Temporal Forward–Backward Consistency, Not Residual Error, Measures the Prediction Accuracy of Extended Dynamic Mode Decomposition

Masih Haseli, *Graduate Student Member, IEEE*, and Jorge Cortés, *Fellow, IEEE*

***Abstract*—Extended Dynamic Mode Decomposition (EDMD) is a popular data-driven method to approximate the action of the Koopman operator on a linear function space spanned by a dictionary of functions. The accuracy of EDMD model critically depends on the quality of the particular dictionary span, specifically on how close it is to being invariant under the Koopman operator. Motivated by the observation that the residual error of EDMD, typically used for dictionary learning, does not encode the quality of the function space and is sensitive to the choice of basis, we introduce the novel concept of consistency index. We show that this measure, based on using EDMD forward and backward in time, enjoys a number of desirable qualities that make it suitable for data-driven modeling of dynamical systems: it measures the quality of the function space, it is invariant under the choice of basis, can be computed in closed form from the data, and provides a tight upper-bound for the relative root mean square error of all function predictions on the entire span of the dictionary.**

***Index Terms*—Dynamical systems, Koopman operator, dynamic mode decomposition, prediction accuracy, data-driven modeling.**

I. Introduction

KOOPMAN operator theory has gained widespread attention in recent years for the study of dynamical systems, chiefly thanks to the linear structure of the operator despite nonlinearities in the system. In fact, the linearity of the Koopman operator leads to computationally efficient formulations to work with data. This provides a theoretically principled and explainable approach to the challenges posed by incorporating data-driven methods in the modeling, analysis, and control of dynamical systems. Given that the Koopman operator is generally infinite-dimensional, finding accurate finite-dimensional approximations for its action is of utmost importance. The first step to do so is to have a proper way to efficiently measure the quality of the subspace. This measure in turn can be used for subspace identification in optimization or neural network-based learning methods. Defining such a measure is, in turn, the goal of this letter.

Literature Review: The Koopman operator [1] provides an alternative representation of the evolution of dynamical systems in terms of observables rather than system trajectories. Despite possible nonlinearities in the system, the eigenfunctions of the Koopman operator evaluated on the system's trajectories have linear temporal evolution, and this leads to efficient numerical methods used in complex system analysis [2], [3], estimation [4], control [5], [6], [7], [8], and robotics [9], [10], to name a few. Despite these appealing applications, the infinite-dimensional nature of the operator makes its direct use on digital computers challenging. Dynamic Mode Decomposition (DMD) [11] and its generalization Extended Dynamic Mode Decomposition (EDMD) [12] are popular data-driven methods to approximate the action of the Koopman operator on finite-dimensional spaces. EDMD in particular uses a dictionary of functions whose span specifies the finite-dimensional space of choice. The work in [13] studies the prediction accuracy of DMD while [14] provides several convergence results for EDMD as the number of data points and dimension of the dictionary go to infinity. The dependence of the EDMD's prediction accuracy on the choice of the dictionary has led to a search for dictionaries whose span is close to being invariant under the Koopman operator [15]. The works in [16], [17] use methods based on neural networks for this task, while [18] directly learns the Koopman eigenfunctions spanning invariant subspaces. Moreover, the works in [19], [20] approximate finite-dimensional Koopman models relying on knowledge about the system's attractors and their stability. In our previous work, we have provided efficient algebraic algorithms to identify exact Koopman-invariant subspaces [21], [22] or approximate them with tunable predefined accuracy [23].

Statement of Contributions: Our starting point[1] is the observation that the residual error of EDMD, typically used for dictionary learning, does not necessarily measure the quality of the subspace spanned by the dictionary and, consequently,

Manuscript received 15 July 2022; revised 15 September 2022; accepted 30 September 2022. Date of publication 12 October 2022; date of current version 24 October 2022. This work was supported in part by ONR Award under Grant N00014-18-1-2828, and in part by NSF Award under Grant IIS-2007141. Recommended by Senior Editor R. S. Smith. *(Corresponding author: Masih Haseli.)*

The authors are with the Department of Mechanical and Aerospace Engineering, University of California at San Diego, La Jolla, CA 92093 USA (e-mail: mhaseli@ucsd.edu; cortes@ucsd.edu).

Digital Object Identifier 10.1109/LCSYS.2022.3214476

[1]We use the following notation. The symbols $\mathbb{N}$, $\mathbb{R}$, and $\mathbb{C}$, represent the sets of natural, real, and complex numbers resp. Given $A \in \mathbb{C}^{m\times n}$, we denote its transpose, pseudo-inverse, conjugate transpose, Frobenius norm and range space by A^T, $A^\dagger$, A^H, $\|A\|_F$, and $\mathcal{R}(A)$ resp. If A is square, we use A^{-1} to denote its inverse. We denote by $\text{spec}(A)$ the spectrum of A. Similarly, $\text{spec}_{\neq 0}(A)$ denotes the set of nonzero eigenvalues of A. Moreover, $\text{sprad}(A) := \max\{|\lambda| \mid \lambda \in \text{spec}(A)\}$ is the spectral radius of A. If $\text{spec}(A) \subset \mathbb{R}$, then $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalues of A. We use I_m and $\mathbf{0}_{m\times n}$ to denote the $m \times m$ identity matrix and $m \times n$ zero matrix (we drop the indices when appropriate). We denote the 2-norm of the vector $v \in \mathbb{C}^n$ by $\|v\|_2$. Given sets S_1 and S_2, their union and intersection are represented by $S_1 \cup S_2$ and $S_1 \cap S_2$. Also, $S_1 \subseteq S_2$ and $S_1 \subsetneq S_2$ resp. mean that S_1 is a subset and proper subset of S_2. Given the vector space $\mathcal{V}$ defined on the field $\mathbb{C}$, $\dim \mathcal{V}$ denotes its dimension. Moreover, given a set $\mathcal{S} \subseteq \mathcal{V}$, $\text{span}(\mathcal{S})$ is a vector space comprised of all linear combinations of elements in $\mathcal{S}$. If vectors $v, w \in \mathbb{R}^n$ and vector spaces $\mathcal{V}, \mathcal{W} \subseteq \mathbb{R}^n$ are orthogonal, we write $v \perp w$ and $\mathcal{V} \perp \mathcal{W}$. Moreover, $\mathcal{V}^\perp$ denotes the orthogonal complement of $\mathcal{V}$. Given functions f and g with appropriate domains and co-domains, $f \circ g$ denotes their composition.

the prediction accuracy of EDMD on the subspace. To illustrate this point, we provide an example showing that one can choose a sequence of dictionaries spanning the same subspace that make the residual error arbitrarily close to zero. This motivates our goal of identifying better measures to assess the EDMD's prediction accuracy and its dictionary's quality. We define the notion of the consistency matrix and its spectral radius, which we term consistency index, which measures the deviation of the EDMD solutions forward and backward in time from being the inverse of each other. This is justified by the fact that if a subspace is Koopman invariant, the EDMD solutions applied forward and backward in time are the inverse of each other. We characterize various algebraic properties of the consistency index and show that it only depends on the data and the space spanned by the dictionary, and is hence invariant under changes of basis. We also establish that the square root of the consistency index provides a tight upper bound on the relative root mean square EDMD prediction error of all functions in the dictionary's span.

II. Preliminaries

We briefly recall basic facts about the Koopman operator [24] and Extended Dynamic Mode Decomposition [12].

Koopman Operator: Consider a dynamical system with state space $\mathcal{M} \subseteq \mathbb{R}^n$

$$x^+ = T(x),\ x \in \mathcal{M}. \tag{1}$$

Let $\mathcal{F}$ be a vector space defined on $\mathbb{C}$ comprised of functions from $\mathcal{M}$ to $\mathbb{C}$ whose composition with T also belong to $\mathcal{F}$. The Koopman operator associated with the dynamics is

$$\mathcal{K}f = f \circ T. \tag{2}$$

The operator is linear. Its eigenfunctions have linear evolution on the trajectories of the system, i.e., given eigenfunction ϕ with eigenvalue λ, $\phi(x^+) = \phi \circ T(x) = \mathcal{K}\phi(x) = \lambda\phi(x)$. This results in a significant property of the Koopman eigendecomposition: given eigenpairs $\{(\lambda_i, \phi_i)\}_{i=1}^{N_k}$, the evolution of $f = \sum_{i=1}^{N_k} c_i\phi_i$ on a trajectory of the system starting from $x_0 \in \mathcal{M}$ is $f(x(k)) = \sum_{i=1}^{N_k} c_i\lambda_i^k\phi_i(x_0)$, for $k \in \mathbb{N}$. This linear property is useful for both prediction and identification, as the linearity always holds even if the system is nonlinear. However, to completely capture the dynamics, one might need the space $\mathcal{F}$ to be infinite dimensional.

Extended Dynamic Mode Decomposition: The infinite-dimensional property of the Koopman operator prevents its direct use in practical data-driven settings. This leads naturally to constructing finite-dimensional approximations, e.g., using Extended Dynamic Mode Decomposition (EDMD). EDMD uses a dictionary $D : \mathcal{M} \to \mathbb{R}^{1\times N_d}$ containing N_d functions, $D(x) = [d_1(x), \ldots, d_{N_d}(x)]$. To capture the dynamic behavior, EDMD uses data $X, Y \in \mathbb{R}^{N\times n}$ containing N data snapshots gathered from system trajectories,

$$y_i = T(x_i),\ \forall i \in \{1, \ldots, N\}, \tag{3}$$

where x_i^T and y_i^T correspond to the ith rows of X and Y. For convenience, we define the action of D on a data matrix as $D(X) = [D(x_1)^T, D(x_2)^T, \ldots, D(x_N)^T]^T$, where x_i^T is the ith row of X. EDMD approximates the action of the Koopman operator on span(D) by solving

$$\underset{K}{\text{minimize}} \|D(Y) - D(X)K\|_F \tag{4}$$

which has the closed-form solution

$$K_{\text{EDMD}} = \text{EDMD}(D, X, Y) \coloneqq D(X)^\dagger D(Y). \tag{5}$$

We rely on the following basic assumption.

Assumption 1 (Full Rank Dictionary Matrices): $D(X)$ and $D(Y)$ have full column rank. □

Assumption 1 implies that the functions in D are linearly independent, i.e., they form a basis for span(D) and the data are diverse enough to distinguish between the elements of D. Assumption 1 ensures that K_{EDMD} is the unique solution for (4). One can use K_{EDMD} to approximate the Koopman eigenfunctions and, more importantly, the action of the operator on span(D). Given $f \in \text{span}(D)$ in the form of $f(\cdot) = D(\cdot)v_f$ for $v_f \in \mathbb{C}^{N_d}$, one defines the EDMD predictor function for $\mathcal{K}f$ as

$$\mathcal{P}_{\mathcal{K}f}(\cdot) = D(\cdot)K_{\text{EDMD}}v_f. \tag{6}$$

The predictor's quality depends on the quality of the dictionary. If span(D) is Koopman-invariant (i.e., $\mathcal{K}g \in \text{span}(D)$ for all $g \in \text{span}(D)$), the predictor (6) is exact (otherwise, the prediction is inexact for some functions in the space).

III. Motivation and Problem Statement

We consider the prediction accuracy of EDMD for all (uncountably) functions in the dictionary span, as opposed to only finitely many functions. Since in general the dynamics is unknown, it is important to use data to learn a proper dictionary tailored to the dynamics. The *residual error* of EDMD, $\|D(Y) - D(X)K_{\text{EDMD}}\|_F$, is commonly used for this purpose as an objective function in optimization and neural network-based learning schemes. However, it is important to note that even though a high quality subspace (close to Koopman-invariant) leads to small residual error, the converse is not true: a dictionary D with small residual error does not necessarily mean that EDMD's prediction is accurate on span(D).

Example 1 (Residual Error Is Not Invariant Under Linear Transformation of Dictionary): Consider the linear system $x^+ = 0.5x$ and the vector space of functions $\mathcal{S} = \text{span}\{x, x^3 - x^2\}$. To apply EDMD, we gather $N = 1000$ data snapshots from trajectories of the system with length of two time steps and initial conditions uniformly selected from $[-2, 2]$, and form data matrices $X, Y \in \mathbb{R}^{N\times 1}$. We consider a family of dictionaries parameterized by $\alpha \in \mathbb{R}\backslash\{0\}$,

$$D_\alpha(x) = [x, x + \alpha(x^3 - x^2)]. \tag{7}$$

Note that each D_α is a basis for $\mathcal{S}$, and all the dictionaries are related by nonsingular linear transformations. We also define two notions of prediction accuracy: the residual error E of EDMD and its normalized version E_{rel},

$$E(\alpha) = \|D_\alpha(Y) - D_\alpha(X)K_\alpha\|_F,\ \ E_{\text{rel}}(\alpha) = \frac{E(\alpha)}{\|D_\alpha(Y)\|_F},$$

where $K_\alpha = \text{EDMD}(D_\alpha, X, Y)$.

Figure 1 shows the aforementioned notions of error versus the value of $\alpha \in [0.01, 100]$. Figure 1 clearly demonstrates the sensitivity of errors to the choice of basis for $\mathcal{S}$ despite the invariance of spec$(K_{\text{EDMD},\alpha})$ under the choice of basis (see, e.g., [23, Lemma 7.1]). By tuning α, one can make both errors arbitrarily close to zero. As a result, using the residual error as a measure to assess the quality of the space or as an objective function in optimization or neural network-based dictionary learning schemes can lead to erroneous results. □

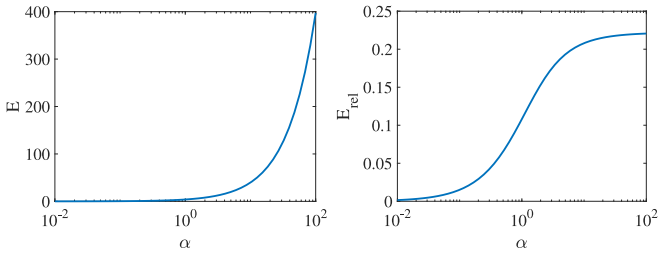

Fig. 1. EDMD's residual error (left) and relative residual error (right) as a function of α in (7).

Remark 1 (Prediction Accuracy of Dictionary Elements versus Dictionary's Span): Some applications only require short-term prediction of *finitely* many observables. In such cases, the residual error of EDMD may be useful: the observables of choice are fixed as elements of the dictionary and the rest of the functions composing the dictionary are learned by minimizing the residual error (which measures the average one time-step prediction error). However, such methods do not necessarily lead to a dictionary which spans an approximate Koopman-invariant subspace. □

The observations in Example 1 prompts us to search for a better measure of the dictionary's quality and therefore the EDMD's prediction accuracy. We formalize this next.

Problem 1 (Characterization of EDMD's Prediction Accuracy and the Dictionary's Quality): Given a dictionary D and data matrices X and Y, under Assumption 1, we aim to provide a data-driven measure of EDMD's accuracy and the dictionary's quality that

(a) only depends on span(D), X, and Y, and hence is invariant under the choice of basis for span(D), i.e., given D' as an alternative basis for span(D), the accuracy measures calculated based on D and D' are equal;
(b) provides a data-driven bound on the distance between $\mathcal{K}f$ and its EDMD prediction $\mathfrak{P}_{\mathcal{K}f}$ for all functions $f \in$ span(D);
(c) can be computed using a closed-form formula (for implementation in optimization solvers). □

IV. Temporal Forward-Backward Consistency

Here, we take the first step towards finding an appropriate measure for EDMD's prediction accuracy by comparing the solutions of EDMD forward and backward in time.[2] Throughout this letter, we use the following notation for forward and backward EDMD matrices

$$K_F = \text{EDMD}(D, X, Y) = D(X)^\dagger D(Y), \tag{8a}$$
$$K_B = \text{EDMD}(D, Y, X) = D(Y)^\dagger D(X). \tag{8b}$$

We rely on the observation that if the dictionary spans a Koopman-invariant subspace, then $K_F K_B = I$. Otherwise, the forward and backward EDMD matrices will not be the inverse of each other, which motivates the next definition.

Definition 1 (Consistency Matrix and Index): Given dictionary D and data matrices X and Y, the *consistency matrix* is $M_C(D, X, Y) = I - K_F K_B$ and the *consistency index* is $\mathcal{I}_C(D, X, Y) = \text{sprad}(M_C(D, X, Y))$. □

[2]The idea of looking forward and backward in time has been considered in the literature for different purposes, such as improving DMD to deal with noisy data [25], [26] and identifying exact Koopman eigenfunctions in our previous work [21] but, to the best of our knowledge, not for formally characterizing EDMD's prediction accuracy.

For convenience, we refer to $M_C(D, X, Y)$ and $\mathcal{I}_C(D, X, Y)$ as M_C and $\mathcal{I}_C$ when the context is clear. Next, we show that the eigenvalues of the consistency matrix are invariant under linear transformations of the dictionary.

Proposition 1 (Consistency Matrix's Spectrum Is Invariant under Linear Transformation of Dictionary): Let D and $\tilde{D}$ be two dictionaries such that $\tilde{D}(\cdot) = D(\cdot)R$, where R is an invertible matrix. Moreover, given data matrices X and Y let Assumption 1 hold. Then,

(a) $M_C(\tilde{D}, X, Y) = R^{-1} M_C(D, X, Y) R$;
(b) $\text{spec}(M_C(D, X, Y)) = \text{spec}(M_C(\tilde{D}, X, Y))$.

Proof: Note that part (b) directly follows from part (a) and the fact that similarity transformations preserve the eigenvalues. To show part (a), define for convenience,

$$K_F = D(X)^\dagger D(Y), \qquad K_B = D(Y)^\dagger D(X),$$
$$\tilde{K}_F = \tilde{D}(X)^\dagger \tilde{D}(Y), \qquad \tilde{K}_B = \tilde{D}(Y)^\dagger \tilde{D}(X).$$

We start by showing that $K_F K_B$ and $\tilde{K}_F \tilde{K}_B$ are similar. By definition, one can write

$$\tilde{K}_F \tilde{K}_B = \tilde{D}(X)^\dagger \tilde{D}(Y) \tilde{D}(Y)^\dagger \tilde{D}(X). \tag{9}$$

Moreover, given Assumption 1 and the definition of $\tilde{D}$,

$$\tilde{D}(X)^\dagger = \left(\tilde{D}(X)^T \tilde{D}(X)\right)^{-1} \tilde{D}(X)^T$$
$$= \left(R^T D(X)^T D(X) R\right)^{-1} R^T D(X)^T = R^{-1} D(X)^\dagger. \tag{10}$$

Equations (9)-(10), in conjunction with the fact that $\tilde{D}(Y)^\dagger \tilde{D}(Y) = D(Y)^\dagger D(Y)$ (cf. Lemma A.1 in the appendix), imply $\tilde{K}_F \tilde{K}_B = R^{-1} K_F K_B R$ directly leading to the required identity following Definition 1. ■

According to Proposition 1, the spectrum of the consistency matrix is a property of the data and the *vector space* spanned by the dictionary, as opposed to the dictionary itself. This property is consistent with the requirement in Problem 1(a). Next, we further investigate the eigendecomposition of the consistency matrix.

Lemma 1 (Consistency Matrix's Properties): Given Assumption 1, the consistency matrix $M_C(D, X, Y)$ satisfies:

(a) it is similar to a symmetric matrix;
(b) it is diagonalizable with a complete set of eigenvectors;
(c) $\text{spec}(M_C(D, X, Y)) \subset [0, 1]$.

Proof: (a) Given Assumption 1, there exists an invertible matrix R such that the columns of $D(X)R$ are orthonormal. Define the dictionary $\tilde{D}(\cdot) = D(\cdot)R$. Note that $\tilde{D}(X)^T \tilde{D}(X) = I_{N_d}$ and hence $\tilde{D}(X)^\dagger = (\tilde{D}(X)^T \tilde{D}(X))^{-1} \tilde{D}(X)^T = \tilde{D}(X)^T$. Hence,

$$M_C(\tilde{D}, X, Y) = I - \tilde{D}(X)^T \tilde{D}(Y) \tilde{D}(Y)^\dagger \tilde{D}(X).$$

Noting that $\tilde{D}(Y)\tilde{D}(Y)^\dagger$ is symmetric, we deduce that $M_C(\tilde{D}, X, Y)$ is symmetric. Then (a) directly follows by the definition of R and Proposition 1(a).

(b) The proof directly follows from part (a) and the fact that symmetric matrices are diagonalizable and have a complete set of eigenvectors.

(c) From part (a), we deduce that $M_C(D, X, Y)$ has real eigenvalues. Since $M_C(D, X, Y) = I - D(X)^\dagger D(Y) D(Y)^\dagger D(X)$, we only need to show

$$\text{spec}\Big(D(X)^\dagger D(Y) D(Y)^\dagger D(X)\Big) \subset [0, 1]. \tag{11}$$

Consider an eigenvector $v \in \mathbb{R}^{N_d} \backslash \{0\}$ with eigenvalue μ, i.e., $D(X)^\dagger D(Y) D(Y)^\dagger D(X) v = \mu v$. Multiplying both sides from the left by $D(X)$ and defining $w = D(X)v$ leads

to $D(X)D(X)^\dagger D(Y)D(Y)^\dagger w = \mu w$. Next, multiplying this equation from the left by w^T,

$$w^T D(X)D(X)^\dagger D(Y)D(Y)^\dagger w = \mu \|w\|_2^2. \tag{12}$$

The fact that $D(X)D(X)^\dagger$ is symmetric and represents the orthogonal projection operator on $\mathcal{R}(D(X))$, in conjunction with $w \in \mathcal{R}(D(X))$, allows us to write $w^T D(X)D(X)^\dagger = w^T$. This, combined with (12), yields $\mu = \frac{w^T D(Y)D(Y)^\dagger w}{\|w\|_2^2}$. Hence, $\lambda_{\min}(D(Y)D(Y)^\dagger) \leq \mu \leq \lambda_{\max}(D(Y)D(Y)^\dagger)$. However, since $D(Y)D(Y)^\dagger$ is an orthogonal projection operator, we have $\text{spec}(D(Y)D(Y)^\dagger) \subset [0, 1]$. Hence, $\mu \in [0, 1]$, leading to (11), concluding the proof. ∎

From Lemma 1, the consistency matrix is similar to a positive semidefinite matrix. The larger the eigenvalues of M_C, the more inconsistent the forward and backward EDMD models get. Also, from Lemma 1, $\mathcal{I}_C = \lambda_{\max}(M_C) \in [0, 1]$. Intuitively, the consistency index determines the quality of the subspace spanned by the dictionary and the prediction accuracy of EDMD on it. This is formalized next.

V. Consistency Index Determines EDMD's Prediction Accuracy on Data

Our main result states that the square root of the consistency index is a tight upper bound for the relative root mean square prediction error of EDMD.

Theorem 1 [$\sqrt{\mathcal{I}_C}$ Bounds the Relative Root Mean Square Error (RRMSE) of EDMD]: For dictionary D and data matrices X, Y, under Assumption 1,

$$\text{RRMSE}_{\max} := \max_{f \in \text{span}(D)} \frac{\sqrt{\frac{1}{N}\sum_{i=1}^N |\mathcal{K}f(x_i) - \mathfrak{P}_{\mathcal{K}f}(x_i)|^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^N |\mathcal{K}f(x_i)|^2}} = \sqrt{\mathcal{I}_C(D, X, Y)}.$$

where x_i is the ith row of X and $\mathfrak{P}_{\mathcal{K}f}$ is defined in (6).

Note that the combination of Definition 1, Lemma 1, and Theorem 1 mean that $\sqrt{\mathcal{I}_C(D, X, Y)}$ satisfies all the requirements in Problem 1.[3] Before proving the result, we first remark its importance regarding function predictions.

Remark 2 ($\sqrt{\mathcal{I}_C}$ Determines the Relative L_2-Norm Error of EDMD's Prediction Under Empirical Measure): Given that the elements of span(D) and their composition with T are measurable and considering the empirical measure $\mu_X = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$, where δ_{x_i} is the Dirac measure defined based on the ith row of X, one can rewrite $\text{RRMSE}_{\max}$ as

$$\text{RRMSE}_{\max} = \max_{f \in \text{span}(D)} \frac{\|\mathcal{K}f - \mathfrak{P}_{\mathcal{K}f}\|_{L_2(\mu_X)}}{\|\mathcal{K}f\|_{L_2(\mu_X)}} = \sqrt{\mathcal{I}_C}. \quad \square$$

To prove Theorem 1, we first provide the following alternative expression of the consistency index.

Theorem 2 (Consistency Index and Difference of Projections): Given Assumption 1,

$$\sqrt{\mathcal{I}_C(D, X, Y)} = \text{sprad}\Big(D(Y)D(Y)^\dagger - D(X)D(X)^\dagger\Big).$$

Proof: From Lemma 1, we have $\mathcal{I}_C = \lambda_{\max}(M_C)$. We use the following notation throughout the proof,

$$\lambda_{\max} = \mathcal{I}_C,\ \mathcal{P}_{D(X)} = D(X)D(X)^\dagger,\ \mathcal{P}_{D(Y)} = D(Y)D(Y)^\dagger.$$

[3]In fact, if one were to plot it as a function of $\alpha \in \mathbb{R}\backslash\{0\}$ in Example 1, one would obtain a constant value (unlike the residual error plotted in Figure 1), showing it correctly encodes the quality of the vector space.

Note that $\mathcal{P}_{D(X)}$ and $\mathcal{P}_{D(Y)}$ are projection operators on $\mathcal{R}(D(X))$ and $\mathcal{R}(D(Y))$, resp. By Definition 1, given an eigenvalue $\lambda \in [0, 1]$ of M_C with eigenvector $v \neq 0$,

$$M_C v = \lambda v \Leftrightarrow K_F K_B v = (1 - \lambda)v. \tag{13}$$

We consider the cases (i) $\lambda_{\max} = 0$, (ii) $\lambda_{\max} = 1$, and (iii) $\lambda_{\max} \in (0, 1)$ separately.

Case (i) ($\lambda_{max} = 0$): In this case, from Lemma 1, we deduce $M_C = 0$. Consequently, $K_F K_B = I$. By multiplying both sides from the left by $D(X)$ and collecting the terms, we have $\mathcal{P}_{D(X)}\mathcal{P}_{D(Y)}D(X) = D(X)$. Hence, one can write

$$\mathcal{P}_{D(X)}\mathcal{P}_{D(Y)}z = z,\ \forall z \in \mathcal{R}(D(X)). \tag{14}$$

Based on [27, summary table in p. 298], we deduce

$$\mathcal{P}_{D(X)}\mathcal{P}_{D(Y)}w = w \Leftrightarrow w \in \mathcal{R}(D(X)) \cap \mathcal{R}(D(Y)). \tag{15}$$

Using (14)-(15), one can write $\mathcal{R}(D(X)) \subseteq \mathcal{R}(D(X)) \cap \mathcal{R}(D(Y))$ and consequently $\mathcal{R}(D(X)) \subseteq \mathcal{R}(D(Y))$. By a similar argument as above and swapping K_F with K_B and $D(X)$ with $D(Y)$, one can also deduce $\mathcal{R}(D(Y)) \subseteq \mathcal{R}(D(X))$. Hence, $\mathcal{R}(D(X)) = \mathcal{R}(D(Y))$. Moreover, since the orthogonal projection on a subspace is unique, we have $\mathcal{P}_{D(Y)} - \mathcal{P}_{D(X)} = 0$, concluding the proof for this part.

Case (ii) ($\lambda_{max} = 1$): By setting $\lambda = \lambda_{\max}$ in (13), multiplying both sides from the left by $D(X)$, defining $w := D(X)v$, we have

$$\mathcal{P}_{D(X)}\mathcal{P}_{D(Y)}w = 0. \tag{16}$$

Hence, noting that $w \neq 0$ (based on Assumption 1 and the fact that $v \neq 0$), we can deduce it is an eigenvector of $\mathcal{P}_{D(X)}\mathcal{P}_{D(Y)}$ with eigenvalue 0. We show next that $w \in \mathcal{R}(D(X)) \cap \mathcal{R}(D(Y))^\perp$. One can write $\mathcal{R}(D(X))$ as the direct sum of the orthogonal subspaces $\mathcal{R}(D(X)) \cap \mathcal{R}(D(Y))$ and $\mathcal{R}(D(X)) \cap \mathcal{R}(D(Y))^\perp$. Hence, we uniquely decompose $w \in \mathcal{R}(D(X))$ as $w = w_{D(Y)} + w_{D(Y)^\perp}$, where $w_{D(Y)} \in \mathcal{R}(D(X)) \cap \mathcal{R}(D(Y))$ and $w_{D(Y)^\perp} \in \mathcal{R}(D(X)) \cap \mathcal{R}(D(Y))^\perp$. Noting that $\mathcal{P}_{D(Y)}w_{D(Y)^\perp} = 0$ and $\mathcal{P}_{D(Y)}w_{D(Y)} = w_{D(Y)}$, we get from (16) that $\mathcal{P}_{D(X)}\mathcal{P}_{D(Y)}w = \mathcal{P}_{D(X)}w_{D(Y)} = 0$. Since $w_{D(Y)} \in \mathcal{R}(D(X))$, we deduce $w_{D(Y)} = 0$ and consequently,

$$w = w_{D(Y)^\perp} \in \mathcal{R}(D(X)) \cap \mathcal{R}(D(Y))^\perp.$$

Therefore, $(\mathcal{P}_{D(Y)} - \mathcal{P}_{D(X)})w = -w$ and, given that $w \neq 0$, we deduce that $\mathcal{P}_{D(Y)} - \mathcal{P}_{D(X)}$ has an eigenvalue equal to -1. Since $\text{spec}(\mathcal{P}_{D(Y)} - \mathcal{P}_{D(X)}) \subset [-1, 1]$, cf. [27, Lemma 1], we conclude $\text{sprad}(\mathcal{P}_{D(Y)} - \mathcal{P}_{D(X)}) = 1$. The proof concludes by noting that $\mathcal{I}_C = \lambda_{\max} = 1$.

Case (iii) [$\lambda_{max} \in (0, 1)$]: Using Lemma A.2 and the closed-form expressions of K_F, K_B, $\mathcal{P}_{D(X)}$, and $\mathcal{P}_{D(Y)}$,

$$\text{spec}_{\neq 0}(K_F K_B) = \text{spec}_{\neq 0}(\mathcal{P}_{D(X)}\mathcal{P}_{D(Y)}). \tag{17}$$

Given $\mu \in (0, 1)$, from [27, Ths. 1-2], $\mu \in \text{spec}_{\neq 0}(\mathcal{P}_{D(X)}\mathcal{P}_{D(Y)})$ if and only if $\{\pm\sqrt{1-\mu}\} \subset \text{spec}_{\neq 0}(\mathcal{P}_{D(Y)} - \mathcal{P}_{D(X)})$. Setting $\mu = 1 - \lambda$, $\lambda \in (0, 1)$, one can use this in conjunction with (13) and (17) to write

$$\lambda \in \text{spec}_{\neq 0}(M_C) \Leftrightarrow \{\pm\sqrt{\lambda}\} \subset \text{spec}_{\neq 0}\big(\mathcal{P}_{D(Y)} - \mathcal{P}_{D(X)}\big).$$

This, in conjunction with [27, Th. 1] and the fact that $\text{sprad}(\mathcal{P}_{D(Y)} - \mathcal{P}_{D(X)}) \leq 1$ (cf. [27, Lemma 1]), shows that if $\text{sprad}(\mathcal{P}_{D(Y)} - \mathcal{P}_{D(X)}) < 1$, then the result holds. To conclude the proof, we need to show that $\text{sprad}(\mathcal{P}_{D(Y)} - \mathcal{P}_{D(X)}) = 1$ is not true. By contradiction, suppose this is the case, then at least one of the following holds:

(a) $\exists w_1 \in \mathbb{R}^N \backslash \{0\};\ (\mathcal{P}_{D(Y)} - \mathcal{P}_{D(X)})w_1 = -w_1$,
(b) $\exists w_2 \in \mathbb{R}^N \backslash \{0\};\ (\mathcal{P}_{D(Y)} - \mathcal{P}_{D(X)})w_2 = w_2$.

For case (a), based on [27, summary table in p. 298],

$$w_1 \in \mathcal{R}(D(X)) \cap \mathcal{R}(D(Y))^{\perp}.$$

Now, consider the vector $p_1 \neq 0$ with $w_1 = D(X)p_1$. Consequently, one can write $K_F K_B p_1 = D(X)^{\dagger}\mathcal{P}_{D(Y)}w_1 = 0$, where in the last equality we have used $w_1 \perp \mathcal{R}(D(Y))$. However, this implies that $M_C p_1 = p_1$, contradicting the fact that $\lambda_{\max} \in (0, 1)$.

For case (b), note that $w_2 \in \mathcal{R}(D(Y)) \cap \mathcal{R}(D(X))^{\perp}$ (see, e.g., [27, summary table in p. 298]). Consider the vector space $\mathcal{S} = \mathcal{R}(D(Y)) \cap \mathcal{R}(w_2)^{\perp}$. Clearly $\dim \mathcal{S} < \dim \mathcal{R}(D(Y)) = \dim \mathcal{R}(D(X))$. Consequently, there exists a non-zero vector $w^* \in \mathcal{R}(D(X))$ such that $w^* \perp \mathcal{S}$. Also, $w^* \perp \mathcal{R}(w_2)$ since $w_2 \perp \mathcal{R}(D(X))$. Hence, by noting that $\mathcal{R}(D(Y))$ is the direct sum of $\mathcal{S}$ and $\mathcal{R}(w_2)$, one can conclude $w^* \in \mathcal{R}(D(X)) \cap \mathcal{R}(D(Y))^{\perp}$ and, as a result, we have $(\mathcal{P}_{D(Y)} - \mathcal{P}_{D(X)})w^* = -w^*$. Since w^* satisfies the identity in case (a), the proof follows by replacing w_1 with w^* in the proof of case (a). ■

Remark 3 (Geometric Connections to Grassmannians): Theorem 2 establishes a link between the consistency index and the spectral radius of the difference of projection matrices. Given proper confinement of subspaces with fixed dimension to a Grassmannian, see, e.g., [28], the consistency index can be viewed as a metric measuring the distance (by encoding angles) between vector spaces of fixed dimension. Similar ideas based on the difference of projections have been used in the context of dynamic mode decomposition [29]. Even if the dimension of the vector spaces is not fixed, the consistency index and difference of projections still can be used through results similar to Theorem 1. This is especially relevant if the dimension of the Koopman-invariant subspace is unknown, see, e.g., [23]. □

We are finally ready to prove Theorem 1.

Proof (Theorem 1): We use the following notation throughout the proof: $\mathcal{P}_{D(X)} = D(X)D(X)^{\dagger}$, $\mathcal{P}_{D(Y)} = D(Y)D(Y)^{\dagger}$, and $s = \operatorname{sprad}(\mathcal{P}_{D(Y)} - \mathcal{P}_{D(X)})$. Note that, from Theorem 2, $s = \sqrt{\mathcal{I}_C(D, X, Y)}$. Given an arbitrary function $f(\cdot) = D(\cdot)v_f \in \operatorname{span}(D)$, with $v_f \in \mathbb{C}^{N_d}$, one can use (2), the predictor (6) with $K_F = K_{\mathrm{EDMD}} = \mathrm{EDMD}(D, X, Y)$, and the relationship between the rows of X and Y in (3) to write

$$\mathrm{RRMSE}_f := \frac{\sqrt{\sum_{i=1}^{N} |D(y_i)v_f - D(x_i)K_F v_f|^2}}{\sqrt{\sum_{i=1}^{N} |D(y_i)v_f|^2}} = \frac{\|D(Y)v_f - D(X)K_F v_f\|_2}{\|D(Y)v_f\|_2} \tag{18}$$

Noting that $D(Y) = \mathcal{P}_{D(Y)}D(Y)$, one can write

$$\|D(Y)v_f - D(X)K_F v_f\|_2 = \|(\mathcal{P}_{D(Y)} - \mathcal{P}_{D(X)})D(Y)v_f\|_2 \leq s\|D(Y)v_f\|_2, \tag{19}$$

where the last inequality holds since the matrix $\mathcal{P}_{D(Y)} - \mathcal{P}_{D(X)}$ is symmetric and therefore its spectral radius is equal to its induced 2-norm. Based on (18)-(19), we have $\mathrm{RRMSE}_f \leq s$. Hence, by definition of $\mathrm{RRMSE}_{\max}$ in the statement of the result, we have

$$\mathrm{RRMSE}_{\max} = \max_{f \in \operatorname{span}(D)} \mathrm{RRMSE}_f \leq s \tag{20}$$

Now, we prove that the equality in (20) holds. We consider three cases: (i) $s = 0$ (ii) $s = 1$ or (iii) $s \in (0, 1)$.

Case (i): Since $\mathrm{RRMSE}_{\max} \geq 0$ by definition, in this case $\mathrm{RRMSE}_{\max} = 0$ follows directly.

Case (ii): In this case, there exists a nonzero vector[4] $p^* \in \mathcal{R}(D(Y)) \cap \mathcal{R}(D(X))^{\perp}$. Let v^* be such that $p^* = D(Y)v^*$. Using (19) for v^* instead of v_f, and the properties of p^*, one can write $\|D(Y)v^* - D(X)K_F v^*\|_2 = \|\mathcal{P}_{D(Y)}D(Y)v^*\|_2 = \|D(Y)v^*\|_2$. Hence, for the function $f^*(\cdot) = D(\cdot)v^* \in \operatorname{span}(D)$, one can use (18) to see that $\mathrm{RRMSE}_{f^*} = 1 = s$. Hence, equality holds in (20).

Case (iii): In this case $s \in (0, 1)$ and based on [27, Th. 1], the matrix $\mathcal{P}_{D(Y)} - \mathcal{P}_{D(X)}$ has two eigenvalues $\pm s$ with corresponding orthogonal eigenvectors $v_{+s}, v_{-s} \in \mathbb{R}^{N_d}$. Moreover, based on [27, Th. 1(a)], $\mathcal{P}_{D(Y)}v_{+s} \in \operatorname{span}\{v_{+s}, v_{-s}\}$. Hence, for some $\alpha, \beta \in \mathbb{R}$, we have

$$q^* := \mathcal{P}_{D(Y)}v_{+s} = \alpha v_{+s} + \beta v_{-s}.$$

Let r^* be such that $q^* = D(Y)r^*$. Now, based on the first part of (19) for r^* instead of v_f, we have

$$\begin{aligned}\|D(Y)r^* - D(X)K_F r^*\|_2 &= \|(\mathcal{P}_{D(Y)} - \mathcal{P}_{D(X)})D(Y)r^*\|_2 \\ &= \|(\mathcal{P}_{D(Y)} - \mathcal{P}_{D(X)})(\alpha v_{+s} + \beta v_{-s})\|_2 = s\|\alpha v_{+s} - \beta v_{-s}\|_2 \\ &= s\|\alpha v_{+s} + \beta v_{-s}\|_2 = s\|D(Y)r^*\|_2,\end{aligned} \tag{21}$$

where in the third and fourth equalities we have used the definition of v_{+s} and v_{-s} and their orthogonality. Now, for the function $g^*(\cdot) = D(\cdot)r^* \in \operatorname{span}(D)$, one can use (18) to see that $\mathrm{RRMSE}_{g^*} = s$. Hence, the equality in (20) holds, and this concludes the proof. ■

Remark 4 (Working With Consistency Matrix Is More Efficient Than the Difference of Projections): According to Theorems 1 and 2, one can use the consistency matrix $M_C \in \mathbb{R}^{N_d \times N_d}$ or the difference of projections $D(Y)D(Y)^{\dagger} - D(X)D(X)^{\dagger} \in \mathbb{R}^{N \times N}$ interchangeably to compute the relative root mean square error. However, the size N_d of the consistency matrix depends on the dictionary, while the size of the difference of projections depends on the size N of data. In practical settings $N \gg N_d$, and hence, working with the consistency matrix is more efficient. In fact, given moderate to large data sets, even saving the difference of projections matrix in the memory may be infeasible. The calculation of the consistency matrix requires solving two least-squares problems, which can be done recursively for large data sets. □

Remark 5 (Efficient Computation of the Consistency Index): The consistency index is defined as the spectral radius of the consistency matrix and can be computed as such. One can also use the following to compute it more efficiently: (i) the consistency index is the maximum eigenvalue of a matrix with nonnegative real eigenvalues (cf. Lemma 1(c)); (ii) given an appropriate change of coordinates making $D(X)^T D(X) = I$ (see proof of Lemma 1(a)), the consistency matrix becomes positive semi-definite. Hence, in optimization-based dictionary learning, one can add a constraint $D(X)^T D(X) = I$ and minimize the 2-norm of the consistency matrix (which is the maximum eigenvalue for positive semi-definite matrices). □

VI. Conclusion

We introduced the concept of consistency index, a data-driven measure that quantifies the accuracy of the EDMD method on a finite-dimensional functional space generated by a dictionary of functions. This measure is invariant under the

[4]The argument for the existence of p^* is similar (by swapping $D(X)$ and $D(Y)$) to the argument used for the existence of vectors w_1 and w^* in the proof of Theorem 2 (Case (iii)). We omit it for space reasons.

choice of basis of the functional space, is computable in closed form, and corresponds to the relative root mean squared error. Future work will build on the consistency index to design algebraic algorithms that find dictionaries with EDMD predictions of a predetermined level of accuracy and use it as an objective for optimization and neural network-based methods to identify dictionaries including the system state that span spaces close to being Koopman-invariant.

APPENDIX
BASIC ALGEBRAIC RESULTS

Here, we recall two results that are used in the proofs.

Lemma A.1: Let $B_1, B_2 \in \mathbb{R}^{m\times n}$ be matrices such that $\mathcal{R}(B_1) = \mathcal{R}(B_2)$. Then $B_1 B_1^\dagger = B_2 B_2^\dagger$.

This follows from the uniqueness of the orthogonal projection operator on a subspace.

Lemma A.2: Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times m}$. Then, $\text{spec}_{\neq 0}(AB) = \text{spec}_{\neq 0}(BA)$.

REFERENCES

[1] B. O. Koopman, "Hamiltonian systems and transformation in Hilbert space," *Proc. Nat. Acad. Sci.*, vol. 17, no. 5, pp. 315–318, 1931.

[2] I. Mezić, "Spectral properties of dynamical systems, model reduction and decompositions," *Nonlinear Dyn.*, vol. 41, nos. 1–3, pp. 309–325, 2005.

[3] S. P. Nandanoori, S. Sinha, and E. Yeung, "Data-driven operator theoretic methods for global phase space learning," in *Proc. Amer. Control Conf.*, 2020, pp. 4551–4557.

[4] M. Netto and L. Mili, "A robust data-driven Koopman Kalman filter for power systems dynamic state estimation," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 7228–7237, Nov. 2018.

[5] M. Korda and I. Mezić, "Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control," *Automatica*, vol. 93, pp. 149–160, Jul. 2018.

[6] S. Peitz and S. Klus, "Koopman operator-based model reduction for switched-system control of PDEs," *Automatica*, vol. 106, pp. 184–191, Aug. 2019.

[7] D. Goswami and D. A. Paley, "Bilinearization, reachability, and optimal control of control-affine nonlinear systems: A Koopman spectral approach," *IEEE Trans. Autom. Control*, vol. 67, no. 6, pp. 2715–2728, Jun. 2022.

[8] V. Zinage and E. Bakolas, "Neural Koopman Lyapunov control," 2022, *arXiv:2201.05098*.

[9] G. Mamakoukas, M. L. Castaño, X. Tan, and T. D. Murphey, "Derivative-based Koopman operators for real-time control of robotic systems," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 2173–2192, Dec. 2021.

[10] L. Shi and K. Karydis, "Enhancement for robustness of Koopman operator-based data-driven mobile robotic systems," 2021, *arXiv:2103.00812*.

[11] P. J. Schmid, "Dynamic mode decomposition of numerical and experimental data," *J. Fluid Mech.*, vol. 656, pp. 5–28, Jul. 2010.

[12] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley, "A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition," *J. Nonlinear Sci.*, vol. 25, no. 6, pp. 1307–1346, 2015.

[13] H. Lu and D. M. Tartakovsky, "Prediction accuracy of dynamic mode decomposition," *SIAM J. Sci. Comput.*, vol. 42, no. 3, pp. A1639–A1662, 2020.

[14] M. Korda and I. Mezić, "On convergence of extended dynamic mode decomposition to the Koopman operator," *J. Nonlinear Sci.*, vol. 28, no. 2, pp. 687–710, 2018.

[15] S. L. Brunton, B. W. Brunton, J. L. Proctor, and J. N. Kutz, "Koopman invariant subspaces and finite linear representations of nonlinear dynamical systems for control," *PLoS ONE*, vol. 11, no. 2, pp. 1–19, 2016.

[16] Q. Li, F. Dietrich, E. M. Bollt, and I. G. Kevrekidis, "Extended dynamic mode decomposition with dictionary learning: A data-driven adaptive spectral decomposition of the Koopman operator," *Chaos*, vol. 27, no. 10, 2017, Art. no. 103111.

[17] N. Takeishi, Y. Kawahara, and T. Yairi, "Learning Koopman invariant subspaces for dynamic mode decomposition," in *Proc. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1130–1140.

[18] M. Korda and I. Mezić, "Optimal construction of Koopman eigenfunctions for prediction and control," *IEEE Trans. Autom. Control*, vol. 65, no. 12, pp. 5114–5129, Dec. 2020.

[19] P. Bevanda, M. Beier, S. Kerz, A. Lederer, S. Sosnowski, and S. Hirche, "Diffeomorphically learning stable Koopman operators," 2021, *arXiv:2112.04085*.

[20] F. Fan, B. Yi, D. Rye, G. Shi, and I. R. Manchester, "Learning stable Koopman embeddings," 2021, *arXiv:2110.06509*.

[21] M. Haseli and J. Cortés, "Learning Koopman eigenfunctions and invariant subspaces from data: Symmetric subspace decomposition," *IEEE Trans. Autom. Control*, vol. 67, no. 7, pp. 3442–3457, Jul. 2022.

[22] M. Haseli and J. Cortés, "Parallel learning of Koopman Eigenfunctions and invariant subspaces for accurate long-term prediction," *IEEE Trans. Control Netw. Syst.*, vol. 8, no. 4, pp. 1833–1845, Dec. 2021.

[23] M. Haseli and J. Cortés, "Generalizing dynamic mode decomposition: Balancing accuracy and expressiveness in Koopman approximations," *Automatica*, *arXiv:2108.03712*.

[24] M. Budišić, R. Mohr, and I. Mezić, "Applied Koopmanism," *Chaos*, vol. 22, no. 4, 2012, Art. no. 47510.

[25] S. T. M. Dawson, M. S. Hemati, M. O. Williams, and C. W. Rowley, "Characterizing and correcting for the effect of sensor noise in the dynamic mode decomposition," *Exp. Fluids*, vol. 57, no. 3, p. 42, 2016.

[26] O. Azencot, W. Yin, and A. Bertozzi, "Consistent dynamic mode decomposition," *SIAM J. Appl. Dyn. Syst.*, vol. 18, no. 3, pp. 1565–1585, 2019.

[27] W. N. Anderson Jr., E. J. Harner, and G. E. Trapp, "Eigenvalues of the difference and product of projections," *Linear Multilinear Algebra*, vol. 17, nos. 3–4, pp. 295–299, 1985.

[28] P. A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton, NJ, USA: Princeton Univ. Press, 2009.

[29] A. Karimi and T. T. Georgiou, "The challenge of small data: Dynamic mode decomposition, redux," in *Proc. IEEE Conf. Decis. Control*, Austin, TX, USA, 2021, pp. 2276–2281.

[30] D. S. Bernstein, *Matrix Mathematics*, 2nd ed. Princeton, NJ, USA: Princeton Univ. Press, 2009.