DySLIM: Dynamics Stable Learning by Invariant Measure for Chaotic Systems

Yair Schiff · Zhong Yi Wan · Jeffrey Parker
Stephan Hoyer · Volodymyr Kuleshov
Fei Sha · Leonardo Zepeda-Nunez

July 2024

ICML
International Conference
On Machine Learning

Background

Dynamical systems: Dynamic model governs how states evolve in time (e.g., weather variables, fluid flows, particles moving in space):

$$\mathbf{u}_k = \mathcal{S}(\mathbf{u}_{k-1}) = \ldots = \mathcal{S}^k(\mathbf{u}_0) \qquad \mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_k \in \mathcal{U}$$

Learning systems from data: We can use historical trajectories to learn a model that mimics system dynamics:

$$\text{Find } \theta \text{ s.t. } \mathcal{S}_\theta(\mathbf{u}_{k-1}) \approx \mathbf{u}_k$$

Trajectory matching: Most approaches rely on minimization of a trajectory matching objective, e.g:

$$\min_\theta \mathcal{L}(\theta) = ||\mathcal{S}_\theta(\mathbf{u}_k) - \mathbf{u}_{k+1}||$$

Background

Chaotic systems are characterized by **states diverging exponentially.**

Inconsistencies between learned models and true dynamics can therefore **become exaggerated.**

Animation credit: https://geoffboeing.com/2016/12/animating-lorenz-attractor-python/

Background

$$\mu^* = \mathcal{S}_{\#}\mu^*$$

Invariant measure

Many systems, even chaotic ones are **ergodic** and remain in a narrow set of possible states known as the **attractor.**

This attractor supports an **invariant measure** that is unchanged by the system dynamics.

Problem statement

❗ Hypothesis: Models learned from data, **do not learn the "right" invariant measure** and fail to remain close to the attractor, resulting in non-physical behavior for long rollouts.

$$\mu_\theta^* \neq \mu^*$$

Learned invariant measure (μ_θ^*)

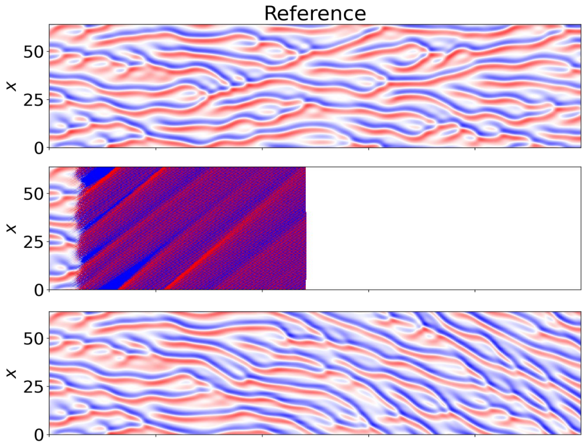
True invariant measure (μ^*)

Problem statement

❗ Hypothesis: Models learned from data, **do not learn the "right" invariant measure** and fail to remain close to the attractor, resulting in non-physical behavior for long rollouts.

$$\mu_\theta^* \neq \mu^*$$

Example failure modes



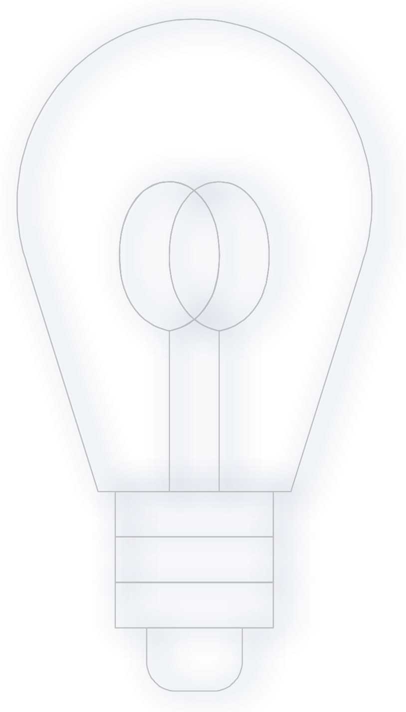
Problem statement

❗ Hypothesis: Models learned from data, **do not learn the "right" invariant measure** and fail to remain close to the attractor, resulting in non-physical behavior for long rollouts.

$$\mu_\theta^* \neq \mu^*$$

❓ Can we achieve **stability** by moving learned model outputs to the correct attractor?

Proposed methodology



Learn a model so that we:

match data trajectories while **preserving the invariant measure.**

$$\min_{\theta} \mathcal{L}(\theta) \quad \text{s.t.} \quad \mu_{\theta}^{*} = \mu^{*}$$

Proposed methodology

$$\min_{\theta} \mathcal{L}(\theta) \quad \text{s.t.} \quad \mu_{\theta}^{*} = \mu^{*}$$

Proposed methodology

$$\min_{\theta} \mathcal{L}(\theta) \quad \text{s.t.} \quad \mu_{\theta}^{*} = \mu^{*}$$

$$\mathcal{L}_{\lambda}^{\mathrm{D}}(\theta) = \mathcal{L}(\theta) + \lambda \mathrm{D}(\mu^{*}, \mu_{\theta}^{*})$$

$$\mathcal{L}_\lambda^{\mathrm{D}}(\theta) = \mathcal{L}(\theta) + \lambda \mathrm{D}(\mu^*, \mu_\theta^*)$$

What is the "right" D to use?

What is the “right” D to use?

- ✓ Respects underlying geometry of state space.
- ✓ Unbiased and sample efficient estimation.
- ✓ Entails convergence properties.
- ✓ Not “cursed” by dimension.

Journal of Machine Learning Research 13 (2012) 723-773 Submitted 4/08; Revised 11/11; Published 3/12

A Kernel Two-Sample Test

Arthur Gretton* ARTHUR.GRETTON@GMAIL.COM
MPI for Intelligent Systems
Spemannstrasse 38
72076 Tübingen, Germany

Karsten M. Borgwardt† KARSTEN.BORGWARDT@TUEBINGEN.MPG.DE
Machine Learning and Computational Biology Research Group
Max Planck Institutes Tübingen
Spemannstrasse 38
72076 Tübingen, Germany

Malte J. Rasch‡ MALTE@MAIL.BNU.EDU.CN
19 XinJieKouWai St.
State Key Laboratory of Cognitive Neuroscience and Learning,
Beijing Normal University,
Beijing, 100875, P.R. China

Bernhard Schölkopf BERNHARD.SCHOELKOPF@TUEBINGEN.MPG.DE
MPI for Intelligent Systems
Spemannstrasse 38
72076, Tübingen, Germany

Alexander Smola§ ALEX@SMOLA.ORG
Yahoo! Research
2821 Mission College Blvd
Santa Clara, CA 95054, USA

Editor: Nicolas Vayatis

Abstract

We propose a framework for analyzing and comparing distributions, which we use to construct statistical tests to determine if two samples are drawn from different distributions. Our test statistic is the largest difference in expectations over functions in the unit ball of a reproducing kernel Hilbert space (RKHS), and is called the *maximum mean discrepancy* (MMD). We present two distribution-free tests based on large deviation bounds for the MMD, and a third test based on the asymptotic distribution of this statistic. The MMD can be computed in quadratic time, although efficient linear time approximations are available. Our statistic is an instance of an integral probability metric, and various classical metrics on distributions are obtained when alternative function classes are used in place of an RKHS. We apply our two-sample tests to a variety of problems, including attribute matching for databases using the Hungarian marriage method, where they perform strongly. Excellent performance is also obtained when comparing distributions over graphs, for which these are the first such tests.

Maximum Mean Discrepancy (MMD)

$$
\begin{aligned}
\mathrm{MMD}^2(\mu^*, \mu^*_\theta) = \;& \mathbb{E}_{\mathbf{u},\mathbf{u}'\sim\mu^*}[\kappa(\mathbf{u},\mathbf{u}')] \\
&+ \mathbb{E}_{\mathbf{v},\mathbf{v}'\sim\mu^*_\theta}[\kappa(\mathbf{v},\mathbf{v}')] \\
&- 2\mathbb{E}_{\mathbf{u}\sim\mu^*,\mathbf{v}\sim\mu^*_\theta}[\kappa(\mathbf{u},\mathbf{v})]
\end{aligned}
$$

Gretton, Arthur, et al. "A kernel two-sample test." The Journal of Machine Learning Research 13.1 (2012): 723-773.

Maximum Mean Discrepancy (MMD)

- ✓ Admits unbiased estimator.
- ✓ Computed in $\mathcal{O}(n^2)$.
- ✓ Entails convergence properties.
- ✓ Enjoys parametric rates of estimation: $\mathcal{O}(1/\sqrt{n})$ sampling error.

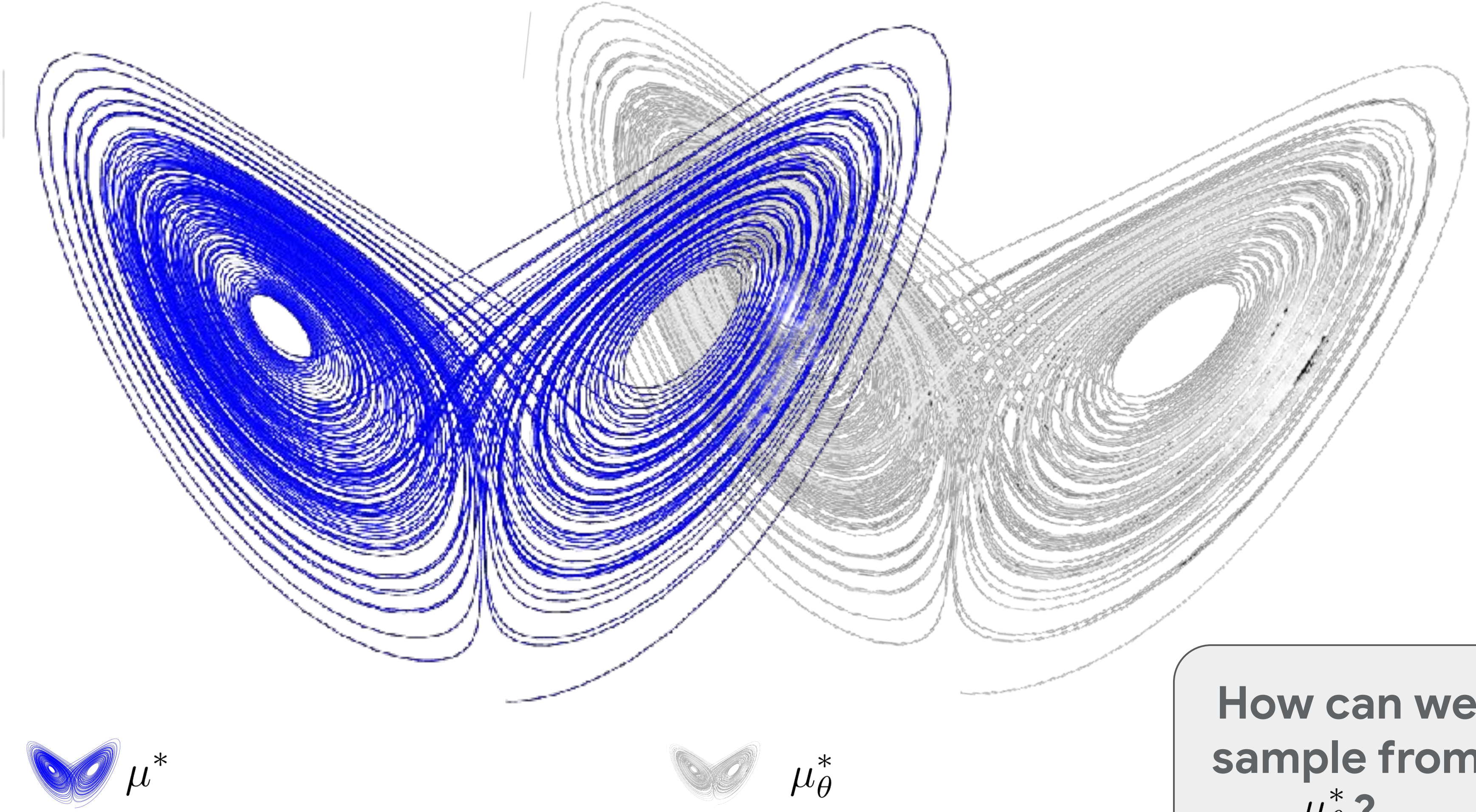
$$\mathcal{L}_{\lambda}^{\mathrm{D}}(\theta) = \mathcal{L}(\theta) + \lambda \mathrm{D}(\mu^*, \mu_{\theta}^*)$$

How can we sample from μ_{θ}^* ?

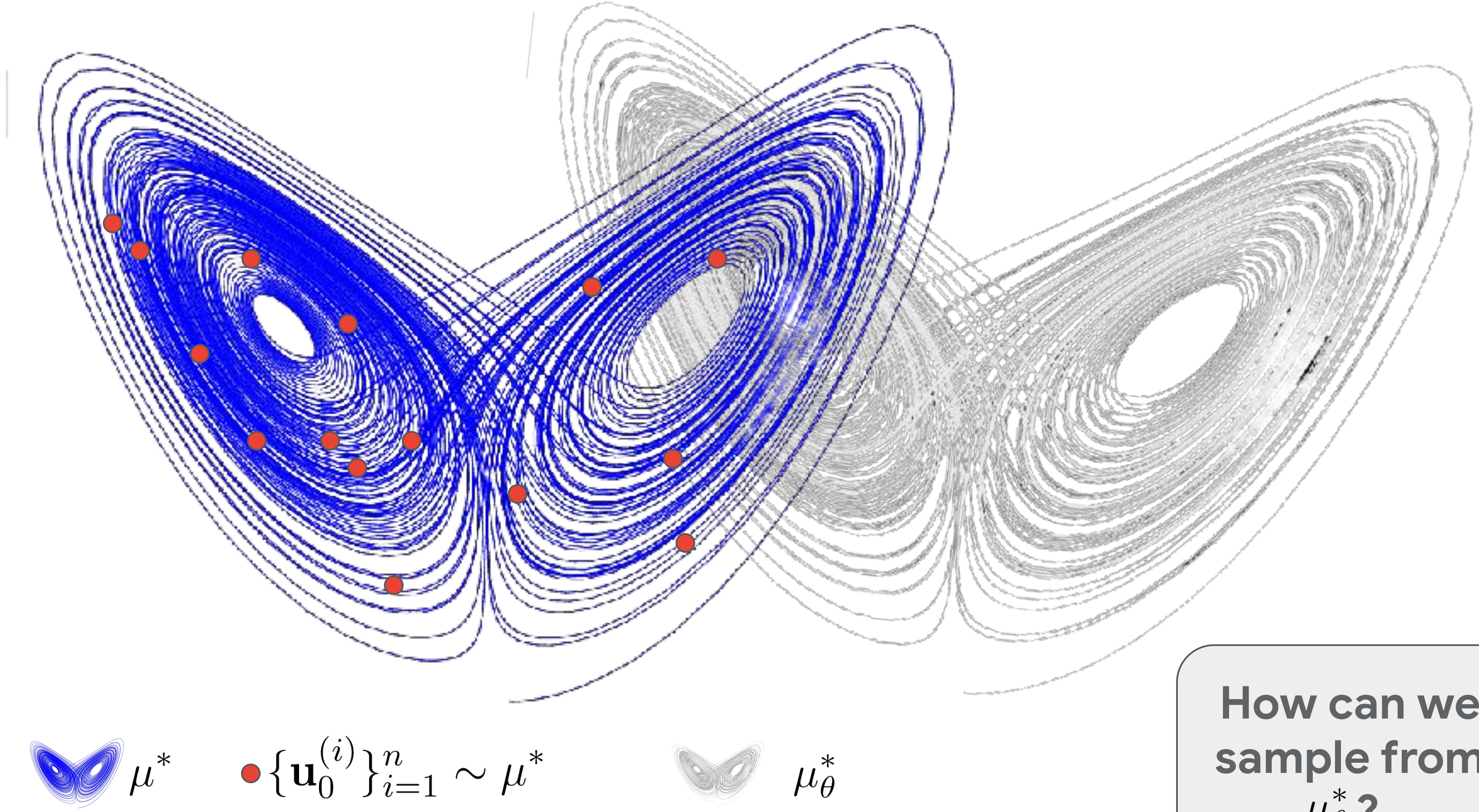
How can we sample from μ_θ^* ?

Assuming the learned model has an attractor, we sample from it by unrolling learned model “far enough” in time:

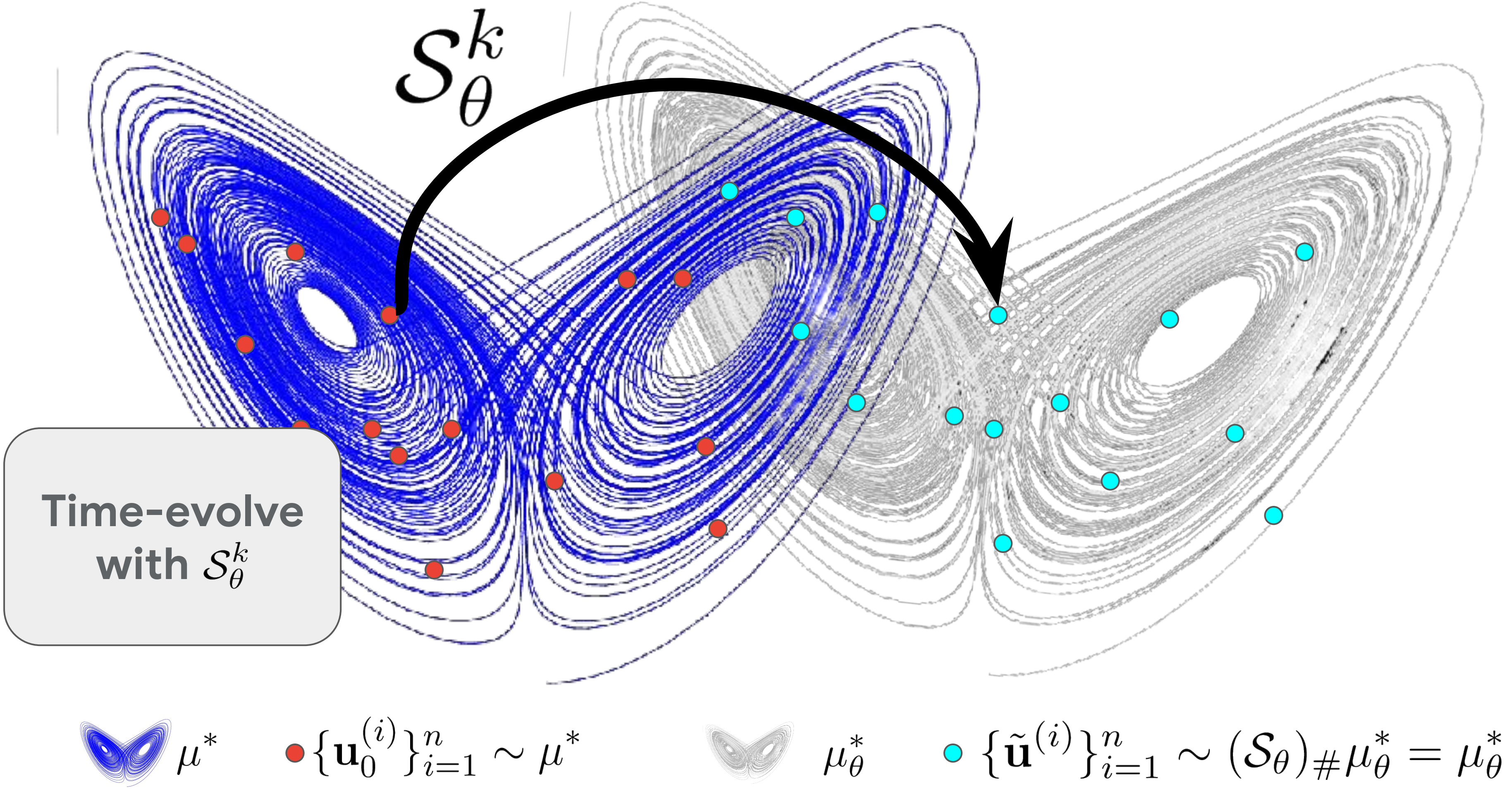
$$(S_\theta^k)_\# \mu^* \approx \mu_\theta^*$$



How can we sample from μ_θ^*?



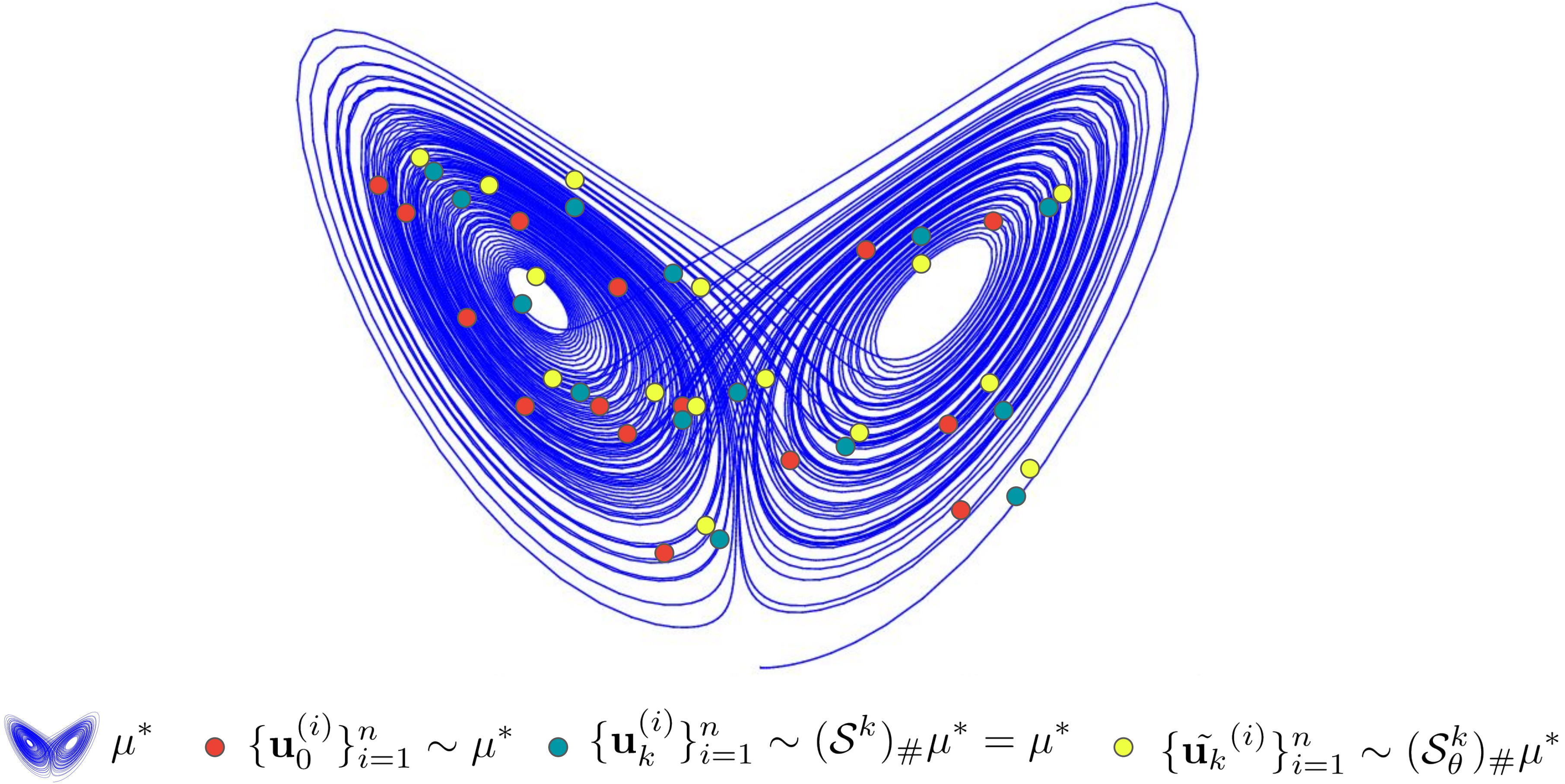
How can we sample from μ_θ^*?


$\mathcal{S}_\theta^k$
Time-evolve with $\mathcal{S}_\theta^k$
μ^*
$\{\mathbf{u}_0^{(i)}\}_{i=1}^n \sim \mu^*$
μ_θ^*
$\{\tilde{\mathbf{u}}^{(i)}\}_{i=1}^n \sim (\mathcal{S}_\theta)_\# \mu_\theta^* = \mu_\theta^*$

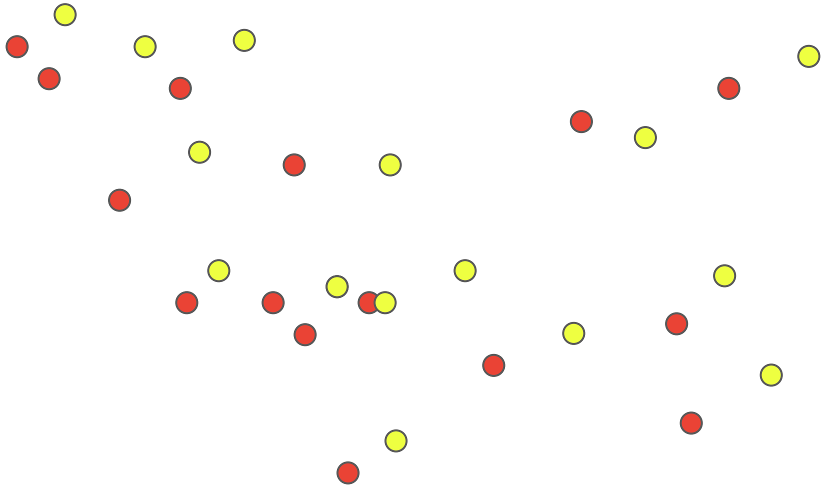
$$\mu^* = (\mathcal{S}^k)_\# \mu^*$$

$$\mu^* = (\mathcal{S}^k)_{\#}\mu^*$$

$$\mathrm{D}(\mu^*, \mu_\theta^*) = \mathrm{D}((\mathcal{S}^k)_{\#}\mu^*, \mu_\theta^*)$$

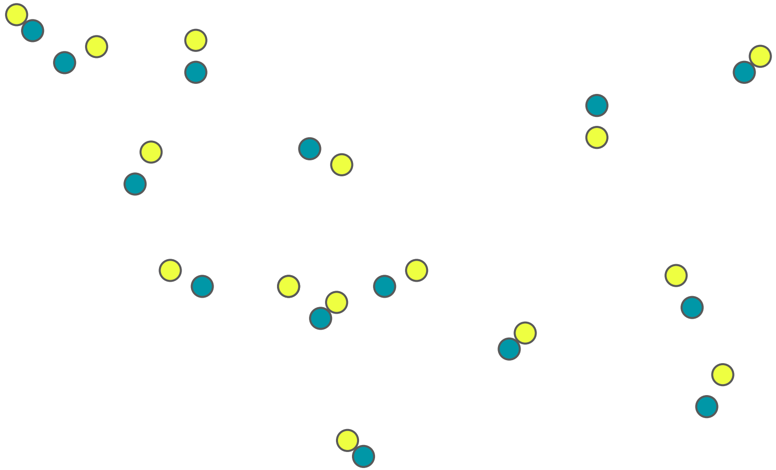

μ^*
$\{\mathbf{u}_0^{(i)}\}_{i=1}^n \sim \mu^*$
$\{\mathbf{u}_k^{(i)}\}_{i=1}^n \sim (\mathcal{S}^k)_\# \mu^* = \mu^*$
$\{\tilde{\mathbf{u}}_k^{(i)}\}_{i=1}^n \sim (\mathcal{S}_\theta^k)_\# \mu^*$

“Unconditional” regularization $\mathrm{D}(\mu^*, \mu_\theta^*)$



$\{\mathbf{u}_0^{(i)}\}_{i=1}^n \sim \mu^*$ $\{\tilde{\mathbf{u}_k}^{(i)}\}_{i=1}^n \sim (\mathcal{S}_\theta^k)_\# \mu^*$

"Conditional" regularization $\mathrm{D}((\mathcal{S}^k)_{\#}\mu^*, \mu_\theta^*)$



● $\{\mathbf{u}_k^{(i)}\}_{i=1}^n \sim (\mathcal{S}^k)_{\#}\mu^* = \mu^*$ ● $\{\tilde{\mathbf{u}}_k{}^{(i)}\}_{i=1}^n \sim (\mathcal{S}_\theta^k)_{\#}\mu^*$

DySLIM

$$\widehat{\mathcal{L}}_{\lambda}^{\mathrm{D}} = \widehat{\mathcal{L}} + \lambda_1 \widehat{\mathrm{D}}(\mu^*, (\mathcal{S}_{\theta}^{k})_{\#}\mu^*) + \lambda_2 \widehat{\mathrm{D}}((\mathcal{S}^{k})_{\#}\mu^*, (\mathcal{S}_{\theta}^{k})_{\#}\mu^*)$$

DySLIM

$$\widehat{\mathcal{L}}_{\lambda}^{\mathrm{D}} = \widehat{\mathcal{L}} + \lambda_1 \widehat{\mathrm{D}}(\mu^*, (\mathcal{S}_\theta^k)_{\#}\mu^*) + \lambda_2 \widehat{\mathrm{D}}((\mathcal{S}^k)_{\#}\mu^*, (\mathcal{S}_\theta^k)_{\#}\mu^*)$$

Trajectory matching

DySLIM

$$\widehat{\mathcal{L}}_{\lambda}^{\mathrm{D}} = \widehat{\mathcal{L}} + \lambda_1 \widehat{\mathrm{D}}(\mu^*, (\mathcal{S}_\theta^k)_\# \mu^*) + \lambda_2 \widehat{\mathrm{D}}((\mathcal{S}^k)_\# \mu^*, (\mathcal{S}_\theta^k)_\# \mu^*)$$

"Unconditional" regularization

DySLIM

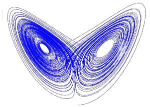
$$\widehat{\mathcal{L}}^{\mathrm{D}}_{\lambda} = \widehat{\mathcal{L}} + \lambda_1 \widehat{\mathrm{D}}(\mu^*, (\mathcal{S}^k_\theta)_\# \mu^*) + \lambda_2 \widehat{\mathrm{D}}((\mathcal{S}^k)_\# \mu^*, (\mathcal{S}^k_\theta)_\# \mu^*)$$

"Conditional" regularization

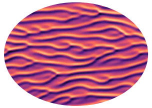
DySLIM

- **System-agnostic** measure-matching regularization.
- **Tractable**, **sample & compute-efficient.**
- Capable of tackling **larger and more complex systems.**

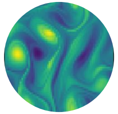
Experiments



Lorenz 63



Kuramoto–Sivashinsky



Kolmogorov Flow

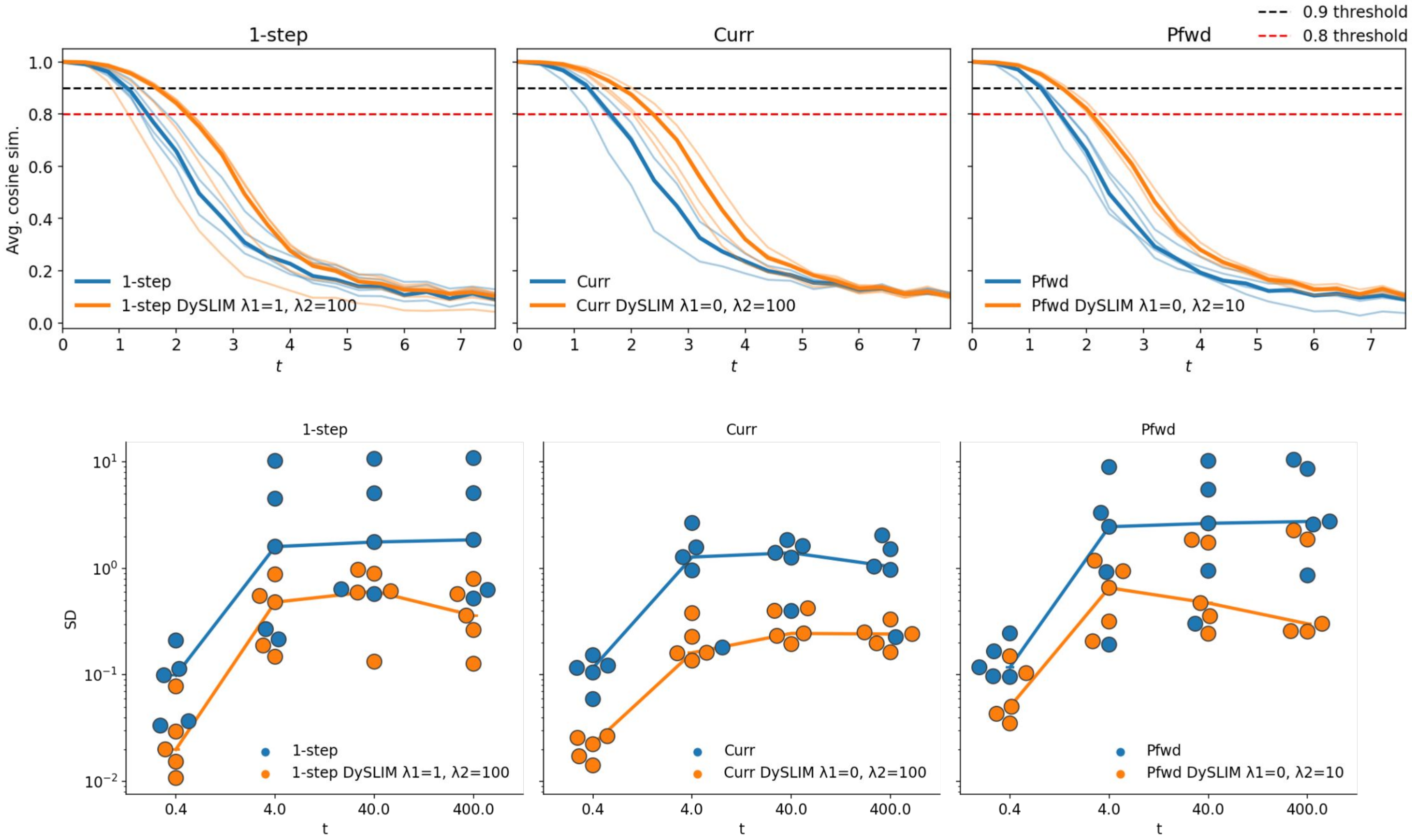
Animation credit: https://johncarlosbaez.wordpress.com/2021/10/17/conjectures-on-the-kuramoto-sevashinsky-equation/

Baselines: Unregularized objectives

1-step	$\|\|\mathcal{S}_\theta(\mathbf{u}) - \mathcal{S}(\mathbf{u})\|\|$
Curriculum	$\sum_{k=1}^{\ell} \|\|\mathcal{S}_\theta^k(\mathbf{u}) - \mathcal{S}^k(\mathbf{u})\|\|$
Pushforward	$\|\|\mathcal{S}_\theta(\mathbf{sg}(\mathcal{S}_\theta^{\ell-1}(\mathbf{u}))) - \mathcal{S}^\ell(\mathbf{u})\|\|$

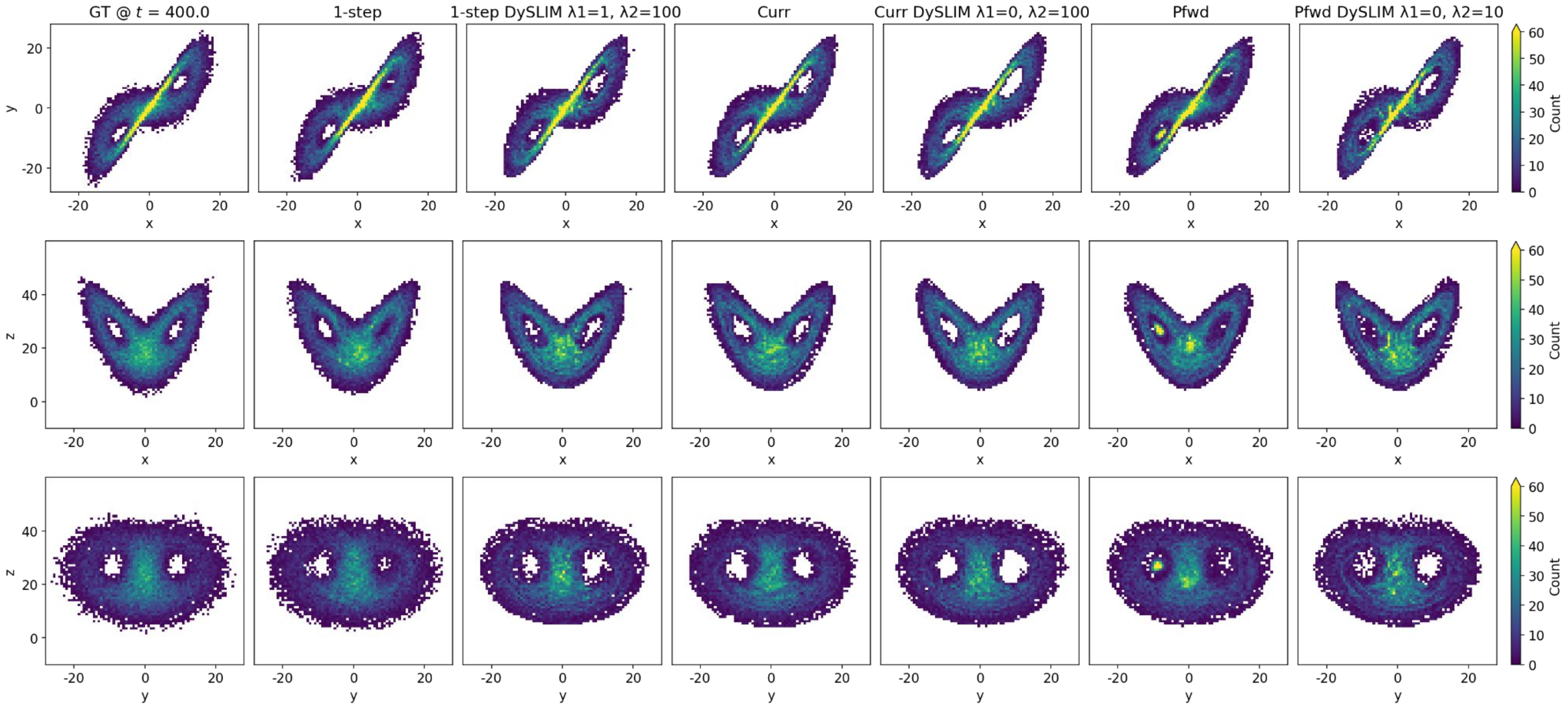
Brandstetter, Johannes, Daniel Worrall, and Max Welling. "Message passing neural PDE solvers." arXiv preprint arXiv:2202.03376 (2022).

Lorenz 63



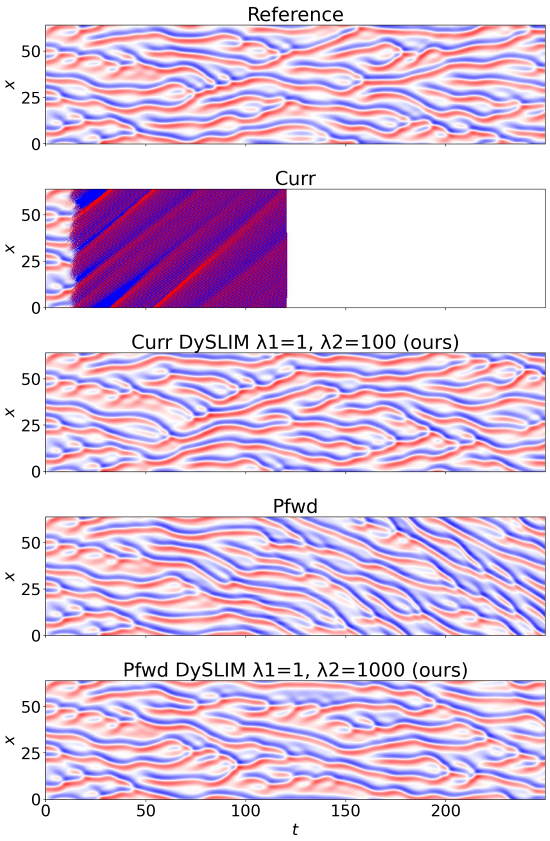
1-step
Curr
Pfwd
0.9 threshold
0.8 threshold
Avg. cosine sim.
t
1-step
1-step DySLIM λ1=1, λ2=100
Curr
Curr DySLIM λ1=0, λ2=100
Pfwd
Pfwd DySLIM λ1=0, λ2=10
SD
0.4
4.0
40.0
400.0

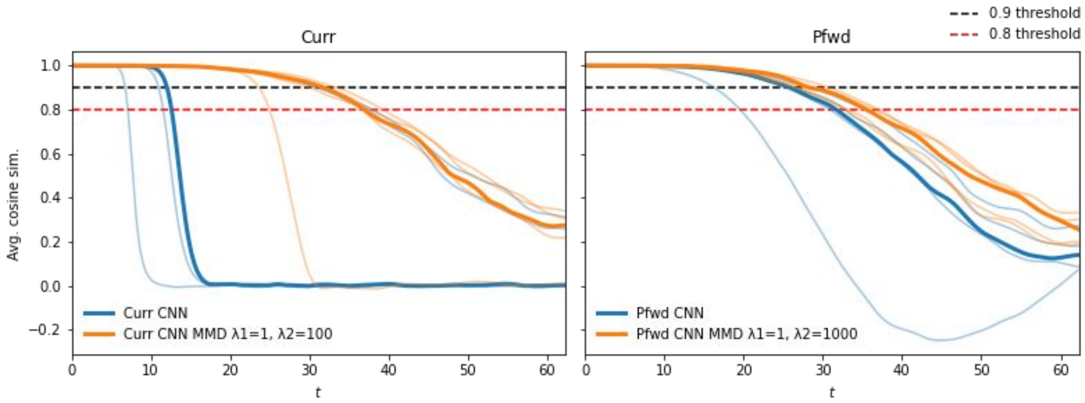
Lorenz 63

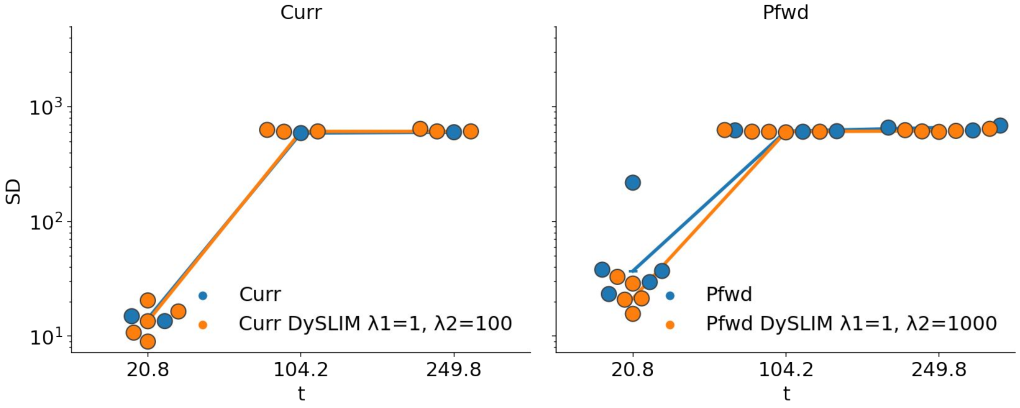


GT @ t = 400.0
1-step
1-step DySLIM λ1=1, λ2=100
Curr
Curr DySLIM λ1=0, λ2=100
Pfwd
Pfwd DySLIM λ1=0, λ2=10
x
y
z
Count

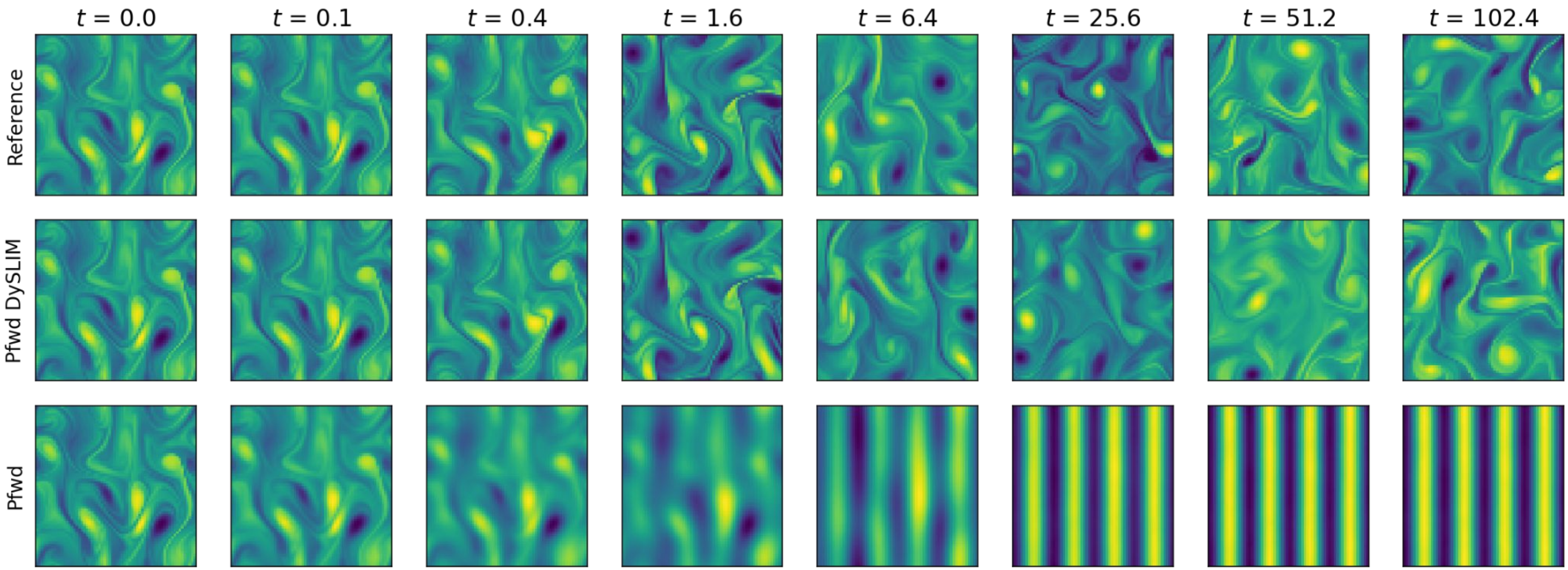
Kuramoto–Sivashinsky







Kolmogorov Flow



Kolmogorov Flow

Table 1. Kolmogorov flow: Metrics for 1-step, curriculum, and pushforward objectives without and with regularization ($\lambda_1 = 0$, $\lambda_2 = 100$). Boldface numbers indicate that the metric is improved by our regularization. All values displayed are in units of $\times 10^{-2}$.

	Batch size	LR	MELR (↓)		MELRw (↓)		covRMSE (↓)		Wass1 (↓)		TCM (↓)	
			Base	DySLIM	Base	DySLIM	Base	DySLIM	Base	DySLIM	Base	DySLIM
1-step	64	5e-4	2.77	**1.84**	0.44	0.85	7.93	**7.30**	16.2	**5.55**	5.39	**2.45**
Curr	64	5e-4	5.35	**1.64**	0.95	**0.45**	8.13	**6.95**	9.66	**4.76**	3.50	**2.83**
Pfwd	128	1e-4	3.19	**2.46**	0.53	0.53	6.81	**6.69**	4.64	**4.51**	3.68	**0.72**

Summary

- **Leveraged key property of ergodic systems**: supporting invariant measure.

- **Introduced DySLIM**: a scalable and system-agnostic measure-matching regularization.

- **Demonstrated that both short-term predictive capabilities and long-term stability can be improved** across a range of well-studied systems. (Lorenz 63, Kuramoto–Sivashinsky, and Kolmogorov Flows.)

Thank you!

Yair Schiff

Zhong Yi Wan

Jeffrey Parker

Stephan Hoyer

Volodymyr Kuleshov

Fei Sha

Leonardo Zepeda-Núñez

Thank you!

Tue 23 Jul 1:30 p.m. — 3 p.m. CEST
Hall C 4-9 #1003

[2402.04467] DySLIM: Dynamics Stable Learning by Invariant Measure for Chaotic Systems

https://github.com/google-research/swirl-dynamics/tree/main/swirl_dynamics/projects/ergodic