# A Sampling Theory Perspective on Activations for Implicit Neural Representations

**Hemanth Saratchandran** [*1] **Sameera Ramasinghe** [*2] **Violetta Shevchenko** [2] **Alexander Long** [2] **Simon Lucey** [1]

## Abstract

Implicit Neural Representations (INRs) have gained popularity for encoding signals as compact, differentiable entities. While commonly using techniques like Fourier positional encodings or non-traditional activation functions (e.g., Gaussian, sinusoid, or wavelets) to capture high-frequency content, their properties lack exploration within a unified theoretical framework. Addressing this gap, we conduct a comprehensive analysis of these activations from a sampling theory perspective. Our investigation reveals that $\mathrm{sinc}$ activations—previously unused in conjunction with INRs—are theoretically optimal for signal encoding. Additionally, we establish a connection between dynamical systems and INRs, leveraging sampling theory to bridge these two paradigms.

## 1. Introduction

Recently, the concept of representing signals as Implicit Neural Representations (INRs) has garnered widespread attention across various problem domains (Mildenhall et al., 2021; Li et al., 2023; Büsching et al., 2023; Peng et al., 2021; Strümpler et al., 2022). This surge in popularity can be attributed to the remarkable capability of INRs to encode high-frequency signals as continuous representations. Unlike conventional neural networks, which typically process and convert sparse, high-dimensional signals (such as images, videos, text) into label spaces (e.g., one-hot encodings, segmentation masks, text corpora), INRs specialize in encoding and representing signals by consuming low-dimensional coordinates.

However, a significant challenge in representing signals using neural networks is the presence of spectral bias (Rahaman et al., 2019). Neural networks inherently tend to favor learning functions with lower frequencies, which can hinder their ability to capture high-frequency information. To address this challenge, a common approach involves projecting low-dimensional coordinates into a higher-dimensional space through positional encodings (Zheng et al., 2022; Tancik et al., 2020). Prior research has demonstrated that incorporating positional encodings allows INRs to achieve high-rank representations, enabling them to capture fine details (Zheng et al., 2022). Nevertheless, positional encodings have a critical limitation – they struggle to maintain smooth gradients, which can be problematic for optimization (Saratchandran et al., 2023). To overcome this limitation, non-traditional activations, such as sinusoids (Sitzmann et al., 2020), Gaussians (Ramasinghe & Lucey, 2022), and wavelets (Saragadam et al., 2023), have emerged as effective alternatives. These unconventional activations facilitate encoding higher frequencies while preserving smooth gradients, and as shown in prior research (Saratchandran et al., 2023; Chng et al., 2024; Saratchandran et al., 2024), they are remarkably stable with respect to various optimization algorithms.

Until now, prior research that delved into the analysis of activations in INRs has primarily been tied to the specific activations proposed in their respective studies. For instance, (Sitzmann et al., 2020) introduced sinusoidal activations and demonstrated their shift invariance and favorable properties for learning natural signals. (Ramasinghe & Lucey, 2022) explored Gaussian activations, showcasing their high Lipschitz constants that enable INRs to capture sharp variations. More recently, wavelet-based activations (Saragadam et al., 2023) were introduced, highlighting their spatial-frequency concentration and suitability for representing images. However, this fragmented approach has obscured the broader picture, making it difficult to draw connections and conduct effective comparisons among these activations. In contrast, our research unveils a unified theory of INR activations through the lens of sampling theory. Specifically, we show that, under mild conditions, activations in INRs can be considered as generator functions that facilitate the reconstruction of a given signal from sparse samples. Leveraging this insight, we demonstrate that activations in the

[*]Equal contribution [1]University of Adelaide, Australia [2]Amazon, Australia. Correspondence to: Hemanth Saratchandran <hemanth.saratchandran@adelaide.edu.au>, Sameera Ramasinghe <sameera.ramasinghe@adelaide.edu.au>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. 

form of $\frac{sin(x)}{x}$ (known as the sinc function) theoretically enable INRs to optimally reconstruct a given signal while preserving smooth gradients. To the best of our knowledge, sinc activations have not been used with INRs previously. Furthermore, we validate these insights in practical scenarios across tasks involving images and neural radiance fields (NeRF).

The proficiency of sinc-activated INRs in signal reconstruction suggests an exciting possibility: the effective modeling of complex dynamical systems using these activations. We explore this idea by focusing on chaotic dynamical systems, noting the similarity between dynamical systems and INRs when approached from a signal processing lens. Dynamical systems can be seen as multi-dimensional signals evolving over time, resembling the task of reconstructing multi-dimensional signals from discrete samples. INRs, with their objective of encoding and reconstructing continuous signals from discrete coordinates and samples, share a similar goal. Drawing inspiration from this connection, we establish parallels between dynamical systems and INRs, using sampling theory to bridge these two paradigms. Our research not only demonstrates the superior performance of sinc-activated INRs in modeling dynamical systems but also provides a theoretical explanation for this advantage.

## 2. Related Work

**INRs.** INRs, pioneered by (Mildenhall et al., 2021), have gained prominence as an effective architecture for signal reconstruction. Traditionally, such architectures employed activations such as ReLU and Sigmoid. However, these activations suffer from spectral bias, limiting their effectiveness in capturing high-frequency content (Rahaman et al., 2019). To overcome this limitation, (Mildenhall et al., 2021) introduced a positional embedding layer to enhance high-frequency modeling. Meanwhile, (Sitzmann et al., 2020) proposed SIREN, a sinusoidal activation that eliminates the need for positional embeddings but exhibits instability with random initializations. In contrast, (Ramasinghe & Lucey, 2022) introduced Gaussian-activated INRs, showcasing robustness to various initialization schemes. More recently, wavelet activations were proposed by (Saragadam et al., 2023) with impressive performance. Yet, the theoretical optimality of these activation functions in the context of signal reconstruction has largely eluded investigation. In this study, we aim to address this gap by examining the selection of activation functions through the lens of sampling theory.

**Data driven dynamical systems modeling.** Numerous approaches have been explored for the data-driven discovery of dynamical systems, employing various techniques. These methodologies include nonlinear regression (Voss et al., 1999), empirical dynamical modeling (Ye et al., 2015), normal form methods (Majda et al., 2009), spectral analysis (Giannakis & Majda, 2012), Dynamic Mode Decomposition (DMD) (Schmid, 2010; Kutz et al., 2016), as well as compressed sensing and sparse regression within a library of candidate models (Reinbold et al., 2021; Wang et al., 2011; Naik & Cochran, 2012; Brunton et al., 2016). Additionally, reduced modeling techniques like Proper Orthogonal Decomposition (POD) (Holmes et al., 2012; Kirby, 2001; Sirovich, 1987; Lumley, 1967), both local and global POD methods (Schmit & Glauser, 2004; Sahyoun & Djouadi, 2013), and adaptive POD methods (Singer & Green, 2009; Peherstorfer & Willcox, 2015) have been widely applied in dynamical system analysis. Koopman operator theory in conjunction with DMD methods has also been utilized for system identification (Budišić et al., 2012; Mezić, 2013).

## 3. A sampling perspective on INRs

In this section, we present our primary theoretical insights, showcasing how sampling theory offers a fresh perspective on understanding the optimality of activations in INRs.

### 3.1. Implicit Neural Representations

We consider INRs of the following form: Consider an $L$-layer network, $F_L$, with widths $\{n_0, \dots, n_L\}$. The output at layer $l$, denoted $f_l$, is given by

$$f_l(x) = \begin{cases} x, & \text{if } l = 0 \\ \phi(W_l F_{l-1} + b_l), & \text{if } l \in \{1, \dots, L-1\} \\ W_{L-1} F_{L-1} + b_L, & \text{if } l = L \end{cases} \quad (1)$$

where $W_l \in \mathbb{R}^{n_l \times n_{l-1}}$, $b_l \in \mathbb{R}^{n_l}$ are the weights and biases respectively of the network, and $\phi$ is a non-linear activation.

### 3.2. Classical sampling theory

Sampling theory considers bandlimited signals, which are characterized by a limited frequency range. Formally, for a continuous signal denoted as $f$, being bandlimited to a maximum frequency of $\Omega$ implies that its Fourier transform, represented as $\widehat{f}(s)$, equals zero for all $|s|$ values greater than $\Omega$. If we have an $\Omega$-bandlimited signal $f$ belonging to the space $L^2(\mathbb{R})$, then the Nyquist-Shannon sampling theorem (as referenced in (Zayed, 2018)) provides a way to represent this signal as $f(x) = \sum_{n=-\infty}^{\infty} f\left(\frac{n}{2\Omega}\right) \text{sinc}\left(2\Omega(x - \frac{n}{2\Omega})\right)$, where $\text{sinc}(x) := \frac{sin(\pi x)}{\pi x}$ for $x \neq 0$ and $\text{sinc}(0) := 1$, and the equality means converges in the $L^2$ sense. Essentially, by sampling the signal at regularly spaced points defined by $\frac{n}{2\Omega}$ for all integer values of $n$, and using shifted sinc functions, we can reconstruct the original signal. However, this requires us to sample at a rate of at least $2\Omega$-Hertz (Zayed, 2018).

In theory, perfect reconstruction necessitates an infinite num-

ber of samples, which is impractical in real-world scenarios. It's crucial to acknowledge that the sampling theorem is an idealization, not universally applicable to real signals due to their non-bandlimited nature. However, as natural signals often exhibit dominant frequency components at lower energies, we can effectively approximate the original signal by projecting it into a finite-dimensional space of bandlimited functions, enabling robust reconstruction.

### 3.3. Optimal Activations via Riesz Sampling

In the previous section, we discussed how an exact reconstruction of a bandlimited signal could be achieved via a linear combination of shifted sinc functions. Thus, it is intriguing to explore if an analogous connection can be drawn to INRs. We will take a general approach and consider spaces of the form

$$V(F) = \left\{ s(x) = \sum_{k\in\mathbb{Z}} a(k)F(x-k) : a \in l^2(\mathbb{R}) \right\}, \quad (2)$$

where $l^2(\mathbb{R})$ denotes the Hilbert space of square summable sequences over the integers $\mathbb{Z}$. The reader who is not familiar with the sequence space $l^2(\mathbb{R})$ can consult App. A.1 for its definition. The space $V(F)$ should be seen as a generalisation of the space of bandlimited functions occurring in the Shannon sampling theorem.

**Definition 3.1.** The family of translates $\{F_k := F(x-k)\}_{k\in\mathbb{Z}}$ is a Riesz basis for $V(F)$ if the following two conditions hold:

1. $A||a||_{l^2}^2 \leq \left|\left| \sum_{k\in\mathbb{Z}} a(k)F_k \right|\right|^2 \leq B||a||_{l^2}^2, \forall a(k) \in l^2(\mathbb{R}).$
2. $\sum_{k\in Z} F(x-k) = 1, \forall x \in \mathbb{R}$ **(PUC)**

where in condition 1 we have that $A$ and $B$ are constants greater than zero. Observe that if $s = \sum_k a(k)F_k = 0$ then the lower inequality in condition 1 implies $a(k) = 0$ for all $k$. In other words, the basis functions $F_k$ are linearly independent, which in turn implies each signal $s \in V(F)$ is uniquely determined by its coefficient sequence $a(k) \in l^2(\mathbb{R})$. The upper inequality in 1. implies that the $L^2$ norm of a signal $s \in V(F)$ is finite, implying that $V(F)$ is a subspace of $L^2(\mathbb{R})$.

Condition 2 in Defn. 3.1 is known as the *partition of unity condition* (**PUC**). It allows the capability of approximating a signal $s \in V(F)$ as closely as possible by selecting a sample step that is sufficiently small. This can be seen as a generalisation of the Nyquist criterion, where in order to reconstruct a bandlimited signal, a sampling step of less than $\frac{\pi}{2\omega_{\max}}$ must be chosen, where $\omega_{\max}$ is the highest frequency present in the signal $s$ (Zayed, 2018; Unser, 2000).

**Definition 3.2.** The family of translates $\{F_k = F(x-k)\}_{k\in\mathbb{Z}}$ is a weak Riesz basis for $V(F)$ if condition 1 from Defn. 3.1 holds but condition 2 does not.

The following proposition considers activations in INRs and the sinc function. Specifically, we show that sinc forms a Riesz basis, Gaussian and wavelets form weak Riesz bases, and ReLU or Sinusoid does not form Riesz/weak Riesz bases. The proof is given in app. A.1.

**Proposition 3.3.** *1. Let $F(x) = \mathrm{sinc}(x) = \frac{\sin(x)}{x}$ then the family $\{\mathrm{sinc}(x-k)\}_{k\in\mathbb{Z}}$ forms a Riesz basis where $V(\mathrm{sinc})$ is the space of signals with bandlimited frequency.*

2. *Let $F(x) = G_s(x) := e^{-x^2/s^2}$, for some fixed $s > 0$, the family $\{G_s(x-k)\}_{k\in\mathbb{Z}}$ forms a weak Riesz basis for the space $V(G_s)$ but not a Riesz basis. In this case $V(G_s)$ can be interpreted as signals whose Fourier transform has Gaussian decay, where the rate of decay will depend on $s$.*

3. *Let $F(x) = \Psi(x)$ denote a wavelet. In general wavelets form a weak Riesz basis but not all form a Riesz basis.*

4. *Let $F(x) = ReLU(x)$, the family $\{ReLU(x-k)\}_{k\in\mathbb{Z}}$ does not form a Riesz/weak Riesz basis as it violates condition 1 from Defn. 3.1.*

5. *Let $F(x) = \sin(\omega x)$, for $\omega$ a fixed frequency parameter, the family $\{sin(\omega(x-k))\}_{k\in\mathbb{Z}}$ does not form a Riesz/weak Riesz basis as it violates condition 1 from Defn. 3.1.*

Interestingly, to fit INRs into the above picture, observe that the elements in $V(F)$ that are finite sums can be represented by INRs with activation $F$ (this is explicitly proved in the next theorem, see also appendix A.1). Thus, it follows that signals in $V(F)$ that have an infinite number of non-zero summands can be approximated by INRs, the proof can be found in the app. A.2.1.

**Theorem 3.4.** *Suppose the family of functions $\{F(x-k)\}_{k\in\mathbb{Z}}$ forms a weak Riesz basis for the space $V(F)$. Let $g$ be a signal in $V(F)$ and let $\epsilon > 0$ be given. Then there exists a 2-layer INR $f$, with a parameter set $\theta$, $F$ as the activation, and $n(\epsilon)$ neurons in the hidden layer, such that*

$$||f(\theta) - g||_{L^2} < \epsilon.$$

The primary limitation of Theorem 3.4 lies in its applicability solely to signals within the domain of $V(F)$. This prompts us to inquire whether Riesz bases can be employed to approximate arbitrary $L^2$-functions, even those outside the confines of $V(F)$. The significance of posing this question lies in the potential revelation that, if affirmed, INRs

can also approximate such signals. This would, in turn, demonstrate the universality of $F$-activated INRs within the space of $L^2(\mathbb{R})$ functions.

We now show, in order to be able to approximate arbitrary signals in $L^2(\mathbb{R})$ the partition of unity condition, condition 2 of Defn. 3.1, plays a key role. To analyse this situation, we introduce the scaled signal spaces.

**Definition 3.5.** For a fixed $\Omega > 0$, let

$$V_\Omega(F) = \left\{ s_\Omega = \sum_{k\in\mathbb{Z}} a_\Omega(k) F(\frac{x}{\Omega} - k) : a \in l^2(\mathbb{R}) \right\}.$$

We call $V_\Omega(F)$ an $\Omega$-scaled signal space.

*Example* 3.6. The canonical example of an $\Omega$-scaled signal space is given by taking $F(x) = \mathrm{sinc}(2\Omega x)$. In this case $V_\Omega(F)$ is the space of $\Omega$-bandlimited signals.

The difference between $V_\Omega(F)$ and $V(F)$ is that in the former the basis functions are scaled by $\Omega$. Previously, we remarked that one of the issues in applying Thm. 3.4 is that it does not provide a means for approximating general signals in $L^2(\mathbb{R})$ by INRs. What we wish to establish now is that given an arbitrary signal $s \in L^2(\mathbb{R})$ and an approximation error $\epsilon > 0$, if $\{F(x-k\}_{k\in\mathbb{Z}}$ is a Riesz basis then there exists a scale $\Omega(\epsilon)$, that depends on $\epsilon$, such that the scaled signal space $V_{\Omega(\epsilon)}(F)$ can approximate $s$ to within $\epsilon$ in the $L^2$-norm. We will follow the approach taken by (Unser, 2000) and give a brief overview of how to proceed. More details can be found in app. A.

In order to understand how we can reconstruct a signal to within a given error using $V_\Omega(F)$, we follow the strategy of (Unser, 2000). We define the approximation operator $A_\Omega : L^2(\mathbb{R}) \to V_\Omega(F)$ by

$$A_\Omega(s(x)) = \sum_{k\in\mathbb{Z}} \left( \int_{\mathbb{R}} s(y) \widetilde{F}(\frac{y}{\Omega} - k) \frac{dy}{\Omega} \right) F(\frac{x}{\Omega} - k) \quad (3)$$

where $\widetilde{F}$ is a suitable analysis function from a fixed test space. We will not go into the details of how to construct $\widetilde{F}$ but for now will simply assume such an $\widetilde{F}$ exists and remark that its definition depends on $F$. For details on how to construct $\widetilde{F}$ we refer the reader to appendix sec. A.1.1. The quantity $\int s(y)\widetilde{F}(\frac{y}{\Omega} - k)\frac{dy}{\Omega}$ is to be thought of as the coefficients $a_\Omega(k)$ in reconstructing $s$.

The approximation error is defined as

$$\epsilon_s(\Omega) = ||s - A_\Omega(s)||_{L^2}. \quad (4)$$

The goal is to understand how we can make the approximation error small by choosing $\Omega$ and the right analysis function $\widetilde{F}$.

The general approach to this problem via sampling theory, see (Unser, 2000) for details, is to proceed via the average approximation error:

$$\bar{\epsilon}_s(\Omega)^2 = \frac{1}{\Omega} \int_0^\Omega ||s(\cdot - \tau) - A_\Omega(s(\cdot - \tau))||^2_{L^2} d\tau. \quad (5)$$

Using Fourier analysis, see (Blu & Unser, 1999), it can be shown that $\bar{\epsilon}_s(\Omega)^2 = \int_{-\infty}^{+\infty} E_{\widetilde{F},F}(\Omega\xi)|\hat{s}(\xi)|^2 \frac{d\xi}{2\pi}$ where $\hat{s}$ denotes the Fourier transform of $s$ and $E_{\widetilde{F},F}$ is the error kernel defined by

$$E_{\widetilde{F},F}(\omega) = |1 - \hat{\widetilde{F}}(\omega)\hat{F}(\omega)|^2 + |\hat{\widetilde{F}}(\omega)|^2 \sum_{k\neq 0} |\hat{F}(\omega + 2\pi k)|^2 \quad (6)$$

where $\hat{F}$ and $\hat{\widetilde{F}}$ denote the Fourier transforms of $F$ and $\widetilde{F}$ respectively. Understanding the approximation properties of the shifted basis functions $F_k = F(x-k)$ comes down to analysing the error kernel $E_{\widetilde{F},F}$. The reason being is that the average error $\bar{\epsilon}_s(\Omega)^2$ is a good predictor of the true error $\epsilon_s(\Omega)^2$. This was shown in (Blu & Unser, 1999) and a statement of their theorem can be found in App. A.1.1 as Thm. A.1

Thm. A.1 shows that the dominant part of the approximation error $\epsilon_s(\Omega)$ is controlled by the average error $\bar{\epsilon}_s(\Omega)$. This means that in order to show that there exists a scale $\Omega$ such that the scaled signal space $V_\Omega(F)$ can be used to approximate $s \in L^2(\mathbb{R})$ up to any given error, it suffices to show that

$$\lim_{\Omega\to 0} E_{\widetilde{F},F} \to 0. \quad (7)$$

This is precisely where the partition of unity condition comes in. As shown in (Blu & Unser, 1999) if a family of shifted basis functions satisfies the partition of unity condition (**PUC**), condition 2 from Defn. 3.1, then the above limit holds meaning the error kernel vanishes. The interested reader can consult App. A.1.1 for details on how the argument proceeds.

From Prop. 3.3, we see that the sinc function has vanishing error kernel as the scale $\Omega \to 0$. However, a Gaussian does not necessarily have vanishing error kernel as $\Omega \to 0$.

We immediately obtain the following approximation result, proof can be found in App. A.2.1.

**Proposition 3.7.** *Let $s \in L^2(\mathbb{R})$ and $\epsilon > 0$. Assume the shifted functions $\{F(x-k)\}_{k\in\mathbb{Z}}$ form a Riesz basis for $V(F)$. Then there exists an $\Omega > 0$ and an $f_\Omega \in V_\Omega(F)$ such that*

$$||s - f_\Omega||_{L^2} < \epsilon. \quad (8)$$

Prop. 3.7 implies that the signal $s$ can be approximated by basis functions given by shifts of $F$ with bandwidth $1/\Omega$.

Using Prop. 3.7 we obtain a universal approximation result for neural networks employing Riesz bases as their activation functions, the proof can be found in app. A.2.1.

**Theorem 3.8.** *Let $s \in L^2(\mathbb{R})$ and $\epsilon > 0$. Assume the shifted functions $\{F(x-k)\}_{k\in\mathbb{Z}}$ form a Riesz basis for $V(F)$. Then there exists a 2-layer INR $\mathcal{N}$, with a parameter set $\theta$, $n(\epsilon)$ neurons in the hidden layer, and an $\Omega > 0$ such that*

$$||\mathcal{N}(\theta) - s||_{L^2} < \epsilon$$

*where $\mathcal{N}(\theta)$ employs $F_\Omega$ as its activation in the hidden layer, where $F_\Omega(x) = F(\frac{1}{\Omega}x)$.*

*Remark* 3.9. Thm. 3.8 shows why sinc being able to generate a Riesz basis is an optimal condition to satisfy for an activation function. Note that in general, any function in $L^2(\mathbb{R})$ that generates a Riesz basis will be optimal in this sense. Furthermore, out of the activations that practitioners in the ML community use, such as sinc, sine, Gaussian, tanh, ReLU, sigmoid, we find sinc is the optimal. For an overview of how the partition of unity condition, condition 2 of Defn. 3.1, plays a role in Thm. 3.8 see App. A.2.

Thm. 3.8 underscores the significance of the activation function within an INR when it comes to signal reconstruction in $L^2(\mathbb{R})$. Specifically, as exemplified in Prop. 3.3, it becomes evident that an INR equipped with a sinc activation function can achieve reconstructions of signals in $L^2(\mathbb{R})$ up to any accuracy, rendering it the optimal choice for the INR architecture. See App. C for the connection of the above analysis to the universal approximation theorem.

## 4. Extensions of the theory

**Results for deep networks:** The theoretical results from the previous section were initially demonstrated in the context of shallow networks to emphasize fundamental techniques. These results naturally extend to deep networks, as discussed in App. B.1.

**Other basis functions:** Various common basis functions are utilized for interpolation in existing literature. To contextualize our results alongside these alternatives, we direct the reader to App. B.2.

**Positional encoding:** For insights into how our theoretical contributions relate to positional encodings, please refer to App. 5.4.

## 5. Experiments

In this section, we aim to compare the performance of different INR activations. First, we focus on image and NeRF reconstructions and later move on to dynamical systems.

### 5.1. Image reconstruction

A critical problem entailed with INRs is that they are sensitive to the hyperparameters in activation functions (Ramasinghe & Lucey, 2022). That is, one has to tune the hyperparameters of the activations to match the spectral properties of the encoded signal. Here, we focus on the robustness of activation parameters when encoding different signals. To this end, we do a grid search and find the *single* best performing hyperparameter setting for *all* the images in a sub-sampled set of the DIV2K dataset (Agustsson & Timofte, 2017) released by (Tancik et al., 2020). For example, for sinc activations, we experiment with different bandwidth parameters, each time fixing it across the entire dataset. Then, we select the bandwidth parameter that produced the best results. Since all the compared activations contain a tunable parameter, we perform the same for all the activations and find the best parameters for each. This dataset contains images of varying spectral properties: 32 images of *Text* and *Natural scenes*, each. We train with different sampling rates and test against the full ground truth image. The PSNR plots are shown in Fig. 1. As depicted, sinc activation performs better or on-par with other activations. We use 4-layer networks with 256 width for these experiments.

| Activation | PSNR | SSIM |
|---|---|---|
| Gaussian | 31.13 | 0.947 |
| Sinusoid | 28.96 | 0.933 |
| Wavelet | 30.33 | 0.941 |
| Sinc | **31.37** | **0.947** |

*Table 1.* **Quantitative comparison in novel view synthesis on the real synthetic dataset (Mildenhall et al., 2021).** sinc-INRs perform on-par with other activations.

### 5.2. Neural Radiance Fields

NeRFs are one of the key applications of INRs, popularized by (Mildenhall et al., 2021). Thus, we evaluate the performance of sinc-INRs in this setting. Table. 1 demonstrates quantitative results. We observed that all the activations perform on-par with NeRF reconstructions with proper hyperparameter tuning, where sinc outperformed the rest marginally.

### 5.3. Dynamical systems

It is intriguing to see if the superior signal encoding properties of sinc-INRs (as predicted by the theory) would translate to a clear advantage in a challenging setting. To this end, we choose dynamical (chaotic) systems as a test bed.

Dynamical systems can be defined in terms of a time dependant state space $\mathbf{x}(t) \in \mathbb{R}^D$ where the time evolution of $\mathbf{x}(t)$ can be described via a differential equation,

$$\frac{d\mathbf{x}(t)}{dt} = f(\mathbf{x}(t), \alpha), \tag{9}$$

| | Rossler | | |
|---|---|---|---|
| Activation | $n = 0.1$ | $n = 0.5$ | $n = 1.$ |
| Baseline | 42.1 | 29.8 | 22.3 |
| Gaussian | 46.6 | 37.1 | 33.6 |
| Sinusoid | 45.1 | 36.6 | 32.1 |
| Wavelet | 40.3 | 35.9 | 30.9 |
| Sinc | **48.9** | **42.8** | **38.5** |
| | Lorenz | | |
| Activation | $n = 0.1$ | $n = 0.5$ | $n = 1.$ |
| Baseline | 43.8 | 28.2 | 20.3 |
| Gaussian | 45.7 | 39.1 | 35.7 |
| Sinusoid | 42.1 | 37.4 | 30.3 |
| Wavelet | 38.2 | 37.3 | 31.8 |
| Sinc | **46.2** | **40.9** | **39.3** |

*Table 2.* SINDy reconstructions (PSNR) with different noise levels ($n$ =standard deviation of the Gaussian noise).

where $f$ is a non-linear function and $\alpha$ are a set of system parameters. The solution to the differential equation 9 gives the time dynamics of the state space $\mathbf{x}(t)$. In practice, we only have access to discrete measurements $[\mathbf{y}(t_1), \mathbf{y}(t_2), \dots \mathbf{y}(t_Q)]$ where $\mathbf{y}(t) = g(\mathbf{x}(t)) + \eta$ and $\{t_n\}_{n=1}^{Q}$ are discrete instances in time. Here, $g(\cdot)$ can be the identity or any other non-linear function, and $\eta$ is noise. Thus, the central challenge in modeling dynamical systems can be considered as recovering the characteristics of the state space from such discrete observations.

**We note that modeling dynamical systems and encoding signals using INRs are analogous tasks. That is, modeling dynamical systems can be interpreted as recovering characteristics of a particular system via measured physical quantities over time intervals. Similarly, INRs are used to recover a signal given discrete samples.**

#### 5.3.1. Discovering the latent dynamics

In practical scenarios, we often encounter limitations in measuring all the variables influencing a system's dynamics. When only partial measurements are available, deriving a closed-form model for the system becomes challenging. However, Takens' Theorem (refer to App. E) offers a significant insight. It suggests that under certain conditions, augmenting partial measurements with delay embeddings can produce an attractor diffeomorphic to the original one. This approach is remarkably powerful, allowing for the discovery of complex system dynamics from a limited set of variables.

**Time Delay Embedding.** To implement this, we start with discrete time samples of an observable variable $y(t)$. We construct a Hankel matrix $\mathbf{H}$ by augmenting these samples as *delay embeddings* in each row:

$$\mathbf{H} = \begin{bmatrix} y_1(t_1) & y_1(t_2) & \dots & y_1(t_n) \\ y_1(t_2) & y_1(t_3) & \dots & y_1(t_{n+1}) \\ \vdots & \vdots & \ddots & \vdots \\ y_1(t_m) & y_1(t_{m+1}) & \dots & y_1(t_{m+n+1}) \end{bmatrix}. \quad (10)$$

According to Takens' Theorem, the dominant eigenvectors of this Hankel matrix encapsulate dynamics that are diffeomorphic to the original attractor. For our experiment, we utilize systems such as the Vanderpol, Limit cycle attractor, Lorenz, and Duffing equations (see F). We generate $5000$ samples, spanning from $0$ to $100$, to form the Hankel matrix. Subsequently, we extract its eigenvectors and plot them to visualize the surrogate attractor that mirrors the original attractor. To assess the method's robustness against noise, we introduce noise into the $y(t)$ samples from a uniform distribution $\eta \sim U(-n, n)$, varying $n$. To demonstrate the efficacy of INRs in this context, we employ a sinc-INR to encode the original measurements as a continuous signal. Initially, we train a sinc-INR using discrete pairs of $t$ and $y(t)$ as inputs and labels. Then, we use the sampled values from the INR as a surrogate signal to create the Hankel matrix, which yields robust results. Interestingly, the continuous reconstruction from the sinc-INR requires sparser samples (with $n\tau = 0.2$)), thus overcoming a restrictive condition typically encountered in this methodology. The results, as depicted in Fig. 2 and Fig. 3, clearly demonstrate that sinc-INRs can accurately recover the dynamics of a system from partial, noisy, random, and sparse observations. In contrast, the performance of classical methods deteriorates under these conditions, underscoring the advantage of the sinc-INR approach in handling incomplete and imperfect data.

#### 5.3.2. Discovering governing equations

The SINDy algorithm (Brunton et al., 2016) is designed to deduce the governing equations of a dynamical system from discrete observations of its variables. Consider observing the time dynamics of a $D$-dimensional variable $\mathbf{y}(t) = [y_1(t), \dots, y_D(t)]$. For observations at $N$ time stamps, we construct the matrix $\mathbf{Y} = [\mathbf{y}(t_1), \mathbf{y}(t_2), \dots \mathbf{y}(t_N)]^T \in \mathbb{R}^{N \times D}$. The initial step in SINDy involves computing $\dot{\mathbf{Y}} = [\mathbf{y}(\dot{t}_1), \mathbf{y}(\dot{t}_2), \dots \mathbf{y}(\dot{t}_N)]^T \in \mathbb{R}^{N \times D}$, achieved either through finite difference or continuous approximation techniques. Subsequently, an augmented library $\Theta(\mathbf{Y})$ is constructed, composed of predefined candidate nonlinear functions of $\mathbf{Y}$'s columns, encompassing constants, polynomials, and trigonometric terms, e.g.,

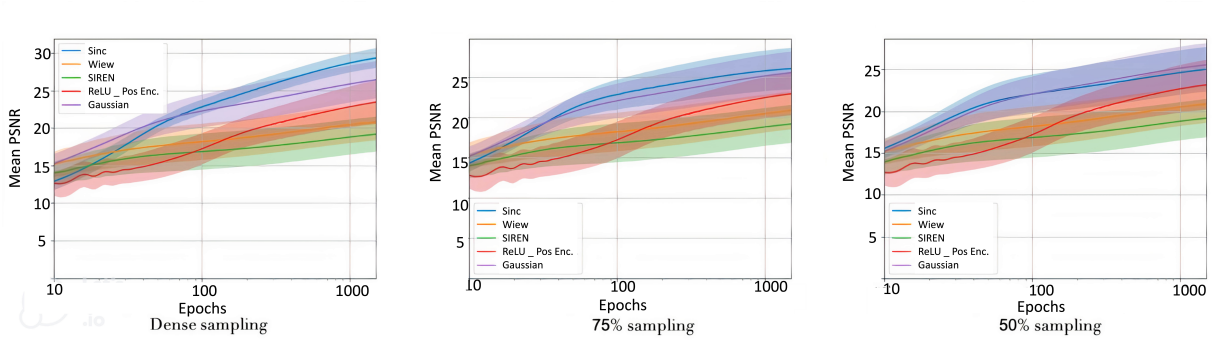


*Figure 1.* **Comparison of Image reconstruction across different INRs over DIVK dataset.** We run a grid search to find the optimal parameters for each INR. Note that a single optimal parameter setting is used for each activation, across all the images in the dataset.

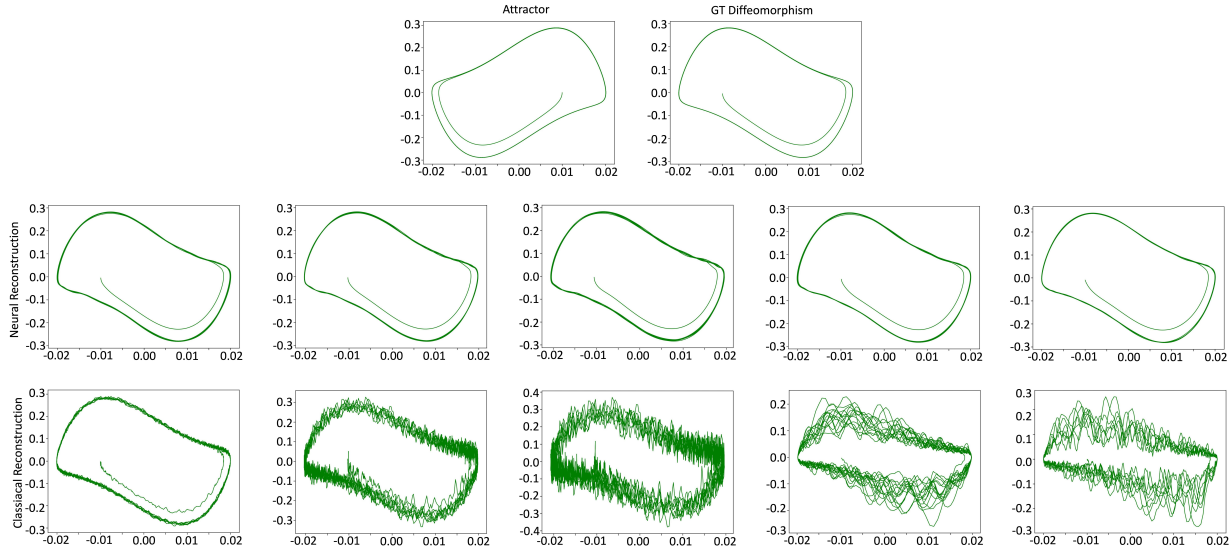


*Figure 2.* Discovering the dynamics from partial observations. We use the Vanderpol system (see App. F) for this illustration. *Top row:* the original attractor and the diffeomorphism obtained by the SVD decomposition of the Hankel matrix (see Sec. 5.3.1) without noise. *Third row:* The same procedure is used to obtain the reconstructions with noisy, random, and sparse samples (the sparsity and the noise increases from left to right). *Second row:* First, a sinc-INR is used to obtain a continuous reconstruction of the signal from discrete samples, which is then used as a surrogate signal to resample measurements. Afterwards, the diffeomorphisms are obtained using those measurements. As shown, sinc-INRs are able to recover the dynamics more robustly with noisy, sparse, and random samples.

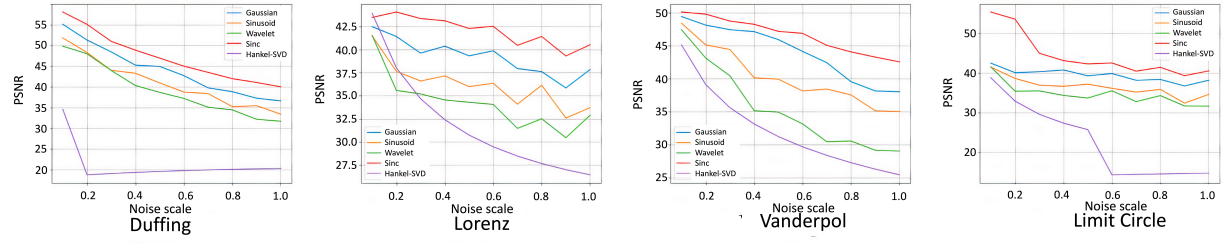


*Figure 3.* Quantitative comparison on discovering the dynamics of latent variables using INRs vs classical methods.

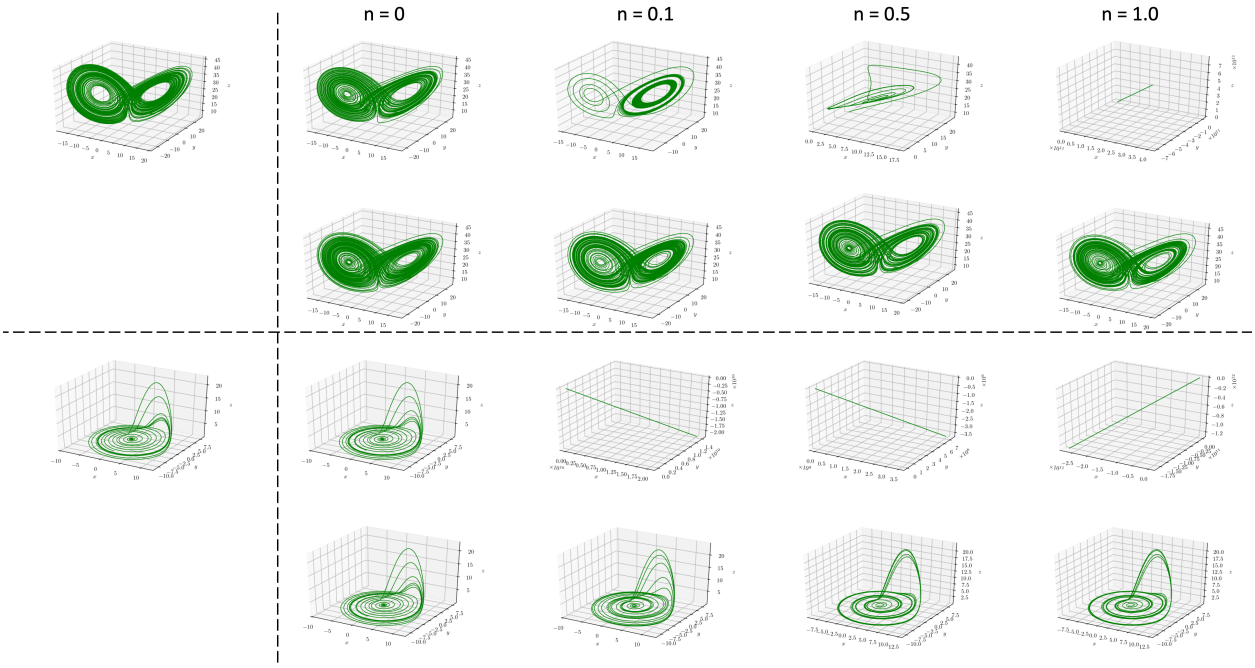


*Figure 4.* We use sinc-INRs to improve the results of the SINDy algorithm. The top block and the bottom block demonstrate experiments on the Lorenz system and the Rossler system, respectively. In each block, the top row and the bottom row represent the results of the baseline SINDy algorithm and the improved version (using INRs). As evident, INRs can be used to obtain significantly robust results.

$$\Theta(\mathbf{Y}) = \begin{bmatrix} \mathbf{A}(t_1) \\ \mathbf{A}(t_2) \\ \vdots \\ \mathbf{A}(t_N) \end{bmatrix}$$

where

$$A(t_i) = \left[ y_1^2(t_i), y_2^2(t_i), \ldots, \sin(y_1(t_i))\cos(y_2(t_i)), \ldots, y_d(t_i)y_2^2(t_i) \right].$$

SINDy then seeks to minimize the loss function:

$$L_S = ||\dot{\mathbf{Y}} - \Theta(\mathbf{Y})\Gamma||_2^2 + \lambda||\Gamma||_1^2, \tag{11}$$

where $\Gamma$ is a sparsity matrix initialized randomly that enforces sparsity.

INRs introduce two significant architectural biases here. When we train an INR using $\{t_n\}_{n=1}^N$ and $\{\mathbf{y}(t_n)\}_{n=1}^N$ as inputs and labels, it allows us to reconstruct a continuous representation of $\mathbf{y}(t)$. By controlling the frequency parameter $\omega$ of sinc functions during training, we can filter out high-frequency noise in $\mathbf{y}$. Additionally, $\dot{\mathbf{y}}$ measurements can be obtained by calculating the Jacobian of the network, taking advantage of the smooth derivatives of sinc-INRs. We then replace $\dot{\mathbf{Y}}$ and $\mathbf{Y}$ in Eq. 11 with values obtained from the INR, keeping the rest of the SINDY algorithm unchanged.

For our experiment, we employ the Lorenz and Rossler systems (refer to App. F for the equations that define these systems), generating 1000 samples from 0 to 100 at intervals of 0.1 to create $\mathbf{Y}$. We introduce noise from a uniform distribution $\eta \sim U(-n, n)$, varying $n$. As a baseline, we compute $\dot{\mathbf{Y}}$ using spectral derivatives, a common method in numerical analysis and signal processing for computing derivatives through spectral methods. This involves translating the function's derivative in the time or space domain to a multiplication by $i\omega$ in the frequency domain. The reason for choosing spectral derivatives is empirical; After evaluating various methods to compute $\dot{\mathbf{Y}}$, including finite difference methods and polynomial approximations, we empirically selected spectral derivatives for the best baseline. As a competing method, for each noise scale, we use a sinc-INR to compute both $\dot{\mathbf{Y}}$ and $\mathbf{Y}$ as described. Utilizing the SINDy algorithm for both scenarios, we obtain the governing equations for each system. The dynamics recovered from these equations are compared in Fig. 4 and Table 2. Remarkably, the sinc-INR approach demonstrates robust results at each noise level, surpassing the baseline. For this experiment, we use 4-layer INRs with each layer having a width of 256.

It is also important to note that the $\omega$ parameter of the sinc activation plays a key role in such signal reconstruction tasks. If $\omega$ is too small, the model will not be able to capture high-frequency informatino. On the other hand, if its too

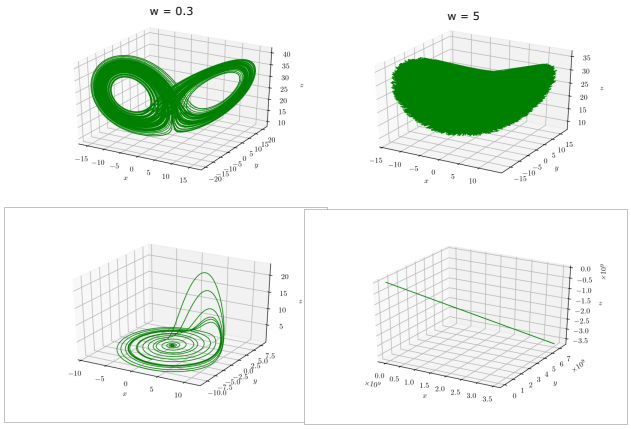


*Figure 5.* The top row and the bottom row depicts the SINDy reconstructions obtained for the Lorenz system and the Rossler system, respectively, using sinc-INRs. As $\omega$ is increased in the sinc function, the INR allows more higher frequencies to be captured, resulting in noisy reconstructions.

high, the INR allows unnecessarily higher frequencies to be captured, resulting in noisy reconstructions. Fig. 5 depicts this phenomenon.

### 5.4. Sinc activations for positional embeddings

We also discuss the possibility of using sinc activations for positional embeddings. Recent research, notably (Zheng et al., 2022), has provided compelling evidence that the effectiveness of positional encodings need not be exclusively tied to a Fourier perspective. They demonstrate that non-Fourier embedding functions, such as shifted Gaussian functions, can be effectively utilized for positional encoding. These functions are characterized by having a sufficiently high Lipschitz constant and the ability to generate high-rank embedding matrices, attributes that are shown to achieve results comparable to Random Fourier Feature (RFF) encodings.

Building on this, (Ramasinghe & Lucey, 2023) further confirmed that shifted Gaussian functions with spatially varying variances can surpass the performance of RFF encodings. Given that sinc functions also exhibit these desirable properties, they can be feasibly employed as shifted basis functions for high-frequency signal encoding.

To explore this, we developed a sinc-based positional embedding layer. For a 2D coordinate $(x_1, x_2)$, each dimension is embedded using sinc functions:

$$\psi_1(x_1) = [\text{sinc}(\|t_1 - x_1\|), \text{sinc}(\|t_2 - x_1\|), \dots, \text{sinc}(\|t_N - x_1\|)] \quad (12)$$

$$\psi_1(x_2) = [\text{sinc}(\|t_1 - x_2\|), \text{sinc}(\|t_2 - x_2\|), \dots, \text{sinc}(\|t_N - x_2\|)] \quad (13)$$

where $t_1, \dots, t_N$ are equidistant samples in $[0, 1]$. Then, these embeddings are concatenated to create the final embedding as,

$$\Psi(x_1, x_2) = [\psi_1(x_1), \psi_1(x_2)]$$

In a comparative study using the DIV2K dataset for image reconstruction, our sinc-based positional embedding layer demonstrated superior performance to an RFF-based layer, as shown Table 3:

| PE layer | PSNR |
|---|---|
| RFF | 23.5 |
| Sinc PE | 26.4 |

*Table 3.* Comparison of the sinc psitional embedding layer against RFF positional embeddings.

This result indicates that sinc-based positional embeddings offer a promising alternative to RFF encodings.

## 6. Conclusion

In this work, we offer a fresh view-point on INRs using sampling theory. Particularly, we focus on proposing a unified framework to analyze the properties of activations of INRs. In this vein, we show that sinc activations are optimal for encoding signals in the context of INRs. We conduct experiments on various modalities including image reconstructions, NeRFs and dynamical systems to showcase that these theoretical predictions hold at a practical level. Further, we discuss the potential of using sinc activations for positional encodings as well.

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

## References


Agustsson, E. and Timofte, R. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

Aldroubi, A., Unser, M., and Aldroubi, A. Sampling procedures in function spaces and asymptotic equivalence with shannon's sampling theory. *Numerical functional analysis and optimization*, 15(1-2):1–21, 1994.

Blu, T. and Unser, M. Approximation error for quasi-interpolators and (multi-) wavelet expansions. *Applied and Computational Harmonic Analysis*, 6(2):219–251, 1999.

Brunton, S. L., Proctor, J. L., and Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016.

Budišić, M., Mohr, R., and Mezić, I. Applied koopmanism. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 22(4):047510, 2012.

Büsching, M., Bengtson, J., Nilsson, D., and Björkman, M. Flowibr: Leveraging pre-training for efficient neural image-based rendering of dynamic scenes. *arXiv preprint arXiv:2309.05418*, 2023.

Chng, S.-F., Saratchandran, H., and Lucey, S. Preconditioners for the stochastic training of implicit neural representations. *arXiv preprint arXiv:2402.08784*, 2024.

Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Giannakis, D. and Majda, A. J. Nonlinear laplacian spectral analysis for time series with intermittency and low-frequency variability. *Proceedings of the National Academy of Sciences*, 109(7):2222–2227, 2012.

Holmes, P., Lumley, J. L., Berkooz, G., and Rowley, C. W. *Turbulence, coherent structures, dynamical systems and symmetry*. Cambridge university press, 2012.

Kennel, M. B., Brown, R., and Abarbanel, H. D. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical review A*, 45 (6):3403, 1992.

Kim, H., Eykholt, R., and Salas, J. Nonlinear dynamics, delay times, and embedding windows. *Physica D: Nonlinear Phenomena*, 127(1-2):48–60, 1999.

Kirby, M. *Geometric data analysis: an empirical approach to dimensionality reduction and the study of patterns*, volume 31. Wiley New York, 2001.

Kutz, J. N., Brunton, S. L., Brunton, B. W., and Proctor, J. L. *Dynamic mode decomposition: data-driven modeling of complex systems*. SIAM, 2016.

Li, Z., Wang, Q., Cole, F., Tucker, R., and Snavely, N. Dynibar: Neural dynamic image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4273–4284, 2023.

Light, W. Ridge functions, sigmoidal functions and neural networks. *Approximation theory VII*, pp. 163–206, 1992.

Lumley, J. L. The structure of inhomogeneous turbulent flows. *Atmospheric turbulence and radio wave propagation*, pp. 166–178, 1967.

Majda, A. J., Franzke, C., and Crommelin, D. Normal forms for reduced stochastic climate models. *Proceedings of the National Academy of Sciences*, 106(10):3649–3653, 2009.

Marks, R. J. I. *Introduction to Shannon sampling and interpolation theory*. Springer Science & Business Media, 2012.

Mezić, I. Analysis of fluid flows via spectral properties of the koopman operator. *Annual Review of Fluid Mechanics*, 45:357–378, 2013.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Naik, M. and Cochran, D. Nonlinear system identification using compressed sensing. In *2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pp. 426–430. IEEE, 2012.

Peherstorfer, B. and Willcox, K. Online adaptive model reduction for nonlinear systems via low-rank updates. *SIAM Journal on Scientific Computing*, 37(4):A2123–A2150, 2015.

Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., and Zhou, X. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9054–9063, 2021.

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., and Courville, A. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pp. 5301–5310. PMLR, 2019.

Ramasinghe, S. and Lucey, S. Beyond periodicity: towards a unifying framework for activations in coordinate-mlps.

In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pp. 142–158. Springer, 2022.

Ramasinghe, S. and Lucey, S. A learnable radial basis positional embedding for coordinate-mlps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2137–2145, 2023.

Reinbold, P. A., Kageorge, L. M., Schatz, M. F., and Grigoriev, R. O. Robust learning from noisy, incomplete, high-dimensional experimental data via physically constrained symbolic regression. *Nature communications*, 12(1):3219, 2021.

Sahyoun, S. and Djouadi, S. Local proper orthogonal decomposition based on space vectors clustering. In *3rd International Conference on Systems and Control*, pp. 665–670. IEEE, 2013.

Saragadam, V., LeJeune, D., Tan, J., Balakrishnan, G., Veeraraghavan, A., and Baraniuk, R. G. Wire: Wavelet implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18507–18516, 2023.

Saratchandran, H., Chng, S.-F., Ramasinghe, S., MacDonald, L., and Lucey, S. Curvature-aware training for coordinate networks. *arXiv preprint arXiv:2305.08552*, 2023.

Saratchandran, H., Ramasinghe, S., and Lucey, S. From activation to initialization: Scaling insights for optimizing neural fields. *arXiv preprint arXiv:2403.19205*, 2024.

Schmid, P. J. Dynamic mode decomposition of numerical and experimental data. *Journal of fluid mechanics*, 656: 5–28, 2010.

Schmit, R. and Glauser, M. Improvements in low dimensional tools for flow-structure interaction problems: using global pod. In *42nd AIAA aerospace sciences meeting and exhibit*, pp. 889, 2004.

Singer, M. A. and Green, W. H. Using adaptive proper orthogonal decomposition to solve the reaction–diffusion equation. *Applied Numerical Mathematics*, 59(2):272–279, 2009.

Sirovich, L. Turbulence and the dynamics of coherent structures. i. coherent structures. *Quarterly of applied mathematics*, 45(3):561–571, 1987.

Sitzmann, V., Martel, J., Bergman, A., Lindell, D., and Wetzstein, G. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020.

Small, M. *Applied nonlinear time series analysis: applications in physics, physiology and finance*, volume 52. World Scientific, 2005.

Stein, E. M. and Shakarchi, R. *Fourier analysis: an introduction*, volume 1. Princeton University Press, 2011.

Strümpler, Y., Postels, J., Yang, R., Gool, L. V., and Tombari, F. Implicit neural representations for image compression. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2022.

Sun, W. and Zhou, X. Irregular wavelet/gabor frames. *Applied and Computational Harmonic Analysis*, 13(1):63–76, 2002.

Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., and Ng, R. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33: 7537–7547, 2020.

Unser, M. Sampling-50 years after shannon. *Proceedings of the IEEE*, 88(4):569–587, 2000.

Voss, H. U., Kolodner, P., Abel, M., and Kurths, J. Amplitude equations from spatiotemporal binary-fluid convection data. *Physical review letters*, 83(17):3422, 1999.

Wang, W.-X., Yang, R., Lai, Y.-C., Kovanis, V., and Grebogi, C. Predicting catastrophes in nonlinear dynamical systems by compressive sensing. *Physical review letters*, 106(15):154101, 2011.

Ye, H., Beamish, R. J., Glaser, S. M., Grant, S. C., Hsieh, C.-h., Richards, L. J., Schnute, J. T., and Sugihara, G. Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling. *Proceedings of the National Academy of Sciences*, 112(13):E1569–E1576, 2015.

Zayed, A. I. *Advances in Shannon's sampling theory*. Routledge, 2018.

Zheng, J., Ramasinghe, S., Li, X., and Lucey, S. Trading positional complexity vs deepness in coordinate networks. In *European Conference on Computer Vision*, pp. 144–160. Springer, 2022.

# A. Proofs of results in section 3.3

## A.1. Preliminaries

We recall the definition of the space of square integrable functions on $\mathbb{R}$, which we denote by $L^2(\mathbb{R})$. This is defined as the vector space of equivalence classes of Lebesgue measurable functions on $\mathbb{R}$ that have finite $L^2$ norm, which is defined via the following inner product

$$\langle f, g \rangle_{L^2} = \int_{\mathbb{R}} f \cdot g. \tag{14}$$

We will also make use of the localized space of square integrable functions on $\mathbb{R}$ denoted $L^2_{loc}(\mathbb{R})$. This space is defined as the vector space of equivalence classes of Lebesgue measurable functions that have finite $L^2$ norm over any compact subset $K \subseteq \mathbb{R}$.

Note that a function that is in $L^2(\mathbb{R})$ is automatically in $L^2_{loc}(\mathbb{R})$ but the converse is in general not true.

We will also need to make use of the Sobolev spaces of order $r$, denoted by $W^r_2(\mathbb{R})$. We define this space as the space of $L^2$-functions that have $r$ weak derivatives that are also in $L^2(\mathbb{R})$.

The space of square integrable sequences will be denoted by $l^2(\mathbb{R})$. This is the Hilbert space of square summable sequences over the integers $\mathbb{Z}$. That is, a sequence $x(k)_{k\in\mathbb{Z}} \in l^2(\mathbb{R})$ if each $x(k) \in \mathbb{R}$ and

$$\sum_{k\in\mathbb{Z}} |x(k)|^2 < \infty. \tag{15}$$

***Proof of prop. 3.3.*** The function $\mathrm{sinc}(x)$ is in $L^2(\mathbb{R})$ and furthermore the translates $\mathrm{sinc}(x-k)$, for $k \in \mathbb{Z}$, form an orthonormal basis of $L^2(\mathbb{R})$. Hence $\mathrm{sinc}(x)$ satisfies the first condition of a Riesz basis with $A = B = 1$.

The next step is to check that the partition of unity condition holds. In order to do this we will make use of the Poisson summation formula (Stein & Shakarchi, 2011) that states that for a function $f \in L^2(\mathbb{R})$ we have

$$\sum_{k\in\mathbb{Z}} f(x+k) = \sum_{n\in\mathbb{Z}} \hat{f}(2\pi n) e^{2\pi i n x}. \tag{16}$$

Using the Poisson summation formula, we can rewrite the partition of unity condition, see cond. 2 in Defn. 3.1, as

$$\sum_{n\in\mathbb{Z}} \hat{f}(2\pi n) e^{2\pi i n x} = 1. \tag{17}$$

We then observe that the Fourier transform of $\mathrm{sinc}(x)$ is given by the characteristic function $\chi_{[-1,1]}$ on the set $[-1,1]$. I.e. $\chi_{[-1,1]}$ takes the value 1 on $[-1,1]$ and 0 elsewhere (Stein & Shakarchi, 2011). The proof now follows by observing that $\chi_{[-1,1]}(2\pi n) = 1$ for $n = 0$ and 0 for $n \neq 0$. We then see that (17) is true for $\mathrm{sinc}(x)$ and thus $\mathrm{sinc}(x)$ forms a Riesz basis.

The proof that the Gaussian $\phi = e^{-x^2/2s^2}$ does not form a Riesz basis and only a weak Riesz basis follows the same strategy as above. The first step is to note that translates of the Gaussian: $\phi_k = e^{-(x-k)^2/2s^2}$ all lie in $L^2(\mathbb{R})$ for any $k \in \mathbb{Z}$. This establishes the upper bound in condition 1 of the Riesz basis definition. To prove the lower bound in condition 1, we use an equivalent definition of condition 1 in the Fourier domain given by

$$A \leq \sum_{k\in\mathbb{Z}} |\hat{\phi}(\xi + 2k\pi)|^2 \leq B \tag{18}$$

where $\hat{\phi}$ denotes the Fourier transform of $\phi$ and $\xi$ the frequency variable in the Fourier domain. The equivalence of (18) with the Riesz basis definition given in Defn. 3.1 follows by noting that Defn. 3.1 is translation invariant, see (Aldroubi et al., 1994) for explicit details. We then observe that in the case of a Gaussian the term $\hat{\phi}(\xi + 2k\pi)$ is given by $e^{-(\xi+2k\pi)^2 s^2/2}$, which follows from the fact that the Fourier transform of a Gaussian is another Gaussian, see (Stein & Shakarchi, 2011). The final observation to make is that the sum

$$\sum_{k\in\mathbb{Z}} |\hat{\phi}(\xi + 2k\pi)|^2 \geq |\hat{\phi}(\xi)|^2 \tag{19}$$

for any $\xi \in \mathbb{R}$ and that we only need to consider $\xi \in [0, 2\pi]$ from the symmetry of the Gaussian about the y-axis and the fact that for any $\xi$ outside of $[0, 2\pi]$, there exists some $k \in \mathbb{Z}$ such that the translate $\xi + 2k\pi$ lies in $[0, 2\pi]$. The lower bound in (18) then follows by taking $0 < A \leq e^{-(2\pi s)^2}$.

In order to show that the Gaussian $\phi$ does not satisfy the partition of unity condition. We go through the formulation (17). In this case this formula reads

$$\sum_{n\in\mathbb{Z}} e^{-(2\pi n)^2 s^2/2} e^{2\pi i n x} = 1. \tag{20}$$

We now observe it is impossible for this equality to hold due to the gaussian decay of the function $e^{-(2\pi n)^2 s^2/2}$. In particular for $x = 0$ the condition becomes

$$\sum_{n\in\mathbb{Z}} e^{-(2\pi n)^2 s^2/2} = 1. \tag{21}$$

The left hand side is clearly greater than 1, and thus we see that the condition cannot hold. This proves that a Gaussian $e^{-x^2/2s^2}$ can only define a weak Riesz basis.

In general, the Fourier transform of a wavelet is localized in phase and frequency, hence as in the case of the Gaussian above, they will be in $L^2(\mathbb{R})$ and form a weak Riesz basis but in general they might not form a Riesz basis. Conditions have been given for a wavelet to form a Riesz basis, see (Sun & Zhou, 2002), though this is outside the scope of this work.

In order to form a Riesz basis $ReLU$ would have to be in $L^2(\mathbb{R})$, which it is not. On the other hand, given $x \in \mathbb{R}$ we have that

$$\sum_{k\in\mathbb{Z}} ReLU(x+k) = \sum_{k\geq -x, k\in\mathbb{Z}} ReLU(x+k) = \sum_{k\geq -x, k\in\mathbb{Z}} (x+k) = \infty$$

showing that there is no way $ReLU$ could satisfy the partition of unity condition.

A similar proof shows that translates of sine cannot form a Riesz/weak Riesz basis. □

### A.1.1. Results on the error kernel and PUC condition

We recall from sec. 3.3 that the understanding of the sampling properties of the shifted basis functions $F_k$ comes down to analysing the error kernel $E_{\widetilde{F},F}$. The reason being was that the average error $\overline{\epsilon}_s(T)^2$ is a good predictor of the true error $\epsilon_s(T)^2$. The main theorem that shows this is the following theorem from (Blu & Unser, 1999)

**Theorem A.1.** *The $L^2$ approximation error $\epsilon_s(\Omega)^2$ can be written as*

$$\epsilon_s(\Omega)^2 = \left( \int_{-\infty}^{\infty} E_{\widetilde{F},F}(\Omega\xi) |\hat{s}(\xi)|^2 \frac{d\xi}{2\pi} \right)^{1/2} + \epsilon_{corr} \tag{22}$$

*where $\epsilon_{corr}$ is a correction term negligible under most circumstances. Specifically, if $f \in W_2^r$ (Sobolev space of order $r$, see appendix A) with $r > \frac{1}{2}$, then $|\epsilon_{corr}| \leq \gamma \Omega^r ||s^{(r)}||_{L^2}$ where $\gamma$ is a known constant and moreover, $|\epsilon_{corr}| = 0$ provided the signal $s$ is bandlimited to $\frac{\pi}{\Omega}$.*

Our goal in this section is to give a sketch of the proof of the following lemma, which relates the partition of unity condition to the vanishing of the error kernel, following the reference (Blu & Unser, 1999).

**Lemma A.2.** *If the family of shifted basis function $\{F(x-k)\}_{k\in\mathbb{Z}}$ satisfies the condition*

$$\sum_{k\in Z} F(x+k) = 1, \forall x \in \mathbb{R} \tag{23}$$

*then $\lim_{\Omega\to 0} E_{\widetilde{F},F} \to 0$, for any $\widetilde{F} \in \overline{\mathcal{S}}$, where $\overline{\mathcal{S}}$ is the space of Schwartz functions $f$ whose Fourier transform satisfies $\hat{f}(0) = 1$.*

We now sketch a proof showing that the vanishing of the error kernel in the limit $T \to 0$ for a suitable test function $\widetilde{F}$ is equivalent to $F$ satisfying the partition of unity condition. We will do this under two assumptions:

A1. The Fourier transform of $F$ is continuous at 0.

A2. The Fourier transform of $\widetilde{F}$ is continuous at 0.

A3. The sampled signal $s$ we wish to reconstruct is contained in $W_2^r$ for some $r > \frac{1}{2}$. This assumption is needed so that the quantity $\epsilon_{corr}$ goes to zero as $T \to 0$.

We remark that an explicit construction of $\widetilde{F}$ will be given after the proof as during the course of the proof we will see what conditions we need to impose for the construction of $\widetilde{F}$ from $F$.

From the definition of the approximation operator, (3), we have that

$$\lim_{T\to 0} ||f - A_T(f)||_{L^2}^2 = \lim_{T\to 0} \int_{-\infty}^{\infty} E_{\widetilde{F},F}(T\omega)|\hat{s}(\omega)|^2 \frac{d\omega}{2\omega} \tag{24}$$

where we remind the reader that the error kernel $E_{\widetilde{F},F}$ is given by equation (6). We now observe that if $\widetilde{F}$ is a function such that $\hat{\widetilde{F}}$ is bounded and $F$ satisfies the first Riesz condition, condition 1 from Defn. 3.1, then by definition it follows that $E_{\widetilde{F},F}$ is bounded. Therefore in the above integral we can apply the dominated convergence theorem and compute

$$\lim_{T\to 0} ||s - A_T(s)||_{L^2}^2 = \int_{-\infty}^{\infty} \lim_{T\to 0} E_{\widetilde{F},F}(T\omega)|\hat{s}(\omega)|^2 \frac{d\omega}{2\omega} \tag{25}$$

$$= E_{\widetilde{F},F}(0) \int_{-\infty}^{\infty} |\hat{s}(\omega)|^2 \frac{d\omega}{2\omega} \tag{26}$$

$$= E_{\widetilde{F},F}(0)||s||^2 \tag{27}$$

where to get the second equality we have used assumptions A1 and A2 above and to get the third equality we have used the fact that the Fourier transform is an isometry from $L^2(\mathbb{R})$ to itself.

We thus see that the statement $\lim_{T\to 0} ||s - Q_T(s)||_{L^2}^2 = 0$ is equivalent to $E_{\widetilde{F},F}(0) = 0$. From (6) this is equivalent to

$$E_{\widetilde{F},F}(0) = |1 - \hat{\widetilde{F}}(0)\hat{F}(0)|^2 + |\hat{\widetilde{F}}(0)|^2 \sum_{k\neq 0} |\hat{F}(2\pi k)|^2 = 0. \tag{28}$$

We see that $E_{\widetilde{F},F}(0)$ is a sum of positive terms and hence will vanish if and only if all the terms in the summands vanish. Looking at the first summand we see that we need $\hat{\widetilde{F}}(0)\hat{F}(0) = 1$, which can hold if and only if both factors are not zero. We normalise the function $F$ so that $\hat{F}(0) = \int F(x)dx = 1$. Thus the conditions that need to be satisfied are

$$\hat{\widetilde{F}}(0) = 1 \text{ and } \sum_{k\neq 0} |\hat{F}(2\pi k)|^2 = 0. \tag{29}$$

We can rewrite the second condition in (29) as

$$\hat{F}(2\pi k) = \delta_k \tag{30}$$

where $\delta$ denotes the Dirac delta distribution. From this viewpoint we then immediately have that the second condition can be written in the form

$$\sum_k F(x + k) = 1 \tag{31}$$

which is precisely the partition of unity condition.

The function $\widetilde{F}$ is easy to choose. Let $\mathcal{S}$ denote Schwartz space of Schwartz functions in $L^2(\mathbb{R})$. It is well known that this space is dense in $L^2(\mathbb{R})$ and that the Fourier transform maps $\mathcal{S}$ onto itself. Therefore, in the Fourier domain let $\widetilde{\mathcal{S}}$ denote the set of Schwartz functions $f$ such that $\hat{f}(0) \neq 0$. Note that $\widetilde{\mathcal{S}}$ is dense in $L^2(\mathbb{R})$ and elements in $\widetilde{\mathcal{S}}$ are continuous at the origin. In order to define $\widetilde{F}$ we simply take any element $f \in \widetilde{\mathcal{S}}$ and let $\widetilde{F} = \frac{1}{\hat{f}(0)} f$. In fact, if we denote the space $\overline{\mathcal{S}}$ to consist of those Schwartz functions $f$ whose Fourier transform satisfies $\hat{f}(0) = 1$, then it is easy to see that $\overline{\mathcal{S}}$ is dense in $L^2(\mathbb{R})$. Thus the space $\overline{\mathcal{S}}$ can be used as a test space for $\widetilde{F}$ and is the defining test space for the approximation operator $A_T$.

### A.2. What does the partition of unity condition mean?

In the previous sec. A.1.1 we saw that the vanishing of the error kernel $E_{\widetilde{F},F}$ in the limit $T \to 0$ was equivalent to the function $F \in L^2(\mathbb{R})$ satisfying the partition of unity condition. In this section we want to explain in a more qualitative manner what the partition of unity condition means for reconstruction in the space $L^2(\mathbb{R})$.

Fix a function $F \in L^2(\mathbb{R})$, we have seen we can create the subspace $V(F) \subseteq L^2(\mathbb{R})$. For the time being let us only assume $F$ satisfies the first condition of being a Riesz basis. Recall this means that:

$$A||a||_{l^2}^2 \leq \left|\left| \sum_{k \in \mathbb{Z}} a(k) F_k \right|\right|^2 \leq B||a||_{l^2}^2, \forall a(k) \in l^2(\mathbb{R}) \tag{32}$$

Given an arbitrary function $g \in V(F)$ the above condition 32 means that when we express

$$g = \sum_{k=-\infty}^{\infty} a(k) F(x-k), \tag{33}$$

the coefficients $a(k)$ are uniquely determined. This follows because condition 32 implies that the translates $F(x-k)$ form a linearly independent set inside $V(F)$. Thus condition 1 is there to tell us how to approximate functions within $V(F)$. It states that we can perfectly reconstruct any function in $V(F)$ using the translates $\{F(x-k\}$.

However, let us now assume that we are given a function $g \in L^2(\mathbb{R}) - V(F)$, that is $g$ is a square integrable function that does not reside in the space $V(F)$. A natural question that arises is can we we still use elements in the space $V(F)$ to approximate $g$? Mathematically, what this question is asking is if we are given a very small $\epsilon > 0$ can we find a function $G \in V(F)$ such that

$$||G - g||_{L^2} < \epsilon? \tag{34}$$

This is precisely where the partition of unity condition comes in:

$$\sum_{k \in Z} F(x+k) = 1, \forall x \in \mathbb{R} (\textbf{PUC}) \tag{35}$$

Mathematically, the reason the partition of unity condition is able to bridge the gap between $V(F)$ and $L^2(\mathbb{R})$ is that if we have an arbitrary function $g \in L^2(\mathbb{R}) - V(F)$, then we can write

$$g = g - G + G \tag{36}$$

for any function $G \in V(F)$. The question now is does there exist a $G \in V(F)$ that makes the quantity $g - G$ very small in the $L^2$-norm? In other words, given a very small $\epsilon > 0$ can we make $g - G$ smaller than $\epsilon$ in the $L^2$-norm.

The way to answer this question is to first note that there is a simple way to try to construct such a $G$. Namely, project $g$ onto the subspace $V(F)$ forming the function $\mathcal{P}(g) \in V(F)$. Then look at the difference

$$g - \mathcal{P}(g) \tag{37}$$

and ask can it be made very small? In general this technique does not work. However, there is another projection. Namely, we can project $g$ onto the $\Omega$-scaled signal space $V_\Omega(F)$ for $\Omega > 0$ forming $\mathcal{P}_\Omega(g)$ and ask if the difference $g - \mathcal{P}_\Omega(g)$ can be made very small. For the definition of the $\Omega$-scaled signal space $V_\Omega(F)$ please see sec. 3.3.

The partition of unity condition says that there exists a $\Omega > 0$ such that the difference

$$g - \mathcal{P}_\Omega(g) \tag{38}$$

can be made very small.

Thus the second condition from the Riesz basis definition, the partition of unity condition, is telling us how to approximate functions outside of $V(F)$ using the translates $\{F(x-k\}$ and the scaled signal spaces $V_\Omega$. It says that we cannot necessarily perfectly reconstruct a function outside of $V(F)$ but we can reconstruct it up to a very small error using the $\Omega$-scaled signal space $V_\Omega(F)$. The partition of unity condition bridges the gap between $V(F)$ and $L^2(\mathbb{R})$ via the scaled signal spaces $V_\Omega(F)$ telling us that reconstruction is possible only in $V_\Omega(F)$ for some $\Omega > 0$.

For a full mathematical proof of how the partition of unity does this we kindly ask the reader to consult sec. A.1.1.

Let us summarize what we have discussed:

1. The first condition of a Riesz basis is there so that we know that translates of $F$ namely $\{F(x-k\}$ can be used to uniquely approximate functions in the signal space $V(F)$. In this case, theoretically the translates $\{F(x-k\}$ provide a perfect reconstruction.

2. The second condition of a Riesz basis, namely the partition of unity condition, is there so that we know how to approximate functions that do not lie in $V(F)$. It says that in order to bridge the gap between $V(F)$ and $L^2(\mathbb{R})$ we need to do so by going through an $\Omega$-scaled signal space $V_\Omega(F)$ for a $\Omega > 0$. In the scaled signal space perfect reconstruction is not possible but we can reconstruct up to a very small error.

### A.2.1. Proofs of main results in section 3.3

***Proof of theorem 3.4.*** We first note that by condition 1 in Defn. 3.1. The space $V(F)$ is a subspace of $L^2(\mathbb{R})$. Therefore, the space $V(F)$ with the induced $L^2$-norm forms a well-defined normed vector space.

Since $g \in V(F)$ we can write $g = \sum_{k=-\infty}^{\infty} a(k)F(x-k)$ in $L^2(\mathbb{R})$. This means that the difference

$$g - \sum_{k=-\infty}^{\infty} a(k)F(x-k) = 0 \in L^2(\mathbb{R}) \tag{39}$$

and in particular that the partial sums

$$S_n := \sum_{k=-n}^{n} a(k)F(x-k) \tag{40}$$

converge in $L^2$ to $g$ as $n \to \infty$. Writing this out, this means that given any $\epsilon > 0$, there exists an integer $k(\epsilon)$ such that

$$\left|\left| g - \sum_{k=-k(\epsilon)}^{k(\epsilon)} a(k)F(x-k) \right|\right|_{L^2} < \epsilon. \tag{41}$$

We can then define a 2-layer neural network $f$ with $n(\epsilon) = 2k(\epsilon) + 1$ neurons as follows: Let the weights in the first layer be the constant vector $[1, \cdots, 1]^T$ and the associated bias to be the vector $[-k(\epsilon), -k(\epsilon)+1, \ldots, k(\epsilon)]^T$. Let the weights associated to the second layer be the vector $[a(-k(\epsilon)), a(-k(\epsilon)+1), \cdots, a(k(\epsilon))]$ and the associated bias be 0. These weights and biases will make up the parameters for the neural network $f$ and in the hidden layer we take $F$ as the non-linearity.

Applying (41) we obtain that

$$||f(\theta) - g||_L^2 < \epsilon. \tag{42}$$

□

***Proof of prop. 3.7.*** The proof of this proposition will be in two steps. The reason for this is that we need to use Thm. A.1 and in doing so we want to know that the error $\epsilon_{corr}$ can be made arbitrarily small. Thm. A.1 shows that if we assume our signal $s \in W_2^1(\mathbb{R})$, then we have the bound

$$\epsilon_{corr} \leq \gamma\Omega||s^{(1)}||_{L^2} \tag{43}$$

where $s^{(1)}$ denotes the first Sobolev derivative, which exists because of the assumption that $s \in W_2^1$.

We thus see that if we choose $\Omega > 0$ sufficiently small we can make $\epsilon_{corr} < \frac{\epsilon}{2}$, by (43). Furthermore, by lemma A.2 we have that the average approximation error $\overline{\epsilon}(\Omega) < \frac{\epsilon}{2}$ for $\Omega$ sufficiently small. Therefore, by taking $f_\Omega = A_\Omega(s) \in V_\Omega(F)$ the proposition follows for the signal $s \in W_2^1$.

As we have only proved the proposition for signals in $s \in W_2^1(\mathbb{R})$ we are not done. We want to prove it for signals $s \in L^2(\mathbb{R})$. This is the second step, which proceeds as follows.

We start by observing that $A_\Omega$ is a bounded operator from $L^2(\mathbb{R})$ into $L^2(\mathbb{R})$, see (Blu & Unser, 1999). Let $T = ||A_\Omega||_{op}$ denote the operator norm of $A_\Omega$. We also use the fact that $C_c^\infty(\mathbb{R})$ is dense in $L^2(\mathbb{R})$, see (Stein & Shakarchi, 2011).

Then by density of $C_c^\infty(\mathbb{R})$ in $L^2(\mathbb{R})$ we can find an $f \in C_c^\infty(\mathbb{R})$ such that

$$||f - s||_{L^2} < \min\left\{\frac{\epsilon}{3T}, \frac{\epsilon}{3}\right\} \tag{44}$$

$||f - s||_{L^2} < \frac{\eta}{3T}$. Furthermore, since $f \in C_c^\infty$ it lies in $W_2^1$. By the above we have that there exists $\Omega > 0$ such that $||f - A_\Omega(f)||_{L^2} < \frac{\epsilon}{3}$. We then estimate:

$$\begin{aligned}
||s - A_\Omega(s)|| &= ||s - f + f - A_\Omega(f) + A_\Omega(f) - A_\Omega(s)||_{L^2} && (45)\\
&\leq ||s - f||_{L^2} + ||f - A_\Omega(f)||_{L^2} + ||A_\Omega(f) - A_\Omega(s)||_{L^2} && (46)\\
&\leq ||s - f||_{L^2} + ||f - A_\Omega(f)||_{L^2} + ||A_\Omega||_{op}||s - f||_{L^2} && (47)\\
&\leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} && (48)\\
&= \epsilon && (49)
\end{aligned}$$

where (46) follows from the triangle inequality and (47) from (44). This completes the proof. □

***Proof of Thm. 3.8.*** By Prop. 3.7 there exists an $\Omega > 0$ sufficiently small and an $f_\Omega \in V_\Omega(F)$ such that

$$||s - f_\Omega||_{L^2} < \frac{\epsilon}{2}. \tag{50}$$

As $f_\Omega$ lies in $V_\Omega(F)$ we can write $f_\Omega = \sum_{k=-\infty}^{\infty} a_\Omega(k)F(\frac{1}{\Omega}(x - \Omega k))$. This implies that the partial sums

$$S_n = \sum_{k=-n}^{n} a_\Omega(k)F(\frac{1}{\Omega}(x - \Omega k)) \tag{51}$$

converge under the $L^2$-norm to $f_\Omega$ as $n \to \infty$. By definition of convergence this means given any $\epsilon > 0$ there exists an integer $k(\epsilon) > 0$ such that

$$\left|\left|f_\Omega - \sum_{k=-k(\epsilon)}^{k(\epsilon)} a_\Omega(k)F\left(\frac{1}{\Omega}(x - \Omega k)\right)\right|\right| < \frac{\epsilon}{2}. \tag{52}$$

We define a neural network $\mathcal{N}$ with $n(\epsilon) = 2k(\epsilon) + 1$ neurons in its hidden layer as follows. The weights in the first layer will be the constant vector $[1, \ldots, 1]^T$ and the associated bias will be the vector $[-\Omega k(\epsilon), -\Omega k(\epsilon) + 1, \ldots, \Omega k(\epsilon)]^T$. The weights associated to the second layer will be $[a(-k(\epsilon)), \ldots, a(k(\epsilon))]$ and the bias for this layer will be 0. These weights and biases will make up the parameters $\theta$ for the neural network. In the hidden layer we take as activation the function $F_\Omega$. With these parameters and activation function, we see that

$$\mathcal{N}(\theta)(x) = \sum_{k=-k(\epsilon)}^{k(\epsilon)} a_\Omega(k)F\left(\frac{1}{\Omega}(x - \Omega k)\right). \tag{53}$$

We then have that (52) implies that

$$||\mathcal{N}(\theta) - f_\Omega||_{L^2} < \frac{\epsilon}{2}. \tag{54}$$

Combining this with (50) we have

$$||\mathcal{N}(\theta) - s||_{L^2} = ||\mathcal{N}(\theta) - f_\Omega + f_\Omega - s||_{L^2} \tag{55}$$
$$\leq ||\mathcal{N}(\theta) - f_\Omega||_{L^2} + ||f_\Omega - s||_{L^2} \tag{56}$$
$$\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} \tag{57}$$
$$= \epsilon \tag{58}$$

where (56) follows from the triangle inequality and (57) follows by using (50) and (54). The theorem has been proved.

□

## B. Extensions of the theory

### B.1. Extending the theory to deep networks

In this section we will extend our main theorem 3.8 to the setting of deep networks. Our results will be proved for signals lying in $L^2(K)$ where $K$ is a compact subset of $\mathbb{R}$. This is not a strong assumption as data sets are always finite and hence are always contained in some compact set $K$. Furthermore, we will be assuming that the Riesz bases we deal with will all be generated by a continuous function $F$. As many practical deep networks employ a continuous activation function this assumption is still useful for such practical deep networks.

We start with the following lemma.

**Lemma B.1.** *Let $F$ be a continuous function that generates a Riesz basis $V(F)$. Given any compact set $K$, let $\phi_K(x)$ denote the function that is the affine map $Ax + b$, for some $A$, $b \in \mathbb{R}$, over $K$ and zero outside. Then for any $\epsilon > 0$, we have that there exists an $\Omega > 0$ and an $f \in V_\Omega(F)$ such that*

$$||f - \phi||_{L^2(K)} < \epsilon.$$

*Proof.* We first observe that $\phi_K \in L^2(\mathbb{R})$. This means we can apply Prop. 3.7 to find an $\Omega > 0$ and an $f \in V(F)$ such that

$$||f - \phi||_{L^2(\mathbb{R})} < \epsilon.$$

However, note that $\phi_K$ vanishes outside of $K$. Thus we must have that

$$||f - \phi||_{L^2(K)} < \epsilon.$$

□

We will need one more lemma before we prove the main theorem for deep networks.

**Lemma B.2.** *Let $F$ be a continuous function that generates a Riesz basis $V(F)$. Fix a compact set $K$ and consider the maps $\phi_n$ defined by $x \mapsto x - n$ over $K$ and is zero outside $K$, where $n \in \mathbb{Z}$. If for a fixed $n \in \mathbb{Z}$ there exists an $\Omega > 0$ and a $f_n \in V_\Omega(F)$ such that*

$$||\phi_n - f_n||_{L^2(\mathbb{R})} < \epsilon$$

*for some $\epsilon > 0$ then for any other $k \in \mathbb{Z}$ such that $k \neq n$, there exists an $f_k \in V_\Omega(F)$ such that*

$$||\phi_k - f_k||_{L^2(\mathbb{R})} < \epsilon$$

*Proof.* The proof of the lemma starts by observing that over the compact set $K$, the graphs of the functions $\phi_k$ for $k \in \mathbb{Z}$ are all parallel. Fix $k \in \mathbb{Z}$ such that $k \neq n$. Write

$$f_n \sum_{j=-\infty}^{\infty} a_j(n) F_\Omega(x - \Omega j).$$

Since $\phi_k$ is parallel to $\phi_n$ over $K$ and equals $\phi_n$ outside of $K$. For those points $x - \Omega j$ that lie inside $K$ we can simply increase or decrease the amplitude $a_j(n)$, depending on whether $phi_k(x)$ is bigger or smaller than $\phi_n(x)$, and obtain a

representation of the basis functions $F_\Omega(x-\Omega j)$ over $K$ that approximate $\phi_k$ over $K$. For those points $x-j$ that lie outside of $K$ we don't change the amplitude $a_j(n)$ of the respective basis function $F_\Omega(x-\Omega j)$. This creates a new function $f_k \in V_\Omega(F)$ such that

$$||\phi_k - f_k||_{L^2(\mathbb{R})} < \epsilon.$$

□

The next step is to prove the an analogue of Thm. 3.8 for a deep network that has two hidden layers. The general case of $k$ hidden layers then follows by an induction argument.

**Theorem B.3.** *Let $K$ be a fixed compact set, $s \in L^2(K)$, and $\epsilon > 0$. Assume the shifted functions $\{F(x-k)\}_{k\in\mathbb{Z}}$ form a Riesz basis for $V(F)$ where $F$ is a continuous function. Then there exists a deep INR $\mathcal{N}$, with two hidden layers with $n(\epsilon)_1$ neurons in the first hidden layer and $n(\epsilon)_2$ neurons in the second hidden layer, parameter set $\theta$, and an $\Omega_1 > 0$, $\Omega_2 > 0$, such that*

$$||\mathcal{N}(\theta) - s||_{L^2(K)} < \epsilon$$

*where $\mathcal{N}(\theta)$ employs $F_{\Omega_1}$ and $F_{\Omega_2}$ as its activation in the first and second hidden layers respectively, where $F_\Omega(x) = F(\frac{1}{\Omega}x)$.*

*Proof.* Extend $s$ by zero to all of $\mathbb{R}$ to obtain a function $\widetilde{s} \in L^2(\mathbb{R})$. We then apply Thm. 3.8 to obtain a shallow $\mathcal{N}_s$ that can approximate $\widetilde{s}$ on $\mathbb{R}$ in the $L^2$ norm with activation $F_{\Omega_2}$ where $\Omega_2 > 0$. That is,

$$||\mathcal{N}_s - \widetilde{s}||_{L^2(\mathbb{R})}.$$

We will denote the number of neurons in the hidden layer of $\mathcal{N}_s$ by $2n_2$ and write

$$\mathcal{N}_s(x) = \sum_{k=-n_2}^{n_2} b_k F_{\Omega_2}(x - \Omega_2 k).$$

Thus the parameters of $\mathcal{N}_s$ are given by: $[1, \cdots, 1]^T$ as the weight matrix of the hidden layer and $[(-\Omega_1)(-n_2), (-\Omega_1)(-n_2+1), \cdots, (-\Omega_1)(n_2)]$ as the bias of the hidden layer, $[b_{-n_2}, \cdots, b_{n_2}]$ as the weight matrix of the final layer and $[0]$ as the bias of the final layer.

We then observe that we can define $2n_2$ functions $\phi_k$ given by $x \mapsto x - \Omega k$ over the compact set $K$ and is zero outside $K$, for $k$ an integer such that $-n_2 \leq k \leq n_2$.

Applying Lem. B.2 for each $k$ and any $\widetilde{\epsilon} > 0$ we can find an $\Omega_1 > 0$ and $f_k \in V_{\Omega_1}(F)$ such that

$$||\phi_k - f_k||_{L^2(\mathbb{R})} < \widetilde{\epsilon}.$$

Each $f_k$ is an infinite series. Therefore, we can find a $n_1$ such that if we let $\widetilde{f_k}$ denote the first $2n_1$ sums of $f_k$ so that

$$\widetilde{f_k} = \sum_{j=-n_1}^{n_1} a_j(k) F_{\Omega_1}(x - \Omega_1 j)$$

and so that

$$||\phi_k - \widetilde{f_k}||_{L^2(\mathbb{R})} < \epsilon.$$

We now observe that since $\widetilde{f_k}$ is $\epsilon$ close to $\phi_k$ in $L^2$ by taking $\epsilon$ even smaller if necessary we have that $\widetilde{f_k}$ is $\epsilon$ close to $\phi_k$ in the pointwise norm. So that we have

$$|\phi_k(x) - \widetilde{f_k}(x)| < \epsilon$$

for all $x \in K$.

We can now build the required deep network with 2 hidden layers. The parameters $\theta$ of the deep network are defined as follows: The first layer of the network will have $2n_1$ neurons. The weight matrix for the first hidden layer will be a $2n_1 \times 1$

matrix given by $W_1 = [1, \cdots, 1]^T$ with bias $b_1 = [-\Omega_1(-n_1), \cdots - \Omega_1(n_1)]$. The activation function in this layer will be $F_{\Omega_1}$.

The second hidden layer will have $2n_2$ neurons. The weight matrix will be a $2n_2 \times 2n_1$ matrix given by

$$W_2 = \begin{bmatrix} a_{-n_1}(-n_2) & a_{-n_1+1}(-n_2) & \cdots & a_{n_1}(-n_2) \\ \vdots & \vdots & \vdots & \vdots \\ a_{-n_1}(n_2) & a_{-n_1+1}(n_2) & \cdots & a_{n_1}(n_2) \end{bmatrix}$$ with bias $b_2 = [0, \cdots, 0]^T$ and activation $F_{\Omega_2}$.

Finally, the final layer has weight matrix a $1 \times n_2$ matrix given by $W_3 = [b_{-n_2}, \cdots, b_{n_2}]$ with bias $b_3 = [0]$.

This defines a deep network $\mathcal{N}$. We can check that it satisfies the theorem. First observe that for $x \in K$ we have $W_1 x + b_1$ is given by $[x - \Omega_1(-n_1), \cdots, x - \Omega_1(n_1)]^T$. Applying the activation $F_{\Omega_1}$ gives the vector $[F_{\Omega_1}(x - \Omega_1(-n_1)), \cdots, F_{\Omega_1}(x - \Omega_1(n_1))]^T$. When we apply $W_2$ and add $b_2$ to this vector we obtain the vector: $\begin{bmatrix} \sum_{j=-n_1}^{n_1} a_j(-n_2) F_{\Omega_1}(x - \Omega_1 j) \\ \vdots \\ \sum_{j=-n_1}^{n_1} a_j(n_2) F_{\Omega_1}(x - \Omega_1 j) \end{bmatrix}$. Now observe that this latter vector is $\epsilon$ close in pointwise norm to the vector $\begin{bmatrix} \phi_{-n_2} \\ \vdots \\ \phi_{n_2} \end{bmatrix}$. Therefore over the compact set $K$ we get that the vector $\begin{bmatrix} F_{\Omega_2}(\sum_{j=-n_1}^{n_1} a_j(-n_2) F_{\Omega_1}(x - \Omega_1 j)) \\ \vdots \\ F_{\Omega_2}(\sum_{j=-n_1}^{n_1} a_j(n_2) F_{\Omega_1}(x - \Omega_1 j)) \end{bmatrix}$ is $\epsilon$ close to $\begin{bmatrix} F_{\Omega_2}(x - \Omega_2(-n_2)) \\ \vdots \\ F_{\Omega_2}(x - \Omega_2(n_2)) \end{bmatrix}$ where we have used the continuity of $F$.

Applying the matrix $W_3$ and adding the bias $b_3$ we obtain that the number $\sum_{k=-n_2}^{n_2} b_k \left( F_{\Omega_2}(\sum_{j=-n_1}^{n_1} a_j(k) F_{\Omega_1}(x - \Omega_1 j)) \right)$ is $\epsilon$ close to the number $\sum_{k=-n_2}^{n_2} b_k F_{\Omega_2}(x - \Omega_2 k)$. As all functions in question are continuous, and the set $K$ is compact it follows that the two functions are $\epsilon$ close in the uniform norm and hence in the $L^2(K)$ norm. It then follows that we have the estimate

$$||\mathcal{N}(x; \theta) - s||_{L^2(K)} < \epsilon$$

and the proof is finished. □

The previous theorem B.3 established that there exists a deep network with 2 hidden layers and activations defined by a Riesz basis function $F$ that can approximate an $L^2$ signal over any compact set $K$.

The idea of the proof was simple. The first step is to approximate $s$ by a shallow network and then to approximate linear functions $x - n$ for some $n \in \mathbb{Z}$ using elements in a scaled signal space. For general deep networks the process repeats itself. For example if we want to approximate the signal $s$ with a deep network with 3 hidden layers, we start by approximating the signal over $K$ with a shallow network, then approximate the linear functions $x - n$ over $K$ with a deep network with 2 hidden layers. Putting together these two approximations, using the continuity of $F$ gives the desired result.

**Theorem B.4.** *Let $K$ be a fixed compact set, $s \in L^2(K)$, and $\epsilon > 0$. Assume the shifted functions $\{F(x - k)\}_{k \in \mathbb{Z}}$ form a Riesz basis for $V(F)$ where $F$ is a continuous function. Then for any $k > 1$, there exists a deep INR $\mathcal{N}$, with $k$ hidden layers with $n(\epsilon)_j$ neurons in the jth hidden layer, parameter set $\theta$, and $\Omega_j > 0$ for $1 \leq j \leq k$ such that*

$$||\mathcal{N}(\theta) - s||_{L^2(K)} < \epsilon$$

*where $\mathcal{N}(\theta)$ employs $F_{\Omega_j}$ and as its activation in the jth hidden layer, where $F_{\Omega_j}(x) = F(\frac{1}{\Omega_j} x)$.*

*Proof.* The proof is by induction with the case $k = 2$ being done in Thm. B.3.

So suppose the theorem is true for deep networks with $k - 1$ hidden layers. As in the proof of Thm. B.3 we start by extending $s$ by zero outside $K$ denoted $\widetilde{s}$ and then constructing a shallow network $\mathcal{N}_s$ that is $\epsilon$ close to $\widetilde{s}$ in the $L^2(\mathbb{R})$ norm.

We will denote the number of neurons in the hidden layer of $\mathcal{N}_s$ by $2n_k$ and write

$$\mathcal{N}_s(x) = \sum_{j=-n_k}^{n_k} b_j F_{\Omega_k}(x - \Omega_k j).$$

The second step is to use Thm. B.3 and the induction hypothesis to build $2n_k$ deep networks with $k_1$ hidden layers that approximate the function $x - j$ over $K$ for each $-n_k \leq j \leq n_k$.

Using the fact that $F$ is continuous we then put these two networks together and obtain the required $k$ hidden layer deep network. □

### B.2. Other basis functions

This work has been considered with the application of sampling theory to the understanding of optimum activation functions for neural network interpolation. Our focus has primarily been on those functions that can generate a Riesz basis or a weak Riesz basis. However, as is well known in interpolation theory there are several other basis functions that can theoretically perform interpolation over various function spaces.

**Hermite basis functions:** An example is given by **Hermite polynomials** which are defined by

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2}.$$

These polynomials form a basis for the space $L^2(\mu)$ where $\mu$ is the Gaussian measure $e^{-x^2}dx$. We tested these basis functions against shifted $\mathrm{sinc}$ basis functions on an INR in an image regression task. We compared a $\mathrm{sinc}$ activated INR with a Hermite activated INR to regress an image from the DIV2K dataset. We found that using a sum of degree $k$ Hermite polynomials for $1 \leq k \leq 4$ performed the best. In this application the $\mathrm{sinc}$ INR outperforms the Hermite INR as shown in table 4.

| | **PSNR (dB)** |
|---|---|
| Sinc | 31.2 |
| Hermite | 25.5 |

*Table 4.* Comparison of a sinc INR with a Hermite INR on an image regression task.

**Fourier basis:** Another example of a common basis function used in the literature is the Fourier basis defined by sums of $sin(n\theta)$ and $cos(n\theta)$. Note that as $cos(x) = sin(x + \pi/2)$ these basis functions are covered by $\mathrm{sin}$ activated INRs which we have already shown do not outperform $\mathrm{sinc}$ in the experiments.

## C. Relation to Universal Approximation

Thms. 3.4 and 3.8 can be interpreted as universal approximation theorems for signals in $L^2(\mathbb{R})$. The classic universal approximation theorems are generally for functions on bounded domains (Cybenko, 1989). In 1992 W. A Light extended those results on bounded domains to a universal approximation for continuous function on $\mathbb{R}^n$ by sigmoid activated networks (Light, 1992). His result can also be made to hold for $\mathrm{sinc}$ activated networks, and since the space of continuous functions is dense in $L^2(\mathbb{R})$ his proof easily extends to give a universal approximation result for $\mathrm{sinc}$ activated 2 layer networks for signals in $L^2(\mathbb{R})$. Thus Thm. 3.8 can be seen as giving a different proof of W.A. Light's result.

Although it seems like such results have been known through classical methods, we would like to emphasize that the importance of Thm. 3.8 comes in how it relates to sampling theory. Given a signal $s \in L^2(\mathbb{R})$ that is bandlimited, the Nyquist-Shannon sampling theorem is a classical sampling theorem, see (Marks, 2012), that allows signal reconstruction using shifted $\mathrm{sinc}$ functions while explicitly specifying the coefficients of these shifted $\mathrm{sinc}$ functions. These coefficients correspond to samples of the signal, represented as $s(n/2\Omega)$. In cases where the signal is not bandlimited, Prop. 3.7 still enables signal reconstruction via shifted $\mathrm{sinc}$ functions, albeit without a closed formula for the coefficients involved. This is precisely where Thm. 3.8 demonstrates its significance. The theorem reveals that the shifted $\mathrm{sinc}$ functions constituting the

approximation can be encoded using a two-layer $\mathrm{sinc}$-activated neural network. Notably, this implies that the coefficients can be learned as part of the neural network's weights, rendering such a $\mathrm{sinc}$-activated network exceptionally suited for signal reconstruction in the $L^2(\mathbb{R})$ space. In fact, Thm. 3.8 shows that one does not need to restrict to $\mathrm{sinc}$ functions and that any activation that forms a Riesz basis will be optimal.

## D. Relation to Filtering

In this section, we place our work within the context of filtering, a common technique for removing noise from signals. Real-world signals are often noisy, so when training a signal with noise using an INR, it is natural to ask if the INR can distinguish between the high-frequency components of the signal and the noise. Each sinc function has a bandwidth hyperparameter that determines the frequencies it contains (via its Fourier transform). Since noise typically consists of very high frequencies, ensuring that the bandwidth hyperparameter of the sinc activation in the INR is not too high will prevent the INR from capturing the noise.

To implement this practically across different signal modalities, the bandwidth hyperparameter can be made learnable. Additionally, a regularizer can be added to the loss function to constrain the learnable bandwidth to stay below a certain threshold frequency corresponding to the noise frequency. The drawback of this approach is that it would require more training iterations, as the specific frequencies present in the noise are not known a priori.

## E. Taken's embedding theorem

Taken's embedding theorem is a delay embedding theorem giving conditions under which the strange attractor of a dynamical system can be reconstructed from a sequence of observations of the phase space of that dynamical system.

The theorem constructs an embedding vector for each point in time

$$x(t_i) = [x(t_i), x_(t_i + n\Delta t), \ldots, x(t_i + (d-1)n\Delta t)]$$

Where $d$ is the embedding dimension and $n$ is a fixed value. The theorem then states that in order to reconstruct the dynamics in phase space for any $n$ the following condition must be met

$$d \geq 2D + l$$

where $D$ is the box counting dimension of the strange attractor of the dynamical system which can be thought of as the theoretical dimension of phase space for which the trajectories of the system do not overlap.

**Drawbacks of the theorem:** The theorem does not provide conditions as to what the best $n$ is and in practise when $D$ is not known it does not provide conditions for the embedding dimension $d$. The quantity $n\Delta t$ is the amount of time delay that is being applied. Extremely short time delays cause the values in the embedding vector to almost be the same, and extremely large time delays cause the value to be uncorrelated random variables. The following papers show how one can find the time delay in practise (Kim et al., 1999; Small, 2005). Furthermore, in practise estimating the embedding dimension is often done by a false nearest neighbours algorithm (Kennel et al., 1992).

Thus in practise time delay embeddings for the reconstruction of dynamics can require the need to carry further experiments to find the best time delay length and embedding dimension.

## F. Dynamical equations

**Lorentz System:** For the Lorenz system we take the parameters, $\sigma = 10$, $\rho = 28$ and $\beta = \frac{8}{3}$. The equations defining the system are:

$$\frac{dx}{dt} = \sigma(-x + y) \tag{59}$$

$$\frac{dy}{dt} = -xz + \rho x - y \tag{60}$$

$$\frac{dz}{dt} = -xy - \beta z \tag{61}$$

**Van der Pol Oscillator:** For the Van der Pol oscillator we take the parameter, $\mu = 1$. The equations defining the system are:

$$\frac{dx}{dt} = \mu(x - \frac{1}{3}x^3 - y) \tag{62}$$

$$\frac{dy}{dt} = \frac{1}{\mu}x \tag{63}$$

**Rössler System:** For the Rössler system we take the parameters, $a = 0.2$, $b = 0.2$ and $c = 5.7$. The equations defining the system are:

$$\frac{dx}{dt} = -(y + z) \tag{64}$$

$$\frac{dy}{dt} = x + ay \tag{65}$$

$$\frac{dz}{dt} = b + z(x - c) \tag{66}$$