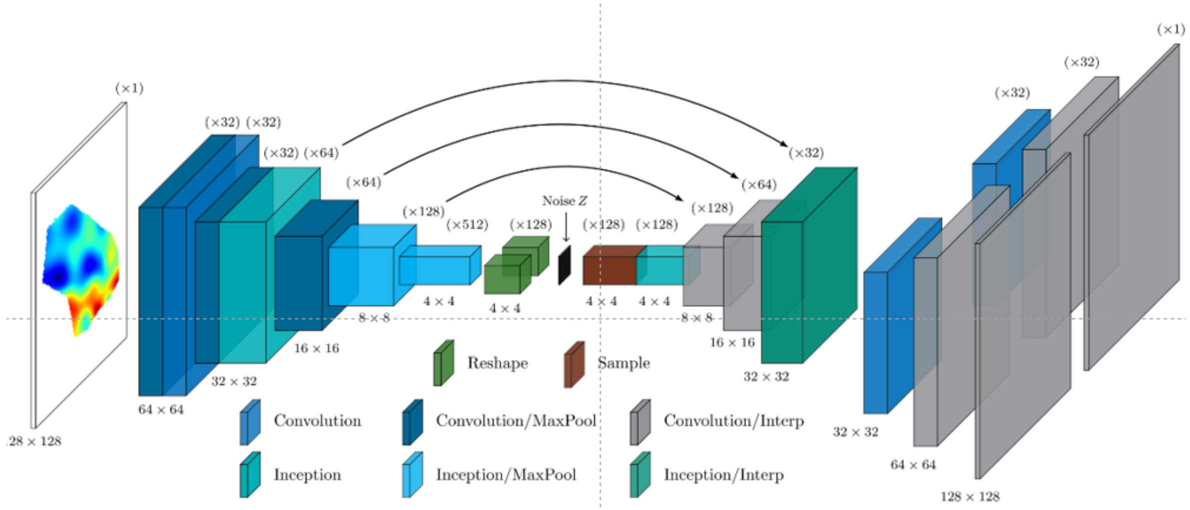
Towards Interpretable, Trustworthy Machine Learning for Science

Guang Lin,

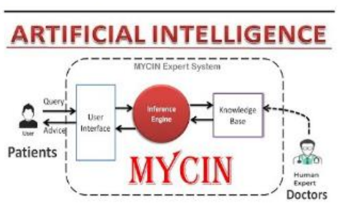
Associate Dean for Research and Innovation, Director of Data Science Consulting Service, Purdue University

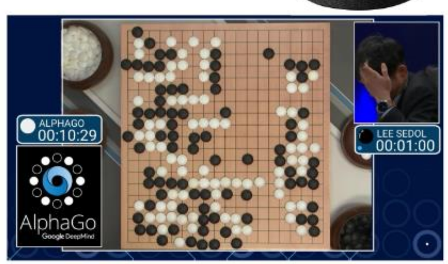


Department of Computer Science, Purdue University, 09/11/2025

The Four Waves of AI

First Wave	Second Wave	Third Wave	Fourth Wave
c. 1970s - 1990s	*c. 2000s - present*	*est. 2020s - 2030s*	*est. 2030s →*
Good at reasoning, but no ability to learn or generalize. • GOFAI - "Good Old Fashioned AI." • Symbolic, heuristic, rule based. • Handcrafted knowledge, "expert systems."	Good at learning and perceiving, but minimal ability to reason or generalize. • Statistical learning, "deep" neural nets, CNNs, RNNs. • Advanced text, speech, language and vision processing.	Excellent at perceiving, learning and reasoning, and able to generalize. • Contextual adaptation, able to explain decisions. • Can converse in natural language. • Requires far fewer data samples for training. • Able to learn and function with minimal supervision.	Able to perform any intellectual task that a human can. • AGI (Artificial General Intelligence), possibly leading to ASI (Artificial Superintelligence) and the "Technological Singularity."

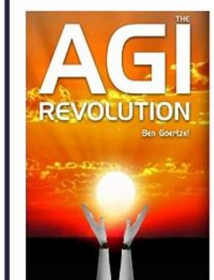


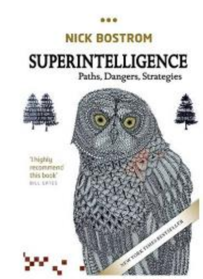


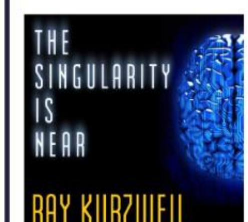
SingularityNET

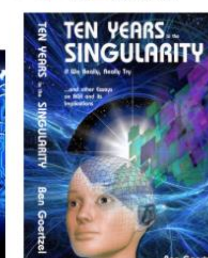
aigo.ai

Pandai









Six Kin Development (adapted from DARPA's "Three Waves of AI")

How Data Science, Artificial Intelligence, and Digital Twins Could Help US Predict the Future

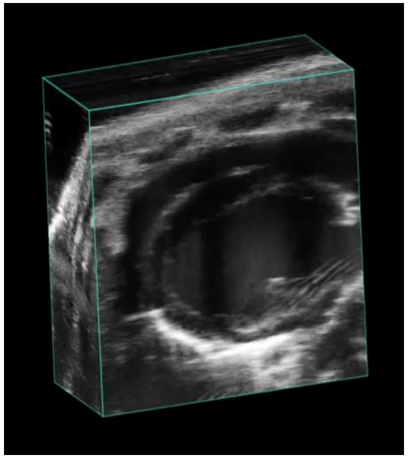
NIH PROJECT ON HEALTH DIGITAL TWIN TO TACKLE PEDIATRIC CARDIOVASCULAR DISEASE

NSF PROJECT ON MULTISCALE DIGITAL TWIN FOR AUTONOMOUS OPERATION OF POWER GRIDS

How to Build Robust AI in Real-World Environment?
-Funded by NSF/Simons Foundation Research Grant on Mathematical and Scientific Foundation of Deep Learning (Scale-MoDL)

Optical Adversarial Attack Can Change the Meaning of Road Signs

Machine Learning Driven Contouring System for High-Frequency Four-Dimensional Cardiac Ultrasound and Photoacoustic Imaging

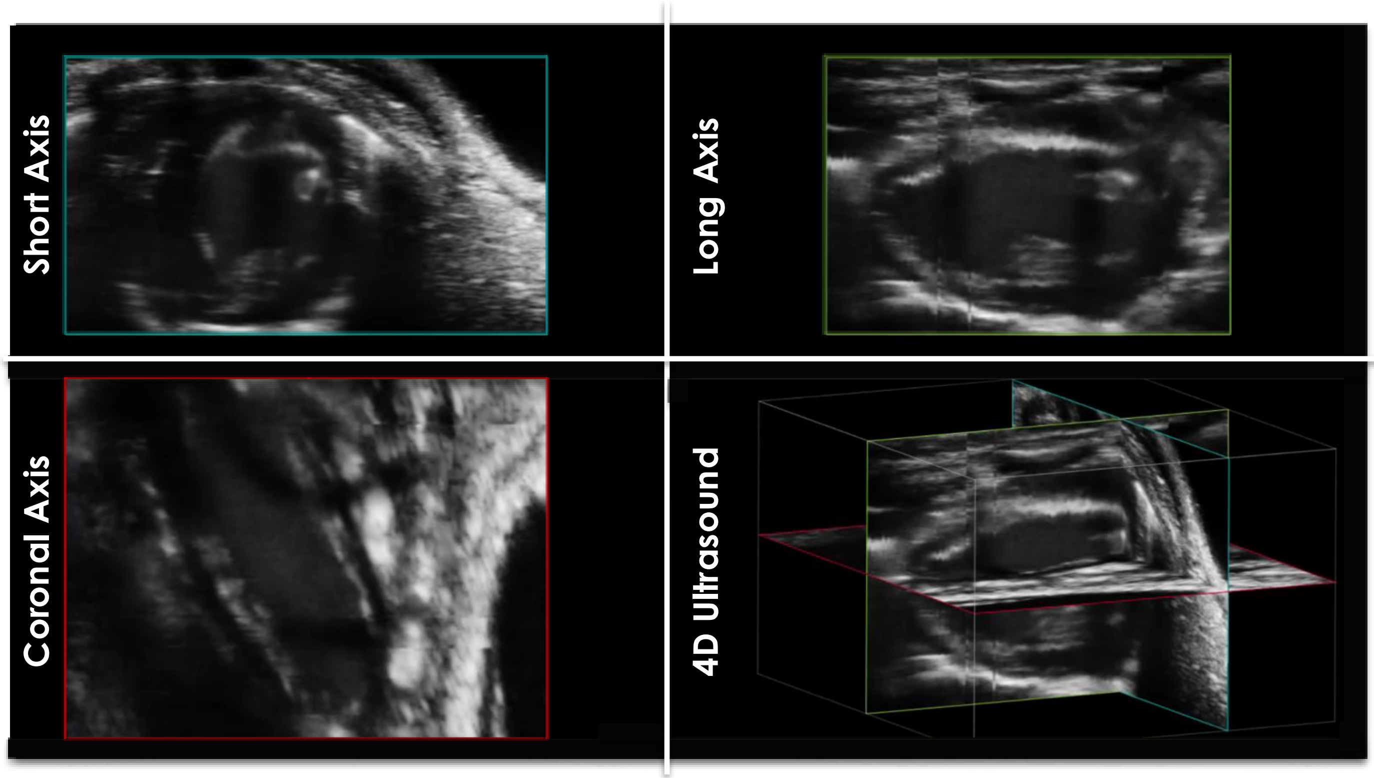


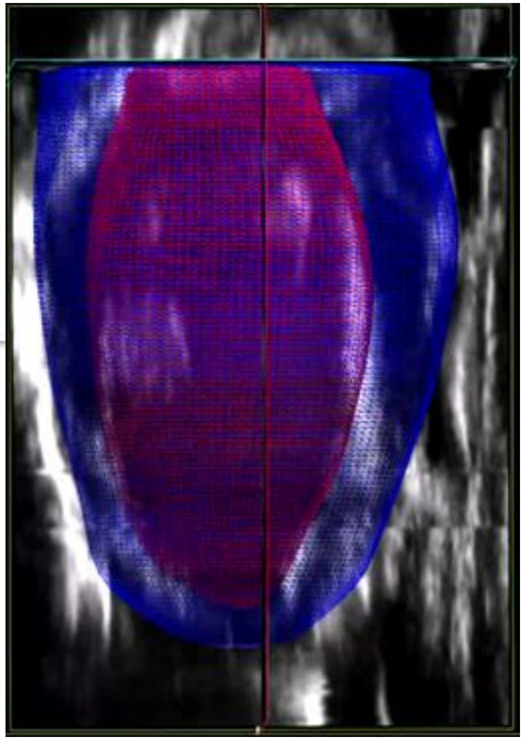
- Guang Lin, Full Professor School of Mechanical Engineering & Department of Mathematics, Purdue University
- Craig J Goergen, Leslie A. Geddes Associate Professor, Weldon School of Biomedical Engineering, Purdue University

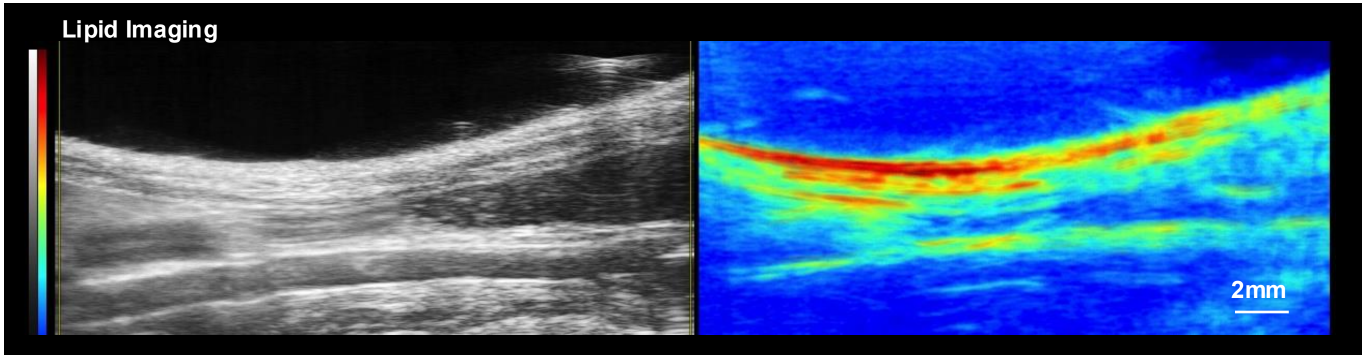
- PRF technology number 69227-02 and 66849

- Trask Grant: Innovation Sparks (Life Science and Medical Devices)

4D Ultrasound: Healthy LV



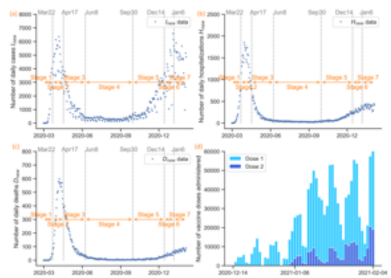




Guang Lin's Group's Main Research

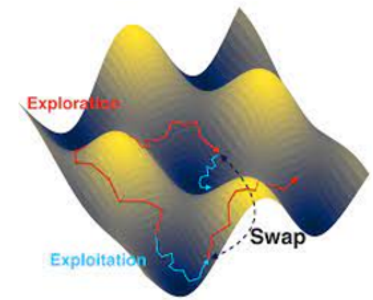
Interpretable AI:
Discovery of Physical Laws from Noisy Data

1. Nature Computational Science, 1-10, 2021
2. Nature Digital Medicine, 2023
3. Proceeding of the Royal Society of London, 2018
4. PlOS Computational Biology, 2021



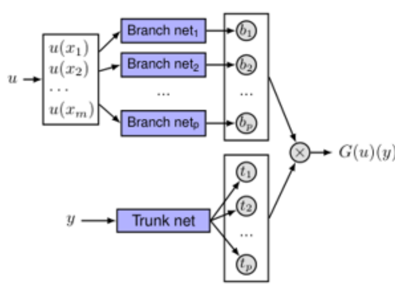
Uncertainty Quantification for Reliable AI

- NeurIPS19, NeurIPS20, ICML20, ICLR21, WSDM21, ICLR22, TMLR22, AAAI-23



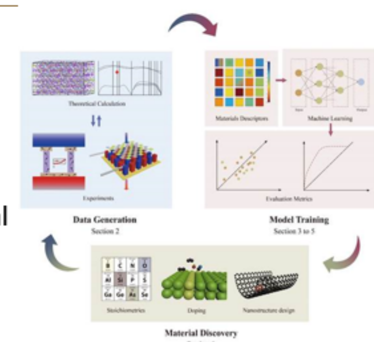
Privacy-preserving AI for Learning from distributed sensitive data:

- Federated Averaging Langevin Dynamics
- Fed-DeepONet



AI for Science & Engineering:
AI for Material Discovery

- Nature Computational Material, 23
- Scientific Report 22



AI for Workforce Development:
IMPACT Data Science Education

PURDUE UNIVERSITY | College of Science

Outline:

- **Incorporate Physics Knowledge and AI to design new interpretable models – Trustworthy Epidemiological Models for COVID-19 Prediction & Intervention**

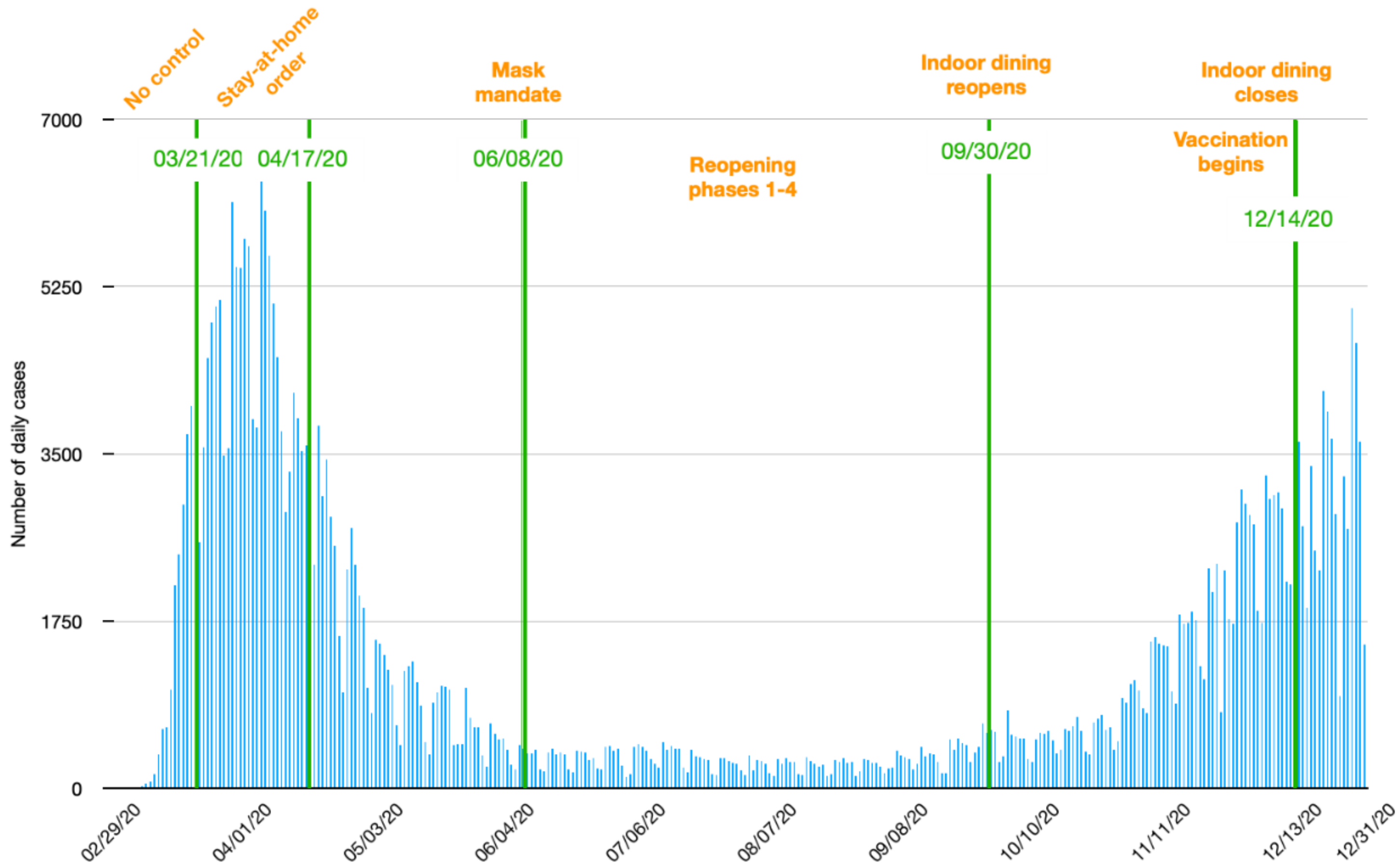
- Interpretable AI enables data-driven scientific discovery with uncertainty quantification capability – ALZHEIMER's Disease Prediction

- Scalable training large-scale Deep Neural Network

How to incorporate Physics Knowledge and AI to design new interpretable models? - Interpretable AI for Science

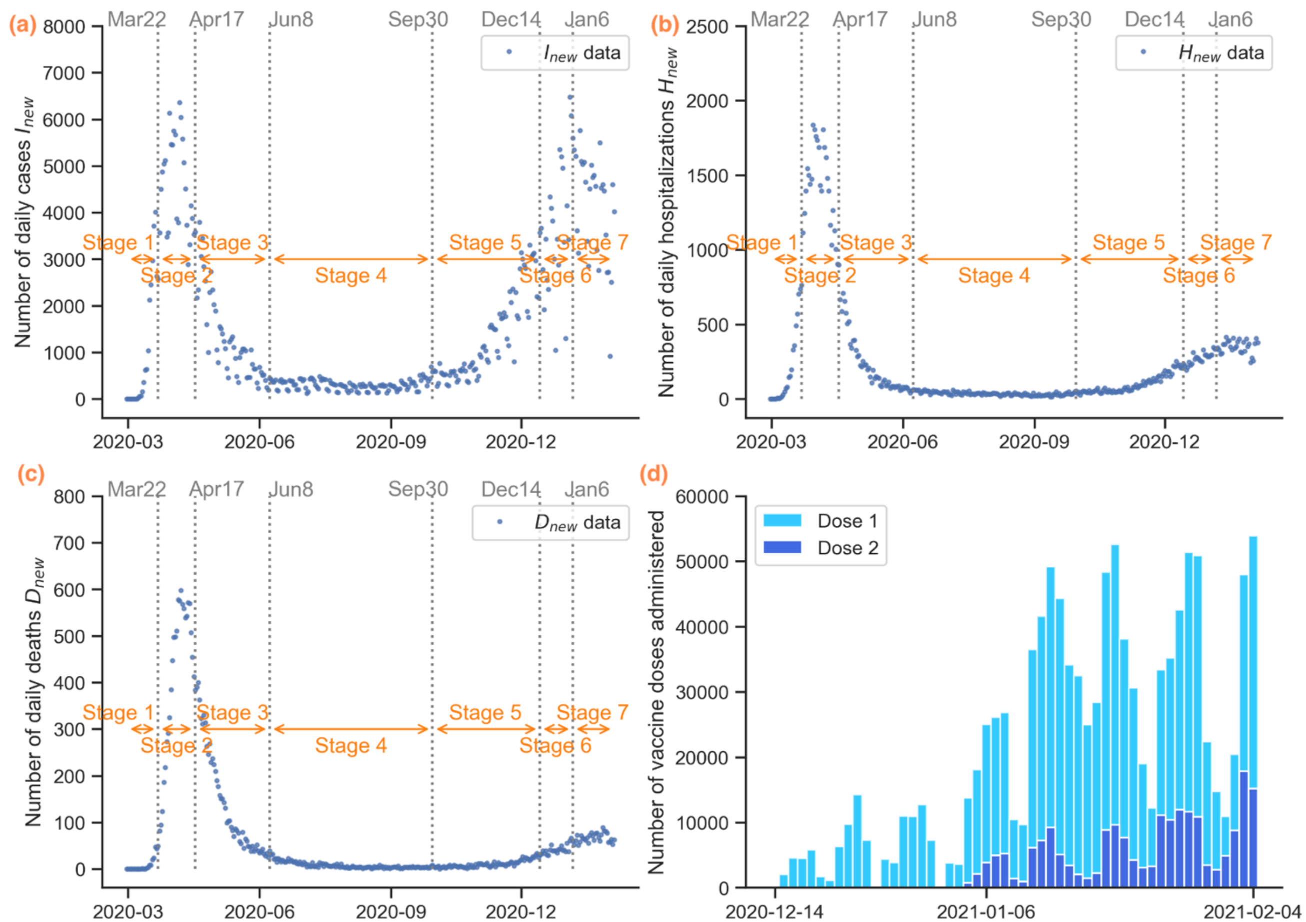
1. Ehsan Kharazmi, Min Cai, Xiaoning Zheng, Guang Lin, George Em Karniadakis, **Identifiability and predictability of integer- and fractional-order epidemiological models using physics-informed neural networks, Nature Computational Science, 1, 744-753, 2021**
2. Sheng Zhang, Joan Ponce, Zhen Zhang, Guang Lin, George Karniadakis, **An integrated framework for building trustworthy data-driven epidemiological models: Application to the COVID-19 outbreak in New York City, *PLoS Computional Biology*** 17(9): e1009334. https://doi.org/10.1371/journal.pcbi.1009334

New York City COVID-19 related Event Timeline



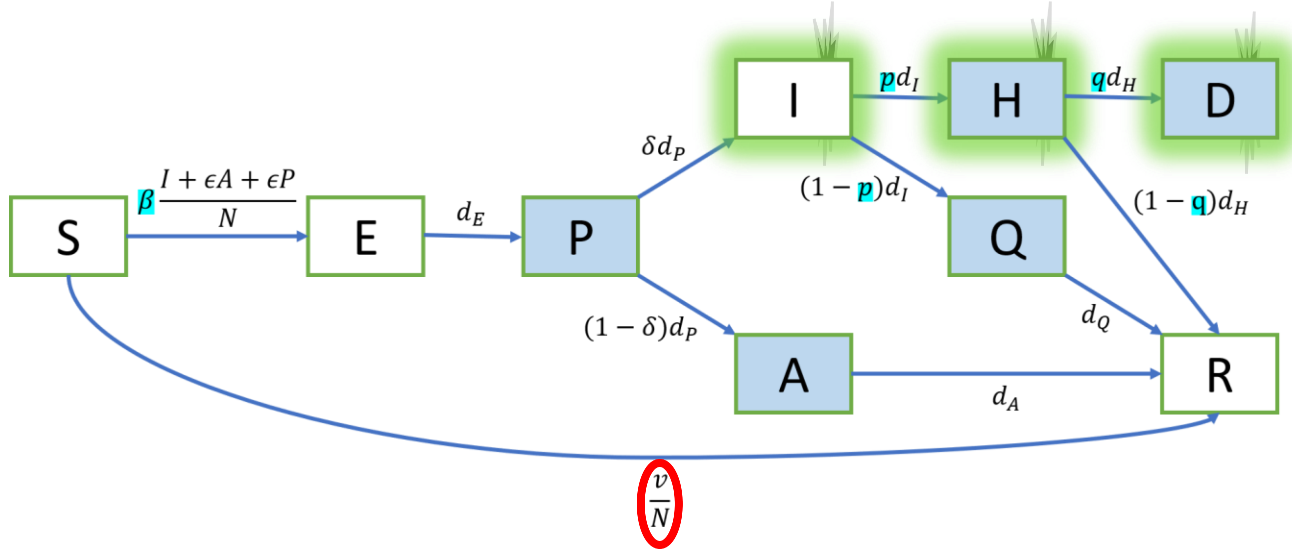
New York City COVID-19 related Event Timeline

Calibrate piecewise-constant model parameters to capture local epidemiological dynamics



Epidemiological Model Development

$$
\begin{cases}
\frac{dS}{dt} = -\beta \frac{I + \epsilon A + \epsilon P}{N} S - \frac{v}{N} S \\
\frac{dE}{dt} = \beta \frac{I + \epsilon A + \epsilon P}{N} S - d_E E \\
\frac{dP}{dt} = d_E E - d_P P \\
\frac{dI}{dt} = \delta d_P P - d_I I \\
\frac{dA}{dt} = (1-\delta) d_P P - d_A A \\
\frac{dH}{dt} = p d_I I - d_H H \\
\frac{dQ}{dt} = (1-p) d_I I - d_Q Q \\
\frac{dD}{dt} = q d_H H \\
\frac{dR}{dt} = d_A A + (1-q) d_H H + d_Q Q + \frac{v}{N} S.
\end{cases}
$$

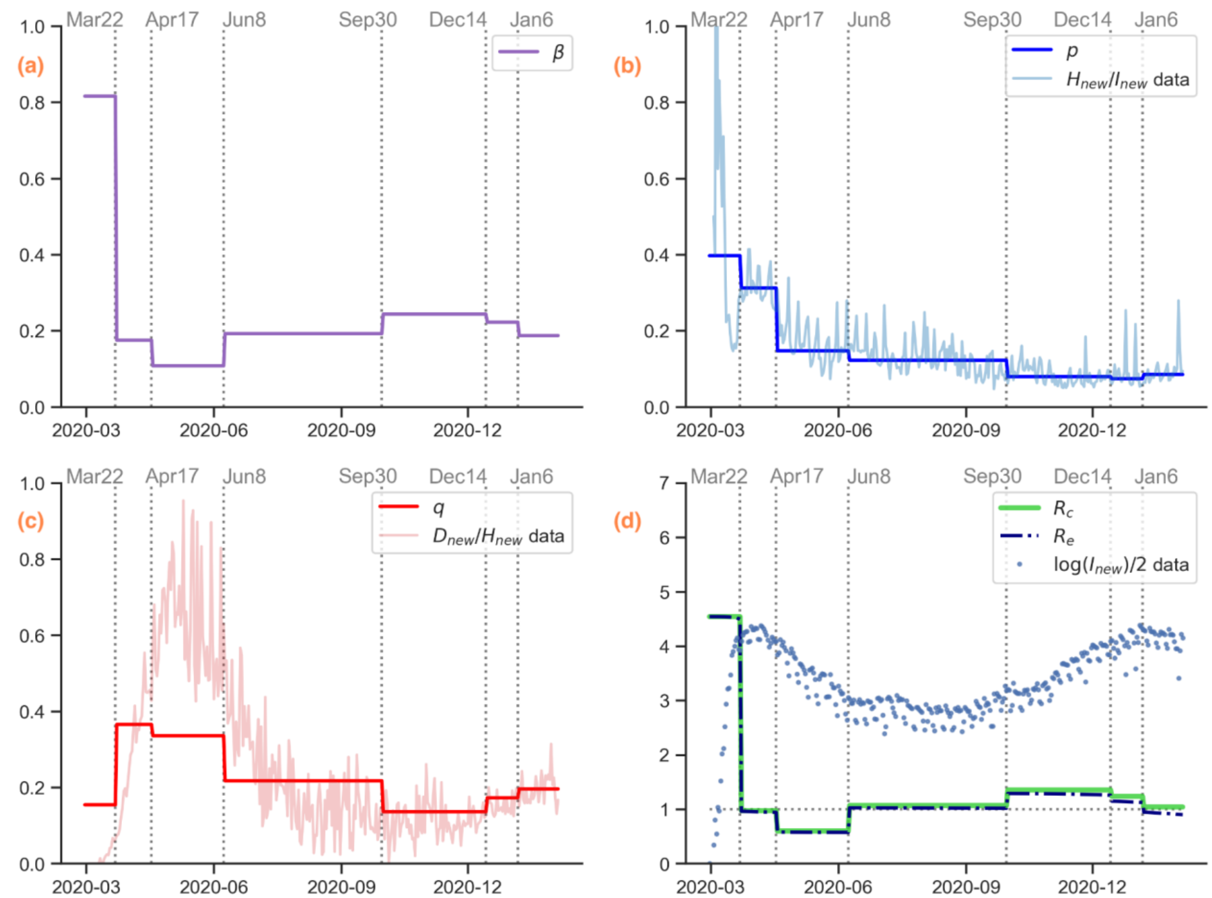


Fixed parameters:
eps = 0.75
delta = 0.6
d_E = 1/2.9
d_P = 1/2.3
d_I = 1/2.9
d_A = 1/7
d_H = 1/6.9
d_Q = 1/10

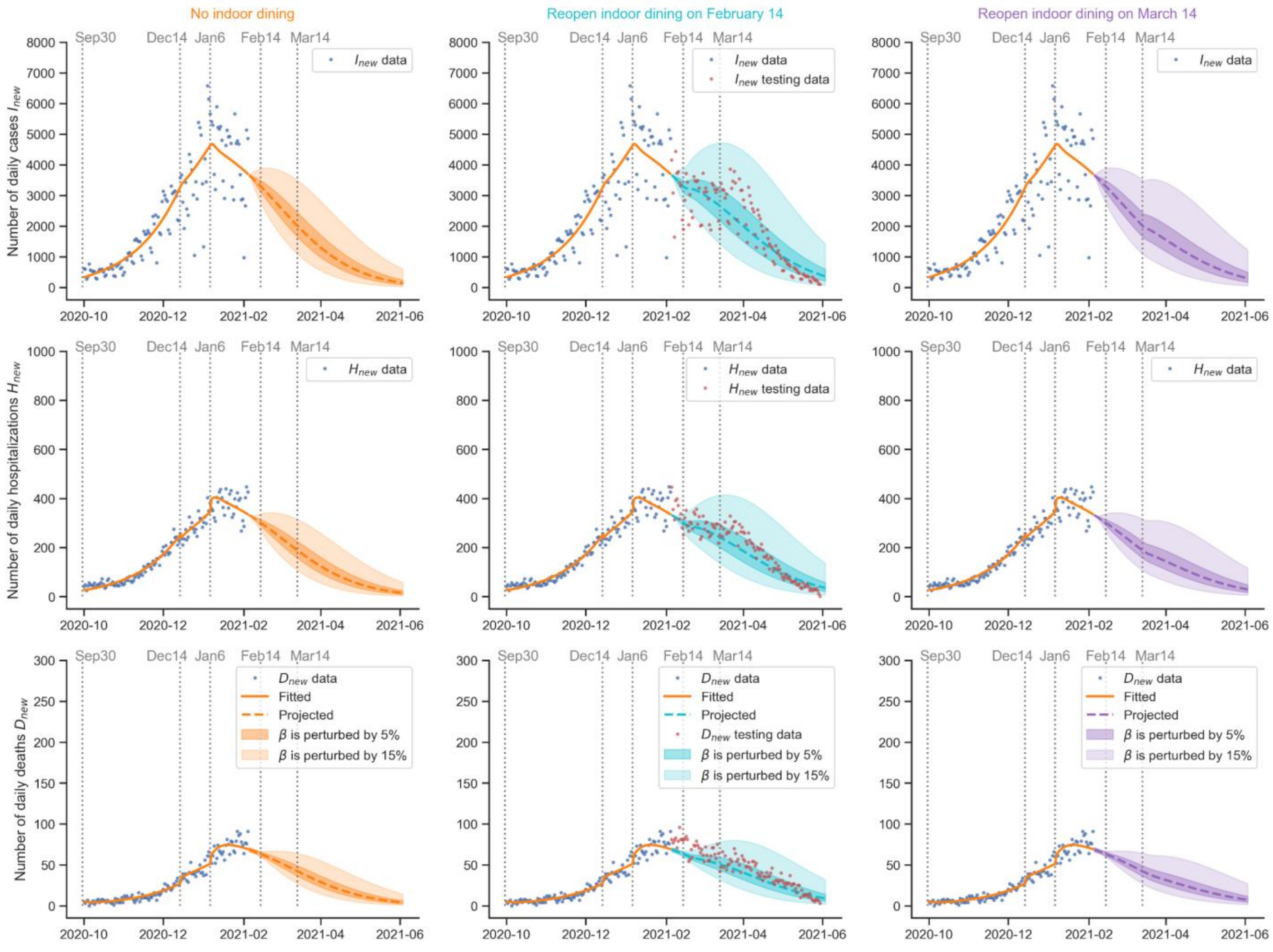
β: Transmission rate
p: Hospitalization rate
q: Death from hospital rate

Calibrated COVID-19 Transmission Rate for New York City

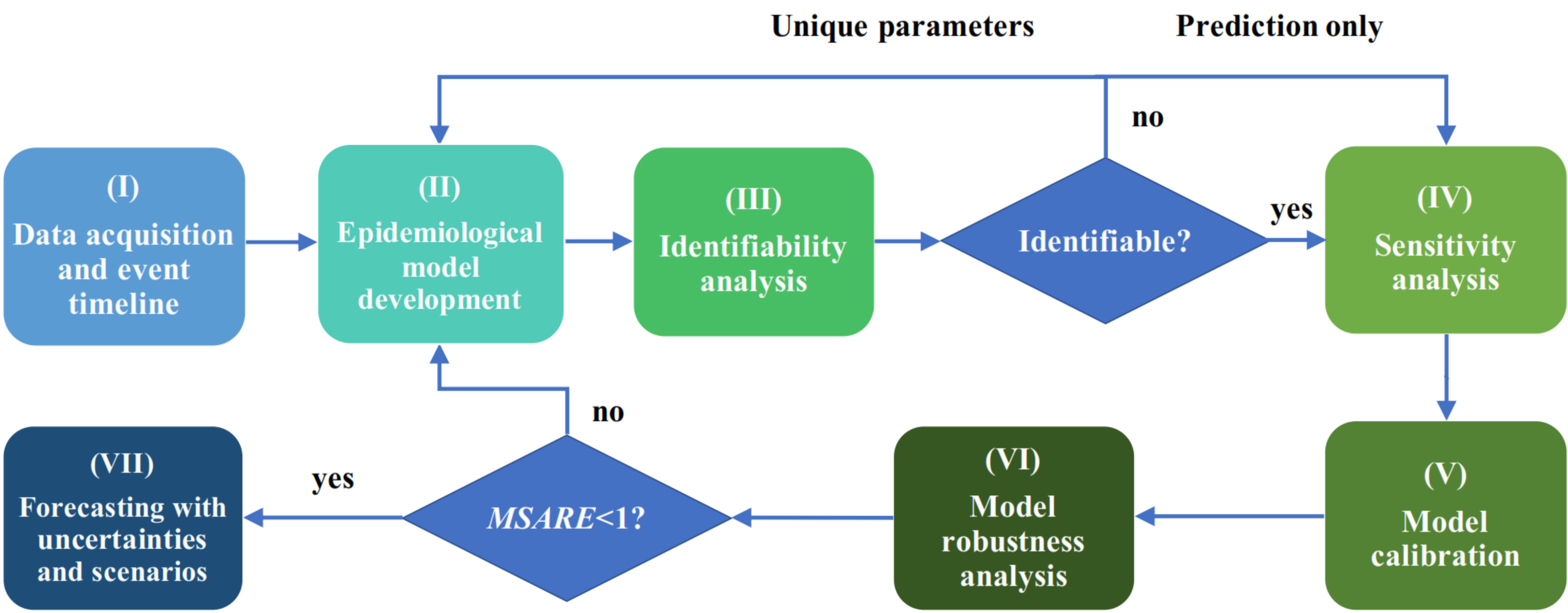
Calibrate piecewise-constant model parameters to capture local epidemiological dynamics



Forecasting with Uncertainties and Scenarios

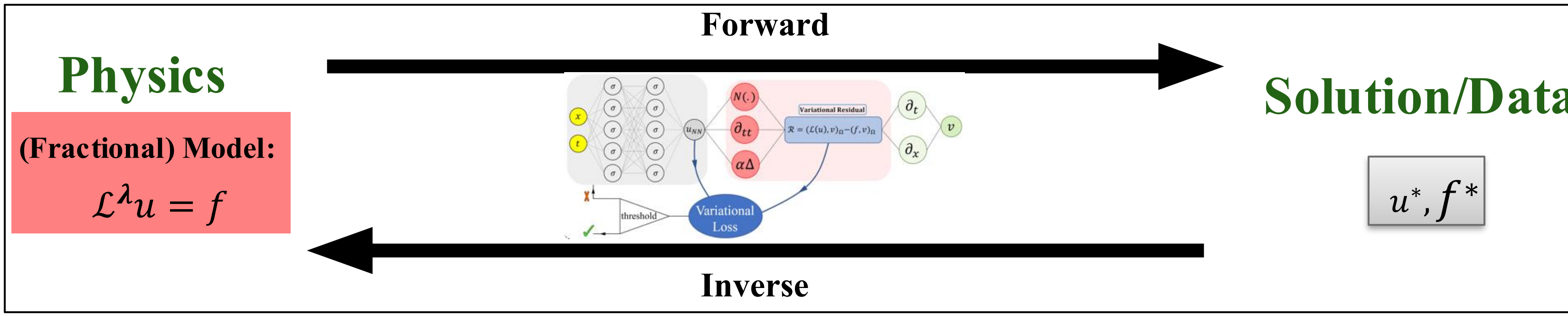


A general framework for building a trustworthy data-driven epidemiological model



Sheng Zhang, Joan Ponce, Zhen Zhang, Guang Lin, George Karniadakis, **An integrated framework for building trustworthy data-driven epidemiological models: Application to the COVID-19 outbreak in New York City,** ***PLoS Computional Biology*** 17(9): e1009334. https://doi.org/10.1371/journal.pcbi.1009334

Physics Informed Neural Networks (PINNs)



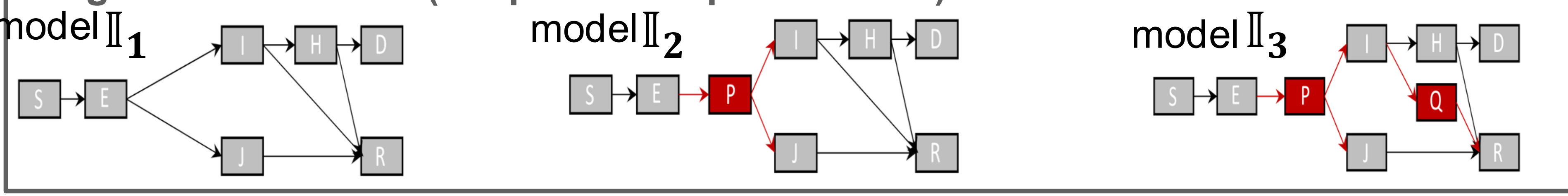
- $\mathcal{L}^{\boldsymbol{\lambda}}$ A (non-local) differential operator with parameters $\boldsymbol{\lambda}$

A flexible **computational** tool to study **model uncertainty**
Incorporate **data** and **different models**
Accurate **fitting** to data
Inferring model parameters and **discovering** unobserved dynamics

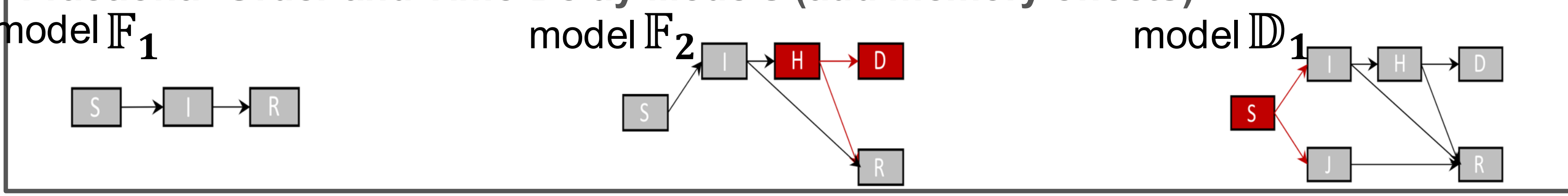
1. Ehsan Kharazmi, Min Cai, Xiaoning Zheng, Guang Lin, George Em Karniadakis, **Identifiability and predictability of integer- and fractional-order epidemiological models using physics-informed neural networks, Nature Computational Science, 1, 744-753, 2021**

Different Epidemiological Models

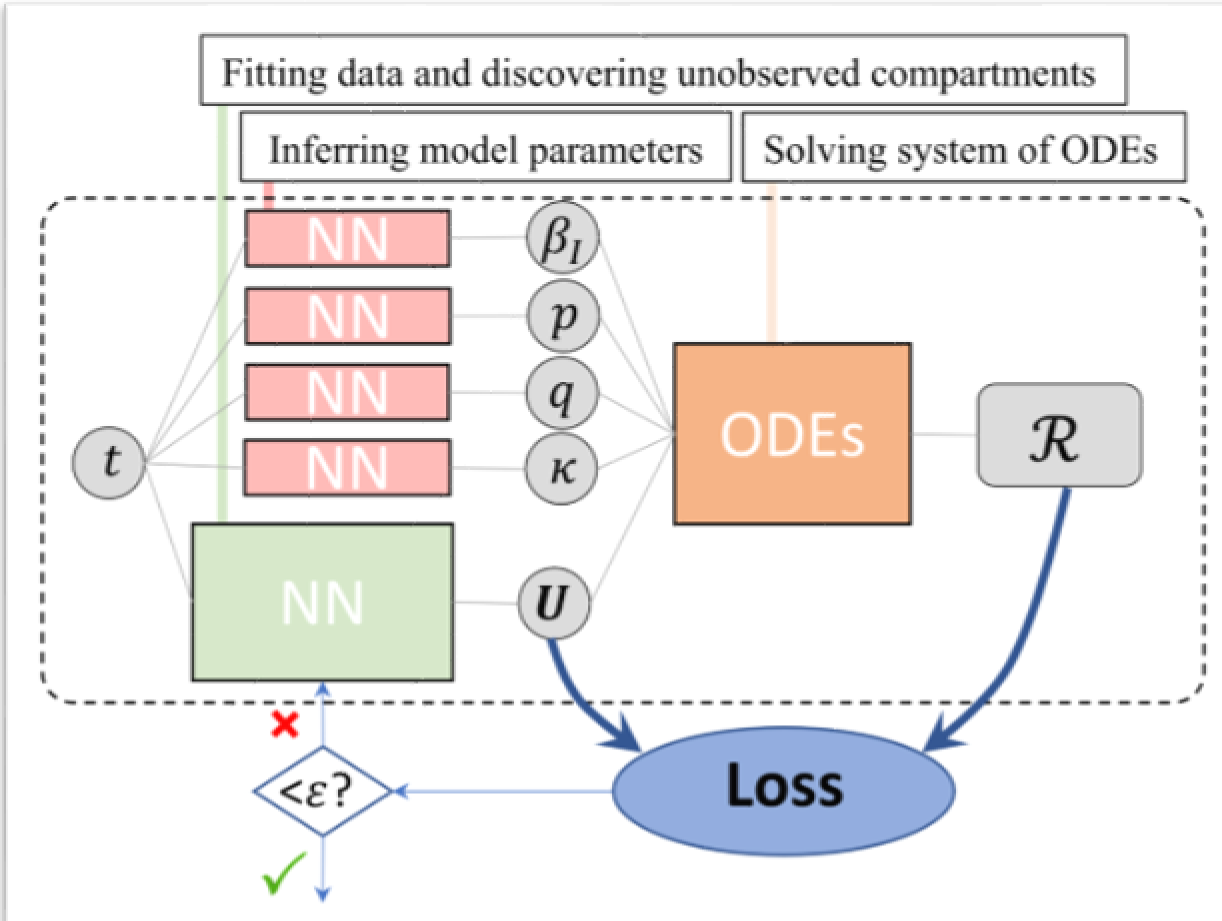
Integer-Order Models (simple to complex models)



Fractional-Order and Time-Delay Models (add memory effects)



PINNs for (Fractional) Epidemiological Models



$$loss =$$

$$\frac{1}{N_u}\sum_{i=1}^{N_u}\left|\boldsymbol{U}(t_i;\boldsymbol{\theta}) - data(t_i)\right|^2$$

Loss u (data/IC points)

$$+\frac{1}{N_r}\sum_{j=1}^{N_r}\left|\mathcal{R}\left(t_j;\boldsymbol{\theta},\boldsymbol{\lambda}\right)\right|^2$$

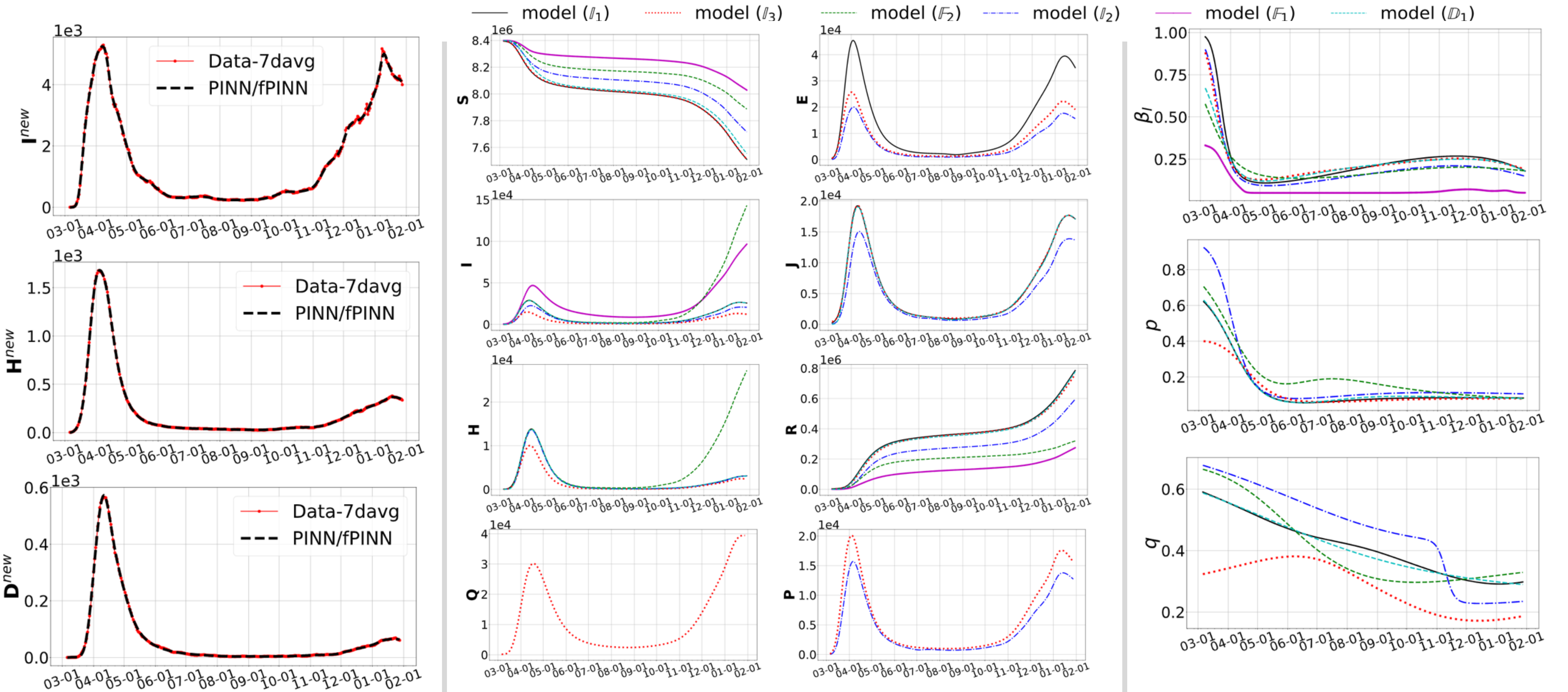
Loss ODE (residual points)

PINN Results: Model Uncertainty based on NYC dataset

Fitting the data accurately

Discovering unobserved dynamics

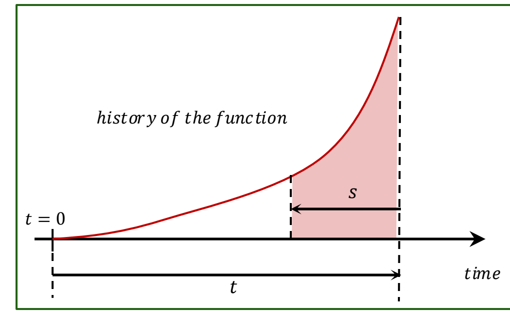
Inferring model parameters



Fractional Order Models Introduce Memory in the Dynamics

Caputo fractional derivative of order $\kappa \in (0,1)$: a **convolution** type **integro-differential** operator

$$\frac{\partial^\kappa}{\partial t^\kappa}u(t) = {}^C_0\mathcal{D}^\kappa_t\, u(t) = \frac{1}{\Gamma(1-\kappa)}\int_0^t \frac{1}{(t-s)^\kappa}\frac{du(s)}{ds}\,ds$$

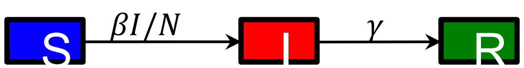


Memory: The derivative at time t depends on the weighted values of the function **from initial point $t = 0$ up to current time t.**

- Fractional order κ is the notion of memory effect
- Smaller κ can induce a delay in the dynamics
- $\kappa = \kappa(t)$ can be time varying

Different Compartments May Have Different Memory Effects!

model $\mathbb{F}_1$

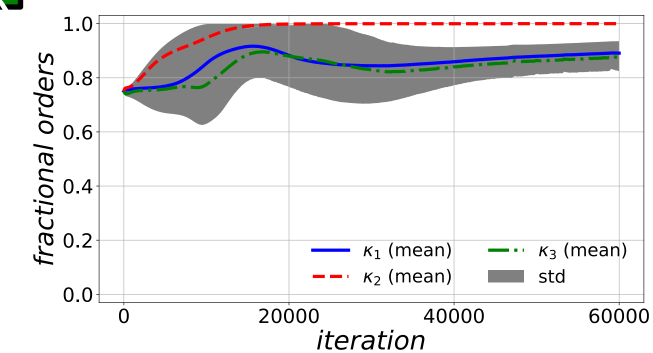


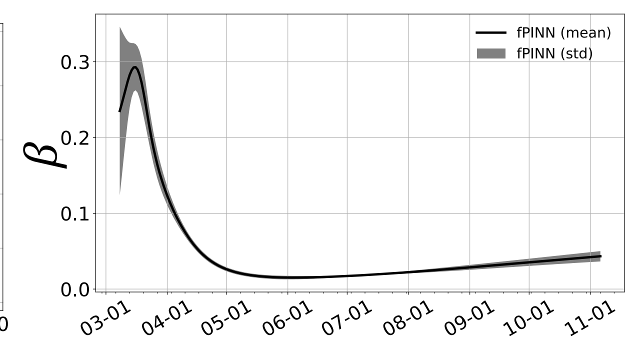
$${}^C_0\mathcal{D}^{\kappa_1}_t\mathbf{S}(t) = -\frac{\beta}{N}\mathbf{I}(t)\mathbf{S}(t),$$

$${}^C_0\mathcal{D}^{\kappa_2}_t\mathbf{I}(t) = \frac{\beta}{N}\mathbf{I}(t)\mathbf{S}(t) - \gamma\mathbf{I}(t),$$

$${}^C_0\mathcal{D}^{\kappa_3}_t\mathbf{R}(t) = \gamma\mathbf{I}(t),$$

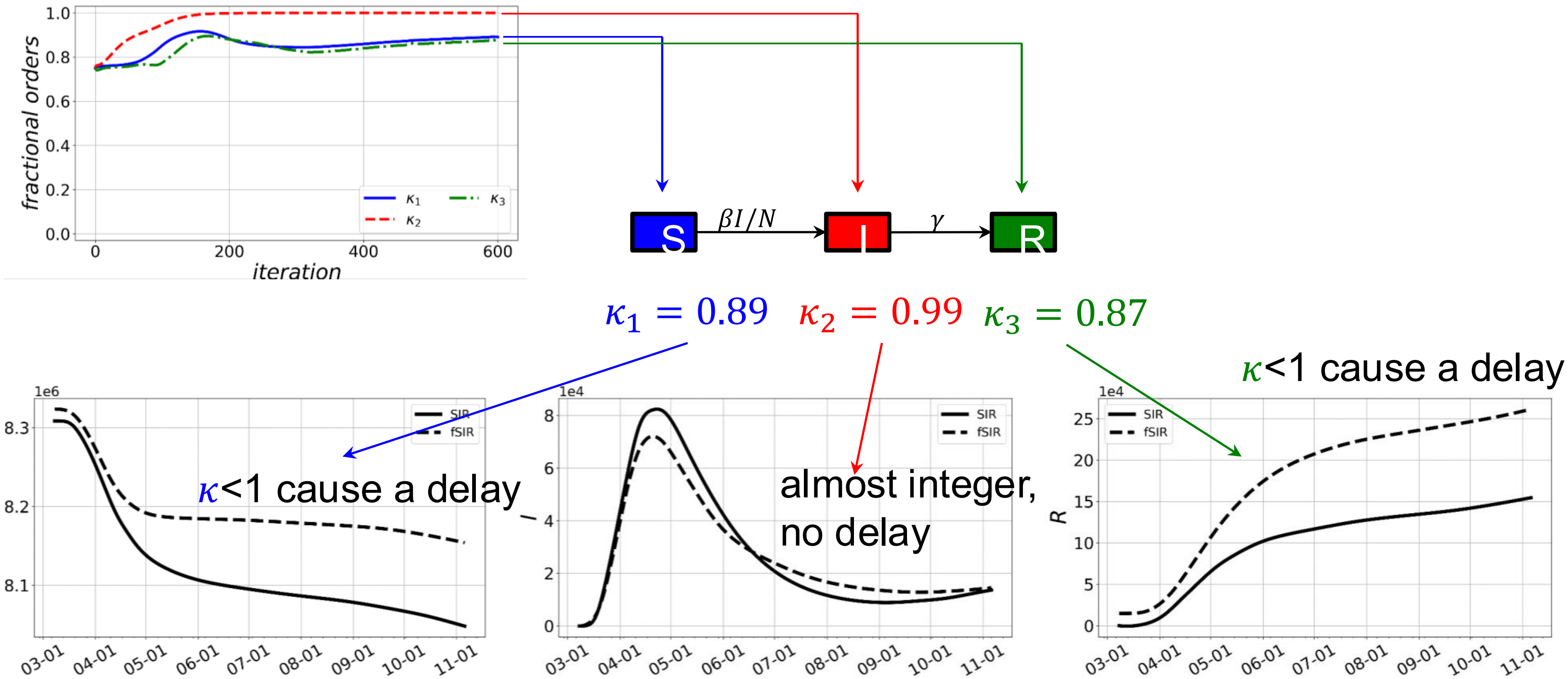
$${}^C_0\mathcal{D}^{\kappa_2}_t\mathbf{I}^c(t) = \frac{\beta}{N}\mathbf{I}(t)\mathbf{S}(t).$$





model $\mathbb{F}_1$

Fractional Order SIR V.S. Integer Order SIR



Summary

This is the first work to employ structural and practical identifiability tools to study COVID-19 model identifiability based on the available data.

A general data-driven epidemiological modeling framework is developed, which seamlessly integrates model identifiability, model sensitivity analysis, model calibration, model prediction with confidence intervals, and evaluating control strategies under uncertainties.

We treat beta (transmission rate), p (proportion of isolated individuals), and q (proportion of disease-related deaths) as time-dependent piece-wise model parameters and calibrate them using the available New York City COVID-19 dataset.

The developed COVID-19 model is employed to evaluate the effects of vaccination deployment scenarios.

We developed a flexible computational framework using physics-informed neural networks (PINNs) to study model uncertainty and discover time-dependent parameters.

Outline:

- Incorporate Physics Knowledge and AI to design new interpretable models – Trustworthy Epidemiological Models for COVID-19 Prediction & Intervention
- **Interpretable AI enables data-driven scientific discovery with uncertainty quantification capability – ALZHEIMER's Disease Prediction**
- Scalable training large-scale Deep Neural Network

Interpretable AI:

Question: Can we use available observation data to discover the physical laws?

Goal: Enable Data-driven Scientific Discovery?

S. Zhang, **G. Lin**, Robust data-driven discovery of governing physical laws with error bars, Proceedings of the Royal Society of London. Series A, mathematical, physical and engineering sciences, in press, 2018.

Jiuhai Chen, Lulu Kang, Guang Lin, Gaussian process assisted active learning of physical laws, Technometrics, in press, 2020.
https://doi.org/10.1080/00401706.2020.1817790

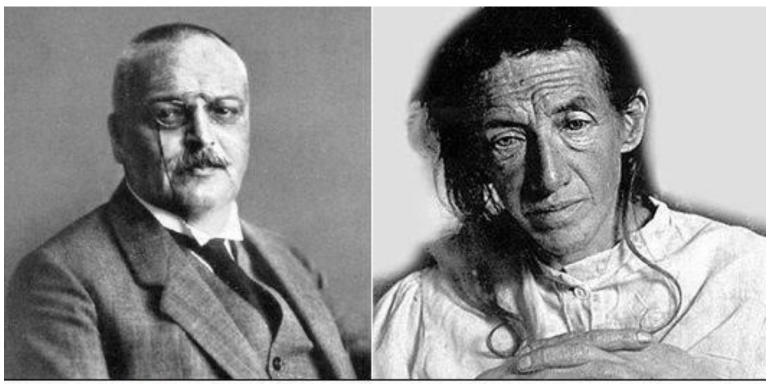
Sheng Zhang, Guang Lin, Robust subsampling-based threshold sparse Bayesian regression to tackle high noise and outliers for data-driven discovery of differential equations, Journal of Computational Physics, 428: 109962, 2021.

ALZHEIMER’S DISEASE PREDICTION

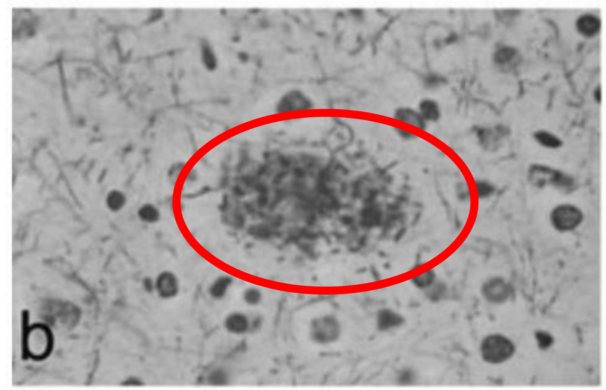
Haoyang Zheng, Jeffrey Petrella, P. Murali Doraiswamy, **Guang Lin***, Wenrui Hao, Data-driven causal model discovery and personalized prediction in Alzheimer’s disease, Nature NPJ Digital Medicine, 5, 137, 2022.

Background - Alzheimer's Disease

Dr. Alois Alzheimer (1864-1915)



Alois Alzheimer — Auguste Deter

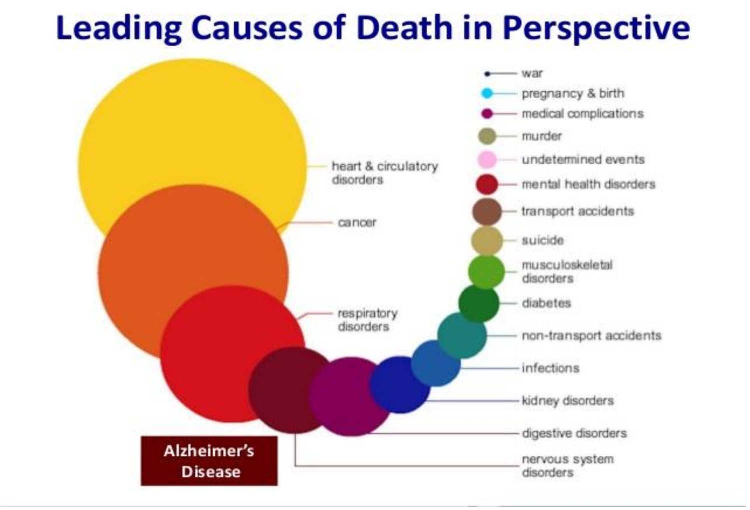


Dr. Alzheimer was the physician who first reported on a patient (Auguste) with dementia, later termed as "Alzheimer's Disease".

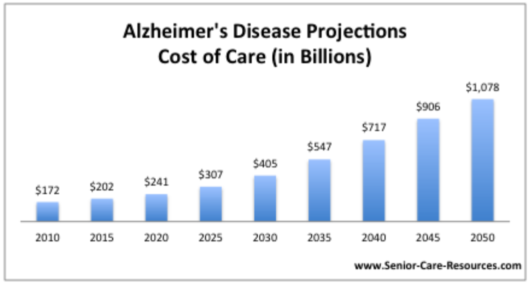
Distinctive plaques and **neurofibrillary tangles** in the brain histology

Zheng Haoyang, et al. "Data-driven causal model discovery and personalized prediction in Alzheimer's disease." NPJ digital medicine 2022

Why AD is important?

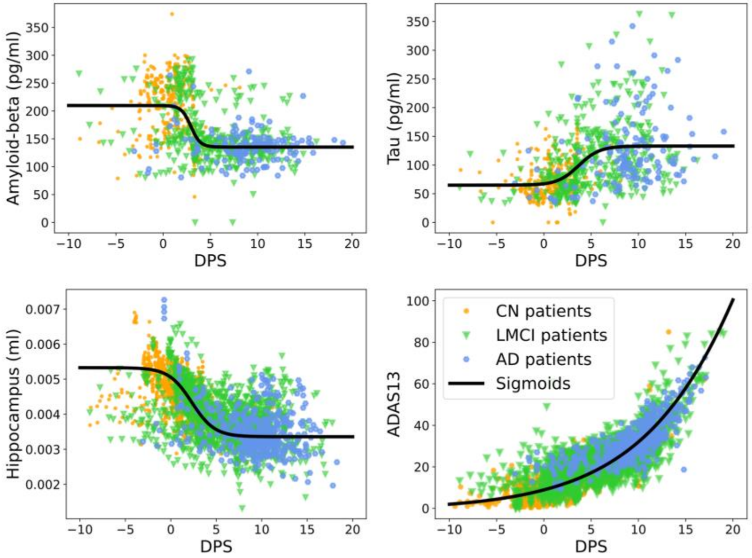


- Most common dementia
- In 2020, over 55 million people have AD
- By 2050, the number could increase to 150 million



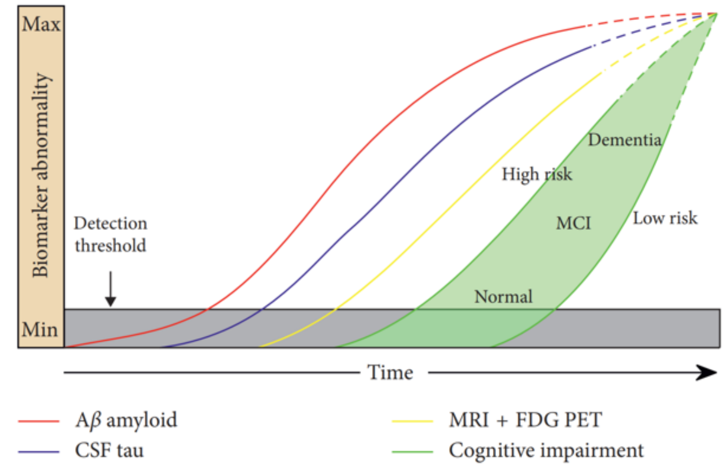
Challenges and Motivation

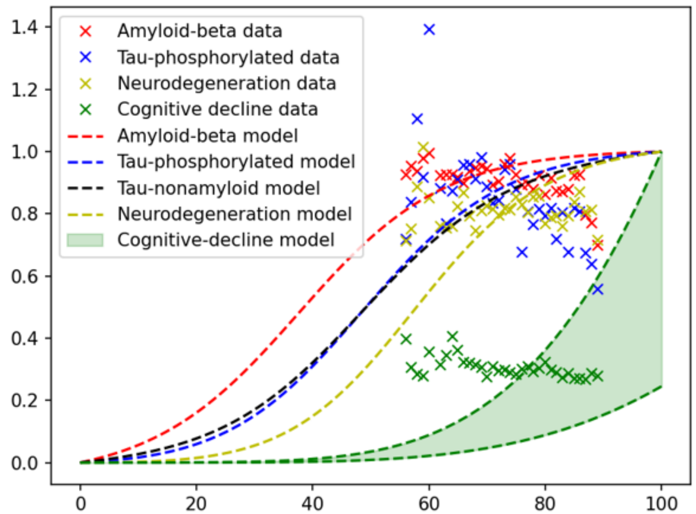
- Can we build data-driven model with ADNI dataset?
- Can we build an interpretable model?
- Can we design personalized model for each treatment?



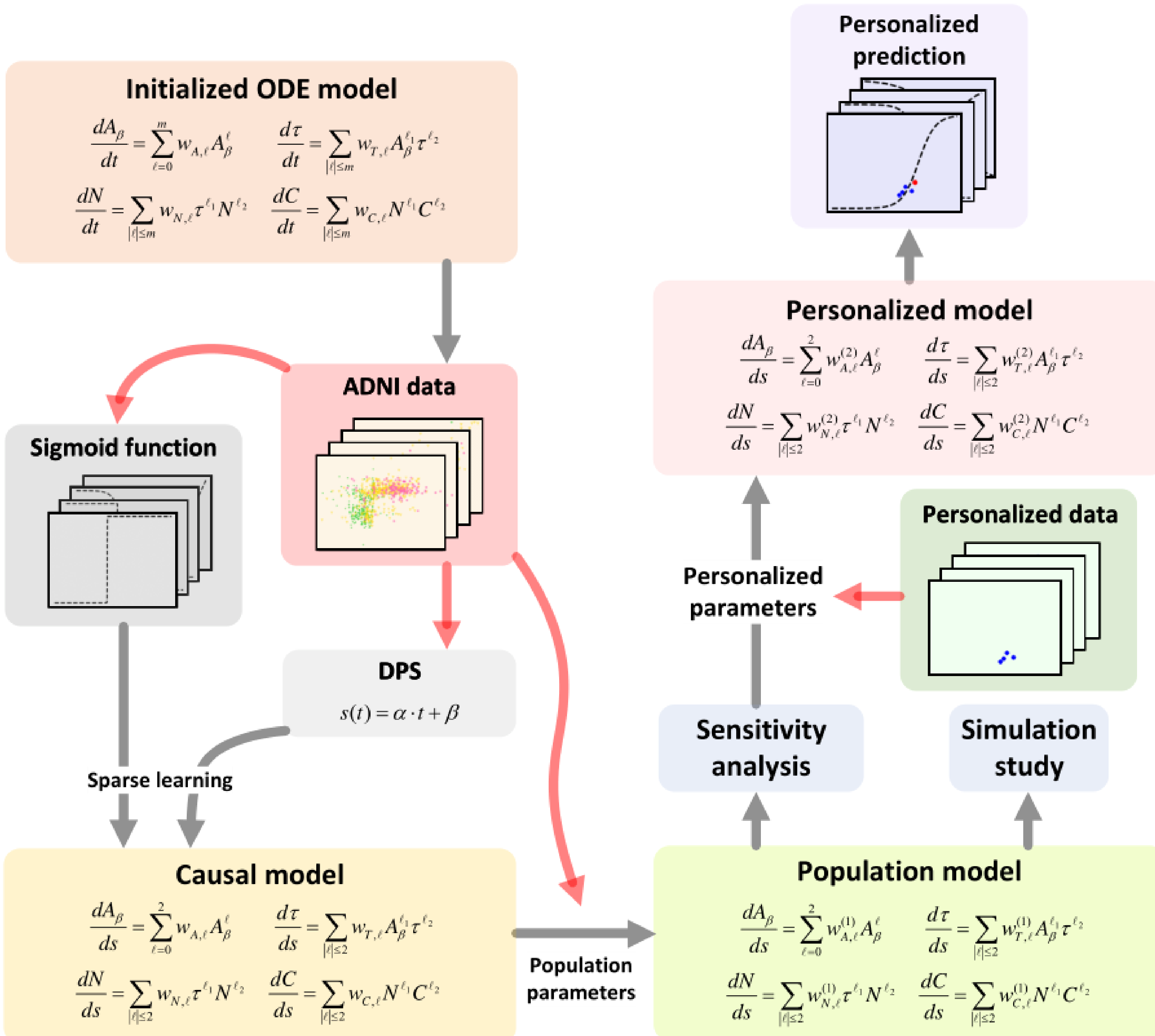
AD prediction

How to apply patient data (ADNI dataset) to optimize ODEs?





PURDUE UNIVERSITY | School of Mechanical Engineering


Initialized ODE model
$\frac{dA_\beta}{dt} = \sum_{\ell=0}^{m} w_{A,\ell} A_\beta^\ell$
$\frac{d\tau}{dt} = \sum_{|\ell|\le m} w_{T,\ell} A_\beta^{\ell_1} \tau^{\ell_2}$
$\frac{dN}{dt} = \sum_{|\ell|\le m} w_{N,\ell} \tau^{\ell_1} N^{\ell_2}$
$\frac{dC}{dt} = \sum_{|\ell|\le m} w_{C,\ell} N^{\ell_1} C^{\ell_2}$
ADNI data
Sigmoid function
DPS
$s(t) = \alpha \cdot t + \beta$
Sparse learning
Causal model
$\frac{dA_\beta}{ds} = \sum_{\ell=0}^{2} w_{A,\ell} A_\beta^\ell$
$\frac{d\tau}{ds} = \sum_{|\ell|\le 2} w_{T,\ell} A_\beta^{\ell_1} \tau^{\ell_2}$
$\frac{dN}{ds} = \sum_{|\ell|\le 2} w_{N,\ell} \tau^{\ell_1} N^{\ell_2}$
$\frac{dC}{ds} = \sum_{|\ell|\le 2} w_{C,\ell} N^{\ell_1} C^{\ell_2}$
Population parameters
Population model
$\frac{dA_\beta}{ds} = \sum_{\ell=0}^{2} w_{A,\ell}^{(1)} A_\beta^\ell$
$\frac{d\tau}{ds} = \sum_{|\ell|\le 2} w_{T,\ell}^{(1)} A_\beta^{\ell_1} \tau^{\ell_2}$
$\frac{dN}{ds} = \sum_{|\ell|\le 2} w_{N,\ell}^{(1)} \tau^{\ell_1} N^{\ell_2}$
$\frac{dC}{ds} = \sum_{|\ell|\le 2} w_{C,\ell}^{(1)} N^{\ell_1} C^{\ell_2}$
Sensitivity analysis
Simulation study
Personalized parameters
Personalized data
Personalized model
$\frac{dA_\beta}{ds} = \sum_{\ell=0}^{2} w_{A,\ell}^{(2)} A_\beta^\ell$
$\frac{d\tau}{ds} = \sum_{|\ell|\le 2} w_{T,\ell}^{(2)} A_\beta^{\ell_1} \tau^{\ell_2}$
$\frac{dN}{ds} = \sum_{|\ell|\le 2} w_{N,\ell}^{(2)} \tau^{\ell_1} N^{\ell_2}$
$\frac{dC}{ds} = \sum_{|\ell|\le 2} w_{C,\ell}^{(2)} N^{\ell_1} C^{\ell_2}$
Personalized prediction

Contribution

- 1. We build the first data-driven cascade model for the Alzheimer's disease
- 2. We design personalized model to calibrate dynamics for each patient and provide personalized treatment.

$$\begin{cases} \frac{dA_\beta}{ds} = w_{A0} + w_{A1}A_\beta + w_{A2}A_\beta^2; \\ \frac{d\tau}{ds} = w_{T0} + w_{T1}\tau + w_{T2}\tau^2 + w_{T3}A_\beta + w_{T4}A_\beta^2 + w_{T5}A_\beta\tau; \\ \frac{dN}{ds} = w_{N0} + w_{N1}N + w_{N2}N^2 + w_{N3}\tau + w_{N4}\tau^2 + w_{N5}\tau N; \\ \frac{dC}{ds} = w_{C0} + w_{C1}C + w_{C2}C^2 + w_{C3}N + w_{C4}N^2 + w_{C5}NC, \end{cases}$$

Zheng Haoyang, et al. "Data-driven causal model discovery and personalized prediction in Alzheimer's disease." NPJ digital medicine 2022

Population Model

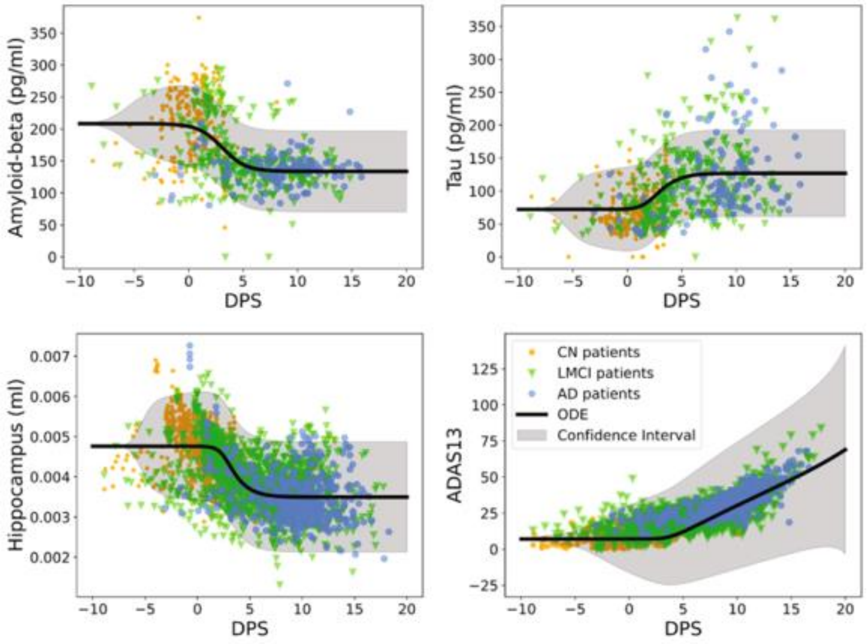
$$\begin{cases} \frac{dA_\beta}{ds} = w_{A0} + w_{A1}A_\beta + w_{A2}A_\beta^2; \\ \frac{d\tau}{ds} = w_{T0} + w_{T1}\tau + w_{T2}\tau^2 + w_{T3}A_\beta + w_{T4}A_\beta^2 + w_{T5}A_\beta\tau; \\ \frac{dN}{ds} = w_{N0} + w_{N1}N + w_{N2}N^2 + w_{N3}\tau + w_{N4}\tau^2 + w_{N5}\tau N; \\ \frac{dC}{ds} = w_{C0} + w_{C1}C + w_{C2}C^2 + w_{C3}N + w_{C4}N^2 + w_{C5}NC, \end{cases}$$

Table 1. Population parameters $w^{(1)}$ of the calibrated causal models based on the ADNI dataset.

Biomarkers	Parameters	Included subjects	
		CN, LMCI, AD	LMCI, AD
A_β	w_{A0}	0	0
	w_{A1}	0.917	0.745
	w_{A2}	−0.873	−0.749
τ	w_{T0}	0	0
	w_{T1}	0.788	0.689
	w_{T2}	−0.246	−0.679
	w_{T3}	0.002	0.000
	w_{T4}	3.066	0.185
	w_{T5}	−3.650	−0.101
N	w_{N0}	0	0
	w_{N1}	1.627	0.899
	w_{N2}	−1.253	−0.927
	w_{N3}	0.018	0.554
	w_{N4}	2.342	1.792
	w_{N5}	−4.015	−2.127
C	w_{C0}	0	0
	w_{C1}	0.159	0.134
	w_{C2}	0.202	−0.067
	w_{C3}	0.010	0.004
	w_{C4}	0.019	0.007
	w_{C5}	−0.176	−0.008

Zheng Haoyang, et al. "Data-driven causal model discovery and personalized prediction in Alzheimer's disease." NPJ digital medicine 2022

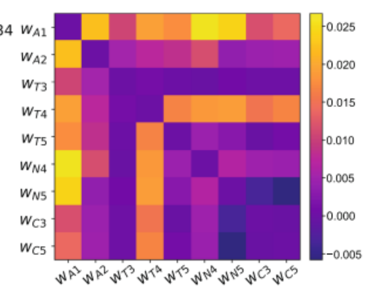
Population Model

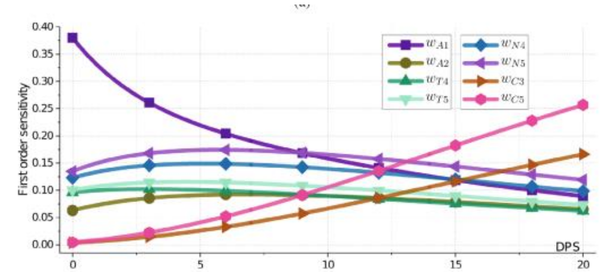


Zheng Haoyang, et al. "Data-driven causal model discovery and personalized prediction in Alzheimer's disease." NPJ digital medicine 2022

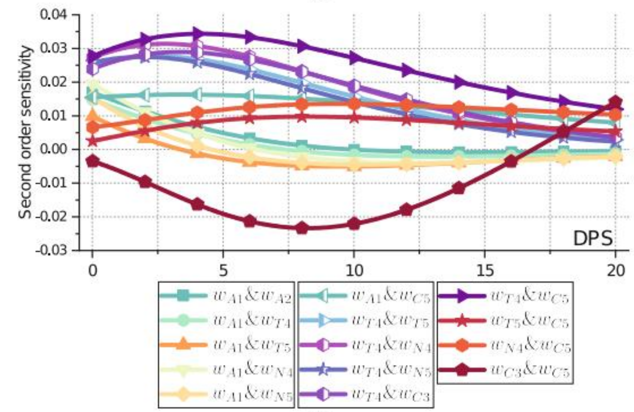
Sensitivity Analysis

$$\begin{cases} \frac{dA_\beta}{ds} = w_{A0} + w_{A1}A_\beta + w_{A2}A_\beta^2; \\ \frac{d\tau}{ds} = w_{T0} + w_{T1}\tau + w_{T2}\tau^2 + w_{T3}A_\beta + w_{T4}A_\beta^2 + w_{T5}A_\beta\tau; \\ \frac{dN}{ds} = w_{N0} + w_{N1}N + w_{N2}N^2 + w_{N3}\tau + w_{N4}\tau^2 + w_{N5}\tau N; \\ \frac{dC}{ds} = w_{C0} + w_{C1}C + w_{C2}C^2 + w_{C3}N + w_{C4}N^2 + w_{C5}NC, \end{cases}$$





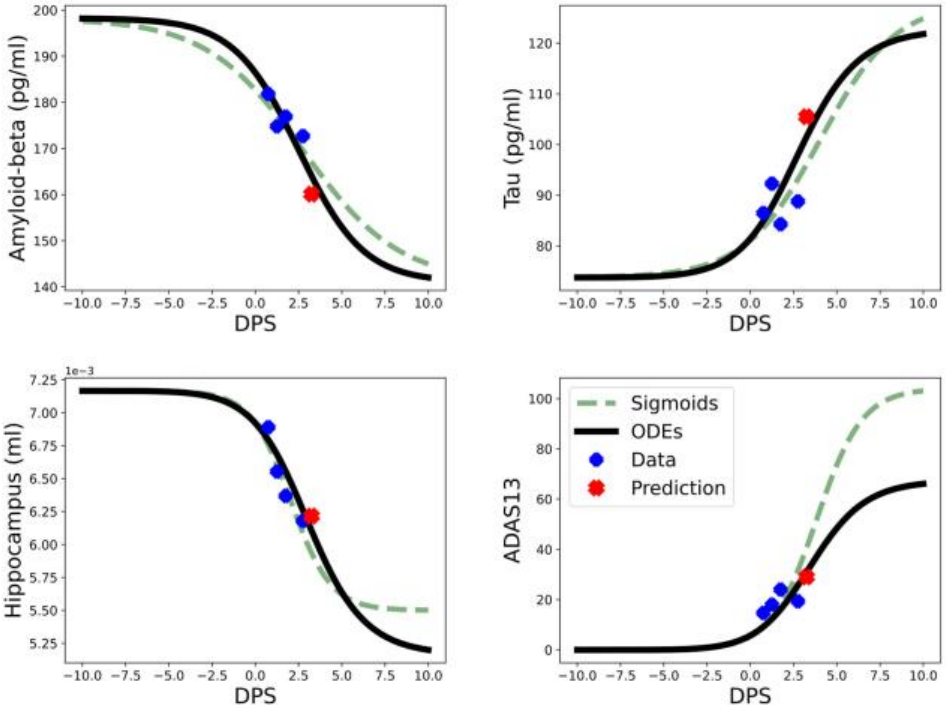
(b)



(c)

Zheng Haoyang, et al. "Data-driven causal model discovery and personalized prediction in Alzheimer's disease." NPJ digital medicine 2022

Personalized Treatment



Zheng Haoyang, et al. "Data-driven causal model discovery and personalized prediction in Alzheimer's disease." NPJ digital medicine 2022

Table 2. The prediction accuracy summary for CN subjects using different numbers of longitudinal biomarker datapoints (*n*).

PseudoIDs (n)	DPS Diff	Model	Accuracy			
			CSF Abeta42	CSF tTau	HIPPv	ADAS13
1 (4)	0.13	ODE	98.3%	93.6%	99.4%	92.6%
		Sigmoid	74.0%	79.8%	70.5%	84.8%
2 (4)	3.00	ODE	99.8%	93.2%	98.7%	93.0%
		Sigmoid	93.9%	61.5%	90.7%	80.4%
3 (5)	0.52	ODE	86.6%	98.8%	95.9%	85.3%
		Sigmoid	90.3%	82.6%	71.1%	56.9%
4 (5)	0.59	ODE	98.8%	96.1%	88.3%	96.6%
		Sigmoid	76.8%	76.9%	86.1%	66.7%
5 (5)	0.39	ODE	97.8%	90.0%	99.7%	94.8%
		Sigmoid	84.3%	79.5%	79.9%	81.9%
6 (4)	0.46	ODE	96.3%	93.6%	90.9%	92.7%
		Sigmoid	75.4%	91.1%	91.2%	84.0%
7 (4)	0.55	ODE	99.8%	88.2%	98.7%	90.3%
		Sigmoid	96.5%	86.0%	92.0%	72.3%
8 (4)	0.63	ODE	95.9%	98.9%	92.0%	92.6%
		Sigmoid	85.8%	86.8%	91.7%	96.6%
9 (4)	0.71	ODE	99.6%	96.1%	97.1%	87.5%
		Sigmoid	89.4%	80.3%	79.2%	69.5%
10 (5)	1.04	ODE	83.4%	81.2%	98.7%	85.5%
		Sigmoid	88.3%	78.4%	74.4%	80.1%
11 (6)	1.04	ODE	98.2%	99.8%	86.5%	85.1%
		Sigmoid	75.7%	76.3%	67.6%	72.6 %
12 (4)	0.40	ODE	94.6%	91.3%	96.5%	91.7%
		Sigmoid	89.7%	81.5%	88.9%	75.1%
13 (6)	0.88	ODE	97.0%	92.8%	96.1%	98.8%
		Sigmoid	97.4%	85.4%	85.3%	84.3%
14 (4)	0.75	ODE	98.4%	99.1%	99.1%	87.1%
		Sigmoid	90.9%	79.7%	88.6%	79.8%
15 (4)	0.55	ODE	99.6%	96.8%	90.9%	81.5%
		Sigmoid	93.8%	95.1%	81.5%	59.2%
Average	0.78 ± 0.64	ODE	96.3% ± 4.9%	94.0% ± 5.0%	95.2% ± 4.4%	90.3% ± 4.8%
		Sigmoid	86.8% ± 7.6%	81.4% ± 7.4%	82.6% ± 8.2%	76.3% ± 10.1%

Our model

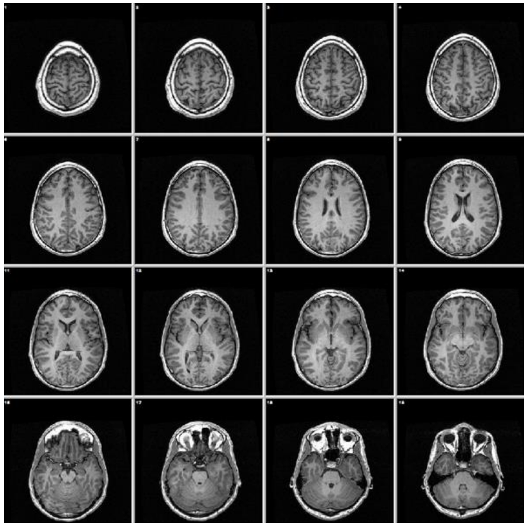
Previous model

Summary

- We learn population model to describe the population dynamics and distinguish patients at different stages
- Sensitivity analysis determine the sensitive parameters, which are calibrated in personalized model
- Personalized model calibrate the parameter to better demonstrate dynamics tailored for each patient

Future Plans

- Data-driven model discovery with spatial-temporal measurements
- Current CSF data summarize from MRI scans to estimate levels of biomarkers
- We hope to directly use MRI scans to learn the diffusion process of biomarkers



PURDUE UNIVERSITY | School of Mechanical Engineering

Outline:

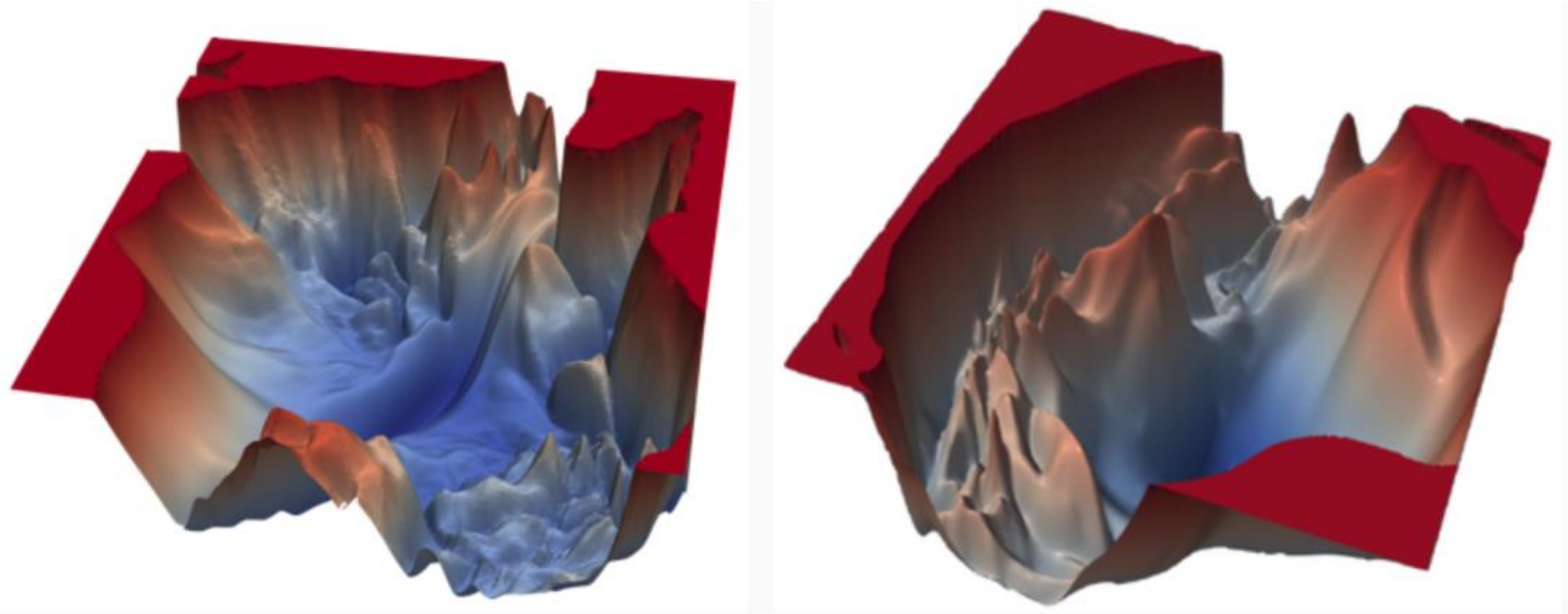
- Incorporate Physics Knowledge and AI to design new interpretable models – Trustworthy Epidemiological Models for COVID-19 Prediction & Intervention

- Incorporate Physics Knowledge into AI to predict multiscale problems: NH-PINN

- Interpretable AI enables data-driven scientific discovery with uncertainty quantification capability – ALZHEIMER's Disease Prediction

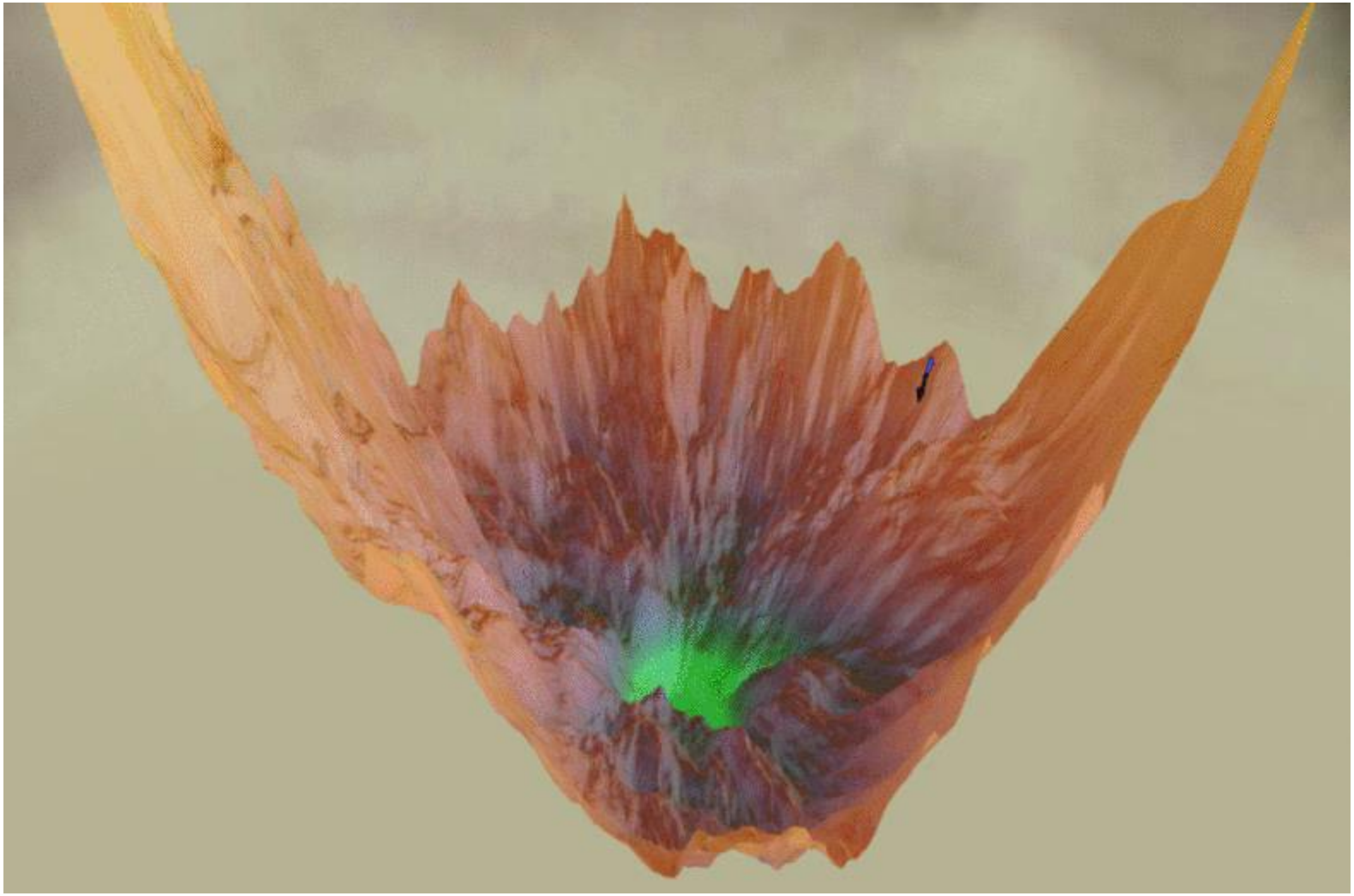
- **Scalable training large-scale Deep Neural Network**

Visualization the Loss Landscape of Deep Neural Nets



The loss landscape of modern deep neural nets [Li et al., 2018]

Gradient Descent Fails



Credit to losslanscape.com

Scalable training large-scale Deep Neural Network:

Question: How can we design efficient optimization/sampling algorithms to train large-scale deep neural networks?

Goal: Enable Fast training large-scale DNN.

W. Deng, X. Zhang, F. Liang, **G. Lin**, An adaptive empirical Bayesian method for sparse deep learning, **2019 Conference on Neural Information Processing Systems (NIPS)**, Dec. 8 – Dec. 14, 2019, Vancouver, Canada.

NeurIPS'19, NeurIPS'20, ICML'20, ICLR'21, JCP'20, JCP'21a, JCP'21b

Scalable algorithms for Bayesian deep learning via Stochastic Gradient Monte Carlo and Beyond

Guang Lin [1]
Joint work with **W. Deng, Y. Wang, Q. Feng, L. Gao, G. Karagiannis, F. Liang**
August 13, 2021
[1]Departments of Mathematics & School of Mechanical Engineering, Purdue University

NeurIPS'19, NeurIPS'20, ICML'20, ICLR'21, JCP'20, JCP'21a, JCP'21b

Markov chain Monte Carlo

Uncertainty quantification is crucial for AI safety problems and reinforcement learning, which draws our attention to **Markov chain Monte Carlo (MCMC)**, which is known for

- Multi-modal sampling $\rightarrow$ Accurate predictive **confidence interval**
- Non-convex optimization $\rightarrow$ Better **point estimate**

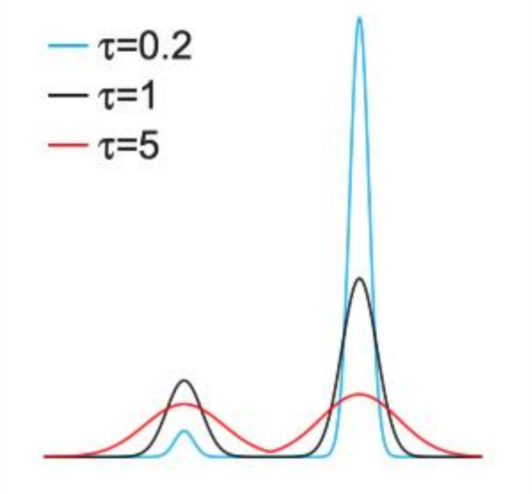
Langevin diffusion

A famous sampling algorithm is called Langevin diffusion.

$$d\boldsymbol{\beta}_t = -\nabla U(\boldsymbol{\beta}_t)dt + \sqrt{2\tau}d\boldsymbol{W}_t,$$

where $\boldsymbol{\beta}_t$ is the parameter at time t, $U(\cdot)$ is the energy function, $\boldsymbol{W}_t$ is a Brownian motion and τ is the temperature.

As $t \to \infty$, $\boldsymbol{\beta}_t$ converges to the stationary Gibbs distribution $Ce^{-\frac{U(\boldsymbol{\beta})}{\tau}}$.



(a) Gibbs measures at three different temperatures τ.

Stochastic gradient Langevin dynamics

However, evaluating gradient in big data problems is **too costly**.

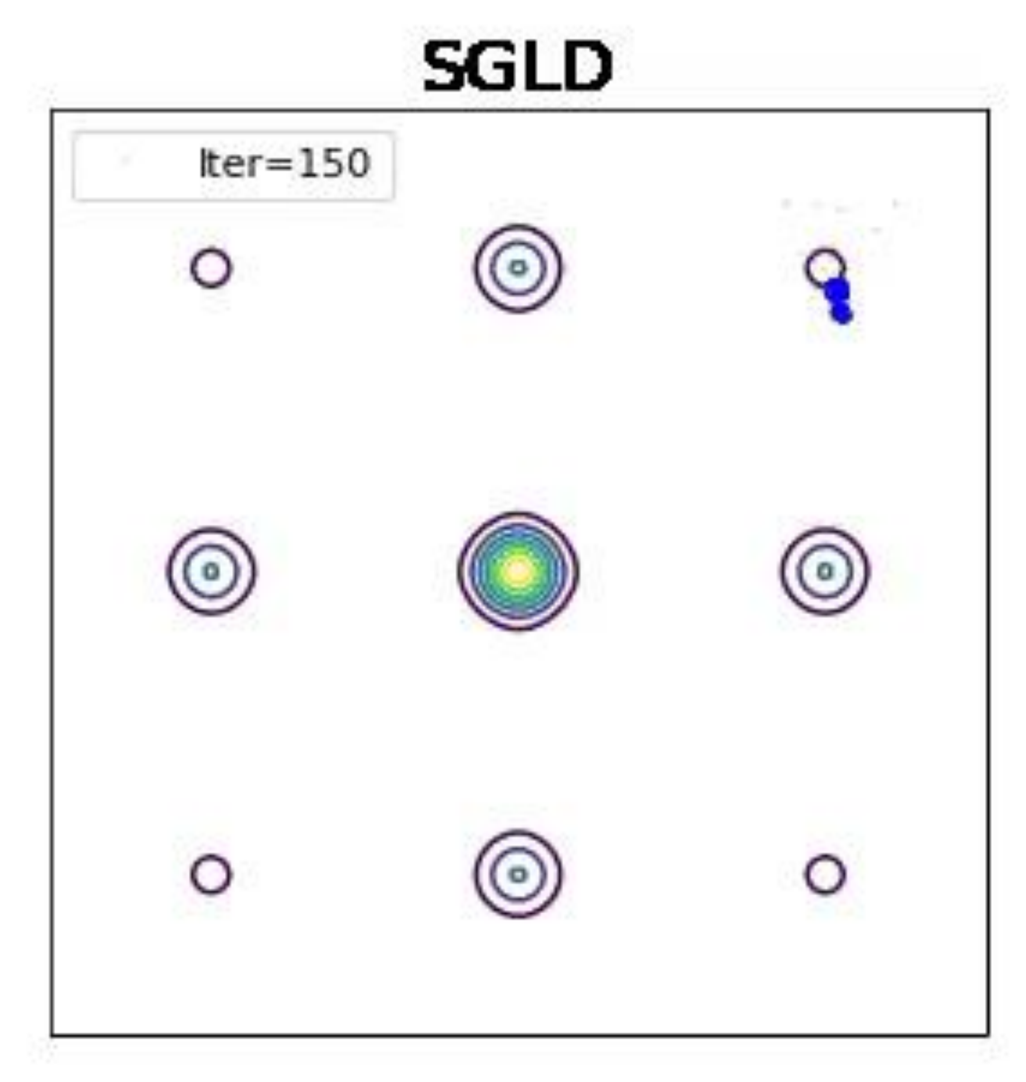
To tackle this issue, Max Welling, etc [Welling and Teh, 2011] proposed the stochastic gradient Langevin dynamics algorithm (SGLD)

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k - \eta \nabla \widetilde{U}(\boldsymbol{\beta}_k) + \mathcal{N}(0, 2\eta\tau \boldsymbol{I}). \tag{1}$$

As $t \to \infty$ and $\eta \to 0$, $\boldsymbol{\beta}_t$ converges weakly to the stationary Gibbs distribution $Ce^{-\frac{U(\beta)}{\tau}}$.

Stochastic gradient Langevin dynamics

Sample from a multi-modal distribution



Acceleration strategies for MCMC

Most popular strategies to *accelerate* MCMC:

- Simulated annealing [Kirkpatrick et al., 1983]
- **Replica exchange MCMC [Swendsen and Wang, 1986]**

Replica Exchange SGLD
Wei Deng, et al., ICML 2020

Replica exchange Langevin diffusion

Consider two Langevin diffusion processes with $\tau_1 > \tau_2$

$$d\boldsymbol{\beta}_t^{(1)} = -\nabla U(\boldsymbol{\beta}_t^{(1)})dt + \sqrt{2\tau_1}d\boldsymbol{W}_t^{(1)}$$
$$d\boldsymbol{\beta}_t^{(2)} = -\nabla U(\boldsymbol{\beta}_t^{(2)})dt + \sqrt{2\tau_2}d\boldsymbol{W}_t^{(2)},$$

Moreover, the positions of the two particles swap with a probability

$$S(\boldsymbol{\beta}_t^{(1)}, \boldsymbol{\beta}_t^{(2)}) := e^{\left(\frac{1}{\tau_1}-\frac{1}{\tau_2}\right)\left(U(\boldsymbol{\beta}_t^{(1)})-U(\boldsymbol{\beta}_t^{(2)})\right)}$$

In other words, a jump process is included in a Markov process

$$\mathbb{P}(\boldsymbol{\beta}_{t+dt} = (\boldsymbol{\beta}_t^{(2)}, \boldsymbol{\beta}_t^{(1)}) | \boldsymbol{\beta}_t = (\boldsymbol{\beta}_t^{(1)}, \boldsymbol{\beta}_t^{(2)})) = rS(\boldsymbol{\beta}_t^{(1)}, \boldsymbol{\beta}_t^{(2)})dt$$
$$\mathbb{P}(\boldsymbol{\beta}_{t+dt} = (\boldsymbol{\beta}_t^{(1)}, \boldsymbol{\beta}_t^{(2)}) | \boldsymbol{\beta}_t = (\boldsymbol{\beta}_t^{(1)}, \boldsymbol{\beta}_t^{(2)})) = 1 - rS(\boldsymbol{\beta}_t^{(1)}, \boldsymbol{\beta}_t^{(2)})dt$$

A demo

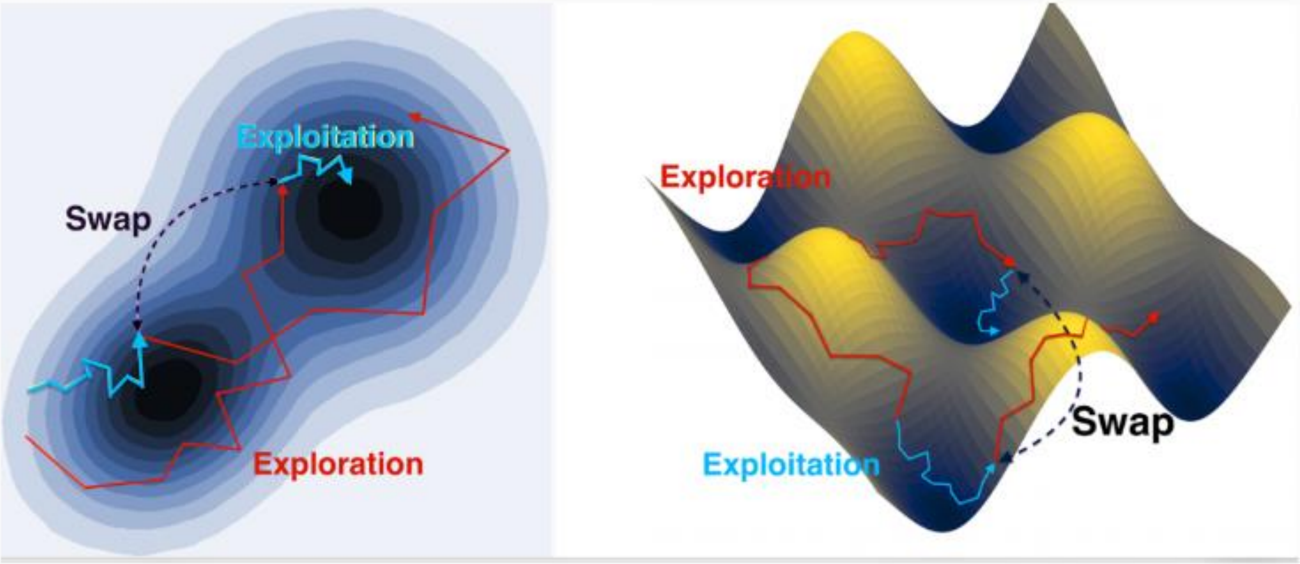


Figure 1: Trajectory plot for replica exchange Langevin diffusion.

Why the naïve numerical algorithm fails

Consider the scalable stochastic gradient Langevin dynamics algorithm [Welling and Teh, 2011]

$$\widetilde{\boldsymbol{\beta}}^{(1)}_{k+1} = \widetilde{\boldsymbol{\beta}}^{(1)}_{k} - \eta_k \nabla \widetilde{L}(\widetilde{\boldsymbol{\beta}}^{(1)}_{k}) + \sqrt{2\eta_k \tau_1}\boldsymbol{\xi}^{(1)}_{k}$$
$$\widetilde{\boldsymbol{\beta}}^{(2)}_{k+1} = \widetilde{\boldsymbol{\beta}}^{(2)}_{k} - \eta_k \nabla \widetilde{L}(\widetilde{\boldsymbol{\beta}}^{(2)}_{k}) + \sqrt{2\eta_k \tau_2}\boldsymbol{\xi}^{(2)}_{k}.$$

Swap the chains with a **naïve** swapping rate $r\mathbb{S}(\widetilde{\boldsymbol{\beta}}^{(1)}_{k+1}, \widetilde{\boldsymbol{\beta}}^{(2)}_{k+1})\eta_k$[§]:

$$\mathbb{S}(\widetilde{\boldsymbol{\beta}}^{(1)}_{k+1}, \widetilde{\boldsymbol{\beta}}^{(2)}_{k+1}) = e^{\left(\frac{1}{\tau_1} - \frac{1}{\tau_2}\right)\left(\widetilde{L}(\widetilde{\boldsymbol{\beta}}^{(1)}_{k+1}) - \widetilde{L}(\widetilde{\boldsymbol{\beta}}^{(2)}_{k+1})\right)}. \tag{2}$$

Exponentiating the unbiased estimators $\widetilde{L}(\widetilde{\boldsymbol{\beta}}^{(\cdot)}_{k+1})$ leads to a **large bias**.

[§] In the implementations, we fix $r\eta_k = 1$ by default.

A corrected algorithm

Assume $\widetilde{L}(\boldsymbol{\theta}) \sim \mathcal{N}(L(\boldsymbol{\theta}), \sigma^2)$ and consider the **geometric Brownian motion** of $\{\widetilde{S}_t\}_{t\in[0,1]}$ in each swap as a Martingale

$$\begin{aligned}\widetilde{S}_t &= e^{\left(\frac{1}{\tau_1}-\frac{1}{\tau_2}\right)\left(\widetilde{L}(\widetilde{\boldsymbol{\beta}}^{(1)})-\widetilde{L}(\widetilde{\boldsymbol{\beta}}^{(2)})-\left(\frac{1}{\tau_1}-\frac{1}{\tau_2}\right)\sigma^2 t\right)} \\ &= e^{\left(\frac{1}{\tau_1}-\frac{1}{\tau_2}\right)\left(L(\widetilde{\boldsymbol{\beta}}^{(1)})-L(\widetilde{\boldsymbol{\beta}}^{(2)})-\left(\frac{1}{\tau_1}-\frac{1}{\tau_2}\right)\sigma^2 t+\sqrt{2}\sigma W_t\right)}.\end{aligned} \tag{3}$$

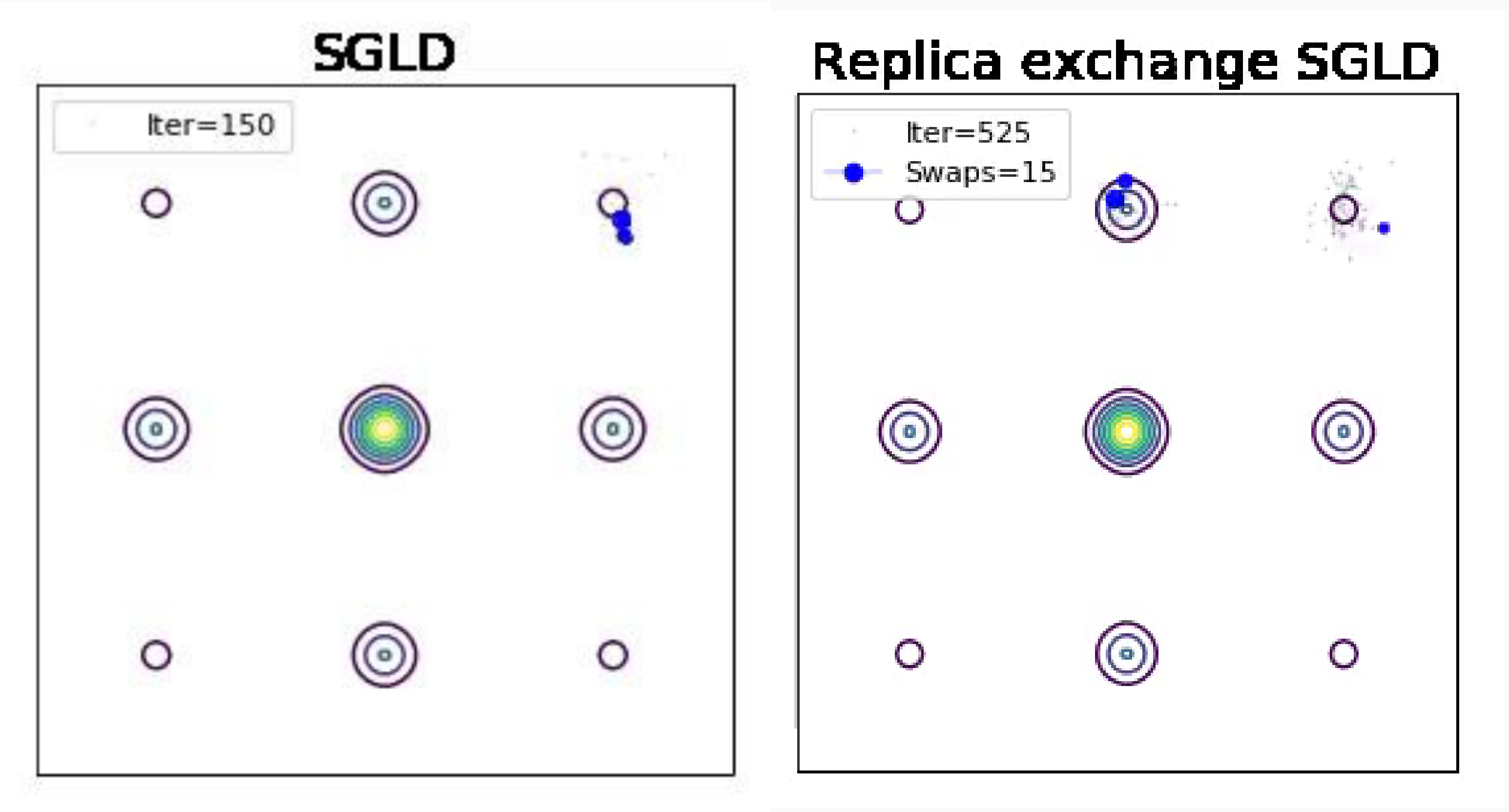
Taking the derivative of $\widetilde{S}_t$ with respect to t and W_t, Itô's lemma gives,

$$d\widetilde{S}_t = \left(\frac{d\widetilde{S}_t}{dt} + \frac{1}{2}\frac{d^2\widetilde{S}_t}{dW_t^2}\right)dt + \frac{d\widetilde{S}_t}{dW_t}dW_t = \sqrt{2}\left(\frac{1}{\tau_1}-\frac{1}{\tau_2}\right)\sigma\widetilde{S}_t dW_t.$$

By fixing $t = 1$ in (3), we have the **suggested unbiased swapping rate**

$$\widetilde{S}_1 = e^{\left(\frac{1}{\tau_1}-\frac{1}{\tau_2}\right)\left(\widetilde{L}(\widetilde{\boldsymbol{\beta}}^{(1)})-\widetilde{L}(\widetilde{\boldsymbol{\beta}}^{(2)})-\left(\frac{1}{\tau_1}-\frac{1}{\tau_2}\right)\boldsymbol{\sigma}^2\right)}.$$

Replica exchange Stochastic gradient Langevin dynamics



Acceleration via replica exchange

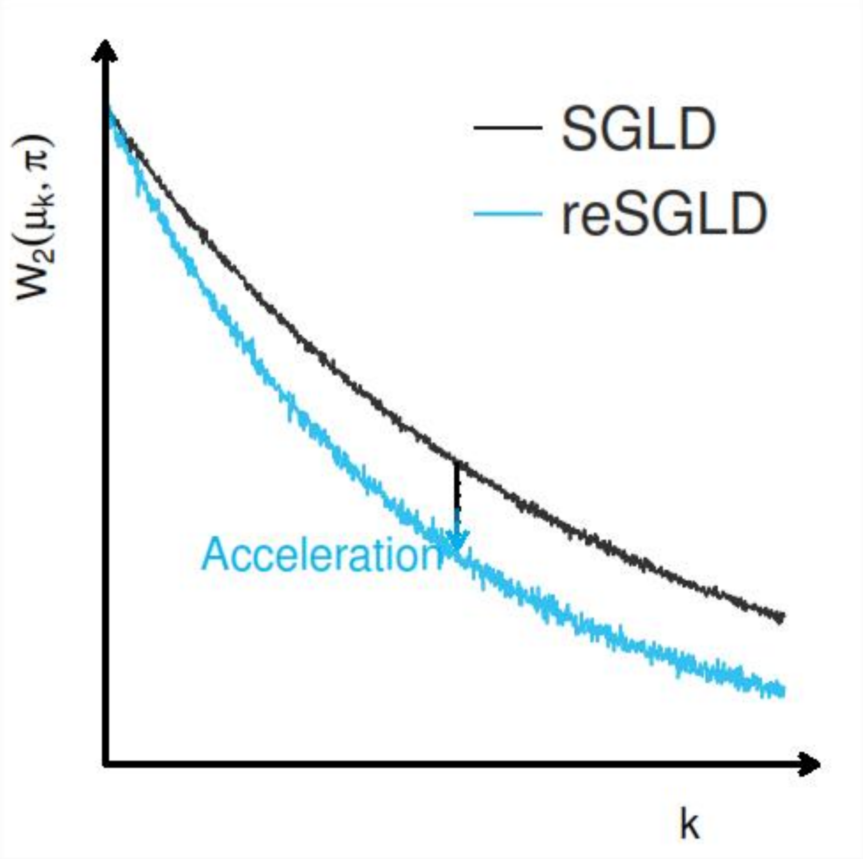


Figure 2: Acceleration via replica exchange (swaps/ interactions)

Accelerating convergence via variance reduction

Can we do better?

Exponential acceleration via variance reduction
Wei Deng et al., ICLR 2021

Accelerating convergence via variance reduction

The desire to obtain more effective swaps drives us to design more efficient energy estimators.

To reduce the variance of the noisy energy estimator $L(B|\boldsymbol{\beta}^{(h)}) = \frac{N}{n}\sum_{i\in B} L(\boldsymbol{x}_i|\boldsymbol{\beta}^{(h)})$ for $h \in \{1,2\}$, we consider an unbiased estimator $L(B|\widehat{\boldsymbol{\beta}}^{(h)})$ for $\sum_{i=1}^{N} L(\boldsymbol{x}_i|\widehat{\boldsymbol{\beta}}^{(h)})$ and a constant c, we see that a new estimator $\widetilde{L}(B|\boldsymbol{\beta}^{(h)})$, which follows

$$\widetilde{L}(B|\boldsymbol{\beta}^{(h)}) = L(B|\boldsymbol{\beta}^{(h)}) + c\left(L(B|\widehat{\boldsymbol{\beta}}^{(h)}) - \sum_{i=1}^{N} L(\boldsymbol{x}_i|\widehat{\boldsymbol{\beta}}^{(h)})\right), \qquad (4)$$

is still the unbiased estimator for $\sum_{i=1}^{N} L(\boldsymbol{x}_i|\boldsymbol{\beta}^{(h)})$.

Accelerating convergence via variance reduction

By decomposing the variance, we have

$$Var(\widetilde{L}(B|\beta^{(h)})) = Var\left(L(B|\beta^{(h)})\right) + c^2 Var\left(L(B|\widehat{\beta}^{(h)})\right) + 2c\text{Cov}\left(L(B|\beta^{(h)}), L(B|\widehat{\beta}^{(h)})\right).$$

In such a case, $Var(\widetilde{L}(B|\beta^{(h)}))$ achieves the minimum variance $(1-\rho^2)Var(L(B|\beta^{(h)}))$ given $c^\star := \frac{-\text{Cov}(L(B|\beta^{(h)}), L(B|\widehat{\beta}^{(h)}))}{Var(L(B|\beta^{(h)}))}$, where $\text{Cov}(\cdot,\cdot)$ denotes the covariance and ρ is the correlation coefficient.

Accelerating convergence via variance reduction

To make variance reduction work, it requires two crucial components.

- To propose a correlated control variate $\widehat{\boldsymbol{\beta}}$
  $\rightarrow$ Update $\widehat{\boldsymbol{\beta}}^{(h)} = \boldsymbol{\beta}^{(h)}_{m\lfloor \frac{k}{m} \rfloor}$ every m iterations
- The optimal c is unknown.
  $\rightarrow$ Set $c = -1$ for highly correlated energy estimators.
  $\rightarrow$ Set adaptive c for the less correlated.

Reduction of Variance

VR-reSGLD may lead to a more efficient energy estimator with a much smaller variance.

Lemma (Variance-reduced energy estimator)

Under the smoothness and dissipativity assumptions, the variance of the variance-reduced energy estimator $\widetilde{L}(B|\boldsymbol{\beta}^{(h)})$*, where* $h \in \{1, 2\}$*, is upper bounded by*

$$Var\left(\widetilde{L}(B|\boldsymbol{\beta}^{(h)})\right) \leq \min\left\{\mathcal{O}\left(\frac{m^2\eta}{n}\right), Var\left(\frac{N}{n}\sum_{i \in B} L(\mathbf{x}_i|\boldsymbol{\beta}^{(1)})\right)\right\},$$

where the detailed $\mathcal{O}(\cdot)$ *constants is shown in the appendix [Deng et al., 2021].*

A smaller variance implies more effective swaps

The variance-reduced energy estimator $\widetilde{L}(B|\beta^{(h)})$ doesn't directly affect $\mathbb{E}[\widetilde{S}_{\eta,m,n}]$ within the support $[0,\infty]$. However, the unbounded support is not appropriate for numerical algorithms, and only the truncated swapping rate $S_{\eta,m,n} = \min\{1, \widetilde{S}_{\eta,m,n}\}$ is considered. As such, the truncated swapping rate becomes significantly smaller.

Lemma (Variance reduction for larger swapping rates)

Given a large enough batch size n, the variance-reduced energy estimator $\widetilde{L}(B_k|\beta_k^{(h)})$ yields a truncated swapping rate that satisfies

$$\mathbb{E}[S_{\eta,m,n}] \approx \min\left\{1, S(\beta^{(1)}, \beta^{(2)})\left(\mathcal{O}\left(\frac{1}{n^2}\right) + e^{-\mathcal{O}\left(\frac{m^2\eta}{n} + \frac{1}{n^2}\right)}\right)\right\}. \quad (5)$$

Acceleration via variance-reduced replica exchange

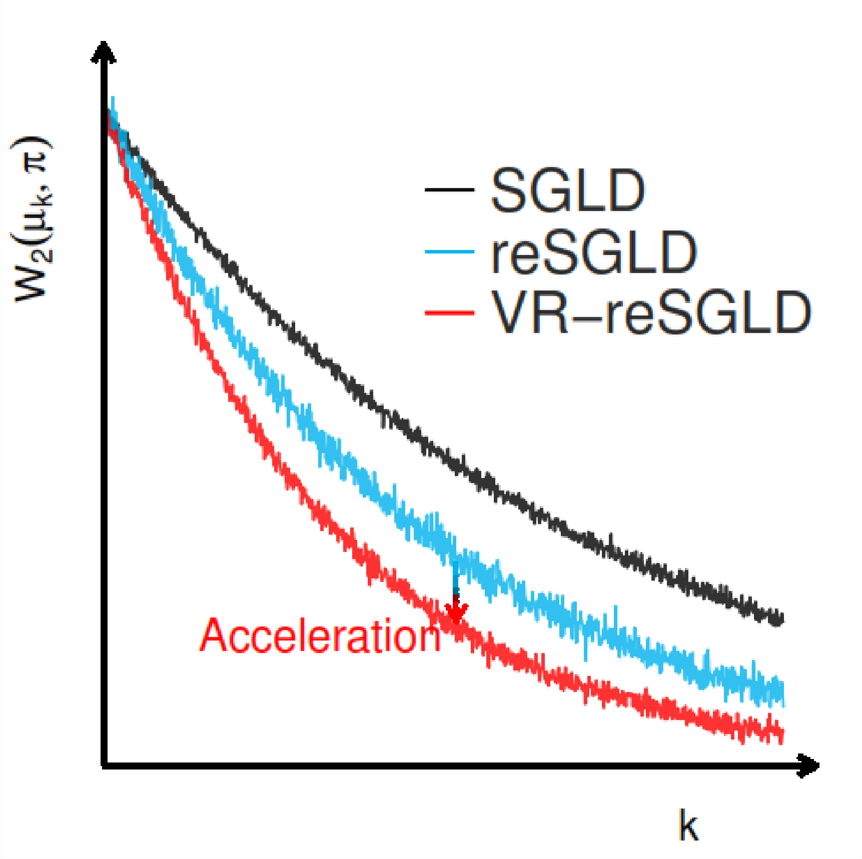
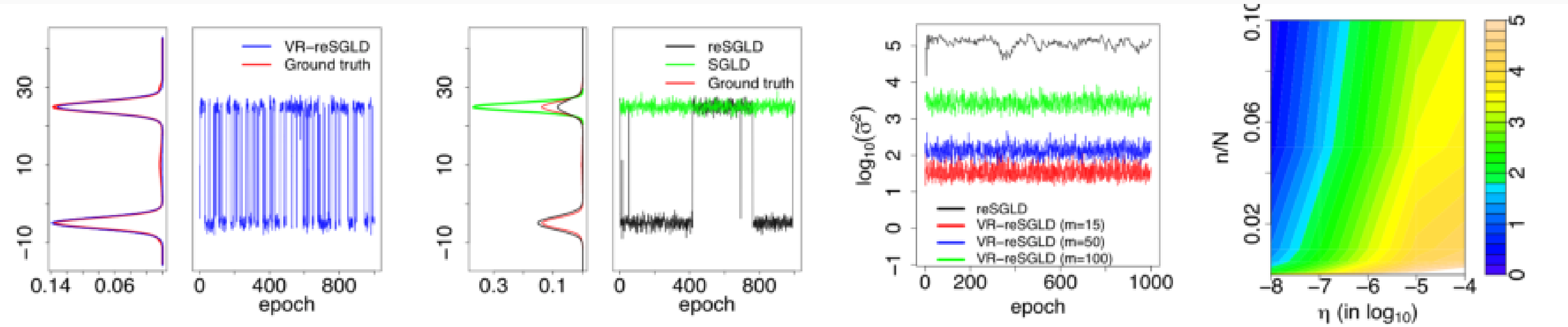


Figure 3: Acceleration via variance-reduced replica exchange.

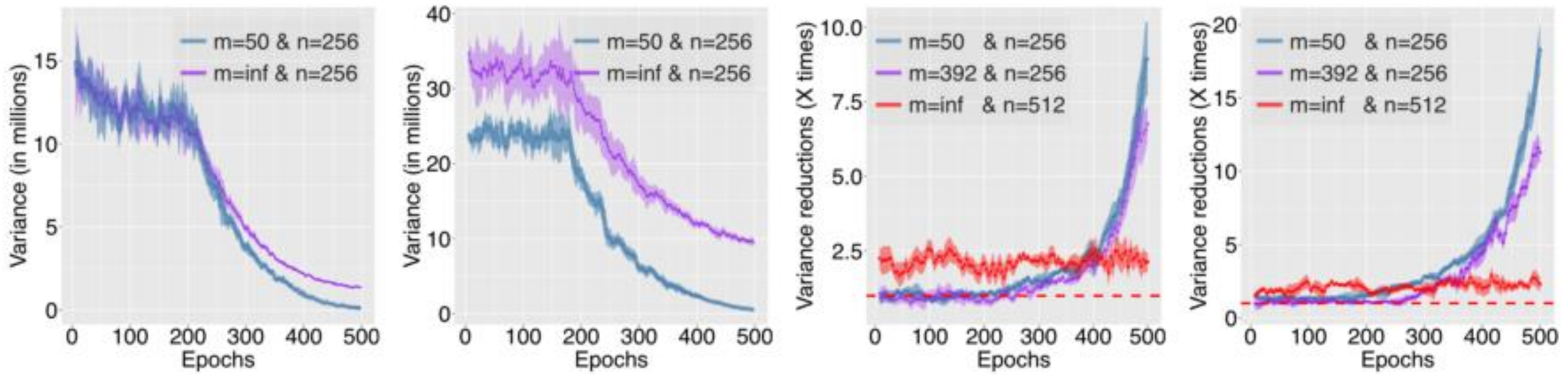
1D simulation of Gaussian mixture



(a) Trace plot for $\boldsymbol{\beta}^{(1)}$ (b) Trace plot for $\boldsymbol{\beta}^{(1)}$ (c) Paths of $\log_{10} \widetilde{\sigma}^2$ (d) Contour of $\log_{10} \widetilde{\sigma}^2$

Figure 2: Trace plots, KDEs of $\boldsymbol{\beta}^{(1)}$, and sensitivity study of $\widetilde{\sigma}^2$ with respect to m, η and n.

Non-convex optimization on CIFAR10 and CIFAR100



(a) CIFAR10: Original v.s. proposed (m=50)
(b) CIFAR100: Original v.s. proposed (m=50)
(c) Variance reduction setups on CIFAR10
(d) Variance reduction setups on CIFAR100

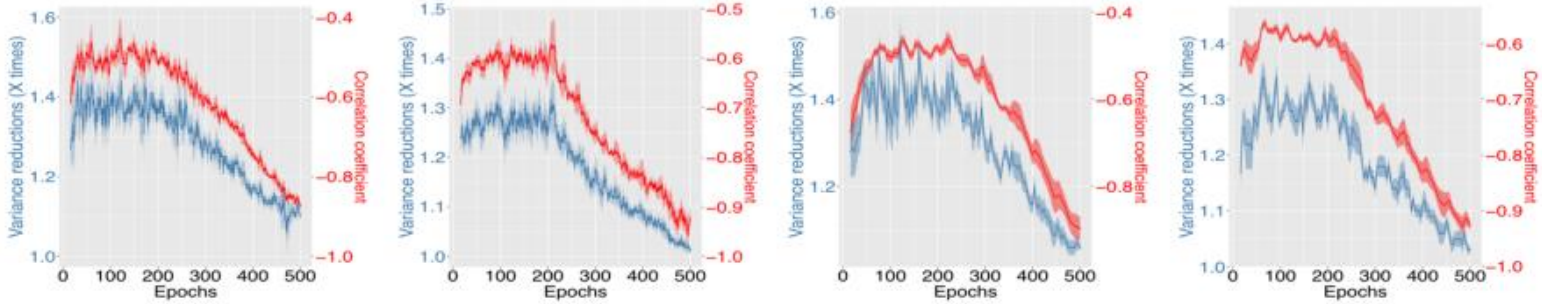
Figure 3: Variance reduction on the noisy energy estimators on CIFAR10 & CIFAR100 datasets.

Non-convex optimization on CIFAR10 and CIFAR100

TABLE 1: PREDICTION ACCURACIES (%) BASED ON BAYESIAN MODEL AVERAGING.

METHOD	CIFAR10			CIFAR100		
	RESNET20	RESNET32	RESNET56	RESNET20	RESNET32	RESNET56
M-SGD	94.07±0.11	95.11±0.07	96.05±0.21	71.93±0.13	74.65±0.20	78.76±0.24
SGHMC	94.16±0.13	95.17±0.08	96.04±0.18	72.09±0.14	74.80±0.19	78.95±0.22
reSGHMC	94.56±0.23	95.44±0.16	96.15±0.17	73.94±0.34	76.38±0.23	79.86±0.26
VR-reSGHMC	**94.84±0.11**	**95.62±0.09**	**96.32±0.15**	**74.83±0.18**	**77.40±0.27**	**80.62±0.22**
cycSGHMC	94.61±0.15	95.56±0.12	96.19±0.17	74.21±0.22	76.60±0.25	80.39±0.21
cVR-reSGHMC	**94.91±0.10**	**95.64±0.13**	**96.36±0.16**	**75.02±0.19**	**77.58±0.21**	**80.50±0.25**

Non-convex optimization on CIFAR10 and CIFAR100



(a) CIFAR10 & m=50 (b) CIFAR100 & m=50 (c) CIFAR10 & m=392 (d) CIFAR100 & m=392

Figure 5: A study of variance reduction techniques using adaptive coefficient and non-adaptive coefficient on CIFAR10 & CIFAR100 datasets.

Summary

- Replica exchange stochastic gradient MCMC shows a potential in exponentially accelerating the convergence in non-convex learning. [Deng et al., 2020]
- Variance reduction of energy estimators yields exponential more effective swaps, which further accelerates the exponential convergence in non-convex learning. [Deng et al., 2021]
- This is the first work to do variance reduction on energy estimators in deep learning, which paves the road for accelerating advanced stochastic gradient MCMC algorithms in non-convex learning.

References i

Deng, W., Feng, Q., Gao, L., Liang, F., and Lin, G. (2020).
Non-Convex Learning via Replica Exchange Stochastic Gradient MCMC.
In *Proc. of the International Conference on Machine Learning (ICML)*.

Deng, W., Feng, Q., Karagiannis, G., Lin, G., and Liang, F. (2021).
Accelerating Convergence of Replica Exchange Stochastic Gradient MCMC via Variance Reduction.
In *Proc. of the International Conference on Learning Representation (ICLR)*.

Kirkpatrick, S., Jr, D. G., and Vecchi, M. P. (1983).
Optimization by Simulated Annealing.
Science, 220(4598):671–680.

In Math We Trust: Interpretable, Trustworthy Machine Learning

"…Because I had worked in the closest possible ways with physicists and engineers, I knew that our data can never be precise…"

Norbert Wiener