CSCI670 Theoretical Thinking - I

18th September, 2021

Theoretical Thinking I assignment for CSCI670 Fall2021 | Prof. Teng Shang-Hua

> Init: 18th Sept, 2021 -> Scale Invariance
> Mod1: 2nd October, 2021 -> Governing equation identification using sparse regression

CSCI 670: Theoretical Thinking Assignment 1

Deadline: 9/29/2021, Wednesday

Research is about identifying open problems/directions or fundamental phenomena—matching one's own interests and insights—that may have impact or potential impact to the world. Now the Web is at our fingertip, we can gather information faster than ever before.

In this assignment, you are asked to write a two-page report on one paper in the field of theoretical computer science, particularly, concerning advanced algorithm design and analysis that you find interesting.

1. How did you find this paper?
2. Give a brief, but clear, discussion of the problem as well as the main result of the paper.
3. Briefly discuss the reason that you find this paper interesting and result(s) important.
4. State one major open question posted or inspired by the paper. What might be your initial idea to address this problem?

> 1. **Discovering governing equations from data by sparse identification of nonlinear dynamical systems**
> 2. **PySINDy: A Python package for the Sparse Identification of Nonlinear Dynamics from Data**
> 3. **Geometric Deep Learning Grids, Groups, Graphs, Geodesics, and Gauges**

Q1. How did you find this paper?

A1. I had ran into the PySINDy paper from this youtube video **Deep Learning of Dynamics and Coordinates with SINDy Autoencoders** on professor **Steve Brunton** youtube channel. I was(still am) subscribed to Prof. Steve's and had watched Prof. **Nathan Kutz's** explanations on PCA form the **AMATH301** channel, who is a co-author on this paper. I ran into the Geometric

Deep Learning paper on the **website of the same name** which I used to check once in a while since I like geometry and deep learning is something I don't feel comfortable with - so geometric deeplearning kinda balance my discomfort and interest. Earlier, the website contained a set of papers all related to geometric deep learning(I had looked into only one or two) but one fine day I found this one paper that put together the unified framework.

Q2. Give a brief, but clear, discussion of the problem as well as the main result of the paper.

A2. The paper basically deals in the domain of learning governing equations of systems from data collected about the system state over time. As such, this looks very similar to the domain of system identification from the control systems discipline. The identification of linear systems where the derivatives of the system state depend on the current state linearly is quite easy. For such a system, $\dot{x} = Ax$, where x is the state variable vector, $\dot{x}$ is the derivative of the state vector with respect to time and A captures the linear dynamics of the system since the derivative components in $\dot{x}$ only depend linearly on the components of the state variable x. It gets a bit challenging when the system dynamics is non-linear, i.e we cannot directly write each component of the vector $\dot{x}$ as some simple constant coefficient combination of the coefficients of the state vector x. This paper addresses the challenge of identifying non-linear system dynamics using data. Since there could be many complicated non-linear models that could explain the data (overfit), the authors are motivated by the observations in physics that most of the systems can be explained by very simple relationships between the physical quantities i.e the physical laws of nature are generally simple. Although the paper doesn't make any reference to **Occam's Razor**, they do use the term parsimonious models. The paper mentions that it is only very recently that sparse regression techniques have been applied to non-linear dynamics identification tasks - this is one of the main contribution of the paper - using sequentially thresholded least-squares regression for sparsity. The paper also points to another succeful method based on symbolic regression - **Discovering Symbolic Models from Deep Learning with Inductive Biases**.

To motivate the method proposed by the paper, I will take the help of two simple examples (not borrowed from the paper). Let us first see the mechanics of the identification of a linear system to better understand the non-linear identification later -

$$\begin{bmatrix} \dot{x_1} \\ \dot{x_2} \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Clearly, if we take a bunch of readings of the system evolving over time, we can estimate the matrix $A = \begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix}$. If we had just one reading of the system state (and also the derivatives, otherwise we need 2 points so that we can estimate the derivatives from the difference) at some time instant (t=0), we would have an underdetermined system with the equation -

$$\begin{bmatrix} \dot{x_1}(0) \\ \dot{x_2}(0) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix}$$

$$\Updownarrow$$

$$\begin{bmatrix}\dot{x_1}[0]\end{bmatrix} = \begin{bmatrix}a_{11} & a_{12}\end{bmatrix}\begin{bmatrix}x_1(0)\\x_2(0)\end{bmatrix} \; and \; \begin{bmatrix}\dot{x_2}(0)\end{bmatrix} = \begin{bmatrix}a_{22} & a_{22}\end{bmatrix}\begin{bmatrix}x_1(0)\\x_2(0)\end{bmatrix}$$

However, if we have measurements at two time points of the system state (and the derivatives), we will have a critically determined system.

$$\begin{bmatrix}\dot{x_1}(0) & \dot{x_1}(1)\\\dot{x_2}(0) & \dot{x_2}(1)\end{bmatrix} = \begin{bmatrix}a_{11} & a_{12}\\a_{21} & a_{22}\end{bmatrix}\begin{bmatrix}x_1(0) & x_1(1)\\x_2(0) & x_2(1)\end{bmatrix}$$

$$\Updownarrow$$

$$\begin{bmatrix}\dot{x_1}(0) & \dot{x_1}(1)\end{bmatrix} = \begin{bmatrix}a_{11} & a_{12}\end{bmatrix}\begin{bmatrix}x_1(0) & x_1(1)\\x_2(0) & x_2(1)\end{bmatrix}$$

$$and$$

$$\begin{bmatrix}\dot{x_2}(0) & \dot{x_2}(1)\end{bmatrix} = \begin{bmatrix}a_{21} & a_{22}\end{bmatrix}\begin{bmatrix}x_1(0) & x_1(1)\\x_2(0) & x_2(1)\end{bmatrix}$$

Now, in the critically determined system, the parameters $a_{11}, a_{12}, a_{21}, a_{22}$ are really sensitive to the noise in our measurements of x and $\dot{x}$ (or estimates of $\dot{x}$ from finite differences taken on x).

However if we take a lot of readings for times $t = 0, 1, 3, ..., n$, we will have an overdetermined system and the estimates of the parameters i.e the matrix A is not that sensitive to the errors in the individual measurements.

We will have the following equations in such a case -

$$\begin{bmatrix}\dot{x_1}(0) & \dot{x_1}(1) & ... & \dot{x_1}(n)\\\dot{x_2}(0) & \dot{x_2}(1) & ... & \dot{x_2}(n)\end{bmatrix} = \begin{bmatrix}a_{11} & a_{12}\\a_{21} & a_{22}\end{bmatrix}\begin{bmatrix}x_1(0) & x_1(1) & ... & x_1(n)\\x_2(0) & x_2(1) & ... & x_2(n)\end{bmatrix}$$

$$\Updownarrow$$

$$\begin{bmatrix}\dot{x_1}(0) & \dot{x_1}(1) & ... & \dot{x_1}(n)\end{bmatrix} = \begin{bmatrix}a_{11} & a_{12}\end{bmatrix}\begin{bmatrix}x_1(0) & x_1(1) & ... & x_1(n)\\x_2(0) & x_2(1) & ... & x_2(n)\end{bmatrix}$$

$$and$$

$$\begin{bmatrix}\dot{x_2}(0) & \dot{x_2}(1) & ... & \dot{x_2}(n)\end{bmatrix} = \begin{bmatrix}a_{21} & a_{22}\end{bmatrix}\begin{bmatrix}x_1(0) & x_1(1) & ... & x_1(n)\\x_2(0) & x_2(1) & ... & x_2(n)\end{bmatrix}$$

Using least squares, we can estimate the parameters of A from the last two equations.

Now let us look at a system with non-linear dynamics -

$$\dot{x_1} = x_2 \centerdot x_1 + sin(x_1) + 3cos(x_2) \; and \; \dot{x_2} = 3cos(x_1) + sin(x_2)$$

$$\Updownarrow$$

$$\begin{bmatrix} \dot{x_1} \\ \dot{x_2} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 3 \\ 0 & 0 & 1 & 3 & 0 \end{bmatrix} \begin{bmatrix} x_1 \cdot x_2 \\ sin(x_1) \\ sin(x_2) \\ cos(x_1) \\ cos(x_2) \end{bmatrix} = Af(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix})$$

$$where,\ f(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}) = \begin{bmatrix} x_1 \cdot x_2 \\ sin(x_1) \\ sin(x_2) \\ cos(x_1) \\ cos(x_2) \end{bmatrix}$$

Now, if we have a bunch of observations about the system state over time $t = 0, 1, 2..., n$, then

$$\dot{X} = Af(X)$$

$$\Updownarrow$$

$$\begin{bmatrix} \dot{x_1}(0) & \dot{x_1}(1) & ... & \dot{x_1}(n) \\ \dot{x_2}(0) & \dot{x_2}(1) & ... & \dot{x_2}(n) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & ... & a_{15} \\ a_{21} & a_{22} & ... & a_{25} \end{bmatrix} \begin{bmatrix} f(\begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix}) & ... & f(\begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix}) \end{bmatrix}$$

$$\Updownarrow$$

$$\begin{bmatrix} \dot{x_1}(0) & \dot{x_1}(1) & ... & \dot{x_1}(n) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & ... & a_{15} \end{bmatrix} \begin{bmatrix} f(\begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix}) & ... & f(\begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix}) \end{bmatrix}$$

$$and$$

$$\begin{bmatrix} \dot{x_2}(0) & \dot{x_2}(1) & ... & \dot{x_2}(n) \end{bmatrix} = \begin{bmatrix} a_{21} & a_{22} & ... & a_{25} \end{bmatrix} \begin{bmatrix} f(\begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix}) & ... & f(\begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix}) \end{bmatrix}$$

Here, we can see that we can estimate the matrix A if we know the funtion f. However, the function f is unkown to us. Hence an approach of guessing the functions called g, from a library of functions is taken. In the second paper about PySINDy, the included library of functions contains polynomials (uni and multivariate) and trigonometric functions with options for introducing custom functions into the library of functions.

Now, we can perform sparse regression to find the matrix A in the equation $\dot{X} = Ag(X)$. Since we are enforcing that the matrix A is sparse, we hope to recover the correct dynamics. Clearly the library of functions (basis functions) affect the extent to which we are able to find a good canditate for our model. There might be multiple solutions if we are not careful enought to choose the library with independent functions. Also, if we have some intuitions and partial prior knowledge of the system, it can be injected by admitting the possible forms of functuions we beleive might be governing the system into the library of functions.

The paper mentions that LASSO is too expensive if the data points are too large (and we actually need large set of data due to higher sampling to estimate the derivatives from the finite differences properly - in practice, we always have to do that since it is really expensive or infeasible to obtain the derivatives of the state directly). Hence they use the sequential thresholded least-squares regression method (which I have yet to look into). In the PySINDy paper, they have also implemented sparse relaxed regularized regression in addition to the sequential threshold least squares.

The results in the paper show that their method is able to recover many standard non-linear dynamic systems like the chaotic Lorenz System.

Q3. Briefly discuss the reason that you find this paper interesting and result(s) important.
A3. General - Since I keep interest in physics, I was just looking at some research at the intersection. I find such work interesting inherently (as long as my small brain can comprehend atleast 20-30%). I am always excited about connections between physics and other disciplines. I hope to be able to pick a few topics in physics like quantum mechanics and statistical physics at USC.

Specifically, the paper is interesting because - 1) I can only recall high school physics (to some extent) and this paper can be understood with only high school physics. 2) the approach is really simple and elegant, 3) they provide open source library to test their method which is very cool as we can build our own dynamic system and let the library figure what the governing equations are and when it actually finds close solutions, we can verify it and feel amazed/happy. 4) They have explicit and detailed equations and nice diagrams which makes the paper easy to read and understand. 5) The paper deals with the philosophy of knowledge discovery, i.e how we humans come up with models just by observing the surroundings and taking measurements. I could have also picked a few other papers but found them too challenging to complete with the available time I had for finishing this assignment. Below is a partial list of papers I had wanted to pick up for completing this assignment -

1. **Rao's Distance Measure, Colin Atkinson and Ann F. S. Mitchell**
2. **Generalized inverse of a matrix and its applications, C. Radhakrishna Rao, Sujit Kumar Mitra**
3. **A Resistive Circuits Analysis Using Graph Spectral Decomposition, Milos Dakovic, Ljubisa Stankovic, Budimir Lutovac, Ervin Sejdic, Tomislav B. Sekara**
4. **Eigenvectors from eigenvalues: A survey of a basic identity in linear algebra,Peter B. Denton, Stephen J. Parke, Terence Tao, and Xining Zhang**
5. **A Training Algorithm for Optimal Margin Classier, Bernhard E Boser, Isabelle M. Guyon, Vladmir N. Vapnik**
6. etc.

Q4. State one major open question posed or inspired by the paper. What might be your initial idea to address this problem?
A4. One major problem which the paper actually hints in the introduction is having a good frame of reference for the observed data. For example, as the paper already mentions, if one takes the position of two other planets as observed from Earth, there is probably no good sparse representation of the motion of the planets. But if we have position data of the Sun and we convert our frame of reference to that of the Sun, immediately the governing equations must get simplified and lead to sparse identification. A challenge is to actually find a good coordinate transform from which we can then run SINDy to get the simple governing equations. Their later paper **Data-driven discovery of coordinates and governing equations** accomplises this using SINDy + Autoencoders.

I think my approach might have been a little bit different initially, although could have converged to the autoencoder method. I would probably have started with very similar approach to the guessing game we played above. If we think about it, how did humans figure out that we should think of Sun as the frame of reference and not the earth - well it took us so long to figure that out. There are stories of many mathematicians and astronomers who used epicycles to describe the complicated motion of the planets as observed from earth. Had the

Sun be not glowing so brightly unlike the planets, one might have had to try all the planets as references to maybe find one of them to be a good frame of reference that simplified the governing equations for the motions of the planets. So, I would have started with some library of candidate coordinate transformations (frames of reference) that seem possible and then try to find a model where we can simultaneuosly find sparsity in the dynamics matrix A and the right frame of reference. the easisest would be to bruteforce all the candidates from the frame of reference library and then measuer the sparsity using some norm of the matrix A. This method might or might not have worked and then I would have probably tried a few other things like Principal Components Analysis (just thinking loudly) before actually moving into neural nets and auto encoders.

Something I guess can be extended is using dimension analysis to narrow down the possibilities further. I have read about linear algebra methods to find relations between physical quantities(upto proportionality) in a linear algebra book (**this one, page 165**). There might be a way to relate dimension analysis to general discovery of physics equations in some setting.

Idempotent and scale separation invariant operators (disparate thoughts induced from paper [3])

> Max is sometimes linear
>
> $$max\{a+K, b+K, c+K, ...\} = max\{a, b, c, ...\} + K$$
>
> $$max\{\gamma a, \gamma b, \gamma c, ...\} = \gamma \times max\{a, b, c, ...\} \forall \gamma \in \mathbb{R}^{+}$$

> Composition of maximum is a maximum, maximum is idempotent
>
> $$max\{max\{max.....max\{a, b, c, ...\}\}\} = max\{a, b, c...\}$$

Max can be applied on finer scales and then reapplied on the outcomes(coarser) of their applications on finer scales (See **Geometric DL-3.4**).

Another function that comes to mind is the center of mass function. Center of mass of a system can be calculated by first applying it to subsystems of the system and then applying again to subsytem COM - hence the fining-coarsing (should it be called scale separation?) of COM function (it outputs both total mass of the system and the COM coordinates). Similarly, the max pooling applied on finer segments/grids and then reapplied on the output of the already max pooled finer grids still gives the same result (classification class here) - so prediction class is invariant to fining-coarsing of the max-pooling operation

$$max\{\alpha_1, \alpha_2, \alpha_3..., \beta_i, \beta_2, \beta_3, ..., \gamma_1, \gamma_2, \gamma_3... \}$$

$$||$$

$$max\{max\{\alpha_i | i \in 1, 2, ...\}, max\{\beta_i | i \in 1, 2, ...\}, max\{\gamma_i | i \in 1, 2, ...\}\}$$

Let $S = \{m_{a1}, m_{a2}, ..., m_{b1}, m_{b2}, ...\}$ be a set of particle masses from a system in 1D. hence, $S = \{m_{ai} | i = 1, 2, ...\} \cup \{m_{bi} | i = 1, 2, ...\} = S_a \cup S_b$. Let the position of m_z be p_z and $P = \{p_{a1}, p_{a2}, ..., p_{b1}, p_{b2}, ...\} = P_a + P_b$. Let Y be a tuple representation of the system such that $Y = \{\langle m_z, p_z \rangle | m_z \in S, p_z \in P\}$.

$$Y = \{\langle m_{a1}, p_{a1} \rangle, \langle m_{a2}, p_{a2} \rangle, ... \langle m_{b1}, p_{b1} \rangle, \langle m_{b2}, p_{b2} \rangle, \} = Y_a + Y_b$$

If Q is tuple representation of a system, define

$$C(Q) = \left\langle \sum_{m_z} m_z, \frac{\sum m_z p_z}{\sum_{m_z}} \right\rangle \; where \; (m_z, p_z) \in Q$$

Hence C is a multivariable function which takes in a tuple representation of a sytem and returns the total mass and the center of mass of the system.Given this definition, the function C behaves just like the function max when we are thinking in terms of fining-coarsing -

$$C(Y_a) = \langle M_a, P_a \rangle$$
$$C(Y_b) = \langle M_a, P_a \rangle$$

$$C(Y) = C(\ \langle M_a, P_a \rangle, \langle M_b, P_b \rangle\ )$$

$$\implies C(Y) = C(\ C(Y_a), C(Y_b)\ )$$

Hence C is comatible with the fining-coarsing operation and under this operation, the lumped tuple representation of the system $\langle M, P \rangle$ of the system is invariant to how the fining and coarsing is done (i.e details of set partitioning for fining is immaterial)

> Fractional part operator $\{x\} = x - \lfloor x \rfloor$ also has a similar property over addition (not over sets).

$$A = x_1 + x_2 + ... + x_n + y_1 + y_2 + ... + y_m$$
$$A = X + Y$$
$$\implies \{A\} = \{X + Y\} = \{\{X\} + \{Y\}\}$$
$$s.t\ \{X\} = \{\{x_1\} + \{x_2\} + ... + \{x_n\}\}$$
$$and\ \{Y\} = \{\{y_1\} + \{y_2\} + ... + \{y_m\}\}$$

> Same goes for the modulo operator %.

$$
\begin{aligned}
A &= x_1 + x_2 + ... + x_n + y_1 + y_2 + ... + y_m \\
A &= X + Y \\
\implies A\%q &= (X + Y)\%q = ((X\%q) + (Y\%q))\%q \\
s.t\ X\%q &= (x_1\%q + x_2\%q + ... + x_n\%q)\%q \\
and\ Y\%q &= (y_1\%q + y_2\%q + ... + y_m\%q)\%q
\end{aligned}
$$

.