Training neural operators to preserve invariant measures of chaotic attractors

THE UNIVERSITY OF CHICAGO

Ruoxi Jiang*, Peter Y. Lu*, Elena Orlova, Rebecca Willett

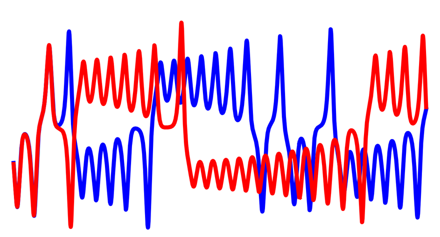
Background: Emulators for chaotic systems

Goal: Consider a chaotic dynamical system $\frac{d\mathbf{u}}{dt} = G(\mathbf{u}, \phi)$ with an unknown governing equation G and a set of parameters ϕ that specify an environment. We aim to approximate the dynamics with a data-driven emulator $\hat{g}_\theta$:

$$\hat{\mathbf{u}}_{t+\Delta t} := \hat{g}_\theta(\hat{\mathbf{u}}_t, \phi).$$

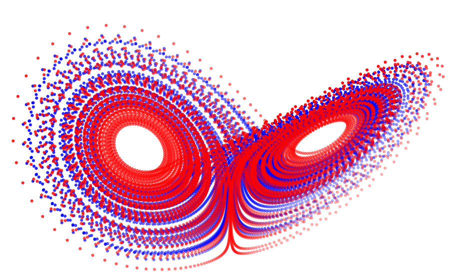
Challenges:

- We are interested in a system where the unknown chaotic G is highly sensitive to initial conditions, and is impossible to exactly predicted over a long term.
- Noise exacerbates this unpredictability, and makes it difficult to train emulators.



Our contributions:

- Optimal-transport (OT) based method to train emulators to match long-term known statistics characteristic of chaotic attractors.
- Contrastive learning (CL) based method to implicitly train emulators to match long-term unknown statistics of chaotic attractors.



Emulators trained with RMSE vs. invariant measures

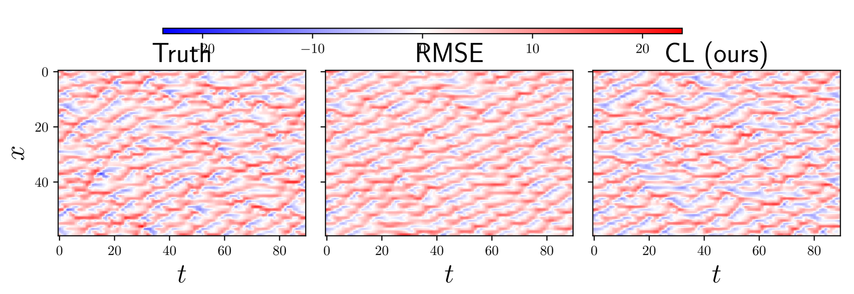


Figure 1. Training with RMSE loss yields emulators with very different statistical properties from the true chaotic system (e.g. more periodic), whereas **training with a contrastive loss preserves the statistical properties of the chaotic system.**

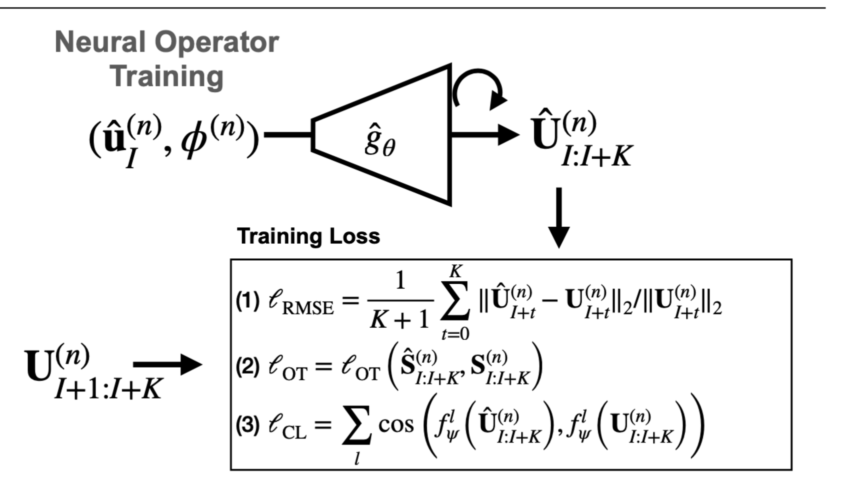


Figure 2. Emulators are trained to take an initial state and output future states in a recurrent fashion.

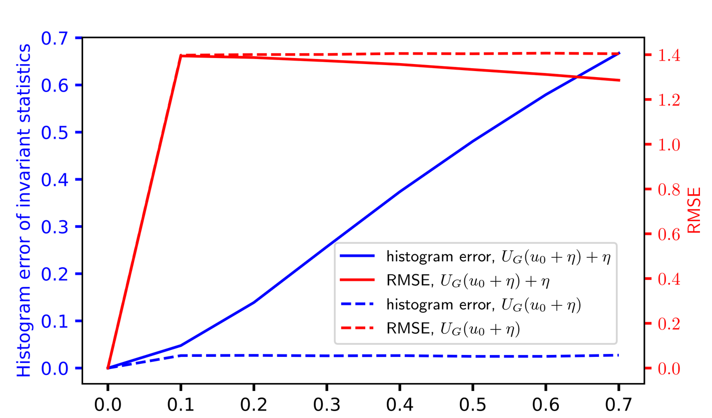


Figure 3. **Impact of noise on various error metrics using ground truth simulations** with increasingly noisy initial conditions $\mathbf{U}_G(\mathbf{u}_0 + \eta)$ and added measurement noise $\mathbf{U}_G(\mathbf{u}_0 + \eta) + \eta$. Here, $\mathbf{U}_G(\cdot)$ refers to the solution given an initial condition.

RMSE loss:

- While widely used for training emulators, RMSE does not capture long-term dynamics;
- Chaos and noise make the predictions of emulators trained with only RMSE degenerate quickly over time.

Invariant measures:

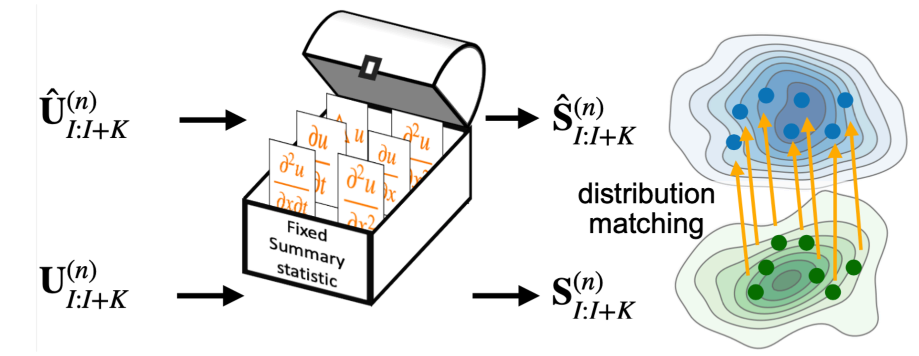
- Can capture time-invariant statistical behavior of the true dynamics;
- Have a much more robust response to noise.

Notations:

- **Sequence of dynamics.** A sequence of $K+1$ consecutive time points on the trajectory coming from environments $n = \{1, \ldots, N\}$ is $\mathbf{U}_{I:I+K}^{(n)} := \{\mathbf{u}_{t_i}^{(n)}\}_{i=I}^{I+K}$, where I is the beginning of the time interval.
- **Time-invariant statistics.** Any time-invariant statistical property $S_{\mathcal{A}}$ of the dynamics on the attractor $\mathcal{A}$ can be written as $S_{\mathcal{A}} = \mathbb{E}_{\mu_{\mathcal{A}}}[s] = \int s(\mathbf{u})\, d\mu_{\mathcal{A}}(\mathbf{u}) = \lim_{T\to\infty} \frac{1}{T}\int^T s(\mathbf{u}_{\mathcal{A}}(t))\, dt$ for some function $s(\mathbf{u})$ where $\mu_{\mathcal{A}}$ is a natural invariant probability measure of trajectroy in the basin of the attractor $\mathcal{A}$.

1st approach: Physics-informed optimal transport

We assume access to expert domain knowledge to define summary statistics $\mathbf{s}(\mathbf{u}_t)$ representing physical property of the dynamical system. We aim to match the distributions of the statistics.



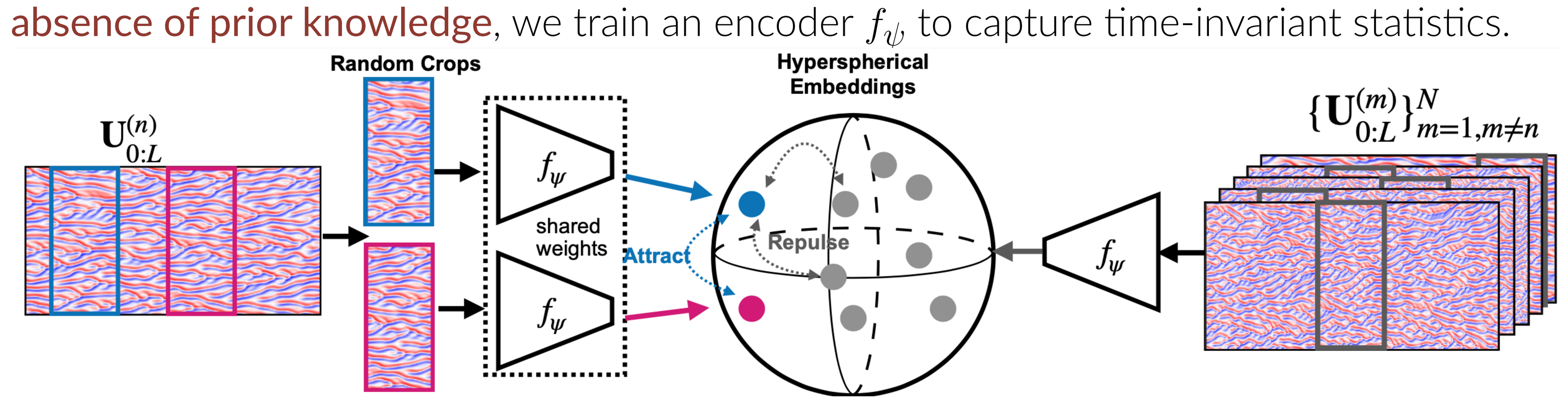
With discrete samples of statistics $\mathbf{S}_{I:I+K} := \{\mathbf{s}(\mathbf{u}_{t_i})\}_{i=I}^{I+K}$, we use the Sinkhorn algorithm [1] to efficiently solve the entropy regularized optimal transport problem,

$$\ell_{\mathrm{OT}}(\mathbf{S}, \hat{\mathbf{S}}) = \frac{1}{2}\left(W^\gamma(\mathbf{S}, \hat{\mathbf{S}})^2 - \frac{W^\gamma(\mathbf{S}, \mathbf{S})^2 + W^\gamma(\hat{\mathbf{S}}, \hat{\mathbf{S}})^2}{2}\right), \quad (1)$$

where $W^\gamma(\mathbf{S}, \hat{\mathbf{S}})^2$ is the Wasserstein distance with an entropy regularization term (of the scale γ) and squared cost matrix. Our final loss is: $\ell(\theta) = \alpha\ell_{\mathrm{OT}} + \ell_{\mathrm{RMSE}}$.

2nd approach: Contrastive feature learning

In absence of prior knowledge, we train an encoder f_ψ to capture time-invariant statistics.



Key Premise:

- Sequences from the same trajectory have the same chaotic attractor, and so should have similar embeddings.
- Sequences from different trajectories corresponding to different chaotic attractors should have dissimilar embeddings.

For **contrastive feature loss,** we use the cosine distance between a series of features of f_ψ [2]:

$$\ell_{\mathrm{CL}}(\mathbf{U}, \hat{\mathbf{U}}; f_\psi) := \sum_l \cos\left(f_\psi^l(\mathbf{U}), f_\psi^l(\hat{\mathbf{U}})\right), \quad (2)$$

where f_ψ^l gives l-th layer feature output. Our final loss is: $\ell(\theta) = \lambda\ell_{\mathrm{CL}} + \ell_{\mathrm{RMSE}}$.

Experiments

Experimental setup. We have noisy observations $\mathbf{u}(t)$ with noise $\eta \sim \mathcal{N}(0, r^2\sigma^2 I)$. **Baselines.** We consider the baseline as training with RMSE. **Backbones.** We use the Fourier neural operator [3].

Evaluation metrics.

- Histogram error: $\mathrm{Err}(\hat{\mathbf{H}}, \mathbf{H}) := \sum_{b=1}^{B} \|c_b - \hat{c}_b\|_1$, where $\mathbf{H}$ is a histogram of the invariant statistics $\mathbf{S}$, and c_b is frequencies of the corresponding values of B bins.
- Energy spectrum error: $\frac{1}{T}\sum_{\mathbf{u}_t, \hat{\mathbf{u}}_t \in \mathbf{U}_{1:T}, \hat{\mathbf{U}}_{1:T}} \frac{\||\mathcal{F}[\mathbf{u}_t]|^2 - |\mathcal{F}[\hat{\mathbf{u}}_t]|^2\|_1}{\||\mathcal{F}[\mathbf{u}_t]|^2\|_1}$, where $\mathcal{F}[\cdot]$ is the spatial FFT.
- Leading Lyapunov exponent (LE) error: The LE (λ) measures how quickly the chaotic system becomes unpredictable. We report the relative absolute error as $|\hat{\lambda} - \lambda|/|\lambda|$.
- Fractal dimension (FD) error: FD (D) is a characterization of the dimension of the attractor. We report the absolute error as $|\hat{D} - D|$.

Results

Lorenz-96 system $\frac{du^i}{dt} = (u^{i+1} - u^{i-2})u^{i-1} - u^i + F$. Let $\mathbf{s}(\mathbf{u}) := \{\frac{du^i}{dt}, (u^{i+1} - u^{i-2})u^{i-1}, u^i\}$. We use 2000 training samples with each $\phi^{(n)} \sim U([10.0, 18.0])$.

Training	Histogram ↓	Energy Spec. ↓	Leading LE ↓	FD ↓
ℓ_{RMSE}	0.215	0.291	0.440	3.580
$\ell_{\mathrm{OT}}+\ell_{\mathrm{RMSE}}$	**0.057**	**0.123**	0.084	3.453
$\ell_{\mathrm{CL}}+\ell_{\mathrm{RMSE}}$	0.132	0.241	**0.064**	**1.894**

Table 1. **Performance on 1500-step predictions with noise scale $r = 0.3$.**

Training stats.	Histogram ↓	Energy Spec. ↓	Leading LE ↓	FD ↓
$\mathbf{S}$ (full)	**0.057**	**0.123**	**0.084**	3.453
$\mathbf{S}_1$ (partial)	0.090	0.198	0.263	3.992
$\mathbf{S}_2$ (minimum)	0.221	0.221	0.276	**3.204**

Table 2. **Performance of OT method for different choices of summary statistics, $r = 0.3$:** (1) full statistics $\mathbf{S}(\mathbf{u}) := \{\frac{du_i}{dt}, (u_{i+1} - u_{i-2})u_{i-1}, u_i\}$; (2) partial statistics $\mathbf{S}_1(\mathbf{u}) := \{(u_{i+1} - u_{i-2})u_{i-1}\}$; or (3) minimum statistics $\mathbf{S}_2(\mathbf{u}) := \{\bar{\mathbf{u}}\}$, where $\bar{\mathbf{u}}$ is the spatial average.

Training	Histogram ↓	Energy Spec. ↓	Leading LE ↓	FD ↓
ℓ_{RMSE}	0.255	0.307	0.459	3.879
$\ell_{\mathrm{OT}}+\ell_{\mathrm{RMSE}}$	**0.055**	**0.124**	0.080	4.015
$\ell_{\mathrm{CL}}+\ell_{\mathrm{RMSE}}$	0.130	0.193	**0.031**	**1.747**

Table 3. **Emulator performance with reduced environment diversity with $r = 0.3$.** We shrink the parameter range for generating the dataset from $[10, 18]$ to $[16, 18]$.

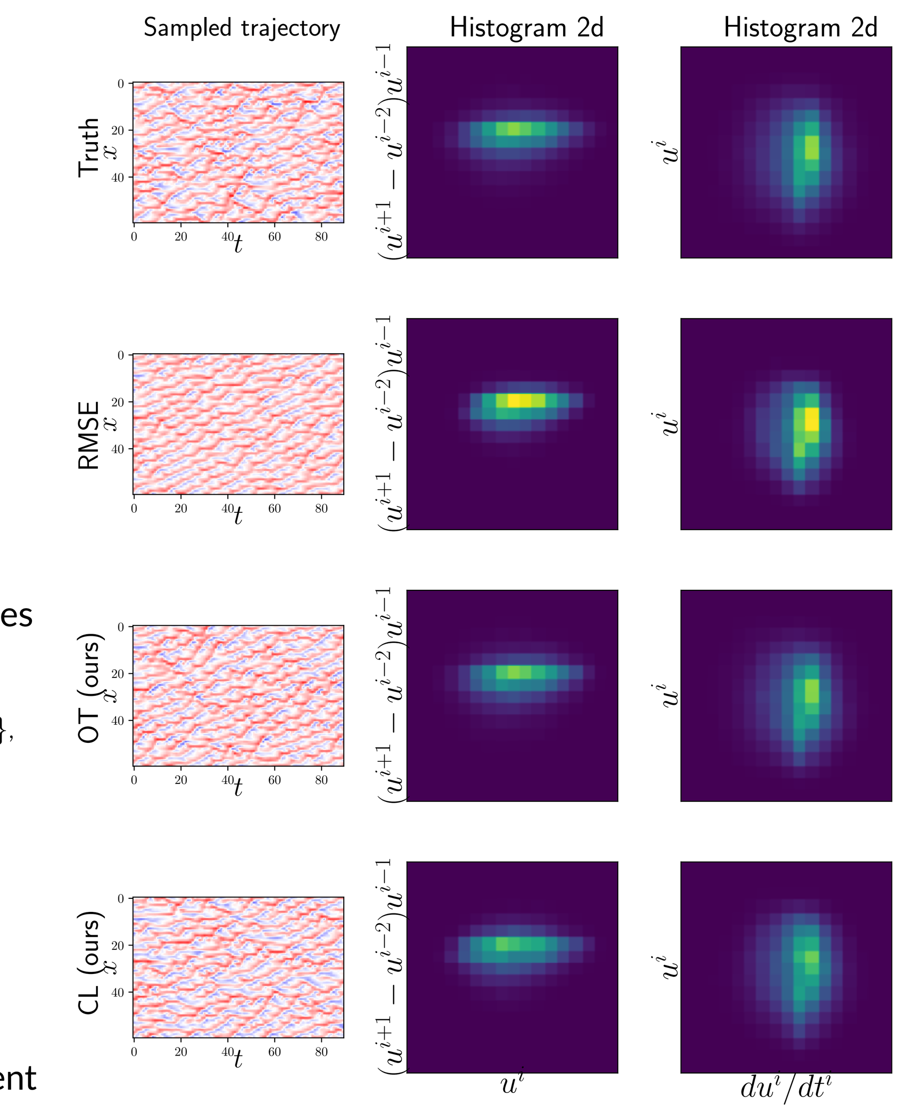


Figure 4. **Sampled dynamics and summary statistics distributions.**

Kuramoto–Sivashinsky (KS) equation $\frac{\partial u}{\partial t} = -u\frac{\partial u}{\partial x} - \phi\frac{\partial^2 u}{\partial x^2} - \frac{\partial^4 u}{\partial x^4}$. We define $\mathbf{s}(u) := \{\frac{\partial u}{\partial t}, \frac{\partial u}{\partial x}, \frac{\partial^2 u}{\partial x^2}\}$. We generate 2000 training samples, with each $\phi^{(n)} \sim U([1.0, 2.6])$.

Training	Histogram ↓	Energy Spec. ↓	Leading LE ↓
ℓ_{RMSE}	0.390	0.290	0.101
$\ell_{\mathrm{OT}}+\ell_{\mathrm{RMSE}}$	**0.172**	0.211	**0.094**
$\ell_{\mathrm{CL}}+\ell_{\mathrm{RMSE}}$	0.193	**0.176**	0.108

Table 4. **Performance on 500-step predictions with noise scale $r = 0.3$.**

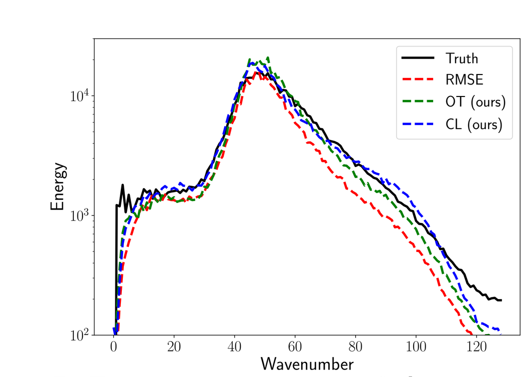


Figure 5. **Energy spectrum of the sampled dynamics.**

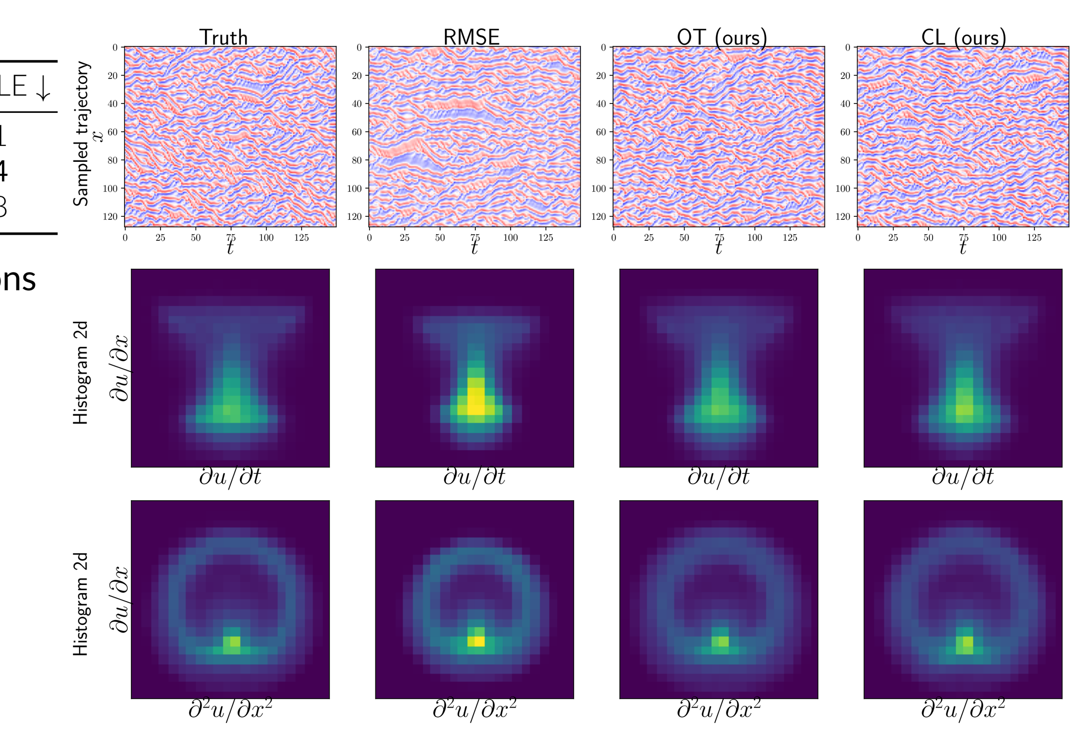


Figure 6. **Sampled dynamics and summary statistics distributions.**

References

[1] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, Red Hook, NY, 2013. Curran Associates, Inc.

[2] Z. Richard et al. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[3] Z. Li et al. Fourier neural operator for parametric partial differential equations. In *ICLR*, 2021.

roxie62/lup/eorlova@uchicago.edu