mathbf{B}_2)(f(\boldsymbol{\theta}_{t-1}) - \mathbf{y}) \\ &= (\mathbf{I} - \eta \mathbf{B}_1 \mathbf{B}_2) \mathbf{r}_{t-1}. \end{aligned} \tag{43}$$

Thus,

$$\begin{aligned} \|\mathbf{r}_t\|_2 &\leq \|\mathbf{I} - \eta \mathbf{B}_1 \mathbf{B}_2\| \|\mathbf{r}_{t-1}\|_2 \\ &\leq \left(\left\|\mathbf{I} - \eta \mathbf{J}\mathbf{J}^T\right\| + \eta \left\|\mathbf{J}\mathbf{J}^T - \mathbf{B}_1 \mathbf{B}_2\right\| \right) \|\mathbf{r}_{t-1}\|_2 \\ &\overset{\text{(i)}}{\leq} \left(1 - 2\eta\alpha^2 + 2\eta\beta(\epsilon_0 + \epsilon)\right) \|\mathbf{r}_{t-1}\|_2 \\ &\overset{\text{(ii)}}{\leq} \left(1 - \eta\alpha^2\right) \|\mathbf{r}_{t-1}\|_2 \end{aligned}$$

For inequality (ii) we used the assumption $2\beta(\epsilon_0 + \epsilon) \leq \alpha^2$, and for inequality (i) we used the bound

$$\begin{aligned} \left\|\mathbf{J}\mathbf{J}^T - \mathbf{B}_1 \mathbf{B}_2\right\| &= \left\|\mathbf{J}\mathbf{J}^T - \mathbf{J}\mathbf{B}_2 + \mathbf{J}\mathbf{B}_2 - \mathbf{B}_1 \mathbf{B}_2\right\| \\ &\leq \|\mathbf{J}\| \left\|\mathbf{J}^T - \mathbf{B}_2\right\| + \|\mathbf{J} - \mathbf{B}_1\| \|\mathbf{B}_2\| \leq 2\beta(\epsilon_0 + \epsilon), \end{aligned} \tag{44}$$

where the last inequality follows from our assumptions, and using that, by the triangle inequality and assumptions 2 and 3, we have

$$\left\|\mathbf{B}_2 - \mathbf{J}^T\right\| = \|\mathcal{J}(\boldsymbol{\theta}_{t-1}) - \mathbf{J}\| \leq \|\mathcal{J}(\boldsymbol{\theta}_{t-1}) - \mathcal{J}(\boldsymbol{\theta}_0)\| + \|\mathcal{J}(\boldsymbol{\theta}_0) - \mathbf{J}\| \leq \epsilon_0 + \epsilon. \tag{45}$$

This establishes that

$$\|\mathbf{r}_t\|_2 \leq (1 - \eta\alpha^2) \|\mathbf{r}_{t-1}\|_2 \leq (1 - \eta\alpha^2)^t \|\mathbf{r}_0\|_2, \tag{46}$$

where in the last inequality we used the induction hypothesis (34). This completes the proof of the bound (34) for iteration t concluding Step II.

Step III: Original and linearized residuals are close. In this step, we bound the deviation of the residuals of the original and linearized problem defined as

$$\mathbf{e}_t \coloneqq \widetilde{\mathbf{r}}_t - \mathbf{r}_t.$$

Specifically, we use the induction hypothesis together with the fact that based on Step I we have $\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t \in \mathcal{B}_R(\boldsymbol{\theta}_0)$, to show that

$$\|\mathbf{e}_t\| \leq 2\beta\eta(\epsilon_0+\epsilon)(1-\eta\alpha^2)^{t-1}t\|\mathbf{r}_0\|_2. \tag{47}$$

Before we prove this however note that for $x \leq 1/2$ we have $(1-x)^{t-1}t \leq \frac{1}{e(\ln 2)x}$ for all $t \geq 0$. Now using this identity with $x = \eta\alpha^2 \leq \frac{\alpha^2}{\beta^2} \leq \frac{1}{2}$ in (47) we conclude that

$$\|\mathbf{e}_t\| \leq \frac{2\beta(\epsilon_0+\epsilon)}{e(\ln 2)\alpha^2}\|\mathbf{r}_0\|_2,$$

completing the proof of (36) for iteration t. Thus, all that remains in this step is to establish (47). To this aim note that from the formulas for the linear and non-linear residuals in (41) and (43), we have that

$$\widetilde{\mathbf{r}}_t = (\mathbf{I}-\eta\mathbf{J}\mathbf{J}^T)\widetilde{\mathbf{r}}_{t-1}.$$

Thus for $\mathbf{e}_t = \widetilde{\mathbf{r}}_t - \mathbf{r}_t$ we have, with the same notation as in step II,

$$\begin{aligned}
\|\mathbf{e}_t\| &= \left\|(\mathbf{I}-\eta\mathbf{J}\mathbf{J}^T)\widetilde{\mathbf{r}}_{t-1} - (\mathbf{I}-\eta\mathbf{B}_1\mathbf{B}_2)\mathbf{r}_{t-1}\right\|_2 \\
&= \left\|(\mathbf{I}-\eta\mathbf{J}\mathbf{J}^T)(\widetilde{\mathbf{r}}_{t-1}-\mathbf{r}_{t-1}) + \eta(\mathbf{B}_1\mathbf{B}_2 - \mathbf{J}\mathbf{J}^T)\mathbf{r}_{t-1}\right\|_2 \\
&\leq \left\|\mathbf{I}-\eta\mathbf{J}\mathbf{J}^T\right\|\|\widetilde{\mathbf{r}}_{t-1}-\mathbf{r}_{t-1}\|_2 + \eta\left\|\mathbf{B}_1\mathbf{B}_2 - \mathbf{J}\mathbf{J}^T\right\|\|\mathbf{r}_{t-1}\|_2 \\
&\leq (1-\eta\alpha^2)\|\mathbf{e}_{t-1}\|_2 + 2\eta\beta(\epsilon_0+\epsilon)(1-\eta\alpha^2)^{t-1}\|\mathbf{r}_0\|_2,
\end{aligned}$$

where the last inequality follows from $\left\|\mathbf{B}_1\mathbf{B}_2 - \mathbf{J}\mathbf{J}^T\right\| \leq 2\beta(\epsilon_0+\epsilon)$, by (44), and from using the fact that $\|\mathbf{r}_{t-1}\|_2 \leq (1-\eta\alpha^2)^{t-1}\|\mathbf{r}_0\|_2$ which holds based on Step II. Finally, plugging in the induction hypothesis $\|\mathbf{e}_{t-1}\|_2 \leq c\xi^{t-2}(t-1)\|\mathbf{r}_0\|_2$ with $\xi := 1-\eta\alpha^2$ and $c := 2\eta\beta(\epsilon_0+\epsilon)$ in the above we conclude that

$$\begin{aligned}
\|\mathbf{e}_t\| \leq& \xi\|\mathbf{e}_{t-1}\| + c\xi^{t-1}\|\mathbf{r}_0\|_2 \\
\leq& c\xi^{t-1}(t-1)\|\mathbf{r}_0\|_2 + c\xi^{t-1}\|\mathbf{r}_0\|_2 \\
=& c\xi^{t-1}t\|\mathbf{r}_0\|_2 \\
=& 2\eta\beta(\epsilon_0+\epsilon)\left(1-\eta\alpha^2\right)^{t-1}t\|\mathbf{r}_0\|_2.
\end{aligned}$$

This concludes the proof of the bound (47) for iteration t, finishing Step III.

Step IV: Original and linearized parameters are close: The difference between the parameter of the original iterate $\boldsymbol{\theta}$ and the linearized iterate $\widetilde{\boldsymbol{\theta}}$ obey

$$\begin{aligned}
\frac{1}{\eta}\left\|\boldsymbol{\theta}_t - \widetilde{\boldsymbol{\theta}}_t\right\|_2 &\leq \left\|\sum_{\tau=0}^{t-1} \nabla\mathcal{L}(\boldsymbol{\theta}_\tau) - \nabla\mathcal{L}_{\text{lin}}(\widetilde{\boldsymbol{\theta}}_\tau)\right\|_2 \\
&= \left\|\sum_{\tau=0}^{t-1} \mathcal{J}^T(\boldsymbol{\theta}_\tau)\mathbf{r}_\tau - \mathbf{J}^T\widetilde{\mathbf{r}}_\tau\right\|_2 \\
&\leq \sum_{\tau=0}^{t-1} \left\|(\mathcal{J}^T(\boldsymbol{\theta}_\tau) - \mathbf{J}^T)\widetilde{\mathbf{r}}_\tau\right\|_2 + \left\|\mathcal{J}^T(\boldsymbol{\theta}_\tau)(\mathbf{r}_\tau - \widetilde{\mathbf{r}}_\tau)\right\|_2 \\
&\overset{\text{(i)}}{\leq} \sum_{\tau=0}^{t-1} (\epsilon_0 + \epsilon)\|\widetilde{\mathbf{r}}_\tau\|_2 + \beta\|\mathbf{e}_\tau\|_2 \\
&\overset{\text{(ii)}}{\leq} \sum_{\tau=0}^{t-1} (\epsilon_0 + \epsilon)(1-\eta\alpha^2)^\tau\|\mathbf{r}_0\|_2 + 2\eta\beta^2(\epsilon_0+\epsilon)(1-\eta\alpha^2)^{\tau-1}\tau\|\mathbf{r}_0\|_2.
\end{aligned}$$

Here, (i) follows from (45) combined with Assumption 1 and (ii) follows from (47) established in step III. We now proceed by using the formulas for low-order polylogarithms to conclude that

$$\begin{aligned}
\frac{1}{\eta}\left\|\boldsymbol{\theta}_t - \widetilde{\boldsymbol{\theta}}_t\right\|_2 &\leq (\epsilon_0+\epsilon)\|\mathbf{r}_0\|_2\left(\frac{1-(1-\eta\alpha^2)^\tau}{\eta\alpha^2} + 2\eta\beta^2\frac{1-t(1-\eta\alpha^2)^{t-1}+(t-1)(1-\eta\alpha^2)^t}{\eta^2\alpha^4}\right) \\
&\leq (\epsilon_0+\epsilon)\|\mathbf{r}_0\|_2\left(\frac{1}{\eta\alpha^2} + 2\eta\beta^2\frac{1}{\eta^2\alpha^4}\right) \\
&\leq (\epsilon_0+\epsilon)\frac{2.5}{\eta\alpha^2}\frac{\beta^2}{\alpha^2}\|\mathbf{r}_0\|_2.
\end{aligned}$$

This concludes the proof of (37) for iteration t, completing Step IV.

Step V: Proof of (38): By the triangle inequality

$$\begin{aligned}
\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2 &\leq \left\|\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0\right\|_2 + \left\|\boldsymbol{\theta}_t - \widetilde{\boldsymbol{\theta}}_t\right\|_2 \\
&\overset{\text{(i)}}{\leq} \left\|\mathbf{J}^\dagger\mathbf{r}_0\right\|_2 + (\epsilon_0+\epsilon)\frac{2.5}{\alpha^2}\frac{\beta^2}{\alpha^2}\|\mathbf{r}_0\|_2 \\
&\overset{\text{(ii)}}{=} R/2.
\end{aligned}$$

Here, inequality (ii) follows from the definition of R in equation (33). Moreover, inequality (i) follows from the bound (37), which we just proved, and the fact that, from equation (40) in Theorem 2,

$$\begin{aligned}
\left\|\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0\right\|_2^2 &= \sum_{i=1}^n \langle\mathbf{w}_i, \mathbf{r}_0\rangle^2 \frac{(1-(1-\eta\sigma_i^2)^t)^2}{\sigma_i^2} \\
&\leq \sum_{i=1}^n \langle\mathbf{w}_i, \mathbf{r}_0\rangle^2 / \sigma_i^2 \\
&= \left\|\mathbf{J}^\dagger\mathbf{r}_0\right\|^2.
\end{aligned}$$

This concludes the proof of (38) for iteration t, completing the proof of Step V and the entire theorem.

D Proofs for neural network generators (proof of Theorem 3)

The proof of Theorem 3 relies on the fact that, in the overparameterized regime, the non-linear least squares problem is well approximated by an associated linearized least-squares problem. Studying the associated linear problem enables us to prove the result.

We apply Theorem 4, which ensures that the associated linear problem is a good approximation of the non-linear least squarest problem, with the non-linear function

$$f(\mathbf{C}) = \mathbf{A}\mathrm{ReLU}(\mathbf{UC})\mathbf{v}$$

and with the parameter given by $\boldsymbol{\theta} = \mathbf{C}$. Recall that $\mathbf{v}$ is a fixed vector with half of the entries $1/\sqrt{k}$, and the other half $-1/\sqrt{k}$. Let $\mathcal{J}(\mathbf{C}) \in \mathbb{R}^{m\times nk}$ be the Jacobian of f. We have that $\mathcal{J}(\mathbf{C}) = \mathbf{A}\mathcal{J}_G(\mathbf{C})$, where $\mathcal{J}_G(\mathbf{C})$ is the Jacobian of the generator G defined in (20). Both f and its Jacobian are random variables because $\mathbf{A}$ is a random matrix. As the reference Jacobian in the associated linear problem, we choose a matrix $\mathbf{J} = \mathbf{A}\mathbf{J}_G \in \mathbb{R}^{m\times nk}$ (specified later) that obeys

$$\mathbf{J}\mathbf{J}^T = \mathbb{E}\left[\mathcal{J}(\mathbf{C})\mathcal{J}^T(\mathbf{C})\right] = \mathbf{A}\underbrace{\mathbb{E}\left[\mathcal{J}_G(\mathbf{C})\mathcal{J}_G^T(\mathbf{C})\right]}_{\boldsymbol{\Sigma}(\mathbf{U})}\mathbf{A}^T.$$

Here, expectation is with respect to $\mathbf{C}$ with iid $\mathcal{N}(0,\omega^2)$ parameters, and *not* with respect to $\mathbf{A}$. We apply Theorem 4 with

$$\alpha = \frac{1}{3\sqrt{2}}\frac{\sqrt{n}}{\sqrt{m}}\sigma_n\left(\boldsymbol{\Sigma}(\mathbf{U})\right), \quad \beta = 2\frac{\sqrt{n}}{\sqrt{m}}\|\mathbf{U}\|, \quad \epsilon_0 = 2\beta\left(\frac{\log(\frac{2n}{\delta})}{k}\right)^{1/4}, \quad \epsilon = \frac{\xi}{16}\frac{\alpha^4}{\beta^3}, \quad \omega = \frac{\xi\|\mathbf{y}\|_2}{\beta\sqrt{m}}.$$

We next verify that the conditions of Theorem 4 are satisfied (specifically, Assumptions 1, 2, 3) by applying a series of Lemmas.

Throughout these proofs we use the fact that for a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ with i.i.d. $\mathcal{N}(0, \frac{1}{m})$ entries, the bounds

$$\sigma_{\min}(\mathbf{A}) \geq \frac{\sqrt{n}-(1+\eta)\sqrt{m}}{\sqrt{m}} \quad \text{and} \quad \|\mathbf{A}\| \leq \frac{\sqrt{n}+(1+\eta)\sqrt{m}}{\sqrt{m}}$$

hold with probability at least $1-2e^{-\frac{\eta^2}{2}m}$ which with $\eta = 1$ in turn implies that for $m \leq \frac{n}{9}$ we have

$$\sigma_{\min}(\mathbf{A}) \geq \frac{1}{3}\frac{\sqrt{n}}{\sqrt{m}} \quad \text{and} \quad \|\mathbf{A}\| \leq 2\frac{\sqrt{n}}{\sqrt{m}} \tag{48}$$

holds with probability at least $1-2e^{-\frac{m}{2}}$. See [Ver12, Corollary 5.35] for a proof of this standard result.

Bound on initial residual: We start with bounding the initial residual by applying the following lemma.

Lemma 1 (Initial residual [HS20, Lemma 6]). *Consider $G(\mathbf{C}) = \mathrm{ReLU}(\mathbf{UC})\mathbf{v}$, and let $\mathbf{C} \in \mathbb{R}^{n\times k}$ be generated at random with i.i.d. $\mathcal{N}(0,\omega^2)$ entries. Suppose half of the entries of $\mathbf{v}$ are $1/\sqrt{k}$ and the other half are $-1/\sqrt{k}$. Then, with probability at least $1-\delta$,*

$$\|G(\mathbf{C})\|_2 \leq \omega\sqrt{8\log(2n/\delta)}\|\mathbf{U}\|_F.$$

With this lemma in place, the initial residual can be upper bounded as follows

$$\begin{aligned}\|\mathbf{r}_0\|_2 &\leq \|\mathbf{y}\|_2 + \|\mathbf{A}G(\mathbf{C}_0)\|_2 \\ &\overset{\text{(i)}}{\leq} \|\mathbf{y}\|_2 + 2\|G(\mathbf{C}_0)\|_2 \\ &\overset{\text{(ii)}}{\leq} 3\|\mathbf{y}\|_2. \end{aligned} \tag{49}$$

Here (i) holds with probability at least $1-e^{-\frac{m}{2}}$ using the fact that $\mathbf{A}$ has i.i.d. Gaussian entries that are independent of $G(\mathbf{C}_0)$, and for (ii) we used that, by Lemma 1,

$$\begin{aligned}\|G(\mathbf{C}_0)\|_2 &\leq \omega\sqrt{8\log(2n/\delta)}\|\mathbf{U}\|_F \\ &\leq \omega\sqrt{8\log(2n/\delta)}\sqrt{n}\|\mathbf{U}\| \\ &\overset{\text{(i)}}{=} \xi\sqrt{2\log(2n/\delta)}\|\mathbf{y}\|_2 \\ &\overset{\text{(ii)}}{\leq} \|\mathbf{y}\|_2, \end{aligned} \tag{50}$$

where (i) follows from $\omega = \frac{\xi\|\mathbf{y}\|_2}{\beta\sqrt{m}} = \frac{\xi\|\mathbf{y}\|_2}{2\sqrt{n}\|\mathbf{U}\|}$ and for (ii) we used the fact that $\xi \leq \frac{1}{\sqrt{2\log(2n/\delta)}}$.

Verifying Assumption 1: Note that

$$\sigma_{\min}(\mathbf{J}) = \sigma_{\min}(\mathbf{A}\mathbf{J}_G) \geq \sigma_{\min}(\mathbf{A})\sigma_{\min}(\mathbf{J}_G) \geq \frac{1}{3}\frac{\sqrt{n}}{\sqrt{m}}\sigma_n(\mathbf{\Sigma}(\mathbf{U})) \geq \alpha\sqrt{2}.$$

We next show that the norm of the reference Jacobian and the Jacobian are bounded, with the lemma below.

Lemma 2 (Spectral norm of Jacobian [HS20, Lemma 5]). *Consider $G(\mathbf{C}) = \mathrm{ReLU}(\mathbf{UC})\mathbf{v}$ with $\mathbf{v} \in \mathbb{R}^k$ and $\mathbf{U} \in \mathbb{R}^{n\times k}$ and associated Jacobian $\mathcal{J}_G(\mathbf{C})$ (20), and let $\mathbf{J}_G$ be any matrix obeying $\mathbf{J}_G\mathbf{J}_G^T = \mathbb{E}\left[\mathcal{J}_G(\mathbf{C})\mathcal{J}_G^T(\mathbf{C})\right]$, where the expectation is over a matrix $\mathbf{C}$ with iid $\mathcal{N}(0,\omega^2)$ entries. Then*

$$\|\mathcal{J}_G(\mathbf{C})\| \leq \|\mathbf{v}\|_2\|\mathbf{U}\| \quad \textit{and} \quad \|\mathbf{J}_G\| \leq \|\mathbf{v}\|_2\|\mathbf{U}\|.$$

By Lemma 2, with $\|\mathbf{v}\|_2 = 1$,

$$\|\mathbf{J}\| = \|\mathbf{A}\mathbf{J}_G\| \leq \|\mathbf{A}\|\|\mathbf{J}_G\| \leq 2\sqrt{n/m}\|\mathbf{U}\| = \beta,$$

where the last inequality follows from Lemma 2, with $\|\mathbf{v}\|_2 = 1$, and by using that, with high probability, $\|\mathbf{A}\| \leq 2\sqrt{n/m}$ per (48). Analogously, we obtain $\|\mathcal{J}(\mathbf{C})\| \leq \beta$, for all $\mathbf{C}$, with high probability. This completes the verification of Assumption 1.

Verifying Assumption 2: To verify the assumption, we first state a concentration lemma from [HS20].

Lemma 3 (Concentration lemma [HS20, Lemma 3]). *Consider* $G(\mathbf{C}) = \mathrm{ReLU}(\mathbf{UC})\mathbf{v}$ *with* $\mathbf{v} \in \mathbb{R}^k$ *and* $\mathbf{U} \in \mathbb{R}^{n\times k}$ *and associated Jacobian* $\mathcal{J}_G(\mathbf{C})$ (20). *Let* $\mathbf{C} \in \mathbb{R}^{n\times k}$ *be generated at random with i.i.d.* $\mathcal{N}(0, \omega^2)$ *entries. Then, with probability at least* $1 - \delta$,

$$\left\|\mathcal{J}_G(\mathbf{C})\mathcal{J}_G^T(\mathbf{C}) - \mathbf{\Sigma}(\mathbf{U})\right\| \leq \|\mathbf{U}\|^2 \sqrt{\log\left(\frac{2n}{\delta}\right)\sum_{\ell=1}^{k} v_\ell^4}.$$

Using the fact that $\sum_\ell^k v_\ell^4 = \frac{1}{k}$ by Lemma 3 we have

$$\left\|\mathcal{J}_G(\mathbf{C}_0)\mathcal{J}_G^T(\mathbf{C}_0) - \mathbf{\Sigma}(\mathbf{U})\right\| \leq \|\mathbf{U}\|^2 \sqrt{\frac{\log\left(2n/\delta\right)}{k}}. \tag{51}$$

To show that (51) implies the condition in (29), we use the following lemma.

Lemma 4 ([Oym+19, Lem. 6.4]). *Let* $\mathbf{X} \in \mathbb{R}^{n\times N}$, $N \geq n$ *and let* $\mathbf{\Sigma}$ *be* $n \times n$ *psd matrix obeying* $\left\|\mathbf{XX}^T - \mathbf{B}\right\| \leq \tilde{\epsilon}^2$, *for a scalar* $\tilde{\epsilon} \geq 0$. *Then there exists a matrix* $\mathbf{J}_G \in \mathbb{R}^{n\times N}$ *obeying* $\mathbf{\Sigma} = \mathbf{J}_G\mathbf{J}_G^T$ *such that*

$$\|\mathbf{J}_G - \mathbf{X}\| \leq \tilde{2}\epsilon.$$

From Lemma 4 combined with equation (51), we have that there exists a matrix $\mathbf{J}_G \in \mathbb{R}^{n\times N}$ that obeys

$$\|\mathbf{J}_G - \mathcal{J}_G(\mathbf{C}_0)\| \leq 2\|\mathbf{U}\| \left(\frac{\log(2n/\delta)}{k}\right)^{1/4}.$$

Using this inequality, as well as that $\|\mathbf{A}\| \leq 2\frac{\sqrt{n}}{\sqrt{m}}$, per (48), we get

$$\begin{aligned}
\|\mathbf{J} - \mathcal{J}(\mathbf{C}_0)\| &= \|\mathbf{A}(\mathbf{J}_G - \mathcal{J}_G(\mathbf{C}_0))\| \\
&\leq \|\mathbf{A}\|\|\mathbf{J}_G - \mathcal{J}_G(\mathbf{C}_0)\| \\
&\leq 2\frac{\sqrt{n}}{\sqrt{m}} 2\|\mathbf{U}\| \left(\frac{\log(2n/\delta)}{k}\right)^{1/4} \\
&\leq 2\beta \left(\frac{\log(2n/\delta)}{k}\right)^{1/4} \\
&= \epsilon_0,
\end{aligned}$$

as desired. This concludes the proof of Assumption 2.

This part of the proof also specifies our choice of the reference Jacobian $\mathbf{J} = \mathbf{A}\mathbf{J}_G$ as a matrix that is ϵ_0 close to the Jacobian at initialization, $\mathcal{J}(\mathbf{C}_0)$, and that exists by Lemma 4 above.

Verifying Assumption 3: Verification of the assumption requires us to control the perturbation of the Jacobian matrix around a random initialization. We begin with the following lemma from [HS20].

Lemma 5 (Jacobian perturbation around initialization [HS20, Lemma 7])**.** *Let* $\mathbf{C}_0$ *be a matrix with i.i.d.* $N(0, \omega^2)$ *entries. Then, for all* $\mathbf{C}$ *obeying*

$$\|\mathbf{C} - \mathbf{C}_0\| \leq \omega \widetilde{R} \quad \textit{with} \quad \widetilde{R} \leq \frac{1}{2}\sqrt{k},$$

the Jacobian mapping (20)*associated with the generator* $G(\mathbf{C}) = \mathrm{ReLU}(\mathbf{UC})\mathbf{v}$ *obeys*

$$\|\mathcal{J}_G(\mathbf{C}) - \mathcal{J}_G(\mathbf{C}_0)\| \leq \|\mathbf{v}\|_\infty 2(k\widetilde{R})^{1/3}\|\mathbf{U}\|,$$

with probability at least $1 - ne^{-\frac{1}{2}\widetilde{R}^{4/3}k^{7/3}}$.

In order to verify Assumption 3, first note that the radius in the theorem, defined in equation (33), obeys

$$\begin{aligned}
R &= 2\left\|\mathbf{J}^\dagger \mathbf{r}_0\right\|_2 + 5\frac{\beta^2}{\alpha^4}(\epsilon_0 + \epsilon)\|\mathbf{r}_0\|_2 \\
&\overset{\text{(i)}}{\leq} \left(\frac{\sqrt{2}}{\alpha} + \frac{5}{8\beta}\right)\|\mathbf{r}_0\|_2 \\
&\overset{\text{(ii)}}{\leq} 9\frac{1}{\alpha}\|\mathbf{y}\|_2 \\
&\overset{\text{(iii)}}{=} 9\omega\frac{\sqrt{m}}{\xi}\frac{\beta}{\alpha} \\
&\overset{\text{(iv)}}{\leq} \omega\frac{1}{(4\cdot 16)^3}\xi^3\frac{\alpha^{12}}{\beta^{12}}\sqrt{k} \\
&:= \omega\widetilde{R}.
\end{aligned}$$

Here, (i) follows from the fact that $\left\|\mathbf{J}^\dagger \mathbf{r}_0\right\|_2 \leq \frac{1}{\alpha\sqrt{2}}\|\mathbf{r}_0\|_2$, and using that $\epsilon_0 + \epsilon \leq 2\epsilon = \frac{1}{8}\xi\frac{\alpha^4}{\beta^3} \leq \frac{1}{8}\xi\frac{\alpha^4}{\beta^3}$ (ii) from $\beta \geq \alpha$ and from the bound on the initial residual (49), (iii) from $\omega = \frac{\xi\|\mathbf{y}\|}{\beta\sqrt{m}}$ and finally (iv) follows from the assumption (21) which is equivalent to

$$k \geq m\xi^{-8}9^2 64^6\left(\frac{\beta}{\alpha}\right)^{26} = m\xi^{-8}9^2 64^6\left(6\sqrt{2}\frac{\|\mathbf{U}\|^2}{\sigma_n^2}\right)^{26} = C^2\frac{\kappa_{\mathbf{u}}^{26}}{\xi^8}m.$$

For this choice of radius by Lemma 5 and by using $\|\mathbf{A}\| \leq 2\frac{\sqrt{n}}{\sqrt{m}}$ (per (48)) we have

$$\begin{aligned}\|\mathbf{A}\mathcal{J}_G(\mathbf{C}) - \mathbf{A}\mathcal{J}_G(\mathbf{C}_0)\| \leq & \|\mathbf{A}\|\|\mathcal{J}_G(\mathbf{C}) - \mathcal{J}_G(\mathbf{C}_0)\| \\ \leq & 2\frac{\sqrt{n}}{\sqrt{m}} 2\|\mathbf{v}\|_\infty (k\widetilde{R})^{1/3}\|\mathbf{U}\| \\ = & 2\beta\frac{1}{\sqrt{k}}(k\widetilde{R})^{1/3} \\ = & \frac{1}{32}\xi\frac{\alpha^4}{\beta^3} \\ = & \frac{\epsilon}{2}\end{aligned}$$

holds with probability at least

$$1 - ne^{-\frac{1}{2}\widetilde{R}^{4/3}k^{7/3}} \overset{\text{(i)}}{\geq} 1 - ne^{-k^2}$$

where in (i) we used (21). Therefore, Assumption 3 holds with high probability by our choice of $\epsilon = \frac{\xi}{30}\frac{\alpha^4}{\beta^3}$.

Concluding the proof of Theorem 3: To begin, let $\mathbf{c}^*$ be a solution to the optimization problem

$$\mathbf{c}^* \in \arg\min_{\mathbf{c}} \quad \frac{1}{2}\|\mathbf{J}_G\mathbf{c} - \mathbf{x}^*\|_2^2.$$

To complete the proof of Theorem 3 let us consider the linearized optimization problem which takes the form

$$\min_{\mathbf{c}} \mathcal{L}_{\text{lin}}(\mathbf{c}) = \frac{1}{2}\|\mathbf{A}G(\mathbf{c}_0) + \mathbf{A}\mathbf{J}_G(\mathbf{c} - \mathbf{c}_0) - \mathbf{y}\|_2^2,$$

with corresponding iterates given by

$$\widetilde{\mathbf{c}}_{t+1} = \widetilde{\mathbf{c}}_t - \eta\nabla\mathcal{L}_{\text{lin}}(\widetilde{\mathbf{c}}_t).$$

Here, $\mathbf{c}$ is the vectorized version of $\mathbf{C}$, with a slight abuse of notation. With this notation, we conclude the proof as

$$\begin{aligned}\|G(\mathbf{C}_\infty) - \mathbf{x}^*\|_2 \leq & \|G(\mathbf{C}_\infty) - \mathbf{J}_G\widetilde{\mathbf{c}}_\infty\|_2 + \|\mathbf{J}_G\widetilde{\mathbf{c}}_\infty - \mathbf{x}^*\|_2 \\ \leq & \xi\|\mathbf{x}^*\|_2 + C\left(\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\langle\mathbf{w}_i, \mathbf{x}^*\rangle^2\right)\sum_{i>2m/3}\sigma_i^2,\end{aligned}$$

where we used the bounds

$$\|G(\mathbf{C}_\infty) - \mathbf{J}_G\widetilde{\mathbf{c}}_\infty\|_2 \leq \xi\|\mathbf{x}^*\|_2 \tag{52}$$

and

$$\|\mathbf{J}_G\widetilde{\mathbf{c}}_\infty - \mathbf{x}^*\|_2 \leq C\left(\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\left\langle \mathbf{w}_i, \mathbf{x}^*\right\rangle^2\right)\sum_{i>2m/3}\sigma_i^2. \tag{53}$$

The bound (53) follows from Theorem 1 by noting that $\mathbf{w}_1, \ldots, \mathbf{w}_n$ are the left singular vectors of $\mathbf{J}_G$ with associated singular values $\sigma_1 \geq \ldots \geq \sigma_n$ (because $\mathbf{J}_G\mathbf{J}_G^T = \mathbf{\Sigma}(\mathbf{U})$).

It remains to prove the bound (52). With $\mathcal{J}_G(\mathbf{a}, \mathbf{b}) = \int_0^1 \mathcal{J}_G(s\mathbf{b} - (1-s)\mathbf{a})ds$, at $t = +\infty$,

$$\begin{aligned}
\|G(\mathbf{C}_\infty) - \mathbf{J}_G\widetilde{\mathbf{c}}_\infty\|_2 &= \|\mathcal{J}_G(\mathbf{C}_\infty, \mathbf{0})\mathrm{vect}(\mathbf{C}_\infty) - \mathbf{J}_G\widetilde{\mathbf{c}}_\infty\|_2 \\
&\leq \|\mathcal{J}_G(\mathbf{C}_\infty, \mathbf{0})(\mathrm{vect}(\mathbf{C}_\infty) - \widetilde{\mathbf{c}}_\infty)\|_2 + \|\mathcal{J}_G(\mathbf{C}_\infty, \mathbf{0})\widetilde{\mathbf{c}}_\infty - \mathbf{J}_G\widetilde{\mathbf{c}}_\infty\|_2 \\
&\leq \|\mathcal{J}_G(\mathbf{C}_\infty, \mathbf{0})\|\|\mathrm{vect}(\mathbf{C}_\infty) - \widetilde{\mathbf{c}}_\infty\|_2 + \|\mathcal{J}_G(\mathbf{C}_\infty, \mathbf{0}) - \mathbf{J}_G\|_2\|\widetilde{\mathbf{c}}_\infty\|_2 \\
&\overset{\text{(i)}}{\leq} \frac{\sqrt{m}}{\sqrt{n}}\frac{\beta}{2}\|\mathrm{vect}(\mathbf{C}_\infty) - \widetilde{\mathbf{c}}_\infty\|_2 + \frac{\sqrt{m}}{2\sqrt{n}}(\epsilon + \epsilon_0)\|\widetilde{\mathbf{c}}_\infty\|_2 \\
&\overset{\text{(ii)}}{\leq} \frac{\beta}{2}\|\mathrm{vect}(\mathbf{C}_\infty) - \widetilde{\mathbf{c}}_\infty\|_2 + \frac{1}{2}(\epsilon + \epsilon_0)\frac{1}{\alpha}\|\mathbf{x}^*\|_2.
\end{aligned}$$

In the above (i) follows from $\|\mathbf{J}_G(\mathbf{C})\| \leq \|\mathbf{U}\| = \frac{\sqrt{m}}{2\sqrt{n}}\beta$ (recall that $\beta = 2\frac{\sqrt{n}}{\sqrt{m}}\|\mathbf{U}\|$) and from the bound

$$\|\mathcal{J}_G(\mathbf{C}_\infty, \mathbf{0}) - \mathbf{J}_G\|_2 \leq \|\mathcal{J}_G(\mathbf{C}_\infty, \mathbf{0}) - \mathcal{J}_G(\mathbf{C}_0)\|_2 + \|\mathcal{J}_G(\mathbf{C}_0) - \mathbf{J}_G\|_2 \leq \frac{\sqrt{m}}{2\sqrt{m}}(\epsilon_0 + \epsilon).$$

Moreover, (ii) follows from $m \leq n$ and $\|\mathbf{c}_\infty\|_2 \leq \|\mathbf{c}^*\|_2 \leq \frac{\|\mathbf{x}^*\|_2}{\sigma_{\min}(\mathbf{J}_G)}$. We can now apply Theorem 4 equation (37) to bound the first term on the right-hand-side above to obtain

$$\begin{aligned}
\|G(\mathbf{C}_\infty) - \mathbf{J}_G\widetilde{\mathbf{c}}_\infty\|_2 &\leq 1.25\frac{\beta^3}{\alpha^4}(\epsilon_0 + \epsilon)\|\mathbf{r}_0\|_2 + \frac{1}{2\alpha}(\epsilon + \epsilon_0)\|\mathbf{x}^*\|_2 \\
&\overset{\text{(i)}}{\leq} 7.5\frac{\beta^3}{\alpha^4}(\epsilon_0 + \epsilon)\|\mathbf{x}^*\|_2 + \frac{1}{2\alpha}(\epsilon + \epsilon_0)\|\mathbf{x}^*\|_2 \\
&\overset{\text{(ii)}}{\leq} 16\frac{\beta^3}{\alpha^4}\epsilon\|\mathbf{x}^*\|_2 \\
&\overset{\text{(iii)}}{=} \xi\|\mathbf{x}^*\|_2.
\end{aligned}$$

Here, (i) follows from $\|\mathbf{r}_0\|_2 \leq 3\|\mathbf{y}\|_2 = 3\|\mathbf{A}\mathbf{x}^*\|_2 \leq 6\|\mathbf{x}^*\|$, where we used (49) combined with the fact that $\|\mathbf{A}\mathbf{x}^*\|_2 \leq 2\|\mathbf{x}^*\|_2$. Moreover, (ii) follows from $\frac{\beta}{\alpha} \geq 1$ and $\epsilon_0 \leq \epsilon$ and finally (iii) from the choice $\epsilon = \frac{\xi}{16}\frac{\alpha^4}{\beta^3}$. This concludes the proof of the bound (52) and the proof of the theorem.