Constant-Expansion Suffices for Compressed Sensing with Generative Priors

Constantinos Daskalakis
MIT
costis@mit.edu

Dhruv Rohatgi
MIT
drohatgi@mit.edu

Manolis Zampetakis
MIT
mzampet@mit.edu

June 29, 2020

Abstract

Generative neural networks have been empirically found very promising in providing effective structural priors for compressed sensing, since they can be trained to span low-dimensional data manifolds in high-dimensional signal spaces. Despite the non-convexity of the resulting optimization problem, it has also been shown theoretically that, for neural networks with random Gaussian weights, a signal in the range of the network can be efficiently, approximately recovered from a few noisy measurements. However, a major bottleneck of these theoretical guarantees is a network *expansivity* condition: that each layer of the neural network must be larger than the previous by a logarithmic factor. Our main contribution is to break this strong expansivity assumption, showing that *constant* expansivity suffices to get efficient recovery algorithms, besides it also being information-theoretically necessary. To overcome the theoretical bottleneck in existing approaches we prove a novel uniform concentration theorem for random functions that might not be Lipschitz but satisfy a relaxed notion which we call "pseudo-Lipschitzness." Using this theorem we can show that a matrix concentration inequality known as the *Weight Distribution Condition (WDC)*, which was previously only known to hold for Gaussian matrices with logarithmic aspect ratio, in fact holds for constant aspect ratios too. Since the WDC is a fundamental matrix concentration inequality in the heart of all existing theoretical guarantees on this problem, our tighter bound immediately yields improvements in all known results in the literature on compressed sensing with deep generative priors, including one-bit recovery, phase retrieval, low-rank matrix recovery and more.

1 Introduction

Compressed sensing is the study of recovering a high-dimensional signal from as few measurements as possible, under some structural assumption about the signal that pins it into a low-dimensional subset of the signal space. The assumption that has driven the most research is sparsity; it is well known that a k-sparse signal from $\mathbb{R}^n$ can be efficiently recovered from only $O(k \log n)$ linear measurements [4]. Numerous variants of this problem have been studied, e.g. tolerating measurement noise, recovering signals that are only approximately sparse, recovering signals from phaseless measurements or one-bit measurements, to name just a few [1].

However, in many applications sparsity in some basis may not be the most natural structural assumption to make for the signal to be reconstructed. Given recent strides in performance of generative neural networks [7, 6, 12, 3, 13], there is strong evidence that data from some domain $\mathcal{D}$, e.g. faces, can be used to identify a deep neural network $G : \mathbb{R}^k \to \mathbb{R}^n$, where $k \ll n$, whose range $G(x)$ over varying "latent codes" $x \in \mathbb{R}^k$ covers well the objects of $\mathcal{D}$. Thus, if we want to perform compressed sensing on signals from this domain $\mathcal{D}$, the machine learning paradigm suggests that a reasonable structural assumption to make is that the signal lies in the range of G, suggesting the following problem, first proposed in [2]:

It has been shown empirically that this problem (and some variants of it) can be solved efficiently [2]. It has also been shown empirically that the quality of the reconstructed signals in the low number of measurements regime might greatly outperform those reconstructed using a sparsity assumption. It has even been shown that the network G need not be trained on data from the domain of interest $\mathcal{D}$ but that

COMPRESSED SENSING WITH A DEEP GENERATIVE PRIOR (CS-DGP)
Given: Deep neural network $G : \mathbb{R}^k \to \mathbb{R}^n$; measurement matrix $A \in \mathbb{R}^{m \times n}$, where $m \ll n$.
Given: $y = AG(x^*) + e \in \mathbb{R}^m$, for some *unknown* latent vector $x^* \in \mathbb{R}^k$, noise vector $e \in \mathbb{R}^m$.
Goal: Recover x^* (or in a different variant of the problem just $G(x^*)$).

a convolutional neural network G with random weights might suffice to regularize the reconstruction well [19, 20].

Despite the non-convexity of the optimization problem, some theoretical guarantees have also emerged [9, 11], when G is a fully-connected ReLU neural network of the following form (where d is the depth):

$$G(x) = \mathsf{ReLU}(W^{(d)}(\dots \mathsf{ReLU}(W^{(2)}(\mathsf{ReLU}(W^{(1)}x)))\dots)), \tag{1}$$

where each $W^{(i)}$ is a matrix of dimension $n_i \times n_{i-1}$, such that $n_0 = k$ and $n_d = n$. These theoretical guarantees mirror well-known results in sparsity-based compressed sensing, where efficient recovery is possible if the measurement matrix A satisfies a certain deterministic condition, e.g. the Restricted Isometry Property. But for arbitrary G, recovery is in general intractable [14], so some assumption about G must also be made. Specifically, it has been shown in prior work that, if the measurement matrix A satisfies a certain *Range Restricted Isometry Condition (RRIC)* with respect to G, and each weight matrix $W^{(i)}$ satisfies a *Weight Distribution Condition (WDC)*, then x^* can be efficiently recovered up to error roughly $||e||$ from $O(k \cdot \log(n) \cdot \text{poly}(d))$ measurements [9, 11]. See Section 3 for a definition of the WDC, and Appendix 6 for a definition of the RRIC.

But it's critical to understand when these conditions are satisfied (for example, in the sparsity setting, the Restricted Isometry Property is satisfied by i.i.d. Gaussian matrices when $m \geq k \log n$). Similarly, the RRIC has been shown to hold when A is i.i.d. Gaussian and $m \geq k \cdot \log(n) \cdot \text{poly}(d)$, which is an essentially optimal measurement complexity if d is constant. However, until this work, the WDC has seemed more onerous. Under the assumption that each $W^{(i)}$ has i.i.d. Gaussian entries, the WDC was previously only known to hold when $n_i \geq cn_{i-1} \log n_{i-1}$: i.e. when every layer of the neural network is larger than the previous by a logarithmic factor. This *expansivity* condition is a major limitation of the prior theory, since in practice neural networks do not expand at every layer.

Our work alleviates this limitation, settling a problem left open in [9, 11] and recently also posed in survey [16]. We show that the WDC holds when $n_i \geq cn_{i-1}$. This proves the following result, where our contribution is to replace $n_i \geq n_{i-1} \cdot \log(n_{i-1}) \cdot \text{poly}(d)$ with $n_i \geq n_{i-1} \cdot \text{poly}(d)$.

Theorem 1.1. *Suppose that each weight matrix has expansion $n_i \geq n_{i-1} \cdot \text{poly}(d)$, and the number of measurements is $m \geq k \cdot \log(n) \cdot \text{poly}(d)$. Suppose that A has i.i.d. Gaussian entries $N(0, 1/m)$ and each $W^{(i)}$ has i.i.d. Gaussian entries $N(0, 1/n_i)$. Then there is an efficient gradient-descent based algorithm which given G, A, and $y = A \cdot G(x^*) + e$, outputs, with probability at least $1 - e^{-k/\text{poly}(d)}$, an estimate $\tilde{x} \in \mathbb{R}^k$ satisfying $\|\tilde{x} - x^*\| \leq O(2^d \|e\|)$ when $\|e\|$ is sufficiently small.*

We note that the dependence in d of the expansivity, number of measurements, and error in our theorem is the same as in [11]. Moreover, the techniques developed in this paper yield several generalizations of the above theorem, stated informally below, and discussed further in Section D.

Theorem 1.2. *Suppose G is a random neural network with constant expansion, and conditions analogous to those of Theorem 1.1 are satisfied. Then the following results also hold with high probability. The Gaussian noise setting admits an efficient algorithm with recovery error $\Theta(k/m)$. Phase retrieval and one-bit recovery with generative prior G have no spurious local minima. And compressed sensing with a two-layer deconvolutional prior has no spurious local minima.*

To see why expansivity plays a role in the first place, we provide some context:

Global landscape analysis. The theoretical guarantees of [9, 11] fall under an emerging method for analyzing non-convex optimization problems called *global landscape analysis* [18]. Given a non-convex function f, the basic goal is to show that f does not have spurious local minima, implying that gradient descent will

(eventually) converge to a global minimum. Stronger guarantees may provide bounds on the convergence rate. In less well-behaved settings, the goal may be to prove guarantees in a restricted region, or show that the local minima inform the global minimum.

In stochastic optimization settings wherein f is a random function, global landscape analysis typically consists of two steps: first, prove that $\mathbb{E}f$ is well-behaved, and second, apply concentration results to prove that, with high probability, f is sufficiently close to $\mathbb{E}f$ that no pathological behavior can arise. The analysis of compressed sensing with generative priors by [9] follows this two-step outline (see Section C for a sketch). The second step requires inducting on the layers of the network. For each layer, it's necessary to prove uniform concentration of a function which takes a weight matrix $W^{(i)}$ as a parameter; this concentration is precisely the content of the WDC. As a general rule, tall random matrices concentrate more strongly, which is why proving the WDC for random matrices requires an expansivity condition (a lower bound on the aspect ratio of each weight matrix).

Concentration of random functions. The uniform concentration required for the above analysis can be abstracted as follows. Given a family of functions $\{f_t : \mathcal{X} \to \mathbb{R}\}_{t\in\Theta}$ on some metric space $\mathcal{X}$, and given a distribution P on Θ, we pick a random function f_t. We seek to show that with high probability, f_t is uniformly near $\mathbb{E}f_t$. A generic approach to solve this problem is via Lipschitz bounds. If f_t is sufficiently Lipschitz for all $t \in \Theta$, and $f_t(x)$ is near $\mathbb{E}f_t(x)$ with high probability for any single $x \in \mathcal{X}$, then by union bounding over an ϵ-net, uniform concentration follows.

However, in the global landscape analysis conducted by [9], the functions f_t have poor Lipschitz constants. Pushing through the ϵ-net argument therefore requires very strong pointwise concentration, and hence the expansivity condition is necessitated.

1.1 Technical contributions

Concentration of Lipschitz random functions is a widely-used tool in probability theory, which has found many applications in global landscape analysis for the purposes of understanding non-convex optimization, as outlined above. For the functions arising in our analysis, however, Lipschitzness is actually *too strong* a property, and leads to suboptimal results. A main technical contribution of our paper is to define a relaxed notion of *pseudo-Lipschitz* functions and derive a concentration inequality for pseudo-Lipschitz random functions. This serves as a central tool in our analysis, and is a general tool that we envision will find other applications in probability and non-convex optimization.

Informally, a function family $\{f_t : \mathcal{X} \to \mathbb{R}\}_{t\in\Theta}$ is *pseudo-Lipschitz* if for every t there is a *pseudo-ball* such that f_t has small deviations when its argument is varied by a vector within the pseudo-ball. If for every t the pseudo-ball is simply a ball, then the function family is plain Lipschitz. But this definition is more flexible; the pseudo-ball can be any small-diameter, convex body with non-negligible volume and, importantly, every t could have a different pseudo-ball of small deviations. We show that uniform concentration still holds; here is a simplified (and slightly specialized) statement of our result (presented in full detail in Section 4 along with a formal definition of pseudo-Lipschitzness):

Theorem 1.3 (Informal-Concentration of pseudo-Lipschitz random functions)**.** *Let θ be a random variable taking values in Θ. Let $\{f_t : \mathcal{X} \to \mathbb{R}\}_{t\in\Theta}$ be a function family, where $\mathcal{X}$ is a subset of $\mathcal{B}$, the unit-ball in $\mathbb{R}^k$. Suppose that:*

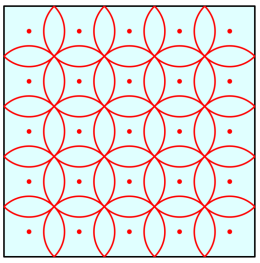
(1) For any fixed x, the random variable $f_\theta(x)$ is well-concentrated around its mean,

(2) $\{f_t\}_{t\in\Theta}$ is pseudo-Lipschitz,

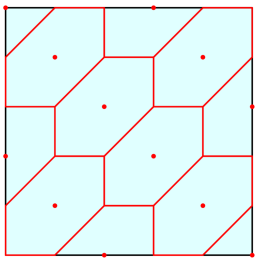
(3) $\mathbb{E}f_\theta(x)$ is Lipschitz in x.

Then f_θ is well-concentrated around its mean, uniformly in x. Quantitatively, the strength of the concentration in (1) only needs to be proportional to the inverse volume of the pseudo-balls of small deviation guaranteed by (2) raised to the power k, i.e. the number of pseudo-balls needed to cover $\mathcal{B}$.

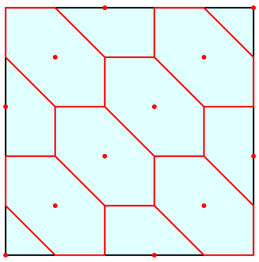
This result achieves asymptotically stronger results than are possible through Lipschitz concentration. Where does the gain come from? For each parameter t, consider the pseudo-ball of ϵ-deviations of f_t, as



(a) Spherical ϵ-Net: for all t, f_t deviates by at most ϵ within each ball.



(b) Aspherical ϵ-Net: for a *specific* t, f_t deviates by at most ϵ within each weirdly-shaped pseudo-ball.



(c) Aspherical ϵ-Net$'$: for a different t', $f_{t'}$ deviates by at most ϵ within each weirdly-shaped pseudo-ball.

Figure 1: A parameter-independent spherical ϵ-net, versus a parameter-dependent aspherical ϵ-net for two specific parameters; in particular, the parameter t determines the rotational angle of the weirdly-shaped pseudo-balls within which f_t changes by at most ϵ. The shaded square is $\mathcal{X}$. Each ball in the leftmost panel is the intersection of weirdly-shaped pseudo-balls with the same center, under all possible rotations. As such the radius of this ball is small, so to cover $\mathcal{X}$ with such balls we need a lot of small balls. Instead, the aspherical parameter-dependent ϵ-nets are more efficient.

guaranteed by the pseudo-Lipschitzness. A standard ϵ-net would be covering the metric space by balls (as exemplified in Figure 1a), each of which would have to lie in the intersection of the pseudo-balls of all parameters t. If for each parameter the pseudo-ball is "wide" in a different direction (see Figures 1b and 1c for a schematic), then the balls of the standard ϵ-net may be very small compared to any pseudo-ball, and the size of the standard ϵ-net could be very large compared to the size of the ϵ-net obtained for any fixed t using pseudo-balls. Hence, the standard Lipschitzness argument may require much stronger concentration in (1) than our result does, in order to union bound over a potentially much larger ϵ-net.

There is an obvious technical obstacle to our proof: the pseudo-balls depend on the parameter t, so an efficient covering of $\mathcal{X}$ by pseudo-balls will necessarily depend on t. It's then unclear how to union bound over the centers of the pseudo-balls (as in the standard Lipschitz concentration argument). We resolve the issue with a decoupling argument. Thus, we ultimately show that under mild conditions, the pseudo-Lipschitz random function is asymptotically as well-concentrated as a Lipschitz random function—even though its Lipschitz constant may be much worse.

Applications. With our new technique, we are able to show that the WDC holds for $n \times k$ Gaussian matrices whenever $n \geq ck$, for some absolute constant c, where previously it was only known to hold if $n \geq ck \log k$. As a consequence, we show Theorem 1.1: that compressed sensing with a random neural network prior does not require the logarithmic expansivity condition.

In addition, there has been follow-up research on variations of the CS-DGP problem described above. The WDC is a critical assumption which enables efficient recovery in the setting of Gaussian noise [10], as well as global landscape analysis in the settings of phaseless measurements [8], one-bit (sign) measurements [17], two-layer convolutional neural networks [15], and low-rank matrix recovery [5]. Moreover, there are currently no known theoretical results in this area—compressed sensing with generative priors—that avoid the WDC: hence, until now, logarithmic expansion was necessary to achieve any provable guarantees. Our result extends the prior work on these problems, in a black-box fashion, to random neural networks with constant expansion. We refer to Appendix D for details about these extensions.

Lower bound. As a complementary contribution, we also provide a simple lower bound on the expansion required to recover the latent vector. This lower bound is strong in several senses: it applies to one-layer neural networks, even in the absence of compression and noise, and it is an information theoretic lower bound. Without compression and noise, the problem is simply inverting a neural network, and it has been shown [14] that inversion is computationally tractable if the network consists of Gaussian matrices with expansion $2 + \epsilon$. In this setting our lower bound is tight: we show that expansion by a factor of 2 is in fact necessary for exact recovery. Details are deferred to Appendix A.

1.2 Roadmap

In Section 2, we introduce basic notation. In Section 3 we formally introduce the Weight Distribution Condition, and present our main theorem about the Weight Distribution Condition for random matrices. In Section 4 we define pseudo-Lipschitz function families, allowing us to formalize and prove Theorem 1.3. In Section 5, we show how uniform concentration of pseudo-Lipschitz random functions implies that Gaussian random matrices with constant expansion satisfy the WDC. Finally, in Section 6 we prove Theorem 1.1.

2 Preliminaries

For any vector $v \in \mathbb{R}^k$, let $\|v\|$ refer to the ℓ_2 norm of x, and for any matrix $A \in \mathbb{R}^{n\times k}$ let $\|A\|$ refer to the operator norm of A. If A is symmetric, then $\|A\|$ is also equal to the spectral norm $\lambda_{\max}(A)$.

Let $\mathcal{B}$ refer to the unit ℓ_2 ball in $\mathbb{R}^k$ centered at 0, and let S^{k-1} refer to the corresponding unit sphere. For a set $S \subseteq \mathbb{R}^k$ let $\alpha S = \{\alpha x : x \in S\}$ and let $\beta + S = \{\beta + x : x \in S\}$.

For a matrix $W \in \mathbb{R}^{n\times k}$ and a vector $x \in \mathbb{R}^k$, let $W_{+,x}$ be the $n \times k$ matrix $W_{+,x} = \mathrm{diag}(Wx > 0)W$. That is, row i of $W_{+,x}$ is equal to W_i if $W_i x > 0$, and is equal to 0 otherwise.

3 Weight Distribution Condition

In the existing literature in compressed sensing, many results for recovery of a sparse signal are based on an assumption on the measurement matrix that is called the Restricted Isometry Property (RIP). Many results then follow the same paradigm: they first prove that sparse recovery is possible under the RIP, and then show that a random matrix drawn from some specific distribution or class of distributions satisfies the RIP. The same paradigm has been followed in the literature of signal recovery under the deep generative prior. In virtually all of these results, the properties that correspond to the RIP property are the combination of the Range Restricted Isometry Condition (RRIC) and the Weight Distribution Condition (WDC). Our main focus in this paper is to improve upon the existing results related to the WDC property. The WDC has the following definition due to [9].

Definition 3.1. A matrix $W \in \mathbb{R}^{n\times k}$ is said to satisfy the *(normalized) Weight Distribution Condition (WDC)* with parameter ϵ if for all nonzero $x, y \in \mathbb{R}^k$ it holds that $\left\| \frac{1}{n} W_{+,x}^T W_{+,y} - Q_{x,y} \right\| \leq \epsilon$, where $Q_{x,y} = \frac{1}{n}\mathbb{E} W_{+,x}^T W_{+,y}$ (with expectation over i.i.d. $N(0,1)$ entries of W).

Remark. Note that the factor $1/n$ in front of $W_{+,x}^T W_{+,y}$ is not present in the actual condition [9], hence the term "normalized". We scale up W by a factor of $\sqrt{n}$ to simplify later notation.

In Appendix C, we provide a detailed explanation of why the WDC arises and we also give a sketch of the basic theory of global landscape analysis for compressed sensing with generative priors.

3.1 Weight Distribution Condition from constant expansion

To prove Theorem 1.1, our strategy is to prove that the WDC holds for Gaussian random matrices with constant aspect ratio:

Theorem 3.2. *There are constants $c, C > 0$ with the following property. Let $\epsilon > 0$ and let $n, k \in \mathbb{N}$. Suppose that $n \geq c \cdot k \cdot \epsilon^{-2} \log(1/\epsilon)$. If $W \in \mathbb{R}^{n\times k}$ is a matrix with independent entries drawn from $N(0,1)$, then W satisfies the normalized WDC with parameter ϵ, with probability at least $1 - e^{-Ck}$.*

Equivalently, if W has entries i.i.d. $N(0, 1/n)$, then it satisfies the unnormalized WDC with high probability. The proof of 3.2 is provided in Section 5. It uses concentration of pseudo-Lipschitz functions, which are introduced formally in Section 4. As shown in Section 6, Theorem 1.1 then follows from prior work.

4 Uniform concentration beyond Lipschitzness

In this section we present our main technical result about uniform concentration bounds. We generalize a folklore result about uniform concentration of Lipschitz functions by generalizing Lipschitzness to a weaker condition which we call *pseudo-Lipschitzness*. This concentration result can be used to prove Theorem 3.2. Moreover we believe that it may have broader applications.

Before stating our result, let us first define the notion of pseudo-Lipschitzness of function families and give some comparison with the classical notion of Lipschitzness. Let $\{f_t : (\mathbb{R}^k)^d \to \mathbb{R}\}$ be a family of functions over matrices parametrized by $t \in \Theta$. We have the following definitions:

Definition 4.1. A set system $\{B_t \subseteq \mathbb{R}^k : t \in \Theta\}$ is (δ, γ)-wide if $B_t = -B_t$, B_t is convex, and

$$\mathrm{Vol}(B_t \cap \delta\mathcal{B}) \geq \gamma \,\mathrm{Vol}(\delta\mathcal{B}) \qquad \text{for all } t \in \Theta.$$

Definition 4.2 (Pseudo-Lipschitzness)**.** Suppose that there exists a (δ, γ)-wide set system $\{B_t \subseteq \mathbb{R}^k : t \in \Theta\}$, such that

$$|f_t(x) - f_t(y)| \leq \epsilon$$

for any $t \in \Theta$ and $x, y \in (\mathbb{R}^k)^d$ with $y_i - x_i \in B_t$ for all $i \in [d]$. Then we say that $\{f_t\}_{t\in\Theta}$ is $(\epsilon, \delta, \gamma)$-pseudo-Lipschitz.

Example 4.3. Let $\Theta = \{w \in \mathbb{R}^k : \|w\| \leq 2\sqrt{k}\}$ and let $\epsilon > 0$. Then the family of functions $\{f_w(x) = w \cdot x\}_{w\in\Theta}$ is only $2\sqrt{k}$-Lipschitz. So to have $|f_w(x) - f_w(y)| \leq \epsilon$ we need $\|x - y\| \leq \epsilon/(2\sqrt{k})$.

On the other hand, it can be seen that the set system $\{B_w \subseteq \mathbb{R}^k : w \in \Theta\}$ defined by $B_w = \{x \in \mathbb{R}^k : |w \cdot x| \leq \epsilon\}$ is $(c\epsilon, 1/2)$-wide for a constant $c > 0$ (by standard arguments about spherical caps). Therefore the family of functions $f_w(x) = w \cdot x$ is $(\epsilon, c\epsilon, 1/2)$-pseudo-Lipschitz.

Our main technical result is that the above relaxation of Lipschitzness suffices to obtain strong uniform concentration of measure results; see Section 4.1 for the proof.

Theorem 4.4. *Let θ be a random variable taking values in Θ. Let $\{f_t : (\mathbb{R}^k)^d \to \mathbb{R} : t \in \Theta\}$ be a function family, and let $g : (\mathbb{R}^k)^d \to \mathbb{R}$ be a function. Let $\epsilon, \gamma, D > 0$ and $\delta \in (0, 1)$. Define the spherical shell $\mathcal{H} = (1 + \delta/2)\mathcal{B} \setminus (1 - \delta/2)\mathcal{B}$ in $\mathbb{R}^k$. Suppose that:*

1. *For any fixed $x \in \mathcal{H}^d$,*
$$\Pr_\theta[f_\theta(x) \leq g(x) + \epsilon] \geq 1 - p,$$
2. *$\{f_t\}_{t\in\Theta}$ is $(\epsilon, \delta, \gamma)$-pseudo-Lipschitz,*
3. *$|g(x) - g(y)| \leq D$ whenever $x \in (S^{k-1})^d$, $y \in (\mathbb{R}^k)^d$, and $\|y_i - x_i\|_2 \leq \delta$ for all $i \in [d]$.*

Then:

$$\Pr_\theta\left[f_\theta(x) \leq g(x) + 2\epsilon + D,\ \forall x \in (S^{k-1})^d\right] \geq 1 - \gamma^{-2d}(4/\delta)^{2kd}p.$$

As a comparison, if the family $\{f_t\}_t$ were simply L-Lipschitz, then uniform concentration would hold with probability $1 - (3L/\epsilon)^{kd}p$ by standard arguments. So the "effective Lipschitz constant" of an $(\epsilon, \delta, \gamma)$-pseudo-Lipschitz family is $O(\epsilon/\delta)$ when $\gamma = \Omega(1)$.

4.1 Proof of Theorem 4.4

Let $\{B_t\}_{t\in\Theta}$ be a family of sets $B_t \subseteq \mathbb{R}^k$ witnessing that $\{f_t\}_{t\in\Theta}$ is $(\epsilon, \delta, \gamma)$-pseudo-Lipschitz—i.e. $\{B_t\}_{t\in\Theta}$ is (δ, γ)-wide, and $|f_t(x) - f_t(y)| \leq \epsilon$ whenever $t \in \Theta$ and $x, y \in (\mathbb{R}^k)^d$ with $y_i - x_i \in B_t$ for all $i \in [d]$. The standard proof technique for uniform concentration is to fix an ϵ-net, and show that with high probability over the randomness θ, every point in the net satisfies the bound. Here instead, we use the pseudo-balls $\{B_t\}_{t\in\Theta}$ to construct a random, aspherical net $\mathcal{N} \subset \mathbb{R}^k$ that depends on θ and additional randomness. We'll show that with high probability over all the randomness, every point in the net satisfies the bound. In particular we use the following process to construct $\mathcal{N}$:

Let $x_0 = (1, 0, \ldots, 0) \in S^{k-1}$. We define $x_1, x_2, \cdots \in S^{k-1}$ iteratively. For $j \geq 0$, define $C_j \subseteq \mathbb{R}^k$ by

$$C_j = \bigcup_{0 \leq i \leq j} \left(x_i + \frac{1}{2}(B_\theta \cap \delta\mathcal{B}) \right).$$

For each $j \geq 0$, if $S^{k-1} \subseteq C_j$ then terminate the process. Otherwise, by some deterministic process, pick

$$x_{j+1} \in S^{k-1} \setminus C_j.$$

Let's say that the process terminates at step j_{last}, producing a sequence of random variables $\{x_0, x_1, \ldots, x_{j_{\text{last}}}\}$ (with randomness introduced by θ). Note that j_{last} is also a random variable.

For each $0 \leq j \leq j_{\text{last}}$, let y_j be the random variable

$$y_j \sim \text{Unif}\left(x_j + \frac{1}{2}(B_\theta \cap \delta\mathcal{B}) \right).$$

That is, y_j is a perturbation of x_j by a uniform sample from the bounded pseudo-ball. Let $\mathcal{N}$ be the set $\{y_0, y_1, \ldots, y_{j_{\text{last}}}\}$. Observe that $\mathcal{N} \subset \mathcal{H}$, and S^{k-1} is covered by the pseudo-balls $\{y_j + B_\theta\}_j$.

By a volume argument, we can upper bound $|\mathcal{N}| = j_{\text{last}} + 1$.

Lemma 4.5. *The size of $\mathcal{N}$ is at most $\gamma^{-1}(5/\delta)^k$.*

Proof. For each $0 \leq j \leq j_{\text{last}}$ define an auxiliary set $C'_j \subseteq \mathbb{R}^k$ by

$$C'_j = x_i + \frac{1}{4}(B_t \cap \delta\mathcal{B}).$$

We claim that the sets $C'_0, \ldots, C'_{j_{\text{last}}}$ are disjoint. Suppose not; then there are some indices $0 \leq j_1 < j_2 \leq j_{\text{last}}$ and some point $z \in \mathbb{R}^k$ such that $z - x_{j_1} \in \frac{1}{4}(B_\theta \cap \delta\mathcal{B})$ and also $z - x_{j_2} \in \frac{1}{4}(B_\theta \cap \delta\mathcal{B})$. It follows from convexity and symmetry of $B_t \cap \delta\mathcal{B}$ that $x_{j_1} - x_{j_2} \in \frac{1}{2}(B_\theta \cap \delta\mathcal{B})$. So $x_{j_2} \in C_{j_1} \subseteq C_{j_2 - 1}$, contradicting the definition of x_{j_2}. So the sets $C'_0, \ldots, C'_{j_{\text{last}}}$ are indeed disjoint.

But each C'_j is a subset of $(1 + \delta/4)\mathcal{B}$. By the volume lower bound on $B_t \cap \delta\mathcal{B}$, we have

$$\frac{\text{Vol}(C'_j)}{\text{Vol}((1 + \delta/4)\mathcal{B})} \geq \frac{\gamma 4^{-k} \text{Vol}(\delta\mathcal{B})}{\text{Vol}((1 + \delta/4)\mathcal{B})} = \gamma(1 + 4/\delta)^{-k}.$$

The lemma follows. □

We now show that with high probability the inequality $f_\theta(a) \leq g(a) + \epsilon$ holds for all $a \in \mathcal{N}^d$ simultaneously. The main idea is that the random perturbations partially decoupled the net $\mathcal{N}$ from θ. Since each point of the net is distributed uniformly over a set of non-negligible volume, the probability that any fixed $a \in \mathcal{N}^d$ fails the concentration inequality can be bounded against the probability that a uniformly random point from the shell $\mathcal{H}^d$ fails.

Lemma 4.6. *We have*

$$\Pr[\exists a \in \mathcal{N}^d : f_\theta(a) > g(a) + \epsilon] \leq \gamma^{-2d}(4/\delta)^{2kd} p.$$

Proof. For any $a \in \mathcal{H}^d$ let $E_{\text{bad}}(\theta, a)$ be the event that $f_\theta(a) > g(a) + \epsilon$. Fix $j \in \mathbb{N}^d$. Let A_j be the event that $j_1, \ldots, j_d \leq j_{\text{last}}$. (Recall that j_{last} is a deterministic function of θ which is random.) Let $X_1, \ldots, X_d \sim \text{Unif}(\mathcal{H})$ be independent uniform random variables over the shell $\mathcal{H}$. Let E_{cont} be the event that $X_i \in x_{j_i} + \frac{1}{2}(B_\theta \cap \delta\mathcal{B})$ for all $i \in [d]$, where for convenience we define $x_j = (1, 0, \ldots, 0) \in S^{k-1}$ for $j > j_{\text{last}}$. For any $t \in \Theta$, consider conditioning on $(\theta = t)$. Then j_{last} and $x_0, \ldots, x_{j_{\text{last}}}$ are deterministic; assume that A_j occurs. The conditional random vector $(y_{j_1}, \ldots, y_{j_d})|(\theta = t)$ has the uniform product distribution

$$\text{Unif}\left(\prod_{i=1}^{d} \left(x_{j_1} + \frac{1}{2}(B_t + \delta\mathcal{B}) \right) \right).$$

This is precisely the distribution of $(X_1, \ldots, X_d)|(E_{\text{cont}}, \theta = t)$. Thus,

$$\begin{aligned}\Pr[E_{\text{bad}}(\theta, (y_{j_1}, \ldots, y_{j_d}))|\theta = t] &= \Pr[E_{\text{bad}}(\theta, (X_1, \ldots, X_d))|E_{\text{cont}}, \theta = t] \\ &\leq \frac{\Pr[E_{\text{bad}}(\theta, (X_1, \ldots, X_d))|\theta = t]}{\Pr[E_{\text{cont}}|\theta = t]}. \end{aligned} \tag{2}$$

Since $(X_1, \ldots, X_d)$ are independent and uniformly distributed over $\mathcal{H}$,

$$\Pr[E_{\text{cont}}|\theta = t] = \left(\frac{\text{Vol}((B_t \cap \delta\mathcal{B})/2)}{\text{Vol}(\mathcal{H})}\right)^d \geq \left(\frac{2^{-k}\gamma \,\text{Vol}(\delta\mathcal{B})}{\text{Vol}((1+\delta/2)\mathcal{B})}\right)^d \geq \gamma^d(\delta/3)^{kd}.$$

Substituting into Equation (2) and integrating over all $\theta \in A_j$,

$$\Pr[E_{\text{bad}}(\theta, (y_{j_1}, \ldots, y_{j_d})) \wedge A_j] \leq \gamma^{-d}(3/\delta)^{kd} \Pr[E_{\text{bad}}(\theta, (X_1, \ldots, X_d)) \wedge A_j].$$

If $(X_1, \ldots, X_d)$ were deterministic then we would have $f_\theta(X_1, \ldots, X_d) \leq g(X_1, \ldots, X_d)]$ with probability at least $1 - p$. They are not deterministic, but they are independent of θ, which suffices to imply the above inequality. So

$$\Pr[E_{\text{bad}}(\theta, (y_{j_1}, \ldots, y_{j_d})) \wedge A_j] \leq \gamma^{-d}(3/\delta)^{kd} p.$$

Finally we take a union bound over j. By Lemma 4.5 we have $j_{\text{last}} < \gamma^{-1}(5/\delta)^k$. So

$$\Pr[\exists a \in \mathcal{N}^d : E_{\text{bad}}(\theta, a)] \leq (\gamma^{-1}(5/\delta)^k)^d \gamma^{-d}(3/\delta)^{kd} p \leq \gamma^{-2d}(4/\delta)^{2kd} p.$$

as claimed. □

We conclude the proof of Theorem 4.4. Suppose that the event of Lemma 4.6 fails; that is, for all $a \in \mathcal{N}^d$ we have $f_\theta(a) \leq g(a) + \epsilon$. Then let $b \in (S^{k-1})^d$. By construction of $\mathcal{N}$, for every $i \in [d]$ there is some $0 \leq j_i \leq j_{\text{last}}$ such that $b_i \in x_{j_i} + \frac{1}{2}(B_\theta \cap \delta\mathcal{B})$. But $y_{j_i} \in x_{j_i} + \frac{1}{2}(B_\theta \cap \delta\mathcal{B})$, so by convexity and symmetry of B_θ, it follows that $b_i - y_{j_i} \in B_\theta$. Hence by assumption, we have $|f_\theta(b) - f_\theta(y_{j_1}, \ldots, y_{j_d})| \leq \epsilon$. Since $(y_{j_1}, \ldots, y_{j_d}) \in \mathcal{N}^d$, we have

$$f_\theta(y_{j_1}, \ldots, y_{j_d}) \leq g(y_{j_1}, \ldots, y_{j_d}) + \epsilon.$$

Thus,

$$f_\theta(b) \leq f_\theta(y_{j_1}, \ldots, y_{j_d}) + \epsilon \leq g(y_{j_1}, \ldots, y_{j_d}) + 2\epsilon \leq g(b) + 2\epsilon + D$$

as desired. By Lemma 4.6, this uniform bound holds with probability at least $1 - \gamma^{-2d}(4/\delta)^{2kd} p$, over the randomness of θ and the net perturbations. So it also holds with at least this probability over just the randomness of θ.

5 Proof of Theorem 3.2

In [9], a weaker version of Theorem 3.2 was proven—it required a logarithmic aspect ratio (i.e. $n \geq \Omega(k \log k)$). The proof was by a standard ϵ-net argument. In Section 5.1 we discuss why Theorem 3.2 cannot be proven by standard arguments, and sketch how Theorem 4.4 yields a proof.

Then, in Section 5.2 we provide the full proof of Theorem 3.2.

Throughout, we let W be an $n \times k$ random matrix with independent rows $w_1, \ldots, w_n \sim N(0,1)^k$.

5.1 Outline

At a high level, the proof of Theorem 3.2 uses an ϵ-net argument, with several crucial twists. The first twist is borrowed from the prior work [9]: we wish to prove concentration for the random function

$$W_{+,x}^T W_{+,y} = \sum_{i=1}^n \mathbb{1}_{w_i^T x > 0} \mathbb{1}_{w_i^T y > 0} w_i w_i^T,$$

but it is not continuous in x and y. So for $\epsilon > 0$, define the continuous functions

$$h_{-\epsilon}(z) = \begin{cases} 0 & z \le -\epsilon, \\ 1 + z/\epsilon & -\epsilon \le z \le 0, \\ 1 & z \ge 0. \end{cases} \quad \text{and} \quad h_{\epsilon}(z) = \begin{cases} 0 & z \le 0, \\ z/\epsilon & 0 \le z \le \epsilon, \\ 1 & z \ge \epsilon. \end{cases}$$

Following [9], we can now define continuous approximations of $W_{+,x}^T W_{+,y}$:

$$G_{W,-\epsilon}(x,y) = \sum_{i=1}^{n} h_{-\epsilon}(w_i \cdot x) h_{-\epsilon}(w_i \cdot y) w_i w_i^T \text{ and } G_{W,\epsilon}(x,y) = \sum_{i=1}^{n} h_{\epsilon}(w_i \cdot x) h_{\epsilon}(w_i \cdot y) w_i w_i^T.$$

Observe that $h_{-\epsilon}$ is an upper approximation to $\mathbb{1}_{z \ge 0}$, and h_{ϵ} is a lower approximation. Thus, it's clear that for all $W \in \mathbb{R}^{n \times k}$ and all nonzero $x, y \in \mathbb{R}^k$ we have the matrix inequality

$$G_{W,\epsilon}(x,y) \preceq W_{+,x}^T W_{+,y} \preceq G_{W,-\epsilon}(x,y). \tag{3}$$

So it suffices to upper bound $G_{W,-\epsilon}(x,y)$ and lower bound $G_{W,\epsilon}(x,y)$. The two arguments are essentially identical, and we will focus on the former. We seek to prove that with high probability over W, for all $x, y \in S^{k-1}$ simultaneously, $G_{W,-\epsilon}(x,y) \preceq nQ_{x,y} + \epsilon n I_k$. At this point, [9] employs a standard ϵ-net argument. This does not suffice for our purposes, because it uses the following bounds:

1. For fixed $x, y, u \in S^{k-1}$, the inequality $\frac{1}{n} u^T G_{W,-\epsilon}(x,y) u \le u^T Q_{x,y} u + \epsilon$ holds with probability $1 - e^{-\Theta(n)}$
2. $\frac{1}{n} u^T G_{W,-\epsilon}(x,y) u^T$ is $\Theta(\sqrt{k})$-Lipschitz.

The second bound means that the ϵ-net needs to have granularity $O(1/\sqrt{k})$, so we must union bound over $\sqrt{k}^{O(k)} = e^{O(k \log k)}$ triples $x, y, u \in S^{k-1}$. Thus, a high probability bound requires $n \ge \Omega(k \log k)$. Moreover, both bounds (1) and (2) are asymptotically optimal. So a different approach is needed.

This is where Theorem 4.4 comes in. Let $f_W(x,y,u) = u^T G_{W,-\epsilon}(x,y) u$. If we center a ball of small constant radius at some point (x,y,u), then for any W there is some point in the ball where f_W differs by $\Theta(\sqrt{k})$. But for each W, at most points f_W only differs by $\Theta(1)$. More formally, it can be shown that f_W is $(\epsilon, c\epsilon^2, 1/2)$-pseudo-Lipschitz (so its "effective Lipschitz parameter" has no dependence on k). The desired concentration is then a corollary of Theorem 4.4.

5.2 Full proof

In this section we prove Theorem 3.2. Fix $\epsilon > 0$. Let $W \in \mathbb{R}^{n \times k}$ have independent rows $w_1, \ldots, w_n \sim N(0,1)^k$. Recall from Section 5 that we have the matrix inequality

$$G_{W,\epsilon}(x,y) \preceq W_{+,x}^T W_{+,y} \preceq G_{W,-\epsilon}(x,y). \tag{4}$$

So it suffices to upper bound $G_{W,-\epsilon}(x,y)$ and lower bound $G_{W,\epsilon}(x,y)$. The two arguments are essentially identical, and we will focus on the former. We seek to prove that with high probability over W, for all $x, y \in S^{k-1}$ simultaneously,

$$G_{W,-\epsilon}(x,y) \preceq nQ_{x,y} + \epsilon n I_k.$$

Moreover we want this to hold whenever $n \ge Ck$ for some constant $C = C(\epsilon)$. We'll use two standard concentration inequalities:

Lemma 5.1. *Suppose that $k \le n$. Then*

1. $\|W\| \le 3\sqrt{n}$ *with probability at least* $1 - e^{-n/2}$.
2. $\max_{i \in [n]} \|w_i\|_2 \le \sqrt{2k}$ *with probability at least* $1 - ne^{-k/8}$.

Proof. See [21] for a reference on the first bound. The second bound is by concentration of chi-squared with k degrees of freedom. □

Let Θ be the set of matrices $M \in \mathbb{R}^{n\times k}$ such that $\|M\| \le 3\sqrt{n}$ and $\max_{i\in[n]} \|M_i\|_2 \le \sqrt{2k}$. Define the random variable $\theta = W|(W \in \Theta)$.

Fix $u \in S^{k-1}$. For any $M \in \Theta$, define

$$f_M(x,y) = \frac{1}{n} u^T G_{M,-\epsilon}(x,y) u$$

and define $g(x,y) = u^T Q_{x,y} u$. We check that f and g satisfy the three conditions of Theorem 4.4 with appropriate parameters.

Lemma 5.2. *For any $x, y \in \mathbb{R}^k$ with $\|x\|_2, \|y\|_2 \ge 1/2$,*

$$\Pr[f_\theta(x,y) \le g(x,y) + 3\epsilon] \ge 1 - 4\exp(-Cn\epsilon^2).$$

Proof. We first consider $f_W(x,y)$. It is shown in the proof of Lemma 12 in [9] that

$$\mathbb{E}[G_{W,-\epsilon}(x,y)] \preceq Q_{x,y} + \left(\frac{\epsilon}{2\|x\|_2} + \frac{\epsilon}{2\|y\|_2}\right) I,$$

which under our assumptions implies that $\mathbb{E}[f_W(x,y)] \le g(x,y) + 2\epsilon$. Now expand

$$f_W(x,y) = \frac{1}{n}\sum_{i=1}^n h_{-\epsilon}(w_i x) h_{-\epsilon}(w_i y)(w_i u)^2.$$

Let $z_i = h_{-\epsilon}(w_i x) h_{-\epsilon}(w_i y)(w_i u)^2$. Since $h_{-\epsilon}(w_i x) h_{-\epsilon}(w_i y)$ is bounded and $w_i u$ is Gaussian, and $\mathbb{E}z_i$ is bounded by a constant, it follows that $z_i - \mathbb{E}z_i$ is subexponential with some constant variance proxy σ. Therefore by Bernstein's inequality, for all $t > 0$,

$$\Pr[f_W(x,y) - \mathbb{E}f_W(x,y) > t] \le 2\exp(-Cn\min(t, t^2))$$

for some constant $C > 0$. Taking $t = \epsilon$, we get that

$$\Pr[f_W(x,y) > g(x,y) + 3\epsilon] \le 2\exp(-Cn\epsilon^2).$$

Finally, since $\Pr[W \in \Theta] \ge 1/2$, it follows that conditioning on Θ at most doubles the failure probability. So

$$\Pr[f_\theta(x,y) > g(x,y) + 3\epsilon] \le 4\exp(-Cn\epsilon^2)$$

as desired. □

Next, we show that $\{f_M\}_{M\in\Theta}$ is $(\epsilon, \delta, \gamma)$-pseudo-Lipschitz where $\delta = \epsilon^2/116$ and $\gamma = 1/2$.

Definition 5.3. For any $M \in \mathbb{R}^{n\times k}$, $t \in \mathbb{R}$, and $u \in \mathbb{R}^k$, let $B_{M,t,u} \subseteq \mathbb{R}^k$ be the set of points $v \in \mathbb{R}^k$ such that

$$\sum_{i=1}^n |M_i v|(M_i u)^2 \le tn.$$

The pseudo-ball $B_{M,t,u}$ captures the directions in which f_M is Lipschitz more effectively than any spherical ball, as the following lemma shows.

Lemma 5.4. *Let $M \in \mathbb{R}^{n\times k}$. Let $x, y, \tilde{x}, \tilde{y} \in \mathbb{R}^k$. If $y - \tilde{y} \in B_{M,\epsilon^2/4,u}$ and $x - \tilde{x} \in \mathcal{B}_{M,\epsilon^2/4,u}$ then*

$$|f_M(x,y) - f_M(\tilde{x},\tilde{y})| \le \epsilon.$$

Proof. We have

$$\begin{aligned}
|f_M(x,y) - f_M(\tilde{x},\tilde{y})| &\le \frac{1}{n}\sum_{i=1}^{n} |h_{-\epsilon}(M_i x)h_{-\epsilon}(M_i y) - h_{-\epsilon}(M_i \tilde{x})h_{-\epsilon}(M_i \tilde{y})|(M_i u)^2 \\
&\le \frac{1}{n}\sum_{i=1}^{n} \big[h_{-\epsilon}(M_i x)|h_{-\epsilon}(M_i y) - h_{-\epsilon}(M_i \tilde{y})| + \\
&\qquad h_{-\epsilon}(M_i \tilde{y})|h_{-\epsilon}(M_i x) - h_{-\epsilon}(M_i \tilde{x})|\big](M_i u)^2 \\
&\le \frac{1}{n}\sum_{i=1}^{n} \left[|h_{-\epsilon}(M_i y) - h_{-\epsilon}(M_i \tilde{y})| + |h_{-\epsilon}(M_i x) - h_{-\epsilon}(M_i \tilde{x})|\right](M_i u)^2 \\
&\le \frac{1}{\epsilon n}\sum_{i=1}^{n} \left[|M_i(y-\tilde{y})| + |M_i(x-\tilde{x})|\right](M_i u)^2 \\
&\le \epsilon
\end{aligned}$$

where the second-to-last inequality uses that $h_{-\epsilon}$ is $1/\epsilon$-Lipschitz, and the last inequality uses the assumptions on $y - \tilde{y}$ and $x - \tilde{x}$. □

Next, we need to lower bound the volume of $B_{M,t,u}$.

Lemma 5.5. *Fix $u \in S^{k-1}$ and $\delta, t > 0$. Then $\{B_{M,t,u}\}_{M\in\theta}$ is (δ,γ)-wide for $\gamma = \delta^k(1 - 41\delta t^{-1})$. Fix $M \in \Theta$ and $u \in S^{k-1}$. For any $\delta, t > 0$,*

$$\mathrm{Vol}(B_{M,t,u} \cap \delta\mathcal{B}) \ge \delta^k(1 - 41\delta t^{-1})\,\mathrm{Vol}(\mathcal{B}).$$

Proof. Fix $M \in \Theta$. It's clear from the definition that $B_{M,t,u}$ is symmetric (i.e. $v \in B_{M,t,u}$ implies $-v \in B_{M,t,u}$) and convex (by the triangle inequality). It remains to show that

$$\mathrm{Vol}(B_{M,t,u} \cap \delta\mathcal{B}) \ge \delta^k(1 - 41\delta t^{-1})\,\mathrm{Vol}(\mathcal{B}).$$

Writing the $\mathrm{Vol}(B_{M,t,u})$ as a probability,

$$\begin{aligned}
\frac{\mathrm{Vol}(B_{M,t,u} \cap \delta\mathcal{B})}{\mathrm{Vol}(\mathcal{B})} &= \delta^k \frac{\mathrm{Vol}(B_{M,t,u} \cap \delta\mathcal{B})}{\mathrm{Vol}(\delta\mathcal{B})} \\
&= \delta^k \Pr_{\eta\in\delta\mathcal{B}}\left[\sum_{i=1}^{n} |M_i\eta|(M_i u)^2 \le tn\right].
\end{aligned}$$

Since increasing $\|\eta\|$ only increases the sum, the probability over $\delta\mathcal{B}$ is at least the probability over the sphere $\{\eta \in \mathbb{R}^k : \|\eta\|_2 = \delta\}$. Then

$$\begin{aligned}
\Pr_{\eta\in\delta\mathcal{B}}\left[\sum_{i=1}^{n} |M_i\eta|(M_i u)^2 \le tn\right] &\ge \Pr_{\|\eta\|_2=\delta}\left[\sum_{i=1}^{n} |M_i\eta|(M_i u)^2 \le tn\right] \\
&= \Pr_{\eta\sim N(0,1)^k}\left[\sum_{i=1}^{n} |M_i\eta|(M_i u)^2 \le tn\,\|\eta\|_2/\delta\right]
\end{aligned}$$

by spherical symmetry of the Gaussian. Now we upper bound the probability of the complementary event

$$\Pr_{\eta\sim N(0,1)^k}\left[\sum_{i=1}^{n} |M_i\eta|(M_i u)^2 > tn\,\|\eta\|_2/\delta\right],$$

with the aim of making the right-hand side of the inequality deterministic:

$$\begin{aligned}
&\Pr_{\eta\sim N(0,1)^k}\left[\sum_{i=1}^n |M_i\eta|(M_iu)^2 > tn\left\|\eta\right\|_2/\delta\right] \\
&= \Pr_{\eta\sim N(0,1)^k}\left[\sum_{i=1}^n |M_i\eta|(M_iu)^2 > tn\left\|\eta\right\|_2/\delta \middle| \left\|\eta\right\|_2 \geq \sqrt{k}/2\right] \\
&= \frac{\Pr_{\eta\sim N(0,1)^k}\left[\sum_{i=1}^n |M_i\eta|(M_iu)^2 > tn\left\|\eta\right\|_2/\delta \wedge \left\|\eta\right\|_2 \geq \sqrt{k}/2\right]}{\Pr[\left\|\eta\right\|_2 \geq \sqrt{k}/2]} \\
&\leq \frac{\Pr_{\eta\sim N(0,1)^k}\left[\sum_{i=1}^n |M_i\eta|(M_iu)^2 > tn\sqrt{k}/(2\delta)\right]}{1-e^{-9k/128}} \\
&\leq 2\Pr_{\eta\sim N(0,1)^k}\left[\sum_{i=1}^n |M_i\eta|(M_iu)^2 > tn\sqrt{k}/(2\delta)\right]
\end{aligned}$$

For each $i \in [n]$, the random variable $M_i\eta$ is distributed as $N(0, \|M_i\|^2)$. Since $M \in \Theta$, we have $\|M_i\| \leq \sqrt{2k}$. Thus, $|M_i\eta|$ has mean at most $\|M_i\|\sqrt{2/\pi} \leq 2\sqrt{k/\pi}$. So by Markov's inequality,

$$\begin{aligned}
\Pr_{\eta\sim N(0,1)^k}\left[\sum_{i=1}^n |M_i\eta|(M_iu)^2 > tn\sqrt{k}/(2\delta)\right] &\leq \frac{\mathbb{E}_{\eta\sim N(0,1)^k}\left[\sum_{i=1}^n |M_i\eta|(M_iu)^2\right]}{tn\sqrt{k}/(2\delta)} \\
&\leq \frac{\sum_{i=1}^n (M_iu)^2 2\sqrt{k/\pi}}{tn\sqrt{k}/(2\delta)} \\
&\leq 36\delta/(t\sqrt{\pi})
\end{aligned}$$

where the last inequality uses the assumption $M \in \Theta$ to get $\|Mu\|^2 \leq 9n\|u\|^2 = 9n$. We conclude that

$$\frac{\mathrm{Vol}(B_{M,t,u} \cap \delta\mathcal{B})}{\mathrm{Vol}(\mathcal{B})} \geq \delta^k(1 - 72\delta/(t\sqrt{\pi}))$$

as desired. □

Taking $t = \epsilon^2$ and $\delta = \epsilon^2/82$ yields

$$\mathrm{Vol}(B_{M,t,u} \cap \delta\mathcal{B}) \geq \frac{1}{2}\mathrm{Vol}(\delta\mathcal{B}).$$

So $\{f_M\}_{M\in\Theta}$ is $(2\epsilon, \epsilon^2/82, 1/2)$-pseudo-Lipschitz. Finally, we have a lemma about the smoothness of g; see Appendix B for the proof.

Lemma 5.6. *Let $x, y, \tilde{x}, \tilde{y} \in (3/2)\mathcal{B} \setminus (1/2)\mathcal{B}$. Let $d = \max(\|x-y\|, \|\tilde{x}-\tilde{y}\|)$. Then*

$$\left\|Q_{x,y} - Q_{\tilde{x},\tilde{y}}\right\| \leq O(\|x-\tilde{x}\|_2 + \|y-\tilde{y}\|_2).$$

Thus, we have $|g(x,y) - g(\tilde{x},\tilde{y})| \leq C\epsilon$ whenever $x, y \in S^{k-1}$ and $\tilde{x}-x, \tilde{y}-y \in \delta\mathcal{B}$. Having shown that the three conditions of Theorem 4.4 are satisfied, we can prove uniform concentration over all $x, y \in S^{k-1}$.

Lemma 5.7. *Let $\epsilon > 0$. Fix $u \in S^{k-1}$. Then with probability at least $1 - (C/\epsilon)^{8k}e^{-cn\epsilon^2}$ (over θ), it holds that for all $x, y \in S^{k-1}$,*

$$f_\theta(x,y) \leq g(x,y) + C\epsilon n.$$

Proof. Apply Theorem 4.4 to the family $\{f_M\}_{M\in\Theta}$ and random variable θ. By Lemma 5.2, we can take $p = \exp(-cn\epsilon^2)$ for a constant $c > 0$. We know that $\{f_{x,y}\}$ is $(2\epsilon, \epsilon^2/82, 1/2)$-pseudo-Lipschitz. By Lemma 5.6, we can take $D = C\epsilon$. The claim follows. □

Now we get a uniform bound over all $u \in S^{k-1}$, via a standard ϵ-net an Lipschitz bound. Adding notation for clarity, define

$$f_M(x, y, u) = \frac{1}{n} u^T G_{M,-\epsilon}(x, y) u.$$

The following lemma shows that f is Lipschitz in u.

Lemma 5.8. *Let $M \in \Theta$. Let $x, y \in \mathbb{R}^k$ and $u, v \in S^{k-1}$. Then*

$$|f_M(x, y, u) - f_M(x, y, v)| \leq 18n \|u - v\|_2$$

Proof. We have

$$\begin{aligned}
|f_M(x, y, u) - f_M(x, y, v)| &\leq \sum_{i=1}^{n} h_{-\epsilon}(M_i x) h_{-\epsilon}(M_i y) \left| (M_i u)^2 - (M_i v)^2 \right| \\
&\leq \sum_{i=1}^{n} |M_i(u - v)| \cdot |M_i(u + v)| \\
&\leq \|M(u - v)\|_2 \|M(u + v)\|_2 \\
&\leq \|M\|^2 \|u - v\|_2 \|u + v\|_2 \\
&\leq 18n \|u - v\|_2
\end{aligned}$$

by an application of Cauchy-Schwarz, and the bound $\|M\| \leq 3\sqrt{n}$. □

The matrix bound on $G_{W,-\epsilon}(x, y)$ follows from applying an ϵ-net together with Lemma 5.8, and finally removing the conditioning on Θ.

Theorem 5.9. *Let $\epsilon > 0$. Then with probability at least $1 - (C/\epsilon)^{8k} e^{-cn\epsilon^2} - ne^{-k/8}$ over W, it holds that for all $x, y \in S^{k-1}$,*

$$\frac{1}{n} G_{W,-\epsilon}(x, y) \preceq Q_{x,y} + \epsilon.$$

Proof. Let $\mathcal{E} \subset S^{k-1}$ be an ϵ-net of size at most $(3/\epsilon)^k$. By a union bound, it holds with probability at least $1 - (3/\epsilon)^k (C/\epsilon)^{8k} e^{-cn\epsilon^2}$ over θ that for all $x, y \in S^{k-1}$ and all $u \in \mathcal{E}$,

$$f_\theta(x, y, u) \leq g(x, y, u) + 2\epsilon.$$

For any $x, y, u \in S^{k-1}$ there is some $v \in \mathcal{E}$ with $\|u - v\|_2 \leq \epsilon$, so by Lemma 5.8,

$$f_\theta(x, y, u) \leq f_\theta(x, y, v) + O(\epsilon) \leq g(x, y, v) + O(\epsilon).$$

But $\|Q_{x,y}\| \leq 1$, so

$$|g(x, y, u) - g(x, y, v)| = |u^T Q_{x,y} u - v^T Q_{x,y} v| \leq 2 \|u - v\|_2 .$$

Thus, $f_\theta(x, y, u) \leq g(x, y, u) + O(\epsilon)$. We conclude that with probability at least $1 - (3/\epsilon)^k (C/\epsilon)^{8k} e^{-cn\epsilon^2}$ over θ we have the desired inequality. But $\Pr[W \in \Theta] \geq 1 - ne^{-k/8} - e^{-n/2}$. So the desired inequality holds with probability at least $1 - (3/\epsilon)^k (C/\epsilon)^{8k} e^{-cn\epsilon^2} - ne^{-k/8} - e^{-n/2}$ over W. □

We turn to the lower bound on $G_{W,\epsilon}(x, y)$. The proof is essentially identical. For fixed x, y there is an analogous bound to Lemma 5.2:

Lemma 5.10. *[9] There are constants c_K, γ_K such that if $n > c_K \epsilon^{-2} k$ then for any fixed $x, y \in S^{k-1}$ we have that*

$$G_\epsilon(x, y) \succeq nQ_{x,y} - 2\epsilon n I_k$$

with probability at least $1 - 2e^{-\gamma_K \epsilon^2 n}$.

The same pseudo-Lipschitz bounds hold for $u^T G_{M,\epsilon}(x, y) u$ as we proved for $u^T G_{M,-\epsilon} u$, and the remaining argument is unchanged. Thus, we have the following theorem.

Theorem 5.11. *There is a constant c such that if $n > ck\epsilon^{-2} \log(\epsilon^{-2})$ then with high probability over W, for all $x, y \in S^{k-1}$,*

$$G_{W,\epsilon}(x, y) \succeq nQ_{x,y} - \epsilon n.$$

Our main result immediately follows.

Proof of Theorem 3.2. Recall that for all $x, y \in S^{k-1}$,

$$G_{W,\epsilon}(x,y) \preceq W_{+,x}^T W_{+,y} \preceq G_{W,-\epsilon}(x,y)$$

It follows from Theorems 5.9 and 5.11 that with high probability over W, for all $x, y \in S^{k-1}$,

$$nQ_{x,y} - \epsilon n \preceq W_{+,x}^T W_{+,y} \preceq nQ_{x,y} + \epsilon n.$$

Since $Q_{x,y}$ and $W_{+,x}^T W_{+,y}$ are both invariant under scaling by a positive constant, this inequality then holds for all nonzero $x, y \in \mathbb{R}^k$. So W satisfies the normalized WDC with parameter ϵ. □

6 Proof of Main Result: Theorem 1.1

In this section, we prove Theorem 1.1. We apply Theorem 3.2 together with a theorem from [11]. To provide the precise statement we need to introduce several more definitions. First, we introduce the *unnormalized* WDC (as defined in [9]):

Definition 6.1. A matrix $W \in \mathbb{R}^{n \times k}$ is said to satisfy the *unnormalized Weight Distribution Condition (WDC)* with parameter ϵ if for all nonzero $x, y \in \mathbb{R}^k$ it holds that

$$\left\| W_{+,x}^T W_{+,y} - Q_{x,y} \right\| \leq \epsilon$$

where $Q_{x,y} = \mathbb{E} W_{+,x}^T W_{+,y}$ (with expectation over i.i.d. $N(0,1)$ entries of W).

Comparing with Definition 3.1, it's clear that W satisfies the unnormalized WDC if and only if $\sqrt{n}W$ satisfies the normalized WDC. In this paper we proved results about the normalized WDC for ease of notation, but for compressed sensing we are concerned that our weight matrices satisfy the unnormalized WDC.

Definition 6.2 ([9])**.** A matrix $A \in \mathbb{R}^{m \times n}$ satisfies the *Range Restricted Isometry Condition* (RRIC) with respect to G and with parameter ϵ if for all $x, y, z, w \in \mathbb{R}^k$, it holds that

$$\begin{aligned} &|(G(x) - G(y))^T A^T A (G(z) - G(w)) - (G(x) - G(y))^T (G(z) - G(w))| \\ &\leq \epsilon \left\| G(x) - G(y) \right\|_2 \left\| G(z) - G(w) \right\|_2 . \end{aligned}$$

Definition 6.3 ([11])**.** The *empirical risk function* is $f : \mathbb{R}^k \to \mathbb{R}$ defined by

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

Let ALGO refer to Algorithm 1 in [11]. This algorithm is a modification of gradient descent on the empirical risk, which given an initial point x_0 computes iterates $x_1, x_2, \ldots$. The result of [11] is a bound on the convergence rate of ALGO, assuming that the (unnormalized) WDC and RRIC are satisfied. We restate it below. We treat network depth d as a constant for simplicity of notation; we denote all upper-bound constants by C.

Theorem 6.4 ([11])**.** *Suppose that $W^{(1)}, \ldots, W^{(d)}$ satisfy the (unnormalized) WDC, and A satisfies the RRIC with respect to G, each with parameter $\epsilon \leq C$. Suppose that the noise e satisfies $\|e\| \leq C \|x^*\|$. Let x_0 be the initial point of* ALGO*. Then after $N \leq Cf(x_0)/\|x^*\| +$ iterations, the iterate x_N satisfies*

$$\left\| x_N - x^* \right\|_2 \leq C(\|x^*\| \sqrt{\epsilon} + \|e\|)$$

Additionally, for any $i \geq N$,

$$\|x_{i+1} - x^*\| \leq \gamma^{i+1-N} \|x_N - x^*\| + C \|e\|$$

where $\gamma \in (0,1)$ is a constant.

So let $\epsilon \in (0, C)$. We only need to show that the (unnormalized) WDC and RRIC are satisfied under the conditions of Theorem 1.1. That is, we have the following assumptions (recall that $k = n_0$ and $n = n_d$) for constants K_2, K_3:

(a) For $i \in [d]$, the weight matrix $W^{(i)}$ has dimension $n_i \times n_{i-1}$ with

$$n_i \geq K_2 n_{i-1} \epsilon^{-2} \log(1/\epsilon^2).$$

Additionally, the entries are drawn i.i.d. from $N(0, 1/n_i)$.

(b) The measurement matrix A has dimension $m \times n$ with $m \geq K_3 \epsilon^{-1} \log(1/\epsilon) dk \log \prod_{i=1}^{d} n_i$. Additionally, the entries are drawn i.i.d. from $N(0, 1/m)$.

By Theorem 3.2, with high probability, the weight matrices $\sqrt{n_1}W^{(1)}, \ldots, \sqrt{n_d}W^{(d)}$ satisfy the normalized WDC, so $W^{(1)}, \ldots, W^{(d)}$ satisfy the unnormalized WDC. To show that A satisfies the RRIC with respect to G, we appeal to Lemma 17 in [9], which proves this exact fact (without any assumptions about the expansion of G).

Thus, Theorem 6.4 is applicable. This completes the proof of Theorem 1.1.

Acknowledgments

We thank Paul Hand and Alex Dimakis for useful discussions. This research was supported by NSF Awards IIS-1741137, CCF-1617730 and CCF-1901292, by a Simons Investigator Award, by the DOE PhILMs project (No. DE-AC05-76RL01830), by the DARPA award HR00111990021, by the MIT Undergraduate Research Opportunities Program, and by a Google PhD Fellowship.

References

[1] Holger Boche, Robert Calderbank, Gitta Kutyniok, and Jan Vybíral. A survey of compressed sensing. In *Compressed sensing and its applications*, pages 1–39. Springer, 2015.

[2] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 537–546. JMLR. org, 2017.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[4] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

[5] Jorio Cocola, Paul Hand, and Vladislav Voroninski. Nonasymptotic guarantees for low-rank matrix recovery with generative priors. *arXiv preprint arXiv:2006.07953*, 2020.

[6] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[8] Paul Hand, Oscar Leong, and Vlad Voroninski. Phase retrieval under a generative prior. In *Advances in Neural Information Processing Systems*, pages 9136–9146, 2018.

[9] Paul Hand and Vladislav Voroninski. Global guarantees for enforcing deep generative priors by empirical risk. *IEEE Transactions on Information Theory*, 66(1):401–418, 2019.

[10] Reinhard Heckel, Wen Huang, Paul Hand, and Vladislav Voroninski. Deep denoising: Rate-optimal recovery of structured signals with a deep prior. 2018.

[11] Wen Huang, Paul Hand, Reinhard Heckel, and Vladislav Voroninski. A provably convergent scheme for compressive sensing under random generative priors. *arXiv preprint arXiv:1812.04176*, 2018.

[12] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[13] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10215–10224, 2018.

[14] Qi Lei, Ajil Jalal, Inderjit S Dhillon, and Alexandros G Dimakis. Inverting deep generative models, one layer at a time. In *Advances in Neural Information Processing Systems*, pages 13910–13919, 2019.

[15] Fangchang Ma, Ulas Ayaz, and Sertac Karaman. Invertibility of convolutional generative networks from partial measurements. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9628–9637. Curran Associates, Inc., 2018.

[16] Gregory Ongie, Ajil Jalal, Christopher A Metzler, Richard G Baraniuk, Alexandros G Dimakis, and Rebecca Willett. Deep learning techniques for inverse problems in imaging. *arXiv preprint arXiv:2005.06001*, 2020.

[17] Shuang Qiu, Xiaohan Wei, and Zhuoran Yang. Robust one-bit recovery via relu generative networks: Improved statistical rates and global landscape analysis. *arXiv preprint arXiv:1908.05368*, 2019.

[18] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018.

[19] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018.

[20] Dave Van Veen, Ajil Jalal, Mahdi Soltanolkotabi, Eric Price, Sriram Vishwanath, and Alexandros G Dimakis. Compressed sensing with deep image prior and learned regularization. *arXiv preprint arXiv:1806.06438*, 2018.

[21] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

A Lower bound: necessity of non-trivial expansivity

Our main result implied that constant expansion is sufficient to efficiently recover a latent parameter, even in the presence of compressive measurements and noise. In this section we prove a lower bound: that a factor-of-2 expansion is necessary, even in the absence of compressive measurements and noise. That is, inverting a neural network with expansion of less than 2 is impossible. To prove this lower bound it suffices to consider a one-layer neural network

$$G(x) = \mathsf{ReLU}(Wx),$$

where $x \in \mathbb{R}^k$ and $W \in \mathbb{R}^{m \times k}$. We show that if $m \leq 2k - 1$ then for any weight matrix W, there is some signal which corresponds to multiple latent vectors. Moreover if the rows of W have bounded ℓ_2 norm, then even approximate recovery is impossible.

Note that this bound is tight, since if the rows of W consist of the $2k$ signed basis vectors $\{\pm e_1, \ldots, \pm e_k\}$ then G is injective.

Proposition A.1. *Suppose that $m \leq 2k-1$ and $\max_{i\in[m]} \|W_i\|_2 \leq B$. Then there are $x, y \in \mathbb{R}^k$ with $G(x) = G(y)$ but $\|x-y\|_2 \geq 1/B$.*

Proof. If $\text{rank}(W) < k$, then there are distinct $x, y \in \mathbb{R}^k$ with $Wx = Wy$, so certainly $G(x) = G(y)$. Otherwise, the rows of W contain a basis for $\mathbb{R}^k$. Without loss of generality, assume that $W_1, \ldots, W_k$ span $\mathbb{R}^k$. Define $S = \{1, \ldots, k\}$ and $T = \{k+1, \ldots, m\}$. Then the square submatrix W_S is invertible. Define

$$x = (W_S)^{-1} \begin{bmatrix} -1 \\ -1 \\ \vdots \\ -1 \end{bmatrix}.$$

Since W_T has at most $k-1$ rows, it has non-trivial null space. Let $v \in \mathbb{R}^k$ be a unit vector with $W_T v = 0$. Let $\lambda = \|W_S v\|_\infty$. Since W_S is invertible it's clear that $\lambda > 0$. On the other hand $\lambda \leq \sup_{i\in[m]} |W_i v| \leq B$. Define

$$y = x + v/\lambda.$$

Certainly $\|x-y\|_2 = \lambda^{-1} \geq 1/B$. For any $i \in T$, it's clear that $W_i x = W_i y$. Moreover, fix any $i \in S$. Then $W_i x = -1$, and

$$W_i y = W_i x + W_i v/\lambda \leq -1 + \|W_S v\|_\infty / \lambda \leq 0.$$

Thus, $\mathsf{ReLU}(W_i x) = \mathsf{ReLU}(W_i y)$ for all $i \in [m]$. So $G(x) = G(y)$. □

B Lipschitz bound for $Q_{x,y}$

In this appendix we show that $Q_{x,y} = \frac{1}{n}\mathbb{E} W_{+,x}^T W_{+,y}$ is Lipschitz (under the operator norm) as a function of x and y.

For $x \in \mathbb{R}^k$ nonzero, let $\hat{x}$ refer to the unit vector $x/\|x\|_2$. For $x, y \in \mathbb{R}^k$ nonzero, define $M_{x,y} \in \mathbb{R}^{k\times k}$ to be the matrix such that

$$M_{x,y}(a\hat{x} + b\hat{y} + z) = a\hat{y} + b\hat{x}$$

for any $a, b \in \mathbb{R}$ and $z \in \mathbb{R}^k$ with $z \perp x$ and $z \perp y$. Also let $\angle(x,y)$ denote the angle between vectors x and y (always between 0 and π).

It can be calculated that

$$Q_{x,y} = \frac{\pi - \angle(x,y)}{2\pi} I_k + \frac{\sin\angle(x,y)}{2\pi} M_{x,y}.$$

This is the expression we'll manipulate. It can be easily shown that $\angle(x,y)$ is Lipschitz, so the first term is Lipschitz. The remaining difficulty is then in showing that $(\sin\theta)M_{x,y}$ is Lipschitz. We start by proving this under some weak assumptions, i.e. essentially that $x, y, \tilde{y}$ are "in general position".

Lemma B.1. *Let $x, y, \tilde{y} \in S^{k-1}$. Let $\theta = \angle(x,y)$ and $\phi = \angle(x,\tilde{y})$. Assume that $\theta, \phi \notin \{0, \pi\}$. Then*

$$\left\|(\sin\theta)M_{x,y} - (\sin\phi)M_{x,\tilde{y}}\right\| \leq \sqrt{79}\,\|y - \tilde{y}\|_2.$$

Proof. Fix an orthonormal basis for $\text{span}\{x, y, \tilde{y}\}$ in which

$$x = (1, 0, 0)$$

$$y = (\cos\theta, \sin\theta, 0)$$

$$\tilde{y} = (\cos\phi, \sin\phi\cos\alpha, \sin\phi\sin\alpha.$$

Let $d = \|y - \tilde{y}\|_2$ and observe that $\sin^2\alpha \leq d^2/\sin^2\phi$. Moreover if $\beta = \angle(y,\tilde{y})$ then $\beta \leq 2\sqrt{2 - 2\cos\beta} = 2d$, so $|\theta - \phi| \leq \beta \leq 2d$. Since sin and cos are 1-Lipschitz it follows that $|\sin\theta - \sin\phi| \leq 2d$ and $|\cos\theta - \cos\phi| \leq 2d$.

Let $u \in S^{k-1}$. It can be checked that in this basis,

$$M_{x,y}u = (u_1 \cos\theta + u_2 \sin\theta, u_1 \sin\theta - u_2 \cos\theta, 0)$$

and

$$\begin{aligned} M_{x,\tilde{y}}u = (&u_1 \cos\phi + u_2 \sin\phi \cos\alpha + u_3 \sin\phi \sin\alpha, \\ &u_1 \sin\phi \cos\alpha - u_2 \cos\phi \cos^2\alpha - u_3 \cos\phi \sin\alpha \cos\alpha, \\ &u_1 \sin\phi \sin\alpha - u_2 \cos\phi \sin\alpha \cos\alpha - u_3 \cos\phi \sin^2\alpha). \end{aligned}$$

Indeed, we see that $M_{x,y}u \in \operatorname{span}\{x, y\}$ and $(M_{x,y}u) \cdot x = u \cdot y$ and $(M_{x,y}u) \cdot y = u \cdot x$ (and similarly for $M_{x,\tilde{y}}u$). But now

$$\begin{aligned} &((\sin\theta)M_{x,y}u - (\sin\phi)M_{x,\tilde{y}}u)_1^2 \\ &= (u_1(\sin\theta\cos\theta - \sin\phi\cos\phi) + u_2(\sin^2\theta - \sin^2\phi\cos\alpha) - u_3\sin^2\phi\sin\alpha) \\ &\leq (\sin\theta\cos\theta - \sin\phi\cos\phi)^2 + (\sin^2\theta - \sin^2\phi\cos\alpha)^2 - \sin^4\phi\sin^2\alpha \\ &\leq (|\sin\theta - \cos\theta| + |\sin\phi - \cos\phi|)^2 + (|\sin^2\theta - \sin^2\phi| + |\sin^2\phi(1 - \cos^2\alpha)|)^2 \\ &\quad + \sin^4\phi\sin^2\alpha \\ &\leq 33d^2. \end{aligned}$$

Similarly,

$$\begin{aligned} &((\sin\theta)M_{x,y}u - (\sin\phi)M_{x,\tilde{y}}u)_2^2 \\ &\leq (\sin^2\theta - \sin^2\phi\cos\alpha)^2 + (\sin\theta\cos\theta - \sin\phi\cos\phi\cos^2\alpha)^2 + \sin^2\phi\cos^2\phi\sin^2\alpha\cos^2\alpha \\ &\leq 16d^2 + (|\sin\theta\cos\theta - \sin\phi\cos\phi| + |\sin\phi\cos\phi\sin\alpha|)^2 + d^2 \\ &\leq 42d^2 \end{aligned}$$

and

$$\begin{aligned} &((\sin\theta)M_{x,y}u - (\sin\phi)M_{x,\tilde{y}}u)_3^2 \\ &\leq \sin^4\phi\sin^2\alpha + \sin^2\phi\cos^2\phi\sin^4\alpha + \sin^2\phi\cos^2\phi\sin^2\alpha\cos^2\alpha \\ &\leq 3d^2. \end{aligned}$$

Therefore

$$\left\|(\sin\theta)M_{x,y}u - (\sin\phi)M_{x,\tilde{y}}u\right\|_2 \leq d\sqrt{79}.$$

Since this holds for all unit vectors u, the lemma follows. □

It immediately follows that beyond relating $M_{x,y}$ to $M_{x,\tilde{y}}$ we can relate $M_{x,y}$ to $M_{\tilde{x},\tilde{y}}$.

Corollary B.2. *Let $x, y, \tilde{x}, \tilde{y} \in S^{k-1}$. Let $\theta = \angle(x, y)$ and $\tilde{\theta} = \angle(\tilde{x}, \tilde{y})$. Assume that $\theta, \tilde{\theta} \notin \{0, \pi\}$. Then*

$$\left\|(\sin\theta)M_{x,y} - (\sin\tilde{\theta})M_{x,\tilde{y}}\right\| \leq \sqrt{79}(\|x - \tilde{x}\|_2 + \|y - \tilde{y}\|_2).$$

Proof. Either $\angle(x, \tilde{y}) \neq 0$ or $\angle(y, \tilde{x}) \neq 0$, since otherwise $M_{x,y} = M_{\tilde{x},\tilde{y}}$. Without loss of generality suppose $\angle(x, \tilde{y}) \neq 0$. Then by the previous lemma, we can relate $M_{x,y}$ to $M_{x,\tilde{y}}$. And we can relate $M_{x,\tilde{y}}$ to $M_{\tilde{x},\tilde{y}}$. We get the claimed bound. □

From this result the full claim is now easily derived.

Lemma B.3. *Let $x, y, \tilde{x}, \tilde{y} \in S^{k-1}$. Let $d = \max(\|x - \tilde{x}\|_2, \|y - \tilde{y}\|_2)$. Let $\theta = \angle(x, y)$ and $\tilde{\theta} = \angle(\tilde{x}, \tilde{y})$. Then*

$$\left\|(\sin\theta)M_{x,y} - (\sin\tilde{\theta})M_{\tilde{x},\tilde{y}}\right\| \leq 2\sqrt{79}d.$$

Proof. If $\sin\theta = \sin\tilde{\theta} = 0$ then the claim is clear. If both are nonzero it follows from the previous lemma. So now suppose without loss of generality that $\sin\tilde{\theta} = 0$ but $\sin\theta > 0$. We've shown in the previous lemma that $|\sin\theta - \sin\tilde{\theta}| \le 4d$. Additionally, $\|M_{x,y}\| \le 1$. Therefore

$$\|(\sin\theta)M_{x,y} - (\sin\tilde{\theta})M_{\tilde{x},\tilde{y}}\| = (\sin\theta)\|M_{x,y}\| \le 4d.$$

This bound is sufficient. □

Lemma B.4. *Let $x, y, \tilde{x}, \tilde{y} \in (3/2)\mathcal{B} \setminus (1/2)\mathcal{B}$. Let $d = \max(\|x-y\|, \|\tilde{x}-\tilde{y}\|)$. Then*

$$\|Q_{x,y} - Q_{\tilde{x},\tilde{y}}\| \le O(\|x-\tilde{x}\|_2 + \|y-\tilde{y}\|_2).$$

Proof. Let $\hat{x} = x/\|x\|$ and $\hat{y} = y/\|y\|$. Similarly define $\hat{\tilde{x}}$ and $\hat{\tilde{y}}$. Note that normalizing at most doubles the distance. Therefore

$$\begin{aligned}\|(\sin\theta)M_{x,y} - (\sin\tilde{\theta})M_{\tilde{x},\tilde{y}}\| &= \left\|(\sin\theta)M_{\hat{x},\hat{y}} - (\sin\tilde{\theta})M_{\hat{\tilde{x}},\hat{\tilde{y}}}\right\| \\ &\le 2\sqrt{79}(2d)\end{aligned}$$

Additionally, let θ and $\tilde{\theta}$ be as defined previously. We have $|\theta - \tilde{\theta}| \le 4(2d) = 8d$. Hence,

$$\begin{aligned}2\pi\|Q_{x,y} - Q_{\tilde{x},\tilde{y}}\| &= \|(\tilde{\theta}-\theta)I_k + (\sin\theta)M_{x,y} - (\sin\tilde{\theta})M_{\tilde{x},\tilde{y}}\| \\ &\le 8d + 4\sqrt{79}d.\end{aligned}$$

Simplifying,

$$\|Q_{x,y} - Q_{\tilde{x},\tilde{y}}\| \le 7d$$

as desired. □

C Global landscape analysis

In this section, we explain why the Weight Distribution Condition arises, by sketching the basic theory of compressed sensing with generative priors. While our work applies to a number of different models in compressed sensing with generative priors (see Section 3.1 for details), we limit the exposition in this section to the *global landscape analysis* of the vanilla model, due to [9].

Let $G : \mathbb{R}^k \to \mathbb{R}^n$ be a fully connected ReLU neural network of the form

$$G(x) = \mathsf{ReLU}(W^{(d)}(\dots\mathsf{ReLU}(W^{(2)}(\mathsf{ReLU}(W^{(1)}x)))\dots)).$$

Let $x^* \in \mathbb{R}^k$ be an unknown latent vector. We wish to recover x^* (or $G(x^*)$) from $m \ll n$ noisy linear measurements of $G(x^*)$. Specifically, for some measurement matrix $A \in \mathbb{R}^{m\times n}$ and noise vector $e \in \mathbb{R}^m$ we observe

$$y = AG(x^*) + e.$$

The scenario of interest is that the number of measurements m is much less than the output dimension n, but is slightly more than the latent dimension k. Each $W^{(i)}$ is a matrix with dimension $n_i \times n_{i-1}$, such that $n_0 = k$ and $n_d = n$. The noise is arbitrary, and recovery bounds will depend on $\|e\|$.

The aim of [9] is to show that the empirical squared-loss

$$f(x) = \frac{1}{2}\|AG(x) - y\|_2^2$$

has no critical points except the true solution x^*, and a rescaled vector $-\rho_d x^*$. Thus, gradient descent would recover x^* up to a global rescaling. This result is strengthened in subsequent work [11], but it contains the main ideas, which we outline now. See Section 2 of [9] for more details.

Ignoring issues of non-differentiability, the gradient is

$$\nabla f(x) = \left(\prod_{i=d}^{1} W^{(i)}_{+,x^{(i)}}\right)^T A^T A \left(\left(\prod_{i=d}^{1} W^{(i)}_{+,x^{(i)}}\right) x - \left(\prod_{i=d}^{1} W^{(i)}_{+,(x^*)^{(i)}}\right) x^*\right)$$

where $x^{(i)} = W^{(i-1)} \cdots W^{(1)} x$ and $(x^*)^{(i)} = W^{(i-1)} \cdots W^{(1)} x^*$. Next, if A satisfies a certain restricted isometry condition, it follows that $A^T A$ is approximately the identity on the range of G, so

$$\nabla f(x) \approx \left(\prod_{i=d}^{1} W^{(i)}_{+,x^{(i)}}\right)^T \left(\left(\prod_{i=d}^{1} W^{(i)}_{+,x^{(i)}}\right) x - \left(\prod_{i=d}^{1} W^{(i)}_{+,(x^*)^{(i)}}\right) x^*\right).$$

Next, we need to show that this approximation concentrates around its mean. Consider the second term in the difference (the first is no more complicated):

$$(W^{(1)}_{+,x^{(1)}})^T \cdots (W^{(d)}_{+,x^{(d)}})^T W^{(d)}_{+,(x^*)^{(d)}} \cdots W^{(1)}_{+,(x^*)^{(1)}} x.$$

The proof that this product concentrates is by induction on d. Each step collapses the innermost pair $(W^{(i)}_{+,x})^T W^{(i)}_{+,y}$ in the product, using the Weight Distribution Condition (which bounds $(W^{(i)}_{+,x})^T W^{(i)}_{+,y} - \mathbb{E}(W^{(i)}_{+,x})^T W^{(i)}_{+,y}$) to replace the pair with their expectation.

Finally, once the approximation for $\nabla f(x)$ has been further approximated by its (deterministic) expectation, the expectation is analyzed algebraically.

D Extensions

D.1 Gaussian noise

The CS-DGP problem (Compressed Sensing with a Deep Generative Prior) can be modified to the Gaussian noise setting, i.e. the noise vector $e \in \mathbb{R}^m$ has distribution $e \sim N(0, \sigma^2)^m$. In this setting it has been shown that there is an efficient algorithm estimating x^* up to $\tilde{O}(\sigma^2 k/m)$ (ignoring logarithmic terms and dependence of the depth of the network), so long as the measurement matrix A is random and the weight matrices satisfy the Weight Distribution Condition [10]. A corollary was that if the neural network was logarithmically expansive and had Gaussian random weights, then efficient recovery was possible. Our result directly yields an improvement, implying that constant expansion suffices.

D.2 One-bit recovery

One-bit recovery with a neural network prior, introduced in [17], has the following formal statement. Let $G : \mathbb{R}^k \to \mathbb{R}^n$ be a neural network of the form

$$G(x) = \mathsf{ReLU}(W^{(d)}(\ldots \mathsf{ReLU}(W^{(2)}(\mathsf{ReLU}(W^{(1)}x)))\ldots)).$$

Let $x^* \in \mathbb{R}^k$ be an unknown latent vector. We wish to recover x^* from m one-bit noisy measurements of $G(x^*)$. Specifically, we observe a sign vector

$$y = \text{sign}(AG(x^*) + \xi + \tau)$$

where $A \in \mathbb{R}^{m \times n}$ is a random measurement matrix, $\xi \in \mathbb{R}^m$ is a noise vector, and $\tau \in \mathbb{R}^k$ is a random quantization threshold. In this setting, global landscape analysis of the loss function can be performed, and it can be shown that if each weight matrix satisfies the WDC then there are no spurious critical points outside a neighborhood of x^*, a neighborhood of some negative scalar multiple of x^*, and a neighborhood of 0 [17].

Once again, our result implies that the analysis can go through when each weight matrix has i.i.d. Gaussian entries and constant expansion in dimension (whereas previously logarithmic expansion was required).

D.3 Phase retrieval

Phase retrieval with a neural network prior, introduced in [8], has the following formal statement. Let $G : \mathbb{R}^k \to \mathbb{R}^n$ be a neural network of the form

$$G(x) = \mathsf{ReLU}(W^{(d)}(\dots \mathsf{ReLU}(W^{(2)}(\mathsf{ReLU}(W^{(1)}x)))\dots)).$$

Let $x^* \in \mathbb{R}^k$ be an unknown latent vector. We wish to recover x^* from m phaseless noisy measurements of $G(x^*)$. Specifically, we observe

$$y = |AG(x^*)|$$

where $A \in \mathbb{R}^{m \times n}$ is a measurement matrix. As in the prior two examples, global landscape analysis can be performed if each weight matrix satisfies the WDC (and A satisfies a certain isometry condition) [8]. Thus, our contribution extends the analysis to Gaussian matrices with constant expansion.

D.4 Deconvolutional neural networks

It is shown in [15] that if G is a two-layer deconvolutional neural network with Gaussian weights and logarithmic expansion in the number of channels, then (under certain other moderate conditions), the empirical risk function is well-behaved. This result again applies the WDC for Gaussian random matrices as a black box, so our result decreases the requisite expansion to a constant.

D.5 Low-rank matrix recovery

Low-rank matrix recovery with a neural network prior was recently introduced in [5]. Let $G : \mathbb{R}^k \to \mathbb{R}^n$ be a neural network of the form

$$G(x) = \mathsf{ReLU}(W^{(d)}(\dots \mathsf{ReLU}(W^{(2)}(\mathsf{ReLU}(W^{(1)}x)))\dots)).$$

Let $x^* \in \mathbb{R}^k$ be an unknown latent vector. In the spiked Wishart model, we are given a matrix

$$Y = uG(x^*)^T + \sigma\mathcal{Z},$$

where $u \sim N(0, I_N)$ and $\mathcal{Z}$ has i.i.d. $N(0,1)$ entries; that is, Y is an $N \times n$ rank-1 matrix perturbed with random noise. In the spiked Wigner model, instead we are given $Y = G(x^*)G(x^*)^T + \sigma\mathcal{H}$ where $\mathcal{H}$ is an $n \times n$ matrix drawn from a Gaussian Orthogonal Ensemble. In either model, we want to recover x^*.

It's shown in [5] that for both models, if each weight matrix of G satisfies the WDC (and the latent dimension is sufficiently small), then the appropriate loss functions enjoy favorable optimization landscapes. Our result improves the requisite expansion for Gaussian random neural networks to a constant.