Subspace Embeddings Under Nonlinear Transformations

Aarshvi Gajjar[1] and Cameron Musco[2]

[1]UMass Amherst, agajjar@umass.edu
[2]UMass Amherst, cmusco@cs.umass.edu

Abstract

We consider low-distortion embeddings for subspaces under *entrywise nonlinear transformations*. In particular we seek embeddings that preserve the norm of all vectors in a space $S = \{y : y = f(x) \text{ for } x \in Z\}$, where Z is a k-dimensional subspace of $\mathbb{R}^n$ and $f(x)$ is a nonlinear activation function applied entrywise to x. When f is the identity, and so S is just a k-dimensional subspace, it is known that, with high probability, a random embedding into $O(k/\epsilon^2)$ dimensions preserves the norm of all $y \in S$ up to $(1 \pm \epsilon)$ relative error. Such embeddings are known as *subspace embeddings*, and have found widespread use in compressed sensing and approximation algorithms.

We give the first low-distortion embeddings for a wide class of nonlinear functions f. In particular, we give additive ϵ error embeddings into $O(\frac{k \log(n/\epsilon)}{\epsilon^2})$ dimensions for a class of nonlinearities that includes the popular Sigmoid SoftPlus, and Gaussian functions. We strengthen this result to give relative error embeddings under some further restrictions, which are satisfied e.g., by the Tanh, SoftSign, Exponential Linear Unit, and many other 'soft' step functions and rectifying units.

Understanding embeddings for subspaces under nonlinear transformations is a key step towards extending random sketching and compressing sensing techniques for linear problems to nonlinear ones. We discuss example applications of our results to improved bounds for compressed sensing via generative neural networks.

1 Introduction

Random sketching and dimensionality reduction methods are an increasingly important tool in working with massive and high-dimensional datasets [3, 29, 30]. These methods attempt to very quickly compress data points into a lower-dimensional space, while still preserving important information about their structure, from which a downstream task (e.g., clustering, regression, PCA) can be solved approximately.

1.1 Low-Distortion Embeddings

Many such approaches are based around the idea of *low-distortion embeddings*, dimension reducing maps which preserve the norm of all vectors in some set.

Definition 1 (Low-Distortion Embedding)**.** A linear map $\Pi : \mathbb{R}^n \to \mathbb{R}^m$ is an (ϵ_1, ϵ_2)-error embedding for $S \subseteq \mathbb{R}^n$ if, for all $y \in S$:

$$(1 - \epsilon_1)\|y\|_2 - \epsilon_2 \leq \|\Pi y\|_2 \leq (1 + \epsilon_1)\|y\|_2 + \epsilon_2,$$

where $\|\cdot\|_2$ is the Euclidean norm. When $\epsilon_2 = 0$, we say that Π is an ϵ_1-relative-error embedding.

When the set S is just a k-dimensional linear subspace of $\mathbb{R}^n$, it is well known that letting $\Pi \in \mathbb{R}^{m \times n}$ be a random map (e.g., an appropriately scaled matrix with i.i.d. sub-Gaussian entries) with $m = O\left(\frac{k}{\epsilon^2}\right)$ will result in Π being an ϵ-relative error embedding for S with high probability. Such an embedding is known as an *oblivious subspace embedding* (OSE) since Π can be chosen from a distribution which is oblivious to the dataset it is applied to. This is a key property e.g., in applications to low-memory streaming and low-communication distributed algorithms. OSE's have found a widespread application in fast algorithms for numerical linear algebra and regression [9, 20, 21, 26, 30], clustering [5, 11], and classification [23].

Despite their widespread success, OSE's only apply to *linear subspaces*. Theoretical results are limited for more general sets, including natural sets arising in the application of nonlinear models such as neural networks and modern graph and work embedding methods.

1.2 Subspace Embeddings Under Nonlinear Transformations

In this work, we study low-distortion embeddings for *subspaces under entrywise nonlinear transformations.* In particular, we study sets of the form:

$$S = \{y : y = f(x) \text{ for } x \in Z\}, \tag{1}$$

where Z is a k-dimensional linear subspace of $\mathbb{R}^n$ and $f(x)$ is a nonlinear activation function applied entrywise to x. It is helpful to think of such a set S as all possible outputs of a two layer neural network, with k inputs and n outputs. If f is a nonlinear activation function applied to each neuron in the output layer, $W \in \mathbb{R}^{n\times k}$ is the weight matrix connecting the first layer to the second layer, and $x \in \mathbb{R}^k$ is any input, then the neural network output will be $f(Wx)$. Since Wx lies in a k-dimensional subspace (the column span of W), the output set is thus of the form given in (1).

Understanding low-distortion embeddings for the output sets of neural networks is a key theoretical tool behind recent results on compressed sensing from generative models [4, 13, 27]. In particular, [4] study the case for which f is piecewise linear with 2 pieces – e.g., the popular ReLU activation function. In this setting, one can see that the set S lies within a union of linear subspaces. Applying an OSE seperately on each of these subspaces and then taking a union bound, yields a relative error embedding on the set S. [4] also study the case for which f is any Lipschitz function. This encompasses nearly all common activation functions. For such functions, one can extend the results for OSEs which are based on embedding all points in a net with bounded cardinality over the subspace. The approximation of this net is preserved under a Lipschitz transformation, and thus the same argument yields low-distortion embedding bounds for entrywise transformed subspaces. However, this approach only results in embeddings with additive (not relative) error and requires an additional restriction – it applies to S of the form:

$$S = \{y : y = f(x) \text{ for } x \in Z \text{ and } \|x\|_2 \leq R\}, \tag{2}$$

where R is a bound on the radius of the input set.

1.3 Our Contributions

We significantly extend the results on low-distortion embeddings for subspaces under nonlinear transformation. Our results, along with prior work, are summarized in Table 1. Our first bound applies to a wide class of nonlinearities which (1) have a bounded second derivative and (2) approach linear asymptotes for large magnitude x. Such nonlinearities include for example, the Sigmoid $f(x) = \frac{1}{1+e^{-x}}$, the SoftPlus $f(x) = \ln(1+e^x)$, and the Gaussian $f(x) = e^{-x^2}$. We show that functions of this type can be approximated to small uniform error via a piecewise linear function with a bounded number of linear regions. Applying embedding results of [4] for piecewise linear functions then yields an additive error embedding for these functions. Formally:

Theorem 1 (Additive Error Embedding). Let $S = \{y : y = f(x) \text{ for } x \in Z\}$, where Z is a k-dimensional subspace of $\mathbb{R}^n$ and let $f : \mathbb{R} \to \mathbb{R}$ be a nonlinearity satisfying for constants $a, b, c, d_1, e_1, d_2, e_2$ and any $\epsilon \in (0, 1]$:

1. Bounded Second Derivative: $\sup_x |f''(x)| \leq a$ and f'' has a finite number of discontinuities.
2. Linear Asymptotes: $\forall x \geq \frac{c}{\epsilon^b}$, $|f(x) - (d_1 x + e_1)| \leq \epsilon$ and $\forall x \leq -\frac{c}{\epsilon^b}$, $|f(x) - (d_2 x + e_2)| \leq \epsilon$.

Then, if $\Pi \in \mathbb{R}^{m \times n}$ has i.i.d entries $\Pi_{ij} \sim \mathcal{N}(0, \frac{1}{m})$, and $m = O\left(\frac{k \log(n/\epsilon_2) + \log(1/\delta)}{\epsilon_1^2}\right)$ for $\epsilon_1, \epsilon_2, \delta \in (0, 1]$, with probability at least $1 - \delta$, Π is an (ϵ_1, ϵ_2)-error embedding for S.

For simplicity we assume Π to be a random Gaussian embedding matrix. However, our results hold more generally for any family of random embedding matrices that yields a subspace embedding for a k-dimensional subspace with probability $1 - \delta$ using $m = O\left(\frac{k + \log(1/\delta)}{\epsilon^2}\right)$. See [30] for a discussion of various embedding matrix distributions, many of which yield matrices that can be multiplied by much more quickly and stored in less space than a dense Gaussian embedding.

Next, we investigate relative error embeddings, which, prior to our work, were only known for linear spaces or unions of linear spaces. These results suffice for f which is piecewise linear, but not for more general functions. We give the first results for a much wider class of nonlinearities that, both satisfy the second derivative and linear asymptote assumptions of Theorem 1, along with an additional property: they are close to linear at the origin. Such nonlinearities include a large number of 'soft' step functions and rectifying units, including Tanh, ArcTan, the SoftSign, the Square Nonlinearity (SQNL), and the Exponential Linear Unit (ELU). The following theorem gives an embedding for this class of functions.

Theorem 2 (Relative Error Embedding)**.** Let $S = \{y : y = f(x) \text{ for } x \in Z\}$, where Z is a k-dimensional subspace of $\mathbb{R}^n$ and $f : \mathbb{R} \to \mathbb{R}$ is a nonlinearity satisfying conditions (1) and (2) of Theorem 1 along with, for some constants g_1, g_2, g_3:

3. Linear Near Origin[1]: For any y with $|y| \leq g_1$, $|g_2 \cdot f^{-1}(y) - y| \leq g_3 \cdot y^2$.

Then, if $\Pi \in \mathbb{R}^{m \times n}$ has i.i.d entries $\Pi_{ij} \sim \mathcal{N}(0, \frac{1}{m})$, and $m = O\left(\frac{k \log(n/\epsilon) + \log(1/\delta)}{\epsilon^2}\right)$ for $\epsilon, \delta \in (0, 1]$, with probability at least $1 - \delta$, Π is an ϵ-relative-error embedding for S.

Nonlinearity Class	**Examples**	**Embedding Dim.**	**Error Type**	**Reference**
Piecewise linear with t pieces	ReLU, Binary Step Leaky ReLU	$O\left(\frac{k \log(nt)}{\epsilon^2}\right)$	relative	[4] See Thm. 3
L-Lipschitz	Nearly all	$O\left(\frac{k \log(LR/\epsilon_2)}{\epsilon_1^2}\right)$	additive, input bounded in radius R	[4]
f'' bounded, linear asymptotes	Sigmoid, SoftPlus, Gaussian	$O\left(\frac{k \log(n/\epsilon_2)}{\epsilon_1^2}\right)$	additive	Thm. 1
Near-linear at origin, f'' bounded, linear asymptotes	Tanh, Arctan, SQNL SoftSign, ELU	$O\left(\frac{k \log(n/\epsilon)}{\epsilon^2}\right)$	relative	Thm. 2

Table 1: Low-distortion embedding results (Def. 1) for k-dimensional subspaces under entrywise nonlinear transformations. For simplicity we hide dependences on the failure probability δ when embedding with a random linear map. Our results (highlighted in rows 3-4) significantly expand the class of nonlinearities for which low-dimensional embeddings are known and give the first relative error results beyond piecewise linear functions.

1.4 Applications

Our primary technical contributions are the embedding results of Theorems 1 and 2. To illustrate the usefulness of these results, in Section 5 we give example applications to compressed sensing from generative models [4, 27]. In this setting, the goal is to recover $x \in \mathbb{R}^n$ from $m \ll n$ noisy linear measurements $y = Ax + \eta$ where $A \in \mathbb{R}^{m \times n}$ is a measurement matrix and $\eta \in \mathbb{R}^m$ is some measurement noise.

Under the assumption that x lies in some set S (e.g., the set of all possible outputs of a generative neural network $G : \mathbb{R}^k \to \mathbb{R}^n$), approximate recovery up to the noise threshold $\|\eta\|_2$ is possible when A is

[1] Note that when f is bi-Lipschitz, this assumption is equivalent to $|f(y) - g_2 \cdot y| \leq g_3' \cdot x^2$ for some constant g_3'.

an (ϵ_1, ϵ_2)-error embedding for S. Thus, our improved embedding results immediately lead to new results here, removing Lipschitzness and bounded input assumptions required by [4] when G has two layers and employs any nonlinearity satisfying Theorem 1.

In the important case when G has $d > 2$ layers, we show how to apply our techniques to remove the bounded input assumption of [4] for any bounded nonlinearity satisfying the assumptions of Theorem 1, including the Sigmoid, Gaussian, Tanh, Arctan, SoftSign, and SQNL.

1.5 Related Work

Low-distortion embeddings are widely studied in the literature on randomized algorithms and compressed sensing. When S is a *finite set*, the Johnson-Lindenstrauss lemma [12, 17] gives that a random $\Pi \in \mathbb{R}^{m \times n}$ is an ϵ-relative-error embedding with high probability when $m = O\left(\frac{\log |S|}{\epsilon^2}\right)$. A majority of the work on infinite sets focuses on the case where S is a linear subspace. As discussed, in this setting, many constructions for relative-error oblivious subspace embeddings (OSEs) are known. See e.g., [18] and [30] for surveys.

The case where S is the union of linear subspaces is also studied widely in the compressed sensing literature. The well known Restricted Isometry Property (RIP) is equivalent to a relative error embedding for the union of linear subspaces arising as the spans of all subsets of a fixed number of columns of a given matrix [6, 14].

Embeddings for nonlinear spaces have been less explored. As discussed, recent work considers low-distortion embeddings for the output sets of neural networks [4, 13] with ReLU nonlinearities and under Lipschitz assumptions. We build on and significantly extend this work – see Table 1 for a summary. [2] considers embeddings on a smooth manifold, although this is different than our nonlinear entrywise transformation setting. A number of approaches consider random projection for linear regression under various loss functions, including the Huber, Tukey, and Orlicz norm losses [1, 8, 10]. These methods prove low-distortion embedding results for the norms induced by these losses. This can be viewed as embedding results for the standard ℓ_1 or ℓ_2 norms, after applying appropriate entrywise nonlinearity, although the goal is find an embedding $\Pi \in \mathbb{R}^{m \times n}$ so that for $W \in \mathbb{R}^{n \times k}$ and all $x \in \mathbb{R}^k$, $\left\| f(\Pi M x) \right\|_2 \approx \left\| f(Mx) \right\|_2$. This is related to but different from our goal, and requires significantly different techniques.

Finally, we note that Gordon's theorem in functional analysis [16] gives that when S is a set of unit vectors with Gaussian mean width $m = \mathbb{E}_{g \sim \mathcal{N}(0,1)} \sup_{x \in S} \langle g, x \rangle$, a random embedding Π into $O\left(\frac{m^2}{\epsilon^2}\right)$ dimensions is an ϵ-relative error embedding with high probability. The Gaussian mean width is equivalent up to logarithmic factors to the Rademacher complexity of S, a quantity widely studied in computational learning theory [28]. A number of Rademacher complexity bounds are known for neural networks [15, 22], although they don't apply directly in our setting since (1) they bound the complexity of the function class corresponding to the network, rather than its output set S and (2) they are parameterized by various quantities in the neural network, such as the norms of its weight matrices. Our bounds are entirely independent of the neural network parameters, depending only on the nonlinearity used. An interesting direction for future work would be to better understand the connections between randomized dimensionality reduction for subspaces under nonlinear transformations and the work in learning theory on neural networks Rademacher complexity.

2 Embeddings under Piecewise Linear Transformations

We begin by showing how to extend OSE results to subspaces under piecewise linear entrywise transformations. The key idea is that such a transformation fragments the subspace into a bounded number of linear regions, each of which can be embedded with an OSE. This idea is applied e.g., by [4] to embed ReLU networks. For completeness, we give a proof in the general case for any piecewise linear function with t linear pieces.

Theorem 3 (Piecewise Linear Embedding)**.** Let $Z \subseteq \mathbb{R}^n$ be a $k-$dimensional linear subspace and $f : \mathbb{R} \to \mathbb{R}$ be piecewise linear with at most t pieces. Let $S = \{y : y = f(x) \text{ for } x \in Z\}$. Then if $\Pi \in \mathbb{R}^{m \times n}$ has i.i.d.

entries $\Pi_{ij} \sim \mathcal{N}(0, \frac{1}{m})$, $m = O\left(\frac{k\log(nt)+\log(1/\delta)}{\epsilon^2}\right)$ for $\epsilon, \delta > 0$, with probability at least $1-\delta$, Π is an ϵ-relative-error embedding for S (Definition 1).

We establish Theorem 3 from the following lemma, which counts the number of $k-$dimensional linear regions in S. We obtain the embedding for S by a union bound over these regions.

Lemma 1. Let $Z \subseteq \mathbb{R}^n$ be a $k-$dimensional linear subspace and $f : \mathbb{R} \to \mathbb{R}$ be piecewise linear with at most t pieces. Let $S = \{y : y = f(x) \text{ for } x \in Z\}$. S lies in the union of $O((tn)^k)$ k-dimensional linear subspaces.

Proof. Any vector $x \in Z$ can be written as Qz for some $z \in \mathbb{R}^k$ where $Q \in \mathbb{R}^{n\times k}$ has columns spanning Z. Any $z \in \mathbb{R}^k$ thus corresponds to a vector $x \in S$. If we fix the pieces of f that the n entries of Qz fall into, then f simply performs a linear transformation of Qz, and so $x = f(Qz)$ lies in a k-dimensional subspace of $\mathbb{R}^n$. Now, each entry of Qz can fall into one of t pieces of f. Fixing which pieces it falls into splits $\mathbb{R}^k$ using $n \cdot (t-1)$ different $k-1$ dimensional hyperplanes, corresponding to the sets $\{z \in \mathbb{R}^k : (Qz)_i > t_j\}$ where t_j is the j^{th} change point of f.

One can show (c.f. [4]) that c hyperplanes split $\mathbb{R}^k$ into $O(c^k)$ regions. Plugging in $c = n \cdot (t-1)$, we have that S is generated by applying a different linear transformation to $O((tn)^k)$ regions of $\mathbb{R}^k$, and thus S lies in the union of $O((tn)^k)$ k-dimensional subspaces. □

Proof of Theorem 3. Let $S_1, S_2 \ldots, S_w$ be the $w = O((tn)^k)$ linear subspaces , the union of which contains S. It is well known (c.f. Theorem 6 of [30]) that if $\Pi \in \mathbb{R}^{n\times m}$ has independent entries $\Pi_{ij} \sim \mathcal{N}(0, \frac{1}{m})$ and $m = O\left(\frac{k+\log(1/\delta)}{\epsilon}\right)$, then with probability $\geq 1-\delta$, Π is an ϵ-relative-error embedding for any k-dimensional subspace of $\mathbb{R}^n$.

Setting $\delta' = \delta/w = O(\delta/(tn)^k)$, and applying a union bound, we have that Π is an ϵ-relative-error embedding for $S_1 \cup \ldots \cup S_w \supseteq S$ with probability at least $1-\delta$ as long as $m = O\left(\frac{k+\log(1/\delta')}{\epsilon}\right) = O\left(\frac{k\log(nt)+\log(1/\delta)}{\epsilon^2}\right)$. This completes the proof. □

3 Additive Error Embeddings

We next show how to extend the result of Theorem 3 to give additive error embeddings for functions that are well approximated by piecewise linear functions with a bounded number of pieces. Such functions include the popular Sigmoid activation function, the SoftPlus, and the Gaussian activation function. More generally, we give a result for any function which (1) has a bounded second derivative and (2) converges at a reasonable rate to linear asymptotes.

Theorem 1. Let $S = \{y : y = f(x) \text{ for } x \in Z\}$, where Z is a k-dimensional subspace of $\mathbb{R}^n$ and $f : \mathbb{R} \to \mathbb{R}$ is a nonlinearity satisfying for constants $a, b, c, d_1, e_1, d_2, e_2$:

1. Bounded Second Derivative: $\sup_x |f''(x)| \leq a$ and f'' has a finite number of discontinuities.
2. Linear Asymptotes: For any $\epsilon \in (0,1]$, $\forall x \geq \frac{c}{\epsilon^b}$, $|f(x) - (d_1 x + e_1)| \leq \epsilon$ and $\forall x \leq -\frac{1}{\epsilon^b}$, $|f(x) - (d_2 x + e_2)| \leq \epsilon$.

Then, if $\Pi \in \mathbb{R}^{m\times n}$ has i.i.d entries $\Pi_{ij} \sim \mathcal{N}(0, \frac{1}{m})$, and $m = O\left(\frac{k\log(n/\epsilon_2)+\log(1/\delta)}{\epsilon_1^2}\right)$ for $\epsilon_1, \epsilon_2, \delta \in (0,1]$, with probability at least $1-\delta$, Π is an (ϵ_1, ϵ_2)-error embedding for S.

The first assumption of bounded second derivative ensures that f is well approximated by a piecewise linear function with sufficiency small pieces. The second ensures that, outside a range of width $O(1/\epsilon^b)$ around the origin, $f(x)$ can be approximated to ϵ error via a single straight line. This is a crucial condition that applies to a large class of functions and ensures that the piecewise linear approximation has a bounded number of pieces. Formally we show:

Lemma 2. Let $f : \mathbb{R} \to \mathbb{R}$ be a function satisfying the conditions of Theorem 1. Then for any $\epsilon \in (0, 1]$, there exists a piecewise linear function $\tilde{f}(x)$ with $t = O(1/\epsilon^{b+1/2})$ pieces so that, $\forall x \in \mathbb{R}$, $|f(x) - \tilde{f}(x)| \le \epsilon$.

Proof. For $i = 0, 1, \ldots, \lceil \frac{2c}{\gamma \cdot \epsilon^b} \rceil$, let $t_i = \frac{-c}{\epsilon^b} + i \cdot \gamma$, where γ is a stepsize we will define later. These t_i divide the interval $\left[-\frac{c}{\epsilon^b}, \frac{c}{\epsilon_b}\right]$ into subintervals of length γ. Let $\tilde{f} : \mathbb{R} \to \mathbb{R}$ be a piecewise linear approximation of f with $\lceil \frac{2c}{\gamma \cdot \epsilon^b} \rceil + 1$ pieces defined by:

$$\tilde{f}(x) = \begin{cases} d_1 x + e_1, & \text{if } x \ge \frac{c}{\epsilon^b} \\ d_2 x + e_2, & \text{if } x \le -\frac{c}{\epsilon^b} \\ f(t_i) + \frac{f(t_{i+1}) - f(t_i)}{\gamma}(x - t_i) & \text{if } x \in [t_i, t_{i+1}] \end{cases}$$

By assumption (2) of Theorem 1 we have $|f(x) - \tilde{f}(x)| \le \epsilon$ for any $x \notin [-\frac{c}{\epsilon^b}, \frac{c}{\epsilon^b}]$. Thus it suffices to focus on $x \in [-\frac{c}{\epsilon^b}, \frac{c}{\epsilon^b}]$. Within this interval, f is approximated by piecewise linear interpolation over intervals of width γ. For any t_i, t_{i+1} and $x \in [t_i, t_{i+1}]$ it is well known that (c.f. [7]) Rolle's theorem yields a bound on the approximation:

$$|f(x) - \tilde{f}(x)| \le \frac{(t_{i+1} - t_i)^2}{8} \cdot \max_{t \in [t_i, t_{i+1}]} |f''(t)| \le \frac{\gamma^2 \cdot a}{8},$$

by our assumed upper bound of $f''(x) \le a$. Setting $\gamma = \sqrt{\frac{8}{a}} \cdot \sqrt{\epsilon}$ we have $|f(x) - \tilde{f}(x)| \le \epsilon$. We note that this bound requires that $f''(x)$ is continuous on the interval $[t_i, t_{i+1}]$. Since we assume $f''(x)$ has a finite number of discontinuities, we can ensure that this is the case by placing an additional break point at each discontinuity. This will increase the number of linear pieces in $\tilde{f}(x)$ by just an additive constant. The proof is now complete: $\tilde{f}(x)$ is a piecewise linear function with $\lceil \frac{2c}{\gamma \cdot \epsilon^b} \rceil + 1 = O\left(\frac{1}{\epsilon^{b+1/2}}\right)$ pieces with $|f(x) - \tilde{f}(x)| \le \epsilon$, $\forall x \in \mathbb{R}$. □

We Lemma 2 in place, we now show how to extend the embedding bound of Theorem 3 to any function that is well approximated by a piecewise linear function.

Lemma 3. Consider a function $f : \mathbb{R} \to \mathbb{R}$ and the set $S = \{y : y = f(x) \text{ for } x \in Z\}$ where Z is a k-dimensional subspace of $\mathbb{R}^n$. Assume that there exists piecewise linear $\tilde{f} : \mathbb{R} \to \mathbb{R}$ with t pieces and $|f(x) - \tilde{f}(x)| \le \frac{\epsilon_2}{n}$ $\forall x \in \mathbb{R}$. Then, if $\Pi \in \mathbb{R}^{m \times n}$ has i.i.d entries $\Pi_{ij} \sim \mathcal{N}(0, \frac{1}{m})$, and $m = O\left(\frac{k \log(nt) + \log(1/\delta)}{\epsilon_1^2}\right)$, with probability at least $1 - \delta$, Π is an (ϵ_1, ϵ_2)-error embedding for S.

Proof. Define $\tilde{S} = \{\tilde{y} : \tilde{y} = \tilde{f}(x) \quad \text{for } x \in Z\}$. By our approximation assumption, for all $x \in Z$, letting $y = f(x)$ and $\tilde{y} = \tilde{f}(x)$, we have: $\|y - \tilde{y}\|_2 \le \frac{\epsilon_2}{n} \cdot \sqrt{n} = \frac{\epsilon_2}{\sqrt{n}}$. Applying Theorem 3 with parameters ϵ_1 and $\delta/2$, we have that with probability at least $1 - \delta/2$, Π is an ϵ_1-relative-error embedding for $\tilde{S}$. Additionally, it is well known (c.f. [25]) that with probability at least $1 - 2e^{-m/2} \ge 1 - \delta/2$, Π's spectral norm is bounded by $\|\Pi\|_2 \le \frac{3\sqrt{n}}{\sqrt{m}} \le 3\sqrt{n}$. Assuming both events occur, which happens with probability $\ge 1 - \delta$, for any $y \in S$ we have:

$$\begin{aligned} \|\Pi y\|_2 &\le \|\Pi \tilde{y}\|_2 + \|\Pi(y - \tilde{y})\|_2 && \text{(triangle inequality)} \\ &\le (1 + \epsilon_1)\|\tilde{y}\|_2 + \|\Pi\|_2 \cdot \frac{\epsilon_2}{\sqrt{n}} && \text{(subspace embedding)} \\ &\le (1 + \epsilon_1)\left(\|y\|_2 + \frac{\epsilon_2}{\sqrt{n}}\right) + 3\epsilon_2 && \text{(spectral norm bound + triangle inequality)} \\ &\le (1 + \epsilon_1)\|y\|_2 + O(\epsilon_2). \end{aligned}$$

Symmetrically, we can prove that $\|\Pi y\|_2 \ge (1 - \epsilon_1)\|y\| - O(\epsilon_2)$. Adjusting constants on m, we have that Π is an (ϵ_1, ϵ_2)-error embedding for S, completing the proof. □

We now combine Lemmas 2 and 3 to prove the additive error embedding result of Theorem 1.

Proof of Theorem 1. By the assumptions of the theorem and Lemma 2, there exists piecewise linear $\tilde{f} : \mathbb{R} \to \mathbb{R}$ with $t = O\left(\frac{n^{b+1/2}}{\epsilon_2^{b+1/2}}\right)$ pieces and $|f(x) - \tilde{f}(x)| \leq \frac{\epsilon_2}{n}$ for all $x \in \mathbb{R}$. Applying Lemma 3, which holds due to the existence of this $\tilde{f}$, we have that Π is an (ϵ_1, ϵ_2)-error embedding for S when:

$$m = O\left(\frac{k\log(nt) + \log(1/\delta)}{\epsilon_1^2}\right) = O\left(\frac{k\log(n/\epsilon_2) + \log(1/\delta)}{\epsilon_1^2}\right).$$

This completes the theorem. □

3.1 Example Nonlinearities

Many common neural network activation functions satisfy the assumptions of Theorem 1. Thus, the theorem provides a bound on the number of dimensions required to embed the output space of a large class of two-layer neural networks. We give some important examples below.

Sigmoid. $f(x) = \frac{1}{1+e^{-x}}$.

- *Condition 1*: We can compute $f''(x) = \frac{2e^{-2x}}{(1+e^{-x})^3} - \frac{e^{-x}}{(1+e^{-x})^2}$. Thus $\sup_x |f''(x)| = \sup_y |p(y)|$ where $p(x) = \frac{2y^2}{(1+y)^3} - \frac{y}{(1+y)^2}$. We can check that this polynomial is maximized at $p(y) = \frac{1}{6\sqrt{3}}$ at $y = 2 + \sqrt{3}$. Thus condition (1) of Theorem 1 is satisfied with $a = \frac{1}{6\sqrt{3}}$.
- *Condition 2*: We can also check that for any $\epsilon \in (0,1]$, when $x < -\frac{1}{\epsilon} < -\ln(1/\epsilon)$, $f(x) \in [0, \epsilon)$. Similarly, when $x > \frac{1}{\epsilon} > \ln(1/\epsilon)$, $f(x) \in [\frac{1}{1+\epsilon}, 1] \subset [1-\epsilon, 1]$. Thus, condition (2) is satisfied with $b = c = 1$, $d_1 = 1$, $d_2 = 0$, and $e_1 = e_2 = 0$.

SoftPlus. $f(x) = \ln(1 + e^x)$.

- *Condition 1*: We can compute $f''(x) = \frac{e^x}{(1+e^x)^2}$. Thus $\sup_x |f''(x)| = \sup_y |p(y)|$ where $p(x) = \frac{y}{(1+y)^2}$. We can check that this polynomial is maximized at $p(y) = \frac{1}{4}$ at $y = 1$. Thus condition (1) of Theorem 1 is satisfied with $a = \frac{1}{4}$.
- *Condition 2*: We can also check that for any $\epsilon \in (0,1]$, when $x > \frac{1}{\epsilon} > \ln(1/\epsilon)$, $f(x) \geq x$ and $f(x) \leq \ln((1+\epsilon)e^x) \leq x + \ln(1+\epsilon) \leq x + \epsilon$. Thus, $|f(x) - x| \leq \epsilon$. Similarly, when $x < -\frac{1}{\epsilon} < \ln(\epsilon)$, $f(x) \geq 0$ and $f(x) \leq \ln(1+\epsilon) \leq \epsilon$. Thus $|f(x)| < \epsilon$. So, condition (2) is satisfied with $b = c = 1$, $d_1 = d_2 = 0$, and $e_1 = 1$ and $e_2 = 0$.

Gaussian. $f(x) = e^{-x^2}$.

- *Condition 1*: We can verify that $f''(x) = e^{-x^2}(4x^2 - 2)$, and has $\sup_x |f''(x)| = |f''(0)| = 2$. Thus condition (1) of Theorem 1 is satisfied with $a = 2$.
- *Condition 2*: We can also check that for any $\epsilon \in (0,1]$, when $|x| \geq \sqrt{\ln(1/\epsilon)} \leq \frac{1}{\epsilon}$, $|f(x)| \leq \epsilon$, and thus condition (2) is satisfied with $b = c = 1$ and $d_1 = d_2 = e_1 = e_2 = 0$.

4 Relative Error Embeddings

We now show that the additive error embedding result of Theorem 1 can be improved to relative error under the additional assumption that the nonlinearity f is close to linear near the origin. This assumption holds for a many 'soft' step functions and rectifying units, including Tanh, ArcTan, SoftSign, Square Nonlinearity (SQNL), and the Exponential Linear Unit (ELU).

Theorem 2. Let $S = \{y : y = f(x) \text{ for } x \in Z\}$, where Z is a k-dimensional subspace of $\mathbb{R}^n$ and $f : \mathbb{R} \to \mathbb{R}$ is a nonlinearity satisfying conditions (1) and (2) of Theorem 1 along with, for some constants g_1, g_2, g_3:

3. Linear Near Origin: For any y with $|y| \leq g_1$, $|g_2 \cdot f^{-1}(y) - y| \leq g_3 \cdot y^2$.

Then, if $\Pi \in \mathbb{R}^{m \times n}$ has i.i.d entries $\Pi_{ij} \sim \mathcal{N}(0, \frac{1}{m})$, and $m = O\left(\frac{k\log(n/\epsilon)+\log(1/\delta)}{\epsilon^2}\right)$ for $\epsilon, \delta \in (0,1]$, with probability at least $1-\delta$, Π is an ϵ-relative-error embedding for S.

Proof. Assume without loss of generality that $\epsilon < g_1$. If it is not, we can replace ϵ with $\min(g_1, \epsilon)$, and since g_1 is a fixed constant, this will affect the bound only by constants. We split S into two sets containing elements with relatively large norms and relatively small norms. Specifically, $S = S_L \cup S_U$ where $S_L = \{y \in S : \|y\|_2 > \epsilon/\sqrt{n}\}$ and $S_U = \{y \in S : \|y\|_2 \leq \epsilon/\sqrt{n}\}$. We then prove that with probability $1-\delta/2$, Π is an ϵ-relative-error embedding for each of S_L and S_U. Via a union bound, this yields the theorem.

Case 1: S_L. Since by assumption f satisfies the requirements of Theorem 1, applying that theorem with $\epsilon_1 = \frac{\epsilon}{2}$ and $\epsilon_2 = \frac{\epsilon^2}{2\sqrt{n}}$ and gives that, for $m = O\left(\frac{k\log(n/\epsilon)+\log(1/\delta)}{\epsilon^2}\right)$, with probability $1-\delta/2$, for all $y \in S_L$:

$$\|\Pi y\|_2 \leq (1+\tfrac{\epsilon}{2})\|y\|_2 + \tfrac{\epsilon^2}{2\sqrt{n}} \leq (1+\epsilon)\|y\|_2,$$

where the second bound holds since for $y \in S_L$, $\|y\|_2 \geq \frac{\epsilon}{\sqrt{n}}$ and thus $\frac{\epsilon^2}{2\sqrt{n}} \leq \frac{\epsilon}{2}\|y\|_2$. Similarly, we have $\|\Pi y\|_2 \geq (1-\epsilon)\|y\|_2$, which completes the bound in this case.

Case 2: S_U. We prove the theorem for S_U using the fact f is close to linear near the origin – i.e., where $\|y\|_2$ is small. Let $\tilde{f}(x) = g_2 \cdot x$ be a linear approximation to f near the origin, i.e. for all x such that $|x| < g_1, \tilde{y} = \tilde{f}(x)$. The approximation to S thus becomes $\tilde{S} = \{\tilde{y} : \tilde{y} = \tilde{f}(x) \quad \text{for } x \in Z\}$. By assumption (3) of the theorem, for $y \in S_U$, $\|y\|_2 \leq \frac{\epsilon}{\sqrt{n}}$ and thus for all $i \in \{1, 2, \ldots n\}$, $|y(i)| \leq \frac{\epsilon}{\sqrt{n}} < g_1$. This gives that:

$$|g_2 \cdot f^{-1}(y(i)) - y(i)| = |\tilde{y}(i) - y(i)| \leq g_3 \cdot y(i)^2 \leq \frac{g_3 \cdot \epsilon}{\sqrt{n}} \cdot y(i).$$

In turn we have:

$$\|y - \tilde{y}\|_2 \leq \frac{g_3 \cdot \epsilon}{\sqrt{n}} \cdot \|y\|_2. \tag{3}$$

Now, note that $\tilde{S}$ is just a k-dimensional linear subspace. As discussed in the proof of Theorem 1, it is well known that for $m = O\left(\frac{k+\log(1/\delta)}{\epsilon^2}\right)$, with probability $\geq 1-\delta/2$, $\|\Pi\|_2 \leq 3\sqrt{n}$ and for all $\tilde{y} \in \tilde{S}$, $(1-\epsilon)\|\tilde{y}\|_2 \leq \|\Pi\tilde{y}\|_2 \leq (1+\epsilon)\|\tilde{y}\|_2$ (i.e., Π is an ϵ-error subspace embedding for $\tilde{S}$). Along with (3), these two conditions give that, for every $y \in S$:

$$\begin{aligned}
\|\Pi y\|_2 &\leq \|\Pi\tilde{y}\|_2 + \|\Pi(y-\tilde{y})\|_2 \\
&\leq (1+\epsilon)\|\tilde{y}\|_2 + \|\Pi\|_2 \cdot \|y-\tilde{y}\|_2 \\
&\leq (1+\epsilon)\|y\|_2 + (1+\epsilon+\|\Pi\|_2) \cdot \|y-\tilde{y}\|_2 \\
&\leq (1+\epsilon)\|y\|_2 + \frac{g_3(1+\epsilon+3\sqrt{n}) \cdot \epsilon}{\sqrt{n}}\|y\|_2 \leq (1+c\epsilon)\|y\|_2,
\end{aligned}$$

for some constant c. Similarly, one can prove that $\|\Pi y\|_2 \geq (1-c\epsilon)\|y\|_2$. Thus, adjusting constants on ϵ by increasing m by a constant gives that, with probability $1-\delta/2$, Π is an ϵ-relative-error embedding for S. Combined with our argument for Case 1 (the set S_L), this completes the proof. □

4.1 Example Nonlinearities

Many common neural network activation functions satisfy the assumptions of Theorem 2. In particular, soft step functions and rectifying units (i.e., soft variants of the ReLU) often have linear asymptotes and are close to linear near the origin. We give two illustrative examples below: Tanh and ELU. Other nonlinearities, including ArcTan, SoftSign and the Square Nonlinearity (SQNL) are described in Appendix A.

Tanh (Hyperbolic Tangent). $f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

- *Condition 1*: We can check that $\sup_x |f''(x)| = \frac{4}{3\sqrt{3}}$, achieved at $x = \frac{1}{2}\ln(2-\sqrt{3})$. Thus, condition (1) of Theorem 1 is satisfied with $a = \frac{4}{3\sqrt{3}}$.
- *Condition 2*: For $x > \frac{1}{\epsilon} > \ln(1/\epsilon)$, we have $f(x) \leq 1$ and $f(x) \geq \frac{1/\epsilon - \epsilon}{1/\epsilon + \epsilon} = \frac{1-\epsilon^2}{1+\epsilon^2} \geq 1 - \epsilon$. So $|f(x) - 1| \leq \epsilon$. Similarly, for $x < -\frac{1}{\epsilon} < \ln(\epsilon)$ we have $f(x) \geq -1$ and $f(x) \leq \frac{\epsilon - 1/\epsilon}{\epsilon + 1/\epsilon} = \frac{-(1-\epsilon^2)}{1+\epsilon^2} \leq 1 - \epsilon$. Thus, $|f(x) + 1| \leq \epsilon$. Thus, condition (2) of Theorem 1 is satisfied with $b = c = 1$.
- *Condition 3*: $f^{-1}(y) = \frac{1}{2}\ln\left(\frac{1+y}{1-y}\right)$. We can check that $\frac{\left|\frac{1}{2}\ln\left(\frac{1+y}{1-y}\right) - y\right|}{y^2} \leq \frac{1}{5}$ for $y \in [-1/2, 1/2]$. Thus, the final condition (3) of Theorem 2 holds with $g_1 = 1/2$, $g_2 = 1$, and $g_3 = 1/5$.

Exponential Linear Unit (ELU). $f(x) = \begin{cases} e^x - 1 \text{ for } x \leq 0 \\ x \text{ for } x \geq 0 \end{cases}.$

- *Condition 1*: For $x \geq 0$ we have $f''(x) = 0$. For $x \leq 0$, we have $f''(x) = e^x \leq 1$. Thus, $\sup_x |f''(x)| \leq 1$ and condition (1) of Theorem 1 is satisfied with $a = 1$.
- *Condition 2*: For $x > \frac{1}{\epsilon}$, we have $f(x) = x$ and thus, $|f(x) - x| = 0$. For $x < -\frac{1}{\epsilon} < -\ln(1/\epsilon)$, we have $|f(x) + 1| \leq \epsilon$. Hence condition (2) of Theorem 1 is satisfied with $b = c = 1$.
- *Condition 3*: We have $f^{-1}(y) = \begin{cases} \ln(1+y) \text{ for } y \leq 0 \\ y \text{ for } y \geq 0 \end{cases}.$

  We can check that $\frac{|f^{-1}(y) - y|}{y^2} \leq 1$ for $y \in [-1/2, 0]$ and $\frac{|f^{-1}(y) - y|}{y^2} = 0$ for $y > 0$. Thus, condition (3) of Theorem 2 holds with $g_1 = 1/2, g_2 = 1$ and $g_3 = 1$.

5 Application: Compressed Sensing from Generative Models

Recently, deep generative models have become an important tool in the recovery of high-dimensional data from limited measurements using compressed sensing techniques [4, 24, 27]. They have found significant success in solving linear inverse problems [19], offering a powerful alternative to the traditional structural assumption of sparsity.

Formally, compressed sensing seeks to recover a signal $x \in \mathbb{R}^n$ from $m \ll n$ linear measurements, $y = Ax + \eta$, where $A \in \mathbb{R}^{m \times n}$ is the measurement matrix and $\eta \in \mathbb{R}^m$ is some measurement noise. Recovering x from y requires solving this underdetermined and noisy linear system – a task which is only possible under structural assumptions on x. Most commonly, in the *sparse recovery* setting, it is assumed that x is sparse in some basis, such as the Fourier or Wavelet basis [14]. Methods based on generative models instead assume that x lies in the output span of some generative neural network $G : \mathbb{R}^k \to \mathbb{R}^n$. That is, x lies in a low-dimensional subspace under a series of linear transformations and entrywise nonlinearities.

[4] extend the well-known restricted eigenvalue condition (REC) from sparse recovery, showing that, under the assumption that x lies in some set S, as long as the objective function $\min_{x \in S} \|y - Ax\|_2$ can be

minimized to small additive error (e.g., via projected gradient descent), x can be approximately recovered from any measurement matrix $A \in \mathbb{R}^{m\times n}$ satisfying the *S-REC property*:

$$\|A(x_1 - x_2)\| \geq (1-\epsilon_1)\|x_1 - x_2\| - \epsilon_2 \quad \forall x_1, x_2 \in S. \tag{4}$$

In turn, [4] consider $S = \{x : x = G(z) \text{ for } z \in \mathbb{R}^k, \|z\|_2 \leq R\}$ – the output span of a generative model G under a bounded input restriction. They show that when $A \in \mathbb{R}^{m\times n}$ has i.i.d. $\mathcal{N}(0, 1/m)$ entries, it satisfies (4) with high probability as long as $m = O\left(\frac{k\log(LR/\epsilon_2)}{\epsilon_1^2}\right)$, where L is the Lipschitzness of G (i.e., for any $z_1, z_2 \in \mathbb{R}^k$, $\|G(z_1) - G(z_2)\| \leq L\|z_1 - z_2\|$). When G uses just ReLU nonlinearities, the bounded radius and Lipschitz assumptions can be removed, $\epsilon_2 = 0$, and $m = O\left(\frac{dk\log n}{\epsilon_1^2}\right)$, where d is the depth of the neural network.

5.1 Our Results

Our improved embedding results immediately apply to the setting of [4], letting us remove the dependence on the Lipschitz constant L and the assumption of a bounded input $\|z\|_2 \leq R$ for two layer neural networks under the nonlinearities discussed in Sections 3 and 4 (including the Sigmoid, Tanh, ELU, Softplus, etc.)

We employ a small modification of Theorem 1, which applies to the *difference of two vectors* generated from a subspace under an entrywise nonlinearity. This theorem is proven essentially identically to Theorem 1.

Theorem 4 (Additive Error Embedding – Distance)**.** Let $S = \{y : y = f(x) \text{ for } x \in Z\}$, where Z is a k-dimensional subspace of $\mathbb{R}^n$ and $f : \mathbb{R} \to \mathbb{R}$ is a nonlinearity satisfying the conditions of Theorem 1. Then, if $\Pi \in \mathbb{R}^{m\times n}$ has i.i.d entries $\Pi_{ij} \sim \mathcal{N}(0, \frac{1}{m})$, and $m = O\left(\frac{k\log(n/\epsilon_2)+\log(1/\delta)}{\epsilon_1^2}\right)$ for $\epsilon_1, \epsilon_2, \delta \in (0, 1]$, with probability at least $1-\delta$, for all $y_1, y_2 \in S$:

$$(1-\epsilon_1)\|y_1 - y_2\|_2 - \epsilon_2 \leq \|\Pi(y_1 - y_2)\|_2 \leq (1+\epsilon_1)\|y_1 - y_2\|_2 + \epsilon_2.$$

Now, let $G : \mathbb{R}^k \to \mathbb{R}^n$ be a two layered generative neural network with $G(z) = f(Wz)$ for some weight matrix $W \in \mathbb{R}^{n\times k}$ and some nonlinearity f satisfying the conditions of Theorem 1. Let S be the output set of G: $S = \{x \in \mathbb{R}^n : x = G(z) \text{ for } z \in \mathbb{R}^k\}$. Then Theorem 4 implies that, when A has random Gaussian entries, it satisfies the restricted eigenvalue condition of (4), and thus, x can be recovered from noisy measurements $y = Ax + \eta$. In comparison to the result of [4], $m = O\left(\frac{k\log(n/\epsilon_2)+\log(1/\delta)}{\epsilon_1^2}\right)$ has no dependence on the Lipschitzness L of $G(z)$. Additionally, the bound holds under the weaker assumption that S is G's full output set, rather than the outputs restricted to the range of bounded diameter inputs.

5.2 Extension to deep networks

Our results apply to depth-2 neural networks, and an important direction for future work is to extend them to general depth-d networks. In this section, we give an example of how our techniques can be applied to deeper networks.

Let $G : \mathbb{R}^k \to \mathbb{R}^n$ be a neural network with d layers and $\leq n$ nodes per non-input layer. The previously mentioned results of [4] show that when $A \in \mathbb{R}^{n\times m}$ has i.i.d entries $A_{ij} \sim \mathcal{N}(0, \frac{1}{m})$, it satisfies the S-REC property of (4) for $S = \{x : x = G(z) \text{ for } z \in \mathbb{R}^k, \|z\|_2 \leq R\}$ and $m = O\left(\frac{k\log(LR/\epsilon_2)}{\epsilon_1^2})\right)$. We extend this result, showing how to remove the norm restriction on the representation z for nonlinearities that satisfy the conditions of Theorem 1 and are bounded in magnitude by some constant u. This includes all soft step functions we consider, such as the Sigmoid, Tanh and SoftStep.

We split G into the composition of two functions: $G_1 : \mathbb{R}^k \to \mathbb{R}^n$ mapping the input layer to the second layer and $G_2 : \mathbb{R}^n \to \mathbb{R}^n$, mapping the second layer to the output. Assume that G_2 is L-Lipschitz Note that $G_1(z) = f(W_1 z)$, where W_1 is the weight matrix of the first layer and f is the nonlinearity.

Let $\tilde{G}_1 : \mathbb{R}^k \to \mathbb{R}^n$ be an approximation to G_1 which uses a piecewise linear approximation $\tilde{f}$ with $|\tilde{f}(x) - f(x)| \leq \frac{\epsilon_2}{nL}\ \forall x$. The existence of $\tilde{f}$ with $t = O\left(\left(\frac{nL}{\epsilon_2}\right)^{b+1/2}\right)$ pieces is guarantee by Lemma 2. We have for any $z \in \mathbb{R}^k$, $\left\|G_1(z) - \tilde{G}_1(z)\right\|_2 \leq \frac{\epsilon_2}{\sqrt{n}L}$.

Let $\tilde{G}(z) = G_2(\tilde{G}_1(z))$. By our Lipschitzness assumption on G_2, for all z,

$$\left\|G(z) - \tilde{G}(z)\right\| = \left\|G_2(G_1(z)) - G_2(\tilde{G}_1(z))\right\|_2 \leq L \cdot \left\|G_1(z) - \tilde{G}_1(z)\right\|_2 \leq \frac{\epsilon_2}{\sqrt{n}}. \tag{5}$$

Additionally, by Lemma 1, the output of $\tilde{G}_1(z)$ lies in the union of $(nt)^k$ k-dimensional linear subspaces. Since we assume $f(x) \leq u$ for all x, $\tilde{f}(x) \leq u + \frac{\epsilon_2}{nL}$ for all x. Thus $\left\|\tilde{G}_1(z)\right\|_2 \leq (u + \frac{\epsilon_2}{nL}) \cdot \sqrt{n} = O(\sqrt{n})$. Thus, the output of $\tilde{G}(z)$ lies in the union of t regions of the form $S = \{G_2(z') : z' \in Z, \left\|z'\right\|_2 = O(\sqrt{n})\}$, where Z is a k-dimensional subspace. We know via the results of [4] and a union bound over these t regions that for $m = O\left(\frac{k \log(Ln/\epsilon_2) + \log 1/\delta}{\epsilon_1^2}\right)$, with probability $\geq 1 - \delta$, for any $\tilde{x}_1, \tilde{x}_2 \in \mathbb{R}^n$ in the approximate output set $\tilde{S} = \{\tilde{G}(z) : z \in \mathbb{R}^k\}$,

$$\left\|A(\tilde{x}_1 - \tilde{x}_2)\right\|_2 \geq (1 - \epsilon_1)\|\tilde{x}_1 - \tilde{x}_2\|_2 - \epsilon_2.$$

For any $x_1, x_2 \in \mathbb{R}^n$ in the true output set $S = \{\tilde{G}(z) : z \in \mathbb{R}^k\}$ via (5) we thus have, following the proof of Lemma 3:

$$\begin{aligned}
\left\|A(x_1 - x_2)\right\|_2 &\geq \left\|A(\tilde{x}_1 - \tilde{x}_2)\right\|_2 - \|A\|_2 \cdot \frac{2\epsilon_2}{\sqrt{n}} && \text{(triangle inequality)}\\
&\geq (1 - \epsilon_1)\|\tilde{x}_1 - \tilde{x}_2\|_2 - \epsilon_2 - O(\epsilon_2) && (\|A\|_2 = O(\sqrt{n}) \text{ with high probability)}\\
&\geq (1 - \epsilon_1)\|x_1 - x_2\|_2 - O(\epsilon_2) && \text{(triangle inequality)}
\end{aligned}$$

Adjusting constants on ϵ_2, this gives us the S-REC property of (4) for $S = \{G(z) : z \in \mathbb{R}^k\}$ when A makes $m = O\left(\frac{k \log(Ln/\epsilon_2) + \log 1/\delta}{\epsilon_1^2}\right)$ measurements. Thus, for any Lipschitz neural network using bounded linearities satisfying the assumptions of Theorem 1, we obtain a similar result to [4] but without the bounded input assumption.

5.3 Conclusions and Future Work

Our paper makes initial steps in building a systematic understanding of randomized dimensionality reduction for subspaces under entrywise nonlinear transformations. An important next step is to extend our results to the output spaces of neural networks with $d > 2$ layers. It is possible to use an argument similar to Theorem 1 to give some bounds here, by approximating all nonlinearities in the neural network via piecewise linear functions. However, due to compounding error at each level, ϵ_2 must be set very small at the first level, leading to relatively weak embedding bounds. Understanding how to avoid this compounding error would be very interesting.

As discussed, it would also be interesting to apply Rademacher and other complexity bounds for learning neural networks to understanding the compressibility of their output spaces and to give low-distortion embedding bounds. This would let us leverage an even richer class of tools in proving embedding bounds.

References

[1] A. Andoni, C. Lin, Y. Sheng, P. Zhong, and R. Zhong. Subspace embedding and linear regression with Orlicz norm. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 224–233, 2018.

[2] R. G. Baraniuk and M. B. Wakin. Random projections of smooth manifolds. *Foundations of Computational Mathematics*, 9(1):51–77, 2009.

[3] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 245–250, 2001.

[4] A. Bora, A. Jalal, E. Price, and A. G. Dimakis. Compressed sensing using generative models. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 537–546, 2017.

[5] C. Boutsidis, A. Zouzias, and P. Drineas. Random projections for k-means clustering. In *Advances in Neural Information Processing Systems 23 (NeurIPS)*, 2010.

[6] E. J. Candes, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

[7] N. Carothers. Approximation theory. *Bowling Green State University, Ohio*, 1998.

[8] K. L. Clarkson, R. Wang, and D. P. Woodruff. Dimensionality reduction for Tukey regression. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[9] K. L. Clarkson and D. P. Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing (STOC)*, 2013.

[10] K. L. Clarkson and D. P. Woodruff. Sketching for M-estimators: A unified approach to robust regression. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 921–939. SIAM, 2014.

[11] M. B. Cohen, S. Elder, C. Musco, C. Musco, and M. Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC)*, 2015.

[12] S. Dasgupta and A. Gupta. An elementary proof of the Johnson-Lindenstrauss lemma. *International Computer Science Institute, Technical Report*, 22(1):1–5, 1999.

[13] M. Dhar, A. Grover, and S. Ermon. Modeling sparse deviations for compressed sensing using generative models. *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[14] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[15] N. Golowich, A. Rakhlin, and O. Shamir. Size-independent sample complexity of neural networks. In *Proceedings of the 31st Annual Conference on Computational Learning Theory (COLT)*, pages 297–299, 2018.

[16] Y. Gordon. On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. In *Geometric Aspects of Functional Analysis*, pages 84–106. Springer, 1988.

[17] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26(189-206):1, 1984.

[18] R. Kannan and S. Vempala. Randomized algorithms in numerical linear algebra. *Acta Numerica*, 26:95, 2017.

[19] M. T. McCann, K. H. Jin, and M. Unser. Convolutional neural networks for inverse problems in imaging: A review. *IEEE Signal Processing Magazine*, 34(6):85–95, 2017.

[20] X. Meng and M. W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing (STOC)*, pages 91–100, 2013.

[21] J. Nelson and H. L. Nguyên. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2013.

[22] B. Neyshabur, R. Tomioka, and N. Srebro. Norm-based capacity control in neural networks. In *Proceedings of the 28th Annual Conference on Computational Learning Theory (COLT)*, pages 1376–1401, 2015.

[23] S. Paul, C. Boutsidis, M. Magdon-Ismail, and P. Drineas. Random projections for support vector machines. In *Artificial Intelligence and Statistics*, pages 498–506, 2013.

[24] J. Rick Chang, C.-L. Li, B. Poczos, B. Vijaya Kumar, and A. C. Sankaranarayanan. One network to solve them all–solving linear inverse problems using deep projection models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5888–5897, 2017.

[25] M. Rudelson and R. Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010)*, pages 1576–1602, 2010.

[26] T. Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 143–152, 2006.

[27] V. Shah and C. Hegde. Solving linear inverse problems using gan priors: An algorithm with provable guarantees. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4609–4613. IEEE, 2018.

[28] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[29] S. S. Vempala. *The Random Projection Method*, volume 65. American Mathematical Society, 2005.

[30] D. P. Woodruff et al. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014.

A Example Nonlinearities for Relative Error Embeddings

We now give a number of other examples of nonlinearities that satisfy the assumptions of our relative error embedding result, Theorem 2.

ArcTan. $f(x) = \tan^{-1}(x)$

- *Condition 1*: . We can check that $\sup_x |f''(x)| = \frac{3\sqrt{3}}{8}$ achieved at $|x| = \frac{1}{\sqrt{3}}$. Thus, condition (1) of Theorem 1 is satisfied with $a = \frac{3\sqrt{3}}{8}$.
- *Condition 2*: We use a series expansion which gives that:

$$\tan^{-1}(x) = \begin{cases} x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + .. \text{ for } |x| \leq 1 \\ \frac{\pi}{2} - \frac{1}{x} + \frac{1}{3x^3} - .. \text{ for } x \geq 1 \\ -\frac{\pi}{2} - \frac{1}{x} + \frac{1}{3x^3} - .. \text{ for } x \leq -1 \end{cases}.$$

  For $x \geq \frac{1}{\epsilon}$, we thus have $f(x) \leq \frac{\pi}{2}$ and $f(x) \geq \frac{\pi}{2} - \epsilon$. Thus $|f(x) - \frac{\pi}{2}| \leq \epsilon$. Similarly, for $x \leq -\frac{1}{\epsilon}$, we have $|f(x) + \frac{\pi}{2}| \leq \epsilon$. Thus, condition (2) of Theorem 1 is satisfied with $b = c = 1$.
- *Condition 3*: $f^{-1}(y) = \tan(y)$ for $y \in (-\frac{\pi}{2}, \frac{\pi}{2})$. We can check that when $|y| \leq 1$, $\frac{|tan(y)-y|}{y^2} \leq \tan(1) - 1 \leq .56$. Thus, condition (3) of Theorem 2 holds with $g_1 = 1, g_2 = 1$ and $g_3 = .56$.

SoftSign. $f(x) = \frac{x}{1+|x|}$.

- *Condition 1*: It can be checked that $f''(x) = \frac{2x}{(1+|x|)^3} - \frac{2|x|}{x(1+|x|)^2}$ and $\sup_x |f''(x)| = 2$, achieved at $x = 0$. Thus, condition (1) of Theorem 1 is satisfied with $a = 2$.
- *Condition 2*: For $x > \frac{1}{\epsilon}$, we have $f(x) \leq 1$ and $f(x) \geq 1 - \frac{1}{1+x} \geq 1 - \epsilon$. Thus, $|f(x) - 1| \leq \epsilon$. Similarly, for $x < -\frac{1}{\epsilon}$, we have $f(x) \geq -1$ and $f(x) \leq -1 + \frac{1}{1-x} \leq -1 + \epsilon$. Thus, $|f(x) + 1| \leq \epsilon$. Hence condition (2) of Theorem 1 is satisfied with $b = c = 1$.
- *Condition 3*: We have $f^{-1}(y) = \frac{y}{1-|y|}$. It can be checked that $\frac{|f^{-1}(y)-y|}{y^2} \leq 2$ when $|y| \leq 1/2$. Thus, condition (3) of Theorem 2 holds for for $g_1 = 1/2$, $g_2 = 1$ and $g_3 = 2$.

Square Nonlinearity (SQNL). Here $f(x) = \begin{cases} 1 \text{ for } x \geq 2 \\ x - \frac{x^2}{4} \text{ for } x \in [0, 2] \\ x + \frac{x^2}{4} \text{ for } x \in [-2, 0] \\ -1 \text{ for } x \leq 2 \end{cases}$.

- *Condition 1*: $f''(x) = 0$ for $x \notin [-2, 2]$, $f''(x) = -\frac{1}{2}$ for $x \in [0, 2]$ and $f''(x) = \frac{1}{2}$ for $x \in [-2, 0]$ Thus, $\sup_x |f''(x)| = \frac{1}{2}$ and so condition (1) of Theorem 1 is satisfied with $a = \frac{1}{2}$.
- *Condition 2*: For $x \geq \frac{1}{\epsilon}$, we have $f(x) = 1$ and hence, $|f(x) - 1| = 0$. For $x \leq -\frac{1}{\epsilon}$, we have $f(x) = -1$ and hence $|f(x) + 1| = 0$. Hence condition (2) of Theorem 1 is satisfied with $b = c = 1$.
- *Condition 3*: $f^{-1}(y) = \begin{cases} 2 - 2\sqrt{1-y} \text{ for } y \in [0, 2] \\ -2 + 2\sqrt{1+y} \text{ for } x \in [-2, 0] \end{cases}$.

  We can check that $\frac{|f^{-1}(y)-y|}{y^2} \leq 1$ for $y \in [-1/2, 1/2]$, which gives that condition (3) of Theorem 2 holds for for $g_1 = 1/2$, $g_2 = 1$ and $g_3 = 1$.