Abstract. It is known that for a non-linear dynamical system, periodic and quasi-periodic attractors can be reconstructed in a discrete sense using time-delay embedding. Following this argument, it has been shown that even chaotic non-linear systems can be represented as a linear system with intermittent forcing. Although it is known that linear models such as those generated by the Hankel Dynamic Mode Decomposition can - in principle - reconstruct any ergodic dynamical system, quantitative details such as the required sampling rate and the number of delays remain unknown. This work addresses fundamental issues related to the structure and conditioning of linear time delayed models of non-linear dynamics on an attractor. First, we prove that, for scalar systems, the minimal number of time delays required for perfect signal recovery is solely determined by the sparsity in the Fourier spectrum. For the vector case, we devise a rank test and provide a geometric interpretation of the necessary and sufficient conditions for the existence of an accurate linear time delayed model. Further, we prove that the output controllability index of a certain associated linear system serves as a tight upper bound on the minimal number of time delays required. An explicit expression for the exact representation of the linear model in the spectral domain is also provided. From a computational implementation perspective, the effect of the sampling rate on the numerical conditioning of the time delayed model is examined. As a natural extension of Bazán’s work, an upper bound on the 2-norm condition number is derived, with the implication that conditioning can be improved with additional time delays and/or decreasing sampling rates. Finally, it is explicitly shown that the underlying dynamics can be accurately recovered using only a partial period of trajectory data.

Key words. time delay embedding, time series modeling, Koopman operator, dynamic mode decomposition

AMS subject classifications. 37M99, 65P99, 93C05

1. Introduction. In a general sense, time delay embedding, also known as delay coordinate embedding, refers to the inclusion of history information in dynamical system models. The idea of time delay embedding has been employed in a wide variety of contexts including time series modeling [19], Koopman operators [3, 4, 33, 14] and closure modeling [48]. The use of delays to construct a “rich” feature space for geometrical reconstruction of non-linear dynamical systems is justified by the Takens embedding theorem [63] which states that by using a delay-coordinate map, one can construct a diffeomorphic shadow manifold from univariate observations of the original system in the generic sense, and its extensions in a measure-theoretic sense [57], filtered memory [57], deterministic/stochastic forcing [60, 61], and multivariate embeddings [22].

Time delay embedding naturally arises in the representation of the evolution of partially observed states in dynamical systems. As an illustrative example, consider a $N$-dimensional linear autonomous discrete dynamical system with $Q$ partially observed (or resolved) states,
where \( \hat{x}^n \in \mathbb{R}^Q, \tilde{x}^n \in \mathbb{R}^{N-Q}, n \in \mathbb{N}, A_{11} \in \mathbb{R}^{Q \times Q}, A_{12} \in \mathbb{R}^{Q \times (N-Q)}, A_{21} \in \mathbb{R}^{(N-Q) \times Q}, A_{22} \in \mathbb{R}^{(N-Q) \times (N-Q)} \). The dynamical evolution of the observed states \( \hat{x} \) is given by:

\[
\hat{x}^{n+1} = A_{11} \hat{x}^n + \sum_{k=0}^{n-1} A_{12} A_{22}^k A_{21} \hat{x}^{n-k} + A_{12} A_{22}^n \tilde{x}^0.
\]

Typically, the last term is of a transient nature, and thus the above equation can be considered to be closed in the observed variables \( \hat{x} \). The second term on the right hand side of (1.2) describes how the time-history of the observed modes affects the dynamics. Thus, Equation (1.2) implies that it is possible to extract the dynamics of the observables \( \hat{x} \) using time delayed observables, i.e., \( \hat{x}^{n+1} = C_0 \hat{x}^n + \sum_{k=1}^{L} C_k \hat{x}^{n-k} \), where \( C_k \in \mathbb{R}^{Q \times Q} \), and \( L \) is the number of time delays. It should, however, be noted that explicit delays might not be necessary if one has access to high order time derivatives [63] or abundant distinct observations [22].

Leveraging delay coordinates to construct predictive models of dynamical systems has been a topic of great interest. As an example, such models have been studied extensively in the time series analysis community via the well-known family of autoregressive and moving average (ARMA) models [9]. In the machine learning community, related ideas are used in feedforward neural networks (FNN) that augment input dimensions with time delays [24], time-delay neural networks (TDNN) [38, 52, 11] that statically perform convolutions in time, and the family of recurrent neural networks (RNN) [27] that dynamically perform non-linear convolutions in time [43]. In a dynamical systems context, time delays are leveraged in higher order or Hankel Dynamic Mode Decomposition [39, 3, 14]. Although in essence, each community relies on approximations with time-delays, the focus is typically on different aspects: the time series community focuses on stochastic problems, and prefer explicit and interpretable models [9]; the machine learning community is typically more performance-driven and focuses on minimizing the error and scalability [52]; the dynamical systems community is focused on the regulated, continuous dynamical system and interpretability of temporal behavior in terms of eigenvalues and eigenvectors [32]. Moreover, the scientific computing community emphasizes very high dimensional settings, as exemplified by fluid dynamics.

A relevant and outstanding question in each of the aforementioned contexts is the following: Given time series data from a non-linear dynamical system, how much memory is required to accurately recover the underlying dynamics, given a model structure? The memory can be characterized by the two hyperparameters, namely the number of time delays and the corresponding data sampling intervals, if uniformly sampled. Takens embedding theorem [63] only proves the generic existence of a time delayed system, given the model has enough non-linearity to approximate the diffeomorphism, but not how to determine the number of delays and the sampling rate.

It is important to note that the required number of time delays depends on the non-linearity of the approximation model. For instance, in a general setting, the number of dominant principal components of the time delay vectors [13, 26] is only a rough indicator of
the underlying dimension of the manifold that embeds the dynamical system. With sufficient non-linearity in the model, the determination of the hyperparameters can be considered equivalent to the problem of phase-space reconstruction [24, 1]. Popular methods include the false nearest neighbor method [34], singular value analysis [12], averaged mutual information [62], saturation of system invariants [1], box counting methods [56], standard model selection techniques [17], and even reinforcement learning [41]. On the other hand, for linear models, criteria based on statistical significance such as the model utility F-test [42] or information theoretic techniques such as AIC/BIC [9] are used. The use of the partial autocorrelation in linear autoregressive (AR) models to determine the number of delays can be categorized as a model selection approach. It should be mentioned that by treating the models as a black-box, a general approach such as cross validation can be leveraged.

When the sampling rate is fixed, the question of the number of time delays required should not be confused with the length of statistical dependency between the present and past states on the trajectory. For example, an AR(2) model can have a long time statistical dependency, but the number of time delays in the model may be very small. Indeed, it has been explicitly shown [48] that for a non-linear dynamical system with dual linear structure, embedding the memory in a dynamic fashion requires a much smaller number of delays compared to a prescribed static model structure [28].

From the viewpoint of discovering the dynamics of a partially observed system, the goal is to determine the non-linear convolution operator [20, 28] or the so-called closure dynamics [48]. It has to be recognized that the number of time delays will also be dependent on the specific structure of the model. The interchangeability between the number of distinct observables and the number of time delays is also reflected in Takens’ original work on the embedding theorem [63]. Such interchangeability with the latent space dimension is also explored in closure dynamics [48, 28, 51] and recurrent neural networks [27]. Since the required number of delays is strongly dependent on the model structure, it is prudent to first narrow down to a specific type of model, and then determine the delays needed.

For an ergodic dynamical system, assuming that the observable belongs to a finite-dimensional Koopman invariant subspace \( \mathcal{H} \), Arbabi and Mezić [3] showed that Hankel-DMD, a linear model, can provide an exact representation of the Koopman eigenvalues and eigenfunctions in \( \mathcal{H} \). This pioneering work, together with several numerical investigations on the application of Hankel-DMD to non-linear dynamical systems [18, 39, 14] and theoretical studies on time-delayed observables using singular value decomposition (SVD) [33] highlight the ability of linear time delayed models to represent non-linear dynamics. From a heuristic viewpoint, SVD has been demonstrated [12, 13, 26] to serve as a practical guide to determine the required number of time delays and sampling rate, for linear models.

It should be noted that much of the literature [64, 59, 15] related to DMD and Hankel-DMD consider SVD projection either in the time delayed dimension (e.g. singular spectrum analysis) or the state dimension. SVD can provide optimal linear coordinates to maximize signal-to-noise ratio [26], and thus promote robustness and efficiency. On the other hand, projection via Fourier transformation enables the possibility of additional theoretical analysis. For instance, Fourier-based analysis of the Navier–Stokes equations include non-linear triadic wave interactions [54] and decomposition into solenoidal and dilatational components [49]. Pertinent to the present work, ergodic systems characterized by periodic or quasi-periodic
attractors have been shown to be well approximated by Fourier analysis [58, 55, 46]. Fourier analysis has also been employed to approximate the transfer function to obtain an intermediate discrete-time reduced order model with stability guarantees for very large scale linear systems [67, 30]. For general phase space reconstruction, asymptotic decay rates from Fourier analysis have been leveraged to infer appropriate sampling intervals and number of delays [40]. We thus leverage a Fourier basis representation to uncover the structure of time delay embeddings in linear models of non-linear dynamical systems. We also address related issues of numerical conditioning. It should be emphasized that this work is purely concerned with deterministic linear models and noise free data. It can also be shown that SVD becomes equivalent to Fourier analysis in the limit of large windows [26].

The manuscript is organized as follows: The problem formulation and model structure is presented in section 2. Following this, the Fourier transformation of the problem and main theoretical results regarding the minimal time delay embedding for both scalar and vector time series together with explicit, exact solutions of the delay transition matrix after Fourier transformation are presented in sections 3 and 4. Numerical implementation and theoretical results related to conditioning issues is presented in section 5. Theoretical results are verified numerically on a scalar signal and the Van der Pol system in section 6. The main contributions of the work are summarized in section 7.

2. Linear model with time-delay embedding. Consider a continuous autonomous dynamical system,

\[
\frac{d}{dt} x = F(x(t)),
\]

on a state space \(M \subset \mathbb{R}^J, J \in \mathbb{N}^+\), where \(x\) is the coordinate vector of the state, \(x \in M\), \(F(\cdot) : M \rightarrow \mathbb{R}^J\) is in \(C^\infty\). Denote \(\phi_t(x_0)\), i.e., the flow generated by (2.1) as the state at time \(t\) of the dynamical system that is initialized as \(x(0) = x_0 \in M\). By uniformly sampling with time interval \(\Delta t\), the trajectory data of the dynamical systems can be obtained as \(\{x_j\}_{j=0}^\infty\), where \(x_j \doteq x(j \Delta t), j \in \mathbb{N}\).

The aforementioned linear model with time-delay embedding order \(L\) assumes that the predicted future state \(\hat{x}_{j+1}\) is a sum of \(L + 1\) linear mappings from the present state \(x_j\) and previous \(L\) states \(\{x_{j-l}\}_{l=1}^L\), \(j \in \mathbb{N}\),

\[
\hat{x}_{j+1} = W_0 x_j + W_1 x_{j-1} + \ldots + W_L x_{j-L},
\]

where \(W_l \in \mathbb{R}^{J \times J}\) is the associated weight matrix for the \(l\)-th time delay snapshot, \(l = 0, \ldots, L\). The goal is to determine the weight matrices that result in the best possible \(\hat{x}_{j+1}\) approximation to the true future state \(x_{j+1}\). The weight matrices are sought by minimizing the residuals in the Frobenius norm from data in the following form:

\[
W_0, \ldots, W_L = \arg\min_{W_0, \ldots, W_L \in \mathbb{R}^{J \times J}} \|x_{j+1} - (W_0 x_j + W_1 x_{j-1} + \ldots + W_L x_{j-L})\|_F^2.
\]

It is important to note that data-driven models such as AR, VAR [9], SSA [66], HAVOK [14], Hankel-DMD [3] or HODMD [39], can be derived from the above setup by introducing stochasticity, analyzing the eigenspectrum on the principal components, or adding intermittent forcing as inputs.
We restrict ourselves to the dynamics on a periodic attractor, where one can determine an arbitrarily close Fourier interpolation in time with a proper uniform sampling rate [5]. In addition, without loss of generality, we assume that the data is zero mean, i.e., \( \frac{1}{\mathbb{R}^r} \mathbf{x}(\tau) d\tau = 0 \). We start with the scalar case, and extend the corresponding results to vector case in section 4. Note that the data is collected from uniformly sampling a \( T \)-periodic time series \( x(t) \in \mathbb{R} \). The number of samples per period is \( M \), with uniform sampling interval \( \Delta t = T/M \).

Without loss of generality, we assume that sampling is initiated at \( t = 0 \), \( x_{tk} = x(tk), \) \( tk = k\Delta t \), \( k \in I_M, I_M = \{0, 1, \ldots, M - 1 \} \) and \( T \) is the smallest positive real number that represents the periodicity.

### 2.1. Projection of the trajectory on a Fourier basis.

With the above conditions, we consider a surrogate signal of \( x(t) \): \( S_M(t) \)

\[
S_M(t) = \sum_{k \in \mathcal{I}_M} a_k e^{-j \frac{2\pi k t}{T}} \quad \text{with} \quad a_i = \frac{1}{M} \sum_{k \in \mathcal{I}_M} x_k e^{j \frac{2\pi i k}{M}} \in \mathbb{C},
\]

where \( j = \sqrt{-1} \) and

\[
\forall k \in \mathcal{I}_M, \quad x_k = x(k\Delta t) = S_M(k\Delta t),
\]

which is obtained by projecting \( x(t) \) on the following linear space \( \mathcal{H}_F \)

\[
\mathcal{H}_F = \text{span}\{1, e^{-j \frac{2\pi t}{T}}, \ldots, e^{-j \frac{2\pi (M-1)t}{T}}\},
\]

which is spanned by the Fourier basis in (2.6) with test functions as delta functions as \( \delta(t - tk) \), \( k \in \mathcal{I}_M \). This process is equivalent to the discrete Fourier transform (DFT).

The above procedure naturally represents the uniformly sampled trajectory in the time domain \( \{x_k \}_{k=0}^{M-1} \) using coefficients in the frequency domain \( \{a_i \}_{i=0}^{M-1} \). Since we consider real signals, \( \{a_i \}_{i=0}^{M-1} \) possess reflective symmetry: \( \forall i \in \mathcal{I}_M, \text{Re}(a_i) = \text{Re}(a_{M-i}), \text{Im}(a_i) + \text{Im}(a_{M-i}) = 0, \) where \( \text{Re} \) and \( \text{Im} \) represent the real and imaginary part of a complex number. In addition, since \( T \) is the smallest period by definition, we must have \( a_1 = a_{\frac{T}{M-1}} \neq 0 \). Further, since \( \mathbf{F} \) is smooth, the flow \( \phi_t(x_0) = x(t) \) is also smooth in \( t \) [47]. Thus, the error in the Fourier interpolation is uniformly bounded by twice the sum of the absolute value of truncated Fourier coefficients [10]. This leads to the uniform convergence

\[
\lim_{M \to \infty} |x(t) - S_M(t)| = 0.
\]

Hence, one can easily approximate the original periodic trajectory uniformly to the desired level of accuracy by increasing \( M \) above a certain threshold.

### 3. The structure of time delay embedding for scalar time series.

Now, we apply the linear model with time-delay embedding ((2.2)) at the locations \( \{x_k \}_{k=0}^{M-1} \). Given \( \{x_k \}_{k=0}^{M-1} \) consider constructing \( L \)-time delays of \( x(t) \), \( L \in \mathbb{N} \). Note that \( L = 0 \) corresponds to no delays considered. To avoid negative indices, we utilize the modulo operation defined in (3.1),

\[
\forall q \in \mathbb{N}, \mathcal{P}(q) \equiv q \text{ (mod } M) = \begin{cases} 
q, & \text{if } q \in \mathcal{I}_M, \\
q - M \lfloor q/M \rfloor, & \text{otherwise}
\end{cases}
\]
to construct the $L$ time-delay vector $Y_k$,

$$Y_k = \begin{bmatrix} x_{p(k)} \\ x_{p(k-1)} \\ \vdots \\ x_{p(k-L)} \end{bmatrix} \in \mathbb{R}^{L+1},$$  

where $k \in \mathcal{I}_M$, $\lfloor \cdot \rfloor$ is the floor function. Considering Fourier interpolation, we have

(3.3) \hspace{1cm} \forall q \in \mathcal{I}_M, \hspace{.2cm} x_{P(q)} = \sum_{i \in \mathcal{I}_M} a_i \omega^{q_i}, \hspace{.2cm} \omega = e^{-j \frac{2\pi q}{M}} \in \mathbb{C},

which is also true for $q \notin \mathcal{I}_M$

(3.4) \hspace{1cm} x_{P(q)} = S_M((q - M \lfloor q/M \rfloor)\Delta t) = \sum_{i \in \mathcal{I}_M} a_i e^{-j \frac{2\pi (q - M \lfloor q/M \rfloor)}{M}} = \sum_{i \in \mathcal{I}_M} a_i \omega^{q_i}.

Using (2.4), we can rewrite the $L$ time-delay vector $Y_k$ in (3.2) in the Fourier basis as

(3.5) \hspace{1cm} Y_k = \Omega_{k,L} a,

where $\forall k \in \mathcal{I}_M$, $\Omega_{k,L} = \begin{bmatrix} 1 & \omega^k & \omega^{2k} & \ldots & \omega^{(M-1)k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{k-L} & \omega^{2(k-L)} & \ldots & \omega^{(M-1)(k-L)} \end{bmatrix}$, $a = \begin{bmatrix} a_0 \\ \vdots \\ a_{M-1} \end{bmatrix} \in \mathbb{C}^{M \times 1}$.

The problem of the minimal time delay required for the linear model with $L$ time delays in (2.2) to perfectly predict the data $\{x_k\}_{k=0}^{M-1}$ is equivalent to the existence of the delay transition matrix $K$ such that,

(3.6) \hspace{1cm} x_{p(k+1)} = K^T Y_k, \hspace{.2cm} \forall k \in \mathcal{I}_M,

where

$$K = \begin{bmatrix} K_0 & K_1 & \ldots & K_L \end{bmatrix}^T \in \mathbb{R}^{(L+1) \times 1},$$

and

(3.7) \hspace{1cm} x_{p(k+1)} = Y_k^T a,

where

(3.8) \hspace{1cm} Y_k = \begin{bmatrix} 1 & \omega^k & \omega^{2k} & \ldots & \omega^{(M-1)k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{k-L} & \omega^{2(k-L)} & \ldots & \omega^{(M-1)(k-L)} \end{bmatrix}^T.

For convenience, we vertically stack (3.6) $\forall k \in \mathcal{I}_M$,

(3.9) \hspace{1cm} Y_M K = x_M,

where $Y_M = \begin{bmatrix} Y_0^T \\ Y_1^T \\ \vdots \\ Y_{M-2}^T \\ Y_{M-1}^T \end{bmatrix}$, $x_M = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{M-1} \\ x_0 \end{bmatrix}.$

In the following subsections, we discuss the minimal number of required time delays, the exact solution of $K$ and the number of samples required on the time domain.
3.1. Minimal number of time delays. Our goal is to determine the minimal number of time delays $L$, such that there exists a matrix $K$ that satisfies the linear system (3.6). Given one period of data, we can transform the system from the time domain to the spectral domain. Consider (3.5) and (3.7), then (3.9) is equivalent to the following, $\forall k \in \mathcal{I}_M$:

\[
\mathbf{a}^T \begin{bmatrix}
1 & \omega & \omega^2 & \cdots & \omega^{(M-1)} \\
\omega^k & \omega^2 & \cdots & \omega^{k-L} \\
\vdots & \vdots & \ddots & \vdots \\
\omega^{(k+1)(M-1)} & \omega^{(M-1)k} & \cdots & \omega^{(M-1)(k-L)}
\end{bmatrix} K = 0.
\]

This can be written as

\[
\mathbf{a}^T \begin{bmatrix}
1 & \omega & \omega^2 & \cdots & \omega^{(M-1)} \\
\omega & \omega^2 & \cdots & \omega^{L-2} \\
\vdots & \vdots & \ddots & \vdots \\
\omega^{(M-1)} & \omega^{(M-1)2} & \cdots & \omega^{(M-1)(L-2)} \\
\end{bmatrix} K = 0.
\]

We define the residual matrix $R$ as,

\[
R := \begin{bmatrix}
1 & 1 & \cdots & 1 \\
\omega & \omega^{-1} & \cdots & \omega^{L-1} \\
\omega^2 & \omega^2 & \cdots & \omega^{2(L-1)} \\
\vdots & \vdots & \ddots & \vdots \\
\omega^{M-1} & \omega^{(M-1)} & \cdots & \omega^{(M-1)(L-1)}
\end{bmatrix} K.
\]

Given one period of data, we vertically stack the above equation for each $k \in \mathcal{I}_M$. Recognizing the non-singular nature of a Vandermonde square matrix with distinct nodes, we have

\[
\begin{bmatrix}
a_0 & a_1 & a_2 & \cdots & a_{M-1} \\
a_0 & \omega a_1 & \omega^2 a_2 & \cdots & \omega^{M-1} a_{M-1} \\
a_0 & \omega^2 a_1 & \omega^4 a_2 & \cdots & \omega^{2(M-1)} a_{M-1} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
a_0 & \omega^{M-1} a_1 & \omega^{2(M-1)} a_2 & \cdots & \omega^{(M-1)(M-1)} a_{M-1}
\end{bmatrix} R = 0.
\]

This gives

\[
\begin{bmatrix}
1 & 1 & 1 & \cdots & 1 \\
\omega & \omega^2 & \omega^3 & \cdots & \omega^{M-1} \\
1 & \omega^2 & \omega^4 & \cdots & \omega^{2(M-1)} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & \omega^{M-1} & \omega^{2(M-1)} & \cdots & \omega^{(M-1)(M-1)}
\end{bmatrix} \begin{bmatrix}
a_0 \\
a_1 \\
a_2 \\
\vdots \\
a_{M-1}
\end{bmatrix} R = 0,
\]

\[
R = 0.
\]
and thus

\[(3.15) \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{M-1} \end{bmatrix} \mathbf{R} = \mathbf{0}.\]

Note the equivalence between (3.15) and (3.9). Now, we consider the case when the Fourier spectrum is sparse with \(P\) non-zero coefficients, \(P \in \mathbb{N}\) and \(P \leq M\). Denote the set of wave numbers associated with non-zero coefficients as,

\[(3.16) I_P = \{i \neq 0 | i \in I_M\} = \{i_p\}_{p=0}^{P-1},\]

with ascending order \(0 \leq i_0 < i_1 < \ldots < i_{P-1} \leq M - 1\), where \(|I_M| = P \in \mathbb{N}\). Note that there is a reflective symmetry restriction on the Fourier spectrum.

The feasibility of using the number of time delays \(L\) to ensure the existence of a real solution \(\mathbf{K}\) for the linear system is equivalent to the existence of the linear system \(\mathbf{R} = \mathbf{0}\) after removing the rows that correspond to zero Fourier modes in \(\mathbf{R}\), denoted as \(\mathbf{R}_{I_M}^P\),

\[(3.17) \mathbf{R}_{I_M}^P = \mathbf{0} \iff \mathbf{A}_{I_M,L}^P \mathbf{K} = \mathbf{b}_{I_M}^P,\]

where

\[(3.18) \mathbf{A}_{I_M,L}^P = \begin{bmatrix} 1 & \omega^{-i_0} & \ldots & \omega^{-Li_0} \\ 1 & \omega^{-i_1} & \ldots & \omega^{-Li_1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{-i_{P-1}} & \ldots & \omega^{-Li_{P-1}} \end{bmatrix} \in \mathbb{C}^{P \times (L+1)}, \quad \text{and} \quad \mathbf{b}_{I_M}^P = \begin{bmatrix} \omega^{i_0} \\ \omega^{i_1} \\ \vdots \\ \omega^{i_{P-1}} \end{bmatrix} \in \mathbb{C}^{P \times 1}.\]

Before presenting the main theorem Theorem 3.4, we define the Vandermonde matrix in Definition 3.1 and introduce Lemma 3.2 and Lemma 3.3.

**Definition 3.1.** Vandermonde matrix with nodes as \(\alpha_0, \alpha_1, \ldots, \alpha_{M-1} \in \mathbb{C}\) of order \(N\) is defined as,

\[
\mathbf{V}_N(\alpha_0, \alpha_1, \ldots, \alpha_{M-1}) = \begin{bmatrix} 1 & \alpha_0 & \ldots & \alpha_0^{N-1} \\ 1 & \alpha_1 & \ldots & \alpha_1^{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \alpha_{M-1} & \ldots & \alpha_{M-1}^{N-1} \end{bmatrix}.
\]

**Lemma 3.2.** \(\forall M, N \in \mathbb{N}\), the Vandermonde matrix \(\mathbf{A} = \mathbf{V}_N(\alpha_0, \alpha_1, \ldots, \alpha_{M-1})\) constructed from distinct \(\{\alpha_i\}_{i \in I_M}, \alpha_i \in \mathbb{C}\), has the two properties,

1. \(\text{rank}(\mathbf{A}) = \min(M, N)\),
2. if \(\mathbf{A}\) has full column rank, \(\forall Q \in \mathbb{N}, Q \leq M\), the rank of the submatrix \(\mathbf{A}'\) by arbitrarily selecting \(Q\) rows is \(\min(Q, N)\).

**Proof.** See Appendix A.1.
Lemma 3.3. \( \forall m, n \in \mathbb{N}, A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^{m \times 1}, \exists x \in \mathbb{C}^{n \times 1} \text{ s.t. } Ax = b \). Further, when the solution is unique, the above still holds and the solution is real.

Proof. See Appendix A.2.

Theorem 3.4. For a uniform sampling of \( S_M(t) \) with length \( M \) and \( P \) non-zero coefficients in the Fourier spectrum, the minimal number of time delays \( L \) for a perfect prediction, i.e., one that satisfies (3.9) is \( P - 1 \). Moreover, when \( L = P - 1 \), the solution is unique.

Proof. Consider the discrete Fourier spectrum of \( S_M(t) \) with \( M \) uniform samples per period. The perfect prediction using a time-delayed linear model requires the existence of a real \( K \) that satisfies (3.9), which is equivalent to (3.15). Therefore, (3.9) and (3.15) share the same solutions in \( \mathbb{C}^{(L+1) \times 1} \). Since the Fourier spectrum contains only \( P \) non-zero coefficients, (3.15) is equivalent to (3.17). The necessary and sufficient condition to have a solution (not necessarily real) \( K \) for (3.17) follows from the Rouché-Capelli theorem [45],

\[
\text{rank} \left( \begin{bmatrix} A_{T_M^L} & b_{T_M^L} \end{bmatrix} \right) = \text{rank} \left( \begin{bmatrix} b_{T_M^L} & A_{T_M^L} \end{bmatrix} \right).
\]

Using the first property in Lemma 3.2, \( \text{rank}(A_{T_M^L}) = \min(P, L+1) \). While for the augmented matrix,

\[
\text{rank} \left( \begin{bmatrix} A_{T_M^L} & b_{T_M^L} \end{bmatrix} \right) = \text{rank} \left( \begin{bmatrix} b_{T_M^L} & A_{T_M^L} \end{bmatrix} \right)
= \text{rank} \left( \begin{bmatrix} \omega^{j_0} & 1 & \omega^{-i_0} & \ldots & \omega^{-(L+1)i_0} \\ \omega^{i_1} & 1 & \omega^{-(L+1)i_1} \\ \omega^{i_2} & 1 & \omega^{-(L+1)i_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \omega^{i_{p-1}} & 1 & \omega^{-(L+1)i_{p-1}} \end{bmatrix} \right)
= \text{rank} \left( \begin{bmatrix} \omega^{j_0} & \omega^{j_1} & \omega^{j_2} & \vdots & \omega^{j_{p-1}} \\ \omega^{i_1} & 1 & \omega^{-(L+1)i_1} \\ \omega^{i_2} & 1 & \omega^{-(L+1)i_2} \\ \vdots & \vdots & \ddots & \vdots \\ \omega^{i_{p-1}} & 1 & \omega^{-(L+1)i_{p-1}} \end{bmatrix} \right)
= \text{rank} \left( \text{diag}(\omega^{j_0}, \ldots, \omega^{i_{p-1}}) V_{L+2}(\omega^{-i_0}, \ldots, \omega^{-i_{p-1}}) \right)
= \text{rank} (V_{L+2}(\omega^{-i_0}, \ldots, \omega^{-i_{p-1}}))
= \min(P, L + 2).
\]

Therefore, if \( L + 2 \leq P \), i.e., \( L \leq P - 2 \), \( \min(P, L + 2) = L + 2 \neq L + 1 = \min(P, L + 1) \). If \( L + 1 \geq P \), i.e., \( L \geq P - 1 \), then \( \min(P, L + 2) = P = \min(P, L + 1) \). So the minimal \( L \) for (3.19) to hold is \( P - 1 \), which makes \( A_{T_M^L} \) an invertible Vandermonde square matrix. Thus the solution is unique in \( \mathbb{C}^{(L+1) \times 1} \). From Lemma 3.3, consider (3.9), the solution is real.

From the above Theorem 3.4, we can easily derive Propositions 3.5 and 3.6 that are intuitive.
Proposition 3.5. If there is only one frequency in the Fourier spectrum of $S_M(t)$, simply one time delay in the linear model is enough to perfectly recover the signal.

Proposition 3.6. If the Fourier spectrum of $S_M(t)$ is dense, then the maximum number of time delays, i.e., over the whole period $M - 1$ is necessary to perfectly recover the signal.

3.2. Exact solution for the delay transition matrix $K$. Two interesting facts have to be brought to the fore:

1. From (3.17), it is clear that $K$ is independent of the quantitative value of the Fourier coefficients, but only depends on the pattern in the Fourier spectrum.

2. For $L = P - 1$, $A_{T^L_p}$ is an invertible Vandermonde matrix, which implies the uniqueness of the solution $K$.

Consider the general explicit formula for the inverse of a Vandermonde matrix [53]. Note that $A_{I_p} = \mathcal{V}_p^1(\omega_{-i_0}, \ldots, \omega_{-i_{P-1}})$. Thus

\begin{align}
\mathcal{V}_p^1(\omega_{-i_0}, \ldots, \omega_{-i_{P-1}})_{mn} &= (-1)^{m+1} \sum_{0 \leq l \leq P-1, l \neq n-1} \omega^{-i_l - \omega^{-i_{n-1}}} \\
&= \sum_{n=1}^P (-1)^{m+1} \sum_{0 \leq k_1 < \ldots < k_{P-m} \leq P-1 \atop k_1, \ldots, k_{P-m} \neq n-1} \omega^{-(i_{k_1} + \ldots + i_{k_{P-m}})} \\
&\quad \times \prod_{0 \leq l \leq P-1, l \neq n-1} \omega^{-i_l - \omega^{-i_{n-1}}} \\
&= \sum_{n=1}^P (-1)^{m+1} \sum_{0 \leq k_1 < \ldots < k_{P-m} \leq P-1 \atop k_1, \ldots, k_{P-m} \neq n-1} e^{i2\pi k_l/M} e^{i2\pi n-1} \\
&\quad \times \prod_{0 \leq l \leq P-1, l \neq n-1} e^{i2\pi l/M} - e^{i2\pi n-1}. 
\end{align}

where $1 \leq m, n \leq P$ and $K_m \equiv K_{m-1}$.

Despite the explicit form, the above expression is not useful in practice. Without loss of generality, considering $P$ is even, the computational complexity at least grows as $(P^2/2)$. As an example, for a moderate system with 50 non-sparse modes, $(50^2/2) \approx 1.2 \times 10^{14}$.

3.3. Eigenstructure of the companion matrix. The eigenstructure of the companion matrix formed with time delays is closely related to the Koopman eigenvalues and eigenfunctions under ergodicity assumptions [3]. From the viewpoint of HAVOK [14], for a general time delay $L$, the corresponding Koopman eigenvalues are eigenvalues of the companion matrix $K_{\text{comp}}$.
defined as,

\[
Y_{k+1} = Y_k K_{\text{comp}}, \quad K_{\text{comp}} = \begin{bmatrix}
K_0 & 1 & 0 & \ldots & 0 \\
K_1 & 0 & 1 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
K_{L-1} & 0 & 0 & \ldots & 1 \\
K_L & 0 & 0 & \ldots & 0
\end{bmatrix} \in \mathbb{R}^{(L+1) \times (L+1)}.
\]

The corresponding eigenvalues satisfy \( \det(\lambda I - K_{\text{comp}}) = 0 \), i.e.,

\[
\lambda^{L+1} - K_0 \lambda^L - \ldots - K_L = 0.
\]

The corresponding eigenstructure is fully determined by the eigenvalues \([23], \lambda_0, \ldots, \lambda_L\), i.e.,

\[
K_{\text{comp}} = Q^{-1} \Lambda Q,
\]

where \( \Lambda = \text{diag}(\lambda_0, \ldots, \lambda_L) \), \( Q = V_{L+1}(\lambda_0, \ldots, \lambda_L) \).

3.3.1. Special case: dense Fourier spectrum. Note that \( \omega^{-M} = 1 \) and \( P = M \). Consider \( L = P - 1 = M - 1 \), so that the last column of \( A_{PM,L} \) becomes

\[
\begin{bmatrix}
1 \\
\omega^{-M} \\
\omega^{-2(M-1)} \\
\vdots \\
\omega^{-(M-1)(M-1)}
\end{bmatrix} = \begin{bmatrix}
1 \\
\omega \\
\omega^2 \\
\vdots \\
\omega^{M-1}
\end{bmatrix} = b_{LM}.
\]

Therefore, the unique solution can be found from observations as

\[
K = \begin{bmatrix} 0 & \ldots & 0 & 1 \end{bmatrix}^T.
\]

The companion matrix \([3]\) associated with the Koopman operator is in the form of a special circulant matrix \([45]\), for which analytical eigenvalues and eigenvectors can be easily determined. In (3.23), we have

\[
K_{\text{comp}} = \begin{bmatrix}
0 & 1 & 0 & \ldots & 0 \\
0 & 0 & 1 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & 1 \\
1 & 0 & 0 & \ldots & 0
\end{bmatrix} \in \mathbb{R}^{M \times M},
\]

which has eigenvalues evenly distributed on the unit circle

\[
\forall i \in \mathbb{Z}_M, \quad \lambda_i = e^{-j \frac{2\pi i}{M}} = \omega^i,
\]

and normalized eigenvectors as

\[
\nu_i = \frac{1}{\sqrt{M}} \begin{bmatrix}
1 & \omega^{-i} & \omega^{-2i} & \ldots & \omega^{-(M-1)i}
\end{bmatrix}^T.
\]
3.4. Analysis in the time domain. Projection of the trajectory onto a Fourier basis implies that at least one period of training data has to be obtained to be able to construct a linear system that has a unique solution corresponding to \( \mathbf{K}^* \). However, we will show that in the time domain, a full period of data is not necessary to determine the solution \( \mathbf{K}^* \) if the Fourier spectrum is sparse.

Denote the number of non-zero Fourier coefficients as \( P \in \mathbb{N} \), and its index set as \( I_M^P \) as before. Instead of having a full period of data, without loss of generality, we consider \( L \) time delays and select the \( Q \) rows in (3.9), for which the index is denoted as \( 0 \leq k_0 < \ldots < k_{Q-1} \leq M-1 \), and \( Q \in \mathbb{N}, L + Q \leq M \). Therefore, we have the following equation in the time domain,

\[
\begin{bmatrix}
Y^T_{k_0} \\
Y^T_{k_1} \\
\vdots \\
Y^T_{k_{Q-2}} \\
Y^T_{k_{Q-1}}
\end{bmatrix}
\begin{bmatrix}
X_P(k_0+1) \\
X_P(k_1+1) \\
\vdots \\
X_P(k_{Q-2}+1) \\
X_P(k_{Q-1}+1)
\end{bmatrix}.
\]

(3.31)

Consider a Fourier transform and recall (3.11). Choosing \( k \) over \( k_0, \ldots, k_{Q-1} \), the above equation can be equivalently rewritten as

\[
\begin{bmatrix}
a_0 & \omega^{k_0}a_1 & \omega^{2k_0}a_2 & \ldots & \omega^{(M-1)k_0}a_{M-1} \\
a_0 & \omega^{k_1}a_1 & \omega^{2k_1}a_2 & \ldots & \omega^{(M-1)k_1}a_{M-1} \\
a_0 & \omega^{k_2}a_1 & \omega^{2k_2}a_2 & \ldots & \omega^{(M-1)k_2}a_{M-1} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
a_0 & \omega^{k_{Q-1}}a_1 & \omega^{2k_{Q-1}}a_2 & \ldots & \omega^{(M-1)k_{Q-1}}a_{M-1}
\end{bmatrix}
\begin{bmatrix}
a_0 \\
a_1 \\
a_2 \\
\vdots \\
a_{Q-1}
\end{bmatrix} = \mathbf{0}.
\]

(3.32)

Recall that only \( P \) Fourier coefficients are non-zero, and thus the above equation that constrains \( \mathbf{K} \) equivalently becomes

\[
\begin{bmatrix}
a_{i_0} & \omega^{k_0}a_{i_1} & \omega^{2k_0}a_{i_2} & \ldots & \omega^{(P-1)k_0}a_{i_{P-1}} \\
a_{i_0} & \omega^{k_1}a_{i_1} & \omega^{2k_1}a_{i_2} & \ldots & \omega^{(P-1)k_1}a_{i_{P-1}} \\
a_{i_0} & \omega^{k_2}a_{i_1} & \omega^{2k_2}a_{i_2} & \ldots & \omega^{(P-1)k_2}a_{i_{P-1}} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
a_{i_0} & \omega^{k_{Q-1}}a_{i_1} & \omega^{2k_{Q-1}}a_{i_2} & \ldots & \omega^{(P-1)k_{Q-1}}a_{i_{P-1}}
\end{bmatrix}
\begin{bmatrix}
a_{i_0} \\
a_{i_1} \\
a_{i_2} \\
\vdots \\
a_{i_{P-1}}
\end{bmatrix} = \mathbf{0}.
\]

(3.33)

\[
\begin{bmatrix}
1 & \omega^{k_0} & \omega^{2k_0} & \ldots & \omega^{(P-1)k_0} \\
1 & \omega^{k_1} & \omega^{2k_1} & \ldots & \omega^{(P-1)k_1} \\
1 & \omega^{k_2} & \omega^{2k_2} & \ldots & \omega^{(P-1)k_2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & \omega^{k_{Q-1}} & \omega^{2k_{Q-1}} & \ldots & \omega^{(P-1)k_{Q-1}}
\end{bmatrix}
\begin{bmatrix}
a_{i_0} \\
a_{i_1} \\
a_{i_2} \\
\vdots \\
a_{i_{P-1}}
\end{bmatrix} = \mathbf{0}.
\]

(3.34)

Since \( \{\omega^k\}_{j=0}^{Q-1} \) are distinct from each other, from Lemma 3.2, \( \text{rank}(\mathbf{V}_P(\omega^{k_0}, \ldots, \omega^{k_{Q-1}})) = \min(P, Q) \). Therefore, if we choose to have training data points no less than the number of
non-zero Fourier coefficients, i.e., \( Q \geq P \), then \( V_P(\omega^{k_0}, \ldots, \omega^{k_{Q-1}}) \) is full rank, which leads to \( R_{I_M} = 0 \). Meanwhile, the solution \( K \) is uniquely determined given \( L = P - 1 \). Therefore, given \( \hat{Q} \geq P \),

\[
\begin{bmatrix}
\tilde{Y}_k^T \\
\vdots \\
\tilde{Y}_{kQ-2}^T \\
\tilde{Y}_{kQ-1}^T
\end{bmatrix}
K =
\begin{bmatrix}
x_P(k_{0}+1) \\
x_P(k_1+1) \\
\vdots \\
x_P(k_{Q-2}+1) \\
x_P(k_{Q-1}+1)
\end{bmatrix}
\iff R_{I_M} = 0 \iff K = K^*.
\]

For the case with minimal number of data samples, i.e., \( Q = P \), a natural choice is to construct \( P \) rows of the future state from the \( P \)-th to \( 2P - 1 \)-th rows in (3.9). In the above setting, in order to construct the linear system in time domain that has the unique solution \( K^* \) of (3.17), we only require access to the first \( 2P \) snapshots of data. The key observation is that when the signal is sparse, instead of constructing the classic unitary DFT matrix ((3.14) to (3.15)), a random choice of \( P \) rows will be sufficient to uniquely determine a real solution \( K^* \).

It has to be mentioned, however, that randomly chosen data points might not be optimal. For example, in (3.33), the particular choice of sampling (i.e. the choice of \( Q \) rows), will determine the condition number of the complex Vandermonde matrix \( V_P(\omega^{k_0}, \ldots, \omega^{k_{Q-1}}) \). The necessary and sufficient condition for perfect conditioning of a Vandermonde matrix is when \( \{\omega^j\}_{j=0}^{Q-1} \) are uniformly spread on the unit circle [7].

The above instance of accurately recovering the dynamical system without using a full period of data on the attractor is also reported elsewhere, for instance in sparse polynomial regression for data-driven modeling of dynamical systems [18]. Indeed, this is one of the key ideas behind SINDy [16]: one can leverage the prior knowledge of the existence of a sparse representation (for instance, in a basis of monomials), such that sparse regression can significantly reduce the amount of data required with no loss of information.

4. Extension of the analysis to the vector case. In this section, we extend the above analysis to the case of a vector dynamical system. Assuming the state vector has \( J \) components, given the time series of \( l \)-th component, \( \{x_k^{(l)}\}_{k=0}^{M-1}, l = 1, \ldots, J \), we have, \( \forall k \in I_M \)

\[
\tilde{x}_P(k+1) =
\begin{bmatrix}
x_P^{(1)}(k+1) \\
x_P^{(2)}(k+1) \\
\vdots \\
x_P^{(J)}(k+1)
\end{bmatrix}
\in \mathbb{R}^{J \times 1},
\]

where \( k \in I_M, \forall 1 \leq l \leq J, l \in \mathbb{N}, x_P^{(l)}(k) \in \mathbb{R}, J \in \mathbb{N} \). Rewrite (3.6) in a vector form:

\[
\tilde{x}_P(k+1) = \tilde{K}^T \tilde{Y}_k, \ \forall k \in I_M,
\]

where \( \tilde{x}_P(k+1) \in \mathbb{R}^J, \tilde{K} \in \mathbb{R}^{J(L+1) \times J} \) and

\[
\tilde{Y}_k =
\begin{bmatrix}
y_k^{(1)} \\
\vdots \\
y_k^{(J)}
\end{bmatrix}
\in \mathbb{R}^{J(L+1) \times 1},
\]
where $\mathbf{Y}_k^{(l)}$ are the $L$ time-delay embeddings defined in (3.2) for the $l$-th component of the state. In the present work, we treat the time-delay uniformly across all components.

Following similar procedures as before, denoting the Fourier coefficient of $l$-th component as $\mathbf{a}^{(l)} \in \mathbb{C}^{M \times 1}$, the following lemma which is an analogy to (3.15) in the scalar case.

**Lemma 4.1.** The necessary and sufficient condition for the existence of a real solution $\tilde{\mathbf{K}}$ in (4.2) is equivalent to the existence of a solution for the following linear system:

$$
\begin{align*}
(\text{4.4}) \quad & \left[ \text{diag}(\mathbf{a}^{(1)}) \quad \ldots \quad \text{diag}(\mathbf{a}^{(J)}) \right] \\
& \times \begin{bmatrix}
\mathbf{b}^M_T \\
\vdots \\
\mathbf{b}^M_T \\
\end{bmatrix} \\
& - \begin{bmatrix}
\mathbf{A}^M_T,L \\
\vdots \\
\mathbf{A}^M_T,L \\
\end{bmatrix} \tilde{\mathbf{K}} = \mathbf{0}.
\end{align*}
$$

The existence of the above solution is equivalent to the following relationship,

$$
\text{(4.5)} \quad \text{rank} \left[ \left[ \text{diag}(\mathbf{a}^{(1)})\mathbf{A}^M_T,L \quad \ldots \quad \text{diag}(\mathbf{a}^{(J)})\mathbf{A}^M_T,L \right] \right]
= \text{rank} \left[ \left[ \text{diag}(\mathbf{a}^{(1)})\mathbf{A}^M_T,L \quad \ldots \quad \text{diag}(\mathbf{a}^{(J)})\mathbf{b}^M_T \right] \right].
$$

**Proof.** See Appendix A.3.

Next, with the introduction of the Krylov subspace in Definition 4.2, we present Remark 4.3 and Remark 4.4 from (4.4) that interprets and reveals the possibility of using less embeddings than the corresponding sufficient condition for the scalar case in Theorem 3.4.

**Definition 4.2 (Krylov subspace).** For $n, r \in \mathbb{N}$, $\mathbf{A} \in \mathbb{C}^{n \times n}$, $\mathbf{b} \in \mathbb{C}^{n \times 1}$, Krylov subspace is defined as

$$
\text{(4.6)} \quad K_r(\mathbf{A}, \mathbf{b}) = \text{span}\{\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}^{r-1}\mathbf{b}\}.
$$

**Remark 4.3 (Geometric interpretation).** For $j = 1, \ldots, J$, define $\mathbf{c}^{(j)} = \text{diag}(\mathbf{a}^{(j)})\mathbf{b}^M_T$, and $\mathcal{E}^{(j)}_L$ as the column space of $\text{diag}(\mathbf{a}^{(j)})\mathbf{A}^M_T,L$. The existence of the solution in (4.4) is then equivalent to

$$
\text{(4.7)} \quad \forall j \in \{1, \ldots, J\}, \mathbf{c}^{(j)} \in \mathcal{W}_L \iff \text{span} \{\mathbf{c}^{(1)}, \ldots, \mathbf{c}^{(J)}\} \subseteq \mathcal{W}_L,
$$

where $\mathcal{W}_L$ is the column space from all components, and $\oplus$ is the direct sum operation between vector spaces. Note that the column space of $\mathbf{A}^M_T,L$ can represented as a Krylov subspace $K_{L+1}(\Lambda^{-1}, \mathbf{e})$, where

$$
\text{(4.8)} \quad \mathbf{e} \equiv \begin{bmatrix} 1 & \ldots & 1 \end{bmatrix}^T,
$$

$$
\text{(4.9)} \quad \Lambda \equiv \text{diag}(\omega^0, \ldots, \omega^{M-1}).
$$

A geometric interpretation of the above expressions is shown in Figure 1: for each $j$, $\mathbf{b}^M_T = \Lambda^{-(M-1)}\mathbf{e}$ and $\mathbf{e}$ are projected, stretched and rotated using the $j$-th Fourier spectrum.
diagonal matrix \( \text{diag}(\mathbf{a}(j)) \) yields \( \mathcal{E}_L^{(j)} \) and its total column subspace \( \mathcal{W}_L \). If all of the projected and stretched \( \mathbf{b}_M \)'s are contained in \( \mathcal{W}_L \), a real solution exists for (4.2). Notice that in (4.7), \( \forall i \neq j, \mathcal{E}_L(i) \) expands the column space \( \mathcal{E}_L^{(j)} \) to include \( \mathbf{c}(j) \). Thus, the minimal number of time delays required in the vector case as in (4.2) can be smaller than that of the scalar case.

**Remark 4.4 (Effect of interplay between the Fourier spectra of each component).**

The vector case involves the interaction between the \( J \) different Fourier spectra corresponding to each component of the state. This complicates the derivation of an explicit result for the minimal number of time delays as in the scalar case (Theorem 3.4). We note two important observations that illustrate the impact of the interplay between the \( J \) Fourier spectra:

- To ensure \( \mathbf{c}(j) \) lies in \( \mathcal{W}_L \), each \( \mathcal{E}_L^{(j)} \) should provide distinct vectors to maximize the

Figure 1. Illustration of the geometrical interpretation of Lemma 4.1.
dimension of $W_L$. If a linear dependency is present in $\{a^{(j)}\}_{j=1}^f$, (4.7) no longer holds.

- Since $c^{(j)}$ is projected using $\text{diag}(a^{(j)})$, if $a^{(j)\top}a^{(j)} = 0$, $c^{(j)}_L$ will not contribute to increasing the dimension of $W_L$.

Drawing insight from the representation of the column space of $A_{IM,LM}$ as the Krylov subspace in Remark 4.3, we present a connection between the output controllability from linear system control theory [35], and the number of time delays required for linear models in general sense.

**Definition 4.5 (Output controllability of a general linear system).** Consider a linear system with state vector $x(t) \in \mathbb{C}^{M \times 1}$, $M \in \mathbb{N}$, $t \in \mathbb{R}^+$,

\begin{align}
(4.10) \quad \dot{x} &= Ax + Bu, \\
(4.11) \quad y &= Cx + Du,
\end{align}

where $A \in \mathbb{C}^{M \times M}$, $B \in \mathbb{C}^{M \times N}$, $C \in \mathbb{C}^{P \times M}$, $D \in \mathbb{C}^{P \times N}$. $y(t) \in \mathbb{C}^{P \times 1}$ is the output vector. The above system is said to be output controllable if for any $y(0), y' \in \mathbb{C}^{P \times 1}$, there exists $t_1 \in \mathbb{R}^+, t_1 < +\infty$ and $u' \in \mathbb{N}^{N \times 1}$, such that under such input and initial conditions, the output vector of the linear system can be transferred from $y(0)$ to $y' = y(t_1)$.

The necessary and sufficient condition [35, 29] for a linear system to be output controllable is given in Definition 4.6. A natural definition for the output controllability index that is similar to the controllability and observability index is given in Definition 4.7. Finally, we note in Theorem 4.9 that the output controllability index minus one is a tight upper bound for the number of time delays required for the linear model in the general sense.

**Definition 4.6 (Output controllability test).** The system in (4.10) and (4.11) is output controllable if and only if

\begin{align}
(4.12) \quad \text{OC}(A,B,C,D;M) = [CB \quad CAB \quad \ldots \quad CA^{M-1}B \quad D] \in \mathbb{C}^{P \times (M+1)N}
\end{align}

is full rank. Note that when $D = 0$, we omit $D$ in the notation.

**Definition 4.7 (Output controllability index).** If the system in (4.10) and (4.11) is output controllable, then the output controllability index is defined as the least integer $\mu$ such that

\begin{align}
(4.13) \quad \text{OC}(A,B,C,D;\mu) \in \mathbb{C}^{P \times (\mu+1)N}
\end{align}

is full rank.

**Lemma 4.8.** For any matrix $A$ that is a horizontal stack of diagonal matrices, the row elimination matrix $E$ that removes any row that is a zero vector leads to a full rank matrix with the rank of original matrix. Moreover, $E^\top EA = A$.

**Proof.** See Appendix A.4.

**Theorem 4.9.** Following definitions in (4.8) and (4.9), consider the following linear dynamical system with output controllability index $\mu$:

\begin{align}
\dot{Z} &= AZ + Bu \\
y &= CZ
\end{align}
with

\[
A = \begin{bmatrix}
A^{-1} & & \\
& \ddots & \\
& & A^{-1}
\end{bmatrix} \in \mathbb{C}^{M J \times M J},
\]

\[
B = \begin{bmatrix}
e & & \\
& \ddots & \\
& & e
\end{bmatrix} \in \mathbb{C}^{M J \times J},
\]

\[
C' = [\text{diag}(a^{(1)}) \ldots \text{diag}(a^{(J)})] \in \mathbb{C}^{M \times JM},
\]

\[
C = EC' \in \mathbb{C}^{P \times JM},
\]

where \( P \) is the number of non-zero row vectors in \( C' \), and \( \text{rank}(C) = \text{rank}(C') = P \) as indicated by Lemma 4.8. Then, \( \mu - 1 \) is a tight upper bound on the minimal number of time delays that ensures the existence of solution of (4.4), and thus a perfect reconstruction of the dynamics.

Proof. Consider

\[
\mathcal{O}(\mathcal{O}(A, B, C; \mu)) = C [B \ AB \ldots A^{\mu-1}B]
\]

\[
= C [I \ A \ldots A^{\mu-1}] \begin{bmatrix}
B \\
& \ddots \\
& & B
\end{bmatrix}
\]

\[
= EC' \begin{bmatrix}
I \\
& \ddots \\
& & I
\end{bmatrix} \begin{bmatrix}
A^{-(\mu-1)} \\
& \ddots \\
& & A^{-(\mu-1)}
\end{bmatrix} \begin{bmatrix}
e \\
& \ddots \\
& & e
\end{bmatrix}
\]

\[
= E \begin{bmatrix}
\text{diag}(a^{(1)})e \\
& \ddots \\
& & \text{diag}(a^{(J)})e
\end{bmatrix} \begin{bmatrix}
\text{diag}(a^{(1)})A^{-(\mu-1)}e \\
& \ddots \\
& & \text{diag}(a^{(J)})A^{-(\mu-1)}e
\end{bmatrix}.
\]

Following Definition 4.5, for any integer \( i \geq \mu \), \( \mathcal{O}(\mathcal{O}(A, B, C; i)) \) is full rank. Thus, \( \forall v \in \mathbb{C}^{P \times 1} \), \( v \) lies in the column space of \( \mathcal{O}(\mathcal{O}(A, B, C; i)) \). Therefore, \( Fv \) should lie in the column space of \( \mathcal{O}(\mathcal{O}(A, B, C; i)) \). Noticing Lemma 4.8 and Remark 4.3, we have

\[
Fv \in \text{Col}(\mathcal{F}\mathcal{O}(A, B, C; i)) = \mathcal{W}_{i-1}.
\]

Now, consider \( \forall j = 1, \ldots, J, v^{(j)} = E \text{diag}(a^{(j)})b_{\mathcal{F}M}^{j} \in \mathbb{C}^{P \times 1} \), from the above, we have

\[
Fv^{(j)} = FE \text{diag}(a^{(j)})b_{\mathcal{F}M}^{j} = \text{diag}(a^{(j)})b_{\mathcal{F}M}^{j} = c^{(j)} \in \mathcal{W}_{i-1}.
\]

Since the minimal \( i \) for \( \mathcal{O}(\mathcal{O}(A, B, C; i)) \) to be full rank is \( \mu \), the output observability index is \( \mu \). Correspondingly, when the number of time delays \( L = \mu - 1 \), a solution exists for (4.4), which makes \( \mu - 1 \) an upper bound for the minimal time delay in Lemma 4.1. Finally, to show that the bounds are tight, consider that when \( J = 1 \), Theorem 4.9 reverts to Theorem 3.4 where \( \mu = P \), and thus \( \mu - 1 = P - 1 \) is essentially the minimal number of time delays required.
5. Numerical conditioning considerations. In practical terms, one can pursue two general formulations to numerically compute the delay transition matrix $K$ in (2.3):

1. **Formulation in time domain**: If all available delay vectors and corresponding future states are stacked, the direct solution of (2.3) is a least square problem in the time domain with the requirement of at least $P$ samples.

2. **Formulation in spectral domain**: In this approach, the Fourier signals from a full period of data is extracted and (3.17) is numerically solved.

5.1. Ill-conditioning due to excessive sampling rate. Consider signals that consist of a finite number of harmonics with the index set of Fourier coefficients as $I^p_M$. Since the first half of the indices $i_0, \ldots, i_{P/2-1}$ is determined by the inherent period of each harmonic, these indices are independent of the number of samples per period $M$, as long as $M$ satisfies the Nyquist condition. It is thus tempting to choose a relatively large sampling rate. However, this may not be favorable from a numerical standpoint. When $L = P - 1$ and the sampling rate is excessive compared to the potentially lower frequency dynamics of the system, each column could become nearly linearly dependent. We will now explore the circumstances under which the corresponding linear system in either the spectral or time domain can become ill-conditioned. It has to also be recognized that the denominator in (3.22) consists of the difference between different nodes on the unit circle, and can therefore impact numerical accuracy.

The condition number of the Vandermonde matrix with complex nodes (3.17) is also pertinent to the present discussion. It is well known that the condition number of a Vandermonde matrix grows exponentially with the order of matrix $n$ when the nodes are real positive or symmetrically distributed with respect to the origin [21]. When the nodes are complex, the numerical conditioning of a Vandermonde matrix can be as perfect as that of a DFT matrix, or as poor as that of the quasi-cyclic sequence [25]. Specifically, it has been shown that a large square Vandermonde matrix is ill-conditioned unless its nodes are nearly uniformly spaced on or about the unit circle [50]. Interestingly, for a rectangular Vandermonde matrix with $n$ nodes and order $N$, i.e., $V_N(z_1, \ldots, z_n)$, Kunis and Nagel [36] provided a lower bound on the 2-norm condition number of the Vandermonde matrix that contains “nearly-colliding” nodes:

$$\kappa_2(V_N(z_1, \ldots, z_n)) \geq \frac{\sqrt{6}}{\pi\tau} \approx \frac{0.77}{\tau}, \text{ for all } \tau \leq 1, \text{ i.e., “nearly colliding”},$$

where $\tau \leq N \min_{j \neq l} |t_j - t_l|_T, |t_j - t_l|_T \leq \min_{r \in \mathbb{Z}} |t_j - t_l + r|$. Applying the above result to (3.17), when $M$ is large enough so that $\tau \leq 1$ is satisfied1, the lower bound of the 2-norm condition number will increase proportionally with the number of samples per period $M$. Thus, the tightly clustered nodes due to excessive sampling will lead to the ill-conditioning of the linear system in (3.17).

5.2. Subsampling. (5.1) shows that the tightly clustering of nodes due to excessive sampling can lead to ill-conditioning. A straightforward fix would thus be to filter out unimportant harmonics, and re-sample the signal at a smaller sampling rate that can still capture the highest frequency retained in the filtering process. In this way, the nodes can be more favorably

---

1since $\tau = O(1/M)$
redistributed on the unit circle. Recall that, if the complex nodes of the Vandermonde matrix are uniformly distributed on a unit circle, then one arrives at a perfect conditioning of the Vandermonde matrix with condition number of one similar to the DFT matrix [50]. Without any loss of generality, we assume the number of samples per period $M$ is even. The wave numbers of sparse Fourier coefficients are denoted by $I^{p}_{M}$. The sorted wave numbers are symmetrical with respect to $M/2$ and recall that the values of the first half of $I^{p}_{M}$, i.e., $i_{0}, \ldots, i_{M-1}$ is independent of $M$, as long as the Nyquist condition is satisfied [37]. Then, a continuous signal $x(t)$ is subsampled uniformly. Due to symmetry, the smallest number of samples per period $M^*$ that preserves the signal is $2(i_{M-1} + 1)$.

6. Numerical experiments. To illustrate the utility and effectiveness of the theoretical results developed in this work, a series of examples are considered.

6.1. A moderately sampled 5-mode sine signal. First, an explicit time series consisting of five frequencies with a long period $T = 100$ is considered:

$$
(6.1) \quad x(t) = 0.3 \cos\left(\frac{2\pi t}{100}\right) + 0.5 \sin\left(\frac{4\pi t}{100}\right) + 0.9 \cos\left(\frac{8\pi t}{100}\right) + 1.6 \sin\left(\frac{16\pi t}{100}\right) + 1.2 \cos\left(\frac{24\pi t}{100}\right).
$$

Such a signal may be realized, for instance, by observing the first component of a 10-dimensional linear dynamical system.

The sampling rate is set at 1 per unit time, which is arbitrary and considered for convenience, and the signal is sampled for two periods from $n = 0$ to $n = 199$. Thus we have a discretely sampled time series of length 200 as $\{x_{n}\}_{n=0}^{199}$, with $x_n = x(t)|_{t=n}$. Only the first 20% of the original signal is used, which is 40% of a full period with around 20 to 30 data points sampled. The variation in the number of data points is due to the fact that we fix the use of first 20% of trajectory, and then reconstruct the signal with a different number of time delays. We solve the least squares problem in the time domain with the iterative least squares solver scipy.linalg.lstsq [31] with lapack driver as gelsd, and cutoff for small singular values as $10^{-15}$.

Figure 2. Left: A posteriori prediction vs ground truth, time delayed linear model with number of delays $L = 9$. Right: A posteriori MSE normalized by standard deviation of $x(t)$ vs number of time delays.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{Left: A posteriori prediction vs ground truth, time delayed linear model with number of delays $L = 9$. Right: A posteriori MSE normalized by standard deviation of $x(t)$ vs number of time delays.}
\end{figure}
The analysis in Theorem 3.4 implies that one can avoid using the full period of data for exact prediction. Numerical results are presented in Figure 2 with number of time delays $L = 9$. These results show that time delayed DMD, unlike non-linear models such as neural networks, avoid the requirement of a full period of data when the dynamics is expressible by a set of sparse harmonics.

From Theorem 3.4, the 5-mode signal has $P = 10$ non-zero Fourier coefficients in the Fourier spectrum, and thus the least number of delays is $L = P - 1 = 9$, which agrees well with Figure 2 which shows the a posteriori mean square error normalized by the standard deviation of the data, between prediction and ground truth. Figure 2 clearly shows that a sharp decrease of a posteriori error when the number of delays $L = 9$.

6.2. Excessively sampled 5-mode sine signal. As explained earlier, linear time delayed models can avoid the use of a full period of data if there is enough information to determine the solution within the first $P$ states. Thus, if one increases the sampling rate, less data will be required to recover an accurate solution. However, one still needs to numerically compute the solution of a linear system, while the condition number grows with increasing sampling rates. As displayed in Figure 4, the condition number increases in both time and spectral domain formulations, with increasing sampling rate.

Using scipy.linalg.lstsq [31] and a time domain formulation, we found that there is no visual difference between the truth and a posteriori prediction when the condition number is below $10^{13}$, i.e., $M \leq 300$ in the spectral domain, or $M \leq 200$ in the time domain. However, as the condition number grows beyond $10^{13}$ (i.e. machine precision noise of even $10^{-16}$ can contaminate digits around 0.001), a posteriori prediction error can accumulate when $M = 400$ (Figure 3).

6.2.1. Effect of sampling rate, formulation domain, and numerical solver on model accuracy. To compare the impact of different solution techniques, we choose several off-the-shelf numerical methods to compute $K$ in either the time domain or spectral domain. These methods include:

(i) mldivide from MATLAB [44], i.e., backslash operator which effectively uses QR/LU
(ii) `scipy.linalg.lstsq` [31], which by default calls `gelsd` from LAPACK [2] to solve the minimum 2-norm least squares solution with SVD, and an algorithm based on divide and conquer;

(iii) Björck & Pereyra (BP) algorithm [8] which is designed to solve the Vandermonde system exactly in an efficient way exploiting the inherent structure. For a $n \times n$ matrix, instead of the standard Gaussian elimination with $O(n^3)$ arithmetic operations and $O(n^2)$ elements for storage, the BP algorithm only requires $n(n+1)(2O_M + 3O_A)/2^2$ for arithmetic operations and no further storage than storing the roots and right hand side of the system.

As shown in Figure 4, the condition number increases exponentially with increasing number of samples per period $M$, leading to a significant deterioration of accuracy. Comparing the time and spectral domain formulations, Figure 4 shows that the solution for the spectral case is more accurate than the time domain solution when the sampling rate is low. This is not unexpected as one would need to perform FFT on a full period of data to find the appropriate Fourier coefficients in the spectral case. When $M > 600$, however, the spectral domain solutions obtained by BP and `mldivide` algorithms blow up, while the time domain solution is more robust in that the error is bounded. Note that the singular value decomposition - in `lstsq` and in `mldivide` that removes the components of the solution in the subspace spanned by less significant right singular vectors - is extremely sensitive to noise. Further, from (3.33), the difference between the formulations in the spectral and time domains can be attributed to $V_P(\omega^{k_0}, \ldots, \omega^{k_0-1})$ and $\text{diag}(a_{i_0}, \ldots, a_{i_{p-1}})$, which could be ill-conditioned. Thus, regularization in the time domain formulation is more effective. Figure 4 also shows that, when the system becomes highly ill-conditioned, i.e., $M > 600$, `lstsq` with thresholding $\epsilon = 10^{-15}$ results in a more stable solution than `mldivide`.

It should be mentioned that the condition number computed in Figure 4 around the inverse of machine precision, i.e., $O(10^{16})$, should be viewed in a qualitative rather than quantitative sense [23].

![Figure 4](image)

**Figure 4.** Left: A posteriori MSE normalized by the standard deviation of $x(t)$ with increasing sampling rate and different numerical solvers. Right: Numerical condition number with increasing sampling rate.

$^2 O_A$ and $O_M$ denote addition/subtraction and multiplication/division.
6.2.2. Effect of increasing the number of time delays. By adding more time delays than the theoretical minimum, the dimension of the solution space grows, along with the features for least squares fitting. Accordingly, the null space becomes more dominant, and thus one should expect non-unique solutions with lower residuals. Note that, for simplicity, the following numerical analysis assumes the scalar case, i.e., \( J = 1 \).

For the complex Vandermonde system in (3.17), following Bazán’s work [6], we discovered very distinct features of the asymptotic behavior of the solution, and the corresponding system in (3.17) when the number of time delays \( L \to \infty \).

(i) The norm of the minimum 2-norm solution of (3.17) \( \| \hat{K}_L \|_2 = 0 \), as shown in Proposition 6.1.

(ii) An upper bound for the convergence rate of \( \| \hat{K}_L \|_2^2 \) is derived in Lemma 6.2.

(iii) An upper bound on the 2-norm condition number of (3.17) is shown in Proposition 6.3 to scale with \( 1 + O(1/\sqrt{L}) \).

**Proposition 6.1.** \( \lim_{L \to \infty} \| \hat{K}_L \|_2 = 0 \), where \( \hat{K}_L \) is the minimum 2-norm solution of (3.17).

**Proof.** To begin with, consider the following under-determined linear system for \( f \in \mathbb{R}^n \), given \( N \geq n \)

(6.2) \[ V_N(z_1, \ldots, z_n)f = \text{diag}(z_1, \ldots, z_n)e, \]

where \( e = [1 \ 1 \ldots \ 1]^T \). Denote \( f_N \) to be the minimum 2-norm solution. Suppose for all nodes, \( i = 1, \ldots, n, \ |z_i| \leq 1 \). Bazán [6] showed that

(6.3) \[ \lim_{N \to +\infty} \| f_N \|_2 = 0. \]

Consider multiplying (3.17) on both sides from the left with \( \text{diag}(\omega^{L_{i_0}}, \ldots, \omega^{L_{i_{P-1}}} ) \). Notice that the diagonal matrix is non-singular for any \( L \in \mathbb{N} \), and the inverse of permutation matrix is its transpose. Then we have

(6.4) \[ \begin{bmatrix} \omega^{L_{i_0}} & \omega^{(L-1)i_0} & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ \omega^{L_{i_{P-1}}} & \omega^{(L-1)i_{P-1}} & \cdots & 1 \end{bmatrix} K = \begin{bmatrix} \omega^{(L+1)i_0} \\ \vdots \\ \omega^{(L+1)i_{P-1}} \end{bmatrix}, \]

(6.5) \[ \begin{bmatrix} 1 & \omega^{i_0} & \cdots & \omega^{L_{i_0}} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{i_{P-1}} & \cdots & \omega^{L_{i_{P-1}}} \end{bmatrix} P^T K = \begin{bmatrix} \omega^{i_0} \\ \vdots \\ \omega^{i_{P-1}} \end{bmatrix} L^+ e, \]

(6.6) \[ V_{L+1}(\omega^{i_0}, \omega^{i_1}, \ldots, \omega^{i_{P-1}})f = (\text{diag}(\omega^{i_0}, \omega^{i_1}, \ldots, \omega^{i_{P-1}}))^{L^++1} e, \]

where \( f \triangleq P^T K, P \in \mathbb{R}^{(L+1) \times (L+1)} \) is the column permutation matrix that reverses the column order in \( A_{I_{L+1}^p}, L \). Note that a solution exists when \( L + 1 = P \) and it is not unique when \( L + 1 > P \). Denote \( f_L \) as the corresponding minimal 2-norm solution of (6.6). From (6.3), consider (6.6) and take \( L \to +\infty, \| f_L \|_2 \to 0 \). The row permutation matrix does not change the 2-norm of a vector, and hence there is a one-to-one correspondence between the solution in (6.6) and (3.17), such that the corresponding minimal 2-norm solution for (3.17) is \( \hat{K}_L \triangleq P f_L \) thus \( \| \hat{K}_L \|_2 \to 0 \).
Lemma 6.2. \( \forall L \geq P - 1 \), denote \( \hat{K}_L \) as the minimum 2-norm solution of \((3.17)\). The following tight upper bound can be derived

\[
(6.7) \quad \|\hat{K}_L\|_2^2 \leq \frac{\|\hat{K}_{P-1}\|_2^2}{1 + \left[\frac{L-P+1}{L-P-1}\right]^2}.
\]

Proof. See Appendix A.5.

Proposition 6.3. Let \( P \) be the number of non-zero Fourier coefficients. \( \forall L \geq P - 1 \), denote \( \hat{K}_{P-1} \) as the unique solution of \((3.17)\). With the minimal number of time delays, the upper bound on the 2-norm condition number of the system is given by

\[
(6.8) \quad \kappa_2(A_{\hat{T}_{P-1}^{\alpha, \beta}}, L) = \kappa_2(V_{L+1}(\omega^{-i_0}, \ldots, \omega^{-i_{P-1}})) \leq 1 + \frac{d}{2} \left[1 + \sqrt{1 + \frac{4}{d}}\right],
\]

where

\[
(6.9) \quad d = P \left(1 + \frac{\|\hat{K}_{P-1}\|_2^2}{(P-1)(1 + \left[\frac{L-P+1}{L-P-1}\right])\delta^2} - 1\right),
\]

\[
(6.10) \quad \delta = \min_{0 \leq j < k \leq P-1} |\omega^{-i_j} - \omega^{-i_k}|.
\]

Further, if \( L \to \infty \), then \( \kappa_2(A_{\hat{T}_{P}^{\alpha, \beta}}, L) \to 1 \), i.e., perfect conditioning is achieved.

Proof. Consider the fact that the Vandermonde matrix \( V_N(z_1, \ldots, z_n) \) with \( n \) distinct nodes \( \{z_i\}_{i=1}^n \), \( z_i \in \mathbb{C} \) of order \( N \), \( N \geq n \), i.e., \( V_N \) is full rank. The Frobenius-norm condition number is defined as \( \kappa_F(V_N) \triangleq \|V_N\|_F\|V_N^\dagger\|_F \), where \( \dagger \) represents Moore-Penrose pseudo-inverse. Bazán [6] showed that if \( \forall i = 1, \ldots, n \), with distinct \( |z_i| \leq 1 \), \( N \geq n \), then

\[
(6.11) \quad \kappa_F(V_N) \leq n \left[1 + \frac{(n-1) + \|f_N\|^2_2 + \prod_{i=1}^n |z_i|^2 - \sum_{i=1}^n |z_i|^2}{(n-1)\delta^2}\right]^{n-1} \phi_N(\alpha, \beta),
\]

where \( \delta = \min_{1 \leq i < j \leq n} |z_i - z_j| \), \( \phi_N(\alpha, \beta) \triangleq \sqrt{\frac{1 + \alpha^2 + \cdots + \alpha^{2(N-1)}}{1 + \beta^2 + \cdots + \beta^{2(N-1)}}}, \alpha \triangleq \max_{1 \leq j \leq n} |z_j|, \beta \triangleq \min_{1 \leq j \leq n} |z_j| \).

The key to understand the behavior of the upper bound of \( \kappa_2(V_N) \), is to estimate the convergence rate of \( \|f_N\|_2 \) which is considered difficult for a general distribution of nodes [6]. For the particular case of \((3.17)\), we can show a tight upper bound in Lemma 6.2. Thus, \( \forall 1 \leq i \leq n, |z_i| = 1 \), \((6.11)\) becomes,

\[
(6.12) \quad \kappa_F(V_N) \leq n \left(1 + \frac{\|f_N\|^2_2}{(n-1)\delta^2}\right)^{n-1} \cdot \frac{1}{n-1}.
\]

Now we note a general inequality between the condition number in the 2-norm and in the Frobenius norm [6] by considering,

\[
(6.13) \quad n - 2 < n - 2 + \kappa_2(V_N) + \kappa_2^{-1}(V_N) \leq \kappa_F(V_N),
\]

\[
(6.14) \quad \kappa_2(V_N) \leq \frac{1}{2} \left[\kappa_F(V_N) - n + 2 + \sqrt{(\kappa_F(V_N) - n + 2)^2 - 4}\right].
\]
The right hand side in (6.14) is monotonically increasing with respect to $\kappa F(V_N)$. Therefore using the upper bound from (6.12) in (6.14), and some algebra we have the following upper bound,

$$\forall N > n,$$

\begin{equation}
\kappa_2(V_N) \leq 1 + \frac{d}{2} \left[ \frac{1}{2} + \sqrt{1 + \frac{4}{d}} \right],
\end{equation}

where

\begin{equation}
d \equiv n \left[ \left( \frac{1}{2} \left( \frac{\|f_N\|^2_2}{(n - 1)^2} \right)^{\frac{n-1}{2}} - 1 \right) \right].
\end{equation}

Finally, note that $d$ monotonically increases with $\|f_N\|^2_2$, and thus with $n = P, N = L + 1, N = \omega^{-l}, l = 0, \ldots, P - 1$ and Lemma 6.2, the desired upper bound is achieved. As $L \to \infty$, $\hat{K}_L \to 0$ and $d \to 0$, and thus it is trivial to show that $\kappa_2(A_{\hat{K}_L}) \to 1$.

**Remark 6.4.** Note that the bound in Proposition 6.3 does not demand a potentially restrictive condition on the number of time delays, i.e., $L + 1 > 2(P - 1)/\delta$ that is required in Bazán’s work, which utilizes the Gershgorin circle theorem for the upper bound of the 2-norm condition number [6]. More recently, this constraint has been defined in the context of the nodes being “well-separated” [36]. Applying such a result to our case, we have

\begin{equation}
\kappa_2(A_{\hat{K}_L}) \leq \sqrt{1 + \frac{2}{\delta^2}} - 1
\end{equation}

since we have an estimation for the convergence rate of the minimal 2-norm solution. However, although our upper bound in Proposition 6.3 holds\(^3\) for all $L \geq P - 1$, it is too conservative compared to Bazán’s upper bound when $L \to \infty$. To see this, denote $k_m = \min_{i,j \in \mathcal{T}_L^P, i \neq j} \{\|k\| = (i - j) \mod M\}$, i.e., the minimal absolute difference between any pair of distinct indices in $\mathcal{T}_L^P$, in the sense of modulo $M$. Assuming that the number of samples per period is large enough so that $M \gg 2\pi k_m$, we have $\delta = \sqrt{2[1 - \cos(2\pi k_m/M)]} \approx 2\pi k_m/M = O(1/M)$. If we assume that the system with time delay $L$ is far from being perfectly conditioned, we have $\kappa_F(V_{L+1}) \gg P + 2$, which leads to the following approximation for our upper bound,

\begin{equation}
\kappa_2(V_N) \leq \frac{1}{2} \left( \kappa_F(V_{L+1}) - P + 2 + \sqrt{\kappa_F(V_{L+1}) - P + 2^2 - 4} \right) \approx \kappa_F(V_{L+1}) - P + 2 \leq d + 2.
\end{equation}

Hence, for an excessively sampled case, if $L$ is small enough such that $\kappa_F(V_{L+1}) \gg \kappa_2(V_{L+1}) \gg P + 2$ holds but large enough such that

\begin{equation}
\frac{\|\hat{K}_{P-1}\|^2_2}{(P - 1)\left(1 + \frac{L - P + 1}{M^2}\right)\delta^2} \ll 1,
\end{equation}

\(^3\)and is more general than Bazán’s upper bound (6.17)
then the approximated upper bound becomes

\[
2 + d = 2 + P \left[ \left( 1 + \frac{\|\hat{K}_{P-1}\|^2}{(P-1)(1+\left[ \frac{L-P+1}{M} \right])\delta^2} \right)^{\frac{P-1}{2}} - 1 \right],
\]

\[
\approx 2 + \frac{P\|\hat{K}_{P-1}\|^2}{2\delta^2(1+\left[ \frac{L-P+1}{M} \right])} \approx 2 + \frac{P\|\hat{K}_{P-1}\|^2}{8\pi^2k_{\min}^2/M^2(1+\left[ \frac{L-P+1}{M} \right])} = 2 + O\left( \frac{M^3}{L} \right).
\]

Meanwhile, when \( L \) is very large, and thus \( \delta(L+1) > 2(P-1) \) is satisfied, Bazán’s bound in (6.17) scales with \( 1 + O\left( \sqrt{M}/\sqrt{L} \right) \) for \( L/M \gg 1 \). Thus, to retain the same upper bound of condition number, one only needs to increase the number of time delays at the same rate as the sampling.

Figure 5 shows that the residuals from the least squares problem in both the time and spectral domains decrease exponentially with the addition of time delays. Further, the a posteriori MSE shows significant improvement with the addition of time delays.

![Figure 5](image_url)

Figure 5. Left: A posteriori MSE normalized by standard deviation of \( x(t) \) with increasing time delays. Right: Sum of squared residuals with increasing time delays.

Figure 6 shows the trend of the 2-norm condition number in both the time and spectral domains. The condition number decays exponentially in the spectral case, but increases in the time domain case. This appears to be contradictory since the condition number is typically reflective of the quality of the solution. However, since SVD regularization is implicit in `scipy.linalg.lstsq` with `gelsd` option, computing the 2-norm condition number in the same way as in the numerical solver \(^4\) is a more relevant measure of the quality of the solution of the SVD truncated system. Thus, the reasons for improved predictive accuracy are due to a) the increasing dimension of the solution space for a potentially under-determined system with more time delays, and b) the well conditioned system after SVD truncation as shown in Figure 6. The large condition number in the time domain with increasing number of delays is a result of the ill-conditioning of \( V_P(\omega^{k_0}, \ldots, \omega^{k_{Q-1}}) \) and \( \text{diag}(a_{i_0}, \ldots, a_{i_{P-1}}) \) in (3.34).

\(^4\)i.e., SVD with the same thresholding \( (\epsilon = 10^{-15}) \) such that any singular value below \( \epsilon \cdot \sigma_{\max} \) is removed
6.2.3. Effect of subsampling. As indicated in Remark 6.4, reducing the number of samples per period $M$ is shown to decrease the upper bound on the condition number. For a given signal, however, there is a restriction on the minimum possible $M$ compared to the number of time delays $L$. In the above case for the 5-mode sine signal, $i_P - 1 = 12$, and thus the minimal sampling per period that one can use to perfectly preserve the original signal in the subsampling is $M = 26$. The condition number with $M$ ranging from 26 to 98 is shown in Figure 7. This shows the effectiveness of subsampling in reducing the condition number significantly. Correspondingly, the a posteriori normalized MSE is also reduced as shown in Figure 7.

The previous two subsections demonstrated the role of numerical conditioning on model performance. We note that explicit stabilization techniques [39, 18] require further investigation.

![Figure 6](image1.png)

**Figure 6.** Condition number as a function of the number of time delays.

![Figure 7](image2.png)

**Figure 7.** Left: Condition number as a function of sampling rate. Right: A posteriori normalized MSE with sampling rate.
6.3. Van der Pol oscillator. Now we consider a non-linear system, the Van der Pol oscillator (VdP) with forward Euler time discretization:

\[
\begin{bmatrix}
    x_1^{n+1} \\
    x_2^{n+1}
\end{bmatrix} = \begin{bmatrix}
    x_1^n \\
    x_2^n
\end{bmatrix} + \Delta t \begin{bmatrix}
    x_1^n - x_2^n - x_1^n \\
    \mu(1 - x_1^n x_2^n) x_2^n - x_1^n
\end{bmatrix},
\]

where \( \mu = 2, x_1^0 = 1, x_2^0 = 0, \Delta t = 0.01 \). After 530 time steps, the system approximately falls on the attractor with an approximate period of 776 steps. Total data is collected after the system falls on the attractor for 4 periods.

6.3.1. Filtered trigonometric representation. The percentage of the \( \ell \)-th cumulative energy in the total energy in the spectrum \( \{a_0, \ldots, a_{M-1}\} \) is defined as,

\[
\xi_\ell = \frac{\sum_{i=0}^{\ell-1} |a_i|^2}{\sum_{i=0}^{M-1} |a_i|^2},
\]

where \( \ell \) is the number of Fourier coefficients included.

For each component of the VdP system, the variation of \( \xi_\ell \) is shown in Figure 8. We take a straightforward strategy of filtering out those frequencies with an amplitude that is 100 times smaller than the maximum over all frequencies. This truncation results in 10 terms for first component and 18 terms for the second component ((6.24) and (6.25)), and as seen in Figure 9, the original and filtered signals are virtually identical. The filtered data will thus be used in the following sections for the purposes of asymptotic analysis. For convenience, the filtered trigonometric expressions for the first and second component are

\[
\begin{align*}
\tilde{x}_1(t) &= 0.8271 \cos(\theta t) + 1.8897 \sin(\theta t) + 0.1289 \cos(3\theta t) - 0.3246 \sin(3\theta t) \\
&\quad + 0.0271 \cos(5\theta t) + 0.1494 \sin(5\theta t) + 0.0007 \cos(7\theta t) + 0.0641 \sin(7\theta t) \\
&\quad - 0.0049 \cos(9\theta t) + 0.0283 \sin(9\theta t),
\end{align*}
\]

\[
\begin{align*}
\tilde{x}_2(t) &= 1.5236 \cos(\theta t) - 0.6759 \sin(\theta t) + 0.9505 \cos(3\theta t) - 0.3246 \sin(3\theta t) \\
&\quad + 0.5987 \cos(5\theta t) - 0.1218 \sin(5\theta t) + 0.3594 \cos(7\theta t) - 0.0140 \sin(7\theta t) \\
&\quad + 0.2042 \cos(9\theta t) + 0.0290 \sin(9\theta t) + 0.1096 \cos(11\theta t) + 0.0378 \sin(11\theta t) \\
&\quad + 0.0555 \cos(13\theta t) + 0.0325 \sin(13\theta t) + 0.0263 \cos(15\theta t) + 0.0236 \sin(15\theta t) \\
&\quad + 0.0114 \cos(17\theta t) + 0.0154 \sin(17\theta t),
\end{align*}
\]

where \( \theta = 2\pi/T \). The number of non-zero Fourier coefficients and the lowest sampling rate, together with theoretical minimal time delay for each component is listed in the Table 1. For the vector case, i.e., considering both components together, from Lemma 4.1, the minimal number of time delays is \( L = 8 \), which is smaller than \( L = 9 \) for \( \tilde{x}_1(t) \) or \( L = 17 \) for \( \tilde{x}_2(t) \).

| \( P \) | \( L \) | \( \Delta \nu_{\ell-1} \) | \( M_{\min} \) |
|---|---|---|---|
| \( \tilde{x}_1(t) \) | 10 | 9 | 20 |
| \( \tilde{x}_2(t) \) | 18 | 17 | 38 |
| \( \tilde{x}_{1,2}(t) \) | 8 | 8 | 38 |
6.3.2. Prediction of moderately subsampled VdP system using minimal time delays: scalar case. From Table 1, it is clear that the smallest number of samples per period is significantly smaller than the original number of samples per period, i.e., $M = 776$. The analysis in the previous section also showed that the choice of a smaller number of samples per period is helpful in reducing the condition number. Thus, we choose a moderately subsampled representation without any loss in reconstruction compared to the filtered representation. Individually treating the first and second components, we choose $M = 200, 100$ with theoretical minimum time delays $L = 9, 17$, respectively.

Numerical results displayed in Figure 10 show that, even using training data that covers less than 25% of the period for the first component, and 50% of the period for the first component, the linear model with minimal time delays is still able to accurately predict the dynamics over the entire time period of the limit cycle. Note that a similar predictive performance is expected for the original (unfiltered) VdP system.
6.3.3. Prediction of moderately subsampled VdP system using minimal time delays: vector case. As given in Table 1, Lemma 4.1 predicts that the consideration of both com-
ponents requires only 8 delays. The effectiveness of the criterion developed in Lemma 4.1 is confirmed to a resounding degree in Figure 11. The left figure shows the predictive performance of the time delayed linear model for the minimum number of delays and the right figure shows the behavior of the a posteriori normalized MSE versus the number of time delays. It should be recognized that in contrast to the scalar case, in which the minimal time delay can be directly inferred from the Fourier spectrum, the vector case requires iterative evaluations of the rank test in Lemma 4.1.

![Figure 11. Left: Prediction vs ground truth with $M = 80$ for filtered VdP. Right: A posteriori MSE normalized by standard deviation with as a function of the number of time delays for the vector case.](image)

7. Conclusions. In summary, this work addressed fundamental questions regarding the structure and conditioning of linear time delay models of non-linear dynamics on an attractor. The following are the main contributions of this work:

1. We proved that for non-linear scalar dynamical systems, the number of time delays required by linear models to perfectly recover limit cycles is determined by the sparsity in the Fourier spectrum.
2. In the vector case, we proved that the minimal number of time delays has a tight upper bound that is precisely the output controllability index of a related linear system.
3. We developed an equivalent representation of the linear time delayed model in the spectral domain and provided the exact solution of the delay transition matrix $K$ for the scalar case.
4. We derived an upper bound on the 2-norm condition number as a function of the sampling rate and the number of time delays. Thus, ill-conditioning can be mitigated by increasing the number of time delays and subsampling the original signal.
5. We explicitly showed that the dynamics over the full period can be perfectly recovered by training the linear time delayed model over just a partial period.
6. Numerical experiments were shown to confirm each of the above theoretical results.

A few observations are pertinent to the above conclusions:

Due to accuracy considerations on the numerical integrator, the sampling rate in the raw data may be excessively high. We believe that instabilities in prediction arise from choices that
lead to poor numerical conditioning. Thus, as an alternate to pursuing explicit stabilization techniques \([39, 18]\), appropriate subsampling and time delays can be employed. Indeed, when noise is present in the data, explicit stabilization may be warranted.

The effectiveness of linear time delayed models of non-linear dynamics is that - by leveraging Fourier interpolation - an arbitrarily close trajectory from a high dimensional linear system can be derived. This also intuitively explains the ability of the model - when the signal has a sparse spectrum - to perform “true” predictions without training on a full period of data. To pursue a more formal explanation in a general sense, one could resort to the concept of inductive bias \([65]\) or model preference.

 image: https://www.deepmind.com/assets/images/2019-Hanin-paper-1.png

Acknowledgments. This work was supported by DARPA under the grant titled Physics Inspired Learning and Learning the Order and Structure Of Physics, (Technical Monitor: Dr. Jim Gimlett).

Appendix A. Proofs of lemmas.

A.1. Proof of Lemma 3.2.

Proof. 

(A.1) \[
\begin{bmatrix}
1 & \alpha_0 & \ldots & \alpha_0^{N-1} \\
1 & \alpha_1 & \ldots & \alpha_1^{N-1} \\
\vdots & \vdots & \ddots & \vdots \\
1 & \alpha_{M-1} & \ldots & \alpha_{M-1}^{N-1}
\end{bmatrix}
\]

If \(M \geq N\), then

(A.2) \[
V_N(\alpha_0, \alpha_1, \ldots, \alpha_{M-1}) = 
\begin{bmatrix}
V_N(\alpha_0, \alpha_1, \ldots, \alpha_{N-1}) \\
V_N(\alpha_N, \ldots, \alpha_{M-1})
\end{bmatrix} \in \mathbb{C}^{M \times N}.
\]

Since \(\{\alpha_i\}_{i \in \mathbb{I}_M}\) are distinct, \(V_N(\alpha_0, \alpha_1, \ldots, \alpha_{N-1})\) is full rank with rank \(N\). Since \(M \geq N\), the row space of \(V_N(\alpha_0, \alpha_1, \ldots, \alpha_{M-1})\) and is fully spanned by the first \(N\) rows, and is thus full rank. Likewise, if \(M < N\),

(A.3) \[
V_N(\alpha_0, \alpha_1, \ldots, \alpha_{M-1}) = 
\begin{bmatrix}
V_M(\alpha_0, \alpha_1, \ldots, \alpha_{M-1})
\end{bmatrix} \in \mathbb{C}^{M \times N}.
\]

Similarly, the first \(M\) columns are full rank and \(V_N(\alpha_0, \alpha_1, \ldots, \alpha_{M-1})\) is also full rank. Thus in either case, \(V_N(\alpha_0, \alpha_1, \ldots, \alpha_{M-1})\) is full rank with rank as \(\min(M, N)\).

To show the the second property, one can simply replace \(\{\alpha_i\}_{i \in \mathbb{I}_M}\) with \(\{\alpha_i\}_{i \in \mathbb{J}}\) in the above arguments. Since \(|\mathbb{J}| = Q\),

(A.4) \[
\text{rank}(V_N(\{\alpha_i\}_{i \in \mathbb{J}})) = \min(Q, N).
\]

A.2. Proof of Lemma 3.3.

Proof. First, let’s prove from left to right. If \(\exists x \in \mathbb{C}^{n \times 1}\), we have \(Ax = b\). Note that \(\bar{A}x = \bar{A}\bar{x} = Ax = b\), then consider \(x' = \frac{x + \bar{x}}{2} \in \mathbb{R}^{n \times 1}\). \(Ax' = (Ax + \bar{A}x)/2 = (b + b)/2 = b\). Second, it is easy to show from right to left. Third, when uniqueness is added, note that \(Ax = b \iff \bar{A}x = b\), it is easy to show both directions since it is impossible to have complex solution being unique and not real.
A.3. Proof of Lemma 4.1.

Proof. Given the definitions in (4.1)–(4.3), note (3.5), we have

\[ \tilde{Y}_k = \begin{bmatrix} \Omega_{k,L} & \cdots & a^{(1)} \\ \vdots & \ddots & \vdots \\ \Omega_{k,L} & \cdots & a^{(J)} \end{bmatrix}. \]

Recall (3.8), note that

\[ \Upsilon_k = \Lambda^b_{M,M}, \]

where \( \Lambda = \begin{bmatrix} 1 & \omega & \cdots & \omega^{(M-1)} \end{bmatrix} \).

Moreover, note that

\[ \Omega^\top_{k,L} = \Lambda^k_{M,M,L}. \]

We rewrite (4.2) for a given \( k \) using (3.7) for the left hand side and (A.5) for the right hand side in (4.2),

\[ \begin{bmatrix} \Upsilon_k^\top \\
\vdots \\
\Upsilon_k^\top 
\end{bmatrix} \begin{bmatrix} \begin{bmatrix} \begin{bmatrix} \Upsilon_k \\
\vdots \\
\Upsilon_k 
\end{bmatrix} \begin{bmatrix} a^{(1)} \\
\vdots \\
a^{(J)} \end{bmatrix} = \tilde{K}^\top \begin{bmatrix} \Omega_{k,L} \\
\vdots \\
\Omega_{k,L} 
\end{bmatrix} \begin{bmatrix} a^{(1)} \\
\vdots \\
a^{(J)} \end{bmatrix}. \]

Using (A.6) and (A.7) for the above, we have

\[ \begin{bmatrix} \begin{bmatrix} a^{(1)} \\
\vdots \\
a^{(J)} \end{bmatrix} \begin{bmatrix} \begin{bmatrix} \Upsilon_k \\
\vdots \\
\Upsilon_k 
\end{bmatrix} - \begin{bmatrix} \Omega_{k,L} \\
\vdots \\
\Omega_{k,L} 
\end{bmatrix} \tilde{K} \end{bmatrix} = 0, \]

\[ \begin{bmatrix} \begin{bmatrix} a^{(1)} \\
\vdots \\
a^{(J)} \end{bmatrix} \begin{bmatrix} \Lambda^k \\
\vdots \\
\Lambda^k \end{bmatrix} \begin{bmatrix} \begin{bmatrix} b_{M,M} \\
\vdots \\
b_{M,M} \end{bmatrix} - \begin{bmatrix} A^L_{M,M,L} \\
\vdots \\
A^L_{M,M,L} \end{bmatrix} \tilde{K} \end{bmatrix} = 0. \]
Considering \( k = 0, 1, \ldots, M - 1 \), we stack

\[
\begin{bmatrix}
\mathbf{a}^{(1)} \\
\vdots \\
\mathbf{a}^{(J)} \\
\end{bmatrix}
\begin{bmatrix}
\Lambda^k \\
\ddots \\
\end{bmatrix}
\]
row by row as

\[
(A.11)
\begin{bmatrix}
\alpha_0^{(1)} & \cdots & \alpha_{M-1}^{(1)} & \cdots & \alpha_{M-1}^{(J)} \\
\alpha_0^{(1)} & \cdots & \omega^{M-1}a_{M-1}^{(1)} & \cdots & \omega^{M-1}a_{M-1}^{(J)} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
\omega^{(M-1)(M-1)}a_{M-1}^{(1)} & \cdots & \omega^{(M-1)(M-1)}a_{M-1}^{(J)} \\
\end{bmatrix}
= \mathbf{V}_M(\{\omega^j\}_{j=0}^{M-1}) \begin{bmatrix} \mathbf{I} & \cdots & \mathbf{I} \end{bmatrix} \text{diag}(\{\mathbf{a}^{(i)}\}_{i=1}^J)
= \mathbf{V}_M(\{\omega^j\}_{j=0}^{M-1}) \begin{bmatrix} \text{diag}(\mathbf{a}^{(1)}) & \cdots & \text{diag}(\mathbf{a}^{(J)}) \end{bmatrix}.
\]

Then plug the above equality into (A.10), and notice the non-singularity of \( \mathbf{V}_M(\{\omega^j\}_{j=0}^{M-1}) \), for \( k = 0, 1, \ldots, M - 1 \), (A.10) can be rewritten as

\[
(A.12)
\begin{bmatrix}
\text{diag}(\mathbf{a}^{(1)}) & \cdots & \text{diag}(\mathbf{a}^{(J)})
\end{bmatrix}
\begin{bmatrix}
\mathbf{b}_{\mathbf{I}}^{M} & \cdots & \mathbf{b}_{\mathbf{I}}^{M} \\
\mathbf{A}_{\mathbf{I},L}^{M} & \cdots & \mathbf{A}_{\mathbf{I},L}^{M} \\
\end{bmatrix}
= \mathbf{K} = 0.
\]

From the Rouché-Capelli theorem [45], the necessary and sufficient condition for the existence of a complex solution to (A.12) is,

\[
(A.13)
\text{rank} \left[ \begin{bmatrix}
\text{diag}(\mathbf{a}^{(1)}) & \mathbf{A}_{\mathbf{I},L}^{M} \\
\vdots & \ddots \\
\text{diag}(\mathbf{a}^{(J)}) & \mathbf{A}_{\mathbf{I},L}^{M} \\
\end{bmatrix} \right]
= \text{rank} \left[ \begin{bmatrix}
\text{diag}(\mathbf{a}^{(1)}) & \mathbf{A}_{\mathbf{I},L}^{M} \\
\vdots & \ddots \\
\mathbf{b}_{\mathbf{I}}^{M} & \mathbf{d}_{\mathbf{I}}^{M} \\
\end{bmatrix} \right].
\]

Note that since the above procedures are can be retained in (4.2), (4.2) and (A.12) share the same solution in \( \mathbb{C}^{(J+1) \times J} \). From Lemma 3.3, (A.13) is also the necessary and sufficient condition for (4.2) to have a real solution.

A4. Proof of Lemma 4.8.

Proof. For \( n, J \in \mathbb{N} \), consider \( J \) diagonal matrices in \( \mathbf{A} \), for \( j = 1, \ldots, J \), with the \( j \)-th diagonal matrices being \( \text{diag}(\mathbf{a}^{(j)}) \in \mathbb{C}^{n \times n} \). \( \mathbf{a}^{(j)} = \begin{bmatrix} a_1^{(j)} & a_2^{(j)} & \ldots & a_n^{(j)} \end{bmatrix}^\top \). Thus

\[
\mathbf{A} = \begin{bmatrix}
\text{diag}(\mathbf{a}^{(1)}) & \text{diag}(\mathbf{a}^{(2)}) & \cdots & \text{diag}(\mathbf{a}^{(J)})
\end{bmatrix} \in \mathbb{C}^{n \times nJ}.
\]

We define the following row index set that describes the row that is not a zero row vector in \( \mathbf{A} \).

\[
(A.14) \quad \Gamma = \{1 \leq \gamma_1 < \gamma_2 < \ldots < \gamma_P \leq n \},
\]

where we further order the index in \( \Gamma \) as

\[1 \leq \gamma_1 < \gamma_2 < \ldots < \gamma_P \leq n,\]
where \( P = |\Gamma| \). Now we construct the row elimination matrix \( E \in \mathbb{C}^{P \times n} \) from \( \Gamma \) with
\[
(A.15) \quad i \in \{1, \ldots, P\}, j \in \{1, \ldots, n\}, E_{ij} = \delta_{\gamma_i,j}.
\]

For \( EA \), since \( E \) only removes the zero row vector, the rank of the matrix \( EA \) is the same as \( A \). To show \( EA \) is full rank, simply consider the following procedure:

From the definition of \( \Gamma \), on each row with row index \( i = 1, \ldots, P \), there are non-zero entries. Start by choosing an entry, denoted as \( a_{ij}^0 \) that is non-zero. Then, one can simply perform column operations that switch the column with index \( j_i \) corresponding to the non-zero entry of \( i \)-th row, with the current \( i \)-th column. These operations can be iteratively performed, after which the following matrix is obtained:
\[
(A.16) \quad EAR = \begin{bmatrix}
a_{j_1}^{i_1} & * \\
a_{j_2}^{i_2} & * \\
& \ddots \\
a_{j_P}^{i_P} & *
\end{bmatrix},
\]

where \( \forall i = 1, \ldots, P, a_{ij_i}^0 \neq 0 \) and \( R \) is the elementary column operation matrix. Thus \( EAR \) is full rank, and \( EA \) is full rank.

Define\(^6\) \( F = E^\top \). Thus
\[
(A.17) \quad i, j \in \{1, \ldots, n\}, G_{ij} = FI_kE_{kj} = \delta_{\gamma_i,i'}\delta_{\gamma_{ij},j} = \sum_{k=1}^P \delta_{\gamma_i,i'}\delta_{\gamma_{ij},j} = \begin{cases} 
1, & i = j \in \Gamma, \\
0, & \text{otherwise}.
\end{cases}
\]

Therefore, \( G \) is simply a diagonal matrix that keeps the row with index in \( \Gamma \) unchanged, but makes the row zero when the index is not in \( \Gamma \). However, the row index that is not in \( \Gamma \) corresponds to a zero row vector, and thus \( GA = A \), i.e., \( E^\top EA = A \). \( \blacksquare \)

### A.5. Proof of Lemma 6.2.

\textbf{Proof.} For \( q \in \mathbb{N} \), denote \( L_q = qM + P - 1 \). Note that in (3.17), when \( L = P - 1 \), the minimal 2-norm solution \( \hat{K}_{P-1} \) is also unique. Specifically we denote \( \hat{K}_{P-1} = \begin{bmatrix} \hat{K}_0 & \cdots & \hat{K}_{P-1} \end{bmatrix} \).

Note that, for any \( L \geq P - 1 \), we can find \( q = \left\lfloor \frac{L-P+1}{M} \right\rfloor \), such that \( L \in T_q = [L_q, L_{q+1}) \). From the definition of the minimal 2-norm solution, we have \( \|\hat{K}_L\|_2 \leq \|\hat{K}_{L_q}\|_2 \).

Consider \( A_{T^P_{L_q}} \) and notice that for \( q = 0 \), i.e., \( L_0 = P - 1 \leq L < L_1 = M + P - 1 \), so \( \|\hat{K}_L\|_2 \leq \|\hat{K}_{L_0}\|_2 = \|\hat{K}_{P-1}\|_2 \); for \( q \geq 1 \), for any \( 1 \leq j \leq P \), the \( j \)-th column of \( A_{T^P_{L_q}} \) is duplicated with the \( (j + kM) \)-th column, \( k = 1, \ldots, q \). For \( q \geq 1 \), \( A_{T^P_{L_q}} \) in (3.17), consider the following easily validated special class of real solutions,
\[
(A.18) \quad K = \begin{bmatrix} K_0 & \cdots & K_{P-1} & 0 & \ldots & 0 & K_M & \cdots & K_{L_1} & 0 & \ldots & 0 & \cdots & K_{qM} & \cdots & K_{L_q} \end{bmatrix} \in \mathbb{R}^{1 \times (L_{q+1})}.
\]

\(^5\)while the choice of \( j_i \) is not unique

\(^6\)\( F_{ij} = \delta_{\gamma_i,j} \)
with the constraint that for any $1 \leq j \leq P$, $\sum_{l=0}^{q} K_{j-1+lM} = K_{j-1}$. To find the minimal 2-norm solution, note that we have

$$\min \| \mathbf{K} \|^2_2 = \sum_{j=1}^{P} \min \left( \sum_{l=0}^{q} K_{j-1+lM}^2 \right).$$

(A.19)

From Jensen’s inequality, $\forall j = 1, \ldots, P$,

$$\sum_{l=0}^{q} K_{j-1+lM}^2 \leq \left( \sum_{l=0}^{q} K_{j-1+lM} \right)^2,$$

(A.20)

$$\sum_{l=0}^{q} K_{j-1+lM}^2 \leq \frac{K_{j-1}^2}{q+1},$$

(A.21)

where the equality holds when $K_{j-1+lM} = K_{j-1}/(q+1)$ for $l = 0, \ldots, q$. Thus $\min \| \mathbf{K} \|^2_2 = \sum_{j=1}^{P} K_{j-1}^2/(q+1) = \| \mathbf{K}_{P-1} \|^2_2/(q+1)$. Since the above minimal norm is found within a special class of solutions in (3.17), the general minimal 2-norm is

$$\| \mathbf{K}_L \|^2_2 \leq \| \mathbf{K}_{L_0} \|^2_2 \leq \| \mathbf{K}_{P-1} \|^2_2/(q+1).$$

Combining both cases for $q = 0$ and $q \geq 1$, we have the desired result.

REFERENCES

[1] H. D. Abarbanel, R. Brown, J. J. Sidorowich, and L. S. Tsimring, The analysis of observed chaotic data in physical systems, Reviews of modern physics, 65 (1993), p. 1331.
[2] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, LAPACK Users’ Guide, Society for Industrial and Applied Mathematics, Philadelphia, PA, third ed., 1999.
[3] H. Arbabi and I. Mezić, Ergodic theory, dynamic mode decomposition, and computation of spectral properties of the koopman operator, SIAM Journal on Applied Dynamical Systems, 16 (2017), pp. 2096–2126.
[4] H. Arbabi and I. Mezić, Study of dynamics in post-transient flows using koopman mode decomposition, Physical Review Fluids, 2 (2017), p. 124402.
[5] E. Attinger, A. Anne, and D. McDonald, Use of fourier series for the analysis of biological systems, Biophysical Journal, 6 (1966), p. 291.
[6] F. S. Bazán, Conditioning of rectangular vandermonde matrices with nodes in the unit disk, SIAM Journal on Matrix Analysis and Applications, 21 (2000), pp. 679–693.
[7] L. Berman and A. Feuer, On perfect conditioning of vandermonde matrices on the unit circle, Electronic Journal of Linear Algebra, 16 (2007), p. 13.
[8] A. Björck and V. Pereyra, Solution of vandermonde systems of equations, Mathematics of Computation, 24 (1970), pp. 893–903.
[9] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Liung, Time series analysis: forecasting and control, John Wiley & Sons, 2015.
[10] J. P. Boyd, Chebyshev and Fourier spectral methods, Courier Corporation, 2001.
[11] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, Signature verification using a “siamese” time delay neural network, in Advances in neural information processing systems, 1994, pp. 737–744.
[12] D. S. Broomhead and R. Jones, Time-series analysis, Proc. R. Soc. Lond. A, 423 (1989), pp. 103–121.
[13] D. S. Broomhead and G. P. King, Extracting qualitative dynamics from experimental data, Physica D: Nonlinear Phenomena, 20 (1986), pp. 217–236.
[14] S. L. Brunton, B. W. Brunton, J. L. Proctor, E. Kaiser, and J. N. Kutz, Chaos as an intermittently forced linear system, Nature communications, 8 (2017), p. 19.
[15] S. L. Brunton, J. L. Proctor, and J. N. Kutz, Compressive sampling and dynamic mode decomposition, arXiv preprint arXiv:1312.5186, (2013).
[16] S. L. Brunton, J. L. Proctor, and J. N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proceedings of the National Academy of Sciences, (2016), p. 201517384.
[17] L. Cao, Practical method for determining the minimum embedding dimension of a scalar time series, Physica D: Nonlinear Phenomena, 110 (1997), pp. 43–50.
[18] K. Champion, S. L. Brunton, and J. N. Kutz, Discovery of nonlinear multiscale systems: Sampling strategies and embeddings, arXiv preprint arXiv:1805.07411, (2018).
[19] S. Chen and S. A. Billings, Representations of non-linear systems: the narmax model, International Journal of Control, 49 (1989), pp. 1013–1032.
[20] A. J. Chorin and O. H. Hald, Estimating the uncertainty in underresolved nonlinear dynamics, Mathematics and Mechanics of Solids, 19 (2014), pp. 28–38.
[21] A. Córdova, W. Gautschi, and S. Ruscheweyh, Vandermonde matrices on the circle: spectral properties and conditioning, Numerische Mathematik, 57 (1990), pp. 577–591.
[22] E. R. Deyle and G. Sugihara, Generalized theorems for nonlinear state space reconstruction, PLoS ONE, (2011), https://doi.org/10.1371/journal.pone.0018295.
[23] Z. Drma ˇc, I. Mezi ´c, and R. Mohr, Data driven koopman spectral analysis in vandermonde-cauchy form via the dft: numerical method and theoretical insights, arXiv preprint arXiv:1808.09557, (2018).
[24] R. J. Frank, N. Davey, and S. P. Hunt, Time series prediction and neural networks, Journal of intelligent and robotic systems, 31 (2001), pp. 91–103.
[25] W. Gautschi, How (un) stable are vandermonde systems, Asymptotic and computational analysis, 124 (1990), pp. 193–210.
[26] J. F. Gibson, J. Doyne Farmer, M. Casdagli, and S. Eubank, An analytic approach to practical state space reconstruction, Physica. D, Nonlinear phenomena, 57 (1992), pp. 1–30.
[27] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, Deep learning, vol. 1, MIT press Cambridge, 2016.
[28] A. Gouasmi, E. J. Parish, and K. Duraisamy, A priori estimation of memory effects in reduced-order models of nonlinear systems using the mori–zwanzig formalism, Proc. R. Soc. A, 473 (2017), p. 20170385.
[29] L. T. Gruyitch, Observability and controllability of general linear systems: Observability and controllability of general linear systems, (2018).
[30] S. Gugercin and K. Willcox, Krylov projection framework for fourier model reduction, Automatica, 44 (2008), pp. 209–215.
[31] E. Jones, T. Oliphant, and P. Peterson, Scipy: open source scientific tools for {Python}, (2014).
[32] E. Kaiser, J. N. Kutz, and S. L. Brunton, Sparse identification of nonlinear dynamics for model predictive control in the low-data limit, Proceedings of the Royal Society A, 474 (2018), p. 20180335.
[33] M. Kamb, E. Kaiser, S. L. Brunton, and J. N. Kutz, Time-delay observables for koopman: Theory and applications, arXiv preprint arXiv:1810.01479, (2018).
[34] M. B. Kennel, R. Brown, and H. D. Abarbanel, Determining embedding dimension for phase-space reconstruction using a geometrical construction, Physical review A, 45 (1992), p. 3403.
[35] E. Kreindler and P. Sarachik, On the concepts of controllability and observability of linear systems, IEEE Transactions on Automatic Control, 9 (1964), pp. 192–209.
[36] S. Kunis and D. Nagel, On the condition number of vandermonde matrices with pairs of nearly-colliding nodes, arXiv preprint arXiv:1812.08645, (2018).
[37] H. Landau, Sampling, data transmission, and the nyquist rate, Proceedings of the IEEE, 55 (1967), pp. 1701–1706.
[38] K. J. Lang, A. H. Waibel, and G. E. Hinton, A time-delay neural network architecture for isolated word recognition, Neural networks, 3 (1990), pp. 23–43.
[39] S. Le Clainche and J. M. Vega, Higher order dynamic mode decomposition, SIAM Journal on Applied
linear applications, SIAM Journal on Scientific Computing, 26 (2005), pp. 944–962.