Machine-learning of model error in dynamical systems

Matthew E. Levine[1] Andrew M. Stuart[1]

[1]Department of Computing and Mathematical Sciences
California Institute of Technology

Workshop on Dynamics and Data Assimilation, Physiology and Bioinformatics
Banff International Research Station
June 1, 2022

Caltech

Introduction

- Machine learning works (with enough data)!

Introduction

- Machine learning works (with enough data)!
- Mechanistic models based on physics work (with enough knowledge and compute)!

Introduction

- Machine learning works (with enough data)!
- Mechanistic models based on physics work (with enough knowledge and compute)!
- In most open prediction problems, we have SOME data and SOME prior knowledge.

Caltech

Introduction

- Machine learning works (with enough data)!
- Mechanistic models based on physics work (with enough knowledge and compute)!
- In most open prediction problems, we have SOME data and SOME prior knowledge.
- The next generation of high-performing prediction models will **hybridize physics-based and data-driven modeling techniques**
- How can we help lay the groundwork for this future?

Caltech

Our problem

True system (ODE):

$$\begin{aligned} \dot{x} &= f^{\dagger}(x, y) \\ \dot{y} &= \frac{1}{\varepsilon} g^{\dagger}(x, y) \end{aligned} \tag{1}$$

- **Relevance:** across disciplines (climatology, physiology, celestial mechanics, etc.).

Caltech

Our problem

True system (ODE):

$$\begin{aligned} \dot{x} &= f^{\dagger}(x, y) \\ \dot{y} &= \frac{1}{\varepsilon} g^{\dagger}(x, y) \end{aligned} \tag{1}$$

- **Relevance:** across disciplines (climatology, physiology, celestial mechanics, etc.).
- **Goal:** Given noisy observations of x, learn predictive model for future x dynamics.

Caltech

Our problem

True system (ODE):

$$\begin{aligned} \dot{x} &= f^\dagger(x, y) \\ \dot{y} &= \frac{1}{\varepsilon} g^\dagger(x, y) \end{aligned} \tag{1}$$

- **Relevance:** across disciplines (climatology, physiology, celestial mechanics, etc.).
- **Goal:** Given noisy observations of x, learn predictive model for future x dynamics.
- **Methodological constraints:**
  - Partial, noisy observations (e.g. observe x, but not y)

Caltech

Our problem

True system (ODE):

$$\dot{x} = f^\dagger(x, y)$$
$$\dot{y} = \frac{1}{\varepsilon} g^\dagger(x, y) \qquad (1)$$

- **Relevance:** across disciplines (climatology, physiology, celestial mechanics, etc.).
- **Goal:** Given noisy observations of x, learn predictive model for future x dynamics.
- **Methodological constraints:**
  - Partial, noisy observations (e.g. observe x, but not y)
  - No knowledge of y, $g^\dagger, \varepsilon$, nor $\dim(y)$

Caltech

Our problem

True system (ODE):

$$\dot{x} = f^{\dagger}(x, y)$$
$$\dot{y} = \frac{1}{\varepsilon} g^{\dagger}(x, y) \tag{1}$$

- **Relevance:** across disciplines (climatology, physiology, celestial mechanics, etc.).
- **Goal:** Given noisy observations of x, learn predictive model for future x dynamics.
- **Methodological constraints:**
  - Partial, noisy observations (e.g. observe x, but not y)
  - No knowledge of y, $g^{\dagger}, \varepsilon$, nor $\dim(y)$
  - Observations may be irregularly spaced and noisy

Caltech

Our problem

True system (ODE):

$$\dot{x} = f^\dagger(x, y)$$
$$\dot{y} = \frac{1}{\varepsilon} g^\dagger(x, y) \qquad (1)$$

- **Relevance:** across disciplines (climatology, physiology, celestial mechanics, etc.).
- **Goal:** Given noisy observations of x, learn predictive model for future x dynamics.
- **Methodological constraints:**
  - Partial, noisy observations (e.g. observe x, but not y)
  - No knowledge of y, $g^\dagger, \varepsilon$, nor $\dim(y)$
  - Observations may be irregularly spaced and noisy
  - Ability to leverage partial knowledge of $f^\dagger$

Caltech

Leveraging partial knowledge of the dynamics

For any f_0 (regardless of its fidelity), there exists an $m^\dagger(x, y)$ such that (1) can be re-written as

$$\dot{x} = f_0(x) + m^\dagger(x, y) \tag{2a}$$

$$\dot{y} = \frac{1}{\varepsilon} g^\dagger(x, y). \tag{2b}$$

Caltech

Leveraging partial knowledge of the dynamics

For any f_0 (regardless of its fidelity), there exists an $m^\dagger(x, y)$ such that (1) can be re-written as

$$\dot{x} = f_0(x) + m^\dagger(x, y) \tag{2a}$$

$$\dot{y} = \frac{1}{\varepsilon} g^\dagger(x, y). \tag{2b}$$

There exists a closure $\mathcal{M}_t^\dagger$ that captures the full effect of the y-system on x:

$$\dot{x}(t) = f_0\big(x(t)\big) + \mathcal{M}_t^\dagger\left(\left\{x(s)\right\}_{s=0}^{t};\ y(0)\right). \tag{3}$$

We say the closure term $\mathcal{M}_t^\dagger$ has **memory.**

Caltech

Memoryless closure

When $\varepsilon \to 0$ and the y dynamics, with x fixed, are sufficiently mixing, then we expect that there exists a closure term $\overline{\mathcal{M}^\dagger}$ that **only depends on** x

$$\lim_{\varepsilon \to 0} \mathcal{M}_t^\dagger\left(\{x(s)\}_{s=0}^{t};\ y(0)\right) =: \overline{\mathcal{M}^\dagger}\big(x(t)\big).$$

Caltech

Memoryless closure

When $\varepsilon \to 0$ and the y dynamics, with x fixed, are sufficiently mixing, then we expect that there exists a closure term $\overline{\mathcal{M}^\dagger}$ that **only depends on** x

$$\lim_{\varepsilon \to 0} \mathcal{M}_t^\dagger \left(\left\{ x(s) \right\}_{s=0}^{t};\ y(0) \right) =: \overline{\mathcal{M}^\dagger}(x(t)).$$

For $\varepsilon \to 0$, eq. (3) reduces to

$$\dot{x}(t) = f_0(x) + \overline{\mathcal{M}^\dagger}(x). \tag{4}$$

(4) is also obtained when no unobserved variable y is present.

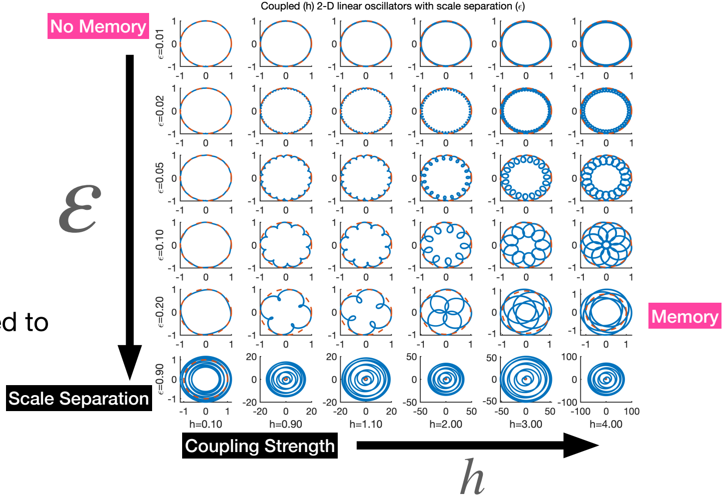
$\overline{\mathcal{M}^\dagger}$ **can be learned with any function approximation technique.**

Caltech

Coupled multi-scale linear oscillator

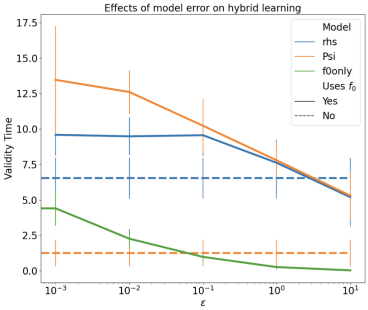
$$\dot{x} = Ax + hy$$
$$\dot{y} = \frac{1}{\varepsilon}Ay$$

- $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$
- $x_0 \sim \mathcal{N}(0,I)$ normalized to unit circle



Hybrid modeling is worthwhile, even when the available physics model appears BAD on its own!!! (Pathak et al. 2018)

Hybrid methods can rescue incorrect models



True Model

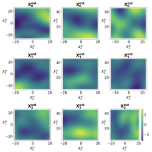
$$f^\dagger := f_{L63}$$

Approximate Model

$$f_\epsilon(x) := f^\dagger(x) + \epsilon\, m^\dagger(x)$$

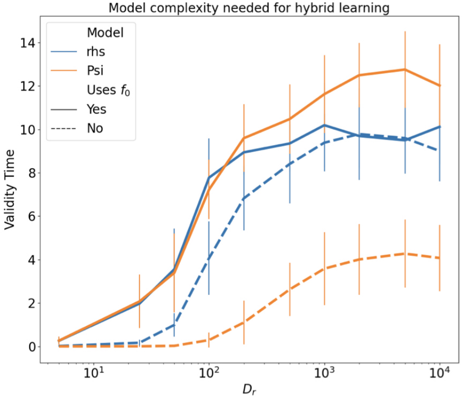
$$\Psi_\epsilon(x) := x + \int_0^{\Delta t} f_\epsilon(x(s))ds$$

$$m^\dagger \sim GP$$



Caltech

Hybrid methods are more parameter efficient



True Model

$$f^{\dagger} := f_{L63}$$

Approximate Model

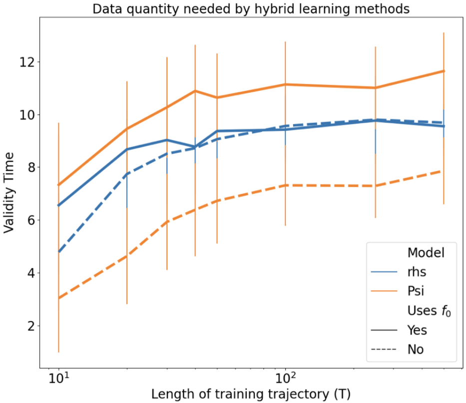
$$f_{\epsilon}(x) := f^{\dagger}(x) + \epsilon\, m^{\dagger}(x)$$

$$\Psi_{\epsilon}(x) := x + \int_0^{\Delta t} f_{\epsilon}(x(s))ds$$

$$\epsilon = 0.05$$

tech

Hybrid methods are less data hungry



True Model

$$f^{\dagger} := f_{L63}$$

Approximate Model

$$f_\epsilon(x) := f^\dagger(x) + \epsilon\, m^\dagger(x)$$

$$\Psi_\epsilon(x) := x + \int_0^{\Delta t} f_\epsilon(x(s))ds$$

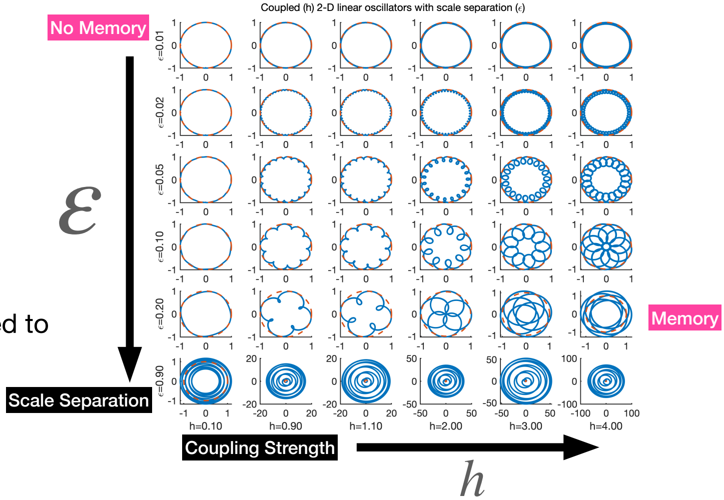
$$\epsilon = 0.05$$

tech

Coupled multi-scale linear oscillator

$$\dot{x} = Ax + hy$$
$$\dot{y} = \frac{1}{\varepsilon}Ay$$

- $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$
- $x_0 \sim \mathcal{N}(0,I)$ normalized to unit circle



Modeling non-Markovian dynamics in continuous-time

- Delay-differential equations:

$$\dot{x} = f_0(x) + f\left(\left\{x(t-\tau)\right\}_\tau;\ \theta\right)$$

  - ✗ Learnt model can be challenging/expensive to solve numerically
  - ✓ Allows for direct supervised training

- **Latent dynamics (re-augment state space):**

$$\begin{aligned}\dot{x} &= f_0(x) + m(x, r;\ \theta)\\ \dot{r} &= g(x, r;\ \theta)\end{aligned}$$

  - ✓ Learnt model is straightforward to solve numerically
  - ✗ **Training is more challenging (Chicken & Egg problem of inferring missing states AND their dynamics)**

Caltech

Learning latent dynamics in continuous-time

$$\dot{x} = f_0(x) + m(x, r;\ \theta) \qquad \iff \qquad \dot{u} = f(u;\ \theta), \quad u = [x, r]^T$$

$$\dot{r} = g(x, r;\ \theta) \qquad\qquad\qquad Hu = x$$

Caltech

Learning latent dynamics in continuous-time

$$\dot{x} = f_0(x) + m(x, r;\ \theta) \qquad \iff \qquad \dot{u} = f(u;\ \theta), \quad u = [x, r]^T$$
$$\dot{r} = g(x, r;\ \theta) \qquad\qquad\qquad Hu = x$$

Assume noisy observations $z = Hu + \eta$.

Caltech

Learning latent dynamics in continuous-time

$$\begin{aligned}\dot{x} &= f_0(x) + m(x, r;\ \theta) \\ \dot{r} &= g(x, r;\ \theta)\end{aligned} \quad \iff \quad \begin{aligned}\dot{u} = f(u;\ \theta), \quad u = [x, r]^T \\ Hu = x\end{aligned}$$

Assume noisy observations $z = Hu + \eta$.
Let $u(t; v, \theta)$ solve $\dot{u} = f(u; \theta),\ u(0) = v$.

Caltech

Learning latent dynamics in continuous-time

$$\dot{x} = f_0(x) + m(x, r;\ \theta) \qquad \Longleftrightarrow \qquad \dot{u} = f(u;\ \theta), \quad u = [x, r]^T$$
$$\dot{r} = g(x, r;\ \theta) \qquad\qquad\qquad Hu = x$$

Assume noisy observations $z = Hu + \eta$.
Let $u(t; v, \theta)$ solve $\dot{u} = f(u; \theta),\ u(0) = v$.

Hard Constraint Idea 1: Infer init. cond. and parameters (Rubanova *et al.* 2019)

$$\underset{\theta, u_0}{\operatorname{argmin}} \int_0^T \|z(t) - Hu(t; u_0, \theta)\|^2 dt.$$

- ✗ Poorly-posed with larger T for chaotic systems with sensitivity to u_0.

Caltech

Learning latent dynamics in continuous-time

$$\dot{x} = f_0(x) + m(x, r;\ \theta) \qquad \Longleftrightarrow \qquad \dot{u} = f(u;\ \theta), \quad u = [x, r]^T$$
$$\dot{r} = g(x, r;\ \theta) \qquad\qquad\qquad\qquad Hu = x$$

Assume noisy observations $z = Hu + \eta$.
Let $u(t; v, \theta)$ solve $\dot{u} = f(u; \theta),\ u(0) = v$.

Let $\hat{m}(t, \tau, \theta_{\mathrm{DYN}}, \theta_{\mathrm{DA}})$ be an estimate of $u(t) \ \mid \{z(t-s)\}_{s=0}^{\tau},\ \theta_{\mathrm{DYN}},\ u(t-\tau) = 0$.

Caltech

Learning latent dynamics in continuous-time

$$\dot{x} = f_0(x) + m(x, r;\ \theta) \qquad \Longleftrightarrow \qquad \dot{u} = f(u;\ \theta), \quad u = [x, r]^T$$
$$\dot{r} = g(x, r;\ \theta) \qquad\qquad\qquad\qquad Hu = x$$

Assume noisy observations $z = Hu + \eta$.

Let $u(t; v, \theta)$ solve $\dot{u} = f(u; \theta),\ u(0) = v$.

Let $\hat{m}(t, \tau, \theta_{\mathrm{DYN}}, \theta_{\mathrm{DA}})$ be an estimate of $u(t) \ \mid \{z(t-s)\}_{s=0}^{\tau},\ \theta_{\mathrm{DYN}},\ u(t-\tau) = 0.$

DA-based inference: Initial conditions can be estimated jointly with parameters

$$\underset{\theta_{\mathrm{DYN}},\ \theta_{\mathrm{DA}}}{\operatorname{argmin}} \sum_{k=1}^{K} \int_0^T \|z^{(k)}(t) - Hu\big(t; \hat{m}(t_k, \tau, , \theta_{\mathrm{DYN}}, \theta_{\mathrm{DA}}), \theta_{\mathrm{DYN}}\big)\|^2 dt.$$

Caltech

Learning latent dynamics in continuous-time

$$\dot{x} = f_0(x) + m(x, r;\ \theta) \qquad \Longleftrightarrow \qquad \dot{u} = f(u;\ \theta), \quad u = [x, r]^T$$

$$\dot{r} = g(x, r;\ \theta) \qquad\qquad\qquad\qquad Hu = x$$

Assume noisy observations $z = Hu + \eta$.

Let $u(t; v, \theta)$ solve $\dot{u} = f(u; \theta),\ u(0) = v$.

Let $\hat{m}(t, \tau, \theta_{\mathrm{DYN}}, \theta_{\mathrm{DA}})$ be an estimate of $u(t) \ \mid \{z(t-s)\}_{s=0}^{\tau},\ \theta_{\mathrm{DYN}},\ u(t-\tau) = 0$.

DA-based inference: Initial conditions can be estimated jointly with parameters

$$\underset{\theta_{\mathrm{DYN}},\ \theta_{\mathrm{DA}}}{\operatorname{argmin}} \sum_{k=1}^{K} \int_0^T \|z^{(k)}(t) - Hu\big(t; \hat{m}(t_k, \tau, , \theta_{\mathrm{DYN}}, \theta_{\mathrm{DA}}), \theta_{\mathrm{DYN}}\big)\|^2 dt.$$

- Here, we perform joint estimation with auto-differentiable 3DVAR
- Chen *et al.* 2021 perform joint estimation with auto-differentiable Ensemble Kalman Filter
- Carassi *et al.* 2021 apply alternating descent (EnKF for $\hat{m}$, supervised SGD for θ)

Caltech

Example 2: Lorenz '63 with partial, noisy observations

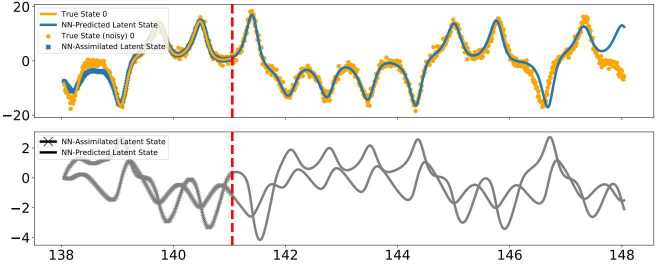


Figure: Accurate short-term forecasts

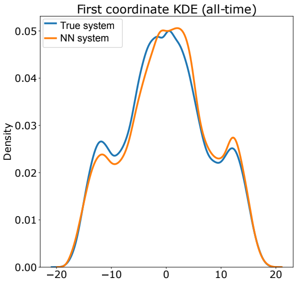


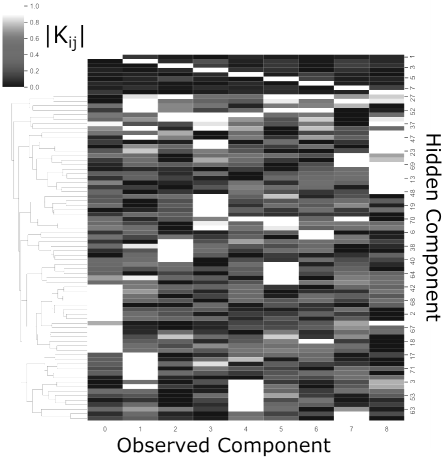
Figure: Accurate long-time statistics (empirically stable for $T = 10^5$)

- **Experimental Setting:** $H = [1, 0, 0]$ (observe first-component only), $T = 1000$, $\Delta t = 0.01$, $\sigma = 1$ (observation noise).
- **Modeling Setting:** $d_r = 2$ (assumed missing dimension), 2-layer NN w/ GeLU activation (width 50).

Caltech

Example 2: Can infer Data Assimilation Parameters

- We can infer θ_{DA} (K for 3DVAR, covariances for EnKF/UKF).
- This can tell us how observables correlate to latent variables (e.g. in clusters)



Caltech

Conclusions

1. Hybrid modeling is often worthwhile
   - Improved predictions, even when physical model is quite bad or nearly perfect
   - Less data hunger, more parameter efficient

Caltech

Conclusions

1. Hybrid modeling is often worthwhile
    - Improved predictions, even when physical model is quite bad or nearly perfect
    - Less data hunger, more parameter efficient
2. Fusing Data Assimilation and machine-learning-based optimization techniques is useful for coping with:
    - Highly non-linear and chaotic systems
    - Noisy and irregularly sampled data
    - Partial observations of large systems
    - Tuning data assimilation schemes

Caltech

Conclusions

1. Hybrid modeling is often worthwhile
   - Improved predictions, even when physical model is quite bad or nearly perfect
   - Less data hunger, more parameter efficient
2. Fusing Data Assimilation and machine-learning-based optimization techniques is useful for coping with:
   - Highly non-linear and chaotic systems
   - Noisy and irregularly sampled data
   - Partial observations of large systems
   - Tuning data assimilation schemes
3. Other things I've learned:
   - Solving ODEs on GPUs in parallel is way fast!
   - Optimizing NNs isn't as bad as you think (often loosely convex), but requires expertise!

Caltech

Future Directions

- **Opportunities** to new problems where decent (or no) models are available, along with data
  - Inferring model errors to improve biological models (need real data)

Caltech

Future Directions

- **Opportunities** to new problems where decent (or no) models are available, along with data
  - Inferring model errors to improve biological models (need real data)
  - Inferring reductions of multi-scale models (simulated and/or real data)

Caltech

Future Directions

- **Opportunities** to new problems where decent (or no) models are available, along with data
  - Inferring model errors to improve biological models (need real data)
  - Inferring reductions of multi-scale models (simulated and/or real data)

Future Directions

- **Opportunities** to new problems where decent (or no) models are available, along with data
  - Inferring model errors to improve biological models (need real data)
  - Inferring reductions of multi-scale models (simulated and/or real data)
- **Challenges**:
  - Limited data $\Longrightarrow$ learn error terms that are 0 away from data and/or provide UQ (as SDE)

Caltech

Future Directions

- **Opportunities** to new problems where decent (or no) models are available, along with data
  - Inferring model errors to improve biological models (need real data)
  - Inferring reductions of multi-scale models (simulated and/or real data)
- **Challenges**:
  - Limited data $\Longrightarrow$ learn error terms that are 0 away from data and/or provide UQ (as SDE)
  - Interpretability $\Longrightarrow$ parsimony/sparsity (ℓ_1 regularization); ensure SMALL corrections
  - Not just for dynamical systems!!!

$$y = Ax + Bx \otimes x + f_{\mathrm{NN}}(x)$$

Caltech

Related Work: Hybrid modeling

- Kaheman, Kadierdan, Eurika Kaiser, Benjamin Strom, J. Nathan Kutz, and Steven L. Brunton. "Learning Discrepancy Models From Experimental Data." ArXiv:1909.08574 [Cs, Eess, Stat], September 18, 2019. http://arxiv.org/abs/1909.08574.
- Rico-Martines, R., I. G. Kevrekidis, M. C. Kube, and J. L. Hudson. "Discrete- vs. Continuous-Time Nonlinear Signal Processing: Attractors, Transitions and Parallel Implementation Issues." In 1993 American Control Conference, 1475–79. San Francisco, CA, USA: IEEE, 1993. https://doi.org/10.23919/ACC.1993.4793116.
- Pathak, Jaideep, Alexander Wikner, Rebeckah Fussell, Sarthak Chandra, Brian R. Hunt, Michelle Girvan, and Edward Ott. "Hybrid Forecasting of Chaotic Processes: Using Machine Learning in Conjunction with a Knowledge-Based Model." Chaos: An Interdisciplinary Journal of Nonlinear Science 28, no. 4 (April 1, 2018): 041101. https://doi.org/10.1063/1.5028373.
- Harlim, J., Jiang, S. W., Liang, S. & Yang, H. Machine learning for prediction with missing dynamics. Journal of Computational Physics 428, 109922 (2021).

Caltech

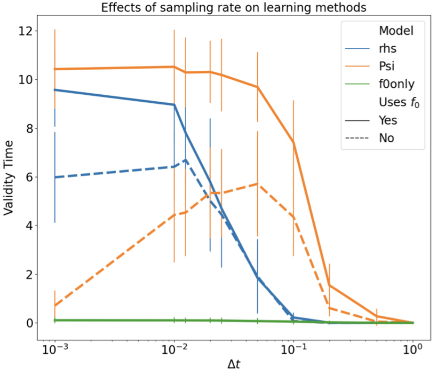
Related Work: Learning dynamics from partial/noisy observations

- Chen, Y., Sanz-Alonso, D. & Willett, R. Auto-differentiable Ensemble Kalman Filters. arXiv:2107.07687 [cs, stat] (2021).
- Ouala, S. et al. Learning latent dynamics for partially observed chaotic systems. Chaos: An Interdisciplinary Journal of Nonlinear Science 30, 103121 (2020).
- Brajard, J., Carrassi, A., Bocquet, M. & Bertino, L. Combining data assimilation and machine learning to infer unresolved scale parametrization. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 379, 20200086 (2021).

Caltech

Example 3: Lorenz '63 with unknown Markovian errors

Timestep informs choice of continuous vs discrete model



True Model

$$f^{\dagger} := f_{L63}$$

Approximate Model

$$f_{\epsilon}(x) := f^{\dagger}(x) + \epsilon\, m^{\dagger}(x)$$

$$\Psi_{\epsilon}(x) := x + \int_0^{\Delta t} f_{\epsilon}(x(s))ds$$

$$\epsilon = 0.05$$

tech

Learning theory for Markovian residuals (no memory)

Model: $\dot{x} = f_0(x) + m(x)$

Trajectory-based loss:

$$\mathcal{I}_T(m) := \frac{1}{T}\int_0^T \|\dot{x}(t) - f_0\big(x(t)\big) - m\big(x(t)\big)\|_2^2 dt.$$

Caltech

Learning theory for Markovian residuals (no memory)

Model: $\dot{x} = f_0(x) + m(x)$

Trajectory-based loss:

$$\mathcal{I}_T(m) := \frac{1}{T}\int_0^T \|\dot{x}(t) - f_0\big(x(t)\big) - m\big(x(t)\big)\|_2^2 dt.$$

A natural loss function

Choose a measure μ on $\mathbb{R}^{d_x}$, let $m^\dagger(x) := \dot{x} - f_0(x)$, and define the loss

$$\mathcal{L}_\mu(m, m^\dagger) := \int_{\mathbb{R}^{d_x}} \|m^\dagger(x) - m(x)\|_2^2 d\mu(x).$$

Assume $m^\dagger$, $x(\cdot)$ is ergodic with invariant density μ. Exchange time/space averages:

$$\mathcal{L}_\mu(m, m^\dagger) = \lim_{T\to\infty} \mathcal{I}_T(m).$$

i.e. Optimizing over a temporal trajectory implicitly optimizes spatially w.r.t. invariant measure.

ch

Learning theory for Markovian residuals (no memory)

Model: $\dot{x} = f_0(x) + m(x)$

Trajectory-based loss:

$$\mathcal{I}_T(m) := \frac{1}{T}\int_0^T \|\dot{x} - f_0(x) - m(x(t))\|_2^2 dt.$$

Assume:

- Linear classes of m (e.g. random feature models, dictionary learning, etc.)
- f_0 is Lipshitz
- *x is ergodic with CLT-like mixing*

Theorem 5.2 (Levine and Stuart, 2021)

- Excess risk and generalization error bounded by $1/\sqrt{T}$ *in distribution.*
- Excess risk and generalization error bounded by $\log\log T/\sqrt{T}$ *almost surely.*

Example 1: Lorenz '96 Multi-Scale closure

Each (slow) variable $X_k \in \mathbb{R}$ is coupled to a subgroup of (fast) variables $Y_k \in \mathbb{R}^J$. We have $X \in \mathbb{R}^K$ and $Y \in \mathbb{R}^{K \times J}$. For $k = 1 \ldots K$ and $j = 1 \ldots J$, we write

$$\dot{X}_k = f_k(X) + h_x \bar{Y}_k \tag{5a}$$

$$\dot{Y}_{k,j} = \frac{1}{\varepsilon} r_j(X_k, Y_k) \tag{5b}$$

$$\bar{Y}_k = \frac{1}{J} \sum_{j=1}^{J} Y_{k,j} \tag{5c}$$

Memoryless closure ($\varepsilon \to 0$)

We apply an averaging hypothesis that assumes

$$\dot{X}_k \approx f_k(X) + m(X_k)$$

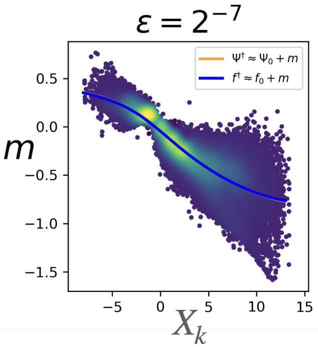
where $m : \mathbb{R} \to \mathbb{R}$ is a random feature model applied component-wise.

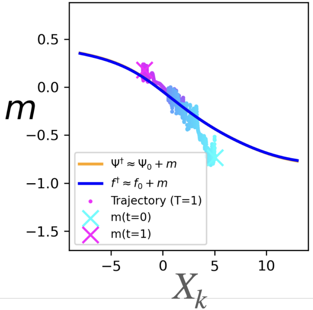
ch

Example 1: Lorenz '96 Multi-Scale closure—scale separated

- At large scale separation ($\varepsilon = 2^{-7}$), the model error $m = f_k - \dot{x}$ is **highly concentrated** around its mean and **oscillates rapidly**.
- Thus, the averaging hypothesis holds and Markovian modeling is sensible.

$$\dot{X}_k = f_k(X) + m(X_k)$$

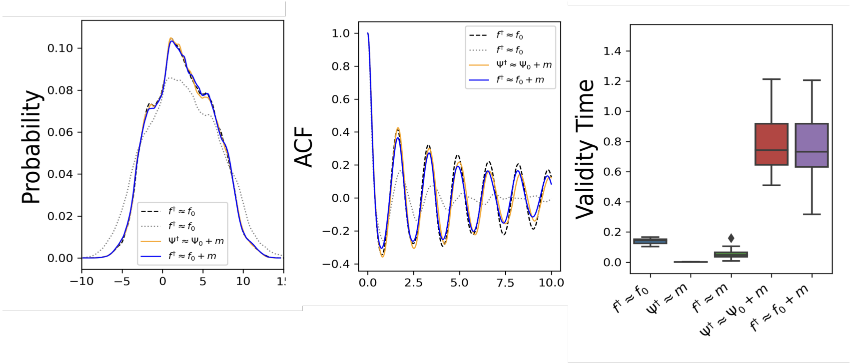




Example 1: Lorenz '96 Multi-Scale closure—scale separated

At large scale separation ($\varepsilon = 2^{-7}$), we can accurately reconstruct the system dynamics and their statistics using a simple Markovian residual on X

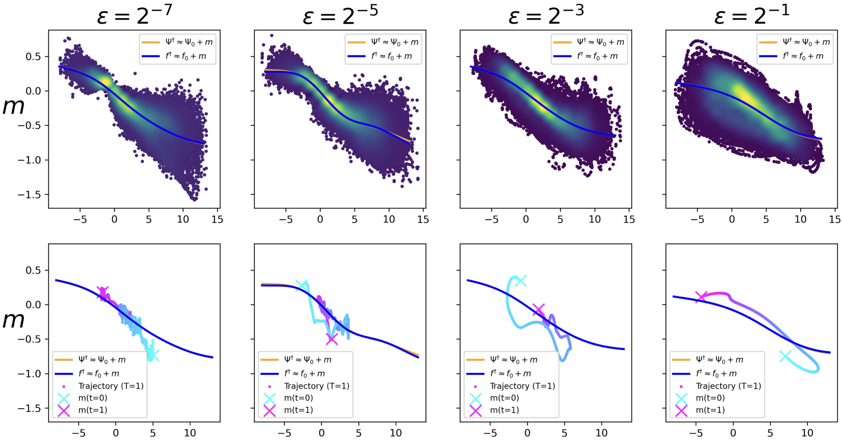
$$\dot{X}_k = f_k(X) + m(X_k)$$



Learning the entire system from scratch did not work (with the data we used)

Example 1: Lorenz '96 Multi-Scale closure beyond scale separation

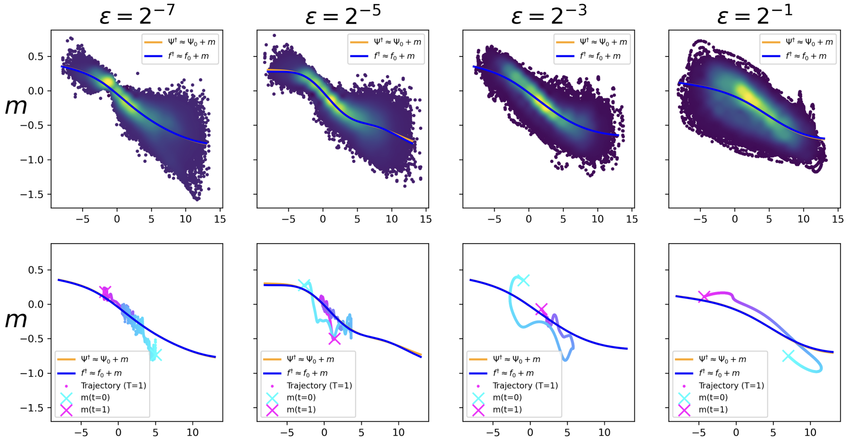
- Consider the model error $m = f_k - \dot{x}$ at different levels of scale separation.
- Less scale separation **increases the variance** of the residuals and **slows their oscillations**.



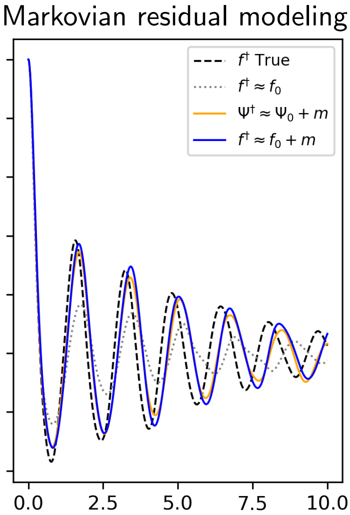
Caltech

Example 1: Lorenz ’96 Multi-Scale closure beyond scale separation

- Consider the model error $m = f_k - \dot{x}$ at different levels of scale separation.
- Less scale separation **increases the variance** of the residuals and **slows their oscillations**.



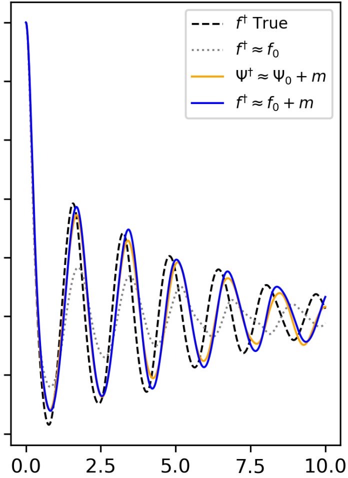
Example 1: Lorenz '96 Multi-Scale closure beyond scale separation ($\varepsilon = 2^{-1}$)



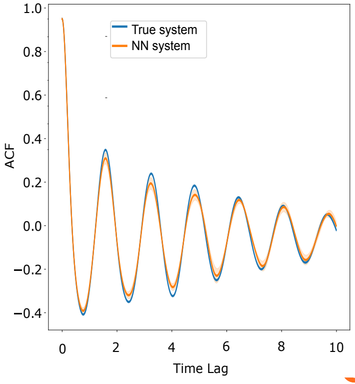
Caltech

Example 1: Lorenz '96 Multi-Scale closure beyond scale separation ($\varepsilon = 2^{-1}$)

Markovian residual modeling



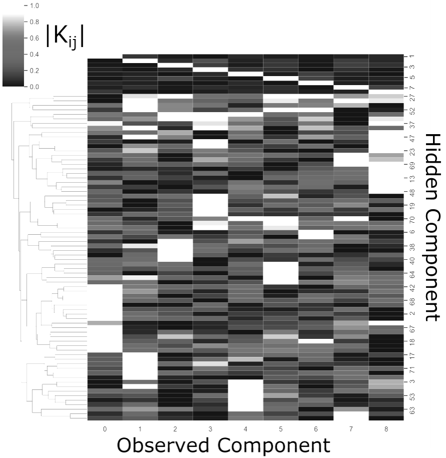
Non-Markovian residual modeling (augmented latent dynamics).



Caltech

Example 1: Lorenz ’96 Multi-Scale closure beyond scale separation ($\varepsilon = 2^{-1}$)

- The true L96MS system has a clustered subgrouping of fast variables—our model has re-discovered this structure, and the DA gain K has learnt to exploit these correlations for improved filtering.



Caltech