Research Statement

Ibrahem AL Jabea

1 Introduction

My research deals with the analysis and the study of complex systems that can be modeled as graphs, simplicial or cell complexes. A central methodological issue underlying my research is finding unifying principles that govern many areas in modern data science and integrate the multiple levels of organization which are commonly found in a range of systems. My research combines a mixture of tools from topological data analysis, geometric data processing, machine learning, signal processing, statistics and topology. Broadly speaking, the research questions that I am interested are divided into two parts:

- Unifying principles on deep learning models and deep learning on higher order complexes is concerned with building deep learning models on generalized spaces such as cell complexes. This question also seeks building a precise mathematical and algorithmic theory that combines between deep learning on complexes and topological data analysis. Applications of deep learning protocols executed on such domains are massive, and they range from non-linear signal processing supported on topological spaces, computational biology and medicine, social science and art.
- Unifying principles on the general notion of data: unifying principle on the notation of data yields more cohesive algorithmic abstraction which ultimately help us to write better, easier to use, more inclusive machine learning packages

1.1 Unifying principles on deep learning models : Deep Learning On Higher Order Complexes

Over the past decade, deep learning [(11)] has been remarkably successful at solving a massive set of problems on data types, including images [(10)] and sequential data [(4)]. This success drove the extension of deep learning to other discrete domains such as sets [(3)], 3D shapes, and discrete manifolds [(5)]. While many of the extended schemes have successfully tackled notable challenges in each particular domain, the plethora of fragmented frameworks have created or resurfaced many long-standing problems in deep learning such as explainability, expressiveness and generalizability. Moreover, theoretical development proven over one discrete domain does not naturally apply to the other domains. Finally, the lack of a cohesive mathematical framework has created many ad hoc and inorganic implementations (9) and ultimately limited the set of practitioners that can potentially benefit from deep learning technologies. The main objective of my research in this area is the construction of a unifying, explainable and general mathematical framework that vastly expands the domains upon which deep learning protocols can be applied, all while maintaining intuitive conceptualization, implementation and relevance to a wide range of practical applications.

A cornerstone in this research question is the choice of domains upon which we build the deep learning protocols defining our unifying mathematical framework. The domains that I choose in my research are rather general type of complexes coming from algebraic topology: *cell complexes*. Cell complexes are generalized spaces that consist of simpler primitives called *cells*. These complexes form a natural generalization of all discrete domains that are important in practice, such as graphs, 3D shapes, and

simplicial complexes. They admit a topological structure, encoded by the way cells are connected with each other. The topological structure of a cell complex naturally models arbitrary higher order abstraction and interactions [(7)] between cells. From a deep learning standpoint, higher order interactions represent the flow of information, and the computations that are executed between these entities. Furthermore, higher order interactions have been empirically related to generalizability and expressiveness of neural networks — making cell complexes natural objects to model these properties. Cell complexes also admit combinatorial structures, which are specified by a collection of adjacency and boundary matrices. From a deep learning perspective, the combinatorial structure naturally extends the general *message passing schemes deep learning protocols,* which encompasses several state-of-the-art graph neural networks [(12; 13)]. This combinatorial description allows for intuitive manipulation, conceptualization, and implementation. Finally, cell complexes can be equipped with discrete or smooth structures — making them ideally suited to model discrete as well as smooth manifolds. Moving forward, this research question would evolve by developing the practice and theory of deep learning models on these domains. This includes but not limited to generative cell complex models and invertible cell complex models. Finally, the applications of these models to geometric data processing and topological data analysis such as mesh classification, learning mesh parameterization and topological filtration learning. This project does only provide a concrete and rigorous link between deep learning and classical objects in topological data analysis (such as filtration learning), but also expands the objects of interests in TDA from classical tools such as the persistence diagram to include other new objects that arise when studying deep learning on these topological/geometric domains.

1.2 Unifying principles on the general notion of data

The common notion of data, which is overwhelmingly preferred by machine learning researchers and practitioners, is rooted in statistical sciences [(6)]. Within this notion, data is described as a sample obtained from a probability density function defined over the domain of interest. Consider, for instance, the dimensionality reduction problem in which one is given a data $X \subset \mathcal{X}$, and the purpose is to find a mapping $f : \mathcal{X} \to \mathcal{Y}$ such that some mathematical structure related to X is preserved via f; $\mathcal{X}$ and $\mathcal{Y}$ denote the data space and the target space, respectively. While the data X is typically sampled from an unknown probability distribution, from the perspective of the dimensionality reduction algorithm, the interesting aspect of the data lies in the mathematical structure the function f is designed to preserve.

The unifying principle on the notation of data yields better algorithmic abstraction which ultimately help us to write better, easier to use, more inclusive machine learning packages. Further, such definitions can solidify our understanding of the relations between models and data.

To provide a concrete definition of data, I want to investigate two main notions. In the first notion, which I refer to as the *categorical notion*, the *invariants* of data are often the interesting object of study. Invariants of data are the structural properties of interest that are preserved under a particular type of *morphisms* (structure-preserving maps). For instance, one invariant can be the metric space structure, encoded in the distance matrix or the persistence diagrams, that is invariant under metric preserving functions (morphism). In the second notion, which I refer to as the *cochain notion*, the domain upon which the data is defined is the main element of interest. Within this perspective, we think of data as a function defined on the elementary building blocks of the domain of interest. This function is acted upon by operators defined on these domains. An example of this type is the solution of a partial differential equation on a surface.

In a signal processing setting, data can be viewed as a function defined in a discrete domain of interest and acted upon via *operators*. Consider, for instance, a discrete domain $\mathcal{M}$, say a graph or a simplicial complex. Roughly speaking, data in this case is defined in the form $f : \mathcal{M} \to \mathcal{X}$, where f associates to every building block in $v \in \mathcal{M}$ a data element $f(v) \in \mathcal{X}$, e.g. a scalar or a vector in some Euclidean space. An important example of this setting is Topological Data Analysis (TDA), where one studies the topological

properties of the domain $\mathcal{M}$ by studying the properties of functions defined on it. Another example is the Graph Signal Processing (GSP), where one is interested in the scalar signal defined on the node or the edge sets of a given graph. Similar to the categorical notion, this notion provides an abstraction that comes with important virtues: (1) it unifies notations in a single cohesive and mathematically elegant perceptive, (2) it becomes evident how to create additional similar structures, and (3) it impacts the way machine learning packages are designed and implemented.

2 Results

Towards the materialization of my objective in unifying principles in TDL, I have been a principal developer in the open-source software packages that democratize the applications of TDL for industrial and academic use. The main results so far have been the following:

1. *Article: TopoX: A Suite of Python Packages for Machine Learning on Topological Domains. Journal of Machine Learning Research (JMLR) 2024.*

2. *Source code: Developer of three transformative Python packages on Topological Deep Learning, focusing on fast robust deep learning computations for graph generalizations such as hypergraphs, simplicial complexes, and cellular complexes. Check out the packages on GitHub: https://github.com/pyt-team.*

    - *TopoNetX (TNX) Developer*
        - *Python package for modeling entities and relationships in higher order networks (meshes, simplicial complexes, etc.).*
    - *TopoModelX (TMX) Developer*
        - *Python package for efficient deep learning models on topological domains (e.g., simplicial and cell complexes).*
    - *TopoEmbedX (TEX) Developer*
        - *Python package for efficient representation learning on relational systems with topological domains (e.g., social networks, proteins).*

3. *Article: An Introduction to Topological Neural Networks. In progress.*

3 Future Research Directions

1. Defining the notion of cochains and discrete exterior calculus on complexes in terms of neural networks will guide the development and generalization of the theory of physics-informed neural networks.

2. Topological Representation Learning. The purpose of this project is to explore topological representation learning in *TopoEmbedX* (TEX) and how it can be applied to represent elements of a topological domain in Euclidean domain.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Image net classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[2] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. IEEE transactions on neural networks, 20(1):61–80, 2008.

[3] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. Salakhutdinov, and A. Smola. Deep sets. arXiv preprint arXiv:1703.06114, 2017.

[4] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In Advances in neural information processing systems, pages 3104–3112, 2014.

[5] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: going beyond euclidean data. IEEE Signal Processing Magazine, 34(4):18–42, 2017M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: going beyond euclidean data. IEEE Signal Processing Magazine, 34(4):18–42, 2017.

[6] J. Friedman, T. Hastie, R. Tibshirani, et al. The elements of statistical learning, volume 1. Springer series in statistics New York, 2001

[7] C. Bick, E. Gross, H. A. Harrington, and M. T. Schaub. What are higher-order networks? arXiv preprint arXiv:2104.11329, 2021.

[8] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. IEEE transactions on neural networks, 20(1):61–80, 2008.

[9] T. Serre. Deep learning: the good, the bad, and the ugly. Annual review of vision science, 5:399–426, 2019.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

[11] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. nature, 521(7553):436–444, 2015.

[12] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, et al. Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261, 2018.

[13] C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan, and M. Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 4602–4609, 2019.

4 Koopman Anaylsis

4.1 Introduction

With the ascent of thermodynamics and statistical mechanics at the end of the 19th century, it was observed that it is impossible to exactly determine the state of each and every particle (position, momentum, etc.) of a gas or a fluid and that is necessary to observe instead how measurement parameters such as temperature or pressure evolve over time. This led to the development of an alternative framework for dynamical systems, based on the "dynamics of observables" picture. The central object of this alternative framework is the Koopman operator: an infinite-dimensional, linear operator that is capable of capturing the dynamics of the observations of the dynamical system by capturing the essence of the evolution of observable functions along the trajectories of dynamical systems, [(2; 3)]. The Koopman operator provides a global linearization of a non-linear system and gives a promising framework for the study of nonlinear dynamical systems based on the tools of linear mathematics.

The main purpose of my Ph.D. dissertation is to study approximation methods for solving nonlinear systems using *Bernhard Koopman's Global Linearization Method* [(2)]. This approach enables the application of linear semigroup methods to a nonlinear system by focusing on the dynamics of the observables of the states rather than directly studying the dynamics of the states. In particular, we will investigate the extent to which eigenvalues and eigenfunctions of the linear Koopman operator, which generates the observations of the underlying nonlinear system, can serve as useful tools for approximating and/or studying the qualitative properties of the underlying nonlinear flow.

4.2 Locally Linear Koopman Semigroup Flows

We outlined a program to approximate locally linear Koopman semigroup flows

$$t \to T(t)g(x) := g(\sigma(t,x)), \tag{1}$$

where $t \to \sigma(t,x) \in \Omega$ is the underlying flow describing the dynamical system, $g \in \mathcal{F}(\Omega, \mathbb{C})$, and $x \in \Omega_t$. To explain the notation, here and throughout my research $\Omega \subset \mathbb{R}^N$ denotes an invariant set, $\mathcal{F}(\Omega, \mathbb{C})$ denotes the vector space of all functions (observations) $g : \Omega \to \mathbb{C}$. The map $t \to \sigma(t,x) \in \Omega$ defined for all $0 \leq t < m(x)$ denotes a flow in Ω with initial value $\sigma(0,x) = x \in \Omega$ and with blow-up time $0 < m(x) \leq \infty$ such that $\sigma(t, \sigma(s,x)) = \sigma(t+s, x)$ for all $0 \leq t, s, t+s < m(x)$. For $g \in \mathcal{F}(\Omega, \mathbb{C})$, the flow properties imply that

$$T(0)g(x) = g(x)$$

for all $x \in \Omega$ and

$$T(t)T(s)g(x) = T(t+s)g(x)$$

for all $x \in \Omega_{t+s}$, where $\Omega_t := \{x \in \Omega : m(x) > t\}$. Moreover, for $x \in \Omega_t$, the map $x \to T(t)g(x)$ is locally linear in the sense that

$$T(t)(ag_1 + g_2)(x) = aT(t)g_1(x) + T(t)g_2(x)$$

for all $g_1, g_2 \in$ and $a \in \mathbb{C}$. Clearly, if the flow is global (i.e., if $m(x) = \infty$ for all $x \in \Omega$), then $T(t)g(x) = g(\sigma(t,x))$ defines a linear operator $T(t)$ with domain and range in $\mathcal{F}(\Omega, \mathbb{C})$. However, if the flow is local (i.e., if $m(x) < \infty$ for some $x \in \Omega$), then a locally linear Koopman semigroup flow $t \to T(t)g(x) = g(\sigma(t,x))$ does not fit into any of the classical semigroup theories because $T(t)$ does not define a linear operator with domain and range in a vector space of functions. However, and this is somewhat surprising, the infinitesimal generator

$$\mathcal{K}g : x \to \lim_{t \to 0^+} \frac{T(t)g(x) - g(x)}{t} = \lim_{t \to 0^+} \frac{g(\sigma(t,x)) - g(x)}{t} \tag{2}$$

is always a well-defined linear operator with domain

$$D(\mathcal{K}) := \{g \in \mathcal{F}(\Omega, \mathbb{C}) : \text{limit in (2) exists for all } x \in \Omega\}$$

and range in $\mathcal{F}(\Omega, \mathbb{C})$. Formally, $\mathcal{K}$ is given by

$$\mathcal{K}g(x) = T'(0)g(x) = g'(x) \cdot \sigma'(0,x). \tag{3}$$

and

$$e^{t\mathcal{K}}g(x) = g(\sigma(x,t))$$

for all $x \in \Omega$ and $0 \leq t < m(x)$.

Let $F : \mathbb{R}^N \to \mathbb{R}^N$ be given by $F = (F_1, ..., F_N)$ where $F_i : \mathbb{R}^N \to \mathbb{R}$ and let $t \to \sigma(t,x) = x(t)$ be the unique classical solution of $\dot{x}(t) = F(x(t))$ with initial value $x(0) = x = (x_1, ..., x_N) \in \mathbb{R}^N$. Let $\Omega \subset \mathbb{R}^N$ be

an open subset of the domain of F such that $t \to \sigma(t,x) = x(t) \in \Omega$ for all $x \in \Omega$ and $0 \leq t \leq m(x)$ and let $\mathcal{M} \subset \mathcal{F}(\Omega, Z) = \{g : \Omega \to Z\}$ be an appropriately chosen space of observations. Then it is easy to see that $t \to \sigma(t,x) := x(t)$ defines an autonomous flow and that the corresponding Koopman operator is formally given by

$$\begin{aligned}\mathcal{K}g(x) :=& \lim_{t\to 0} \frac{g(\sigma(t,x)) - g(x)}{t} = \left.\frac{dg(\sigma(t,x))}{dt}\right|_{t=0} \\ =& \dot{g}(\sigma(t,x)) \cdot \dot{\sigma}(t,x)\Big|_{t=0} = \dot{g}(x(0)) \cdot F(x(0)) = \dot{g}(x) \cdot F(x) \\ =& \sum_{i=1}^{N} \frac{\partial g}{\partial x_i}(x) \cdot F_i(x) = \sum_{i=1}^{N} \mathcal{K}_i g(x),\end{aligned} \tag{4}$$

where $\mathcal{K}_i g(x) = \frac{\partial g}{\partial x_i}(x) \cdot F_i(x)$ generates the semigroup

$$e^{t\mathcal{K}_i} g(x) = g(x_1, ..., x_{i-1}, \sigma_i(t, x_i), x_{i+1}, ..., x_N), \tag{5}$$

and where $\sigma_i(t, x_i)$ solves the one-dimensional, separable equation

$$\dot{u}(t) = F_i(x_1, x_2, ..., x_{i-1}, u(t), x_{i+1}, ..., x_N) \text{ with } u(0) = x_i, \tag{6}$$

for flows $t \to \sigma(t,x)$ induced by solutions of ordinary differential equation $\dot{x}(t) = F(x(t))$, $x(0) = x \in \Omega$, the induced pointwise linear semigroup flow $t \to T(t)g(x) = g(\sigma(t,x))$ has a formal Lie-generator $\mathcal{K}$ that splits into a sum of one-dimensional Lie-generators $\mathcal{K}_i$, i.e,

$$\mathcal{K} = \mathcal{K}_1 + \mathcal{K}_2 + ... + \mathcal{K}_N.$$

Since formally,

$$T(t)g(x) = e^{t\mathcal{K}} g(x) = e^{t(\mathcal{K}_1 + \mathcal{K}_2 + ... + \mathcal{K}_N)} g(x),$$

it is a natural approach to try to compute $T(t)g(x) = g(\sigma(t,x)) = e^{t\mathcal{K}} g(x)$ by using operator splitting formulas like the Lie-Trotter product formula

$$(e^{\frac{t}{n}\mathcal{K}_1} e^{\frac{t}{n}\mathcal{K}_2} ... e^{\frac{t}{n}\mathcal{K}_N})^n$$

as $n \to \infty$. This will be one of the approaches we take to make precise in what way the pointwise semigroup flow $t \to T(t)g(x) = e^{t\mathcal{K}} g(x)$ can be approximated in terms of the Lie-generator $\mathcal{K}$.

We suggested an alternative way to compute $e^{t\mathcal{K}} g(x) = g(\sigma(t,x))$ via finite-dimensional invariant Koopman subspaces. Instead of working on $\mathcal{F}(\Omega, \mathbb{C})$ or a "large" subspace thereof, we aim to construct a "small," finite-dimensional subspace $\mathcal{M} \subset D(\mathcal{K}) \subset \mathcal{F}(\Omega, \mathbb{C})$, spanned by linearly independent observables $\{g_1, g_2, ..., g_n\} \in \mathcal{F}(\Omega, \mathbb{C})$, such that $\mathcal{K}\mathcal{M} \subset \mathcal{M}$. This approach allows us to lift the nonlinear dynamical system to a linear one via the observables $\{g_i\}_{i=1}^n$ and obtain information on the underlying flow $\sigma(t,x)$ describing the dynamical system.

4.3 Koopman Eigenfunctions

Also, we will indicate a particular simple way to compute $e^{t\mathcal{K}} g(x)$ for observables g that are eigenfunctions of $\mathcal{K}$ to eigenvalues λ; i.e., functions g_λ that satisfy $\mathcal{K} g_\lambda(x) = \lambda g_\lambda(x)$ for all $x \in \Omega$. In this case,

$$e^{t\mathcal{K}} g_\lambda(x) = g_\lambda(\sigma(t,x)) = e^{t\lambda} g_\lambda(x),$$

and therefore, at least formally, $\sigma(t,x) = g_\lambda^{-1}(e^{\lambda t} g_\lambda(x))$. We will provide several examples that show how $\sigma(t,x)$ can be determined by the knowledge of $g_\lambda(\sigma(t,x)) = e^{t\lambda} g_\lambda(x)$ for sufficiently many eigenfunctions g_λ. Also, as shown in [(7; 8)], the eigenvalues of the Koopman operator $\mathcal{K}$ provide information that is useful for the stability analysis of the underlying dynamical system; e.g., in certain spaces $\mathcal{M}$ of observations, finite-time blow-up occurs if and only if $\mathcal{K}$ has a positive eigenvalue. Therefore, we are proposing, among others, to study eigenvalues and eigenvectors of Koopman operators is detail.

In this setting, it is important to notice that we have the freedom to appropriately chose the subspace $\mathcal{M} \subset \mathcal{F}(\Omega, \mathbb{C})$ of observations. For example, if one can find a finite-dimensional space $\mathcal{M} = \text{span}\{g_1, g_2, ..., g_n\}$ that is invariant under $\mathcal{K}$, then the associated finite dimensional representation $\mathcal{K}|_{\mathcal{M}} = \mathbf{K}$ of the Koopman operator $\mathcal{K}$ will have an eigenvalue with an eigenfunction in $\mathcal{M}$ and the matrix exponential $e^{t\mathbf{K}} g$ can be computed explicitly for all $g \in \mathcal{M}$. **Our research focuses on the existence of finite-dimensional invariant subspaces. $\mathcal{M} \subset \mathcal{F}(\Omega, \mathbb{C})$ of observations, and, in particular, identifying such subspaces for nonlinear systems of ode's.**

4.4 Data-Driven Approach

Computing the Koopman eigenfunctions is a challenging task that depends on the complexity of the system. For instance, the Lorenz system or the Van der Pol equation do not possess an explicit eigenfunction. Therefore, a data-driven approach has been proposed to approximate the eigenfunctions. The most popular methods include Dynamic Mode Decomposition [(10)], Extended Dynamic Mode Decomposition [(13)] or employing Koopman deep learning processes [(6)].

We used a data-driven approach to approximate the Koopman eigenfunctions for specific systems. This method is particularly useful when dealing with complex systems where analytical solutions may be impossible. This approach not only enables a more practical implementation but also facilitates the study of the underlying dynamics in systems that may lack straightforward analytical solutions.

4.5 Results

Article: Approximations of Koopman Operator Semigroups. In progress.

4.6 Future Research Directions

In recent work on the Kolmogorov-Arnold algorithm, we will demonstrate how it enables greater flexibility in learnable activation functions. Consequently, we plan to extend our data-driven approach by applying the Kolmogorov-Arnold algorithm to improve the approximation of Koopman eigenfunctions. This research will guide our next article, titled Deep Learning for Universal Linear Embeddings of Nonlinear Dynamics Using the Arnold-Kolmogorov Algorithm.

References

[1] W. Arendt, A. Grabosch, G. Greiner, U. Moustakas, R. Nagel, U. Schlotterbeck, U. Groh, H. P. Lotz, F. Neubrander, One-parameter Semigroups of Positive Operators, SpringerVerlag, (1986).

[2] B. O. Koopman and J. V. Neuman, Dynamical Systems of Continuous Spectra, Proc. Nat. Acad. Sci. No. 18 278-286, 1932.

[3] O. Koopman, Hamiltonian Systems and Transformation in Hilbert Space, Proc. Nat. Acad. Sci. 1931 May; 17(5): 315–318.

[4] J. Goldstein, Semigroups of linear operators and applications, Oxford Univ. Press (1985).

[5] M. Korda and I. Mezic. On convergence of extended dynamic mode decomposition to the Koopman operator. Journal of Nonlinear Science, 28(2):687–710, 2018.

[6] Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. Nov. 2018.

[7] J.W. Neuberger, J.M. Neuberger, and James Swift, A linear condition determining global or local existence for nonlinear problems, Central European Journal of Mathematics, 11(8) (2013), 1361-1374.

[8] J. W. Neuberger, A complete theory for jointly continuous nonlinear semigroups on a complete separable metric space, J. Applicable Analysis 78 (2001), 223-231.

[9] A. Pazy, Semigroups of Linear Operators and Applications to Partial Differential Equations, Springer, New York, 1983. MR710486 (85g:47061).

[10] P. J. Schmid. Dynamic mode decomposition for numerical and experimental data. J. Fluid. Mech.

[11] H. Trotter, On the product of semigroups of operators. Proceedings of the American Mathematical Society, 10(4), 545-551.

[12] M. Thalhammer, High-order exponential operator splitting methods for time-dependent Schrödinger equations, SIAM J. Numer. Anal. 46 (2008), 2022–2038.

[13] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley. A data-driven approximation of the Koopman operator: extending dynamic mode decomposition. Journal of Nonlinear Science, 6:1307–1346, 2015.