

---

# Global Guarantees for Blind Demodulation with Generative Priors

---

**Paul Hand**

Dept. of Mathematics and College of Computer Science and Information  
Northeastern University, MA  
p.hand@northeastern.edu

**Babhru Joshi**

Dept. of Mathematics  
University of British Columbia, BC  
b.joshi@math.ubc.ca

## Abstract

We study a deep learning inspired formulation for the blind demodulation problem, which is the task of recovering two unknown vectors from their entrywise multiplication. We consider the case where the unknown vectors are in the range of known deep generative models,  $\mathcal{G}^{(1)} : \mathbb{R}^n \rightarrow \mathbb{R}^\ell$  and  $\mathcal{G}^{(2)} : \mathbb{R}^p \rightarrow \mathbb{R}^\ell$ . In the case when the networks corresponding to the generative models are expansive, the weight matrices are random and the dimension of the unknown vectors satisfy  $\ell = \Omega(n^2 + p^2)$ , up to log factors, we show that the empirical risk objective has a favorable landscape for optimization. That is, the objective function has a descent direction at every point outside of a small neighborhood around four hyperbolic curves. We also characterize the local maximizers of the empirical risk objective and, hence, show that there does not exist any other stationary points outside of these neighborhood around four hyperbolic curves and the set of local maximizers. We also implement a gradient descent scheme inspired by the geometry of the landscape of the objective function. In order to converge to a global minimizer, this gradient descent scheme exploits the fact that exactly one of the hyperbolic curve corresponds to the global minimizer, and thus points near this hyperbolic curve have a lower objective value than points close to the other spurious hyperbolic curves. We show that this gradient descent scheme can effectively remove distortions synthetically introduced to the MNIST dataset.

## 1 Introduction

We study the problem of recovering two unknown vectors  $\mathbf{x}_0 \in \mathbb{R}^\ell$  and  $\mathbf{w}_0 \in \mathbb{R}^\ell$  from observations  $\mathbf{y}_0 \in \mathbb{R}^\ell$  of the form

$$\mathbf{y}_0 = \mathbf{w}_0 \odot \mathbf{x}_0, \tag{1}$$

where  $\odot$  is entrywise multiplication. This bilinear inverse problem (BIP) is known as the blind demodulation problem. BIPs, in general, have been extensively studied and include problems such as blind deconvolution/demodulation [Ahmed et al., 2014, Stockham et al., 1975, Kundur and Hatzinakos, 1996, Aghasi et al., 2016, 2019], phase retrieval [Fienup, 1982, Candès and Li, 2012, Candès et al., 2013], dictionary learning [Tosic and Frossard, 2011], matrix factorization [Hoyer, 2004, Lee and Seung, 2001], and self-calibration [Ling and Strohmer, 2015]. A significant challenge of BIP is the ambiguity of solutions. These ambiguities are challenging because they cause the set of solutions to be non-convex.

A common ambiguity, also shared by the BIP in (1), is the scaling ambiguity. That is any member of the set  $\{c\mathbf{w}_0, \frac{1}{c}\mathbf{x}_0\}$  for  $c \neq 0$  solves (1). In addition to the scaling ambiguity, this BIP is difficult to solve because the solutions are non-unique, even when excluding the scaling ambiguity. For example,  $(\mathbf{w}_0, \mathbf{x}_0)$  and  $(\mathbf{1}, \mathbf{w}_0 \odot \mathbf{x}_0)$  both satisfy (1). This structural ambiguity can be solved by assuming a prior model of the unknown vectors. In past works relating to blind deconvolution and blind demodulation [Ahmed et al., 2014, Aghasi et al., 2019], this structural ambiguity issue was addressed by assuming a subspace prior, i.e. the unknown signals belong to known subspaces. Additionally, in many applications, the signals are compressible or sparse with respect to a basis like a wavelet basis or the Discrete Cosine Transform basis, which can address this structural ambiguity issue.

In contrast to subspace and sparsity priors, we address the structural ambiguity issue by assuming the signals  $\mathbf{w}_0$  and  $\mathbf{x}_0$  belong to the range of known generative models  $\mathcal{G}^{(1)} : \mathbb{R}^n \rightarrow \mathbb{R}^\ell$  and  $\mathcal{G}^{(2)} : \mathbb{R}^p \rightarrow \mathbb{R}^\ell$ , respectively. That is, we assume that  $\mathbf{w}_0 = \mathcal{G}^{(1)}(\mathbf{h}_0)$  for some  $\mathbf{h}_0 \in \mathbb{R}^n$  and  $\mathbf{x}_0 = \mathcal{G}^{(2)}(\mathbf{m}_0)$  for some  $\mathbf{m}_0 \in \mathbb{R}^p$ . So, to recover the unknown vectors  $\mathbf{w}_0$  and  $\mathbf{x}_0$ , we first recover the latent code variables  $\mathbf{h}_0$  and  $\mathbf{m}_0$  and then apply  $\mathcal{G}^{(1)}$  and  $\mathcal{G}^{(2)}$  on  $\mathbf{h}_0$  and  $\mathbf{m}_0$ , respectively. Thus, the blind demodulation problem under generative prior we study is:

$$\text{find } \mathbf{h} \in \mathbb{R}^n \text{ and } \mathbf{m} \in \mathbb{R}^p, \text{ up to the scaling ambiguity, such that } \mathbf{y}_0 = \mathcal{G}^{(1)}(\mathbf{h}) \odot \mathcal{G}^{(2)}(\mathbf{m}).$$

In recent years, advances in generative modeling of images [Karras et al., 2017] has significantly increased the scope of using a generative model as a prior in inverse problems. Generative models are now used in speech synthesis [van den Oord et al., 2016], image in-painting [Iizuka et al., 2017], image-to-image translation [Zhu et al., 2017], superresolution [Sønderby et al., 2017], compressed sensing [Bora et al., 2017, Lohit et al., 2018], blind deconvolution [Asim et al., 2018], blind ptychography [Shamshad et al., 2018], and in many more fields. Most of these papers empirically show that using generative model as a prior to solve inverse problems outperform classical methods. For example, in compressed sensing, optimization over the latent code space to recover images from its compressive measurements have been empirically shown to succeed with 10x fewer measurements than classical sparsity based methods [Bora et al., 2017]. Similarly, the authors of Asim et al. [2018] empirically show that using generative priors in image deblurring inverse problem provide a very effective regularization that produce sharp deblurred images from very blurry images.

In the present paper, we use generative priors to solve the blind demodulation problem (1). The generative model we consider is the an expansive, fully connected, feed forward neural network with Rectified Linear Unit (ReLU) activation functions and no bias terms. Our main contribution is we show that the empirical risk objective function, for a sufficiently expansive random generative model, has a landscape favorable for gradient based methods to converge to a global minimizer. Our result implies that if the dimension of the unknown signals satisfy  $\ell = \Omega(n^2 + p^2)$ , up to log factors, then the landscape is favorable. In comparison, classical sparsity based methods for similar BIPs like sparse blind demodulation [Lee et al., 2017] and sparse phase retrieval [Li and Voroninski, 2013] showed that exact recovery of the unknown signals is possible if the number of measurements scale quadratically, up to a log factor, w.r.t. the sparsity level of the signals. While we show a similar scaling of the number of measurements w.r.t. the latent code dimension, the latent code dimension can be smaller than the sparsity level for the same signal, and thus recovering the signal using generative prior would require less number of measurements.

## 1.1 Main results

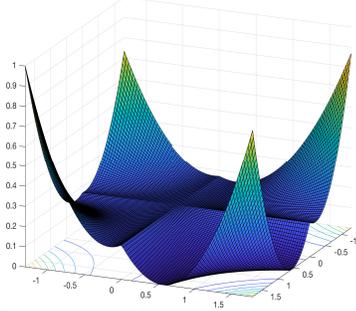
We study the problem of recovering two unknown signals  $\mathbf{w}_0$  and  $\mathbf{x}_0$  in  $\mathbb{R}^\ell$  from observations  $\mathbf{y}_0 = \mathbf{w}_0 \odot \mathbf{x}_0$ , where  $\odot$  denotes entrywise product. We assume, as a prior, that the vectors  $\mathbf{w}_0$  and  $\mathbf{x}_0$  belong to the range of  $d$ -layer and  $s$ -layer neural networks  $\mathcal{G}^{(1)} : \mathbb{R}^n \rightarrow \mathbb{R}^\ell$  and  $\mathcal{G}^{(2)} : \mathbb{R}^p \rightarrow \mathbb{R}^\ell$ , respectively. The task of recovering  $\mathbf{w}_0$  and  $\mathbf{x}_0$  is reduced to finding the latent codes  $\mathbf{h}_0 \in \mathbb{R}^n$  and  $\mathbf{m}_0 \in \mathbb{R}^p$  such that  $\mathcal{G}^{(1)}(\mathbf{h}_0) = \mathbf{w}_0$  and  $\mathcal{G}^{(2)}(\mathbf{m}_0) = \mathbf{x}_0$ . More precisely, we consider the generative networks modeled by  $\mathcal{G}^{(1)}(\mathbf{h}) = \text{relu}(\mathbf{W}_d^{(1)} \dots \text{relu}(\mathbf{W}_2^{(1)} \text{relu}(\mathbf{W}_1^{(1)} \mathbf{h})) \dots)$  and  $\mathcal{G}^{(2)}(\mathbf{m}) = \text{relu}(\mathbf{W}_s^{(2)} \dots \text{relu}(\mathbf{W}_2^{(2)} \text{relu}(\mathbf{W}_1^{(2)} \mathbf{m})) \dots)$ , where  $\text{relu}(\mathbf{x}) = \max(\mathbf{x}, \mathbf{0})$  applies entrywise,  $\mathbf{W}_i^{(1)} \in \mathbb{R}^{n_i \times n_{i-1}}$  for  $i = 1, \dots, d$  with  $n = n_0 < n_1 < \dots < n_d = \ell$ , and  $\mathbf{W}_i^{(2)} \in \mathbb{R}^{p_i \times p_{i-1}}$  for  $i = 1, \dots, s$  with  $p = p_0 < p_1 < \dots < p_s = \ell$ . The blind demodulation problem we consider is:

$$\text{Let: } \mathbf{y}_0 \in \mathbb{R}^\ell, \mathbf{h}_0 \in \mathbb{R}^n, \mathbf{m}_0 \in \mathbb{R}^p \text{ such that } \mathbf{y}_0 = \mathcal{G}^{(1)}(\mathbf{h}_0) \odot \mathcal{G}^{(2)}(\mathbf{m}_0),$$

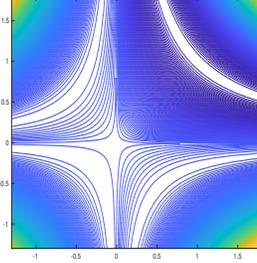
Given:  $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}$  and measurements  $\mathbf{y}_0$ ,  
 Find:  $\mathbf{h}_0$  and  $\mathbf{m}_0$ , up to the scaling ambiguity.

In order to recover  $\mathbf{h}_0$  and  $\mathbf{m}_0$ , up to the scaling ambiguity, we consider the following empirical risk minimization program:

$$\underset{\mathbf{h} \in \mathbb{R}^n, \mathbf{m} \in \mathbb{R}^p}{\text{minimize}} \quad f(\mathbf{h}, \mathbf{m}) := \frac{1}{2} \left\| \mathcal{G}^{(1)}(\mathbf{h}_0) \odot \mathcal{G}^{(2)}(\mathbf{m}_0) - \mathcal{G}^{(1)}(\mathbf{h}) \odot \mathcal{G}^{(2)}(\mathbf{m}) \right\|_2^2. \quad (2)$$



(a) Landscape of the empirical risk function.



(b) Note the four hyperbolic branches visible.

Figure 1: Plots showing the landscape of the objective function with  $\mathbf{h}_0 = 1$  and  $\mathbf{m}_0 = 1$ .

Figures 1a and 1b show the landscape of the objective function in the case when  $\mathbf{h}_0 = \mathbf{m}_0 = 1$ ,  $s = d = 2$ , the networks are expansive, and the weight matrices  $\mathbf{W}_i^{(1)}$  and  $\mathbf{W}_i^{(2)}$  contain i.i.d. Gaussian entries. Clearly, the objective function in (2) is non-convex and, as a result, there does not exist a prior guarantee that gradient based methods will converge to a global minima. Additionally, the objective function does not contain any regularizer which are generally be used to resolve the scaling ambiguity, and thus every point in  $\{(c\mathbf{h}_0, \frac{1}{c}\mathbf{m}_0) | c > 0\}$  is a global optima of (2). Nonetheless, we show that under certain conditions on the networks, the minimizers of (2) are in the neighborhood of four hyperbolic curves, one of which is the hyperbolic curve containing the global minimizers.

In order to define these hyperbolic neighborhoods, let

$$\mathcal{A}_{\epsilon, (\tilde{\mathbf{h}}, \tilde{\mathbf{m}})} = \left\{ (\mathbf{h}, \mathbf{m}) \in \mathbb{R}^{n \times p} \mid \exists c > 0 \text{ s.t. } \left\| (\mathbf{h}, \mathbf{m}) - \left( c\tilde{\mathbf{h}}, \frac{1}{c}\tilde{\mathbf{m}} \right) \right\|_2 \leq \epsilon \left\| \left( c\tilde{\mathbf{h}}, \frac{1}{c}\tilde{\mathbf{m}} \right) \right\|_2 \right\}, \quad (3)$$

where  $(\tilde{\mathbf{h}}, \tilde{\mathbf{m}}) \in \mathbb{R}^{n \times p}$  is fixed. This set is an  $\epsilon$ -neighborhood of the hyperbolic set  $\{(c\tilde{\mathbf{h}}, \frac{1}{c}\tilde{\mathbf{m}}) | c > 0\}$ . We show that the minimizers of (2) are contained in the four hyperbolic sets given by  $\mathcal{A}_{\epsilon, (\mathbf{h}_0, \mathbf{m}_0)}$ ,  $\mathcal{A}_{\epsilon, (-\rho_d^{(1)}\mathbf{h}_0, \mathbf{m}_0)}$ ,  $\mathcal{A}_{\epsilon, (\mathbf{h}_0, -\rho_s^{(2)}\mathbf{m}_0)}$ , and  $\mathcal{A}_{\epsilon, (-\rho_d^{(1)}\mathbf{h}_0, -\rho_s^{(2)}\mathbf{m}_0)}$ . Here,  $\epsilon$  depends on the expansivity and number of layers in the networks, and both  $\rho_d^{(1)}$  and  $\rho_s^{(2)}$  are positive constants close to 1. We also show that the points in the set  $\{(\mathbf{h}, \mathbf{0}) | \mathbf{h} \in \mathbb{R}^n\} \cup \{(\mathbf{0}, \mathbf{m}) | \mathbf{m} \in \mathbb{R}^p\}$  are local maximizers. This result holds for networks with the following assumptions:

- A1. The weight matrices are random.
- A2. The weight matrices of inner layers satisfy  $n_i \geq cn_{i-1} \log n_{i-1}$  for  $i = 1, \dots, d-1$  and  $p_i \geq cp_{i-1} \log p_{i-1}$  for  $i = 1, \dots, s-1$ .
- A3. The weight matrices of the last layer for each generator satisfy  $\ell \geq c((n_{d-1} \log n_{d-1})^2 + (p_{s-1} \log p_{s-1})^2)$ .

In the above assumptions,  $c$  is a constant that depends polynomially on the expansivity parameter of  $\mathcal{G}^{(1)}$  and  $\mathcal{G}^{(2)}$ . Figures 1a and 1b show the landscape of the objective function and corroborate our findings. In the paper we provide two deterministic conditions that are sufficient to characterize the landscape of the objective function, and show that Gaussian matrices satisfy these conditions. In essence, we only require approximate Gaussian matrices. We also note that the state-of-the-art literature for provable convergence of the training of neural networks for regression and classification admit proofs only in the case that the final trained weights are close to their random initialization.

Thus, our neural network assumptions are consistent with the best known cases for which networks can be provably trained.

**Theorem 1 (Informal).** *Let*

$$\mathcal{A} = \mathcal{A}_{\epsilon, (\mathbf{h}_0, \mathbf{m}_0)} \cup \mathcal{A}_{\epsilon, (-\rho_d^{(1)} \mathbf{h}_0, \mathbf{m}_0)} \cup \mathcal{A}_{\epsilon, (\mathbf{h}_0, -\rho_s^{(2)} \mathbf{m}_0)} \cup \mathcal{A}_{\epsilon, (-\rho_d^{(1)} \mathbf{h}_0, -\rho_s^{(2)} \mathbf{m}_0)},$$

where  $\epsilon > 0$  depends on the expansivity of our networks and  $\rho_d^{(1)}, \rho_s^{(2)} \rightarrow 1$  as  $d, s \rightarrow \infty$ , respectively. Suppose the networks are sufficiently expansive such that the number of neurons in the inner layers and the last layers satisfy assumptions A2 and A3, respectively. Then there exist a descent direction, given by one of the one-sided partial derivative of the objective function in (2), for every  $(\mathbf{h}, \mathbf{m}) \notin \mathcal{A} \cup \{(\mathbf{h}, \mathbf{0}) | \mathbf{h} \in \mathbb{R}^n\} \cup \{(\mathbf{0}, \mathbf{m}) | \mathbf{m} \in \mathbb{R}^p\}$  with high probability. In addition, elements of the set  $\{(\mathbf{h}, \mathbf{0}) | \mathbf{h} \in \mathbb{R}^n\} \cup \{(\mathbf{0}, \mathbf{m}) | \mathbf{m} \in \mathbb{R}^p\}$  are local maximizers.

Our main result states that the objective function in (2) does not have any spurious minimizers outside of the four hyperbolic neighborhoods. Thus, a gradient descent algorithm will converge to a point inside the four neighborhoods, one of which contains the global minimizers of (2). However, it may not guarantee convergence to a global minimizer and it may not resolve the inherent scaling ambiguity present in the problem. So, in order to converge to a global minimizer, we implement a gradient descent scheme that exploits the landscape of the objective function. That is, we exploit the fact that points near the hyperbolic curve corresponding to the global minimizer have a lower objective value than points that are close to the remaining three spurious hyperbolic curves. Second, in order to resolve the scaling ambiguity, we promote solutions that have equal  $\ell_2$  norm by normalizing the estimates in each iteration of the gradient descent scheme (See Section 2). In principle, a convergence result to a global minimizer by gradient descent is possible, and would require showing a convexity-like property around the hyperbola. We leave this for possible future work.

Theorem 1 also provides a global guarantee of the landscape of the objective function in (2) if the dimension of the unknown signals scale quadratically w.r.t. to the dimension of the latent codes, i.e.  $\ell = \Omega(n^2 + p^2)$ , up to log factors. Our result, which we get by enforcing generative priors may enjoy better sample complexity than classical priors like sparsity because: i) existing recovery guarantee of unstructured signals require number of measurements that scale quadratically with the sparsity level, and ii) a signal can have a latent code dimension with respect to a GAN that is smaller than its sparsity level with respect to a wavelet basis. For example, consider a set of images that correspond to a single train going down a single track. This set of images form a one dimensional sub-manifold of the manifold of natural images. If properly parameterized by a generative model, then it would have a latent dimensionality of approximately 1, whereas the number of wavelet coefficients needed to describe any of those images is much greater. The work in Bora et al. [2017] shows that compressed sensing can be done with 5-10x fewer measurements than sparsity models. This provides evidence for the more economical representation of generative models than of sparsity models. Additionally, it is more natural to view the natural signal manifold as a low-dimensional manifold, as opposed to being the combinatorially-many union of low dimensional spaces. Performance gains are provided by the fact that the natural signal manifold can be directly exploited, whereas the union of subspaces can only be indirectly exploited via convex relaxations. Thus, our result may be less limiting in terms of sample complexity.

## 1.2 Prior work on problems related to blind demodulation

A common approach of solving the BIP in (1) is to assume a subspace or sparsity prior on the unknown vectors. In these cases the unknown vectors  $\mathbf{w}_0$  and  $\mathbf{x}_0$  are assumed to be in the range of known matrices  $\mathbf{B} \in \mathbb{R}^{\ell \times n}$  and  $\mathbf{C} \in \mathbb{R}^{\ell \times p}$ , respectively. In Ahmed et al. [2014], the authors assumed a subspace prior and cast the BIP as a linear low rank matrix recovery problem. They introduced a semidefinite program based on nuclear norm minimization to recover the unknown matrix. For the case where the rows of  $\mathbf{B}$  and  $\mathbf{C}$  are Fourier and Gaussian vectors, respectively, they provide a recovery guarantee that depend on the number of measurements as  $\ell = \Omega(n + p)$ , up to log factors. However, because this method operates in the space of matrices, it is computationally prohibitively expensive. Another limitation of the lifted approach is that recovering a low rank and sparse matrix efficiently from linear observation of the matrix has been challenging. Recently, Lee et al. [2017] provided a recovery guarantee with near optimal sample complexity for the low rank and sparse matrix recovery problem using an alternating minimization method for a class of signals that satisfy a peakiness condition. However, for general signals the same work established a recovery result for the case where the number of measurements scale quadratically with the sparsity level.

In order to address the computational cost of working in the lifted case, a recent theme has been to introduce convex and non-convex programs that work in the natural parameter space. For example, in Bahmani and Romberg [2016], Goldstein and Studer [2016], the authors introduced PhaseMax, which is a convex program for phase retrieval that is based on finding a simple convex relaxation via the convex hull of the feasibility set. The authors showed that PhaseMax enjoys rigorous recovery guarantee if a good anchor is available. This formulation was extended to the sparse case in Hand and Voroninski [2016], where the authors considered SparsePhaseMax and provided a recovery guarantee with optimal sample complexity. The idea of formulating a convex program using a simple convex relaxation via the convex hull of the feasibility set was used in the blind demodulation problem as well [Aghasi et al., 2019, 2018]. In particular, Aghasi et al. [2018] introduced a convex program in the natural parameter space for the sparse blind demodulation problem in the case where the sign of the unknown signals are known. Like in Lee et al. [2017], the authors in Aghasi et al. [2019] provide a recovery guarantee with optimal sample complexity for a class of signals. However, the result does not extend to signals with no constraints. Other approaches that operate in the natural parameter space are methods based on Wirtinger Flow. For example, in Candès et al. [2015], Wang et al. [2016], Li et al. [2016], the authors use Wirtinger Flow and its variants to solve the phase retrieval and the blind deconvolution problem. These methods are non-convex and require a good initialization to converge to a global solution. However, they are simple to solve and enjoys rigorous recovery guarantees.

### 1.3 Other related work

In this paper, we consider the blind demodulation problem with the unknown signals assumed to be in the range of known generative models. Our work is motivated by experimental results in deep compressed sensing and deep blind deconvolution presented in Bora et al. [2017], Asim et al. [2018] and theoretical work in deep compressed sensing presented in Hand and Voroninski [2017]. In Bora et al. [2017], the authors consider the compressed sensing problem where, instead of a sparsity prior, a generative prior is considered. They used an empirical risk optimization program over the latent code space to recover images and empirical showed that their method succeeds with 10x fewer measurements than previous sparsity based methods. Following the empirical successes of deep compressed sensing, the authors in Hand and Voroninski [2017] provided a theoretical understanding for these successes by characterizing the landscape of the empirical risk objective function. In the random case with the layers of the generative model sufficiently expansive, they showed that every point outside of a small neighborhood around the true solution and a negative multiple of the true solution has a descent direction with high probability. Another instance where generative model currently outperforms sparsity based methods is in sparse phase retrieval Hand et al. [2018]. In sparse phase retrieval, current algorithms that enjoy a provable recovery guarantee of an unknown  $n$ -dimensional  $k$ -sparse signal require at least  $O(k^2 \log n)$  measurements; whereas, when assuming the unknown signal is an output of a known  $d$ -layer generator  $\mathcal{G} : \mathbb{R}^k \rightarrow \mathbb{R}^n$ , the authors in Hand et al. [2018] showed that, under favorable conditions on the generator and with at least  $O(kd^2 \log n)$  measurements, the empirical risk objective enjoys a favorable landscape.

Similarly, in Asim et al. [2018], the authors consider the blind deconvolution problem where a generative prior over the unknown signal is considered. They empirically showed that using generative priors in the image deblurring inverse problem provide a very effective regularization that produce sharp deblurred images from very blurry images. The algorithm used to recovery these deblurred images is an alternating minimization approach which solves the empirical risk minimization with  $\ell_2$  regularization on the unknown signals. The  $\ell_2$  regularization promotes solution with least  $\ell_2$  norm and resolves the scaling ambiguity present in the blind deconvolution problem. We consider a related problem, namely the blind demodulation problem with a generative prior on the unknown signals, and show that under certain conditions on the generators the empirical risk objective has a favorable landscape.

### 1.4 Notations

Vectors and matrices are written with boldface, while scalars and entries of vectors are written in plain font. We write  $\mathbf{1}$  as the vector of all ones with dimensionality appropriate for the context. Let  $\mathbb{S}^{n-1}$  be the unit sphere in  $\mathbb{R}^n$ . We write  $\mathbf{I}_n$  as the  $n \times n$  identity matrix. For  $\mathbf{x} \in \mathbb{R}^K$  and  $\mathbf{y} \in \mathbb{R}^N$ ,  $(\mathbf{x}, \mathbf{y})$  is the corresponding vector in  $\mathbb{R}^K \times \mathbb{R}^N$ . Let  $\text{relu}(\mathbf{x}) = \max(\mathbf{x}, \mathbf{0})$  apply entrywise for  $\mathbf{x} \in \mathbb{R}^n$ . Let  $\text{diag}(\mathbf{W}\mathbf{x} > 0)$  be the diagonal matrix that is 1 in the  $(i, i)$  entry if  $(\mathbf{W}\mathbf{x})_i > 0$  and 0 otherwise. Let  $\mathbf{A} \preceq \mathbf{B}$  mean that  $\mathbf{B} - \mathbf{A}$  is a positive semidefinite matrix. We will write  $\gamma = O(\delta)$  to mean that

there exists a positive constant  $C$  such that  $\gamma \leq C\delta$ , where  $\gamma$  is understood to be positive. Similarly we will write  $c = \Omega(\delta)$  to mean that there exists a positive constant  $C$  such that  $c \geq C\delta$ . When we say that a constant depends polynomially on  $\epsilon^{-1}$ , that means that it is at most  $C\epsilon^{-k}$  for some positive  $C$  and positive integer  $k$ . For notational convenience, we will write  $\mathbf{a} = \mathbf{b} + O_1(\epsilon)$  if  $\|\mathbf{a} - \mathbf{b}\| \leq \epsilon$ , where the norm is understood to be absolute value for scalars, the  $\ell_2$  norm for vectors, and the spectral norm for matrices.

## 2 Algorithm

In this section, we propose a gradient descent scheme that solves (2). The gradient descent scheme exploits the global geometry present in the landscape of the objective function in (2) and avoids regions containing spurious minimizers. The gradient descent scheme is based on two observations. The first observation is that the minimizers of (2) are close to four hyperbolic curves given by  $\{(c\mathbf{h}_0, \frac{1}{c}\mathbf{m}_0) | c > 0\}$ ,  $\{(-c\rho_d^{(1)}\mathbf{h}_0, \frac{1}{c}\mathbf{m}_0) | c > 0\}$ ,  $\{(c\mathbf{h}_0, -\frac{\rho_d^{(2)}}{c}\mathbf{m}_0) | c > 0\}$ , and  $\{(-c\rho_d^{(1)}\mathbf{h}_0, -\frac{\rho_s^{(2)}}{c}\mathbf{m}_0) | c > 0\}$ , where  $\rho_d^{(1)}$  and  $\rho_s^{(2)}$  are close to 1. The second observation is that  $f(c\mathbf{h}_0, \frac{1}{c}\mathbf{m}_0)$  is less than  $f(-c\mathbf{h}_0, \frac{1}{c}\mathbf{m}_0)$ ,  $f(c\mathbf{h}_0, -\frac{1}{c}\mathbf{m}_0)$ , and  $f(-c\mathbf{h}_0, -\frac{1}{c}\mathbf{m}_0)$  for any  $c > 0$ . This is because the curve  $\{(c\mathbf{h}_0, \frac{1}{c}\mathbf{m}_0) | c > 0\}$  corresponds to the global minimizer of (2).

We now introduce some quantities which are useful in stating the gradient descent algorithm. For any  $\mathbf{h} \in \mathbb{R}^n$  and  $\mathbf{W} \in \mathbb{R}^{l \times n}$ , define  $\mathbf{W}_{+,h} = \text{diag}(\mathbf{W}\mathbf{h} > 0)\mathbf{W}$ . That is,  $\mathbf{W}_{+,h}$  zeros out the rows of  $\mathbf{W}$  that do not have a positive inner product with  $\mathbf{h}$  and keeps the remaining rows. We will extend the definition of  $\mathbf{W}_{+,h}$  to each layer of weights  $\mathbf{W}_i^{(1)}$  in our neural network. For  $\mathbf{W}_i^{(1)} \in \mathbb{R}^{n_1 \times n}$  and  $\mathbf{h} \in \mathbb{R}^n$ , define  $\mathbf{W}_{1+,h}^{(1)} := (\mathbf{W}_1^{(1)})_{+,h} = \text{diag}(\mathbf{W}_1^{(1)}\mathbf{h} > 0)\mathbf{W}_1^{(1)}$ . For each layer  $i > 1$ , define

$$\mathbf{W}_{i+,h}^{(1)} = \text{diag}(\mathbf{W}_i^{(1)}\mathbf{W}_{i-1+,h}^{(1)} \cdots \mathbf{W}_{2+,h}^{(1)}\mathbf{W}_{1+,h}^{(1)}\mathbf{h} > 0)\mathbf{W}_i^{(1)}.$$

Lastly, define  $\Lambda_{d+,h}^{(1)} := \prod_{i=d}^1 \mathbf{W}_{i+,h}^{(1)}$ . Using the above notation,  $\mathcal{G}^{(1)}(\mathbf{h})$  can be compactly written as  $\Lambda_{d+,h}^{(1)}\mathbf{h}$ . Similarly, we may write  $\mathcal{G}^{(2)}(\mathbf{m})$  compactly as  $\Lambda_{s+,m}^{(2)}\mathbf{m}$ .

The gradient descent scheme is an alternating descent direction algorithm. We first pick an initial iterate  $(\mathbf{h}_1, \mathbf{m}_1)$  such that  $\mathbf{h}_1 \neq \mathbf{0}$  and  $\mathbf{m}_1 \neq \mathbf{0}$ . At each iteration  $i = 1, 2, \dots$ , we first compare the objective value at  $(\mathbf{h}_1, \mathbf{m}_1)$ ,  $(-\mathbf{h}_1, \mathbf{m}_1)$ ,  $(\mathbf{h}_1, -\mathbf{m}_1)$ , and  $(-\mathbf{h}_1, -\mathbf{m}_1)$  and reset  $(\mathbf{h}_1, \mathbf{m}_1)$  to be the point with least objective value. Second we descend along a direction. We compute the descent direction  $\tilde{\mathbf{g}}_{1,(h,m)}$ , given by the partial derivative of  $f$  in (2) w.r.t.  $\mathbf{h}$ ,

$$\tilde{\mathbf{g}}_{1,(h,m)} := \Lambda_{d+,h}^{(1)\top} \left( \text{diag}(\Lambda_{s+,m}^{(2)})^2 \Lambda_{d+,h}^{(1)} \mathbf{h} - \text{diag}(\Lambda_{s+,m}^{(2)} \odot \Lambda_{s+,m_0}^{(2)} \mathbf{m}_0) \Lambda_{d+,h_0}^{(1)} \mathbf{h}_0 \right)$$

and take a step along this direction. Next, we compute the descent direction  $\tilde{\mathbf{g}}_{2,(h,m)}$ , given by the partial derivative of  $f$  in w.r.t.  $\mathbf{m}$ ,

$$\tilde{\mathbf{g}}_{2,(h,m)} := \Lambda_{s+,m}^{(2)\top} \left( \text{diag}(\Lambda_{d+,h}^{(1)})^2 \Lambda_{s+,m}^{(2)} \mathbf{m} - \text{diag}(\Lambda_{d+,h}^{(1)} \mathbf{h} \odot \Lambda_{d+,h_0}^{(1)} \mathbf{h}_0) \Lambda_{s+,m_0}^{(2)} \mathbf{m}_0 \right).$$

and again take a step along this direction. Lastly, we normalize the iterate so that at each iteration  $i$   $\|\mathbf{h}_i\|_2 = \|\mathbf{m}_i\|_2$ . We repeat this process until convergence. Algorithm 1 outlines this process.

---

### Algorithm 1 Alternating descent algorithm for (2)

---

- Input: Weight matrices,  $\mathbf{W}_i^{(1)}$  and  $\mathbf{W}_i^{(2)}$ , observation  $\mathbf{y}_0$  and step size  $\eta > 0$ .  
Output: An estimate of a global minimizer of (2)
- 1: Choose an arbitrary point  $(\mathbf{h}_1, \mathbf{m}_1)$  such that  $\mathbf{h}_1 \neq \mathbf{0}$  and  $\mathbf{m}_1 \neq \mathbf{0}$
  - 2: **for**  $i = 1, 2, \dots$  **do**:
  - 3:    $(\mathbf{h}_i, \mathbf{m}_i) \leftarrow \arg \min(f(\mathbf{h}_i, \mathbf{m}_i), f(-\mathbf{h}_i, \mathbf{m}_i), f(\mathbf{h}_i, -\mathbf{m}_i), f(-\mathbf{h}_i, -\mathbf{m}_i))$
  - 4:    $\mathbf{h}_{i+1} \leftarrow \mathbf{h}_i - \eta \tilde{\mathbf{g}}_{1,(h_i, m_i)}$ ,    $\mathbf{m}_{i+1} \leftarrow \mathbf{m}_i - \eta \tilde{\mathbf{g}}_{2,(h_{i+1}, m_i)}$
  - 5:    $c \leftarrow \sqrt{\|\mathbf{h}_{i+1}\|_2 / \|\mathbf{m}_{i+1}\|_2}$ ,    $\mathbf{h}_{i+1} \leftarrow \mathbf{h}_{i+1}/c$ ,    $\mathbf{m}_{i+1} \leftarrow \mathbf{m}_{i+1} \cdot c$
  - 6: **end for**
-

### 3 Proof Outline

We now present our main results which states that the objective function has a descent direction at every point outside of four hyperbolic regions. In order to state these directions, we first note that the partial derivatives of  $f$  at a differentiable point  $(\mathbf{h}, \mathbf{m})$  are

$$\nabla_{\mathbf{h}} f(\mathbf{h}, \mathbf{m}) = \tilde{\mathbf{g}}_{1,(\mathbf{h},\mathbf{m})} \text{ and } \nabla_{\mathbf{m}} f(\mathbf{h}, \mathbf{m}) = \tilde{\mathbf{g}}_{2,(\mathbf{h},\mathbf{m})}.$$

The function  $f$  is not differentiable everywhere because of the behavior of the RELU activation function in the neural network. However, since  $\mathcal{G}^{(1)}$  and  $\mathcal{G}^{(2)}$  are piecewise linear,  $f$  is differentiable at  $(\mathbf{h}, \mathbf{m}) + \delta \mathbf{w}$  for all  $(\mathbf{h}, \mathbf{m})$  and  $\mathbf{w}$  and sufficiently small  $\delta$ . The directions we consider are  $g_{1,(\mathbf{h},\mathbf{m})} \in \mathbb{R}^{n+p}$  and  $g_{2,(\mathbf{h},\mathbf{m})} \in \mathbb{R}^{n+p}$ , where

$$g_{1,(\mathbf{h},\mathbf{m})} = \begin{bmatrix} \lim_{\delta \rightarrow 0^+} \nabla_{\mathbf{h}} f((\mathbf{h}, \mathbf{m}) + \delta \mathbf{w}) \\ \mathbf{0} \end{bmatrix}, \quad g_{2,(\mathbf{h},\mathbf{m})} = \begin{bmatrix} \mathbf{0} \\ \lim_{\delta \rightarrow 0^+} \nabla_{\mathbf{m}} f((\mathbf{h}, \mathbf{m}) + \delta \mathbf{w}) \end{bmatrix}, \text{ and} \quad (4)$$

$\mathbf{w}$  is fixed. Let  $D_{\mathbf{g}} f(\mathbf{h}, \mathbf{m})$  be the unnormalized one-sided directional derivative of  $f(\mathbf{h}, \mathbf{m})$  in the direction of  $\mathbf{g}$ :  $D_{\mathbf{g}} f(\mathbf{h}, \mathbf{m}) = \lim_{t \rightarrow 0^+} \frac{f((\mathbf{h}, \mathbf{m}) + t\mathbf{g}) - f(\mathbf{h}, \mathbf{m})}{t}$ .

**Theorem 2.** Fix  $\epsilon > 0$  such that  $K_1(d^7 s^2 + d^2 s^7) \epsilon^{1/4} < 1$ ,  $d \geq 2$ , and  $s \geq 2$ . Assume the networks satisfy assumptions A2 and A3. Assume  $\mathbf{W}_i^{(1)} \sim \mathcal{N}(\mathbf{0}, \frac{1}{n_i} \mathbf{I}_{n_{i-1}})$  for  $i = 1, \dots, d-1$  and  $i$ th row of  $\mathbf{W}_d^{(1)}$  satisfies  $(\mathbf{w}_d^{(1)})_i^\top = \mathbf{w}^\top \cdot \mathbf{1}_{\|\mathbf{w}\|_2 \leq 3\sqrt{n_{d-1}/\ell}}$  with  $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \frac{1}{\ell} \mathbf{I}_{n_{d-1}})$ . Similarly, assume  $\mathbf{W}_i^{(2)} \sim \mathcal{N}(\mathbf{0}, \frac{1}{p_i} \mathbf{I}_{p_{i-1}})$  for  $i = 1, \dots, s-1$  and  $i$ th row of  $\mathbf{W}_s^{(2)}$  satisfies  $(\mathbf{w}_s^{(2)})_i^\top = \mathbf{w}^\top \cdot \mathbf{1}_{\|\mathbf{w}\|_2 \leq 3\sqrt{p_{s-1}/\ell}}$  with  $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \frac{1}{\ell} \mathbf{I}_{p_{s-1}})$ . Let  $\mathcal{K} = \{(\mathbf{h}, \mathbf{0}) \in \mathbb{R}^{n \times p} \mid \mathbf{h} \in \mathbb{R}^n\} \cup \{(\mathbf{0}, \mathbf{m}) \in \mathbb{R}^{n \times p} \mid \mathbf{m} \in \mathbb{R}^p\}$  and  $\mathcal{A} = \mathcal{A}_{K_2 d^3 s^3 \epsilon^{\frac{1}{4}}, (\mathbf{h}_0, \mathbf{m}_0)} \cup \mathcal{A}_{K_2 d^8 s^3 \epsilon^{\frac{1}{4}}, (-\rho_d^{(1)} \mathbf{h}_0, \mathbf{m}_0)} \cup \mathcal{A}_{K_2 d^3 s^8 \epsilon^{\frac{1}{4}}, (\rho_s^{(2)} \mathbf{h}_0, -\mathbf{m}_0)} \cup \mathcal{A}_{K_2 d^8 s^8 \epsilon^{\frac{1}{4}}, (-\rho_d^{(1)} \rho_s^{(2)} \mathbf{h}_0, -\mathbf{m}_0)}$ . Then on an event of probability at least  $1 - \sum_{i=1}^d \tilde{c} n_i e^{-\gamma n_{i-1}} - \sum_{i=1}^s \tilde{c} p_i e^{-\gamma p_{i-1}} - \tilde{c} e^{-\gamma \ell / (n_{d-1} \log n_{d-1} + p_{s-1} \log p_{s-1})}$  we have the following: for  $(\mathbf{h}_0, \mathbf{m}_0) \neq (\mathbf{0}, \mathbf{0})$ , and

$$(\mathbf{h}, \mathbf{m}) \notin \mathcal{A} \cup \mathcal{K}$$

the one-sided directional derivative of  $f$  in the direction of  $\mathbf{g} = g_{1,(\mathbf{h},\mathbf{m})}$  or  $\mathbf{g} = g_{2,(\mathbf{h},\mathbf{m})}$ , defined in (4), satisfy  $D_{-\mathbf{g}} f(\mathbf{h}, \mathbf{m}) < 0$ . Additionally, elements of the set  $\mathcal{K}$  are local maximizers. Here,  $\rho_d^{(k)}$  are positive numbers that converge to 1 as  $d \rightarrow \infty$ ,  $c$  and  $\gamma^{-1}$  are constants that depend polynomially on  $\epsilon^{-1}$  and  $\tilde{c}$ ,  $K_1$ , and  $K_2$  are absolute constants.

We prove Theorem 2 by showing that neural networks with random weights satisfy two deterministic conditions. These conditions are the Weight Distributed Condition (WDC) and the joint Weight Distributed Condition (joint-WDC). The WDC is a slight generalization of the WDC introduced in Hand and Voroninski [2017]. We say a matrix  $\mathbf{W} \in \mathbb{R}^{\ell \times n}$  satisfies the WDC with constants  $\epsilon > 0$  and  $0 < \alpha \leq 1$  if for all nonzero  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^k$ ,

$$\left\| \sum_{i=1}^{\ell} \mathbf{1}_{\mathbf{w}_i \cdot \mathbf{x} > 0} \mathbf{1}_{\mathbf{w}_i \cdot \mathbf{y} > 0} \cdot \mathbf{w}_i \mathbf{w}_i^\top - \alpha \mathbf{Q}_{\mathbf{x}, \mathbf{y}} \right\| \leq \epsilon, \text{ with } \mathbf{Q}_{\mathbf{x}, \mathbf{y}} = \frac{\pi - \theta_0}{2\pi} \mathbf{I}_n + \frac{\sin \theta_0}{2\pi} \mathbf{M}_{\hat{\mathbf{x}} \leftrightarrow \hat{\mathbf{y}}}, \quad (5)$$

where  $\mathbf{w}_i \in \mathbb{R}^n$  is the  $i$ th row of  $\mathbf{W}$ ;  $\mathbf{M}_{\hat{\mathbf{x}} \leftrightarrow \hat{\mathbf{y}}} \in \mathbb{R}^{n \times n}$  is the matrix such that  $\hat{\mathbf{x}} \rightarrow \hat{\mathbf{y}}$ ,  $\hat{\mathbf{y}} \rightarrow \hat{\mathbf{x}}$ , and  $\mathbf{z} \rightarrow \mathbf{0}$  for all  $\mathbf{z} \in \text{span}(\{\mathbf{x}, \mathbf{y}\})^\perp$ ;  $\hat{\mathbf{x}} = \mathbf{x} / \|\mathbf{x}\|_2$  and  $\hat{\mathbf{y}} = \mathbf{y} / \|\mathbf{y}\|_2$ ;  $\theta_0 = \angle(\mathbf{x}, \mathbf{y})$ ; and  $\mathbf{1}_S$  is the indicator function on  $S$ . If  $\mathbf{w}_i \sim \mathcal{N}(\mathbf{0}, \frac{1}{\ell} \mathbf{I}_n)$  for all  $i$ , then an elementary calculation shows that  $\mathbb{E} \left[ \sum_{i=1}^{\ell} \mathbf{1}_{\mathbf{w}_i \cdot \mathbf{x} > 0} \mathbf{1}_{\mathbf{w}_i \cdot \mathbf{y} > 0} \cdot \mathbf{w}_i \mathbf{w}_i^\top \right] = \mathbf{Q}_{\mathbf{x}, \mathbf{y}}$  and if  $\mathbf{x} = \mathbf{y}$  then  $\mathbf{Q}_{\mathbf{x}, \mathbf{y}}$  is an isometry up to a factor of  $1/2$ . Also, note that if  $\mathbf{W}$  satisfies WDC with constants  $\epsilon$  and  $\alpha$ , then  $\frac{1}{\sqrt{\alpha}} \mathbf{W}$  satisfies WDC with constants  $\epsilon/\alpha$  and  $1$ .

We now state the joint Weight Distributed Condition. We say that  $\mathbf{B} \in \mathbb{R}^{\ell \times n}$  and  $\mathbf{C} \in \mathbb{R}^{\ell \times p}$  satisfy joint-WDC with constants  $\epsilon > 0$  and  $0 < \alpha \leq 1$  if for all nonzero  $\mathbf{h}, \mathbf{x} \in \mathbb{R}^n$  and nonzero  $\mathbf{m}, \mathbf{y} \in \mathbb{R}^p$ ,

$$\left\| \mathbf{B}_{+, \mathbf{h}}^\top \text{diag}(\mathbf{C}_{+, \mathbf{m}} \mathbf{m} \odot \mathbf{C}_{+, \mathbf{y}} \mathbf{y}) \mathbf{B}_{+, \mathbf{x}} - \frac{\alpha}{\ell} \mathbf{m}^\top \mathbf{Q}_{\mathbf{m}, \mathbf{y}} \mathbf{y} \cdot \mathbf{Q}_{\mathbf{h}, \mathbf{x}} \right\| \leq \frac{\epsilon}{\ell} \|\mathbf{m}\|_2 \|\mathbf{y}\|_2, \text{ and} \quad (6)$$

$$\left\| C_{+,m}^\top \text{diag} (B_{+,h} \mathbf{h} \odot B_{+,x} \mathbf{x}) C_{+,y} - \frac{\alpha}{\ell} \mathbf{h}^\top \mathbf{Q}_{h,x} \mathbf{x} \cdot \mathbf{Q}_{m,y} \right\| \leq \frac{\epsilon}{\ell} \|\mathbf{h}\|_2 \|\mathbf{x}\|_2 \quad (7)$$

We analyze networks  $\mathcal{G}^{(1)}$  and  $\mathcal{G}^{(2)}$  where the weight matrices corresponding to the inner layers satisfy the WDC with constants  $\epsilon > 0$  and 1 and for the two matrices corresponding to the outer layers, we assume that one of them satisfies WDC with constants  $\epsilon$  and  $0 < \alpha_1 \leq 1$  and the other satisfies WDC with constants  $\epsilon$  and  $0 < \alpha_2 \leq 1$ . We also assume that the two outer layer matrices satisfy joint-WDC with constants  $\epsilon > 0$  and  $\alpha = \alpha_1 \cdot \alpha_2$ . We now state the main deterministic result:

**Theorem 3.** Fix  $\epsilon > 0$ ,  $0 < \alpha_1 \leq 1$  and  $0 < \alpha_2 \leq 1$  such that  $K_1(d^7 s^2 + d^2 s^7) \epsilon^{1/4} / (\alpha_1 \alpha_2) < 1$ ,  $d \geq 2$ , and  $s \geq 2$ . Let  $\mathcal{K} = \{(\mathbf{h}, \mathbf{0}) \in \mathbb{R}^{n \times p} \mid \mathbf{h} \in \mathbb{R}^n\} \cup \{(\mathbf{0}, \mathbf{m}) \in \mathbb{R}^{n \times p} \mid \mathbf{m} \in \mathbb{R}^p\}$ . Suppose that  $\mathbf{W}_i^{(1)} \in \mathbb{R}^{n_i \times n_{i-1}}$  for  $i = 1, \dots, d-1$  and  $\mathbf{W}_i^{(2)} \in \mathbb{R}^{p_i \times p_{i-1}}$  for  $i = 1, \dots, s-1$  satisfy the WDC with constant  $\epsilon$  and 1. Suppose  $\mathbf{W}_d^{(1)} \in \mathbb{R}^{\ell \times n_{d-1}}$  satisfy WDC with constants  $\epsilon$  and  $\alpha_1$ , and  $\mathbf{W}_s^{(2)} \in \mathbb{R}^{\ell \times p_{s-1}}$  satisfy WDC with constants  $\epsilon$  and  $\alpha_2$ . Also, suppose  $(\mathbf{W}_d^{(1)}, \mathbf{W}_s^{(2)})$  satisfy joint-WDC with constants  $\epsilon$ ,  $\alpha = \alpha_1 \cdot \alpha_2$ . Let  $\mathcal{K} = \{(\mathbf{h}, \mathbf{0}) \in \mathbb{R}^{n \times p} \mid \mathbf{h} \in \mathbb{R}^n\} \cup \{(\mathbf{0}, \mathbf{m}) \in \mathbb{R}^{n \times p} \mid \mathbf{m} \in \mathbb{R}^p\}$  and  $\mathcal{A} = \mathcal{A}_{K_2 d^3 s^3 \epsilon^{\frac{1}{4}} \alpha^{-1}, (\mathbf{h}_0, \mathbf{m}_0)} \cup \mathcal{A}_{K_2 d^8 s^3 \epsilon^{\frac{1}{4}} \alpha^{-1}, (-\rho_d^{(1)} \mathbf{h}_0, \mathbf{m}_0)} \cup \mathcal{A}_{K_2 d^3 s^8 \epsilon^{\frac{1}{4}} \alpha^{-1}, (\rho_s^{(2)} \mathbf{h}_0, -\mathbf{m}_0)} \cup \mathcal{A}_{K_2 d^8 s^8 \epsilon^{\frac{1}{4}} \alpha^{-1}, (-\rho_d^{(1)} \rho_s^{(2)} \mathbf{h}_0, -\mathbf{m}_0)}$ . Then, for  $(\mathbf{h}_0, \mathbf{m}_0) \neq (\mathbf{0}, \mathbf{0})$ , and

$$(\mathbf{h}, \mathbf{m}) \notin \mathcal{A} \cup \mathcal{K}$$

the one-sided directional derivative of  $f$  in the direction of  $\mathbf{g} = \mathbf{g}_{1,(\mathbf{h}, \mathbf{m})}$  or  $\mathbf{g} = \mathbf{g}_{2,(\mathbf{h}, \mathbf{m})}$  satisfy  $D_{-\mathbf{g}} f(\mathbf{h}, \mathbf{m}) < 0$ . Additionally, elements of the set  $\mathcal{K}$  are local maximizers. Here,  $\rho_d^{(k)}$  are positive numbers that converge to 1 as  $d \rightarrow \infty$ , and  $K_1$ , and  $K_2$  are absolute constants.

We prove the theorems by showing that the descent directions  $\mathbf{g}_{1,(\mathbf{h}, \mathbf{m})}$  and  $\mathbf{g}_{2,(\mathbf{h}, \mathbf{m})}$  concentrate around its expectation and then characterize the set of points where the corresponding expectations are simultaneously zero. The outline of the proof is:

- The WDC and joint-WDC imply that the one-sided partial directional derivatives of  $f$  concentrate uniformly for all non-zero  $\mathbf{h}$ ,  $\mathbf{h}_0 \in \mathbb{R}^n$  and  $\mathbf{m}$ ,  $\mathbf{m}_0 \in \mathbb{R}^p$  around continuous vectors  $\mathbf{t}_{(\mathbf{h}, \mathbf{m}), (\mathbf{h}_0, \mathbf{m}_0)}^{(1)}$  and  $\mathbf{t}_{(\mathbf{h}, \mathbf{m}), (\mathbf{h}_0, \mathbf{m}_0)}^{(2)}$ , respectively, defined in equations (10) and (11) in the Appendix.
- Direct analysis show that  $\mathbf{t}_{(\mathbf{h}, \mathbf{m}), (\mathbf{h}_0, \mathbf{m}_0)}^{(1)}$  and  $\mathbf{t}_{(\mathbf{h}, \mathbf{m}), (\mathbf{h}_0, \mathbf{m}_0)}^{(2)}$  are simultaneously approximately zero around the four hyperbolic sets  $\mathcal{A}_{\epsilon, (\mathbf{h}_0, \mathbf{m}_0)}$ ,  $\mathcal{A}_{\epsilon, (-\rho_d^{(1)} \mathbf{h}_0, \mathbf{m}_0)}$ ,  $\mathcal{A}_{\epsilon, (\mathbf{h}_0, -\rho_s^{(2)} \mathbf{m}_0)}$ , and  $\mathcal{A}_{\epsilon, (-\rho_d^{(1)} \mathbf{h}_0, -\rho_s^{(2)} \mathbf{m}_0)}$ , where  $\epsilon$  depends on the expansivity and number of layers in the networks, and both  $\rho_d^{(1)}$  and  $\rho_s^{(2)}$  are positive constants close to 1 and depends on the number of layers in the two neural networks as well.
- Using sphere covering arguments, Gaussian and truncated Gaussian matrices with appropriate dimensions satisfy the WDC and joint-WDC conditions.

The full proof of Theorem 3 is provided in the Appendix.

## 4 Numerical Experiment

We now empirically show that Algorithm 1 can remove distortions present in the dataset. We consider the image recovery task of removing distortions that were synthetically introduced to the MNIST dataset. The distortion dataset contain 8100 images of size  $28 \times 28$  where the distortions are generated using a 2D Gaussian function,  $g(x, y) = e^{-\frac{(x-c)^2 + (y-c)^2}{\sigma}}$ , where  $c$  is the center and  $\sigma$  controls its tail behavior. For each of the 8100 image, we fix  $c$  and  $\sigma$ , which vary uniformly in the intervals  $[-3, 3]$  and  $[20, 35]$ , respectively, and  $x$  and  $y$  are in the interval  $[-5, 5]$ . Prior to training the generators, the images in the MNIST dataset and the distortion dataset were resized to  $64 \times 64$  images. We used DCGAN [Radford et al., 2016] with a learning rate of 0.0002 and latent code dimension of 50 to train a generator,  $\mathcal{G}^{(2)}$ , for the distortion images. Similarly, we used the DCGAN with learning rate of 0.0002 and latent code dimension of 100 to train a generator,  $\mathcal{G}^{(1)}$ , for the MNIST images. Finally,

a distorted image  $y_0$  is generated via the pixelwise multiplication of an image  $w_0$  from the MNIST dataset and an image  $x_0$  from the distortion dataset, i.e.  $y_0 = w_0 \odot x_0$ .

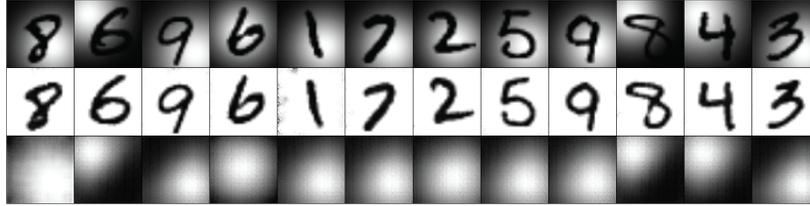


Figure 2: The figure shows the result removing distortion in an image by solving (2) using Algorithm 1. The top row corresponds to the input distorted image. The second and third row corresponds to the images recovered using empirical risk minimization.

Figure 2 shows the result of using Algorithm 1 to remove distortion from  $y_0$ . In the implementation of Algorithm 1,  $\tilde{g}_{1,(h_i,m_i)}$  and  $\tilde{g}_{1,(h_i,m_i)}$  corresponds to the partial derivatives of  $f$  with the generators as  $\mathcal{G}^{(1)}$  and  $\mathcal{G}^{(2)}$ . We used the Stochastic Gradient Descent algorithm with the step size set to 1 and momentum set to 0.9. For each image in the first row of Figure 2, the corresponding images in the second and third rows are the output of Algorithm 1 after 500 iterations.

## References

- Ali Ahmed, Benjamin Recht, and Justin Romberg. Blind deconvolution using convex programming. *IEEE Trans. Inform. Theory*, 60(3):1711–1732, 2014.
- Thomas G Stockham, Thomas M Cannon, and Robert B Ingebretsen. Blind deconvolution through digital signal processing. *Proceedings of the IEEE*, 63(4):678–692, 1975.
- Deepa Kundur and Dimitrios Hatzinakos. Blind image deconvolution. *IEEE signal processing magazine*, 13(3):43–64, 1996.
- Alireza Aghasi, Barmak Heshmat, Albert Redo-Sanchez, Justin Romberg, and Ramesh Raskar. Sweep distortion removal from terahertz images via blind demodulation. *Optica*, 3(7):754–762, 2016.
- Alireza Aghasi, Ali Ahmed, and Paul Hand. Branchhull: Convex bilinear inversion from the entrywise product of signals with known signs. *Applied Computational and Harmonic Analysis*, 2019. doi: <https://doi.org/10.1016/j.acha.2019.03.002>.
- James R Fienup. Phase retrieval algorithms: a comparison. *Applied optics*, 21(15):2758–2769, 1982.
- E. Candès and X. Li. Solving quadratic equations via phaselift when there are about as many equations as unknowns. *Found. Comput. Math.*, pages 1–10, 2012.
- E. Candès, T. Strohmer, and V. Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Commun. Pure Appl. Math.*, 66(8):1241–1274, 2013.
- Ivana Tasic and Pascal Frossard. Dictionary learning. *IEEE Signal Processing Magazine*, 28(2): 27–38, 2011.
- Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(Nov):1457–1469, 2004.
- Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.
- Shuyang Ling and Thomas Strohmer. Self-calibration and biconvex compressive sensing. *Inverse Problems*, 31(11):115002, 2015.
- Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017. URL <http://arxiv.org/abs/1710.10196>.

- Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *CoRR*, abs/1609.03499, 2016. URL <http://arxiv.org/abs/1609.03499>.
- Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Trans. Graph.*, 36(4):107:1–107:14, July 2017. ISSN 0730-0301. doi: 10.1145/3072959.3073659. URL <http://doi.acm.org/10.1145/3072959.3073659>.
- Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017. URL <http://arxiv.org/abs/1703.10593>.
- Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised MAP inference for image super-resolution. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL <https://openreview.net/forum?id=S1RP6GL1e>.
- Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G. Dimakis. Compressed sensing using generative models. 2017. URL <https://arxiv.org/abs/1703.03208>.
- S. Lohit, K. Kulkarni, R. Kerviche, P. Turaga, and A. Ashok. Convolutional neural networks for noniterative reconstruction of compressively sensed images. *IEEE Transactions on Computational Imaging*, 4(3):326–340, Sep. 2018. ISSN 2333-9403. doi: 10.1109/TCI.2018.2846413.
- Muhammad Asim, Fahad Shamshad, and Ali Ahmed. Solving bilinear inverse problems using deep generative priors. *CoRR*, abs/1802.04073, 2018. URL <http://arxiv.org/abs/1802.04073>.
- Fahad Shamshad, Farwa Abbas, and Ali Ahmed. Deep ptych: Subsampled fourier ptychography using generative priors. *CoRR*, abs/1812.11065, 2018.
- Kiryung Lee, Yihing Wu, and Yoram Bresler. Near optimal compressed sensing of a class of sparse low-rank matrices via sparse power factorization. *arXiv preprint arXiv:1702.04342*, 2017.
- Xiaodong Li and Vladislav Voroninski. Sparse signal recovery from quadratic measurements via convex programming. *SIAM Journal on Mathematical Analysis*, 45(5):3019–3033, 2013.
- Sohail Bahmani and Justin Romberg. Phase retrieval meets statistical learning theory: A flexible convex relaxation. *arXiv preprint arXiv:1610.04210*, 2016.
- Tom Goldstein and Christoph Studer. Phasemax: Convex phase retrieval via basis pursuit. *arXiv preprint arXiv:1610.07531*, 2016.
- Paul Hand and Vladislav Voroninski. Compressed sensing from phaseless gaussian measurements via linear programming in the natural parameter space. *CoRR*, abs/1611.05985, 2016. URL <http://arxiv.org/abs/1611.05985>.
- Alireza Aghasi, Ali Ahmed, Paul Hand, and Babhru Joshi. A convex program for bilinear inversion of sparse vectors. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8548–8558. Curran Associates, Inc., 2018. URL <http://papers.nips.cc/paper/8074-a-convex-program-for-bilinear-inversion-of-sparse-vectors.pdf>.
- Emmanuel Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Trans. Inform. Theory*, 61(4):1985–2007, 2015.
- Gang Wang, Georgios B Giannakis, and Yonina C Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *arXiv preprint arXiv:1605.08285*, 2016.
- Xiaodong Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *arXiv preprint arXiv:1606.04933*, 2016.
- Paul Hand and Vladislav Voroninski. Global guarantees for enforcing deep generative priors by empirical risk. *CoRR*, abs/1705.07576, 2017. URL <http://arxiv.org/abs/1705.07576>.

- Paul Hand, Oscar Leong, and Vladislav Voroninski. Phase retrieval under a generative prior. *CoRR*, abs/1807.04261, 2018. URL <http://arxiv.org/abs/1807.04261>.
- Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.
- R. Vershynin. *Compressed sensing: theory and applications*. Cambridge University Press, 2012.
- Halyun Jeong, Xiaowei Li, Yaniv Plan, and Ozgur Yilmaz. Non-gaussian random matrices on sets: Optimal tail dependence and applications. In *13th International conference on Sampling Theory and Applications*, 2019.