

## A PROOF OF THEOREM 1

First, we prove the following lemma.

**Lemma 1.** *Let  $d < D$  and  $p$  be positive integers, and let  $L, M > 0$  and  $\rho \in (0, 1)$  be constants. Let  $\mathcal{S} \subset \mathbb{R}^D$  be a bounded subset for which there exists a  $\rho$ -JL embedding  $\mathbf{A} \in \mathbb{R}^{d \times D}$  of  $\mathcal{S}$  into  $[-M, M]^d$ . Then, for any  $L$ -Lipschitz function  $f : \mathcal{S} \rightarrow \mathbb{R}^p$ , there exists an  $\frac{L}{1-\rho}$ -Lipschitz function  $g : [-M, M]^d \rightarrow \mathbb{R}^p$  such that  $g(\mathbf{A}\mathbf{x}) = f(\mathbf{x})$  for all  $\mathbf{x} \in \mathcal{S}$ .*

*Proof.* For any  $\mathbf{x}, \mathbf{x}' \in \mathcal{S}$ , if  $\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x}'$  then since  $\mathbf{A}$  is a  $\rho$ -JL embedding of  $\mathcal{S}$ , we have that  $\|\mathbf{x} - \mathbf{x}'\|_2 \leq \frac{1}{1-\rho} \|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}'\|_2 = \frac{1}{1-\rho} \|\mathbf{0}\|_2 = 0$ , and so,  $\|\mathbf{x} - \mathbf{x}'\|_2 = 0$ , i.e.,  $\mathbf{x} = \mathbf{x}'$ . Therefore, the map  $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$  from  $\mathcal{S}$  to  $\mathbf{A}(\mathcal{S}) := \{\mathbf{A}\mathbf{x} : \mathbf{x} \in \mathcal{S}\}$  is invertible. We define  $A^{-1} : \mathbf{A}(\mathcal{S}) \rightarrow \mathcal{S}$  to be the inverse of the map  $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ .

Now, for any  $L$ -Lipschitz function  $f : \mathcal{S} \rightarrow \mathbb{R}^p$ , we define  $\tilde{g} : \mathbf{A}(\mathcal{S}) \rightarrow \mathbb{R}^p$  by  $\tilde{g} = f \circ A^{-1}$ . Then, for any  $\mathbf{y}, \mathbf{y}' \in \mathbf{A}(\mathcal{S})$ , we have

$$\begin{aligned} \|\tilde{g}(\mathbf{y}) - \tilde{g}(\mathbf{y}')\|_2 &= \|f(A^{-1}(\mathbf{y})) - f(A^{-1}(\mathbf{y}'))\|_2 && \text{since } g = f \circ A^{-1} \\ &\leq L \|A^{-1}(\mathbf{y}) - A^{-1}(\mathbf{y}')\|_2 && \text{since } f \text{ is } L\text{-Lipschitz} \\ &\leq \frac{L}{1-\rho} \|\mathbf{A}A^{-1}(\mathbf{y}) - \mathbf{A}A^{-1}(\mathbf{y}')\|_2 && \text{since } \mathbf{A} \text{ is a } \rho\text{-JL embedding of } \mathcal{S} \\ &= \frac{L}{1-\rho} \|\mathbf{y} - \mathbf{y}'\|_2. && \text{since } A^{-1} \text{ is the inverse of } \mathbf{x} \mapsto \mathbf{A}\mathbf{x} \end{aligned}$$

Therefore,  $\tilde{g} : \mathbf{A}(\mathcal{S}) \rightarrow \mathbb{R}^p$  is  $\frac{L}{1-\rho}$ -Lipschitz. Then, since  $\mathbf{A}(\mathcal{S}) \subset [-M, M]^d$ , by the Kirschbraun theorem (Schwartz, 1969), there exists a  $\frac{L}{1-\rho}$ -Lipschitz extension of  $\tilde{g}$  to  $[-M, M]^d$ , i.e., a function  $g : [-M, M]^d \rightarrow \mathbb{R}^p$  which is  $\frac{L}{1-\rho}$ -Lipschitz on  $[-M, M]^d$  and satisfies  $g(\mathbf{y}) = \tilde{g}(\mathbf{y})$  for all  $\mathbf{y} \in \mathbf{A}(\mathcal{S})$ . Finally, for any  $\mathbf{x} \in \mathcal{S}$ , we have  $\mathbf{A}\mathbf{x} \in \mathbf{A}(\mathcal{S})$ , and so,  $g(\mathbf{A}\mathbf{x}) = \tilde{g}(\mathbf{A}\mathbf{x}) = f(A^{-1}(\mathbf{A}\mathbf{x})) = f(\mathbf{x})$ , as required.  $\square$

**Remark:** In (Azagra *et al.*, 2021), the authors give an explicit formula for the Lipschitz extension of a given Lipschitz function.

With Lemma 1, we can now prove each of the four parts of Theorem 1. As a reminder, we assume that  $\mathcal{S} \subset \mathbb{R}^D$  is a bounded set for which there exists a  $\rho$ -JL embedding  $\mathbf{A} \in \mathbb{R}^{d \times D}$  of  $\mathcal{S}$  into  $[-M, M]^d$ .

a) Let  $f : \mathcal{S} \rightarrow \mathbb{R}^p$  be an  $L$ -Lipschitz function. By Lemma 1, there exists a  $\frac{L}{1-\rho}$ -Lipschitz function  $g : [-M, M]^d \rightarrow \mathbb{R}^p$  such that  $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$  for all  $\mathbf{x} \in \mathcal{S}$ . By assumption,  $g$  can be  $\epsilon$ -approximated by a feedforward neural network with at most  $\mathcal{N}$  nodes,  $\mathcal{E}$  edges, and  $\mathcal{L}$  layers. In other words, there exists a function  $\hat{g}$  such that  $\|\hat{g}(\mathbf{y}) - g(\mathbf{y})\|_\infty \leq \epsilon$  for all  $\mathbf{y} \in [-M, M]^d$ , and  $\hat{g}$  can be implemented by a feedforward neural network with at most  $\mathcal{N}$  nodes,  $\mathcal{E}$  edges, and  $\mathcal{L}$  layers.

Define another function  $\hat{f} = \hat{g} \circ \mathbf{A}$ , i.e.,  $\hat{f}(\mathbf{x}) = \hat{g}(\mathbf{A}\mathbf{x})$  for all  $\mathbf{x} \in \mathcal{S}$ . Since  $\mathbf{A}(\mathcal{S}) \subset [-M, M]^d$  by assumption, we have that  $\mathbf{A}\mathbf{x} \in [-M, M]^d$  for all  $\mathbf{x} \in \mathcal{S}$ . Then,  $\|\hat{f}(\mathbf{x}) - f(\mathbf{x})\|_\infty = \|\hat{g}(\mathbf{A}\mathbf{x}) - g(\mathbf{A}\mathbf{x})\|_\infty \leq \epsilon$  for all  $\mathbf{x} \in \mathcal{S}$ , i.e.,  $\hat{f}$  is an  $\epsilon$ -approximation of  $f$ .

Furthermore, we can construct a feedforward neural network to implement  $\hat{f} = \hat{g} \circ \mathbf{A}$  by having a linear layer to implement the map  $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ , and then feeding this into the neural network implementation of  $\hat{g}$ . The map  $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$  can be implemented with  $D$  nodes for the input layer, and  $Dd$  edges between the input nodes and first hidden layer. By assumption,  $\hat{g}$  can be implemented by a feedforward neural network with at most  $\mathcal{N}$  nodes,  $\mathcal{E}$  edges, and  $\mathcal{L}$  layers. Hence,  $\hat{f} = \hat{g} \circ \mathbf{A}$  can be implemented by a feedforward neural network with at most  $\mathcal{N} + D$  nodes,  $\mathcal{E} + Dd$  edges, and  $\mathcal{L} + 1$  layers, as desired.

b) If the same feedforward neural network architecture  $\epsilon$ -approximates every  $\frac{L}{1-\rho}$ -Lipschitz function  $g : [-M, M]^d \rightarrow \mathbb{R}^p$ , then our construction of a feedforward neural network that implements

$\hat{f} = \hat{g} \circ \mathbf{A}$  has the same architecture for every  $L$ -Lipschitz function  $f : \mathcal{S} \rightarrow \mathbb{R}^p$ . Hence, the same bounds on the number of nodes, edges, and layers hold.

c) In a similar manner as 1a), we form a CNN that can approximate  $f = g \circ \mathbf{A}$  by first implementing the linear map  $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$  with a CNN and feeding this into a CNN that approximates  $g$ .

The JL matrix  $\mathbf{A} = \mathbf{M}\mathbf{D}$  can be represented by a Resnet-CNN structure as follows. Let  $\mathbf{x}$  be the input of the network, then  $\mathbf{D}\mathbf{x}$ , the random sign flip of the input can be realized by setting the weight/kernel  $w_1$  of the first two layers to be the delta function, and the bias vectors to take large values at the location where  $\mathbf{D}$  has a 1, and small values where  $\mathbf{D}$  has a  $-1$ . Then with the help of the ReLU activation, we can successfully flip the signs. More explicitly, set  $T = \sup_{\mathbf{x} \in \mathcal{S}} \|\mathbf{x}\|_\infty$  so that  $\mathbf{x} \in [-T, T]^D$  for all  $\mathbf{x} \in \mathcal{S}$ . Let  $\mathbf{b}_i$  be the bias to be added to the  $i$ th coordinate of the input. We design a 2 layer Resnet-CNN,  $L(\mathbf{x})$ , as follows

$$L(\mathbf{x})_i = \text{ReLU}(2\mathbf{x}_i + \mathbf{b}_i) - \text{ReLU}(\mathbf{b}_i) - \mathbf{x}_i, \quad i = 1, \dots, D.$$

The bias  $\mathbf{b}_i$  are chosen to realize the sign flip as follows. If  $\mathbf{D}_{ii}$  contains a 1, then we set  $\mathbf{b}_i = 2T$ , which will make  $L(\mathbf{x})_i = \mathbf{x}_i$ . If  $\mathbf{D}_{ii}$  contains a  $-1$ , then we set  $\mathbf{b}_i = -2T$ , which will make  $L(\mathbf{x})_i = -\mathbf{x}_i$ , thus realizing the sign-flip. This architecture can also be used to realize a mask (i.e., setting certain entries of  $\mathbf{x}$  to 0). For the application of  $\mathbf{M}$  to  $\mathbf{D}\mathbf{x}$ , it is a simply a convolution with a mask, therefore again can be achieved by 2 layers of Resnet-CNN.

This Resnet-CNN that implements  $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$  requires  $2D$  nodes,  $D$  parameters, and 2 layers to apply  $\mathbf{D}$  to  $\mathbf{x}$ , and an additional  $D$  nodes,  $D + d$  parameters, and 2 layers to apply  $\mathbf{M}$  to  $\mathbf{D}\mathbf{x}$ . By adding this to the  $\mathcal{N}$  nodes,  $\mathcal{P}$  parameters, and  $\mathcal{L}$  layers needed for a CNN to approximate the  $\frac{L}{1-\rho}$ -Lipschitz function  $g : [-M, M]^d \rightarrow \mathbb{R}^p$ , we obtain that the  $L$ -Lipshitz function  $f : \mathcal{S} \rightarrow \mathbb{R}^p$  can be approximated by a Resnet-CNN with  $\mathcal{N} + 3D$  nodes,  $\mathcal{P} + 2D + d$  parameters, and  $\mathcal{L} + 4$  layers.

d) If the same convolutional neural network architecture  $\epsilon$ -approximates every  $\frac{L}{1-\rho}$ -Lipschitz function  $g : [-M, M]^d \mapsto \mathbb{R}^p$ , then our construction of a convolutional neural network that implements  $\hat{f} = \hat{g} \circ \mathbf{A}$  has the same architecture for every  $L$ -Lipschitz function  $f : \mathcal{S} \rightarrow \mathbb{R}^p$ . Hence, the same bounds on the number of nodes, edges, and layers hold.

## B PROOF OF PROPOSITION 1

Consider a covering of  $U_{\mathcal{S}}$  by  $\mathcal{N}(U_{\mathcal{S}}, \|\cdot\|_2, \delta)$  balls of radius  $\delta$ . Each ball must intersect  $U_{\mathcal{S}}$  as otherwise we could remove that ball from the covering and obtain a covering of  $U_{\mathcal{S}}$  with only  $\mathcal{N}(U_{\mathcal{S}}, \|\cdot\|_2, \delta) - 1$  balls of radius  $\delta$ , which contradicts the definition of  $\mathcal{N}(U_{\mathcal{S}}, \|\cdot\|_2, \delta)$ . Enumerate these balls  $i = 1, \dots, \mathcal{N}(U_{\mathcal{S}}, \|\cdot\|_2, \delta)$ . For each  $i$ , pick a point  $\mathbf{u}_i \in U_{\mathcal{S}}$  which is also in the  $i$ -th ball, and then pick points  $\mathbf{x}_i, \mathbf{x}'_i \in \mathcal{S}$  with  $\mathbf{x}_i \neq \mathbf{x}'_i$  such that  $\frac{\mathbf{x}_i - \mathbf{x}'_i}{\|\mathbf{x}_i - \mathbf{x}'_i\|_2} = \mathbf{u}_i$ . Then, set  $\mathcal{S}_1 = \{\mathbf{x}_i\}_i \cup \{\mathbf{x}'_i\}_i$  so  $|\mathcal{S}_1| \leq 2\mathcal{N}(U_{\mathcal{S}}, \|\cdot\|_2, \delta)$ .

Suppose  $\mathbf{A} \in \mathbb{R}^{d \times D}$  is a  $\rho$ -JL embedding of  $\mathcal{S}_1$ . Then, by definition of a  $\rho$ -JL embedding,

$$(1 - \rho)\|\mathbf{x}_i - \mathbf{x}'_i\| \leq \|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}'_i\| \leq (1 + \rho)\|\mathbf{x}_i - \mathbf{x}'_i\|, \quad \text{for } i = 1, \dots, \mathcal{N}(U_{\mathcal{S}}, \|\cdot\|_2, \delta)$$

Now, for any two points  $\mathbf{y}, \mathbf{y}' \in \mathcal{S}$  with  $\mathbf{y} \neq \mathbf{y}'$ , there exists an index  $i$  such that  $\frac{\mathbf{y} - \mathbf{y}'}{\|\mathbf{y} - \mathbf{y}'\|_2} \in U_{\mathcal{S}}$  lies in the  $i$ -th ball of our covering of  $U_{\mathcal{S}}$ . Since  $\frac{\mathbf{x}_i - \mathbf{x}'_i}{\|\mathbf{x}_i - \mathbf{x}'_i\|_2}$  is also in the  $i$ -th ball, we have that

$$\left\| \frac{\mathbf{x}_i - \mathbf{x}'_i}{\|\mathbf{x}_i - \mathbf{x}'_i\|} - \frac{\mathbf{y} - \mathbf{y}'}{\|\mathbf{y} - \mathbf{y}'\|} \right\| \leq 2\delta.$$

For simplicity of notation, we set  $a = \|\mathbf{x}_i - \mathbf{x}'_i\|$ ,  $b = \|\mathbf{y} - \mathbf{y}'\|$ . Then we immediately have

$$\begin{aligned} \|\mathbf{A}(\mathbf{y} - \mathbf{y}')\| &= \left\| \frac{b}{a}\mathbf{A}(\mathbf{x}_i - \mathbf{x}'_i) + \mathbf{A}(\mathbf{y} - \mathbf{y}') - \frac{b}{a}\mathbf{A}(\mathbf{x}_i - \mathbf{x}'_i) \right\| \\ &\leq \frac{b}{a}\|\mathbf{A}(\mathbf{x}_i - \mathbf{x}'_i)\| + \|\mathbf{A}(\mathbf{y} - \mathbf{y}') - \frac{b}{a}(\mathbf{x}_i - \mathbf{x}'_i)\| \\ &\leq \frac{b}{a}(1 + \rho)\|\mathbf{x}_i - \mathbf{x}'_i\| + \|\mathbf{A}\| \|\mathbf{y} - \mathbf{y}' - \frac{b}{a}(\mathbf{x}_i - \mathbf{x}'_i)\| \\ &\leq (1 + \rho)b + 2\|\mathbf{A}\|\delta b = (1 + \rho + 2\|\mathbf{A}\|\delta)\|\mathbf{y} - \mathbf{y}'\|, \end{aligned}$$

where the second inequality used the previous two formulae. The other side of the bi-lipschitz formula can be proved similarly. Hence,  $\mathbf{A}$  is also a  $(\rho + 2\|\mathbf{A}\|\delta)$ -JL embedding of  $\mathcal{S}$ .

## C PROOF OF PROPOSITION 4

a) By Proposition 1, there exists a finite set  $\mathcal{S}_1$  with at most  $|\mathcal{S}_1| \leq 2\mathcal{N}(U_{\mathcal{S}}, \|\cdot\|_2, \frac{\rho}{4\sqrt{3D}})$  points such that any  $\frac{\rho}{2}$ -JL embedding of  $\mathcal{S}_1$  is also a  $(\frac{\rho}{2} + \|\mathbf{A}\|\frac{\rho}{2\sqrt{3D}})$ -JL embedding of  $\mathcal{S}$ .

We now show that there exists a matrix  $\mathbf{A} \in \mathbb{R}^{d \times D}$  with  $\|\mathbf{A}\| \leq \sqrt{3D}$  which is  $\frac{\rho}{2}$ -JL embedding of  $\mathcal{S}_1$  by generating a random  $\mathbf{A}$  and showing that the probability of  $\|\mathbf{A}\| \leq \sqrt{3D}$  and  $\mathbf{A}$  is a  $\frac{\rho}{2}$ -JL embedding of  $\mathcal{S}_1$  both occurring is greater than zero.

Let  $\mathbf{A} \in \mathbb{R}^{d \times D}$  be a random matrix whose entries are i.i.d. from a subgaussian distribution with mean 0 and variance  $\frac{1}{d}$ . Since

$$\mathbb{E}\|\mathbf{A}\|_F^2 = \sum_{i=1}^d \sum_{j=1}^D \mathbb{E}A_{i,j}^2 = \sum_{i=1}^d \sum_{j=1}^D \frac{1}{d} = D,$$

we have that

$$\mathbb{P}\{\|\mathbf{A}\|_F^2 \geq 3D\} \leq \frac{\mathbb{E}\|\mathbf{A}\|_F^2}{3D} = \frac{1}{3}.$$

Furthermore, since

$$d \gtrsim \rho^{-2} \log \mathcal{N}(U_{\mathcal{S}}, \|\cdot\|_2, \frac{\rho}{4\sqrt{3D}}) \gtrsim \left(\frac{\rho}{2}\right)^{-2} \log(3|\mathcal{S}_1|),$$

by Proposition 2,  $\mathbf{A}$  is a  $\frac{\rho}{2}$ -JL embedding of  $\mathcal{S}_1$  with probability at least  $1 - \frac{1}{3} = \frac{2}{3}$ . Therefore,  $\mathbf{A}$  is both a  $\frac{\rho}{2}$ -JL embedding of  $\mathcal{S}_1$  and satisfies  $\|\mathbf{A}\| \leq \sqrt{3D}$  with probability at least  $\frac{2}{3} - \frac{1}{3} = \frac{1}{3} > 0$ .

Hence, there exists a matrix  $\mathbf{A} \in \mathbb{R}^{d \times D}$  such that  $\mathbf{A}$  is a  $\frac{\rho}{2}$ -JL embedding of  $\mathcal{S}_1$  and satisfies  $\|\mathbf{A}\| \leq \sqrt{3D}$ . Finally, by Proposition 1, since  $\mathbf{A}$  is a  $\frac{\rho}{2}$ -JL embedding of  $\mathcal{S}_1$ , it is also a  $(\frac{\rho}{2} + \|\mathbf{A}\|\frac{\rho}{2\sqrt{3D}})$ -JL embedding of  $\mathcal{S}$ . Since  $\|\mathbf{A}\| \leq \sqrt{3D}$ , we have  $\frac{\rho}{2} + \|\mathbf{A}\|\frac{\rho}{2\sqrt{3D}} \leq \rho$ , and thus,  $\mathbf{A}$  is a  $\rho$ -JL embedding of  $\mathcal{S}$ , as desired.

b) Again, by Proposition 1, there exists a finite set  $\mathcal{S}_1$  with at most  $|\mathcal{S}_1| \leq 2\mathcal{N}(U_{\mathcal{S}}, \|\cdot\|_2, \frac{\rho}{4\sqrt{3D}})$  points such that any  $\frac{\rho}{2}$ -JL embedding of  $\mathcal{S}_1$  is also a  $(\frac{\rho}{2} + \|\mathbf{A}\|\frac{\rho}{2\sqrt{3D}})$ -JL embedding of  $\mathcal{S}$ .

Let  $\mathbf{A} \in \mathbb{R}^{d \times D}$  be a random matrix of the form  $\mathbf{M}\mathbf{D}$  where  $\mathbf{D} \in \mathbb{R}^{D \times D}$  is a diagonal matrix whose entries are independent Rademacher random variables, and  $\mathbf{M} \in \mathbb{R}^{d \times D}$  is a random circulant matrix whose entries are Gaussian random variables with mean 0 and variance  $\frac{1}{d}$  and entries in different diagonals are independent. Again, we can show that

$$\mathbb{E}\|\mathbf{A}\|_F^2 = \sum_{i=1}^d \sum_{j=1}^D \mathbb{E}A_{i,j}^2 = \sum_{i=1}^d \sum_{j=1}^D \mathbb{E}M_{i,j}^2 D_{j,j}^2 = \sum_{i=1}^d \sum_{j=1}^D \mathbb{E}M_{i,j}^2 = \sum_{i=1}^d \sum_{j=1}^D \frac{1}{d} = D,$$

and so,

$$\mathbb{P}\{\|\mathbf{A}\|_F^2 \geq 3D\} \leq \frac{\mathbb{E}\|\mathbf{A}\|_F^2}{3D} = \frac{1}{3}.$$

Now, set  $\alpha = \log(\log|\mathcal{S}_1|)/\log(\log(4D + 4d))$  so that  $\log^\alpha|\mathcal{S}_1| = \log(4D + 4d)$ . Then, since

$$d \gtrsim \rho^{-2} \log(4D + 4d) \log \mathcal{N}(U_{\mathcal{S}}, \|\cdot\|_2, \frac{\rho}{4\sqrt{3D}}) \gtrsim \left(\frac{\rho}{2}\right)^{-2} \log^{1+\alpha}|\mathcal{S}_1|,$$

by Proposition 3,  $\mathbf{A}$  is a  $\frac{\rho}{2}$ -JL embedding of  $\mathcal{S}_1$  with probability at least

$$\frac{2}{3} \left(1 - (D + d)e^{-\log^\alpha|\mathcal{S}_1|}\right) = \frac{2}{3} \left(1 - (D + d)e^{-\log(4D + 4d)}\right) = \frac{2}{3} \left(1 - \frac{1}{4}\right) = \frac{1}{2}.$$

Therefore,  $\mathbf{A}$  is both a  $\frac{\rho}{2}$ -JL embedding of  $\mathcal{S}_1$  and satisfies  $\|\mathbf{A}\| \leq \sqrt{3D}$  with probability at least  $\frac{1}{2} - \frac{1}{3} = \frac{1}{6} > 0$ .

Hence, there exists a matrix  $\mathbf{A} \in \mathbb{R}^{d \times D}$  such that  $\mathbf{A}$  is a  $\frac{\rho}{2}$ -JL embedding of  $\mathcal{S}_1$  and satisfies  $\|\mathbf{A}\| \leq \sqrt{3D}$ . Again, by Proposition 1, since  $\mathbf{A}$  is a  $\frac{\rho}{2}$ -JL embedding of  $\mathcal{S}_1$ , it is also a  $(\frac{\rho}{2} + \|\mathbf{A}\| \frac{\rho}{2\sqrt{3D}})$ -JL embedding of  $\mathcal{S}$ . Since  $\|\mathbf{A}\| \leq \sqrt{3D}$ , we have  $\frac{\rho}{2} + \|\mathbf{A}\| \frac{\rho}{2\sqrt{3D}} \leq \rho$ , and thus,  $\mathbf{A}$  is a  $\rho$ -JL embedding of  $\mathcal{S}$ , as desired.

## D PROOF OF PROPOSITION 6

We first construct a function  $\hat{g}$  that is an  $\epsilon$ -approximation of  $g$ . To do this, we first define a compactly supported “spike” function  $\phi : \mathbb{R}^d \rightarrow [0, 1]$  by

$$\phi(\mathbf{z}) = \max \{1 + \min \{z_1, \dots, z_d, 0\} - \max \{z_1, \dots, z_d, 0\}, 0\}.$$

Then, for any positive integer  $N$ , define an approximation  $\hat{g} : [-M, M]^d \rightarrow \mathbb{R}^p$  to  $g$  by

$$\hat{g}(\mathbf{y}) := \sum_{\mathbf{n} \in \{-N, \dots, N\}^d} g\left(\frac{M\mathbf{n}}{N}\right) \phi\left(\frac{N\mathbf{y}}{M} - \mathbf{n}\right).$$

Similarly to what was done in (Yarotsky, 2018), it can be shown that the scaled and shifted spike functions  $\{\phi(\frac{N\mathbf{y}}{M} - \mathbf{n})\}_{\mathbf{n} \in \{-N, \dots, N\}^d}$  form a partition of unity, i.e.

$$\sum_{\mathbf{n} \in \{-N, \dots, N\}^d} \phi\left(\frac{N\mathbf{y}}{M} - \mathbf{n}\right) = 1 \quad \text{for all } \mathbf{y} \in [-M, M]^d.$$

Trivially,  $\phi(\mathbf{y}) \geq 0$  for all  $\mathbf{y} \in \mathbb{R}^d$ . Also, one can check that  $\text{supp}(\phi) \subseteq [-1, 1]^d$ , and thus,  $\phi(\frac{N\mathbf{y}}{M} - \mathbf{n}) = 0$  for all  $\mathbf{n}$  such that  $\|\frac{N\mathbf{y}}{M} - \mathbf{n}\|_\infty > 1$ . Furthermore, for any  $\mathbf{n}$  such that  $\|\frac{N\mathbf{y}}{M} - \mathbf{n}\|_\infty \leq 1$ , we have

$$\|g(\mathbf{y}) - g\left(\frac{M\mathbf{n}}{N}\right)\|_2 \leq L\|\mathbf{y} - \frac{M\mathbf{n}}{N}\|_2 \leq L\sqrt{d}\|\mathbf{y} - \frac{M\mathbf{n}}{N}\|_\infty \leq \frac{LM\sqrt{d}}{N}.$$

Hence, we can bound the approximation error for any  $\mathbf{y} \in [-M, M]^d$  as follows:

$$\begin{aligned} \|\hat{g}(\mathbf{y}) - g(\mathbf{y})\|_2 &= \left\| \sum_{\mathbf{n} \in \{-N, \dots, N\}^d} g\left(\frac{M\mathbf{n}}{N}\right) \phi\left(\frac{N\mathbf{y}}{M} - \mathbf{n}\right) - g(\mathbf{y}) \right\|_2 \\ &= \left\| \sum_{\mathbf{n} \in \{-N, \dots, N\}^d} \left(g\left(\frac{M\mathbf{n}}{N}\right) - g(\mathbf{y})\right) \phi\left(\frac{N\mathbf{y}}{M} - \mathbf{n}\right) \right\|_2 \\ &\leq \sum_{\mathbf{n} \in \{-N, \dots, N\}^d} \left\| g\left(\frac{M\mathbf{n}}{N}\right) - g(\mathbf{y}) \right\|_2 \phi\left(\frac{N\mathbf{y}}{M} - \mathbf{n}\right) \\ &= \sum_{\left\| \frac{N\mathbf{y}}{M} - \mathbf{n} \right\|_\infty \leq 1} \left\| g\left(\frac{M\mathbf{n}}{N}\right) - g(\mathbf{y}) \right\|_2 \phi\left(\frac{N\mathbf{y}}{M} - \mathbf{n}\right) \\ &\leq \sum_{\left\| \frac{N\mathbf{y}}{M} - \mathbf{n} \right\|_\infty \leq 1} \frac{LM\sqrt{d}}{N} \phi\left(\frac{N\mathbf{y}}{M} - \mathbf{n}\right) \\ &\leq \sum_{\mathbf{n} \in \{-N, \dots, N\}^d} \frac{LM\sqrt{d}}{N} \phi\left(\frac{N\mathbf{y}}{M} - \mathbf{n}\right) \\ &= \frac{LM\sqrt{d}}{N}. \end{aligned}$$

So by choosing  $N = \left\lceil \frac{LM\sqrt{d}}{\epsilon} \right\rceil$ , we can obtain  $\|\hat{g}(\mathbf{y}) - g(\mathbf{y})\|_\infty \leq \|\hat{g}(\mathbf{y}) - g(\mathbf{y})\|_2 \leq \epsilon$  for all  $\mathbf{y} \in [-M, M]^d$ , i.e.,  $\hat{g}$  is an  $\epsilon$ -approximation of  $g$ .

We now focus on constructing a ReLU NN architecture which can implement the  $\epsilon$ -approximation  $\hat{g}$  for any  $L$ -Lipschitz function  $g$ . We do this by first constructing a ReLU NN that is independent

of  $g$  which implements the map  $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}^{(2N+1)^d}$  defined by  $(\Phi(\mathbf{y}))_{\mathbf{n}} = \phi(\frac{N\mathbf{y}}{M} - \mathbf{n})$ . Then, we add a final layer which outputs the appropriate linear combination of the  $\phi(\frac{N\mathbf{y}}{M} - \mathbf{n})$ 's.

**Lemma 2.** *For any integers  $N, d \geq 1$ , the maps  $m_d : \mathbb{R}^d \rightarrow \mathbb{R}^{(2N+1)^d}$  and  $M_d : \mathbb{R}^d \rightarrow \mathbb{R}^{(2N+1)^d}$  defined by*

$$(m_d(\mathbf{y}))_{\mathbf{n}} := \min \left\{ \frac{N\mathbf{y}_1}{M} - \mathbf{n}_1, \dots, \frac{N\mathbf{y}_d}{M} - \mathbf{n}_d, 0 \right\} \quad \text{for } \mathbf{n} \in \{-N, \dots, N\}^d$$

and

$$(M_d(\mathbf{y}))_{\mathbf{n}} := \max \left\{ \frac{N\mathbf{y}_1}{M} - \mathbf{n}_1, \dots, \frac{N\mathbf{y}_d}{M} - \mathbf{n}_d, 0 \right\} \quad \text{for } \mathbf{n} \in \{-N, \dots, N\}^d,$$

can both be implemented by a ReLU NN with  $O((2N+1)^d)$  weights,  $O((2N+1)^d)$  nodes, and  $\lceil \log_2(d+1) \rceil$  layers.

*Proof.* First, we note that we can write

$$(m_d(\mathbf{y}))_{\mathbf{n}} = \min \left\{ \min \left\{ \frac{N\mathbf{y}_1}{M} - \mathbf{n}_1, \dots, \frac{N\mathbf{y}_{\lceil d/2 \rceil}}{M} - \mathbf{n}_{\lceil d/2 \rceil} \right\}, \right. \\ \left. \min \left\{ \frac{N\mathbf{y}_{\lceil d/2 \rceil+1}}{M} - \mathbf{n}_{\lceil d/2 \rceil+1}, \dots, \frac{N\mathbf{y}_d}{M} - \mathbf{n}_d, 0 \right\} \right\}$$

and

$$(M_d(\mathbf{y}))_{\mathbf{n}} = \max \left\{ \max \left\{ \frac{N\mathbf{y}_1}{M} - \mathbf{n}_1, \dots, \frac{N\mathbf{y}_{\lceil d/2 \rceil}}{M} - \mathbf{n}_{\lceil d/2 \rceil} \right\}, \right. \\ \left. \max \left\{ \frac{N\mathbf{y}_{\lceil d/2 \rceil+1}}{M} - \mathbf{n}_{\lceil d/2 \rceil+1}, \dots, \frac{N\mathbf{y}_d}{M} - \mathbf{n}_d, 0 \right\} \right\}$$

In (Arora *et al.*, 2016), it is shown that for any positive integer  $k$ , the maps  $(z_1, \dots, z_k) \mapsto \min\{z_1, \dots, z_k\}$  and  $(z_1, \dots, z_k) \mapsto \max\{z_1, \dots, z_k\}$  can be implemented by a ReLU NN with at most  $c_1 k$  edges,  $c_2 k$  nodes, and  $\lceil \log_2 k \rceil$  layers, where  $c_1, c_2 > 0$  are universal constants. So to construct the map  $m_d$ , we first implement the  $(2N+1)^{\lceil d/2 \rceil}$  maps

$$(\mathbf{y}_1, \dots, \mathbf{y}_{\lceil d/2 \rceil}) \mapsto \min \left\{ \frac{N\mathbf{y}_1}{M} - \mathbf{n}_1, \dots, \frac{N\mathbf{y}_{\lceil d/2 \rceil}}{M} - \mathbf{n}_{\lceil d/2 \rceil} \right\} \quad (5)$$

for  $(\mathbf{n}_1, \dots, \mathbf{n}_{\lceil d/2 \rceil}) \in \{-N, \dots, N\}^{\lceil d/2 \rceil}$ . Implementing each of these maps requires  $c_1 \lceil \frac{d}{2} \rceil$  edges,  $c_2 \lceil \frac{d}{2} \rceil$  nodes, and  $\lceil \log_2 \lceil \frac{d}{2} \rceil \rceil$  layers. Next, we implement the  $(2N+1)^{\lfloor d/2 \rfloor}$  maps

$$(\mathbf{y}_{\lceil d/2 \rceil+1}, \dots, \mathbf{y}_d) \mapsto \min \left\{ \frac{N\mathbf{y}_{\lceil d/2 \rceil+1}}{M} - \mathbf{n}_{\lceil d/2 \rceil+1}, \dots, \frac{N\mathbf{y}_d}{M} - \mathbf{n}_d, 0 \right\} \quad (6)$$

for  $(\mathbf{n}_{\lceil d/2 \rceil+1}, \dots, \mathbf{n}_d) \in \{-N, \dots, N\}^{\lfloor d/2 \rfloor}$ . Implementing each of these maps requires  $c_1(\lfloor \frac{d}{2} \rfloor + 1)$  edges,  $c_2(\lfloor \frac{d}{2} \rfloor + 1)$  nodes, and  $\lceil \log_2(\lfloor \frac{d}{2} \rfloor + 1) \rceil$  layers. After placing these  $(2N+1)^{\lceil d/2 \rceil} + (2N+1)^{\lfloor d/2 \rfloor}$  maps in parallel, we construct one final layer as follows. For each  $\mathbf{n} = (\mathbf{n}_1, \dots, \mathbf{n}_d) \in \{-N, \dots, N\}^d$ , we combine the output of the  $(\mathbf{n}_1, \dots, \mathbf{n}_{\lceil d/2 \rceil})$ -th map of the form in Equation 5 and the output of the  $(\mathbf{n}_{\lceil d/2 \rceil+1}, \dots, \mathbf{n}_d)$ -th map of the form in Equation 6 by using them as inputs to a ReLU NN that implements the map  $(a, b) \mapsto \min\{a, b\}$ . Each of these requires at most  $2c_1$  edges and  $2c_2$  nodes.

The total number of edges used to implement  $m_d$  is

$$\begin{aligned} & c_1 \lceil \frac{d}{2} \rceil (2N+1)^{\lceil d/2 \rceil} + c_1(\lfloor \frac{d}{2} \rfloor + 1)(2N+1)^{\lfloor d/2 \rfloor} + 2c_1(2N+1)^d \\ & \leq c_1(\lceil \frac{d}{2} \rceil + \lfloor \frac{d}{2} \rfloor + 1)(2N+1)^{\lceil d/2 \rceil} + 2c_1(2N+1)^d \\ & = c_1(d+1)(2N+1)^{\lceil d/2 \rceil} + 2c_1(2N+1)^d \\ & = c_1 \left( (d+1)(2N+1)^{-\lfloor d/2 \rfloor} + 2 \right) (2N+1)^d \\ & \leq c_1 \left( (d+1) \cdot 3^{-\lfloor d/2 \rfloor} + 2 \right) (2N+1)^d \\ & \leq 4c_1(2N+1)^d, \end{aligned}$$

where we have used the fact that  $N \geq 1$  by definition, and the easily verifiable inequality  $(d+1) \cdot 3^{-\lfloor d/2 \rfloor} \leq 2$  for all positive integers  $d$ .

A nearly identical calculation shows that the total number of nodes used to implement  $m_d$  is at most  $4c_2(2N+1)^d$ . Finally, since the  $(2N+1)^{\lceil d/2 \rceil}$  maps of the form in Equation 5 and the  $(2N+1)^{\lfloor d/2 \rfloor}$  maps of the form in Equation 6 are in parallel, the total number of layers used to implement  $m_d$  is

$$\max \left\{ \lceil \log_2 \lceil \frac{d}{2} \rceil \rceil, \lceil \log_2 (\lfloor \frac{d}{2} \rfloor + 1) \rceil \right\} + 1 = \lceil \log_2(d+1) \rceil.$$

Hence, the map  $m_d$  can be implemented by a ReLU NN with at most  $C_1(2N+1)^d$  edges,  $C_2(2N+1)^d$  nodes, and  $\lceil \log_2(d+1) \rceil$  layers, as desired. The proof for  $M_d$  is identical, except with min replaced by max.  $\square$

Next, we note that

$$(\Phi(\mathbf{y}))_{\mathbf{n}} = \phi\left(\frac{N\mathbf{y}}{M} - \mathbf{n}\right) = \max \{1 + (m_d(\mathbf{y}))_{\mathbf{n}} - (M_d(\mathbf{y}))_{\mathbf{n}}, 0\} \quad \text{for all } \mathbf{n} \in \{-N, \dots, N\}^d.$$

So to construct a ReLU NN which implements  $\Phi$ , we first place a ReLU NN that implements  $m_d$  in parallel with a ReLU NN that implements  $M_d$ . Then, we add an extra layer which has  $(2N+1)^d$  nodes, where the  $\mathbf{n}$ -th node of this layer has two edges, one from the  $\mathbf{n}$ -th node of  $m_d$  and one from the  $\mathbf{n}$ -th node of  $M_d$ . Since  $m_d$  and  $M_d$  are in parallel and each can each be implemented with ReLU NNs with  $O((2N+1)^d)$  edges,  $O((2N+1)^d)$  nodes, and  $\lceil \log_2(d+1) \rceil$  layers, and the last layer has  $2(2N+1)^d$  edges and  $(2N+1)^d$  nodes, the ReLU NN which implements  $\Phi$  has  $O((2N+1)^d)$  edges,  $O((2N+1)^d)$  nodes, and  $\lceil \log_2(d+1) \rceil + 1$  layers.

Finally, we can construct a ReLU NN which implements

$$\hat{g}(\mathbf{x}) := \sum_{\mathbf{n} \in \{-N, \dots, N\}^d} g\left(\frac{M\mathbf{n}}{N}\right) \phi\left(\frac{N\mathbf{x}}{M} - \mathbf{n}\right)$$

by using the ReLU NN which implements  $\Phi$ , followed by a linear layer which computes the weighted sum for  $\hat{g}$ . This last layer has  $p$  nodes, and  $p(2N+1)^d$  edges. So the ReLU NN that implements  $\hat{g}$  has  $(p+C_1)(2N+1)^d$  edges,  $C_2(2N+1)^d + p$  nodes, and  $\lceil \log_2(d+1) \rceil + 2$  layers, as desired.

## E PROOF OF THEOREM 2

By combining Proposition 4a and Proposition 5, we have that there exists a  $\rho$ -JL embedding  $\mathbf{A} \in \mathbb{R}^{d \times D}$  of  $\mathcal{S}$  with

$$d \gtrsim \min \left\{ \rho^{-2} \log \mathcal{N}(U_{\mathcal{S}}, \|\cdot\|_2, \frac{\rho}{4\sqrt{3D}}), \rho^{-2} (\omega(U_{\mathcal{S}}))^2 \right\}.$$

Let  $M = \sup_{\mathbf{x} \in \mathcal{S}} \|\mathbf{A}\mathbf{x}\|_{\infty}$  so that  $\mathbf{A}(\mathcal{S}) \subset [-M, M]^d$ , and so,  $\mathbf{A}$  is a  $\rho$ -JL embedding of  $\mathcal{S}$  into  $[-M, M]^d$ . By Proposition 6, there exists a ReLU NN architecture with at most

$$\mathcal{E} = (p + C_1) \left( 2 \left\lceil \frac{LM\sqrt{d}}{(1-\rho)\epsilon} \right\rceil + 1 \right)^d \text{ edges,}$$

$$\mathcal{N} = C_2 \left( 2 \left\lceil \frac{LM\sqrt{d}}{(1-\rho)\epsilon} \right\rceil + 1 \right)^d + p \text{ nodes}$$

$$\text{and } \mathcal{L} = \lceil \log_2(d+1) \rceil + 2 \text{ layers}$$

which can  $\epsilon$ -approximate any  $\frac{L}{1-\rho}$ -Lipschitz function  $g : [-M, M]^d$ . Finally, by applying Theorem 1b, we have that there exists a ReLU NN architecture with at most

$$\mathcal{E} + Dd = (p + C_1) \left( 2 \left\lceil \frac{LM\sqrt{d}}{(1-\rho)\epsilon} \right\rceil + 1 \right)^d \text{ edges,}$$

$$\mathcal{N} + D = C_2 \left( 2 \left\lceil \frac{LM\sqrt{d}}{(1-\rho)\epsilon} \right\rceil + 1 \right)^d + p + D \text{ nodes}$$

$$\text{and } \mathcal{L} + 1 = \lceil \log_2(d+1) \rceil + 3 \text{ layers}$$

which can  $\epsilon$ -approximate any  $L$ -Lipschitz function  $f : \mathcal{S} \rightarrow \mathbb{R}^p$ , as desired.

## F PROOF OF THEOREM 3

Let  $f : \mathcal{S} \rightarrow \mathbb{R}^p$  be the target function to approximate. By Proposition 4b, we have that there exists a matrix  $\mathbf{A} \in \mathbb{R}^{d \times D}$  in the form  $\mathbf{M}\mathbf{D}$  where  $\mathbf{M}$  is a partial circulant matrix and  $\mathbf{D}$  is a diagonal matrix with  $\pm 1$  entries such that  $\mathbf{A}$  is a  $\rho$ -JL embedding of  $\mathcal{S}$  with

$$d \gtrsim \rho^{-2} \log(4D + 4d) \log \mathcal{N}(U_{\mathcal{S}}, \|\cdot\|_2, \frac{\rho}{4\sqrt{3D}}).$$

Let  $M = \sup_{\mathbf{x} \in \mathcal{S}} \|\mathbf{A}\mathbf{x}\|_{\infty}$  so that  $\mathbf{A}(\mathcal{S}) \subset [-M, M]^d$ , and so,  $\mathbf{A}$  is a  $\rho$ -JL embedding of  $\mathcal{S}$  into  $[-M, M]^d$ .

By Lemma 1, there exists an  $\frac{L}{1-\rho}$ -Lipschitz function  $g : [-M, M]^d \rightarrow \mathbb{R}^p$  such that  $g(\mathbf{A}\mathbf{x}) = f(\mathbf{x})$  for all  $\mathbf{x} \in \mathcal{S}$ . Let  $g_i : [-M, M]^d \rightarrow \mathbb{R}$  be the  $i$ -th coordinate of  $g$ . Let  $\tilde{g}_i : [-1, 1]^d \rightarrow \mathbb{R}$  be defined by  $\tilde{g}_i(\mathbf{y}) = g_i(M\mathbf{y})$  for all  $\mathbf{y} \in [-1, 1]^d$ . Note that each  $g_i$  is  $\frac{L}{1-\rho}$ -Lipschitz, and so, each  $\tilde{g}_i$  is  $\frac{LM}{1-\rho}$ -Lipschitz,

Then, by Proposition 7, for each  $\tilde{g}_i$ , there exists a CNN  $\tilde{g}_i^{(CNN)}$  with  $O(N)$  residual blocks, each of which has depth  $O(\log N)$  and  $O(1)$  channels, and whose filter size is at most  $K$  such that  $\|\tilde{g}_i - \tilde{g}_i^{(CNN)}\|_{\infty} \leq \tilde{O}(N^{-1/d})$ .

Now, we construct a CNN to approximate  $f$  as follows. First, we implement the map  $\mathbf{x} \mapsto \frac{1}{M}\mathbf{A}\mathbf{x}$  using the same 4 layer Resnet CNN described in the proof of Theorem 1c. Then, we pass the output of that Resnet CNN into  $p$  parallel CNNs which implement  $\tilde{g}_i^{(CNN)}$  for  $i = 1, \dots, p$ . The output of the  $i$ -th of these parallel CNNs is  $\tilde{g}_i^{(CNN)}(\frac{1}{M}\mathbf{A}\mathbf{x})$ , which is an  $\tilde{O}(N^{-1/d})$ -approximation of  $\tilde{g}_i(\frac{1}{M}\mathbf{A}\mathbf{x}) = g_i(\mathbf{A}\mathbf{x}) = f_i(\mathbf{x})$ . Hence, the constructed CNN is a  $\tilde{O}(N^{-1/d})$ -approximation of  $f$ .

The CNN which implements the map  $\mathbf{x} \mapsto \frac{1}{M}\mathbf{A}\mathbf{x}$  needs  $O(1)$  residual blocks, each of which has depth  $O(1)$  and  $O(1)$  channels. Each of the  $p$  parallel CNNs which implement the  $\tilde{g}_i^{(CNN)}$ 's have  $O(N)$  residual blocks, each of which has depth  $O(\log N)$  and  $O(1)$  channels. So the overall network to approximate  $f$  has  $O(pN)$  residual blocks, each of which has depth  $O(\log N)$  and  $O(1)$  channels.

## G PROOF OF PROPOSITION 9

By the sin  $\Theta$  theorem (Wedin, 1972), we have

$$\left\| \frac{x_1}{\|x_1\|} - \frac{x_2}{\|x_2\|} \right\| \leq \frac{\|x_1 k_1^T - x_2 k_2^T\|}{\|x_1\| \|k_1\|} \quad (7)$$

and

$$\left\| \frac{k_1}{\|k_1\|} - \frac{k_2}{\|k_2\|} \right\| \leq \frac{\|x_1 k_1^T - x_2 k_2^T\|}{\|x_2\| \|k_2\|}. \quad (8)$$

Let us find a set whose covering number is easy to compute while containing the unit secant  $U_{\mathcal{Y}}$  as a subset

$$\begin{aligned} & \left\{ \frac{y_1 - y_2}{\|y_1 - y_2\|}, y_1, y_2 \in \mathcal{Y} \right\} = \left\{ \frac{x_1 \otimes k_1 - x_2 \otimes k_2}{\|x_1 \otimes k_1 - x_2 \otimes k_2\|}, x_i = \Phi u_i, k_i = \Psi v_i, i = 1, 2 \right\} \\ &= \left\{ \frac{x_1 \otimes k_1 - (\frac{\|x_1\|}{\|x_2\|} x_2) \otimes k_1}{\|x_1 \otimes k_1 - x_2 \otimes k_2\|} + \frac{(\frac{\|x_1\|}{\|x_2\|} x_2) \otimes k_1 - x_2 \otimes k_2}{\|x_1 \otimes k_1 - x_2 \otimes k_2\|}, x_i = \Phi u_i, k_i = \Psi v_i, i = 1, 2 \right\} \\ &\subseteq \left\{ \frac{x_1 \otimes k_1 - (\frac{\|x_1\|}{\|x_2\|} x_2) \otimes k_1}{\|x_1 \otimes k_1 - x_2 \otimes k_2\|}, x_i = \Phi u_i, k_i = \Psi v_i, i = 1, 2 \right\} \\ &+ \left\{ \frac{(\frac{\|x_1\|}{\|x_2\|} x_2) \otimes k_1 - x_2 \otimes k_2}{\|x_1 \otimes k_1 - x_2 \otimes k_2\|}, x_i = \Phi u_i, k_i = \Psi v_i, i = 1, 2 \right\}. \end{aligned}$$

For the first set in the sum, by using (3) and (7), we have

$$\begin{aligned}
& \left\{ \frac{x_1 \otimes k_1 - (\frac{\|x_1\|}{\|x_2\|} x_2) \otimes k_1}{\|x_1 \otimes k_1 - x_2 \otimes k_2\|}, x_i = \Phi u_i, k_i = \Psi v_i, i = 1, 2 \right\} \\
& \subseteq \left\{ t \cdot \frac{x_1 \otimes k_1 - (\frac{\|x_1\|}{\|x_2\|} x_2) \otimes k_1}{\|x_1 k_1^T - x_2 k_2^T\|}, t \in [0, L], x_i = \Phi u_i, k_i = \Psi v_i, i = 1, 2 \right\} \\
& \subseteq \left\{ t \cdot \frac{x_1 \otimes k_1 - (\frac{\|x_1\|}{\|x_2\|} x_2) \otimes k_1}{\|x_1 - \frac{\|x_1\|}{\|x_2\|} x_2\| \|k_1\|}, t \in [0, L], x_i = \Phi u_i, k_i = \Psi v_i, i = 1, 2 \right\} \\
& \subseteq \left\{ \left( \sqrt{t} \cdot \frac{x_1 - \frac{\|x_1\|}{\|x_2\|} x_2}{\|x_1 - \frac{\|x_1\|}{\|x_2\|} x_2\|} \right) \otimes \left( \sqrt{t} \cdot \frac{k_1}{\|k_1\|} \right), t \in [0, L], x_i = \Phi u_i, k_i = \Psi v_i, i = 1, 2 \right\}
\end{aligned}$$

The covering number with  $\epsilon$  balls of the set  $\left\{ \sqrt{t} \cdot \frac{x_1 \otimes k_1 - (\frac{\|x_1\|}{\|x_2\|} x_2) \otimes k_1}{\|x_1 \otimes k_1 - (\frac{\|x_1\|}{\|x_2\|} x_2) \otimes k_1\|}, t \in [0, L] \right\}$  is  $\left( \frac{3\sqrt{L}}{\epsilon} \right)^n$ , and that for the set  $\left\{ \sqrt{t} \cdot \frac{k_1}{\|k_1\|}, t \in [0, L] \right\}$  is  $\left( \frac{3\sqrt{L}}{\epsilon} \right)^m$ . So the covering number with  $\epsilon$  balls of  $S$  is

$$\left( \frac{6L}{\epsilon} \right)^n + \left( \frac{6L}{\epsilon} \right)^m.$$

The same argument holds for the second set in the sum.

## H PROOF OF PROPOSITION 10

By definition,

$$U_{\mathcal{Y}} = \left\{ \frac{y_1 - y_2}{\|y_1 - y_2\|}, y_1, y_2 \in \mathcal{Y} \right\} = \left\{ \frac{P_{\Omega}(X_1 - X_2)}{\|P_{\Omega}(X_1 - X_2)\|}, X_1, X_2 \in \mathcal{Y} \right\}$$

Since  $y_1 - y_2 = P_{\Omega}(X_1 - X_2)$  and  $X$

$$\left\{ \frac{P_{\Omega}(X_1 - X_2)}{\|P_{\Omega}(X_1 - X_2)\|}, X_1, X_2 \in \mathcal{Y} \right\} \subseteq \left\{ t \cdot P_{\Omega} \frac{(X_1 - X_2)}{\|X_1 - X_2\|_F}, t \in [0, L], X_1, X_2 \in \mathcal{Y} \right\}$$

Notice that  $\frac{(X_1 - X_2)}{\|X_1 - X_2\|_F}$  are matrices of unit Frobenius norm with rank at most  $2r$ . By Lemma 3.1 in (Candes & Plan, 2011), they form a set whose covering number is at most  $\left( \frac{9}{\delta} \right)^{r(m+n+1)}$ .