Global Guarantees for Enforcing Deep Generative Priors by Empirical Risk

Paul Hand and Vladislav Voroninski

***Abstract*—We examine the theoretical properties of enforcing priors provided by generative deep neural networks via empirical risk minimization. In particular we consider two models, one in which the task is to invert a generative neural network given access to its last layer and another in which the task is to invert a generative neural network given only compressive linear observations of its last layer. We establish that in both cases, in suitable regimes of network layer sizes and a randomness assumption on the network weights, that the non-convex objective function given by empirical risk minimization does not have any spurious stationary points. That is, we establish that with high probability, at any point away from small neighborhoods around two scalar multiples of the desired solution, there is a descent direction. Hence, there are no local minima, saddle points, or other stationary points outside these neighborhoods. These results constitute the first theoretical guarantees which establish the favorable global geometry of these non-convex optimization problems, and they bridge the gap between the empirical success of enforcing deep generative priors and a rigorous understanding of non-linear inverse problems.**

***Index Terms*—Information theory, Learning, Optimization, Probability, Generative Models.**

I. Introduction

EXPLOITING the structure of images and natural signals has proven to be a fruitful endeavor across many domains of science. For instance, the wavelet transform, discovered by Daubechies and others [1], led to the observation that natural images are sparse in the wavelet basis, enabling compression algorithms such as JPEG 2000 to tame the storage and transfer of the modern deluge of image and video data. Principles of wavelet based image compression, combined with surprising advances in convex relaxation, have also opened the door to greatly improved signal acquisition strategies, which unlocked critical applications throughout the imaging sciences. In particular, breaking with the dogma of the Nyquist sampling theorem, which stems from worst-case analysis, Candes and Tao, and Donoho [2], [3], [4], provided a theory and practice of compressed sensing (CS), which exploits the sparsity of natural signals in the wavelet basis to design acquisition strategies with drastically lower sample complexity — that on par with the sparsity level of the signal at hand. In particular, using the standard basis in lieu of the wavelet basis without loss of generality, they established that to recover a vector $x \in \mathbb{R}^n$ with $k < n$ non-zero entries from $m = O(k \log n)$ observations $\langle x, a_i \rangle, i = 1, 2 \ldots, m$, where a_i are i.i.d Gaussian, it suffices to minimize $\|x\|_1$ subject to the observations with high probability. On a practical level, compressed sensing has lead to significant reduction in the sample complexity of signal acquisition of natural images, for instance speeding up MRI imaging by an order of magnitude [5]. Beyond MRI, compressed sensing has impacted many if not all imaging sciences, by providing a general tool to exploit the parsimony of natural signals to improve acquisition speed, increase SNR and reduce sample complexity. More broadly, the principled use of sparsity as a prior has led to the development of the field of matrix completion [6], breakthroughs in phase retrieval [7], [8] and blind deconvolution [9]; and is at this point routinely utilized across applied mathematics and machine learning.

Meanwhile, the advent of practical deep learning [10] has significantly improved machine understanding of image and audio data. For instance, deep learning techniques are now the state of the art across most of computer vision and have taken the field far beyond where it stood just a few years prior. The success of deep learning ostensibly stems from its ability to exploit the hierarchical nature of images and other natural signals without explicit hand-engineering. There are many techniques and add-on architectural choices associated with deep learning, but many of them are non-essential from a theoretical and, to a large extent, practical perspective, with simple convolutional deep nets with Rectified Linear Units (ReLUs) achieving close to the state of the art performance on many tasks [11]. The class of functions represented by such deep networks is readily interpretable as hierarchical compression schemes with exponentially many linear filters, each being a linear combination of filters in earlier layers. Constructing such compression schemes by hand would be quite tedious, if not impossible, and the biggest surprise and advantage of deep learning is

P. Hand is with the Department of Mathematics and the Khoury College of Computer Sciences at Northeastern University, Boston, MA.
V. Voroninski is with Helm.ai, Menlo Park, CA.
Manuscript received March 26, 2018; revised April 29, 2019.
This work was presented in part at Conference on Learning Theory (COLT) 2018.

that simple stochastic gradient descent (SGD) allows one to efficiently traverse this class of functions subject to potentially highly non-convex learning objectives. While this latter property has been empirically established in an impressive number of applications, it has so far eluded a completely satisfactory theoretical explanation.

In essence, compressive sensing, and its numerous extensions, consist of enforcing a sparsity prior to regularize the solution of an inverse problem. Thus, improvements in the state of the art of compressed sensing can come from better reconstruction algorithms, better design of signal measurements, or more sophisticated priors. Virtually all of the tens of thousands of research articles in the umbrella field of compressive sensing have focused on the first two directions, taking the linear sparsity model as the de-facto prior for regularization. Those two directions are fundamentally limited in that no approach at recovering a k-sparse signal with respect to a basis could succeed with fewer than k measurements.

Meanwhile, there have been great strides in generative modeling of images in modern machine learning that go well beyond linear sparsity models. Such improvements in priors on natural images beyond wavelet based approaches, when properly enforced, should enable more aggressive regularization of inverse problems, leading to lower sample complexity and higher SNR than traditional compressed sensing approaches. In order to understand the potential for improving upon traditional compressive sensing, broadly speaking, as a function of advances in generative modeling, it is useful to reinterpret compressive sensing from the perspective of the field of generative modeling, a popular framework in machine learning which strives to sample from the probability distribution of natural images and other signals. Note that there is a duality between generative modeling and compression. Any compression scheme implicitly defines a generative model, by its inverse, and vice versa. In particular, wavelet based compression schemes implicitly define a generative model which attempts to sample from natural images via random sparse linear combinations of wavelet basis images. This wavelet-based generative model is clearly too loose to capture the rich hierarchical structure of natural images, making it a sufficiently expressive yet very naive prior.

Generative modeling has a rich history in machine learning, but only recent deep neural network based approaches to generative modeling have enabled the generation of realistic synthetic images in a variety of domains, for example by training generative adversarial networks (GANs) to find a Nash equilibrium of a non-convex game [12], [13]; by training variational auto-encoders (VAEs) [14], [15]; and by training autoregressive models like PixelCNN [16], which generate pixels one-at-a-time by sampling from appropriate conditional probability distributions. GANs and VAEs map a low dimensional latent code space to a higher dimensional embedding space of images or other natural signals. For instance, if we equip the latent code space with a Gaussian distribution, the goal of generative adversarial training is to produce a deep neural network generator whose push-forward distribution is the distribution of natural images or another class of natural signals. Impressively, in-between the original posting of this paper and the current version of the manuscript, deep generative modeling has advanced to the point of producing high-resolution synthetic, yet extremely photorealistic, images of celebrity faces [17]. Further, continuous motion in the latent code space of the associated deep generative models has allowed for interpolation and continuous deformation of the resulting faces, even exhibiting equivariant properties with arithmetic operations in the latent code space corresponding to semantically meaningful image variations [18].

The scope of application of deep generative modeling to regularizing inverse problems is vast. These more sophisticated priors are recently emerging in empirical applications of many fields of imaging, such as medical imaging [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], microscopy [33], inpainting [34], [35], superresolution [36], [37], [38], compressed sensing [39], [40], [41], [42], image manipulation [43], and many more. See [44] for a review of deep learning for inverse problems in imaging. Importantly, approaches that regularize inverse problems using deep generative models, have empirically been shown to improve over sparsity-based approaches, advancing the state of the art in several fields. For instance, in Magnetic Resonance Imaging, deep generative networks have enabled image reconstruction that is qualitatively of higher diagnostic quality and higher SNR than traditional compressive sensing allows, and is additionally two orders of magnitude faster than sparsity-based approaches due to the utilization of GPUs in applying convolutional neural networks [45], [46]. This development is significant because of the tremendous potential clinical applications of diagnostic-quality real-time MRI visualization. Deep generative models have also empirically been used directly for compression [47]. In the case of compressed sensing, optimization of an empirical risk objective over the latent code space has been empirically shown to recover images from 10x fewer linear compressive measurements than sparsity-based approaches [48].

As the quality and reach of deep generative modeling continues to increase, signal recovery in many scenarios will benefit analogously.

As with the rest of machine learning, in the field of deep generative modeling for regularizing inverse problems, or as we refer to it the field of deep compressive

sensing, empirics is far ahead of the state of theoretical justification. In this paper we initiate the rigorous study of enforcing deep generative models as priors on the solutions to inverse problems, by providing a theory of compressive sensing that goes beyond linear sparsity and into the realm of applying deep neural network based generative priors. In particular we show that under suitable randomness assumptions on the weights of a neural network and successively expansive hidden layer sizes, the empirical risk objective for recovering a latent code in $\mathbb{R}^k$ from m linear observations of the last layer of a generative network, where m is proportional to k up to log factors, has no spurious local minima or saddle points, in that there is a descent direction everywhere except possibly small neighborhoods around two scalar multiples of the desired solution. Our descent direction analysis is constructive; based on deterministic conditions on the neural network weights and the measurements; and relies on novel concentration bounds of certain random matrices, uncovering some interesting geometric properties of the landscapes of empirical risk objective functions for random generative multilayer networks with ReLU activations. For a generative network that achieves a greater degree of compression, the proposed scheme would enable lower sample complexity and higher SNR. If a generative model can compress a signal to a latent code dimensionality k much less then the signal's sparsity level, then compressed sensing with the generative prior may significantly outperform compressed sensing with sparsity prior in terms of sample complexity.

A. Related theoretical work

Latent code space optimizations after neural network training, and the optimization over the weights of a neural network during training, may both be interpreted as inverse problems [49]. The tools developed in this paper, such as the novel nonasymptotic concentration results for high dimensional Gaussians followed by a ReLU, may be of independent interest, in particular being amenable for establishing global non-asymptotic analysis regarding convergence of SGD for training deep neural networks. Our work also relates to recent trends in optimization. Traditionally, rigorous understanding of inverse problems has been limited to the simpler setting in which the optimization objective is convex. More recently, there has been progress in understanding non-convex optimization objectives for inverse problems, in albeit analytically simpler situations than those involving multilayer neural networks. For instance, the authors of [50], [51] provide a global analysis of non-convex objectives for phase retrieval and community detection, respectively, ruling out adversarial geometries in these scenarios for the purposes of optimization. Additionally, rigorous guarantees of nonconvex recovery include other results in phase retrieval [52], [53], blind deconvolution [54], [55], [56], robust subspace recovery [57], discrete joint alignment [58], and more.

In related work, the authors of [48] also study inverting compressive linear observations under generative priors, by proving a restricted eigenvalue condition on the range of the generative neural network. However, they only provide a guarantee that is local in nature, in showing the global optimum of empirical risk is close to the desired solution. The work provides no guarantees about why the global minimum of the non-convex problem can be reached. In addition, [59] studied inverting neural networks given access to the last layer using an analytical formula that approximates the inverse mapping of a neural network. The results of [59] are in a setting where the neural net is not generative, and their procedure is at only approximate, and, since it requires observation of the last layer, it is not readily extendable to the compressive linear observation setting. Meanwhile, the optimization problem we study can yield exact recovery, which we observe empirically via gradient descent. Most importantly, in contrast to [48], [59], we provide a global analysis of the non-convex empirical risk objective function and constructively exhibit a descent direction at every point outside a neighborhood of the desired solution and a negative scalar multiple of it. Our guarantees are non-asymptotic, and to the best of our knowledge the first of their kind.

B. Notation

Before we present the main result, we now introduce notation that will be used throughout this paper. Let $[n] = \{1, \ldots, n\}$. Let e_i is the ith standard basis element for $i \in [n]$. Let $\mathrm{relu}(x) = \max(x, 0)$ apply entrywise for $x \in \mathbb{R}^n$. Let $\mathrm{diag}(Wx > 0)$ be the diagonal matrix that is 1 in the (i,i)th entry if $(Wx)_i > 0$, and 0 otherwise. Let $\mathcal{B}(x, r)$ be the Euclidean ball of radius r centered at x. Let $\Pi_{i=d}^{1} W_i = W_d W_{d-1} \cdots W_1$. Let I_n be the $n \times n$ identity matrix. Let $A \preceq B$ mean that $B - A$ is a positive semidefinite matrix. For matrices A, let $\|A\|$ be the spectral norm of A. Let S^{k-1} be the unit sphere in $\mathbb{R}^k$. For any nonzero $x \in \mathbb{R}^n$, let $\hat{x} = x/\|x\|_2$. For a set S, let $|S|$ denote its cardinality. We will write $\gamma = O(\delta)$ to mean that there exists a positive constant C such that $\gamma \leqslant C\delta$, when γ is understood to be positive. Similarly we will write $c = \Omega(\delta)$ to mean that there exists a positive constant C such that $c \geqslant C\delta$. When we say that a constant depends polynomially on ϵ^{-1}, that means that it is at most $C\epsilon^{-k}$ for some positive C and positive integer k. Let $\theta_0 = \angle(x, x_0)$ and $\overline{\theta}_1 = g(\theta_0)$ where g is given by (3). For notational convenience, we will write $a = b + O_1(\epsilon)$ if $\|a - b\| \leqslant \epsilon$, where the norm is understood to be absolute value for scalars,

the ℓ_2 norm for vectors, and the spectral norm for matrices. Write $g^{\circ d}$ to denote the composition of g with itself d times. Let $1_S = 1$ if S and 0 otherwise. For nonzero v, let $D_v f(x)$ be the (normalized) one-sided directional derivative of f at x in the direction of v: $D_v f(x) = \lim_{t\to 0^+} \frac{f(x+tv)-f(x)}{t\|v\|_2}$.

C. Main Results

We consider the inverse problem of recovering a vector $y_0 \in \mathbb{R}^n$ from $m \ll n$ linear measurements. To resolve the inherent ambiguity from undersampling, we assume, as a prior, that the vector belongs to the range of a d-layer generative neural network $G : \mathbb{R}^k \to \mathbb{R}^n$, with $k < n$. To recover the vector $y_0 = G(x_0)$, we attempt to find the latent code $x_0 \in \mathbb{R}^k$ corresponding to it. We consider a generative network modeled by $G(x) = \text{relu}(W_d \dots \text{relu}(W_2 \,\text{relu}(W_1 x_0)) \dots)$, where $\text{relu}(x) = \max(x, 0)$ applies entrywise, $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$, n_i is the number of neurons in the ith layer, and $k = n_0 < n_1 < \cdots < n_d = n$. We consider linear measurements of $G(x_0)$ given by the sampling matrix $A \in \mathbb{R}^{m\times n}$ and consider $k < m \ll n$. The problem at hand is:

Let: $x_0 \in \mathbb{R}^k, A \in \mathbb{R}^{m\times n}, W_i \in \mathbb{R}^{n_i \times n_{i-1}}$ for $i \in [d]$,
$G(x) = \text{relu}(W_d \dots \text{relu}(W_2 \,\text{relu}(W_1 x_0)) \dots)$,
$y_0 = G(x_0)$,
Given: $W_1 \dots W_d, A,$ and observations Ay_0,
Find: x_0.

This problem can be viewed in two ways: (1) as above, given compressive measurements of a vector with the prior information that it belongs to the output of a generative neural network, find that vector; or (2), given compressive observations of the output of a generative neural network, find the latent code corresponding to the network's output by inverting the neural network and compression simultaneously.

As a way to solve the above problem, we consider minimizing the empirical risk objective

$$f(x) := \frac{1}{2}\left\|AG(x) - Ay_0\right\|_2^2. \tag{1}$$

As this objective is nonconvex, there is no *a priori* guarantee of efficiently finding the global minimum [60]. Approaches such as gradient descent could in principle get stuck in local minima, instead of finding the desired global minimizer x_0.

In this paper, we consider a fully-connected generative network $G : \mathbb{R}^k \to \mathbb{R}^n$ with Gaussian weights and no bias term, along with a Gaussian sampling matrix $A \in \mathbb{R}^{m\times n}$. We show that under appropriate conditions and with high probability, f has a strict descent direction everywhere outside two small neighborhoods of x_0 and a negative multiple of x_0. We assume that the network is sufficiently *expansive* at each layer, $n_i = \Omega(n_{i-1} \log n_{i-1})$, and that there are a sufficient number of measurements, $m = \Omega(kd \log(n_1 \cdots n_d))$. Let $D_v f(x)$ be the (normalized) one-sided directional derivative of f at x in the direction of v: $D_v f(x) = \lim_{t\to 0^+} \frac{f(x+tv)-f(x)}{t\|v\|_2}$. Let $\mathcal{B}(x, r)$ be the Euclidean ball of radius r centered at x. Our main result is as follows:

Theorem 1. *Fix $\epsilon > 0$ such that $K_1 d^8 \epsilon^{1/4} \leqslant 1$, and let $d \geqslant 2$. Assume $n_i \geqslant c n_{i-1} \log n_{i-1}$ for all $i = 1 \dots d$ and $m > cdk \log \Pi_{i=1}^d n_i$. Assume that for each i, the entries of W_i are i.i.d. $\mathcal{N}(0, 1/n_i)$, and the entries of A are i.i.d. $\mathcal{N}(0, 1/m)$ and independent from $\{W_i\}$. Then, on an event of probability at least $1 - \sum_{i=1}^d \tilde{c} n_i e^{-\gamma n_{i-1}} - \tilde{c} e^{-\gamma m}$, we have the following. For all nonzero x and x_0, there exists $v_{x,x_0} \in \mathbb{R}^k$ such that the one-sided directional derivatives of f satisfy*

$$D_{-v_{x,x_0}} f(x) < 0, \quad \forall x \notin \{0\} \cup \mathcal{B}(x_0, K_2 d^3 \epsilon^{1/4} \|x_0\|_2) \cup \mathcal{B}(-\rho_d x_0, K_2 d^{13} \epsilon^{1/4} \|x_0\|_2),$$
$$D_v f(0) < 0, \qquad \forall v \neq 0,$$

where ρ_d is a positive number that converges to 1 as $d \to \infty$. Here, c and γ^{-1} are constants that depend polynomially on ϵ^{-1}, and $\tilde{c}, K_1, K_2$ are universal constants.

This theorem states that for a network of fixed depth d, with high probability there is always a descent direction outside of two specified, sufficiently small neighborhoods, provided that the network is Gaussian and sufficiently expansive. Further, for such networks, zero is a local maximizer. We note that the linear dependence of sample complexity with respect to k, for fixed d, is optimal. The fixed d regime is realistic to applications because many deep learning networks in the wild have d on the order of only 10. We also note that the theorem's scalings with respect to ϵ, d, and n_i are all polynomial, and not exponential, though the dependence on each of these variables could likely be improved. While the sample complexity scaling appears to get worse for larger d, we note that larger d allows for the possibility of generative models with lower values of k. This is because the number of piecewise linear pieces in G grows exponentially in d. Also, note that while the weights of any layer of the network are assumed to be i.i.d. Gaussian, there is no assumption on the independence between W_i and W_j for $i \neq j$.

The descent direction v_{x,x_0} is given by the gradient of f:

$$v_{x,x_0} = \begin{cases} \nabla f(x) & G \text{ is differentiable at x,} \\ \lim_{\delta \downarrow 0} \nabla f(x + \delta w) & \text{otherwise,} \end{cases}$$

where w can be arbitrarily chosen such that G is differentiable at $x+\delta w$ for sufficiently small δ. Such a w exists

by the piecewise linearity of G, and could be generated randomly with probability 1. An explicit formula for $\nabla f(x)$, where it exists, is given by (4) in Section II. This expression for v_{x,x_0} is in a form that can be computed for any x, even for points of nondifferentiability, as part of a gradient based algorithm.

This theorem will be proven by showing the sufficiency of two deterministic conditions on G and A, and then by showing that Gaussian G and A of appropriate sizes satisfy these conditions with the appropriate probability. The first deterministic condition is on the spatial arrangement of the network weights within each layer.

Definition 2. *We say that the matrix $W \in \mathbb{R}^{n\times k}$ satisfies the* Weight Distribution Condition *with constant ϵ if for all nonzero $x, y \in \mathbb{R}^k$,*

$$\left\|\sum_{i=1}^{n} 1_{w_i\cdot x>0}1_{w_i\cdot y>0}\cdot w_i w_i^t - Q_{x,y}\right\| \leqslant \epsilon, \text{ with} \quad (2)$$
$$Q_{x,y} = \frac{\pi-\theta_0}{2\pi}I_k + \frac{\sin\theta_0}{2\pi}M_{\hat{x}\leftrightarrow\hat{y}},$$

where $w_i \in \mathbb{R}^k$ is the ith row of W; $M_{\hat{x}\leftrightarrow\hat{y}} \in \mathbb{R}^{k\times k}$ is the matrix[1] *such that $\hat{x} \mapsto \hat{y}$, $\hat{y} \mapsto \hat{x}$, and $z \mapsto 0$ for all $z \in \operatorname{span}(\{x,y\})^\perp$; $\hat{x} = x/\|x\|_2$ and $\hat{y} = y/\|y\|_2$; $\theta_0 = \angle(x,y)$; and 1_S is the indicator function on S.*

The norm on the left hand side of (2) is the spectral norm. Note that an elementary calculation[2] gives that $Q_{x,y} = \mathbb{E}[\sum_{i=1}^n 1_{w_i\cdot x>0}1_{w_i\cdot y>0}\cdot w_i w_i^t]$ for $w_i \sim \mathcal{N}(0, I_k/n)$. As the rows w_i correspond to the neural network weights of the ith neuron in a layer given by W, the WDC provides a deterministic property under which the set of neuron weights within the layer given by W are distributed approximately like a Gaussian. The WDC could also be interpreted as a deterministic property under which the neuron weights are distributed approximately like a uniform random variable on a sphere of a particular radius. Note that if $x = y$, $Q_{x,y}$ is an isometry up to a factor of $1/2$.

The second deterministic condition is that the compression matrix acts like an isometry on pairs of differences of vectors in the range of $G : \mathbb{R}^k \to \mathbb{R}^n$.

Definition 3. *We say that the compression matrix $A \in \mathbb{R}^{m\times n}$ satisfies the* Range Restricted Isometry Condition (RRIC) *with respect to G with constant ϵ if for all $x_1, x_2, x_3, x_4 \in \mathbb{R}^k$,*

$$\Big|\Big\langle A\big(G(x_1)-G(x_2)\big), A\big(G(x_3)-G(x_4)\big)\Big\rangle - \Big\langle G(x_1)-G(x_2), G(x_3)-G(x_4)\Big\rangle\Big| \leqslant \epsilon\|G(x_1)-G(x_2)\|_2\|G(x_3)-G(x_4)\|_2.$$

We can now state our main deterministic result.

Theorem 4. *Fix $\epsilon > 0$ such that $K_1 d^8\epsilon^{1/4} \leqslant 1$, and let $d \geqslant 2$. Suppose that G is such that W_i has the WDC with constant ϵ for all $i = 1\dots d$. Suppose A satisfies the RRIC with respect to G with constant ϵ. Then, for all nonzero x and x_0, there exists $v_{x,x_0} \in \mathbb{R}^k$ such that the one-sided directional derivatives of f satisfy*

$$D_{-v_{x,x_0}}f(x) < -K_3\frac{\sqrt{\epsilon}d^3}{2^d}\max(\|x\|_2, \|x_0\|_2),$$
$$D_y f(0) < -\frac{1}{8\pi 2^d}\|x_0\|_2,$$
$$\forall y \neq 0, x \notin \{0\} \cup \mathcal{B}(x_0, K_2 d^3\epsilon^{1/4}\|x_0\|_2) \cup \mathcal{B}(-\rho_d x_0, K_2 d^{13}\epsilon^{1/4}\|x_0\|_2),$$

where ρ_d is a positive number that converges to 1 as $d \to \infty$, and K_1, K_2, and K_3 are universal constants.

Note that the 2^d scaling in the bounds is an artifact of the scaling of the problem and does not indicate a vanishingly small derivative. Roughly speaking, the relu activation functions zero out roughly half of its arguments. Hence, while W_i has spectral norm approximately 1, the rows of W_i that are retained by the relu will have spectral norm approximately $1/2$. Thus, $f(x)$ itself is on the order of 2^{-d} under the RRIC and WDC for appropriately small ϵ.

In the case that $A = I_n$, the RRIC is trivially satisfied, and we get the following corollary about inverting multilayer neural networks.

Corollary 5 (Approximate Invertibility of Multilayer Neural Networks)**.** *If G is a d-layer neural network such that W_i satisfies the WDC with constant ϵ for all $i = 1\dots d$, then the function $f(x) = \|G(x) - G(x_0)\|_2$ has no stationary points outside of a neighborhood around x_0 and $-\rho_d x_0$.*

In the case of a Gaussian network with Gaussian measurements, the WDC and RRIC are satisfied with high probability if the network is sufficiently expansive and there are a sufficient number of measurements.

Proposition 6. *Fix $0 < \epsilon < 1$. Assume $n_i \geqslant cn_{i-1}\log n_{i-1}$ for all $i = 1\dots d$ and $m > cdk\log\prod_{i=1}^d n_i$. Assume the entires of W_i are i.i.d. $\mathcal{N}(0, 1/n_i)$, and the entries of A are i.i.d. $\mathcal{N}(0, 1/m)$. Then, W_i satisfies the WDC with constant ϵ for all i and A satisfies the RRIC with respect to G with constant ϵ*

[1] A formula for $M_{\hat{x}\leftrightarrow\hat{y}}$ is as follows. If $\theta_0 = \angle(\hat{x},\hat{y}) \in (0,\pi)$ and R is a rotation matrix such that $\hat{x}$ and $\hat{y}$ map to e_1 and $\cos\theta_0\cdot e_1 + \sin\theta_0\cdot e_2$ respectively, then $M_{\hat{x}\leftrightarrow\hat{y}} = R^t\begin{pmatrix}\cos\theta_0 & \sin\theta_0 & 0\\ \sin\theta_0 & -\cos\theta_0 & 0\\ 0 & 0 & 0_{k-2}\end{pmatrix}R$, where 0_{k-2} is a $k-2\times k-2$ matrix of zeros. If $\theta_0 = 0$ or π, then $M_{\hat{x}\leftrightarrow\hat{y}} = \hat{x}\hat{x}^t$ or $-\hat{x}\hat{x}^t$, respectively.

[2] To do this calculation, take $x = e_1$ and $y = \cos\theta_0\cdot e_1 + \sin\theta_0\cdot e_2$ without loss of generality. Then each entry of the matrix can be determined analytically by an integral that factors in polar coordinates.

with probability at least $1 - \sum_{i=1}^{d} \tilde{c} n_i e^{-\gamma n_{i-1}} - \tilde{c} e^{-\gamma m}$. *Here,* c *and* γ^{-1} *are constants that depend polynomially on* ϵ^{-1}*, and* $\tilde{c}$ *is a universal constant.*

As stated after Theorem 1, no assumption is made on the independence between W_i and W_j for $i \neq j$. While Proposition 6 is stated for $A \in \mathbb{R}^{m \times n}$ with i.i.d. Gaussian entries, it also applies in the case of any random matrix that satisfies the following concentration of measure condition:

$$\mathbb{P}\big(|\|Ax\|_2^2 - \|x\|_2^2| \geqslant \epsilon \|x\|_2^2\big) \leqslant 2e^{-m c_0(\epsilon)},$$

for any fixed $x \in \mathbb{R}^n$, where $c_0(\epsilon)$ is a positive constant depending only on ϵ. In particular, Proposition 6 and hence Theorem 1 extends to the case of where the entries of A are independent Bernoulli random variables (and the entries of W_i are Gaussian). See [61] for more.

D. Discussion

In this paper, we provide the first rigorous global analysis of the efficacy of enforcing generative neural network priors. We show that if a generative neural network has Gaussian weights and is sufficiently expansive at each layer, then, with high probability, the empirical risk objective applied to the network output has no spurious local minima or saddle points outside two small neighborhoods around the global optimum and a negative reflection of it. Further, if the output of the network is subject to random Gaussian compressive measurements, then the same conclusion holds with information theoretically optimal sample complexity with respect to the latent code dimensionality. That is, a convergent gradient descent scheme will approximately invert the generative network, even in the presence of a sufficient number of compressive measurements.

As this theoretical work is the first of its kind, it leaves open many important questions deserving further research. Because any particular generative network is unlikely to contain an observed image *exactly*, it is important to establish a similar guarantee to Theorem 1 in the case that the observed image is not in the range of the generative network. This line of work includes establishing noise tolerance and robustness to outliers, both of which have been established for sparsity-based compressed sensing. Such results would provide even further theoretical support for several empirical observations about enforcing generative priors via an optimization over latent code space [48], [43], including empirical robustness of inverting generative models [62]. In particular, it could help explain the significant observation that generative priors can mitigate against adversarial examples [63], [64], which are minor and sometimes imperceptible modifications to images that lead to catastrophic misclassification by neural networks [65]. Robustness against adversarial examples is important for the security of machine learning systems [66], [67], in particular those that will be part of self-driving cars.

In this work, we establish our rigorous global analysis under the assumption that network weights are approximately distributed like Gaussians. There is both empirical and theoretical motivation for this assumption. Empirical analysis reveals that in some trained networks, such as AlexNet, the weight parameters have statistics that are consistent with being approximately Gaussian [59]. Regarding theoretical motivation, recent papers have show that gradient methods for sufficiently overparameterized neural networks provably converge to a point of small or zero loss under various conditions on the training data [68], [69], [70], [71]. A central property unifying these results is that they work in a regime where initially random weights are only slightly perturbed during the training process. Thus, the deterministic condition of our analysis (the WDC) is consistent both with empirical observations and with the neural networks for which current theory is able to establish convergence during the training process.

Additionally, we establish our global analysis under the architectural assumptions that the network is expansive and fully connected. Expansiveness is a natural condition given that generators map low-dimensional, highly-compressed representations to high-dimensional signals with substantial redundancy. In terms of network architecture, fully connected nets are used in a variety of applications, though nets with convolutional structure are much more common in networks that manipulate images. The results of this paper, while directly stated for fully connected nets have already been extended to the case of convolutional generative nets, as proven in [72]. We anticipate further extensions to other realistic architectures are also possible.

We provide in this paper a theoretical framework for studying the enforcement of deep generative priors via empirical risk as a means of regularization on inverse problems. Besides compressive sensing with linear measurements, there are a myriad of inverse problems that may benefit from such an approach. One particularly exciting example is the field of phase retrieval, which is critical in the biological sciences for X-ray crystallography and modern techniques like XFEL-imaging, which is a promising approach that may lead to breakthroughs in understanding of proteins and other molecular structures. Phase retrieval involves recovering vectors from quadratic observations, and enforcing linear sparsity priors subject to such quadratic measurements has been met with potentially fundamental limitations of polynomial time algorithms. In particular, while the theoretically optimal sample complexity of sparse phase retrieval is

$O(s \log n)$, where s is the sparsity of the signal, it is potentially unobtainable via polynomial time algorithms [73], which have so far only produced $O(s^2 \log n)$ efficient reconstruction schemes [74]. This bottleneck in sample complexity makes improvements in signal priors critical for the field of phase retrieval to advance. The present work indicates that it may be possible to use generative priors for problems such as phase retrieval. In fact, the work has already been extended to phase retrieval, where recovery is possible with $O(k)$ measurements, where k is the latent code dimensionality [75]. This could beat sparsity based approaches both because the scaling is linear in the signal's latent dimensionality of the representation, and because k can be smaller than s for the very same signal.

More significantly, the regime of using generative modeling as a means of regularization opens new doors for improving the workflow of biological scientists. In modern phase retrieval, modeling assumptions which aid in lowering sample complexity and increasing SNR are all hand-coded, making the process extremely tedious. In contrast, deep generative modeling simply requires obtaining a dataset of previously reconstructed molecular structures, which are easily available in extensive databases amassed over the years of practice of crystallography [76]. One may envision training generative models on such datasets and using the resulting neural network priors to regularize the inverse problem of phase retrieval, tabula rasa, and potentially more effectively than hand-modeling ever could, as has been witnessed in the field of computer vision. This makes possible recovering the structure of biological molecules without explicit modeling, freeing up scientists to focus on innovating on new imaging modalities instead of grappling with the tedium of hand-coding their prior knowledge to solve the resulting inverse problems. More broadly, combining the power of deep generative modeling with modern methods of optimization and signal recovery, allows potentially paradigm shifting improvements to the empirical sciences, by taking a data-driven artificial intelligence approach to signal recovery.

II. Proofs

The theorems are proven by a concentration argument. We show that $v_{x,x_0} \in \mathbb{R}^k$ concentrates around a particular $h_{x,x_0} \in \mathbb{R}^k$ that is a continuous function of nonzero x, x_0 and is zero only at $x = x_0$ and $x = -\rho_d x_0$. Before we sketch the proof below, we introduce some useful quantities.

In order to analyze which rows of a matrix W are active when computing $\mathrm{relu}(Wx)$, we let

$$W_{+,x} = \mathrm{diag}(Wx > 0)W.$$

For a fixed W, the matrix $W_{+,x}$ zeros out the rows of W that do not have a positive dot product with x. Alternatively put, $W_{+,x}$ contains weights from only the neurons that are active for the input x. We also define $W_{1,+,x} = (W_1)_{+,x} = \mathrm{diag}(W_1 x > 0)W_1$ and

$$W_{i,+,x} = \mathrm{diag}(W_i W_{i-1,+,x} \cdots W_{2,+,x} W_{1,+,x} x > 0)W_i.$$

The matrix $W_{i,+,x}$ consists only of the neurons in the ith layer that are active if the input to the first layer is x. Additionally, it will be useful to control how the operator $x \mapsto W_{+,x}x$ distorts angles. In order to study this, we define

$$g(\theta) := \cos^{-1}\Big(\frac{(\pi - \theta)\cos\theta + \sin\theta}{\pi}\Big). \tag{3}$$

We now specify the choice of v_{x,x_0} as follows. At any $x \in \mathbb{R}^k$ such that G is differentiable at x,

$$\begin{aligned}\nabla f(x) = &(\Pi_{i=d}^1 W_{i,+,x})^t A^t A (\Pi_{i=d}^1 W_{i,+,x})x \\ &- (\Pi_{i=d}^1 W_{i,+,x})^t A^t A (\Pi_{i=d}^1 W_{i,+,x_0})x_0.\end{aligned} \tag{4}$$

Let $w \in \mathbb{R}^k$ be such that G is differentiable at $x + \delta w$ for sufficiently small δ. Such a w exists by the piecewise linearity of G. Let

$$v_{x,x_0} = \begin{cases} \nabla f(x) & G \text{ is differentiable at x,} \\ \lim_{\delta \to 0^+} \nabla f(x + \delta w) & \text{otherwise.} \end{cases} \tag{5}$$

Note that the first part of the definition of v_{x,x_0} can be viewed as the special case of the second part of the definition with $w = 0$. When G is not differentiable at x, multiple values of v_{x,x_0} are consistent with the above definition. These values correspond to the multiple choices of w. The concentration analysis applies simultaneously for all appropriate w because of uniformity in the concentration results below.

A sketch of the proof is as follows:

- The WDC and RRIC imply that

  $$\begin{aligned}v_{x,x_0} \approx &(\Pi_{i=d}^1 W_{i,+,x})^t (\Pi_{i=d}^1 W_{i,+,x})x \\ &- (\Pi_{i=d}^1 W_{i,+,x})^t (\Pi_{i=d}^1 W_{i,+,x_0})x_0 =: \overline{v}_{x,x_0},\end{aligned}$$

  uniformly over nonzero x and x_0. See the proof of Theorem 4 in Section II-C.
- The WDC implies that

  $$\begin{aligned}\overline{v}_{x,x_0} \approx &-\frac{1}{2^d}\Big(\prod_{i=0}^{d-1}\frac{\pi - \overline{\theta}_i}{\pi}\Big)x_0 \\ &+ \frac{1}{2^d}\left[x - \sum_{i=0}^{d-1}\frac{\sin\overline{\theta}_i}{\pi}\Big(\prod_{j=i+1}^{d-1}\frac{\pi - \overline{\theta}_j}{\pi}\Big)\frac{\|x_0\|_2}{\|x\|_2}x\right] =: h_{x,x_0},\end{aligned}$$

  uniformly over nonzero x and x_0, where $\overline{\theta}_i = g(\overline{\theta}_{i-1})$, $\overline{\theta}_0 = \angle(x, x_0)$, and h_{x,x_0} is continuous for nonzero x, x_0. See Sections II-A and II-B.
- Direct analysis shows that $h_{x,x_0} \approx 0$ only within a neighborhood of x_0 and $-\rho_d x_0$. See Section II-D.

- Arguments from probabilistic concentration theory establish that the WDC and RRIC with high probability for Gaussian matrices of appropriate dimensions. See Sections II-E and II-F, respectively. These together establish Proposition 6.

Theorem 1 is the combination of Theorem 4 and Proposition 6.

The proof capitalizes on the structure of the relu nonlinearities because it considers points in the range of the generator G to lie in the union of finitely many subspaces, each given by the range of all possible matrices $W_{i,+,x}$. Probabilistic concentration of these matrices requires a bound on the maximum number of such subspaces. These bounds would be worse or possibly infinite for nonlinearities other than relu. While the nondeterministic result is stated for Gaussian weight matrices W_i, the same analysis would extend to weight matrices whose rows are given by a uniform distribution over a sphere of appropriate radius, as the proof capitalizes on rotational invariance of the neuronal weights.

A. Approximate angle contraction property $W_{+,x}$

In the concentration result of the next section, we will make use of the fact that the angle between $W_{+,x}x$ and $W_{+,y}y$ is approximately $g\big(\angle(x,y)\big)$ if the WDC holds. As Figure II-A shows, g is monotonic and less than the identity. Thus, the mapping $x \mapsto W_{+,x}x$ has an approximate angle contraction property in the sense of the following lemma.

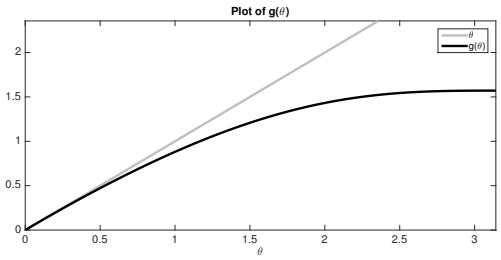

Fig. 1. A plot of $g(\theta)$ from equation (3).

Lemma 7. *Fix $0 < \epsilon < 0.1$. Let $W \in \mathbb{R}^{n\times k}$ satisfy the WDC with constant ϵ. We have $\theta_1 := \angle(W_{+,x}x, W_{+,y}y)$ is well-defined for all $x \neq 0, y \neq 0$, and*

$$|\theta_1 - g(\theta_0)| \leqslant 4\sqrt{\epsilon},$$

where $\theta_0 = \angle(x,y)$ and g is defined by (3).

Proof. It suffices to establish that for all $x \neq 0, y \neq 0$,

$$\left|\cos\theta_1 - \frac{(\pi-\theta_0)\cos\theta_0 + \sin\theta_0}{\pi}\right| \leqslant 5\epsilon.$$

Without loss of generality, consider only $x, y \in S^{k-1}$. By the WDC, $\|W_{+,x}^t W_{+,y} - Q_{x,y}\| \leqslant \epsilon$ for all $x, y \in S^{k-1}$. Let

$$\begin{aligned}\delta_1 &= \langle x, (W_{+,x}^t W_{+,y} - Q_{x,y})y\rangle\\ \delta_2 &= \langle x, (W_{+,x}^t W_{+,x} - I_k/2)x\rangle\\ \delta_3 &= \langle y, (W_{+,y}^t W_{+,y} - I_k/2)y\rangle.\end{aligned}$$

We have $\max(|\delta_1|,|\delta_2|,|\delta_3|) \leqslant \epsilon$ for all $x, y \in S^{k-1}$. As $|\delta_2| \leqslant \epsilon$, $|\delta_3| \leqslant \epsilon$, and $\epsilon < 1/2$, $W_{+,x}x \neq 0$ and $W_{+,y}y \neq 0$. Thus, θ_1 is well-defined. We have

$$\begin{aligned}\cos\theta_1 &= \frac{\langle W_{+,x}x, W_{+,y}y\rangle}{\|W_{+,x}x\|_2\|W_{+,y}y\|_2}\\ &= \frac{\langle x, W_{+,x}^t W_{+,y}y\rangle}{\sqrt{\langle x, W_{+,x}^t W_{+,x}x\rangle\langle y, W_{+,x}^t W_{+,x}y\rangle}}\\ &= \frac{\langle x, Q_{x,y}y\rangle + \delta_1}{\frac{1}{2}\sqrt{(1+2\delta_2)(1+2\delta_3)}}\end{aligned}$$

Thus,

$$\begin{aligned}&|\cos\theta_1 - 2\langle x, Q_{x,y}y\rangle|\\ &\leqslant 2|\langle x, Q_{x,y}y\rangle|\left|1 - \frac{1}{\sqrt{(1+2\delta_2)(1+2\delta_3)}}\right|\\ &\quad + 2\delta_1\frac{1}{\sqrt{(1+2\delta_2)(1+2\delta_3)}}\\ &\leqslant \left|1 - \frac{1}{(1-2\epsilon)}\right| + 2\epsilon\frac{1}{(1-2\epsilon)}\\ &\leqslant 5\epsilon\end{aligned}$$

where the second line follows as $2|\langle x, Q_{x,y}y\rangle| \leqslant 2\|Q_{x,y}\| \leqslant 1$ and $\max(|\delta_1|,|\delta_2|,|\delta_3|) \leqslant \epsilon$, and the last line follows because $\epsilon < 0.1$. The proof is concluded by noting that $2\langle x, Q_{x,y}y\rangle = \frac{1}{\pi}\big[(\pi-\theta_0)\cos\theta_0 + \sin\theta_0\big]$. □

B. Concentration of terms without compression

At points of differentiability x, we have

$$\begin{aligned}v_{x,x_0} =&(\Pi_{i=d}^1 W_{i,+,x})^t A^t A(\Pi_{i=d}^1 W_{i,+,x})x\\ &- (\Pi_{i=d}^1 W_{i,+,x})^t A^t A(\Pi_{i=d}^1 W_{i,+,x_0})x_0.\end{aligned}$$

In this section, we prove that the WDC establishes concentration for the terms in v_{x,x_0} uniformly in x and x_0 in the compressionless case of $A = I_n$. The case with compression will use the concentration of these terms too.

Lemma 8. *Fix $0 < \epsilon < d^{-4}/(16\pi)^2$ and $d \geqslant 2$. Suppose that $W_i \in \mathbb{R}^{n_i\times n_{i-1}}$ satisfies the WDC with constant ϵ for $i = 1\ldots d$. Define $\tilde{h}_{x,y}$ as*

$$\frac{1}{2^d}\left[\Big(\prod_{i=0}^{d-1}\frac{\pi-\overline{\theta}_i}{\pi}\Big)y + \sum_{i=0}^{d-1}\frac{\sin\overline{\theta}_i}{\pi}\Big(\prod_{j=i+1}^{d-1}\frac{\pi-\overline{\theta}_j}{\pi}\Big)\frac{\|y\|_2}{\|x\|_2}x\right],$$

where $\bar{\theta}_i = g(\bar{\theta}_{i-1})$ *for* g *given by* (3) *and* $\bar{\theta}_0 = \angle(x, y)$. *For all* $x \neq 0$ *and* $y \neq 0$,

$$\|(\Pi_{i=d}^1 W_{i,+,x})^t (\Pi_{i=d}^1 W_{i,+,y}) y - \tilde{h}_{x,y}\|_2 \leqslant 24 \frac{d^3\sqrt{\epsilon}}{2^d} \|y\|_2, \tag{6}$$

$$\left\langle (\Pi_{i=d}^1 W_{i,+,x}) x, (\Pi_{i=d}^1 W_{i,+,y}) y \right\rangle \geqslant \frac{1}{4\pi} \frac{1}{2^d} \|x\|_2 \|y\|_2. \tag{7}$$

Proof.

Part I: Assembling some useful bounds.

Define $x_0 = x$, $y_0 = y$,

$$\begin{aligned} x_d &:= \left(\Pi_{i=d}^1 W_{i,+,x}\right) x \\ &= (W_{d,+,x} W_{d-1,+,x} \dots W_{1,+,x}) x \\ &= W_{d,+,x} x_{d-1} = (W_d)_{+,x_{d-1}} x_{d-1}, \end{aligned}$$

and analogously $y_d = \left(\Pi_{i=d}^1 W_{i,+,y}\right) y$, where $(W_i)_{+,x} = \operatorname{diag}(W_i x > 0) W_i$.

By the WDC, we have for all $i = 1 \dots d$, for all $x \neq 0, y \neq 0$,

$$\|(W_i)_{+,x}^t (W_i)_{+,y} - Q_{x,y}\| \leqslant \epsilon, \tag{8}$$

In particular, $\|W_{i,+,x}^t W_{i,+,y} - Q_{x_{i-1},y_{i-1}}\| \leqslant \epsilon$. We now detail several bounds that follow from the WDC. Most immediately, we have that for all $x \neq 0$ and for all $i = 1 \dots d$,

$$\left\| W_{i,+,x}^t W_{i,+,x} - \frac{1}{2} I_{n_{i-1}} \right\| \leqslant \epsilon, \tag{9}$$

and consequently,

$$\frac{1}{2} - \epsilon \leqslant \|W_{i,+,x}\|^2 \leqslant \frac{1}{2} + \epsilon.$$

Hence,

$$\begin{aligned} &\|\Pi_{i=d}^1 W_{i,+,x}\| \|\Pi_{i=d}^1 W_{i,+,y}\| \\ &\qquad \leqslant \frac{1}{2^d}(1+2\epsilon)^d = \frac{1}{2^d} e^{d \log(1+2\epsilon)} \\ &\qquad \leqslant \frac{1+4\epsilon d}{2^d}, \end{aligned} \tag{10}$$

where we used that $\log(1+z) \leqslant z$, $e^z \leqslant 1 + 2z$ for $z < 1$, and $2d\epsilon < 1$. We also have for all $x \neq 0$ that

$$\sqrt{\frac{1}{2} - \epsilon} \|x_{i-1}\|_2 \leqslant \|x_i\|_2 \leqslant \sqrt{\frac{1}{2} + \epsilon} \|x_{i-1}\|_2. \tag{11}$$

Thus, we have that for all $x, y \neq 0$,

$$\left(\frac{1-2\epsilon}{1+2\epsilon}\right)^{d/2} \frac{\|y\|_2}{\|x\|_2} \leqslant \frac{\|y_d\|_2}{\|x_d\|_2} \leqslant \left(\frac{1+2\epsilon}{1-2\epsilon}\right)^{d/2} \frac{\|y\|_2}{\|x\|_2}.$$

Note that if $4\epsilon d < 1$,

$$\begin{aligned} \left(\frac{1+2\epsilon}{1-2\epsilon}\right)^{d/2} &\leqslant (1+8\epsilon)^{d/2} = e^{\frac{d}{2}\log(1+8\epsilon)} \\ &\leqslant e^{4d\epsilon} \leqslant 1 + 8d\epsilon \end{aligned}$$

where we used that $\frac{1+z}{1-z} \leqslant 1 + 4z$ for $0 \leqslant z \leqslant \frac{1}{2}$, $e^z \leqslant 1 + 2z$ for $0 \leqslant z \leqslant 1$, and $4d\epsilon < 1$. As $1 - z \leqslant (1+z)^{-1}$ for $z > 0$, we also have

$$\left(\frac{1-2\epsilon}{1+2\epsilon}\right)^{d/2} \geqslant 1 - 8d\epsilon,$$

and thus, if $4\epsilon d < 1$,

$$(1 - 8d\epsilon) \frac{\|y\|_2}{\|x\|_2} \leqslant \frac{\|y_d\|_2}{\|x_d\|_2} \leqslant (1 + 8d\epsilon) \frac{\|y\|_2}{\|x\|_2}. \tag{12}$$

By Lemma 7, the WDC implies that $\theta_{V_{+,x}x, V_{+,y}y}$ is well-defined for all nonzero x and y, and for $V = W_1, \dots, W_d$, and for all $x \neq 0, y \neq 0, V = W_1, \dots, W_d$,

$$|\theta_{V_{+,x}x, V_{+,y}y} - g(\theta_{x,y})| \leqslant \delta, \tag{13}$$

where $g(\theta)$ is given by (3), and $\delta := 4\sqrt{\epsilon}$. Define $\theta_i := \angle(x_i, y_i) \in [0, \pi]$. Note that by (13), we have[3] $\theta_d = g(\theta_{d-1}) + O_1(\delta)$ for all d, and thus $\theta_d = g(g(\cdots g(g(\theta_0) + O_1(\delta)) + O_1(\delta) \cdots) + O_1(\delta)) + O_1(\delta)$. Because $|g'(\theta)| \leqslant 1$ for all θ and because $\bar{\theta}_d = g(g(\cdots g(\theta_0) \cdots)) = g^{\circ d}(\theta_0)$, we have

$$|\theta_d - \bar{\theta}_d| \leqslant d\delta = 4d\sqrt{\epsilon}. \tag{14}$$

Part II: Establishing (7)

We have that $\cos\theta_d \geqslant 3/(4\pi)$ by combining (14), $\bar{\theta}_d \leqslant \cos^{-1}(\frac{1}{\pi})$ for $d \geqslant 2$, and $4\pi d 4\sqrt{\epsilon} \leqslant 1$. Additionally, by (11), we have $\|x_d\|_2 \|y_d\|_2 \geqslant \|x\|_2 \|y\|_2 \left(\frac{1}{2} - \epsilon\right)^d \geqslant \|x\|_2 \|y\|_2 \frac{1-2d\epsilon}{2^d}$, where the last inequality holds as $\epsilon \leqslant 1/2$. If $2d\epsilon \leqslant 2/3$, we have

$$\begin{aligned} \left\langle (\Pi_{i=d}^1 W_{i,+,x}) x, (\Pi_{i=d}^1 W_{i,+,y}) y \right\rangle &= \cos(\theta_d) \|x_d\|_2 \|y_d\|_2 \\ &\geqslant \frac{1}{4\pi} \frac{1}{2^d} \|x\|_2 \|y\|_2, \end{aligned}$$

which establishes (7).

Part III: Establishing (6)

This proof will proceed by setting up and solving a recurrence relation. We will use that

$$\begin{aligned} &\Gamma_d = s_d \Gamma_{d-1} + r_d, \ \Gamma_0 = y \\ &\quad \Rightarrow \quad \Gamma_d = \left(\prod_{i=1}^d s_i\right) y + \sum_{i=1}^d \left(r_i \prod_{j=i+1}^d s_j\right). \end{aligned} \tag{15}$$

First, we derive a recurrence relation and solve for $(\Pi_{i=d}^1 W_{i,+,x})^t (\Pi_{i=d}^1 W_{i,+,x})$. We have

$$\begin{aligned} M_d &:= (\Pi_{i=d}^1 W_{i,+,x})^t (\Pi_{i=d}^1 W_{i,+,x}) \\ &= (\Pi_{i=d-1}^1 W_{i,+,x})^t \left(\frac{1}{2} I_{n_{d-1}} + O_1(\epsilon)\right) (\Pi_{i=d-1}^1 W_{i,+,x}) \\ &= \frac{1}{2} (\Pi_{i=d-1}^1 W_{i,+,x})^t (\Pi_{i=d-1}^1 W_{i,+,x}) \\ &\quad + O_1(\epsilon \Pi_{i=1}^{d-1} \|W_{i,+,x}\|^2) \\ &= \frac{1}{2} M_{d-1} + O_1\left(\epsilon \frac{1 + 4\epsilon(d-1)}{2^{d-1}}\right), \end{aligned}$$

[3] See Section I-B for the meaning of O_1.

where the first equality follows by (9), and the third equality follows from (10), as $2\epsilon d \leqslant 1$. Solving this recurrence relation with $M_0 = I_{n_0}$ by (15), we get that if $4d\epsilon \leqslant 1$, then

$$\begin{aligned}
&(\Pi_{i=d}^{1} W_{i,+,x})^t (\Pi_{i=d}^{1} W_{i,+,x}) \\
&= \frac{1}{2^d} I_{n_0} + \sum_{i=1}^{d} O_1\Big(\epsilon \frac{1+4\epsilon(i-1)}{2^{i-1}}\Big) \frac{1}{2^{d-i}} \\
&= \frac{1}{2^d} I_{n_0} + \frac{4\epsilon d}{2^d} O_1(1). \qquad (16)
\end{aligned}$$

Thus,

$$(\Pi_{i=d}^{1} W_{i,+,x})^t (\Pi_{i=d}^{1} W_{i,+,x}) x = \frac{1}{2^d} x + O_1\Big(\frac{4\epsilon d}{2^d}\Big) \|x\|_2 \qquad (17)$$

Next, we derive a recurrence relation for $\Gamma_d := (\Pi_{i=d}^{1} W_{i,+,x})^t (\Pi_{i=d}^{1} W_{i,+,y}) y$.

$$\begin{aligned}
\Gamma_d &= (\Pi_{i=d-1}^{1} W_{i,+,x})^t (W_{d,+,x}^t W_{d,+,y}) (\Pi_{i=d-1}^{1} W_{i,+,y}) y \\
&= (\Pi_{i=d-1}^{1} W_{i,+,x})^t \Big(\frac{\pi - \theta_{d-1}}{2\pi} I_{n_{d-1}} \\
&\quad + \frac{\sin\theta_{d-1}}{2\pi} M_{\hat{x}_{d-1} \leftrightarrow \hat{y}_{d-1}} + O_1(\epsilon) \Big) \left(\Pi_{i=d-1}^{1} W_{i,+,y}\right)(y) \\
&= \frac{\pi - \theta_{d-1}}{2\pi} \Gamma_{d-1} \\
&\quad + \frac{\sin\theta_{d-1}}{2\pi} \frac{\|y_{d-1}\|_2}{\|x_{d-1}\|_2} (\Pi_{i=d-1}^{1} W_{i,+,x})^t (\Pi_{i=d-1}^{1} W_{i,+,x}) x \\
&\quad + \epsilon \Big(\frac{1+4\epsilon d}{2^{d-1}} \Big) \|y\|_2 O_1(1) \\
&= \frac{\pi - \theta_{d-1}}{2\pi} \Gamma_{d-1} + \frac{\sin\theta_{d-1}}{2\pi} \frac{\|y_{d-1}\|_2}{\|x_{d-1}\|_2} \frac{x}{2^{d-1}} \\
&\quad + \frac{1}{2\pi} \frac{\|y_{d-1}\|_2}{\|x_{d-1}\|_2} \frac{4d\epsilon}{2^{d-1}} \|x\|_2 O_1(1) + \epsilon \Big(\frac{1+4\epsilon d}{2^{d-1}} \Big) \|y\|_2 O_1(1) \\
&= \frac{\pi - \theta_{d-1}}{2\pi} \Gamma_{d-1} + \frac{\sin\theta_{d-1}}{\pi} \frac{\|y_{d-1}\|_2}{\|x_{d-1}\|_2} \frac{x}{2^d} \\
&\quad + \frac{1+8d\epsilon}{2\pi} \frac{4d\epsilon}{2^{d-1}} \|y\|_2 O_1(1) + \epsilon \Big(\frac{2}{2^{d-1}} \Big) \|y\|_2 O_1(1) \\
&= \frac{\pi - \theta_{d-1}}{2\pi} \Gamma_{d-1} + \frac{\sin\theta_{d-1}}{\pi} \frac{\|y_{d-1}\|_2}{\|x_{d-1}\|_2} \frac{x}{2^d} \\
&\quad + \epsilon \Big(\frac{3}{2\pi} \frac{2 \cdot 4d}{2^d} + \frac{4}{2^d} \Big) \|y\|_2 O_1(1) \\
&= \frac{\pi - \theta_{d-1}}{2\pi} \Gamma_{d-1} + \frac{\sin\theta_{d-1}}{\pi} \frac{\|y_{d-1}\|_2}{\|x_{d-1}\|_2} \frac{x}{2^d} \\
&\quad + \frac{8}{2^d} d\epsilon \|y\|_2 O_1(1) \\
&= \frac{\pi - \theta_{d-1}}{2\pi} \Gamma_{d-1} + \frac{\sin\theta_{d-1}}{\pi} \frac{\|y\|_2}{\|x\|_2} \frac{x}{2^d} \\
&\quad + \frac{1}{2^d} \frac{8d\epsilon}{\pi} \|y\|_2 O_1(1) + \frac{8}{2^d} d\epsilon \|y\|_2 O_1(1) \\
&= \frac{\pi - \theta_{d-1}}{2\pi} \Gamma_{d-1} + \frac{\sin\theta_{d-1}}{\pi} \frac{\|y\|_2}{\|x\|_2} \frac{x}{2^d} \\
&\quad + 11 d\epsilon \frac{\|y\|_2}{2^d} O_1(1) \\
&= \Big(\frac{\pi - \overline{\theta}_{d-1}}{2\pi} + O_1\big(\frac{d\delta}{2\pi}\big) \Big) \Gamma_{d-1} + \frac{\sin\overline{\theta}_{d-1}}{\pi} \frac{\|y\|_2}{\|x\|_2} \frac{x}{2^d} \\
&\quad + 2d\delta \frac{\|y\|_2}{2^d} O_1(1) \qquad (18)
\end{aligned}$$

where the second line follows from (8), the definition $W_{d,+,y} = (W_d)_{+,y_{d-1}}$, and the definition of Q_{x_d,y_d} in (2); the third line follows by the definition of $M_{\hat{x}_{d-1} \leftrightarrow \hat{y}_{d-1}}$ in (2), the bound (10), and the definition of x_{d-1}; the fourth line uses (16); the fifth line follows from $4\epsilon d \leqslant 1$ and (12); the eighth line follows from (12); and the last line follows from (14), $\delta = 4\sqrt{\epsilon}$, and $11\sqrt{\epsilon} \leqslant 4$.

Solving the recurrence relation (18) using (15), we get that

$$\begin{aligned}
\Gamma_d &= \prod_{i=1}^{d} \Big[\frac{\pi - \overline{\theta}_{i-1}}{2\pi} + O_1\big(\frac{i\delta}{2\pi}\big) \Big] y \\
&\quad + \sum_{i=1}^{d} \Big(\frac{\sin\overline{\theta}_{i-1}}{\pi} \frac{\|y\|_2}{\|x\|_2} \frac{x}{2^i} + \frac{2i\delta}{2^i} \|y\|_2 O_1(1) \Big) \\
&\quad \times \prod_{j=i+1}^{d} \Big(\frac{\pi - \overline{\theta}_{j-1}}{2\pi} + O_1\big(\frac{j\delta}{2\pi}\big) \Big) \qquad (19)
\end{aligned}$$

First, we control the first term of Γ_d in (19). We have

$$\begin{aligned}
&\left| \prod_{i=1}^{d} \Big[\frac{\pi - \overline{\theta}_{i-1}}{2\pi} + O_1\big(\frac{i\delta}{2\pi}\big) \Big] y - \prod_{i=1}^{d} \Big[\frac{\pi - \overline{\theta}_{i-1}}{2\pi} \Big] y \right| \\
&\leqslant \Big(\prod_{i=1}^{d} \Big[\frac{1}{2} + \frac{i\delta}{2\pi} \Big] - \frac{1}{2^d} \Big) \|y\|_2 \\
&\leqslant \frac{1}{2^d} \Big[\Big(1 + \frac{d\delta}{\pi} \Big)^d - 1 \Big] \|y\|_2 \\
&\leqslant \frac{1}{2^d} \Big[e^{d \log(1 + d\delta/\pi)} - 1 \Big] \|y\|_2 \\
&\leqslant \frac{1}{2^d} \Big[e^{d^2\delta/\pi} - 1 \Big] \|y\|_2 \\
&\leqslant \frac{2}{2^d} \frac{d^2\delta}{\pi} \|y\|_2 \leqslant \frac{d^2\delta}{2^d} \|y\|_2, \qquad (20)
\end{aligned}$$

where the first inequality follows by Lemma 9 and as $\frac{\pi - \overline{\theta}_i}{2\pi} \in [0, 1/2]$; and fifth inequality follows because $e^{d^2\delta/\pi} \leqslant 1 + 2\frac{d^2\delta}{\pi}$ if $d^2\delta/\pi \leqslant 1$.

Next, we control the second term of Γ_d in (19). Similar to the calculation above, we have

$$\left| \prod_{j=i+1}^{d} \Big[\frac{\pi - \overline{\theta}_{j-1}}{2\pi} + O_1\big(\frac{j\delta}{2\pi}\big) \Big] - \prod_{j=i+1}^{d} \frac{\pi - \overline{\theta}_{j-1}}{2\pi} \right| \leqslant \frac{d^2\delta}{2^{d-i}} \qquad (21)$$

if $d^2\delta/\pi \leqslant 1$. We now get that the second term of (19) is

$$\begin{aligned}
&\sum_{i=1}^{d}\Big[\frac{\sin\overline{\theta}_{i-1}}{\pi}\frac{\|y\|_2}{\|x\|_2}\frac{x}{2^i}+\frac{2i\delta}{2^i}\|y\|_2 O_1(1)\Big]\\
&\qquad\times\prod_{j=i+1}^{d}\left(\frac{\pi-\overline{\theta}_{j-1}}{2\pi}+O_1\Big(\frac{j\delta}{2\pi}\Big)\right)\\
&=\sum_{i=1}^{d}\Big[\frac{\sin\overline{\theta}_{i-1}}{\pi}\frac{\|y\|_2}{\|x\|_2}\frac{x}{2^i}+\frac{2i\delta}{2^i}\|y\|_2 O_1(1)\Big]\\
&\qquad\times\Big[\Big(\prod_{j=i+1}^{d}\frac{\pi-\overline{\theta}_{j-1}}{2\pi}\Big)+O_1\Big(\frac{d^2\delta}{2^{d-i}}\Big)\Big]\\
&=\left[\sum_{i=1}^{d}\frac{\sin\overline{\theta}_{i-1}}{\pi}\frac{\|y\|_2}{\|x\|_2}\frac{x}{2^d}\prod_{j=i+1}^{d}\frac{\pi-\overline{\theta}_{j-1}}{\pi}\right]\\
&\qquad+O_1\Big(\frac{5d^3\delta}{2^d}\Big)\|y\|_2 \qquad (22)
\end{aligned}$$

where the first equality follows by (21) and $d^2\delta/\pi \leqslant 1$, and the second equality follows by expanding the terms and using $d^2\delta \leqslant 1$.

Combining (20) and (22), we get

$$\begin{aligned}
\Gamma_d &= \frac{1}{2^d}\left[\prod_{i=1}^{d}\frac{\pi-\overline{\theta}_{i-1}}{\pi}\right]y\\
&+\frac{1}{2^d}\left[\sum_{i=1}^{d}\frac{\sin\overline{\theta}_{i-1}}{\pi}\Big(\prod_{j=i+1}^{d}\frac{\pi-\overline{\theta}_{j-1}}{\pi}\Big)\frac{\|y\|_2}{\|x\|_2}x\right]\\
&+O_1\Big(\frac{6d^3\delta}{2^d}\Big)\|y\|_2.
\end{aligned}$$

We complete the proof of (6) by combining this equality with (17), and $\delta = 4\sqrt{\epsilon}$.

□

In the proof of the previous lemma, we used the following technical result.

Lemma 9. *Let $d \in \mathbb{N}$ and let $0 \leqslant r_i \leqslant r_{max}$ for $i = 1 \dots d$. We have*

$$\left|\prod_{i=1}^{d}(r_i+t_i)-\prod_{i=1}^{d}r_i\right| \leqslant \prod_{i=1}^{d}(r_{max}+|t_i|)-r_{max}^d.$$

Proof. First, we establish that

$$\left|\prod_{i=1}^{d}(r_i+t_i)-\prod_{i=1}^{d}r_i\right| \leqslant \prod_{i=1}^{d}(r_i+|t_i|)-\prod_{i=1}^{d}r_i.$$

This follows by noting that for $\delta \in [0,1]$,

$$\begin{aligned}
\left|\frac{d}{d\delta}\Big[\prod_{i=1}^{d}(r_i+\delta t_i)\Big]\right| &= \left|\sum_{i=1}^{d}t_i\prod_{j\neq i}(r_j+\delta t_j)\right|\\
&\leqslant \sum_{i=1}^{d}|t_i|\prod_{j\neq i}(r_j+\delta|t_j|)\\
&= \frac{d}{d\delta}\Big(\prod_{i=1}^{d}(r_i+\delta|t_i|)\Big),
\end{aligned}$$

and integrating over $\delta \in [0,1]$. Next, we establish that

$$\prod_{i=1}^{d}(r_i+|t_i|)-\prod_{i=1}^{d}r_i \leqslant \prod_{i=1}^{d}(r_{\max}+|t_i|)-r_{\max}^d.$$

This follows by noting that for all $k = 1 \dots d$,

$$\begin{aligned}
&\frac{\partial}{\partial r_k}\Big[\prod_{i=1}^{d}(r_i+|t_i|)-\prod_{i=1}^{d}r_i\Big]\\
&\quad=\prod_{j\neq k}(r_j+|t_j|)-\prod_{j\neq k}r_j \geqslant 0.
\end{aligned}$$

□

C. Proof of Deterministic Theorem

The proof of Theorem 4 follows the outline provided in Section II.

Proof of Theorem 4. Recall that

$$v_{x,x_0} = \begin{cases}\nabla f(x) & G \text{ is differentiable at x,}\\ \lim_{\delta\to 0^+}\nabla f(x+\delta w) & \text{otherwise,}\end{cases}$$

where G is differentiable at $x+\delta w$ for sufficiently small δ. Such a w exists by the piecewise linearity of G, and any such w can be selected arbitrarily. Also, recall that

$$\begin{aligned}
\nabla f(x) &= (\Pi_{i=d}^{1}W_{i,+,x})^t A^t A(\Pi_{i=d}^{1}W_{i,+,x})x\\
&\quad-(\Pi_{i=d}^{1}W_{i,+,x})^t A^t A(\Pi_{i=d}^{1}W_{i,+,x_0})x_0.
\end{aligned}$$

Let

$$\begin{aligned}
\overline{v}_{x,x_0} &= (\Pi_{i=d}^{1}W_{i,+,x})^t(\Pi_{i=d}^{1}W_{i,+,x})x\\
&\quad-(\Pi_{i=d}^{1}W_{i,+,x})^t(\Pi_{i=d}^{1}W_{i,+,x_0})x_0,\\
h_{x,x_0} &= -\frac{1}{2^d}\Big(\prod_{i=0}^{d-1}\frac{\pi-\overline{\theta}_i}{\pi}\Big)x_0\\
&\quad+\frac{1}{2^d}\left[x-\sum_{i=0}^{d-1}\frac{\sin\overline{\theta}_i}{\pi}\Big(\prod_{j=i+1}^{d-1}\frac{\pi-\overline{\theta}_j}{\pi}\Big)\frac{\|x_0\|_2}{\|x\|_2}x\right],\\
S_{\epsilon,x_0} &= \Big\{x\in\mathbb{R}^k \mid \|h_{x,x_0}\|_2 \leqslant \frac{1}{2^d}\epsilon\max(\|x\|_2,\|x_0\|_2)\Big\},
\end{aligned}$$

where $\overline{\theta}_i = g(\overline{\theta}_{i-1})$ and $\overline{\theta}_0 = \angle(x,x_0)$. For brevity of notation, write $v_x = v_{x,x_0}, \overline{v}_x = \overline{v}_{x,x_0}$, and $h_x = h_{x,x_0}$.

The WDC implies that for all $x \neq 0$ and for all $i = 1 \dots d$,

$$\|W_{i,+,x}\|^2 \leqslant \frac{1}{2} + \epsilon. \tag{23}$$

Now, we establish that for all differentiable points $x \in \mathbb{R}^k$,

$$\|\nabla f(x) - \overline{v}_x\| \leqslant 2\epsilon\Big(\frac{1}{2} + \epsilon\Big)^d \max(\|x\|_2, \|x_0\|_2). \tag{24}$$

At $x \in \mathbb{R}^k$ such that G is differentiable at x, the local linearity of G gives that $G(x+z) - G(x) = (\Pi_{i=d}^1 W_{i,+,x})z$ for any sufficiently small $z \in \mathbb{R}^k$. By the RRIC, we have

$$\begin{aligned}
&|\langle A\Pi_{i=d}^1 W_{i,+,x} z, A\Pi_{i=d}^1 W_{i,+,y}\tilde{z}\rangle \\
&\quad - \langle \Pi_{i=d}^1 W_{i,+,x} z, \Pi_{i=d}^1 W_{i,+,y}\tilde{z}\rangle| \\
&\leqslant \epsilon \Pi_{i=1}^d \|W_{i,+,x}\|\|W_{i,+,y}\|\|z\|_2\|\tilde{z}\|_2.
\end{aligned} \tag{25}$$

for all $z, \tilde{z}$, which, together with (23), implies (24).

Similarly, the WDC implies by Lemma 8 and $\epsilon < 1/(16\pi d^2)^2$ that for all nonzero $x, x_0, y \in \mathbb{R}^k$,

$$\|\overline{v}_x - h_x\|_2 \leqslant K\frac{d^3\sqrt{\epsilon}}{2^d}\max(\|x\|_2, \|x_0\|_2), \text{ and} \tag{26}$$

$$\big\langle (\Pi_{i=d}^1 W_{i,+,x})x, (\Pi_{i=d}^1 W_{i,+,y})y\big\rangle \geqslant \frac{1}{4\pi}\frac{1}{2^d}\|x\|_2\|y\|_2. \tag{27}$$

Thus, we have, for all $x \neq 0, x_0 \neq 0$,

$$\begin{aligned}
\|v_x - h_x\|_2 a &= \lim_{\delta\to 0^+}\|\nabla f(x+\delta w) - h_{x+\delta w}\|_2 \\
&\leqslant \lim_{\delta\to 0^+}\big(\|\nabla f(x+\delta w) - \overline{v}_{x+\delta w}\|_2 + \|\overline{v}_{x+\delta w} - h_{x+\delta w}\|_2\big) \\
&\leqslant \sqrt{\epsilon}\Big(2\frac{(1+2\epsilon)^d}{2^d} + K\frac{d^3}{2^d}\Big)\max(\|x\|_2, \|x_0\|_2) \\
&\leqslant \sqrt{\epsilon}\tilde{K}\frac{d^3}{2^d}\max(\|x\|_2, \|x_0\|_2),
\end{aligned} \tag{28}$$

for some universal constant $\tilde{K}$, where the first inequality follows by the definition of v_x and the continuity of h_x for nonzero x; the second inequality follows by combining (23), (24), (26), and as $2d\epsilon \leqslant 1 \Rightarrow (1+2\epsilon)^d \leqslant e^{2\epsilon d} \leqslant 1 + 4\epsilon d$.

Note that the one-sided directional derivative of f in the direction of $y \neq 0$ at x is $D_y f(x) = \lim_{t\to 0^+}\frac{f(x+ty)-f(x)}{t}$. Due to the continuity and piecewise linearity of the function

$$G(x) = \mathrm{relu}(W_d \dots \mathrm{relu}(W_2\,\mathrm{relu}(W_1 x))),$$

we have that for any $x, y \neq 0$ that there exists a sequence $\{x_n\} \to x$ such that f is differentiable at each x_n and $D_y f(x) = \lim_{n\to\infty}\nabla f(x_n)\cdot y$. Thus, as $\nabla f(x_n) = v_{x_n}$,

$$D_{-v_x} f(x) = -\lim_{n\to\infty} v_{x_n}\cdot\frac{v_x}{\|v_x\|}.$$

Now, we write

$$\begin{aligned}
v_{x_n}\cdot v_x &= h_{x_n}\cdot h_x + (v_{x_n} - h_{x_n})\cdot h_x \\
&\quad + h_{x_n}\cdot(v_x - h_x) + (v_{x_n} - h_{x_n})\cdot(v_x - h_x) \\
&\geqslant h_{x_n}\cdot h_x - \|v_{x_n} - h_{x_n}\|_2\|h_x\|_2 \\
&\quad - \|h_{x_n}\|_2\|v_x - h_x\|_2 - \|v_{x_n} - h_{x_n}\|_2\|v_x - h_x\|_2 \\
&\geqslant h_{x_n}\cdot h_x - \sqrt{\epsilon}\tilde{K}\frac{d^3}{2^d}\max(\|x_n\|_2, \|x_0\|_2)\|h_x\|_2 \\
&\quad - \sqrt{\epsilon}\tilde{K}\frac{d^3}{2^d}\max(\|x\|_2, \|x_0\|_2)\|h_{x_n}\|_2 \\
&\quad - \epsilon\Big(\tilde{K}\frac{d^3}{2^d}\Big)^2\max(\|x_n\|_2, \|x_0\|_2)\max(\|x\|_2, \|x_0\|_2),
\end{aligned}$$

where the first inequality follows from the triangle inequality, and the second inequality follows by (28). As h_x is continuous in x for all nonzero x, we have for any $x \in S^c_{8\sqrt{\epsilon}\tilde{K}d^3}$,

$$\begin{aligned}
\lim_{n\to\infty} v_{x_n}\cdot v_x &\geqslant \|h_x\|_2^2 - 2\sqrt{\epsilon}\tilde{K}\frac{d^3}{2^d}\|h_x\|_2\max(\|x\|_2, \|x_0\|_2) \\
&\quad - \epsilon\Big[\tilde{K}\frac{d^3}{2^d}\Big]^2\max(\|x\|_2, \|x_0\|_2)^2 \\
&= \frac{\|h_x\|_2}{2}\Big[\|h_x\|_2 - 4\sqrt{\epsilon}\Big(\tilde{K}\frac{d^3}{2^d}\Big)\max(\|x\|_2, \|x_0\|_2)\Big] \\
&\quad + \frac{1}{2}\Big[\|h_x\|^2 - 2\cdot\epsilon\Big(\tilde{K}\frac{d^3}{2^d}\Big)^2\max(\|x\|_2, \|x_0\|_2)^2\Big] \\
&\geqslant \frac{\|h_x\|_2}{2}4\sqrt{\epsilon}\Big(\tilde{K}\frac{d^3}{2^d}\Big)\max(\|x\|_2, \|x_0\|_2) \\
&\geqslant \frac{7/8\|v_x\|_2}{2}4\sqrt{\epsilon}\Big(\tilde{K}\frac{d^3}{2^d}\Big)\max(\|x\|_2, \|x_0\|_2),
\end{aligned}$$

where the last inequality uses (28) and the definition of $S_{8\sqrt{\epsilon}\tilde{K}d^3}$. We conclude $D_{-v_x} f(x) < 0$ for all nonzero $x \in S^c_{8\sqrt{\epsilon}\tilde{K}d^3}$.

It remains to prove that $\forall x \neq 0$, $D_x f(0) < 0$. We compute that

$$\begin{aligned}
D_x f(0)\cdot\|x\|_2 &= -\langle A(\Pi_{i=d}^1 W_{i,+,x})x, A(\Pi_{i=d}^1 W_{i,+,x_0})x_0\rangle \\
&= -\langle x, (\Pi_{i=d}^1 W_{i,+,x})^t A^t A(\Pi_{i=d}^1 W_{i,+,x_0})x_0\rangle \\
&= -\langle x, (\Pi_{i=d}^1 W_{i,+,x})^t(A^tA - I_{n_2})(\Pi_{i=d}^1 W_{i,+,x_0})x_0\rangle \\
&\quad - \langle(\Pi_{i=d}^1 W_{i,+,x})x, (\Pi_{i=d}^1 W_{i,+,x_0})x_0\rangle \\
&\leqslant \epsilon\frac{(1+2\epsilon)^d}{2^d}\|x\|_2\|x_0\|_2 - \frac{1/(4\pi)}{2^d}\|x\|_2\|x_0\|_2 \\
&\leqslant \frac{2\epsilon}{2^d}\|x\|_2\|x_0\|_2 - \frac{1/(4\pi)}{2^d}\|x\|_2\|x_0\|_2
\end{aligned}$$

where the first inequality holds by (25), (23), and (27); and the second inequality follows from $4\epsilon d \leqslant 1$. Thus, for $\epsilon < \frac{1}{16\pi}$, $D_x f(0) < -\frac{1}{8\pi 2^d}\|x_0\|_2$.

The proof is finished by applying Lemma 10 and $8\pi d^6\sqrt{8\sqrt{\epsilon}\tilde{K}d^3} \leqslant 1$ to get

$$S_{8\sqrt{\epsilon}\tilde{K}d^3} \subset \mathcal{B}(x_0, 56d\sqrt{8\sqrt{\epsilon}\tilde{K}d^3}\|x_0\|_2) \cup \mathcal{B}(-\rho_d x_0, 500d^{11}\sqrt{8\sqrt{\epsilon}\tilde{K}d^3}\|x_0\|_2).$$

□

D. Control of the zeros of h_{x,x_0}

We now show that h_{x,x_0} is away from zero outside of a neighborhood of x_0 and $-\rho_d x_0$.

Lemma 10. *Suppose $8\pi d^6\sqrt{\epsilon} \leqslant 1$. Let S_{ϵ,x_0} be*

$$\left\{x \neq 0 \in \mathbb{R}^k \mid \|h_{x,x_0}\|_2 \leqslant \frac{1}{2^d}\epsilon \max\left(\|x\|_2, \|x_0\|_2\right)\right\},$$

where d is an integer greater than 1 and let h_{x,x_0} be defined by

$$\frac{h_{x,x_0}}{\|x_0\|_2} = -\frac{1}{2^d}\Big(\prod_{i=0}^{d-1}\frac{\pi-\overline{\theta}_i}{\pi}\Big)\hat{x}_0 + \frac{1}{2^d}\Big[\frac{\|x\|_2}{\|x_0\|_2} - \sum_{i=0}^{d-1}\frac{\sin\overline{\theta}_i}{\pi}\prod_{j=i+1}^{d-1}\frac{\pi-\overline{\theta}_j}{\pi}\Big]\hat{x}, \tag{29}$$

where $\overline{\theta}_0 = \angle(x, x_0)$ and $\overline{\theta}_i = g(\overline{\theta}_{i-1})$ for g given by (3). Define

$$\rho_d := \sum_{i=0}^{d-1}\frac{\sin\breve{\theta}_i}{\pi}\left(\prod_{j=i+1}^{d-1}\frac{\pi-\breve{\theta}_j}{\pi}\right),$$

where $\breve{\theta}_0 = \pi$ and $\breve{\theta}_i = g(\breve{\theta}_{i-1})$. If $x \in S_{\epsilon,x_0}$, then we have that either

$$|\overline{\theta}_0| \leqslant 2\sqrt{\epsilon} \quad \textit{and} \quad |\|x\|_2 - \|x_0\|_2| \leqslant 18d\sqrt{\epsilon}\|x_0\|_2$$

or

$$|\overline{\theta}_0 - \pi| \leqslant 8\pi d^4\sqrt{\epsilon} \quad \textit{and} \quad |\|x\|_2 - \|x_0\|_2\rho_d| \leqslant 200d^7\sqrt{\epsilon}\|x_0\|_2.$$

In particular, we have

$$S_{\epsilon,x_0} \subset \mathcal{B}(x_0, 56d\sqrt{\epsilon}\|x_0\|_2) \cup \mathcal{B}(-\rho_d x_0, 500d^{11}\sqrt{\epsilon}\|x_0\|_2). \tag{30}$$

Additionally, $\rho_d \to 1$ as $d \to \infty$.

Proof. Without loss of generality, let $\|x_0\|_2 = 1$, $x_0 = e_1$ and $x = r\cos\overline{\theta}_0 \cdot e_1 + r\sin\overline{\theta}_0 \cdot e_2$ for $\overline{\theta}_0 \in [0, \pi]$. Let $x \in S_{\epsilon,x_0}$.

First we introduce some notation for convenience. Let

$$\xi = \prod_{i=0}^{d-1}\frac{\pi-\overline{\theta}_i}{\pi}, \qquad \zeta = \sum_{i=0}^{d-1}\frac{\sin\overline{\theta}_i}{\pi}\prod_{j=i+1}^{d-1}\frac{\pi-\overline{\theta}_j}{\pi},$$
$$r = \|x\|_2, \qquad M = \max(r, 1).$$

Thus, $h_{x,x_0} = -\frac{1}{2^d}\xi\hat{x}_0 + \frac{1}{2^d}(r-\zeta)\hat{x}$. By inspecting the components of h_{x,x_0}, we have that $x \in S_{\epsilon,x_0}$ implies

$$|-\xi + \cos\overline{\theta}_0(r-\zeta)| \leqslant \epsilon M \tag{31}$$
$$|\sin\overline{\theta}_0(r-\zeta)| \leqslant \epsilon M \tag{32}$$

Now, we record several properties. We have:

$$\overline{\theta}_i \in [0, \pi/2] \text{ for } i \geqslant 1$$
$$\overline{\theta}_i \leqslant \overline{\theta}_{i-1} \text{ for } i \geqslant 1$$
$$|\xi| \leqslant 1 \tag{33}$$
$$|\zeta| \leqslant \min\Big(\frac{d}{\pi}, \frac{d}{\pi}\overline{\theta}_0\Big) \tag{34}$$
$$\breve{\theta}_i \leqslant \frac{3\pi}{i+3} \text{ for } i \geqslant 0 \tag{35}$$
$$\breve{\theta}_i \geqslant \frac{\pi}{i+1} \text{ for } i \geqslant 0 \tag{36}$$
$$\xi = \prod_{i=0}^{d-1}\frac{\pi-\overline{\theta}_i}{\pi} \geqslant \frac{\pi-\overline{\theta}_0}{\pi}d^{-3} \tag{37}$$
$$\overline{\theta}_0 = \pi + O_1(\delta) \Rightarrow \overline{\theta}_i = \breve{\theta}_i + O_1(i\delta) \tag{38}$$
$$\overline{\theta}_0 = \pi + O_1(\delta) \Rightarrow |\xi| \leqslant \frac{\delta}{\pi} \tag{39}$$
$$\overline{\theta}_0 = \pi + O_1(\delta) \Rightarrow \zeta = \rho_d + O_1(3d^3\delta) \text{ if } \frac{d^2\delta}{\pi} \leqslant 1 \tag{40}$$

We now establish (35). Observe $0 < g(\theta) \leqslant \left(\frac{1}{3\pi} + \frac{1}{\theta}\right)^{-1} =: \tilde{g}(\theta)$ for $\theta \in (0, \pi]$. As g and $\tilde{g}$ are monotonic increasing, we have $\breve{\theta}_i = g^{\circ i}(\breve{\theta}_0) = g^{\circ i}(\pi) \leqslant \tilde{g}^{\circ i}(\pi) = \left(\frac{i}{3\pi} + \frac{1}{\pi}\right)^{-1} = \frac{3\pi}{i+3}$. Similarly, $g(\theta) \geqslant (\frac{1}{\pi} + \frac{1}{\theta})^{-1}$ implies that $\breve{\theta}_i \geqslant \frac{\pi}{i+1}$, establishing (36).

We now establish (37). Using (35) and $\overline{\theta}_i \leqslant \breve{\theta}_i$, we have

$$\prod_{i=1}^{d-1}\Big(1 - \frac{\overline{\theta}_i}{\pi}\Big) \geqslant \prod_{i=1}^{d-1}\Big(1 - \frac{3}{i+3}\Big) \geqslant d^{-3},$$

where the last inequality can be established by showing that the ratio of consecutive terms with respect to d is greater for the product in the middle expression than for d^{-3}.

We establish (38) by using the fact that $|g'(\theta)| \leqslant 1$ for all $\theta \in [0, \pi]$ and using the same logic as for (14).

We now establish (40). As $\overline{\theta}_0 = \pi + O_1(\delta)$, we have $\overline{\theta}_i = \breve{\theta}_i + O_1(i\delta)$. Thus, if $\frac{d^2\delta}{\pi} \leqslant 1$,

$$\prod_{j=i+1}^{d-1}\frac{\pi-\overline{\theta}_j}{\pi} = \prod_{j=i+1}^{d-1}\Big(\frac{\pi-\breve{\theta}_j}{\pi} + O_1(\frac{i\delta}{2\pi})\Big) = \Big(\prod_{j=i+1}^{d-1}\frac{\pi-\breve{\theta}_j}{\pi}\Big) + O_1(d^2\delta)$$

So

$$\zeta = \sum_{i=0}^{d-1}\Big(\frac{\sin\breve{\theta}_i}{\pi} + O_1(\frac{i\delta}{\pi})\Big)\Big[\Big(\prod_{j=i+1}^{d-1}\frac{\pi-\breve{\theta}_j}{\pi}\Big) + O_1(d^2\delta)\Big] \tag{41}$$

$$= \rho_d + O_1\Big(d^2\delta/\pi + d^3\delta/\pi + d^4\delta^2/\pi\Big) \tag{42}$$

$$= \rho_d + O_1(3d^3\delta). \tag{43}$$

Thus (40) holds.

Next, we establish that $x \in S_{\epsilon,x_0} \Rightarrow r \leqslant 4d$, and thus $M \leqslant 4d$. Suppose $r > 1$. At least one of the following holds: $|\sin\overline{\theta}_0| \geqslant 1/\sqrt{2}$ or $|\cos\overline{\theta}_0| \geqslant 1/\sqrt{2}$. If $|\sin\overline{\theta}_0| \geqslant 1/\sqrt{2}$ then (32) implies that $|r-\zeta| \leqslant \sqrt{2}\epsilon r$. Using (34), we get $r \leqslant \frac{d/\pi}{1-\sqrt{2}\epsilon} \leqslant d/2$ if $\epsilon < 1/4$. If $|\cos\overline{\theta}_0| \geqslant 1/\sqrt{2}$, then (31) implies that $|r-\zeta| \leqslant \sqrt{2}(\epsilon r + |\xi|)$. Using (33), (34), and $\epsilon < 1/4$, we get $r \leqslant \frac{\sqrt{2}|\xi|+\zeta}{1-\sqrt{2}\epsilon} \leqslant \frac{\sqrt{2}+d}{1-\sqrt{2}\epsilon} \leqslant 4d$. Thus, we have $x \in S_\epsilon \Rightarrow r \leqslant 4d \Rightarrow M \leqslant 4d$.

Next, we establish that we only need to consider the small angle case ($\overline{\theta}_0 \approx 0$) and the large angle case ($\overline{\theta}_0 \approx \pi$). Exactly one of the following holds: $|r-\zeta| \geqslant \sqrt{\epsilon}M$ or $|r-\zeta| < \sqrt{\epsilon}M$. If $|r-\zeta| \geqslant \sqrt{\epsilon}M$, then by (32), we have $|\sin\overline{\theta}_0| \leqslant \sqrt{\epsilon}$. Hence $\overline{\theta}_0 = O_1(2\sqrt{\epsilon})$ or $\overline{\theta}_0 = \pi + O_1(2\sqrt{\epsilon})$, as $\epsilon < 1$. If $|r-\zeta| \leqslant \sqrt{\epsilon}M$, then by (31) we have $|\xi| \leqslant 2\sqrt{\epsilon}M$. Using (37), we get $\overline{\theta}_0 = \pi + O_1(2\pi d^3\sqrt{\epsilon}M)$. Thus, we only need to consider the small angle case, $\overline{\theta}_0 = O_1(2\sqrt{\epsilon})$ and the large angle case $\overline{\theta}_0 = \pi + O_1(8\pi d^4\sqrt{\epsilon})$, where we have used $M \leqslant 4d$.

Small Angle Case. Assume $\overline{\theta}_0 = O_1(2\sqrt{\epsilon})$. As $\overline{\theta}_i \leqslant \overline{\theta}_0 \leqslant 2\sqrt{\epsilon}$ for all i, we have $\xi \geqslant (1-\frac{2\sqrt{\epsilon}}{\pi})^d = 1 + O_1(\frac{4d\sqrt{\epsilon}}{\pi})$ provided $2d\sqrt{\epsilon} \leqslant 1/2$. By (34), we also have $\zeta = O_1(\frac{d}{\pi}2\sqrt{\epsilon}) = O_1(d\sqrt{\epsilon})$. By (31), we have

$$|-\xi + \cos\overline{\theta}_0(r-\zeta)| \leqslant \epsilon M.$$

Thus, as $\cos\overline{\theta}_0 = 1 + O_1(\overline{\theta}_0^2/2) = 1 + O_1(2\epsilon)$,

$$-\Big(1 + O_1(4d\sqrt{\epsilon})\Big) + (1+O_1(2\epsilon))(r + O_1(d\sqrt{\epsilon})) = O_1(4d\epsilon),$$

and thus,

$$r - 1 = O_1(4d\sqrt{\epsilon} + 2\epsilon 4d + d\sqrt{\epsilon} + 2d\epsilon^{3/2} + 4\epsilon d) \tag{44}$$

$$= O_1(18d\sqrt{\epsilon}). \tag{45}$$

Large Angle Case. Assume $\theta_0 = \pi + O_1(\delta)$ where $\delta = 8\pi d^4\sqrt{\epsilon}$. By (39) and (40), we have $\xi = O_1(\delta/\pi)$, and we have $\zeta = \rho_d + O_1(3d^3\delta)$ if $8d^6\sqrt{\epsilon} \leqslant 1$. By (31), we have

$$|-\xi + \cos\theta_0(r-\zeta)| \leqslant \epsilon M,$$

so, as $\cos\theta_0 = 1 - O_1(\theta_0^2/2)$,

$$O_1(\delta/\pi) + (1+O_1(\delta^2/2))(r - \rho_d + O_1(3d^3\delta)) = O_1(\epsilon M),$$

and thus, using $r \leqslant 4d$, $\rho_d \leqslant d$, and $\delta = 8\pi d^4\sqrt{\epsilon} \leqslant 1$,

$$r - \rho_d = O_1(\epsilon M + \delta/\pi + 3d^3\delta + \frac{5}{2}\delta^2 d + \frac{3}{2}d^3\delta^3) \tag{46}$$

$$= O_1\Big(4\epsilon d + \delta(\frac{1}{\pi} + 3d^3 + \frac{5}{2}d + \frac{3}{2}d^3)\Big) \tag{47}$$

$$= O_1(200d^7\sqrt{\epsilon}) \tag{48}$$

To conclude the proof of (30), we use the fact that

$$\|x - x_0\|_2 \leqslant \big|\|x\|_2 - \|x_0\|_2\big| + (\|x_0\|_2 + \big|\|x\|_2 - \|x_0\|_2\big|)\overline{\theta}_0.$$

This fact simply says that if a 2d point is known to have magnitude within Δr of some r and is known to be within angle $\Delta\theta$ from 0, then its Euclidean distance to the point of polar coordinates $(r, 0)$ is no more than $\Delta r + (r + \Delta r)\Delta\theta$.

Finally, we establish that $\rho_d \to 1$ as $d \to \infty$. Note that $\rho_{d+1} = (1 - \frac{\breve{\theta}_d}{\pi})\rho_d + \frac{\sin\breve{\theta}_d}{\pi}$ and $\rho_0 = 0$. It suffices to show $\tilde{\rho}_d \to 0$, where $\tilde{\rho}_d := 1 - \rho_d$. The following recurrence relation holds: $\tilde{\rho}_d = (1 - \frac{\breve{\theta}_{d-1}}{\pi})\tilde{\rho}_{d-1} + \frac{\breve{\theta}_{d-1} - \sin\breve{\theta}_{d-1}}{\pi}$, with $\tilde{\rho}_0 = 1$. Using the recurrence formula (15) and the fact that $\breve{\theta}_0 = \pi$, we get that

$$\tilde{\rho}_d = \sum_{i=1}^{d}\frac{\breve{\theta}_{i-1} - \sin\breve{\theta}_{i-1}}{\pi}\prod_{j=i+1}^{d}\big(1 - \frac{\breve{\theta}_{j-1}}{\pi}\big) \tag{49}$$

using (36), we have that

$$\prod_{j=i+1}^{d}\Big(1 - \frac{\breve{\theta}_{j-1}}{\pi}\Big) \leqslant \prod_{j=i+1}^{d}\Big(1 - \frac{1}{j}\Big) = \exp\Big(-\sum_{j=i+1}^{d}\frac{1}{j}\Big) \leqslant \exp\Big(-\int_{i+1}^{d+1}\frac{1}{s}ds\Big) = \frac{i+1}{d+1}$$

Using (35) and the fact that $\breve{\theta}_{i-1} - \sin\breve{\theta}_{i-1} \leqslant \breve{\theta}_{i-1}^3/6$, we have that $\tilde{\rho}_d \leqslant \sum_{i=1}^{d}\frac{\breve{\theta}_{i-1}^3}{6\pi}\cdot\frac{i+1}{d+1} \to 0$ as $d \to \infty$. □

E. Proof of WDC for Gaussian Matrices

In this section, we establish a bound on the probability that a Gaussian matrix satisfies the WDC, provided it corresponds to a sufficiently expansive layer of a neural network.

Lemma 11. *Fix $0 < \epsilon < 1$. Let $W \in \mathbb{R}^{n\times k}$ have i.i.d. $\mathcal{N}(0, 1/n)$ entries. If $n > ck\log k$, then with probability at least $1 - 8ne^{-\gamma k}$, W satisfies the WDC with constant ϵ. Here c, γ^{-1} are constants that depend only polynomially on ϵ^{-1}.*

The WDC with constant ϵ can be written as

$$\|W_{+,x}^t W_{+,y} - Q_{x,y}\| \leqslant \epsilon \tag{50}$$

for all nonzero $x, y \in \mathbb{R}^k$. A common way to establish concentration of a random function simultaneously over

an infinite number of values of (x, y), like (50), is as follows:

- Show concentration of the quantity with high probability for a fixed (x, y).
- Bound the Lipschitz constant of the quantity with respect to (x, y).
- Take a union bound over a net whose size is given by the Lipschitz constant.

This argument does not apply in the present case because $W_{+,x}^t W_{+,y}$ is not continuous with respect to (x, y). To deal with this lack of continuity, we form two continuous variants that are greater and less than $W_{+,x}^t W_{+,y}$, respectively, with respect the semidefinite ordering. We now introduce some notation in order to state these bounds. Let

$$h_{-\epsilon}(z) = \begin{cases} 0 & z \leqslant -\epsilon, \\ 1 + \frac{z}{\epsilon} & -\epsilon \leqslant z \leqslant 0, \\ 1 & z \geqslant 0, \end{cases}$$

and

$$h_{\epsilon}(z) = \begin{cases} 0 & z \leqslant 0, \\ \frac{z}{\epsilon} & 0 \leqslant z \leqslant \epsilon, \\ 1 & z \geqslant \epsilon. \end{cases}$$

When applied to a vector, let $h_{-\epsilon}$ and h_ϵ act componentwise. Let w_i^t be the ith row of W. Note that $W_{+,x}^t W_{+,y} = \sum_{i=1}^n 1_{w_i \cdot x > 0} 1_{w_i \cdot y > 0} \cdot w_i w_i^t$, and define

$$G_{-\epsilon}(x, y) := \sum_{i=1}^n h_{-\epsilon}(w_i \cdot x) h_{-\epsilon}(w_i \cdot y) w_i w_i^t$$
$$G_{\epsilon}(x, y) := \sum_{i=1}^n h_{\epsilon}(w_i \cdot x) h_{\epsilon}(w_i \cdot y) w_i w_i^t.$$

As $h_{-\epsilon}(z) \geqslant 1_{z>0}(z)$ and $h_\epsilon(z) \leqslant 1_{z<0}(z)$ for all $z \in \mathbb{R}$, we have that for all nonzero x, y that $G_\epsilon(x, y) \preceq W_{+,x}^t W_{+,y} \preceq G_{-\epsilon}(x, y)$. Thus, it suffices to establish a matrix upper bound on $G_{-\epsilon}$ and a matrix lower bound on G_ϵ.

First we establish a matrix upper bound on $G_{-\epsilon}(x, y)$ uniformly over all nonzero x, y. For ease of exposition, in the next two Lemmas, we will take the entries of W to be i.i.d. $\mathcal{N}(0, 1)$. In this case, $\mathbb{E}[W_{+,x}^t W_{+,y}] = nQ_{x,y}$.

Lemma 12. *Fix $0 < \epsilon < 1$. Let $W \in \mathbb{R}^{n \times k}$ have i.i.d. $\mathcal{N}(0, 1)$ entries. If $n > ck \log k$, then with probability at least $1 - 4ne^{-\gamma k}$,*

$$\forall x \neq 0, y \neq 0, \quad G_{-\epsilon}(x, y) \preceq nQ_{x,y} + 3\epsilon n I_k.$$

Here, c and γ^{-1} are constants that depend only polynomially on ϵ^{-1}.

Proof. Note that the entries of W are assumed to have $\mathcal{N}(0, 1)$ entries, and not $\mathcal{N}(0, 1/n)$ entries like in most of this paper. In this proof, the values of the constants c and γ may change from line to line, but they are all bounded above and below, respectively, by some ϵ-dependent constant. Without loss of generality, let $x, y \in S^{k-1}$.

First, we bound $\mathbb{E}[G_{-\epsilon}(x, y)]$ for fixed $x, y \in S^{k-1}$. Noting that $h_{-\epsilon}(z) \leqslant 1_{z \geqslant -\epsilon}(z) = 1_{z>0}(z) + 1_{-\epsilon \leqslant z \leqslant 0}(z)$, we have

$$\begin{aligned}
\mathbb{E}[G_{-\epsilon}(x, y)] &\preceq \mathbb{E}\Big[\sum_{i=1}^n 1_{w_i \cdot x \geqslant -\epsilon} 1_{w_i \cdot y \geqslant -\epsilon} \cdot w_i w_i^t\Big] \\
&\leq \mathbb{E}\Big[\sum_{i=1}^n (1_{w_i \cdot x > 0} 1_{w_i \cdot y > 0} \\
&\qquad + 1_{-\epsilon \leqslant w_i \cdot x \leqslant 0} + 1_{-\epsilon \leqslant w_i \cdot y \leqslant 0}) \cdot w_i w_i^t\Big] \\
&= nQ_{x,y} + n\mathbb{E}[1_{-\epsilon \leqslant w_i \cdot x \leqslant 0} \cdot w_i w_i^t] \\
&\qquad + n\mathbb{E}[1_{-\epsilon \leqslant w_i \cdot y \leqslant 0} \cdot w_i w_i^t].
\end{aligned}$$

We now bound $\mathbb{E}[1_{-\epsilon \leqslant w_i \cdot x \leqslant 0} \cdot w_i w_i^t]$. For deriving this bound, we may take $x = e_1$ without loss of generality. We have for $\epsilon < 1$,

$$\begin{aligned}
\mathbb{E}[1_{-\epsilon \leqslant w_i \cdot x \leqslant 0} \cdot w_i w_i^t] &= \begin{pmatrix} \zeta_1 & 0 \\ 0 & \zeta_2 \cdot I_{n-1} \end{pmatrix}, \\
0 \leqslant \zeta_1 &\leqslant \int_{-\epsilon}^0 z^2 \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \leqslant \frac{\epsilon}{2}, \\
0 \leqslant \zeta_2 &\leqslant \int_{-\epsilon}^0 \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \leqslant \frac{\epsilon}{2}.
\end{aligned}$$

Thus, $\mathbb{E}[1_{-\epsilon < w_i \cdot x < 0} \cdot w_i w_i^t] \preceq \frac{\epsilon}{2} I_k$ for any $x \neq 0$, resulting in

$$\mathbb{E}[G_{-\epsilon}(x, y)] \preceq nQ_{x,y} + \epsilon n \cdot I_k. \tag{51}$$

Second, we show concentration of $G_{-\epsilon}(x, y)$ for fixed $x, y \in S^{k-1}$. Let $\xi_i = \sqrt{h_{-\epsilon}(w_i \cdot x)} \sqrt{h_{-\epsilon}(w_i \cdot y)} w_i$. We have

$$\begin{aligned}
&G_{-\epsilon}(x, y) - \mathbb{E}[G_{-\epsilon}(x, y)] \\
&= \sum_{i=1}^n \Big(h_{-\epsilon}(w_i \cdot x) h_{-\epsilon}(w_i \cdot y) w_i w_i^t \\
&\qquad - \mathbb{E}\big[h_{-\epsilon}(w_i \cdot x) h_{-\epsilon}(w_i \cdot y) w_i w_i^t\big] \Big) \qquad (52) \\
&= \sum_{i=1}^n (\xi_i \xi_i^t - \mathbb{E}\xi_i \xi_i^t). \qquad (53)
\end{aligned}$$

Note that ξ_i is sub-Gaussian for all i and that the sub-Gaussian norm of ξ_i is bounded above by an absolute constant, which we will call K. By the first part of Remark 5.40 in [77], there exist constants c_K and γ_K such that for all $t \geqslant 0$, with probability at least $1 - 2e^{-\gamma_K t^2}$,

$$\|G_{-\epsilon}(x, y) - \mathbb{E}G_{-\epsilon}(x, y)\| \leqslant \max(\delta, \delta^2) n,$$

where $\delta = c_K \sqrt{\frac{k}{n}} + \frac{t}{\sqrt{n}}$. If $n > (2/\epsilon)^2 c_K^2 k$, $t = \epsilon\sqrt{n}/2$, and $\epsilon < 1$, we have that with probability at least $1 - 2e^{-\gamma_K \epsilon^2 n/4}$,

$$\|G_{-\epsilon}(x, y) - \mathbb{E}G_{-\epsilon}(x, y)\| \leqslant \epsilon n. \tag{54}$$

Third, we bound the Lipschitz constant of $G_{-\epsilon}$. For $\tilde{x}, \tilde{y} \in \mathbb{R}^k$ we have

$$\begin{aligned}
&G_{-\epsilon}(x,y) - G_{-\epsilon}(\tilde{x},\tilde{y}) \\
&= \sum_{i=1}^{n} \big[h_{-\epsilon}(w_i\cdot x)h_{-\epsilon}(w_i\cdot y) \\
&\qquad - h_{-\epsilon}(w_i\cdot \tilde{x})h_{-\epsilon}(w_i\cdot \tilde{y})\big] w_i w_i^t \\
&= \sum_{i=1}^{n} \big[h_{-\epsilon}(w_i\cdot x)\big(h_{-\epsilon}(w_i\cdot y) - h_{-\epsilon}(w_i\cdot \tilde{y})\big) \\
&\qquad + h_{-\epsilon}(w_i\cdot \tilde{y})\big(h_{-\epsilon}(w_i\cdot x) - h_{-\epsilon}(w_i\cdot \tilde{x})\big)\big] w_i w_i^t \\
&= W^t\big[\operatorname{diag}\big(h_{-\epsilon}(Wx)\big)\operatorname{diag}\big(h_{-\epsilon}(Wy) - h_{-\epsilon}(W\tilde{y})\big) \\
&\qquad + \operatorname{diag}\big(h_{-\epsilon}(W\tilde{y})\big)\operatorname{diag}\big(h_{-\epsilon}(Wx) - h_{-\epsilon}(W\tilde{x})\big)\big] W.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\|G_{-\epsilon}(x,y) - G_{-\epsilon}(\tilde{x},\tilde{y})\| \\
&\leqslant \|W\|^2\Big[\|h_{-\epsilon}(Wx)\|_\infty \|h_{-\epsilon}(Wy) - h_{-\epsilon}(W\tilde{y})\|_\infty \\
&\qquad + \|h_{-\epsilon}(W\tilde{y})\|_\infty \|h_{-\epsilon}(Wx) - h_{-\epsilon}(W\tilde{x})\|_\infty\Big] \\
&\leqslant \|W\|^2\Big[\max_{i\in[n]} \big|h_{-\epsilon}(w_i\cdot y) - h_{-\epsilon}(w_i\cdot \tilde{y})\big| \\
&\qquad + \max_{i\in[n]} \big|h_{-\epsilon}(w_i\cdot x) - h_{-\epsilon}(w_i\cdot \tilde{x})\big|\Big] \\
&\leqslant \|W\|^2\Big[\max_{i\in[n]} \frac{1}{\epsilon}\big|w_i\cdot(x-\tilde{x})\big| + \max_{i\in[n]} \frac{1}{\epsilon}\big|w_i\cdot(y-\tilde{y})\big|\Big] \\
&\leqslant \|W\|^2 \frac{1}{\epsilon}\Big(\max_{i\in[n]} \|w_i\|_2\Big)\big(\|x-\tilde{x}\|_2 + \|y-\tilde{y}\|_2\big)
\end{aligned}$$

where the second inequality follows because $|h_{-\epsilon}(z)| \leqslant 1$ for all z, and the third inequality follows because $h_{-\epsilon}$ is $1/\epsilon$-Lipschitz.

Let E_1 be the event that $\|W\| \leqslant 3\sqrt{n}$. By Corollary 5.35 in [77], we have that $\mathbb{P}(E_1) \geqslant 1 - 2e^{-n/2}$, if $n \geqslant k$. Let E_2 be the event that $\max_{i\in[n]} \|w_i\|_2 \leqslant 2\sqrt{k}$. By a single-tailed variant of Corollary 5.17 in [77], there exists a constant γ_0 such that for fixed i, $\|w_i\|_2 \leqslant 2\sqrt{k}$ with probability at least $1 - e^{-\gamma_0 k}$. Thus, $\mathbb{P}(E_2) \geqslant 1 - ne^{-\gamma_0 k}$.

On $E_1 \cap E_2$, we have

$$\begin{aligned}
&\|G_{-\epsilon}(x,y) - G_{-\epsilon}(\tilde{x},\tilde{y})\| \\
&\leqslant \frac{18n\sqrt{k}}{\epsilon}\big(\|x-\tilde{x}\|_2 + \|y-\tilde{y}\|_2\big). \qquad (55)
\end{aligned}$$

for all $x, y, \tilde{x}, \tilde{y} \in S^{k-1}$.

Finally, we complete the proof by a covering argument. Let $\mathcal{N}_\delta$ be a δ-net on S^{k-1} such that $|\mathcal{N}_\delta| \leqslant (3/\delta)^k$. Take $\delta = \frac{\epsilon^2}{36\sqrt{k}}$. Combining (54) and (51), we have

$$\begin{aligned}
\forall x, y \in \mathcal{N}_\delta, \quad G_{-\epsilon}(x,y) &\leq \mathbb{E}G_{-\epsilon}(x,y) + \epsilon n I_k \\
&\leq nQ_{x,y} + 2\epsilon n I_k.
\end{aligned}$$

with probability at least $1 - 2|\mathcal{N}_\delta| e^{-\gamma_K \epsilon^2 n/4} \geqslant 1 - 2(\frac{3}{\delta})^k e^{-\gamma_K \epsilon^2 n/4} \geqslant 1 - 2e^{-\gamma_K \epsilon^2 n/4 + k\log(3\cdot 36\sqrt{k}/\epsilon^2)}$. If $n > \tilde{c}k\log k$, for some $\tilde{c} = \Omega(\epsilon^{-2}\log\epsilon^{-1})$, then this probability is at least $1 - 2e^{-\tilde{\gamma} n}$ for some $\tilde{\gamma} = O(\epsilon^2)$. For any $x, y \in S^{k-1}$, let $\tilde{x}, \tilde{y} \in \mathcal{N}_\delta$ be such that $\|x - \tilde{x}\|_2 < \delta$ and $\|y - \tilde{y}\|_2 < \delta$. By (55), we have that

$$\begin{aligned}
\forall x \neq 0, y \neq 0, \quad G_{-\epsilon}(x,y) &\leq G_{-\epsilon}(\tilde{x},\tilde{y}) + \frac{18n\sqrt{k}}{\epsilon} 2\delta I_k \\
&\leq nQ_{x,y} + 3\epsilon n I_k.
\end{aligned}$$

In conclusion, the result of this lemma holds if $n > (2/\epsilon)^2 c_K^2 k$ and $n > \tilde{c}k\log k$, with probability at least $1 - 2e^{-\gamma_K \epsilon^2 n/4} - 2e^{-n/2} - ne^{-\gamma_0 k} - 2e^{-\tilde{\gamma} n} > 1 - 4ne^{-\gamma k}$ for some $\gamma = O(\epsilon^2)$ and $\tilde{c} = \Omega(\epsilon^{-2}\log\epsilon^{-1})$. □

Lemma 13. *Fix $0 < \epsilon < 1$. Let $W \in \mathbb{R}^{n\times k}$ have i.i.d. $\mathcal{N}(0,1)$ entries. If $n > ck\log k$, then with probability at least $1 - 4ne^{-\gamma k}$,*

$$\forall x \neq 0, y \neq 0, \quad G_\epsilon(x,y) \geq nQ_{x,y} - 3\epsilon n I_k.$$

Here, c and γ^{-1} are constants that depend only polynomially on ϵ^{-1}.

Proof. The proof follows that of Lemma 12 exactly.

First, we bound $\mathbb{E}[G_\epsilon(x,y)]$ for fixed $x, y \in S^{k-1}$. For deriving this bound, we take $x = e_1$ without loss of generality. Noting that $h_\epsilon(z) \geqslant 1_{z>0}(z) - 1_{0\leqslant z\leqslant\epsilon}(z)$ for all z, we have

$$\begin{aligned}
\mathbb{E}[G_\epsilon(x,y)] \geq nQ_{x,y} &- n\mathbb{E}[1_{0\leqslant w_i\cdot x\leqslant\epsilon}\cdot w_i w_i^t] \\
&- n\mathbb{E}[1_{0\leqslant w_i\cdot y\leqslant\epsilon}\cdot w_i w_i^t].
\end{aligned}$$

We have

$$\begin{aligned}
\mathbb{E}[1_{0\leqslant w_i\cdot x\leqslant\epsilon}\cdot w_i w_i^t] &= \begin{pmatrix}\zeta_1 & 0\\ 0 & \zeta_2\cdot I_{n-1}\end{pmatrix}, \\
0 \leqslant \zeta_1 &\leqslant \int_0^\epsilon z^2 \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \leqslant \frac{\epsilon}{2}, \\
0 \leqslant \zeta_2 &\leqslant \int_0^\epsilon \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \leqslant \frac{\epsilon}{2}.
\end{aligned}$$

Thus, $\mathbb{E}[1_{0\leqslant w_i\cdot x\leqslant\epsilon}\cdot w_i w_i^t] \preceq \frac{\epsilon}{2} I_k$ for any $x \neq 0$, resulting in

$$\mathbb{E}[G_\epsilon(x,y)] \geq nQ_{x,y} - \epsilon n\cdot I_k.$$

Second, the same argument as in Lemma 12 provides that for fixed $x, y \in S^{k-1}$, if $n > (2/\epsilon)^2 c_K^2 k$, then we have that with probability at least $1 - 2e^{-\gamma_K \epsilon^2 n/4}$,

$$\|G_\epsilon(x,y) - \mathbb{E}G_\epsilon(x,y)\| \leqslant \epsilon n.$$

Third, the same argument as in Lemma 12 provides that on the event $E_1 \cap E_2$, we have

$$\|G_\epsilon(x,y) - G_\epsilon(\tilde{x},\tilde{y})\| \leqslant \frac{18n\sqrt{k}}{\epsilon}\big[\|x-\tilde{x}\|_2 + \|y-\tilde{y}\|_2\big].$$

for all $x, y, \tilde{x}, \tilde{y} \in S^{k-1}$.

Finally, we complete the proof by an identical covering argument as in Lemma 12. We have that if $n > c_0 k \log k$ then with probability at least $1 - 4ne^{-\gamma k}$,

$$\forall x, y \in S^{k-1}, G_\epsilon(x,y) \succeq nQ_{x,y} - 3\epsilon n I_k.$$

□

We may now prove Lemma 11.

Proof of Lemma 11. It suffices to show

$$\forall x, y \in \mathbb{R}^k, \quad \|W_{+,x}^t W_{+,y} - nQ_{x,y}\| \leqslant 3n\epsilon,$$

where A has i.i.d. $\mathcal{N}(0,1)$ entries. The result is immediate if $x = 0$ or $y = 0$. For all nonzero x and y , we have $G_\epsilon(x,y) \preceq W_{+,x}^t W_{+,y} \preceq G_{-\epsilon}(x,y)$. The lemma then follows directly from Lemmas 12 and 13. □

F. Proof of RRIC

We will make use of a standard concentration result used in proving the Restricted Isometry Property from compressed sensing [61].

Lemma 14 (Variant of Lemma 5.1 in [61])**.** *Let $A \in \mathbb{R}^{m\times n}$ have i.i.d. $\mathcal{N}(0, 1/m)$ entries. Fix $0 < \epsilon < 1, k < m$. Fix a subspace $T \subset \mathbb{R}^n$ of dimension k. With probability at least $1 - (c_1/\epsilon)^k e^{-\gamma_1 \epsilon m}$,*

$$(1-\epsilon)\|x\|_2^2 \leqslant \|Ax\|_2^2 \leqslant (1+\epsilon)\|x\|_2^2, \quad \forall x \in T,$$

and

$$|\langle Ax, Ay\rangle - \langle x, y\rangle| \leqslant \epsilon \|x\|_2 \|y\|_2, \quad \forall x, y \in T.$$

Let $V = \bigcup_{i=1}^M V_i$ and $W = \bigcup_{j=1}^N W_j$, where V_i and W_j are subspaces of $\mathbb{R}^n$ of dimension at most k for all i, j. Then,

$$|\langle Ax, Ay\rangle - \langle x, y\rangle| \leqslant \epsilon \|x\|_2 \|y\|_2, \quad \forall x \in V, y \in W,$$

with probability at least $1 - MN(c_1/\epsilon)^{2k} e^{-\gamma_1 \epsilon m}$. Here, c_1 and γ_1 are universal constants.

Proof. This proof is an immediate extension of Lemma 5.1 in [61]. As A is Gaussian, we may take T to be the span of k standard basis vectors and directly apply the lemma in that paper. The inner product form follows by a standard argument based on the parallelogram identity. The last inequality holds by applying the second inequality to all subspaces of the form $\operatorname{span}(V_i, W_j)$, which have dimension at most $2k$, and by applying a union bound. □

In order to apply Lemma 14, we now provide an upper bound for the number of subspaces that arise from the objective f.

Lemma 15. *Let V be a subspace of $\mathbb{R}^k$. Let $W \in \mathbb{R}^{n\times k}$ have i.i.d. $\mathcal{N}(0, 1/n)$ entries. With probability 1,*

$$|\{\operatorname{diag}(Wv > 0)W \mid v \in V\}| \leqslant 10 n^{\dim V}. \tag{56}$$

Proof. Let $\ell = \dim V$. By rotational invariance of Gaussians, we may take $V = \operatorname{span}(e_1, \ldots, e_\ell)$ without loss of generality. Without loss of generality, we may let W have dimensions $n \times \ell$ and take $V = \mathbb{R}^\ell$.

We will appeal to a classical result from sphere covering [78]. If n hyperplanes in $\mathbb{R}^\ell$ contain the origin and are such that the normal vectors to any subset of ℓ of those hyperplanes are independent, then the complement of the union of these hyperplanes is partitioned into at most

$$2\sum_{i=0}^{\ell-1} \binom{n-1}{i}$$

disjoint regions. Note that for fixed W, $|\{\operatorname{diag}(Wv > 0)W \mid v \in \mathbb{R}^\ell\}|$ equals the number of binary vectors of the form $(1_{w_i \cdot v > 0})_{i\in[n]}$ for $v \in S^{\ell-1}$. Each such binary vector corresponds uniquely to one of the disjoint regions given by partitioning the unit sphere in $\mathbb{R}^\ell$ by the n half-spaces going through the origin with normal vectors $\{w_i\}_{i\in[n]}$. With probability 1, any subset of ℓ rows of W are linearly independent, and thus,

$$\begin{aligned}|\{\operatorname{diag}(Wv > 0)W \mid v \in \mathbb{R}^\ell\}| &= 2\sum_{i=0}^{\ell-1} \binom{n-1}{i} \\ &\leqslant 2\ell\Big(\frac{en}{\ell}\Big)^\ell \\ &\leqslant 10 n^\ell.\end{aligned}$$

where the first inequality uses the fact that $\binom{n}{\ell} \leqslant (en/\ell)^\ell$ and the second inequality uses that $2\ell(e/\ell)^\ell \leqslant 10$ for all $\ell \geqslant 1$.

□

Lemma 16. *Let $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$ have i.i.d. $\mathcal{N}(0, 1/n_i)$ entries for $i = 1, \ldots, d$. Let $k = n_0$. Then, with probability 1,*

$$|\{W_{i,+,x} \mid x \neq 0\}| \leqslant 10^i n_1^k n_2^k \cdots n_i^k. \tag{57}$$

Proof. Recall that $W_{1,+,x} = \operatorname{diag}(W_1 x > 0)W_1$ and $W_{i,+,x} = \operatorname{diag}(W_i W_{i-1,+,x} \cdots W_{1,+,x} x > 0)W_i$. The case of $i = 1$ holds with probability 1 by applying Lemma 15 with $V = \mathbb{R}^k$.

Next, we establish the $i = 2$ case. Let $\mathcal{W}_+ = \{W_{+,x} \mid x \neq 0\}$. Note that $\forall x \neq 0$,

$$W_{2,+,x} \in \bigcup_{\hat{W}\in\mathcal{W}_+} \{\operatorname{diag}(W_2 v > 0)W_2 \mid v \in \operatorname{range}\hat{W}\}.$$

Let the random variable $X_{\hat{W},W_2} = |\{\operatorname{diag}(W_2 v > 0)W_2 \mid v \in \operatorname{range}\hat{W}\}|$. Note that by Lemma 15, for any fixed $\hat{W}$, $\mathbb{P}_{W_2}(X_{\hat{W},W_2} \leqslant 10n_2^k) = 1$. Hence, condi-

tioned on the probability 1 event $E := \{|\mathcal{W}_+| \leqslant 10n_1^k\}$, we have $\mathbb{P}_{W_2}(\sum_{\hat{W}\in\mathcal{W}_+} X_{\hat{W},W_2} \leqslant 10^2 n_1^k n_2^k) = 1$. Thus,

$$\begin{aligned}
&\mathbb{P}_{W_1,W_2}\left(\sum_{\hat{W}\in\mathcal{W}_+} X_{\hat{W},W_2} \leqslant 10^2 n_1^k n_2^k\right)\\
&= \int_{\Omega_1} \mathbb{P}_{W_2}\left(\sum_{\hat{W}\in\mathcal{W}_+} X_{\hat{W},W_2} \leqslant 10^2 n_1^k n_2^k\right) d\mu_1\\
&= \int_{E} \mathbb{P}_{W_2}\left(\sum_{\hat{W}\in\mathcal{W}_+} X_{\hat{W},W_2} \leqslant 10^2 n_1^k n_2^k\right) d\mu_1\\
&= 1,
\end{aligned}$$

where Ω_1 and μ_1 are the state space and probability measure for the random variable W_1. Hence, $|\{W_{2,+,x} \mid x \neq 0\}| \leqslant 10^2 n_1^k n_2^k$ with probability 1.

The case of larger i follows by repeating the logic above.

□

We can now show concentration of v_{x,x_0} to its expectation with respect to C.

Lemma 17. *Fix $0 < \epsilon < 1$. Let $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$, have i.i.d. $\mathcal{N}(0, 1/n_i)$ entries for $i = 1, \ldots, d$. Let $A \in \mathbb{R}^{m\times n_d}$ have i.i.d. $\mathcal{N}(0, 1/m)$ entries that are independent from all W_i. If $m > cdk\log(n_1 n_2 \cdots n_d)$, then with probability at least $1 - e^{-\gamma m}$,*

$$\begin{aligned}
\forall x, y \in \mathbb{R}^k, \quad &\|(\Pi_{i=d}^1 W_{i,+,x})^t A^t A(\Pi_{i=d}^1 W_{i,+,y})\\
&\quad - (\Pi_{i=d}^1 W_{i,+,x})^t(\Pi_{i=d}^1 W_{i,+,y})\|\\
&\leqslant \epsilon \Pi_{i=1}^d \|W_{i,+,x}\|\|W_{i,+,y}\|. \qquad (58)
\end{aligned}$$

Here, c and γ^{-1} are constants that depend only polynomially on ϵ^{-1}.

Proof. For pedagogical purposes, we first establish the lemma in the $d = 2$ case. It suffices to show that $\forall x, y, w, v \in S^{k-1}$,

$$\begin{aligned}
&|\langle AW_{2,+,x}W_{1,+,x}w, AW_{2,+,y}W_{1,+,y}v\rangle\\
&\quad - \langle W_{2,+,x}W_{1,+,x}w, W_{2,+,y}W_{1,+,y}v\rangle|\\
&\leqslant \epsilon\|W_{1,+,x}\|\|W_{2,+,x}\|\|W_{1,+,y}\|\|W_{2,+,y}\|.
\end{aligned}$$

In order to apply Lemma 14, we will show that $\{W_{2,+,x}W_{1,+,x}w \mid x, w \in S^{k-1}\}$ is a subset of a union of at most $10^3(n_1^2 n_2)^k$ subspaces of dimension at most k. For fixed W_1, W_2, let $\mathcal{A}_+ = \{W_{1,+,x} \mid x \neq 0\}$ and $\mathcal{B}_+ = \{W_{2,+,x} \mid x \neq 0\}$. By Lemma 16, there exists a probability 1 event, E, over (W_1, W_2) on which $|\mathcal{A}_+| \leqslant 10n_1^k$ and $|\mathcal{B}_+| \leqslant 10^2 n_1^k n_2^k$. On E,

$$|\{W_{2,+,x}W_{1,+,x} \mid x \neq 0\}| \leqslant 10^3(n_1^2 n_2)^k.$$

Note that $\dim \operatorname{range}(W_{2,+,x}W_{1,+,x}) \leqslant k$ for all $x \neq 0$. Hence $\{W_{2,+,x}W_{1,+,x}w \mid x, w \in S^{k-1}\} \subset V$, where V is a union of at most $10^3(n_1^2 n_2)^k$ subspaces of dimensionality at most k.

By applying the second half of Lemma 14 to the sets V and V, we get that for fixed W_1, W_2,

$$\begin{aligned}
&|\langle AW_{2,+,x}W_{1,+,x}w, AW_{2,+,y}W_{1,+,y}v\rangle\\
&\quad - \langle W_{2,+,x}W_{1,+,x}w, W_{2,+,y}W_{1,+,y}v\rangle|\\
&\leqslant \epsilon\|W_{2,+,x}W_{1,+,x}w\|_2\|W_{2,+,y}W_{1,+,y}v\|_2,\\
&\forall x, y, w, v \in S^{k-1} \qquad (59)
\end{aligned}$$

with probability at least $1 - 10^3(c_1 n_1^2 n_2/\epsilon)^{2k} e^{-\gamma_1 \epsilon m} \geqslant 1 - e^{-\gamma_2 m}$, provided $m \geqslant \tilde{c}k\log(n_1 n_2)$, for universal constants c_1, γ_1 and for some $\gamma_2 = \frac{\gamma_1\epsilon}{2}, \tilde{c} = \Omega(\epsilon^{-1}\log\epsilon^{-1})$.

Integrating over the probability space of (W_1, W_2), independence of A and (W_1, W_2) implies that (59) holds for random (W_1, W_2) with the same probability bound. Continuing from (59), we have

$$\begin{aligned}
&|\langle AW_{2,+,x}W_{1,+,x}w, AW_{2,+,y}W_{1,+,y}v\rangle\\
&\quad - \langle W_{2,+,x}W_{1,+,x}w, W_{2,+,y}W_{1,+,y}v\rangle|\\
&\leqslant \epsilon\|W_{1,+,x}\|_2\|W_{1,+,y}\|_2\|W_{2,+,x}\|_2\|W_{2,+,y}\|_2
\end{aligned}$$

$\forall x, y, w, v \in S^{k-1}$ with probability at least $1 - e^{-\gamma m}$. for some $\gamma > 0$.

For the case of $d \geqslant 2$, the lemma follows similarly. We have

$$|\{\Pi_{i=d}^1 W_{i,+,x}x \mid x \neq 0\}| \leqslant 10^{(d^2)}(n_1^d n_2^{d-1}\cdots n_{d-1}^2 n_d)^k$$

on the probability 1 event. The analogous bound to (59) holds with probability at least $1 - 10^{(d^2)}(c_1 n_1^d n_2^{d-1}\cdots n_{d-1}^2 n_d/\epsilon)^{2k} e^{-\gamma_1\epsilon m} \geqslant 1 - e^{-\gamma_2 m}$, provided $m \geqslant \tilde{c}dk\log(n_1 n_2\cdots n_d)$, for some $\gamma_2 = \frac{\gamma_1\epsilon}{2}, \tilde{c} = \Omega(\epsilon^{-1}\log\epsilon^{-1})$.

□

G. Proof of Theorem 6

Theorem 6 can be proved by combining Lemmas 11 and 17.

Acknowledgment: During the course of this work, PH was partially supported by NSF Grant DMS-1464525 and NSF CAREER DMS-1848087.

References

[1] I. Daubechies, *Ten lectures on wavelets*. Siam, 1992, vol. 61.

[2] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006. [Online]. Available: http://dx.doi.org/10.1002/cpa.20124

[3] D. Donoho, "For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics*, vol. 59(6), 2006.

[4] D. Donoho and J. Tanner, "Counting faces of randomly projected polytopes when the projection radically lowers dimension," *Journal of the American Mathematical Society*, vol. 22, no. 1, pp. 1–53, 2009.

[5] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse mri: The application of compressed sensing for rapid mr imaging," *Magnetic Resonance in Medicine*, vol. 58, no. 6, pp. 1182–1195, 2007. [Online]. Available: http://dx.doi.org/10.1002/mrm.21391

[6] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *CoRR*, vol. abs/0805.4471, 2008. [Online]. Available: http://arxiv.org/abs/0805.4471

[7] E. J. Candès, T. Strohmer, and V. Voroninski, "Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming," *Comm. Pure Appl. Math.*, vol. 66, no. 8, pp. 1241–1274, 2013.

[8] E. Candès, Y. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM J. Imaging Sci.*, vol. 6, no. 1, pp. 199–225, 2013. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/110848074

[9] A. Ahmed, B. Recht, and J. K. Romberg, "Blind deconvolution using convex programming," *CoRR*, vol. abs/1211.5608, 2012. [Online]. Available: http://arxiv.org/abs/1211.5608

[10] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[11] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller, "Striving for simplicity: The all convolutional net," *CoRR*, vol. abs/1412.6806, 2014. [Online]. Available: http://arxiv.org/abs/1412.6806

[12] I. J. Goodfellow, J. Pouget-Abadie, B. Mirza, Mehdi; Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv:1406.2661*, 2014.

[13] Y. Hong, U. Hwang, J. Yoo, and S. Yoon, "How generative adversarial nets and its variants work: An overview of gan," *arXiv preprint arXiv:1711.05914*, 2017.

[14] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[15] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," *arXiv preprint arXiv:1401.4082*, 2014.

[16] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," *arXiv preprint arXiv:1601.06759*, 2016.

[17] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[18] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[19] G. Yang, S. Yu, H. Dong, G. Slabaugh, P. L. Dragotti, X. Ye, F. Liu, S. Arridge, J. Keegan, Y. Guo *et al.*, "Dagan: Deep de-aliasing generative adversarial networks for fast compressed sensing mri reconstruction," *IEEE Transactions on Medical Imaging*, 2017.

[20] B. Kelly, T. P. Matthews, and M. A. Anastasio, "Deep learning-guided image reconstruction from incomplete data," *arXiv preprint arXiv:1709.00584*, 2017.

[21] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll, "Learning a variational network for reconstruction of accelerated mri data," *Magnetic resonance in medicine*, 2017.

[22] T. M. Quan, T. Nguyen-Duc, and W.-K. Jeong, "Compressed sensing mri reconstruction with cyclic loss in generative adversarial networks," *arXiv preprint arXiv:1709.00753*, 2017.

[23] S. U. H. Dar and T. Çukur, "A transfer-learning approach for accelerated mri using deep neural networks," *arXiv preprint arXiv:1710.02615*, 2017.

[24] J. Adler and O. Öktem, "Learned primal-dual reconstruction," *IEEE Transactions on Medical Imaging*, 2018.

[25] P. Moeskops, M. Veta, M. W. Lafarge, K. A. Eppenhof, and J. P. Pluim, "Adversarial training and dilated convolutions for brain mri segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017, pp. 56–64.

[26] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "Generative adversarial networks for noise reduction in low-dose ct," *IEEE transactions on medical imaging*, vol. 36, no. 12, pp. 2536–2545, 2017.

[27] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with context-aware generative adversarial networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 417–425.

[28] F. Mahmood, R. Chen, and N. J. Durr, "Unsupervised reverse domain adaption for synthetic medical images via adversarial training," *arXiv preprint arXiv:1711.06606*, 2017.

[29] S. Kohl, D. Bonekamp, H.-P. Schlemmer, K. Yaqubi, M. Hohenfellner, B. Hadaschik, J.-P. Radtke, and K. Maier-Hein, "Adversarial networks for the detection of aggressive prostate cancer," *arXiv preprint arXiv:1702.08014*, 2017.

[30] D. Mahapatra, "Retinal vasculature segmentation using local saliency maps and generative adversarial networks for image super resolution," *arXiv preprint arXiv:1710.04783*, 2017.

[31] W. Dai, J. Doyle, X. Liang, H. Zhang, N. Dong, Y. Li, and E. P. Xing, "Scan: Structure correcting adversarial network for chest x-rays organ segmentation," *arXiv preprint arXiv:1703.08770*, 2017.

[32] Y. Xue, T. Xu, H. Zhang, R. Long, and X. Huang, "Segan: Adversarial network with multi-scale l_1 loss for medical image segmentation," *arXiv preprint arXiv:1706.01805*, 2017.

[33] Y. Rivenson, Z. Göröcs, H. Günaydin, Y. Zhang, H. Wang, and A. Ozcan, "Deep learning microscopy," *Optica*, vol. 4, no. 11, pp. 1437–1443, 2017.

[34] X.-J. Mao, C. Shen, and Y.-B. Yang, "Image restoration using convolutional auto-encoders with symmetric skip connections. arxiv preprint," *arXiv preprint arXiv:1606.08921*, vol. 2, 2016.

[35] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5485–5493.

[36] C. K. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Huszár, "Amortised map inference for image super-resolution," *arXiv preprint arXiv:1610.04490*, 2016.

[37] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," *arXiv preprint*, 2016.

[38] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711.

[39] S. Lohit, K. Kulkarni, R. Kerviche, P. Turaga, and A. Ashok, "Convolutional neural networks for non-iterative reconstruction of compressively sensed images," *arXiv preprint arXiv:1708.04669*, 2017.

[40] Y. Liu, F. Li, L. Xin, J. Fu, and P. Huang, "High efficient optical remote sensing images acquisition for nano-satellite: reconstruction algorithms," in *Image and Signal Processing for Remote Sensing XXIII*, vol. 10427. International Society for Optics and Photonics, 2017, p. 104271Y.

[41] A. Mousavi, A. B. Patel, and R. G. Baraniuk, "A deep learning approach to structured signal recovery," in *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*. IEEE, 2015, pp. 1336–1343.

[42] A. Mousavi and R. G. Baraniuk, "Learning to invert: Signal recovery via deep convolutional networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2272–2276.

[43] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *European Conference on Computer Vision*. Springer, 2016, pp. 597–613.

[44] A. Lucas, M. Iliadis, R. Molina, and A. K. Katsaggelos, "Using deep neural networks for inverse problems in imaging: Beyond analytical methods," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 20–36, 2018.

[45] M. Mardani, E. Gong, J. Y. Cheng, S. Vasanawala, G. Zaharchuk, M. Alley, N. Thakur, S. Han, W. Dally, J. M. Pauly *et al.*,

"Deep generative adversarial networks for compressed sensing automates mri," *arXiv preprint arXiv:1706.00051*, 2017.

[46] M. Mardani, H. Monajemi, V. Papyan, S. Vasanawala, D. Donoho, and J. Pauly, "Recurrent generative adversarial networks for proximal learning and automated compressive image recovery," *arXiv preprint arXiv:1711.10046*, 2017.

[47] O. Rippel and L. Bourdev, "Real-time adaptive image compression," *arXiv preprint arXiv:1705.05823*, 2017.

[48] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," *arXiv:1703.03208*, 2017.

[49] S. Mallat, "Group invariant scattering," *Comm. Pure Appl. Math.*, vol. 65, no. 10, p. 13311398, 2012.

[50] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," *CoRR*, vol. abs/1602.06664, 2016. [Online]. Available: http://arxiv.org/abs/1602.06664

[51] Bandeira, Boumal, and Voroninski, "On the low-rank approach for semidefinite programs arising in synchronization and community detection," *JMLR*, vol. 49, pp. 1–22, 2016.

[52] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via wirtinger flow: Theory and algorithms," *Information Theory, IEEE Transactions on*, vol. 61, no. 4, pp. 1985–2007, 2015.

[53] Y. Chen and E. Candes, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," in *Advances in Neural Information Processing Systems*, 2015, pp. 739–747.

[54] X. Li, S. Ling, T. Strohmer, and K. Wei, "Rapid, robust, and reliable blind deconvolution via nonconvex optimization," *arXiv preprint arXiv:1606.04933*, 2016.

[55] C. Ma, K. Wang, Y. Chi, and Y. Chen, "Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution," *arXiv preprint arXiv:1711.10467*, 2017.

[56] W. Huang and P. Hand, "Blind deconvolution by a steepest descent algorithm on a quotient manifold," *arXiv preprint arXiv:1710.03309*, 2017.

[57] T. Maunu, T. Zhang, and G. Lerman, "A well-tempered landscape for non-convex robust subspace recovery," *arXiv preprint arXiv:1706.03896*, 2017.

[58] Y. Chen and E. Candes, "The projected power method: An efficient algorithm for joint alignment from pairwise differences," *arXiv preprint arXiv:1609.05820*, 2016.

[59] S. Arora, Y. Liang, and T. Ma, "Why are deep nets reversible: A simple theory, with implications for training," *CoRR*, vol. abs/1511.05653, 2015. [Online]. Available: http://arxiv.org/abs/1511.05653

[60] K. G. Murty and S. N. Kabadi, "Some np-complete problems in quadratic and nonlinear programming," *Mathematical Programming*, vol. 39, no. 2, pp. 117–129, Jun 1987.

[61] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.

[62] Z. C. Lipton and S. Tripathi, "Precise recovery of latent vectors from generative adversarial networks," *arXiv preprint arXiv:1702.04782*, 2017.

[63] A. Ilyas, A. Jalal, E. Asteri, C. Daskalakis, and A. G. Dimakis, "The robust manifold defense: Adversarial training using generative models," *arXiv preprint arXiv:1712.09196*, 2017.

[64] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman, "Pixeldefend: Leveraging generative models to understand and defend against adversarial examples," *arXiv preprint arXiv:1710.10766*, 2017.

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

[66] X. Yuan, P. He, Q. Zhu, R. R. Bhat, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *arXiv preprint arXiv:1712.07107*, 2017.

[67] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *arXiv preprint arXiv:1801.00553*, 2018.

[68] Y. Li and Y. Liang, "Learning overparameterized neural networks via stochastic gradient descent on structured data," in *Advances in Neural Information Processing Systems*, 2018, pp. 8157–8166.

[69] S. S. Du, X. Zhai, B. Poczos, and A. Singh, "Gradient descent provably optimizes over-parameterized neural networks," *arXiv preprint arXiv:1810.02054*, 2018.

[70] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parameterization," *arXiv preprint arXiv:1811.03962*, 2018.

[71] S. Oymak and M. Soltanolkotabi, "Towards moderate overparameterization: global convergence guarantees for training shallow neural networks," *arXiv preprint arXiv:1902.04674*, 2019.

[72] F. Ma, U. Ayaz, and S. Karaman, "Invertibility of convolutional generative networks from partial measurements," in *Advances in Neural Information Processing Systems*, 2018, pp. 9628–9637.

[73] B. Barak, S. B. Hopkins, J. Kelner, P. K. Kothari, A. Moitra, and A. Potechin, "A nearly tight sum-of-squares lower bound for the planted clique problem," *arXiv preprint arXiv:1604.03084*, 2016.

[74] X. Li and V. Voroninski, "Sparse signal recovery from quadratic measurements via convex programming," *SIAM Journal on Mathematical Analysis*, vol. 45, no. 5, pp. 3019–3033, 2013.

[75] P. Hand, O. Leong, and V. Voroninski, "Phase retrieval under a generative prior," in *Advances in Neural Information Processing Systems*, 2018, pp. 9136–9146.

[76] S. Velankar, C. Best, B. Beuth, C. Boutselakis, N. Cobley, A. Sousa Da Silva, D. Dimitropoulos, A. Golovin, M. Hirshberg, M. John *et al.*, "Pdbe: protein data bank in europe," *Nucleic acids research*, vol. 38, no. suppl_1, pp. D308–D317, 2009.

[77] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," in *Compressed Sensing: Theory and Applications*, Y. Eldar and G. Kutyniok, Eds. Cambridge University Press, 2012.

[78] J. G. Wendel, "A problem in geometric probability," *Math. Scand*, vol. 11, pp. 109–111, 1962.