Recent Advances in Learning from High-Dim Dynamical Systems

Ming Zhong

Department of Mathematics
University of Houston

April 14, 2025

High Dim Data Analaysis Group

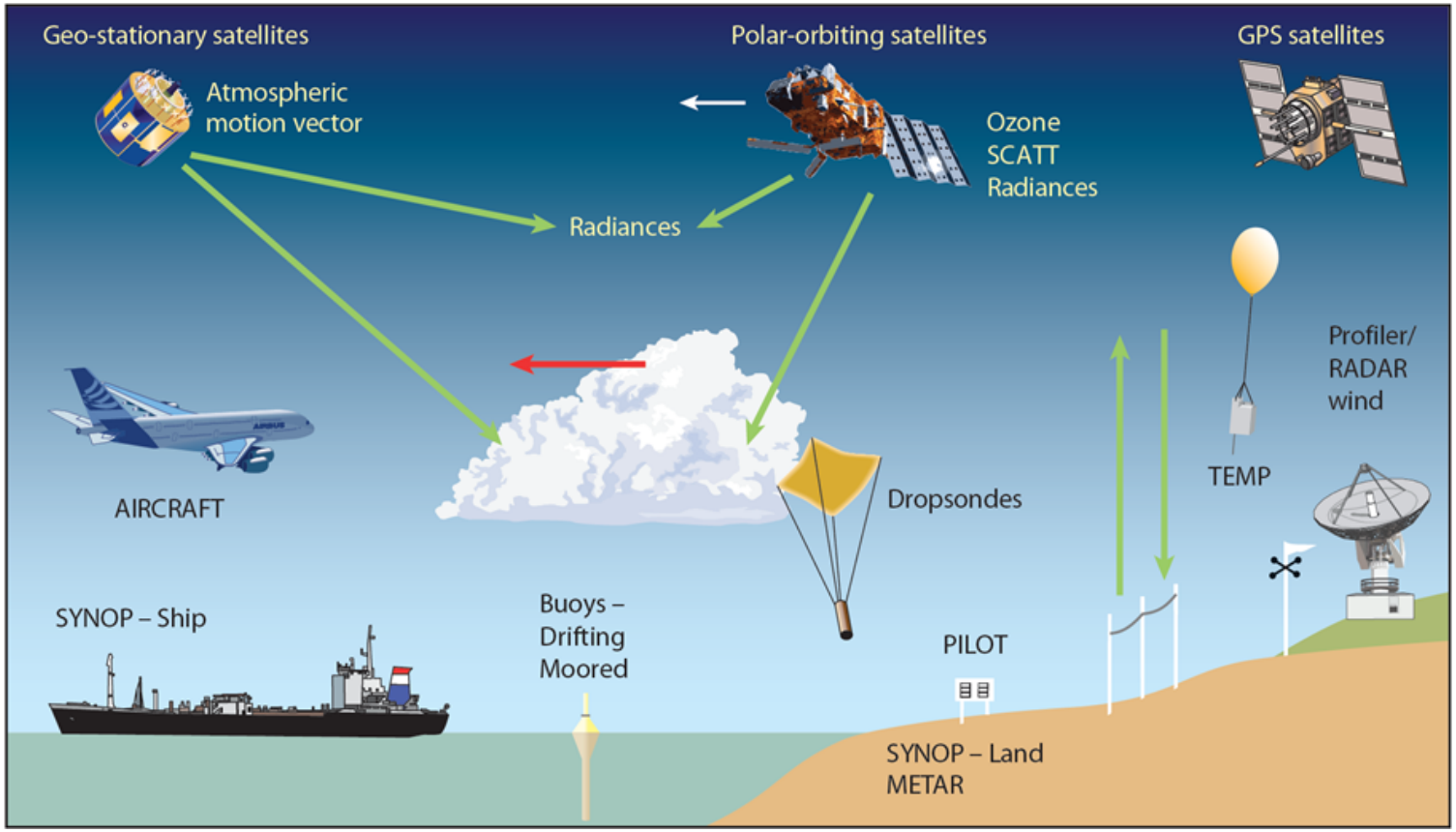
Table of Contents

Time Series Data

Important Applications

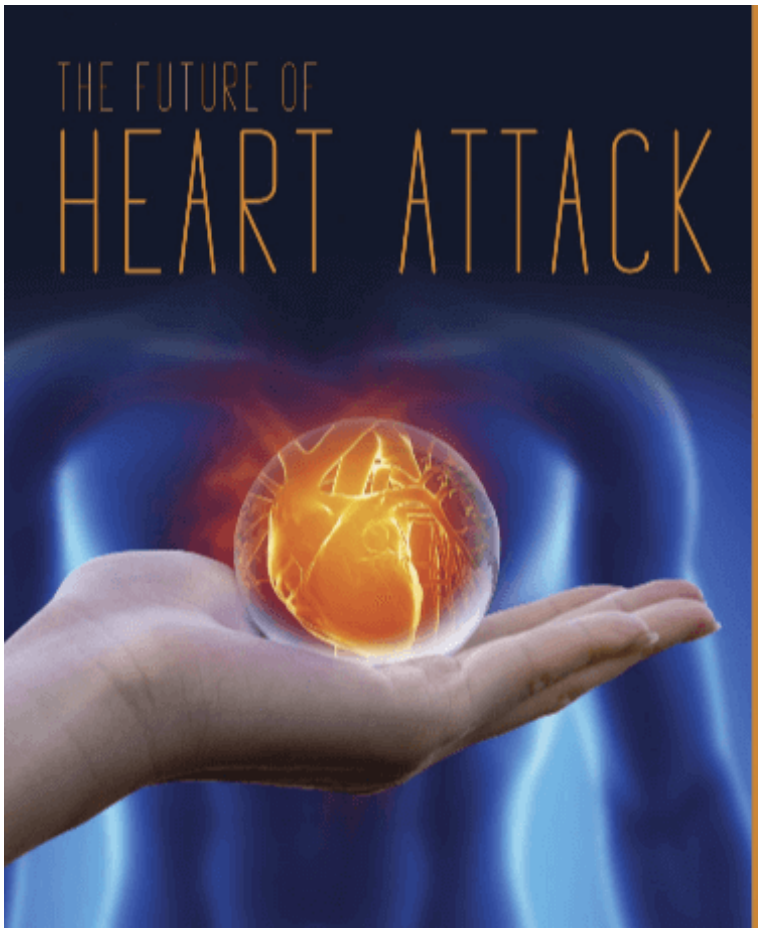
Time Series Data

Important Applications



Time Series Data

Important Applications



PREDICTION

Technology and medical advances are helping doctors foresee cardiac events earlier and save more lives. By Nicole Foster

Science is constantly marching forward, and modern technologies mean that march toward the future is accelerating all of the time. Fortunately, some doctors and researchers are putting that forward momentum to good use when it comes to heart attack prediction. These scientists are working tirelessly to find ways to predict heart attacks years before any symptoms arise — because early prediction means early intervention, and early intervention can save lives.

Tried-and-True Standards

Physicians have been trying to predict heart attacks for as long as there have been heart attacks. Traditionally, they have relied on standard assessments of cholesterol, blood pressure, lifestyle factors and health conditions such as diabetes to predict whether a patient is likely to suffer a heart attack.

Age, lipid levels, obesity, lack of activity and stress can all contribute to blocked arteries, preventing blood flow to the heart, says Dr. Ashed Qayyumi, professor at Emory University School of Medicine. By gathering data, test results and patient information, cardiologists like Qayyumi can generate a score that indicates a patient's heart attack risk.

"It's not an exact prediction," says Qayyumi. "We use the scores to reduce risk and to prevent disease, heart attack or sudden cardiac death. They're now developing new technologies which might get us to better predict [cardiovascular disease]."

Some of those technologies are even coming from corporations. Apple partnered with Stanford Medicine to use data from Apple Watches to power a massive heart study. Fitbit, TomTom, Garmin and countless other wearable technology companies build heart-monitoring tools into their products. Toyota is working with researchers at the University of Michigan to study how their cars can detect when drivers are having a cardiac event; initial results of that study are expected in 2020.

Meanwhile, cardiologists are looking to artificial intelligence (AI), retinal scans and qualitative research for new, predictive strategies.

The Synergizing Epidemiologist

Dr. Stephen Weng is an assistant professor at University of Nottingham and a research fellow for the National Institute for Health Research. He and his team have developed an algorithm that marries artificial intelligence and face-to-face time with your doctor. His system can

> "In the sample size of 83,000 records... an additional 355 lives could have been saved using the AI compared to standard prediction methods."

Time Series Data

Prediction

Given data in the form

$$\{\mathbf{z}_{t_1}, \mathbf{z}_{t_2}, \cdots, \mathbf{z}_{t_L}\}, \quad \mathbf{z}_{t_l} \in \mathbb{R}^D \, (D \gg 1),$$

with $0 = t_1 < t_2 < \cdots < t_L = T$, can we predict

$$\mathbf{z}_{t_{L+1}}, \mathbf{z}_{t_{L+2}}, \cdots, \quad t_{L+1} = T + \delta t$$

Traditional Methods

- Autoregressive (AR), moving average (MA), ARMA and ARIMA models

Machine Learning methods

- Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), Transformers

Time Series Data

Dynamical Data

Basic assumption (Markov Chain)

$$\begin{aligned}\Pr(\mathbf{z}_{L+1} = z_{L+1} \,|\, \mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2, \cdots, \mathbf{z}_L = z_L) \\ = \Pr(\mathbf{z}_{L+1} = z_{L+1} \,|\, \mathbf{z}_L = z_L)\end{aligned}$$

Even further (dynamic structure)

$$\mathbf{z}_{L+1} = \mathbf{z}_L + \mathbf{h}(\mathbf{z}_L)\Delta t + \boldsymbol{\sigma}(\mathbf{z}_L)\Delta \mathbf{w}_L,$$

where

- $\mathbf{h} : \mathbb{R}^D \to \mathbb{R}^D$; $\boldsymbol{\sigma} : \mathbb{R}^D \to \mathbb{R}^{D\times D}$.
- $\Delta \mathbf{w}_L \in \mathbb{R}^D$ is a Brownian motion.

Time Series Data

Dynamics Assumption

Continuously

$$\mathrm{d}\mathbf{z}_t = \mathbf{h}(\mathbf{z}_t)\,\mathrm{d}t + \boldsymbol{\sigma}(\mathbf{z}_t)\,\mathrm{d}\mathbf{w}_t,$$

Learning and Prediction

Given data $\{\mathbf{z}_{t_1}, \mathbf{z}_{t_2}, \cdots, \mathbf{z}_{t_L}\}$, can we predict $\mathbf{z}_{L+1}$, assuming

$$\mathrm{d}\mathbf{z}_t = \mathbf{h}(\mathbf{z}_t)\,\mathrm{d}t + \boldsymbol{\sigma}(\mathbf{z}_t)\,\mathrm{d}\mathbf{w}_t,$$

with h and σ being unknown.

Established approaches

- SINDy, Neural ODE, PINN/PIGP.

Time Series Data

Dynamics Assumption

These methods use the normal regression setup

$$\mathcal{E}(\tilde{\mathbf{h}}) = \mathbb{E}\Big[\frac{1}{T}\int_0^T \Big|\frac{\mathrm{d}\mathbf{z}_t}{\mathrm{d}t} - \tilde{\mathbf{h}}(\mathbf{z}_t)\Big|^2_{\mathbb{R}^D}\, \mathrm{d}t.\Big].$$

Remark:

- $\frac{\mathrm{d}\mathbf{z}_t}{\mathrm{d}t}$ is understood loossely.
- $\hat{\mathbf{h}} = \operatorname{argmin}_{\tilde{\mathbf{h}}\in\mathcal{H}} \mathcal{E}(\tilde{\mathbf{h}})$ for some space $\mathcal{H}$.

Since $D \gg 1$,

- SINDy assumes $\mathbf{h}(\mathbf{z}) \approx \sum_{\eta=1}^{n} \psi_n(\mathbf{z})$ with sparse coefficients, and pre-defined dictionary ψ_n.
- Neural ODE/PINN use deep neural network with over-parametrization and stochastic gradient descent.

Time Series Data

Dynamics Assumption

However

- For SIDNy: the dicionary is chosen to be large enough to contain enough assumptions (non-linearity, derivatives).
  - Sparsity is enforced to reduce the search time.
  - Derivatives can be weaken using weak-SINDy.
  - But "large" is very subjective.
- Deep learning
  - Estimators might have good prediction capability.
  - Estimators do not have clear physical structures (i.e. Gravity is $1/r^2$).
  - Training can be an issue; Setup of the hyperparameters are bit engineering-like.

Table of Contents

Dynmical Times Series Data

Our Approach

Recall $\mathrm{d}\mathbf{z}_t = \mathbf{h}(\mathbf{z}_t)\,\mathrm{d}t + \boldsymbol{\sigma}(\mathbf{z}_t)\,\mathrm{d}\mathbf{w}_t$, we consider

- $\boldsymbol{\sigma} = 0 * \mathbf{I}_{D\times D}$, determinstic system, $\mathbf{h}$ is a high-dim function.
- $\boldsymbol{\sigma} = 0 * \mathbf{I}_{D\times D}$, determinstic system, $\mathbf{h}$ is a differential operator.
- $\boldsymbol{\sigma} = \boldsymbol{\sigma}(\mathbf{z})$ is a SPD matrix, $\mathbf{h}$ is a high-dim function or a differential operator.
- $\boldsymbol{\sigma} = \boldsymbol{\sigma}(\mathbf{z})$ is a singular matrix, $\mathbf{h}$ is a high-dim function.

Learning and Prediction

We build different estimating algorithms when both $\mathbf{h}$ and $\boldsymbol{\sigma}$ are unknown based on the scenarios, especially on how to handle the curse of dimensionality.

Dynmical Times Series Data

Case 1

Interacting agent systems for $\mathbf{x}_i \in \mathbb{R}^d$,

$$\frac{\mathrm{d}\mathbf{x}_i}{\mathrm{d}t} = \frac{1}{N}\sum_{j=1, j\neq i}^{N} \phi(\left|\mathbf{x}_j - \mathbf{x}_i\right|)(\mathbf{x}_j - \mathbf{x}_i), \quad i = 1, \cdots, N,$$

With the setup

$$\mathbf{z} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{bmatrix}, \ \mathbf{h}_\phi(\mathbf{z}) = \begin{bmatrix} \vdots \\ \frac{1}{N}\sum_{j=1, j\neq i}^{N} \phi(\left|\mathbf{x}_j - \mathbf{x}_i\right|)(\mathbf{x}_j - \mathbf{x}_i) \\ \vdots \end{bmatrix}.$$

Then $\dot{\mathbf{z}} = \mathbf{h}_\phi(\mathbf{z})$ with $\mathbf{z} \in \mathbb{R}^D$ with $D = Nd$.

Dynmical Times Series Data

Case 1

Instead of learning $\mathbf{h}_\phi$, we learn ϕ instead from

$$\mathcal{E}(\varphi) = \mathbb{E}\Big[\frac{1}{T}\int_0^T \big|\dot{\mathbf{z}}(t) - \mathbf{h}_\phi(\mathbf{z}(t))\big|_N^2 \, \mathrm{d}t\Big].$$

- $\hat{\phi} = \operatorname{argmin}_{\varphi\in\mathcal{H}} \mathcal{E}(\varphi) \approx \phi$.
- 1D leanring rate.
- $\mathbf{h}_\phi$ has a structure and phisycal meaning (interaction).
- When $\mathcal{H}$ is finite dimensional, the learning problem becomes solving a linear system.
- Review: Learning Collective Behaviors from Observation, Explorations in the Mathematics of Data Science, 2024.

Dynmical Times Series Data

Case 2

Next we consider

$$\dot{\mathbf{z}} = \mathbf{h}(\mathbf{z}) = \mathcal{P}(\mathbf{z}), \quad \mathcal{P} \text{ is a differential operator.}$$

For example, compressible Euler system

$$\frac{\partial}{\partial t}\begin{bmatrix}\rho\\ \mathbf{u}\\ e\end{bmatrix} + \begin{bmatrix}\mathbf{u}\cdot\nabla\rho\\ \mathbf{u}\cdot\nabla\mathbf{u}\\ \mathbf{u}\cdot\nabla e\end{bmatrix} + \begin{bmatrix}\rho\nabla\cdot\mathbf{u}\\ \nabla p/\rho\\ p/\rho\nabla\cdot\mathbf{u}\end{bmatrix} = \begin{bmatrix}0\\ \mathbf{G}\\ 0\end{bmatrix}.$$

- $\rho(t,\mathbf{x})$ (density), $\mathbf{u}(t,\mathbf{x})$ (fluid velocity), $e(t,\mathbf{x})$ (specific internal energy), $p(t,\mathbf{x})$ (presure), $\mathbf{G}(t,\mathbf{x})$ (gravitational acceleration).

Dynmical Times Series Data

Case 2

Consider for $(t, \mathbf{x}) \in \Omega$

$$\dot{\mathbf{z}} = \mathbf{h}(\mathbf{z}) = \mathcal{P}(\mathbf{z})$$
$$\text{subject to } \mathbf{z}(0, \mathbf{x}) = \mathbf{z}_0(\mathbf{x}) \text{ and } \mathcal{B}(\mathbf{z}) = \mathbf{g}(t, \mathbf{x}) \text{ on } \partial\Omega.$$

Learn $\mathbf{z}(t, \mathbf{x})$ for $(t, \mathbf{x}) \in \Omega$ through

$$\text{Loss} = \text{Data Loss} + \lambda * \text{Physics Loss}.$$

where

$$\text{Physics Loss} = \text{PDE/ODE/Integral Loss} + \text{IC Loss} + \text{BC Loss}.$$

Regularized Leanring (Supervised (Data) + Unsupervised (Physics)).

Dynmical Times Series Data

Case 2

Continue on

$$\text{PDE Loss} = \frac{1}{|\Omega|}\int_{\Omega} \left|\frac{\partial \mathbf{z}_{NN}}{\partial t} - \mathcal{P}(\mathbf{z}_{NN})\right|^2_{L^2(\Omega)} \, \mathrm{d}\Omega.$$

and

$$\text{IC Loss} = \frac{1}{|\Omega_x|}\int_{\Omega_x} \left|\mathbf{z}_{NN} - \mathbf{z}_0\right|^2_{L^2(\Omega_x)} \, \mathrm{d}\Omega_x.$$

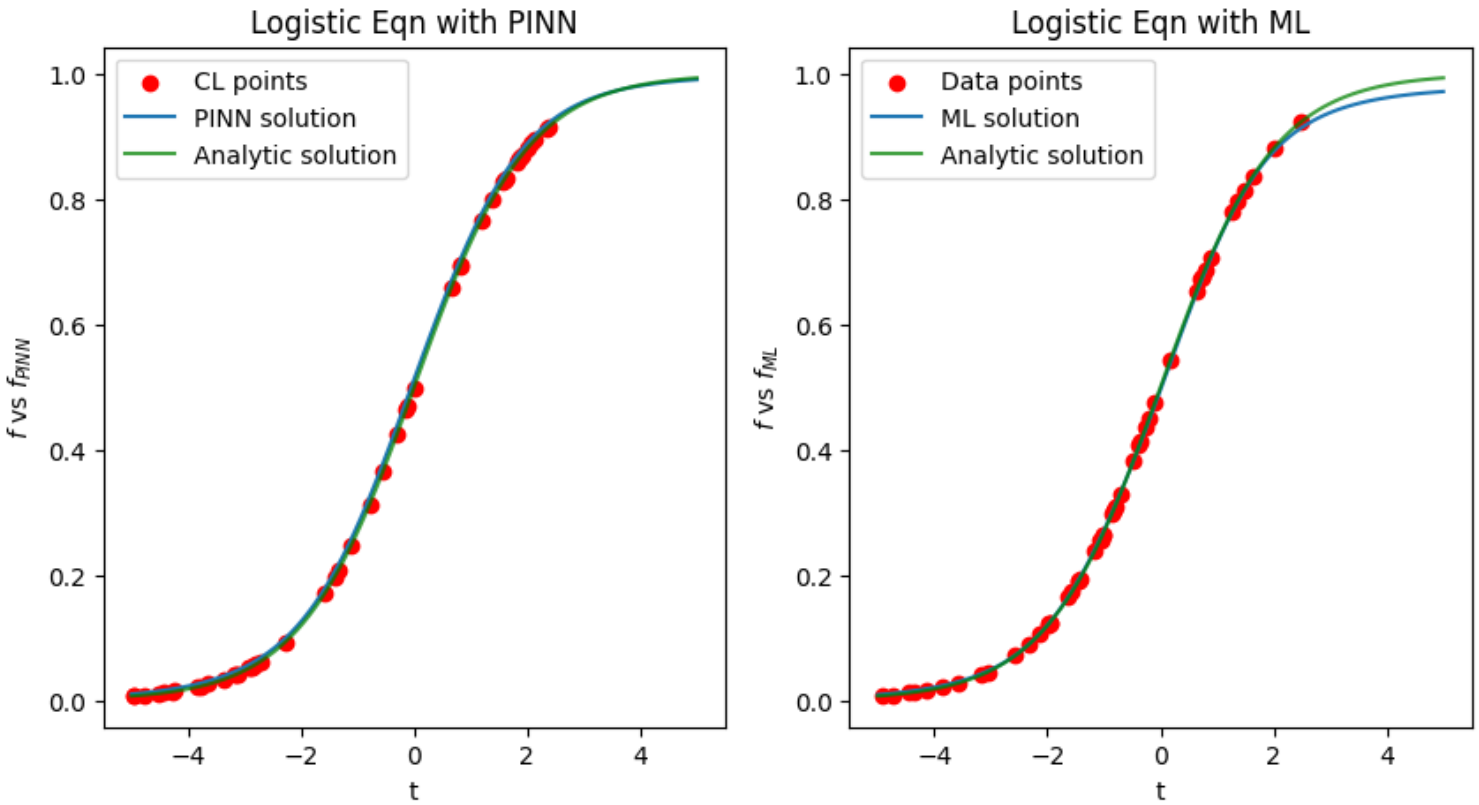
lastly

$$\text{BC Loss} = \frac{1}{|\partial\Omega|}\int_{\partial\Omega} \left|\mathcal{B}(\mathbf{z}_{NN}) - \mathbf{g}\right|^2_{L^2(\partial\Omega)} \, \mathrm{d}\partial\Omega.$$

The addiitional information from Physics improves prediction capability.

Dynmical Times Series Data

Case 2



Dynmical Times Series Data

Case 2

Remarks:

- With additional data in Ω, one can also infer $\mathcal{P}$ (or some parametric structure of $\mathcal{P}$).
- IC loss can also considered as a type of data loss (initial data).
- BC loss can be considered as data loss if BC type is Dirichlet.
- Many types of losses lead to multi-objective optimization.
  - PDE loss has larger (in magnitude) gradient, forcing the minimizer to minimize PDE first, ignoring IC and BC, difficult at solving stiff PDEs.
  - In the case of Hyperbolic PDEs (capturing shocks), how to capture the physically meaningful weak solution?
  - Classic formulation and L^2-norm are used, leadning to other problems.
  - Meshless loss leads to loss of time flow.

Dynmical Times Series Data

Case 2

Some remedies

- Adding data-driven artificial viscosity map to Hyperbolic PDEs: Physics-informed neural networks with adaptive localized artificial viscosity, JCP, 2023.
- Adding hard constrain or building BC into the architecture: Structure Preserving PINN for Solving Time Dependent PDEs with Periodic Boundary, arXiv, 2024.
- Mixture of Experts (MOE), label propagation: Label Propagation Training Schemes for Physics-Informed Neural Networks and Gaussian Processes, arXiv, 2024.

Physics Informed Transformer with Michael Holland on 4/28.

Dynmical Times Series Data

Case 3

Back to the stochastic case

$$\mathrm{d}\mathbf{z}_t = \mathbf{h}(\mathbf{z}_t)\,\mathrm{d}t + \boldsymbol{\sigma}(\mathbf{z}_t)\,\mathrm{d}\mathbf{w}_t, \quad \mathbf{h} \text{ and } \boldsymbol{\sigma} \text{ are unknown.}$$

But we assume $\boldsymbol{\sigma} : \mathbb{R}^D \to \mathbb{R}^{D\times D}$ is Symmetric Positive Definite for all $\mathbf{z}$.

- It is possible to learn $\mathbf{h}$ in the regression setting, since $\mathrm{d}\mathbf{z}_t \approx \mathbf{h}(\mathbf{z}_t)\,\mathrm{d}t$, we can

$$\mathbb{E}\left[|\,\mathrm{d}\mathbf{z}_t - \mathbf{h}(\mathbf{z}_t)\,\mathrm{d}t|_{\ell_2}\right].$$

However it does not use any of the information from the noise matrix, and misses the interaction of the noise and the drift.

Dynmical Times Series Data

Case 3

Hence, we design the loss function as

$$\mathcal{E}(\tilde{\mathbf{h}}) = \mathbb{E}\Big[\frac{1}{2}\Big(\int_0^T < \tilde{\mathbf{h}}(\mathbf{z}_t), \Sigma^\dagger \tilde{\mathbf{h}}(\mathbf{z}_t) > \mathrm{d}t - 2 < \tilde{\mathbf{h}}(\mathbf{z}_t), \Sigma^\dagger \, \mathrm{d}\mathbf{z}_t > \Big)\Big].$$

where

- $\Sigma = \boldsymbol{\sigma}\boldsymbol{\sigma}^\top$ and $\Sigma^\dagger$ is the pseudo-inverse of Σ (in this case, just the normal inverse).
- $\hat{\mathbf{h}} = \mathrm{argmin}_{\tilde{\mathbf{h}} \in \mathcal{H}} \mathcal{E}(\tilde{\mathbf{h}}) \approx \mathbf{h}$.
- One can think of this is almost

$$|\, \mathrm{d}\mathbf{z}_t - \mathbf{h}(\mathbf{z}_t)\, \mathrm{d}t|_{\boldsymbol{\sigma}^\dagger}$$

Taking into consideration of the effect of the noise.

Dynmical Times Series Data

Case 3

When $\Sigma = \sigma * \mathbf{I}_{D\times D}$, the loss simplfies to

$$\mathcal{E}(\tilde{\mathbf{h}}) = \mathbb{E}\Big[\frac{1}{2\sigma^2}\Big(\int_0^T < \tilde{\mathbf{h}}(\mathbf{z}_t), \tilde{\mathbf{h}}(\mathbf{z}_t) > \mathrm{d}t - 2 < \tilde{\mathbf{h}}(\mathbf{z}_t), \mathrm{d}\mathbf{z}_t >\Big)\Big].$$

Then it is very similar to regression.

- When $\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_D \end{bmatrix}$, the components of $\mathbf{h}$ de-couples, so regression on each component of $\mathbf{h}$ is comparable.
- Our method is more effective when $\boldsymbol{\sigma}$ is not-diagonal.

Dynmical Times Series Data

Case 3

We have two papers

- Learning Stochastic Dynamics from Data, Guo, Cialenco, Zhong, ICLR, 2024.
- Noise Guided Structural Learning from Observing Stochastic Dynamics, Guo, Cialenco, Zhong, revision, 2025.
- Noise is learned through a modified Qudratic Variation method.

Henry Guo will discuss more about this method on 4/28 on noise interaction and Stochastic Heat Equation.

Dynmical Times Series Data

Case 4

Lately, we have

$$\mathrm{d}\mathbf{z}_t = \mathbf{h}(\mathbf{z}_t)\,\mathrm{d}t + \boldsymbol{\sigma}(\mathbf{z}_t)\,\mathrm{d}\mathbf{w}_t, \quad \boldsymbol{\sigma} \text{ is singular.}$$

WLOG, we assume

$$\boldsymbol{\sigma} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\sigma}_y \end{bmatrix},$$

Remark: if not, we juse do SVD, and rotate $\mathbf{z}_t$ accordingly. Hence

$$\begin{aligned} \mathrm{d}\mathbf{x}_t &= \mathbf{f}(\mathbf{x}_t, \mathbf{y}_t)\,\mathrm{d}t, \\ \mathrm{d}\mathbf{y}_t &= \mathbf{g}(\mathbf{x}_t, \mathbf{y}_t)\,\mathrm{d}t + \boldsymbol{\sigma}_y(\mathbf{y}_t)\,\mathrm{d}\mathbf{w}_t^y. \end{aligned}$$

Here $\boldsymbol{\sigma}_y$ is SPD.

Dynmical Times Series Data

Case 4

When we set

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \quad \mathbf{h}(\mathbf{z}) = \begin{bmatrix} \mathbf{f}(\mathbf{x}, \mathbf{y}) \\ \mathbf{g}(\mathbf{x}, \mathbf{y}) \end{bmatrix}$$

and

$$\boldsymbol{\sigma} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\sigma}_y \end{bmatrix},$$

It goes back to the original high-dim SDE but $\boldsymbol{\sigma}$ is singular if the normal loss

$$\mathcal{E}(\tilde{\mathbf{h}}) = \mathbb{E}\big[\frac{1}{2\sigma^2}\big(\int_0^T < \tilde{\mathbf{h}}(\mathbf{z}_t), \tilde{\mathbf{h}}(\mathbf{z}_t) > \mathrm{d}t - 2 < \tilde{\mathbf{h}}(\mathbf{z}_t), \mathrm{d}\mathbf{z}_t > \big)\big].$$

is used, the learning will collapse down to only learning $\mathbf{g}$.

Dynmical Times Series Data

Case 4

Hence, we treat them separately

$$\mathcal{E}_f(\tilde{\mathbf{f}}) = \mathbb{E}\big[\frac{1}{T}\int_0^T \big|\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} - \tilde{\mathbf{f}}(\mathbf{x}_t, \mathbf{y}_t)\big|_{\ell_2}^2\big].$$

and

$$\mathcal{E}_g(\tilde{\mathbf{g}}) = \mathbb{E}\big[\frac{1}{2}\big(\int_0^t < \tilde{\mathbf{g}}(\mathbf{x}_t, \mathbf{y}_t), \Sigma_y^\dagger \tilde{\mathbf{g}}(\mathbf{x}_t, \mathbf{y}_t) > \mathrm{d}t - 2 < \tilde{\mathbf{g}}(\mathbf{x}_t, \mathbf{y}_t), \Sigma_y^\dagger \,\mathrm{d}\mathbf{y}_t > \big)\big].$$

We will learn $\boldsymbol{\sigma}_y$ from $\mathbf{y}_t$.

Dynmical Times Series Data

Case 4

A unified algorithm

- Given $\{\mathbf{z}_t\}_{t\in[0,T]}$, we use quadratic variation on $\mathbf{z}_t$ to figure out $\boldsymbol{\sigma}$.
- If $\boldsymbol{\sigma} = 0$, we learn it deterministically.
- If $\boldsymbol{\sigma}$ is non-singular, we learn it stochasticaly
- If $\boldsymbol{\sigma}$ is singular, we perform SVD and find out the x and y direction, then we learn it mixed.

Henry will also discuss that on 4/28.