2022 TRIPODS Winter School Workshop

Yinling Zhang

Lecture 1

Universal Approximation of Sequence-to-Sequence Transformations by Temporal Convolutional Nets

Background: A wide variety of problems in machine learning involve sequence-to-sequence transformations. In the past time such maps have been modeled using discrete-time recurrent neural nets. However, There has been a recent shift in sequence-to-sequence modeling from recurrent network architectures to autoregressive convolutional network architectures.
Main point: Discuss TCNs from the perspective of universal approximation for causal and time-invariant input-output maps that have approximately finite memory.
Input-output maps and approximately finite memory
Let $\mathcal{S}$ denote the set of all real-valued sequences $\boldsymbol{u} = (u_t)_{t\in\mathbb{Z}_+}$, where $\mathbb{Z}_+ := \{0, 1, 2, \ldots\}$. An input-output map (or i/o map, for short) is a nonlinear operator $\mathrm{F} : \mathcal{S} \to \mathcal{S}$ that maps an input sequence $\boldsymbol{u} \in \mathcal{S}$ to an output sequence $\boldsymbol{y} = \mathrm{F}\boldsymbol{u} \in \mathcal{S}$. Denote the application and the composition of i/o maps by concatenation.
An i/o map F has approximately finite memory on a set of inputs $\mathcal{M} \subseteq \mathcal{S}$ if for any $\varepsilon > 0$ there exists $m \in \mathbb{Z}_+$, such that

$$\sup_{\boldsymbol{u}\in\mathcal{M}} \sup_{t\in\mathbb{Z}_+} \left|(\mathrm{F}\boldsymbol{u})_t - \left(\mathrm{F}\mathrm{W}_{t,m}\boldsymbol{u}\right)_t\right| \le \varepsilon$$

where $\mathrm{W}_{t,m} : \mathcal{S} \to \mathcal{S}$ is the windowing operator $\left(\mathrm{W}_{t,m}\boldsymbol{u}\right)_\tau := u_\tau \mathbf{1}_{\{\max\{t-m,0\}<\tau<t\}}$.
The universal approximation theorem
Any causal and time-invariant i/o map that has approximately finite memory and satisfies an additional continuity condition can be approximated arbitrarily well by a ReLU temporal convolutional net. Consider i/o maps with uniformly bounded inputs, i.e., inputs in the set

$$\mathcal{M}(R) := \left\{ \boldsymbol{u} \in \mathcal{S} : \|\boldsymbol{u}\|_\infty := \sup_{t\in\mathbb{Z}_+} |u_t| \le R \right\} \quad \text{for some } R > 0$$

For any $t \in \mathbb{Z}_+$and any $\boldsymbol{u} \in \mathcal{N}(R)$, the finite subsequence $\boldsymbol{u}_{0:t} = (u_0, \ldots, u_t)$ is an element of the cube $[-R, R]^{t+1} \subset \mathbb{R}^{t+1}$; conversely, any vector $\boldsymbol{x} \in [-R, R]^{t+1}$ can be embedded into $\mathcal{M}(R)$ by setting $u_s = x_s \mathbf{1}_{\{0\le s\le t\}}$. To any causal and time-invariant i/o map F we can associate the nonlinear functional $\tilde{\mathrm{F}}_t : \mathbb{R}^{t+1} \to \mathbb{R}$ defined in the obvious way: for any $\boldsymbol{x} = (x_0, x_1, \ldots, x_t) \in \mathbb{R}^{t+1}$,

$$\tilde{\mathrm{F}}_t(\boldsymbol{x}) := (\mathrm{F}\boldsymbol{u})_t,$$

where $\boldsymbol{u} \in \mathcal{S}$ is any input such that $u_s = x_s$ for $s \in \{0, 1, \ldots, t\}$ (the values of u_s for $s > t$ can be arbitrary by causality).

Lecture 2

Policy Gradient Descent for Control: Global Optimality via Convex Parameterization

Part 1: First, examine the convergence and optimality of these methods for the infinite-horizon Linear Quadratic Regulator (LQR), where despite nonconvexity (with respect to policy parameters), gradient descent converges to the optimal policy under mild assumptions.
Part 2: Make a connection between classical convex parameterizations in control theory on one hand, and the gradient dominance property of the nonconvex cost function, on the other.

Lecture 3

Control-Theoretic Tools in Analysis and Synthesis of Neural Network Driven Systems with Performance Guarantees

Background: The nonlinear and large-scale nature of neural networks makes people hard to analyze. Therefore, they are mostly used as black-box models without formal guarantees. This issue becomes even more complicated when NNs are used in learning-enabled closed-loop systems.
Part 1: Present a convex optimization framework, inspired by robust control, that can provide useful certificates of stability, safety, and robustness for NN-driven systems.
Part 2: Address the problem of incorporating the safety guarantees of robust control into NN-driven uncertain dynamical systems.

Lecture 4

On the convergence and implicit bias of overparametrized linear networks

Question: Neural networks trained via gradient descent with random initialization and without any regularization enjoy good generalization performance in practice despite being highly overparametrized. How to explain this phenomenon?
Main content: Present a novel analysis of single-hidden-layer linear networks trained under gradient flow, which connects initialization, optimization, and overparametrization.
Convergence Analysis for Gradient Flow on Single-hidden-layer Linear Networks Given n training samples $\left\{x^{(i)}, y^{(i)}\right\}_{i=1}^n$, where $x^{(i)} \in \mathbb{R}^D$, $y^{(i)} \in \mathbb{R}^m$, we aim to solve the linear regression problem

$$\min_{\Theta\in\mathbb{R}^{D\times m}} \mathcal{L} = \frac{1}{2}\sum_{i=1}^n \left(y^{(i)} - \Theta^T x^{(i)}\right)^2.$$

Rewrite it as:

$$\mathcal{L}(V, U) = \frac{1}{2}\sum_{i=1}^n \left(y^{(i)} - VU^T x^{(i)}\right)^2 = \frac{1}{2}\left\|Y - XUV^T\right\|_F^2,$$

where $Y = \left[y^{(1)}, \cdots, y^{(n)}\right]^T$ and $X = \left[x^{(1)}, \cdots, x^{(n)}\right]^T$.
There are infinitely many solutions Θ^* that achieve optimal loss $\mathcal{L}^*$ of the first equation. Then he shows that under certain conditions, the trajectory of the loss function $\mathcal{L}(t) = \mathcal{L}(V(t), U(t))$ under gradient flow, i.e.,

Lecture 4

$$\dot{V}(t) = -\frac{\partial\mathcal{L}}{\partial V}(V(t), U(t)), \dot{U}(t) = -\frac{\partial\mathcal{L}}{\partial U}(V(t), U(t)),$$

converges to $\mathcal{L}^*$ exponentially.
Implicit Bias of Gradient Flow on Single-hidden-layer Linear Network
In this section, he shows that proper initialization constrains the dynamics of the network parameters to lie within an invariant set.
Assuming that $D > r = \mathrm{rank}(X)$, the regression problem in the last section has infinitely many solutions Θ^* that achieve optimal loss. Among all these solutions, one that is of particular interest in high-dimensional linear regression is the minimum norm solution

$$\begin{aligned}\hat{\Theta} &= \operatorname*{arg\,min}_{\Theta\in\mathbb{R}^{D\times m}} \left\{\|\Theta\|_2 : \|Y - X\Theta\|_F^2 = \min_\Theta \|Y - X\Theta\|_F^2\right\} \\ &= X^T\left(XX^T\right)^\dagger Y,\end{aligned}$$

which has near-optimal generalization error for suitable data models. Here we study conditions under which our trained network is equal or close to the min-norm solution by showing how the initialization explicitly controls the trajectory of the training parameters to be exactly (or approximately) confined within some low-dimensional invariant set. In turn, minimizing the loss over this set leads to the min-norm solution. (Sorry I couldn't understand this part well.)
Experimental Simulation
He shows that large hidden layer width, together with (properly scaled) random initialization, ensures proximity to such an invariant set during training, allowing us to derive a novel non-asymptotic upper-bound on the distance between the trained network and the min-norm solution.

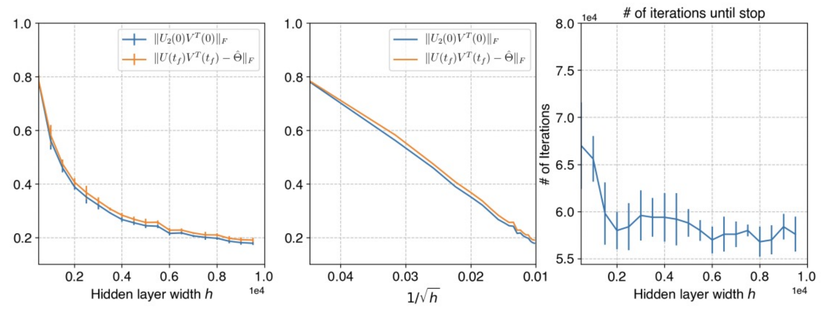


Figure 1. Implicit bias of wide single-hidden-layer linear network under random initialization. The line is plotting the average over 5 runs for each h, and the error bar shows the standard deviation. The gradient descent stops at iteration t_f.

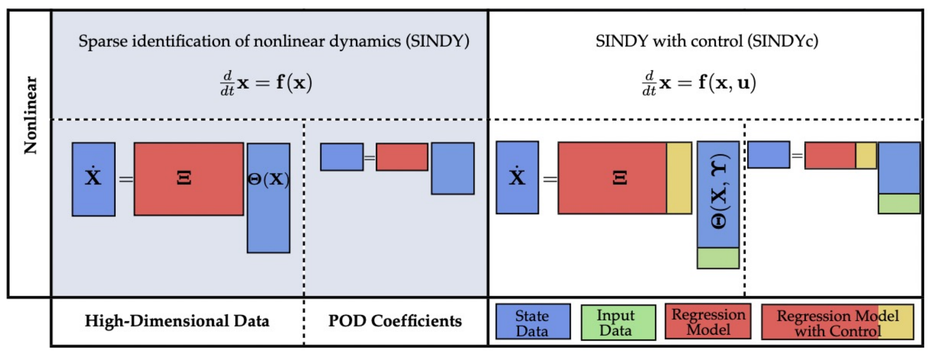
Lecture 5

Machine Learning for Scientific Discovery, with Examples in Fluid Mechanics

Main Content: This work describes how machine learning may be used to develop accurate and efficient nonlinear dynamical systems models for complex natural and engineered systems.
The sparse identification of nonlinear dynamics (SINDy) algorithm

$$\frac{d}{dt}\mathbf{x} = \mathbf{f}(\mathbf{x}, \mathbf{u})$$

The SINDY algorithm is readily generalized to include actuation, as this merely requires building a larger library $\Theta(\mathbf{x}, \mathbf{u})$ of candidate functions that include $\mathbf{u}$. It balances model complexity with accuracy, avoiding overfitting.



It's important to learn effective coordinate systems in which the dynamics may be expected to be sparse. This sparse modeling approach will be demonstrated on a range of challenging modeling problems in fluid dynamics. And he also discuss how to incorporate these models into existing model-based control efforts. Because fluid dynamics is central to transportation, health, and defense systems, then he emphasizes the importance of machine learning solutions that are interpretable, explainable, generalizable, and that respect known physics.

Lecture 6

Uncertainty Quantification in Learning Spatiotemporal Dynamics

Background: Applications such as public health, transportation, and climate science often require learning complex dynamics from large-scale spatiotemporal data. While deep learning has shown tremendous success in these domains, prior works have mostly focused on point estimates without quantifying the uncertainty of the predictions.
A systematic study of UQ for deep spatiotemporal forecasting
She analyzes UQ methods from both the Bayesian and the frequentist point of view, casting in a unified framework.

Method	Bootstrap	Quantile	SQ	MIS	MC Dropout	SG-MCMC
Computation	25	1	1	1	1	25
Small sample	×	×	×	×	×	✓
Consistency	✓	×	×	×	×	✓
Accuracy	✓	✓	✓	✓	✓	✓✓
Uncertainty	×	✓	×	✓✓	×	✓

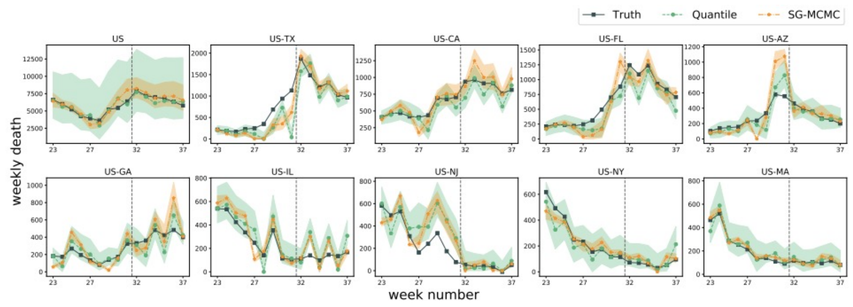
2022 TRIPODS Winter School Workshop

Yinling Zhang

Lecture 6

Interactive Neural Process (INP)

It's a Bayesian active learning framework that can significantly accelerate stochastic simulation. She demonstrates the method on the use cases of COVID-19 forecasting and scenario creation.



Lecture 7

Approximating functions, functionals and operators with neural networks for diverse applications

Background: Review physics-informed neural network and summarize available extensions for applications in computational mechanics and beyond.

Universal Approximation Theorem for Operator

Suppose that σ is a continuous nonpolynomial function, X is a Banach Space, $K_1 \subset X, K_2 \subset \mathbb{R}^d$ are compact, V is a compact set in $C(K_1)$, G is a nonlinear continuous operator. Then for any $\epsilon > 0$, there are positive integers n, p, m, constants $c_i^k, \xi_{ij}^k, \theta_i^k, \zeta_k \in \mathbb{R}, w_k \in \mathbb{R}^d, x_j \in K_1$ such that

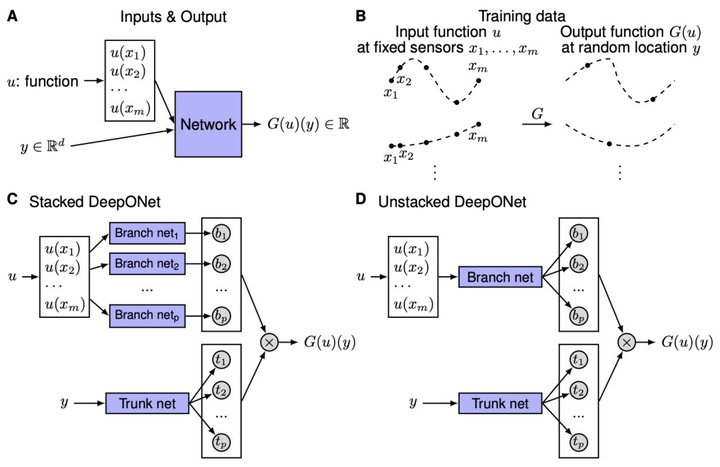
$$|G(u)(y) - \sum_{k=1}^{p} \underbrace{\sum_{i=1}^{n} c_i^k \sigma \left(\sum_{j=1}^{m} \xi_{ij}^k u(x_j) + \theta_i^k \right)}_{\text{branch}} \underbrace{\sigma(w_k \cdot y + \zeta_k)}_{\text{trunk}}| < \epsilon$$

holds for all $u \in V$ and $y \in K_2$.

This approximation theorem indicates the potential application of neural networks to learn nonlinear operators from data, i.e., similar to what the deep learning community is currently doing, that is learning functions from data.

Deep operator networks

Apply the universal theorem to design a new composite NN with small generalization error, the deep operator network (DeepONet), consisting of a NN for encoding the discrete input function space (branch net) and another NN for encoding the domain of the output functions (trunk net).



DeepONet can learn various explicit operators, e.g., integrals, Laplace transforms and fractional Laplacians, as well as implicit operators that represent deterministic and stochastic differential equations. More generally, DeepOnet can learn multiscale operators spanning across many scales and trained by diverse sources of data simultaneously.

Lecture 8

Continuous Network Models for Sequential Predictions

Background: Data-driven machine learning methods such as those based on deep learning are important in many areas of science and engineering for modeling time series. However, deep neural networks are known to be sensitive to various adversarial environments, and thus out of the box models and methods are often not suitable for mission critical applications.

Main Content: 1. Discuss deep dynamic autoencoders and argue that integrating physics-informed energy terms into the learning process can help to improve the generalization performance as well as robustness with respect to input perturbations.

2. Discuss novel continuous-time recurrent neural networks that are more robust and accurate than other traditional recurrent units.

3. Discuss extensions such as multiscale ordinary differential equations for learning long-term sequential dependencies and a connection between recurrent neural networks and stochastic differential equations.

Lecture 9

Games, Flocks, and Cognition

Main Content: Show how optimal control theory guides them in formulating a notion of cognitive cost (of predation avoidance) in a natural collective. They apply these ideas to data on flocking behavior of starlings, using geometric and computational techniques.

Lecture 10

Entrywise Estimation of Singular Vectors of Low-Rank Matrices with Heteroskedasticity and Dependence

Introduce: Propose an estimator for the singular vectors of high-dimensional low-rank matrices corrupted by additive subgaussian noise, where the noise matrix is allowed to have dependence within rows and heteroskedasticity between them.

Study the statistical properties for the individual entries of our estimator, and they apply these results to high-dimensional mixture models.

Results: The main result clearly shows the geometric relationship between the signal matrix, the covariance structure of the noise, and the distribution of the errors in the singular vector estimation task.