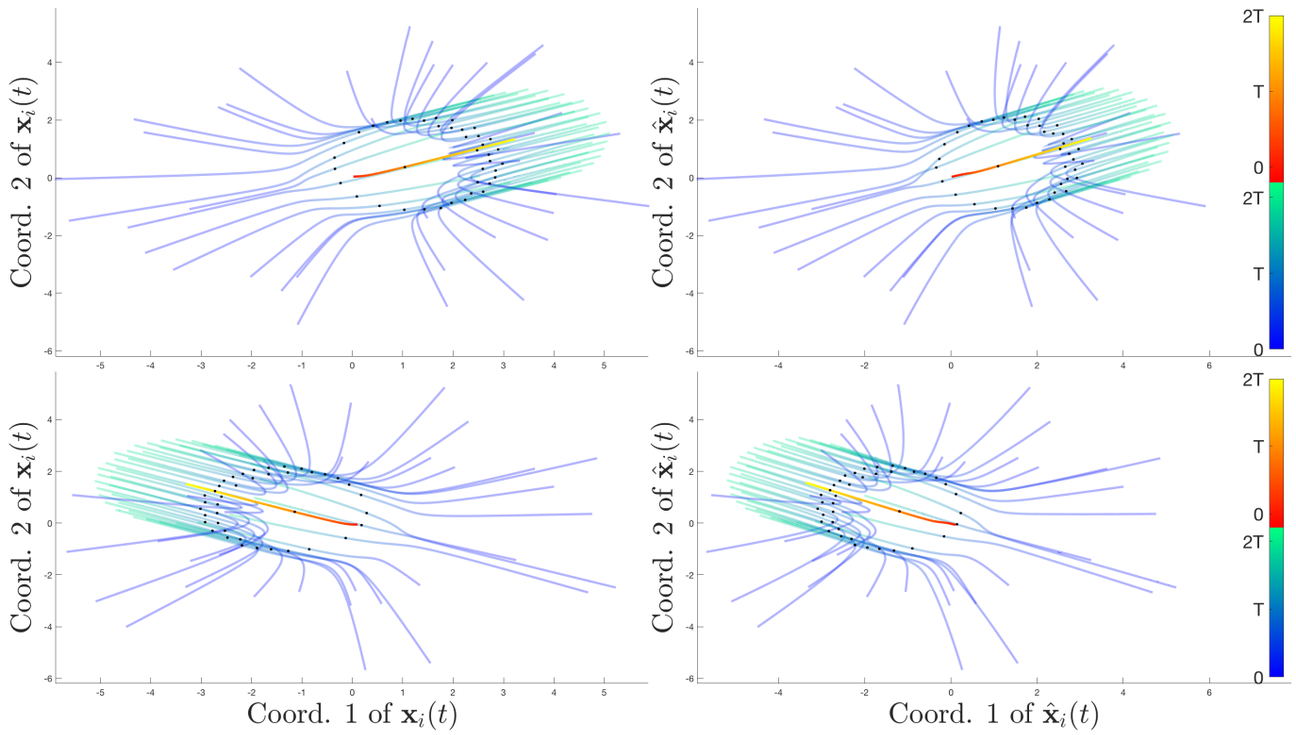
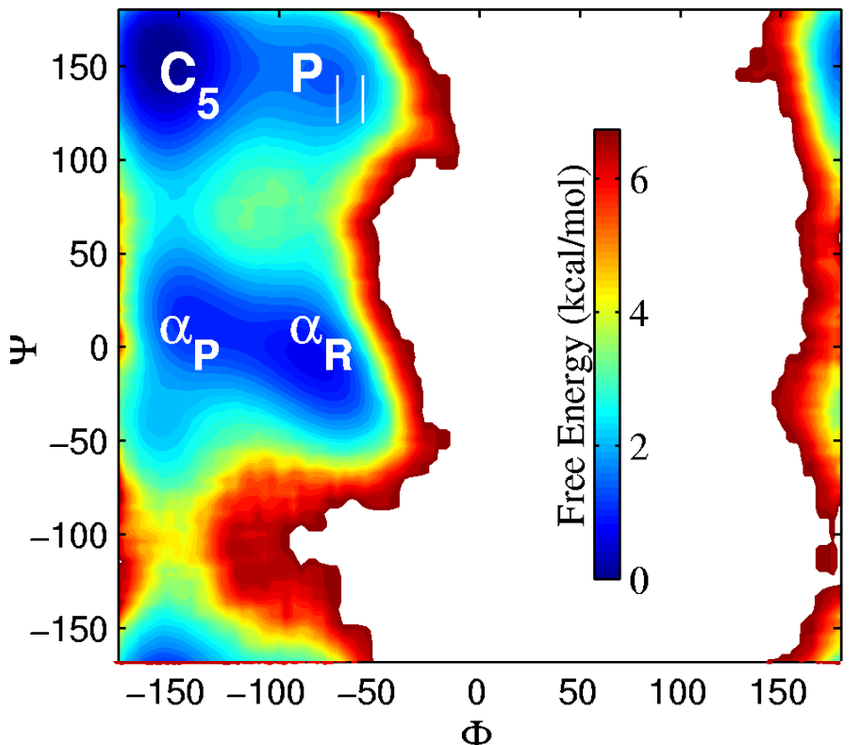
Learning and Geometry for Dynamical Systems:
Agent-based Systems & Diffusions on manifolds

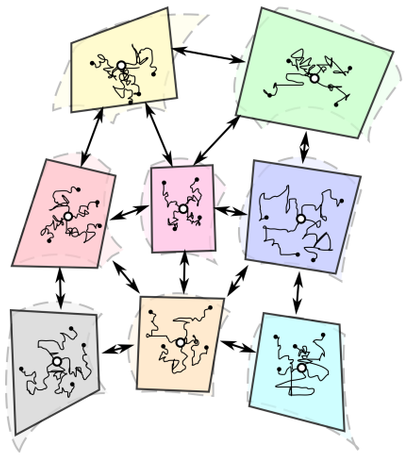
Mauro Maggioni

Johns Hopkins University

CECAM - *Coarse-graining with Machine Learning in molecular dynamics*







NSF, AFOSR

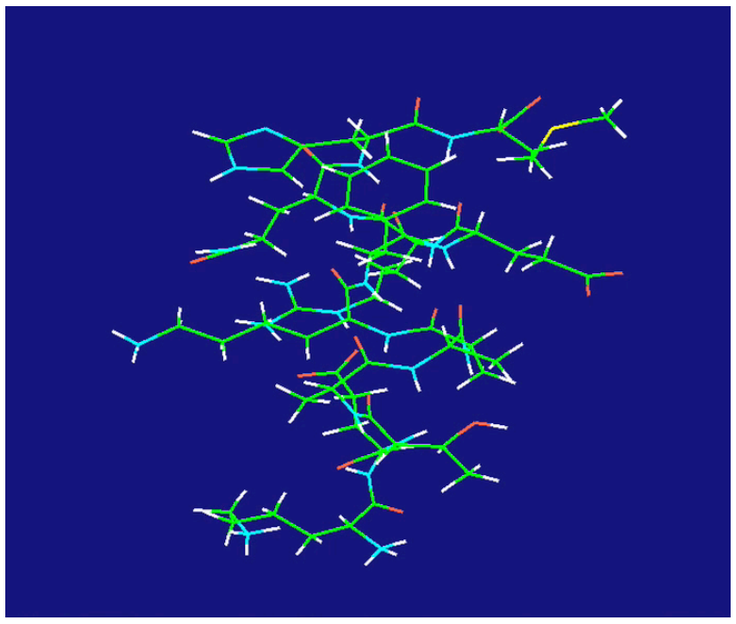
Machine Learning & Physical Systems

Many physical systems have very high-dimensional state spaces, are governed by a very large number of ODE's (or SDE's), which make them difficult to analyze. Homogenization, mean-field approximation, renormalization theory etc...are techniques to simplify such systems.

Physical systems with high-dimensional state spaces (e.g. many-particle systems) may exhibit behavior that is complex and high dimensional. Can Machine Learning help extract useful reductions?

Challenges for ML: learning principles that transfer across physical systems; incorporate existing physical knowledge or constraints.

From https://www.youtube.com/watch?v=bb9ZTbYGRdc



Two stories

Learning interaction kernels of agent-based systems. Given trajectories of a system of interacting agents, that may exhibit emergent behavior (e.g. flocking), can we learn interaction kernels, in a flexible non-parametric fashion, without being cursed by the high dimension of the state space?

Model reduction for stochastic processes (diffusions and Lengevin dynamics) on manifolds. Given the ability of sampling inital conditions and short paths, learn a stochastic process on a manifold (with both the process and the manifold being unknown) that approximates the original not only at short time scales, but also at long time scales. Combination of manifold learning and learning of SDE's.

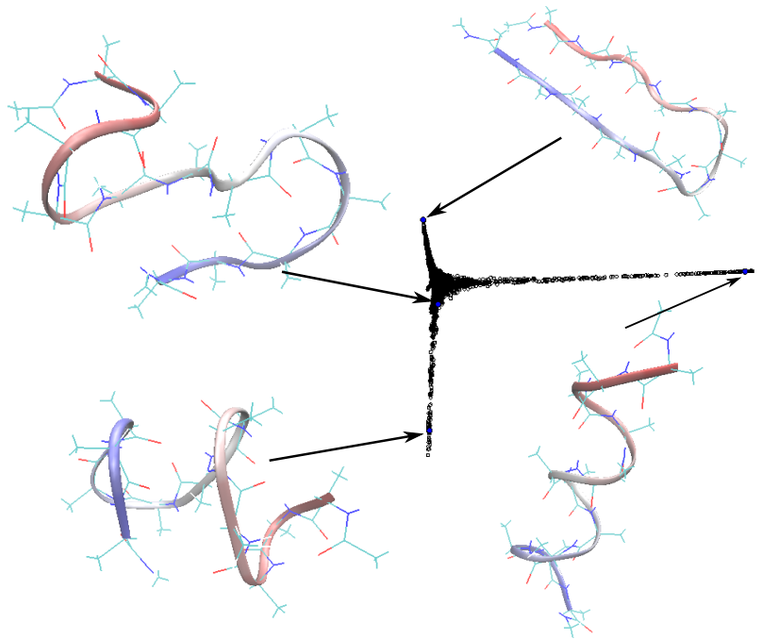
From https://www.youtube.com/watch?v=bb9ZTbYGRdc

Model Reduction

Many physical systems have very high-dimensional state spaces, are governed by a very large number of ODE's (or SDE's), which make them difficult to analyze. Homogenization, mean-field approximation, renormalization theory etc...are techniques to simplify such systems.

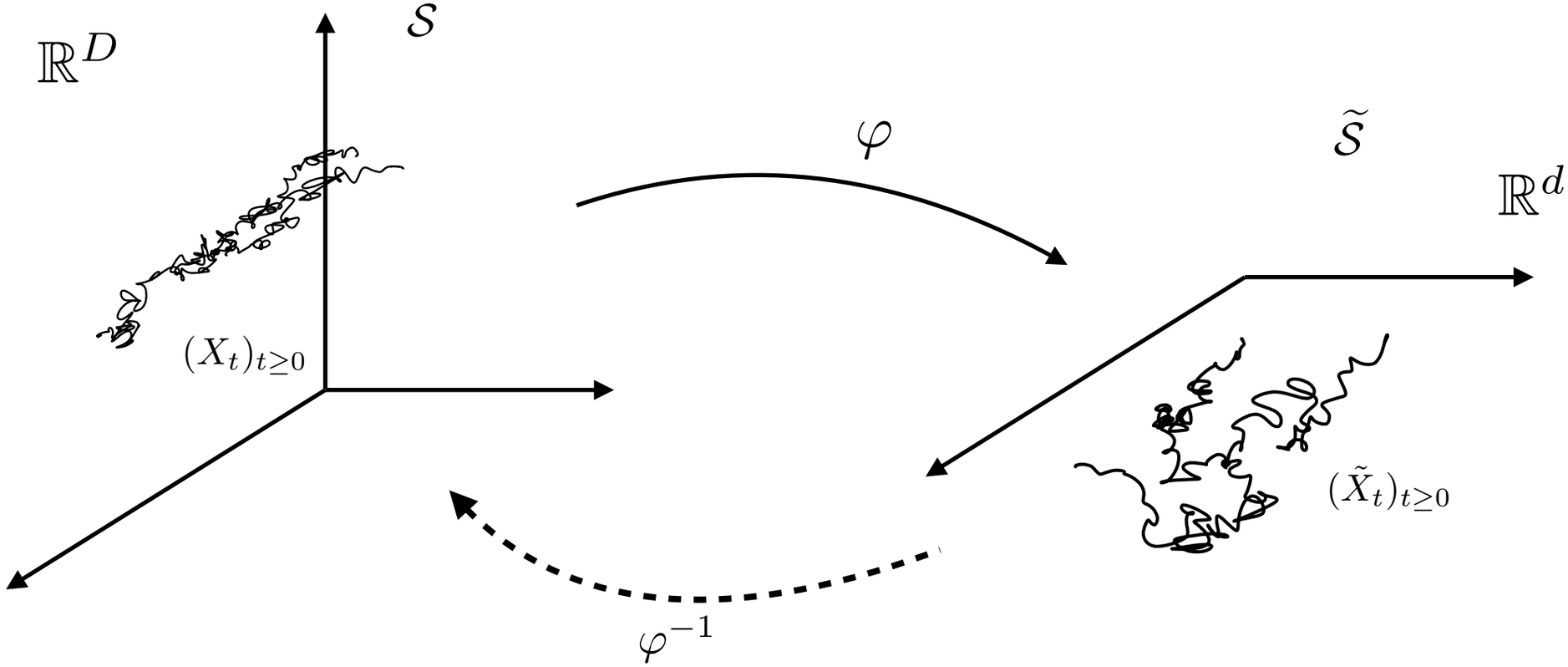
Noise, stochasticity, ergodicity, separation of time scales, multiscale characteristics help.

Wide variety of applications, from weather and atmosphere modeling, to molecular dynamics, to quantum control



Model Reduction

Problem: given observations of the system $(\mathbf{X}_t)_{t\geq 0} \subseteq \mathbb{R}^D$, whose trajectories are determined by an unknown system $\mathcal{S}$ of ODE's (or SDE's) construct a map $\varphi : \mathbb{R}^D \to \mathbb{R}^d$, a system $\widetilde{\mathcal{S}}$ in $\mathbb{R}^d$ such that trajectories $(\tilde{X}_t)_{t\geq 0}$ of $\widetilde{\mathcal{S}}$, when lifted by φ^{-1}, are "close" to those of $\mathcal{S}$.

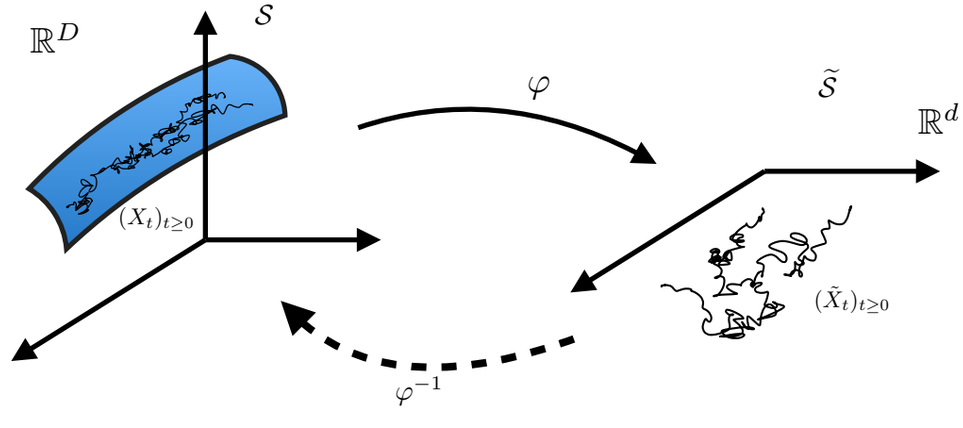


(Partial) list of obstructions

- High-dimension D of state space.
- Complex dynamics.
- Metastability.
- Roughness.

Curse of dimensionality: estimation is difficult in D dimensions, for D large. E.g. to approximate a Lip 1 function f on the unit cube $[0,1]^D$ to accuracy ϵ one needs at least $O(1)$ points in each little cube of size ϵ, and there $O(\epsilon^{-D})$ such (disjoint) cubes.

Assumption: "important aspects" of the dynamics, at least at time scales not too small, are captured by a low-dimensional subspace/manifold in the state space. Think: attractors, inertial manifolds...



(Partial) list of obstructions

- High-dimension D of state space.
- Complex dynamics.
- Metastability.
- Roughness.

Dynamics may be complex, chaotic, highly dependent on initial conditions. In that case even simulating the original system $\mathcal{S}$ is hard, except perhaps for very short times.

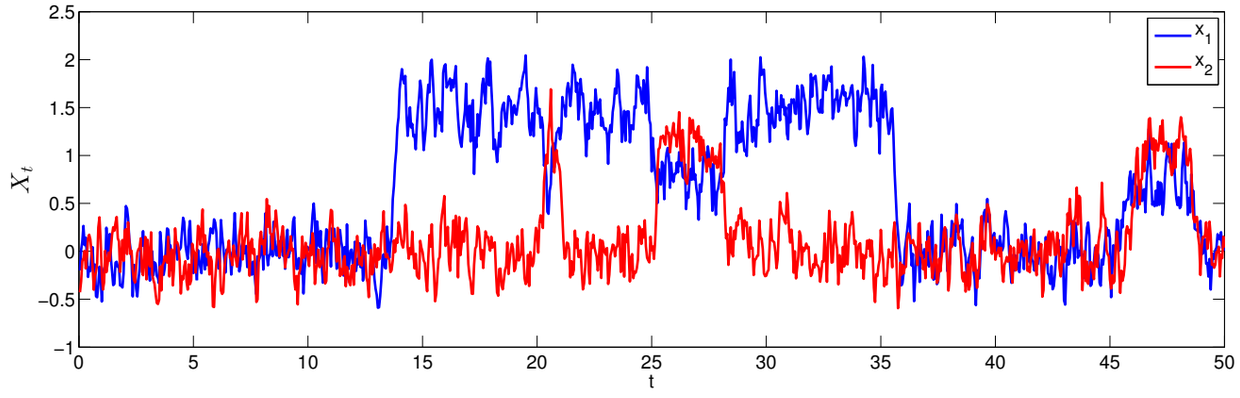
Goal: interested in smooth, large time averages and functionals.

(Partial) list of obstructions

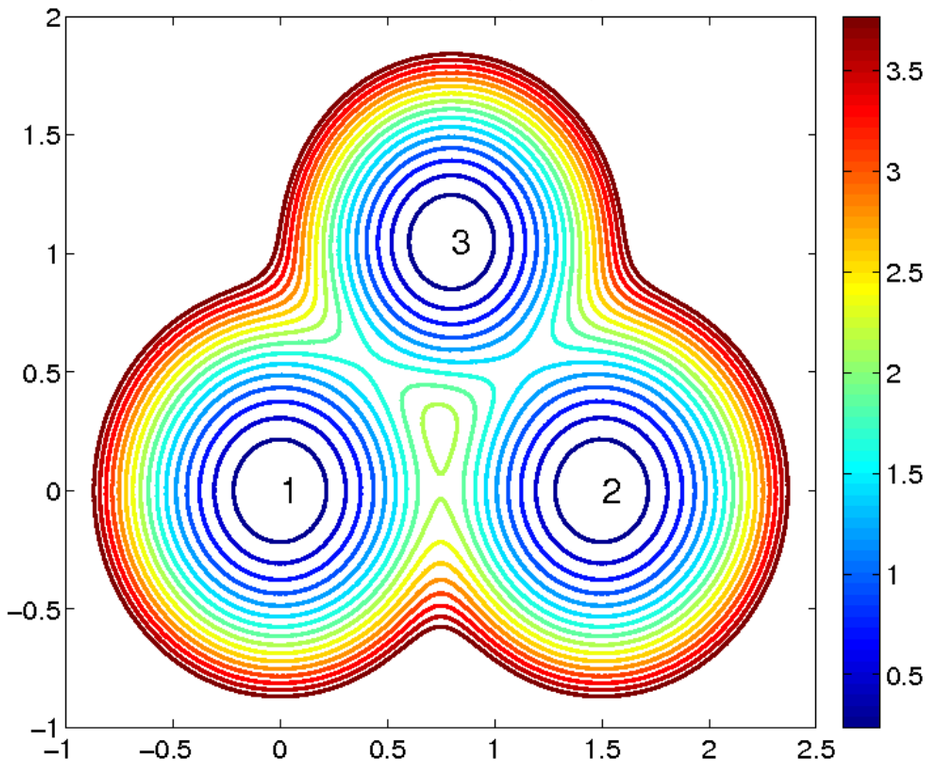
- High-dimension D of state space.
- Complex dynamics.
- **Metastability.**
- Roughness.

System $\mathcal{S}$ trapped for long times before transitioning to different metastable states.

Goal: make $\mathcal{S}'$ really fast to simulate.



$$dX_t = -\nabla U(X_t) + dB_t$$

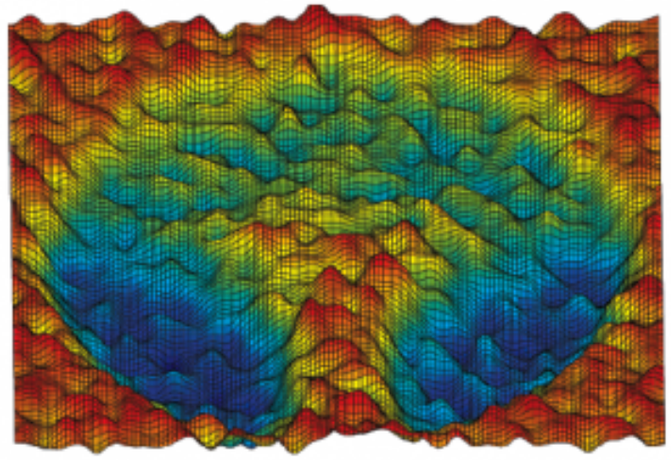


(Partial) list of obstructions

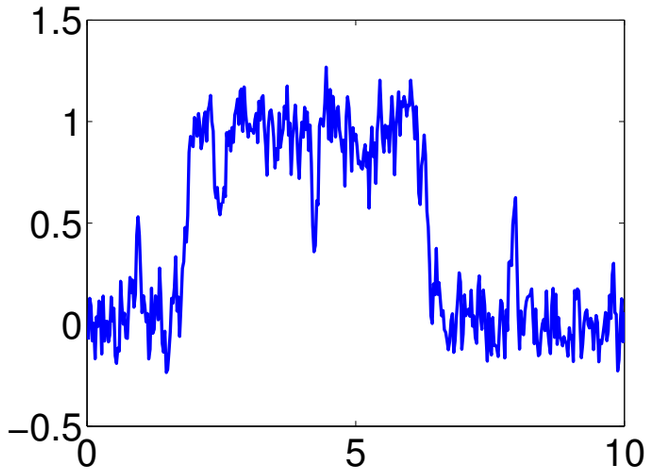
- High-dimension D of state space.
- Complex dynamics.
- Metastability.
- Roughness.

Roughness forces small times in simulator unless implicit schemes (expensive) are used.

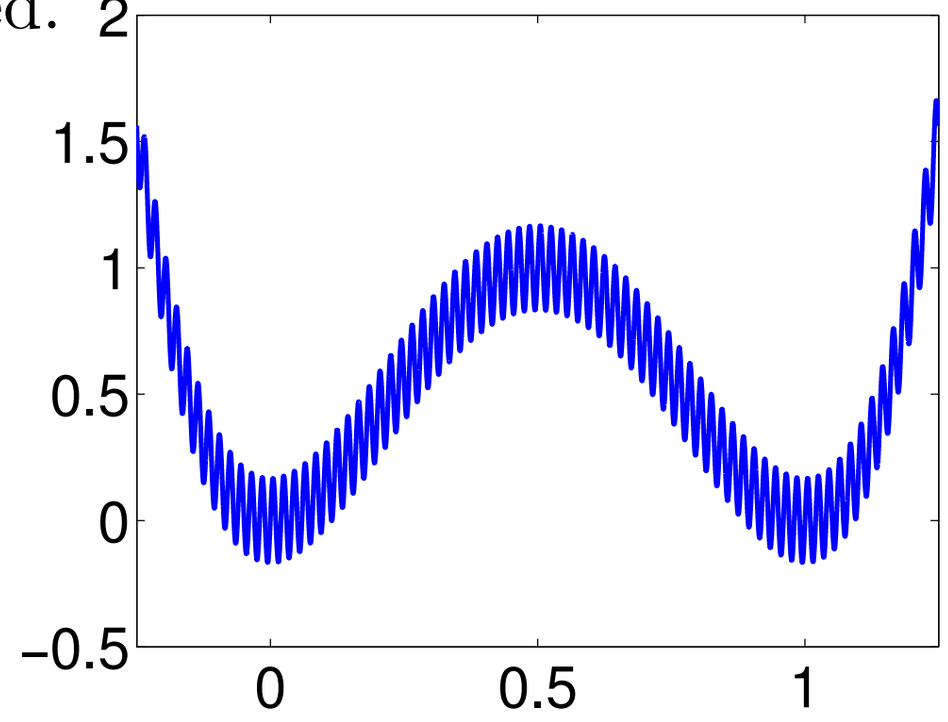
Goal: (empirical) homogenization above fine time scales can help.



source: IPAM



$$dX_t = -\nabla U(X_t) + dB_t$$



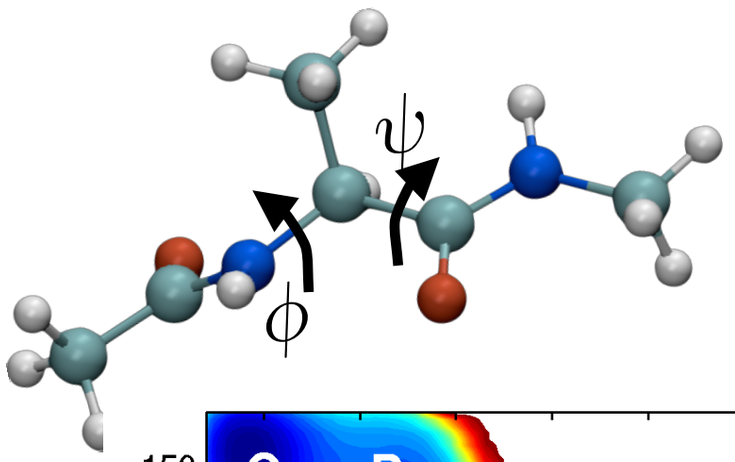
Example: Molecular Dynamics

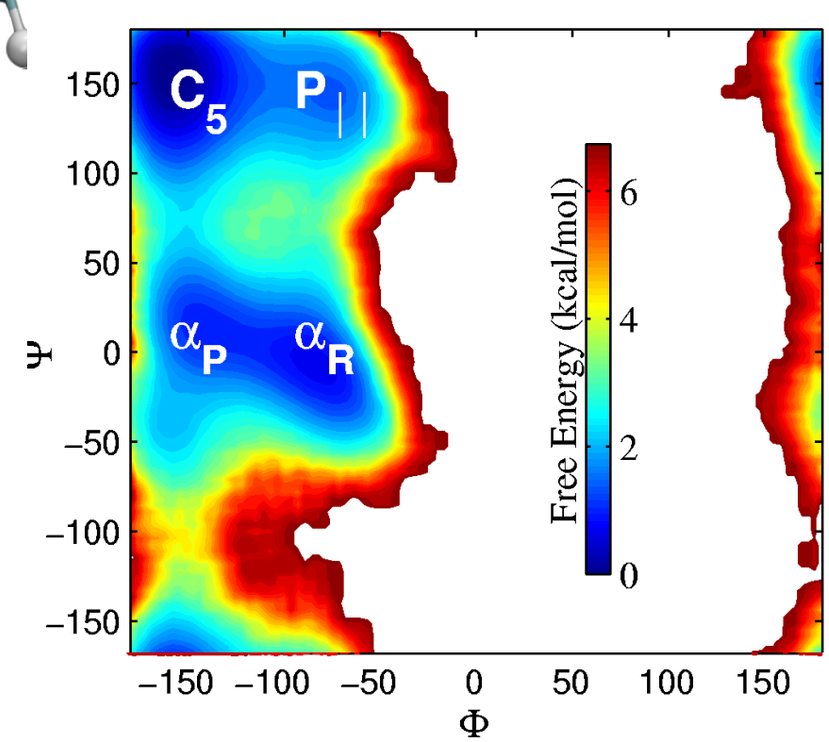

M. Rohrdanz, W. Zheng, MM,
C. Clementi, *Journ. Chem. Phys.*

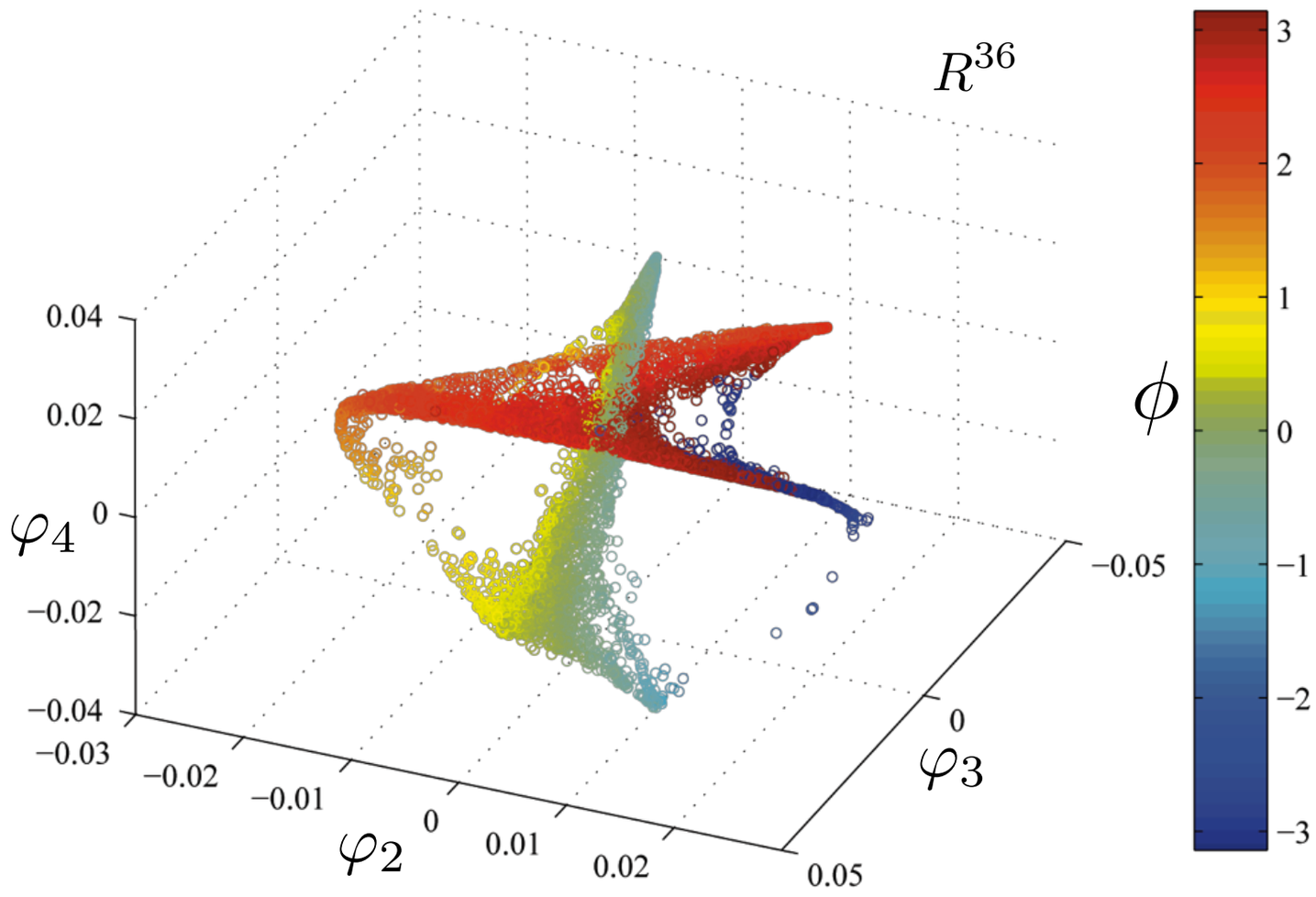
The dynamics of a small peptide (12 atoms with H-atoms removed) in a bath of water molecules, is approximated by a Langevin system of stochastic equations

$$\dot{x} = -\nabla U(x) + \dot{w}$$

The set of configurations is a point cloud in $\mathbb{R}^{12\times 3}$.







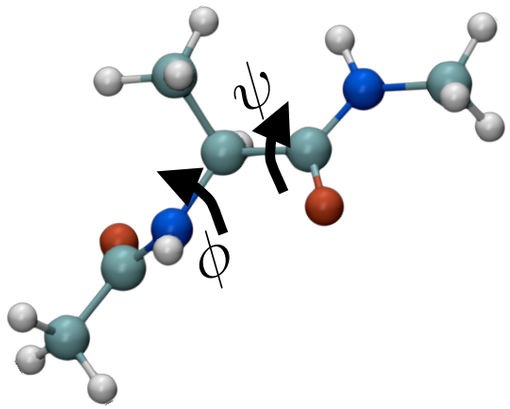
Molecular Dynamics data for Alanine

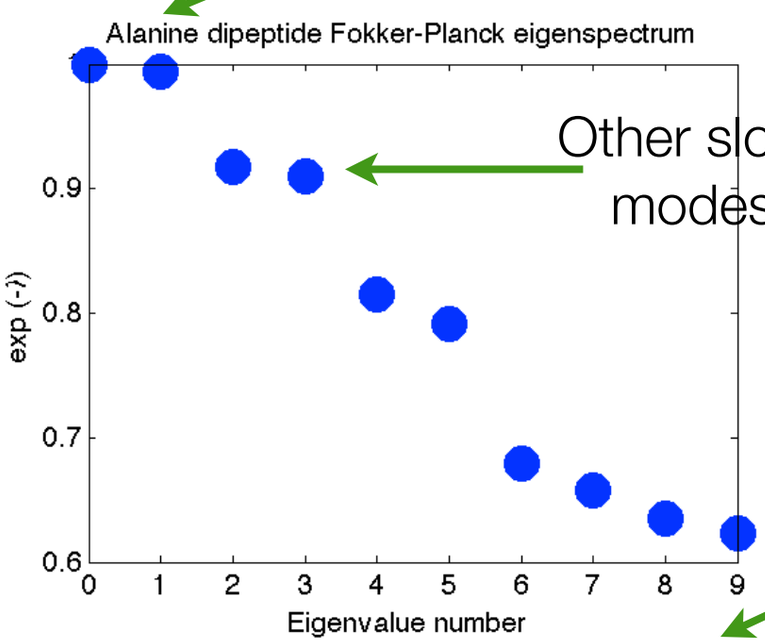
Given simulated data, for very long trajectories, we construct an empirical approximation to the generator of the Fokker-Planck, and compute its eigenvalues/vectors to obtain a low-dimensional embedding and reaction coordinates.

R. Coifman, S. Lafon, MM, B. Nadler, Y. Kevrekidis, *PNAS, JMMS, ACHA*
M. Rohrdanz, W. Zheng, MM, C. Clementi, *JCP*

$$\dot{x} = -\nabla U(x) + \sqrt{\frac{2}{\beta}}\dot{w}$$

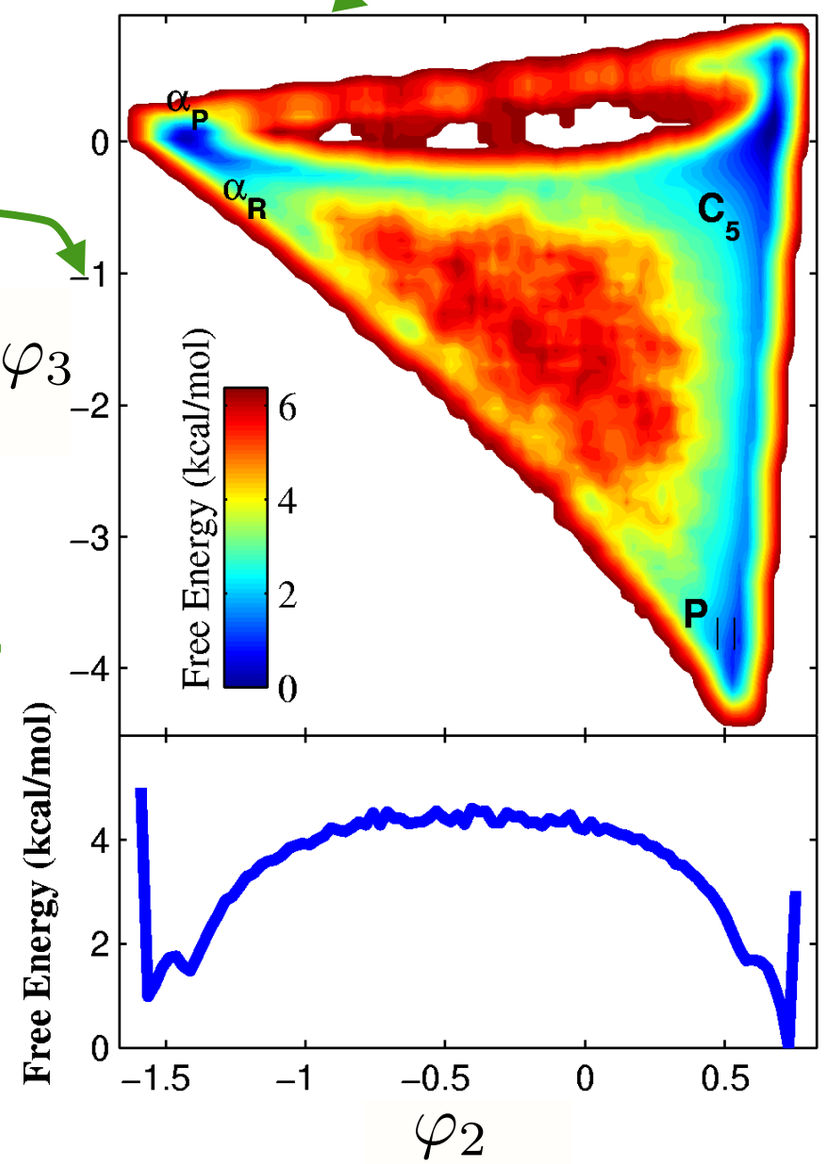
$$\frac{\partial p}{\partial t} = -\sum_{i}^{3N} \frac{\partial}{\partial x_i}\left(\frac{1}{\beta}\frac{\partial}{\partial x_i} + \frac{\partial E}{\partial x_i}\right) p = -\mathbf{H}_{\mathrm{FP}} p$$





Slowest mode

Other slow modes



Coordinate	$C_5, P_{\|} \to \alpha_R\ \alpha_P$	$\alpha_R\ \alpha_P \to C_5, P_{\|}$
From simulation[b]	0.023	0.047
1st DC	0.023 ± 0.001	0.048 ± 0.003
Ψ	0.020 ± 0.001	0.040 ± 0.003

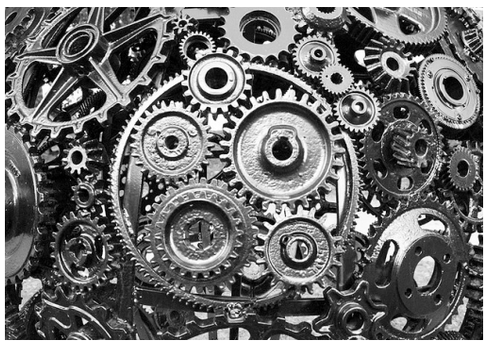
Reduced model with reasonably good rates

$$\text{rate} = \left(\int_{\text{barrier}} \frac{e^{\beta F(x)}}{D(x)} dx \int_{\text{well}} e^{-\beta F(x')} dx'\right)^{-1}$$

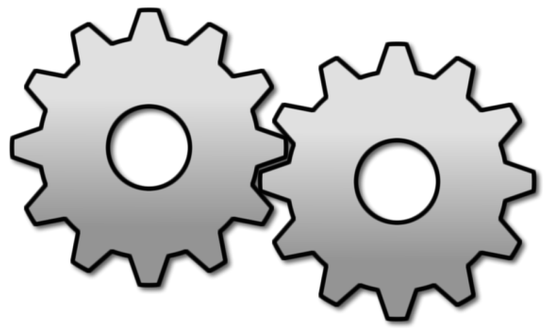
G. Hummer, 2005

Learning a Reduced Model

- Stochastic systems of interest: high dimensional, ergodic, well-approximated by first order Langevin equations on a low-dimensional manifold, at least at a certain time/spatial scale.
- Examples may include MCMC samplers, systems with fast variables, molecular dynamics.
- **Given**: ability to sample the effective state space; ability to obtain *short* paths from a given initial condition; a spatial scale parameter for homogenization,
- **Output**: a *fast simulator* of the system, with large time guarantees.



Complex, expensive simulator,
at very fine (time)scale

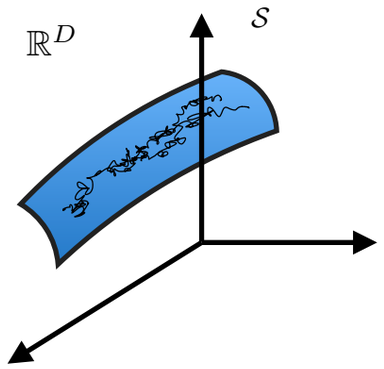


Simple, fast simulator, accurate
at coarse/long (time)scale

Objectives

Miles Crosskey

Assumption: SDE (Ito diffusion) on manifold $\mathcal{M}$



Given:

(i) ability to sample from some measure μ_0 on $\mathcal{M}$

(ii) ability to call a stochastic simulator for the system $\mathcal{S}$ concentrated around a manifold $\mathcal{M}$ of dimension d

(iii) parameter $\delta > 0$ representing the smallest spatial scale of interest

(iv) distance function ρ on the state space

Return:

(i) a fast, continuos-time and continuous-space simulator (ATLAS) with accuracy at scale δ and and with accuracy $\tilde{O}(\delta)$ for large times

(ii) ATLAS is constructed using $O(d/\delta^4)$ paths of length $O(\delta)$ from $\mathcal{S}$, collected in parallel, starting at $O(\delta^{-d})$ locations sampled according to μ_0

(iii) efficient storage mechanism for all paths

Large time guarantees

Theorem [M. Crosskey, MM]. Let $\mathcal{M}$ be a closed compact manifold, and suppose X_t is a stochastic process on $\mathcal{M}$ satisfying

$$dX_t = b(X_t)dt + \Sigma(X_t)dB_t$$

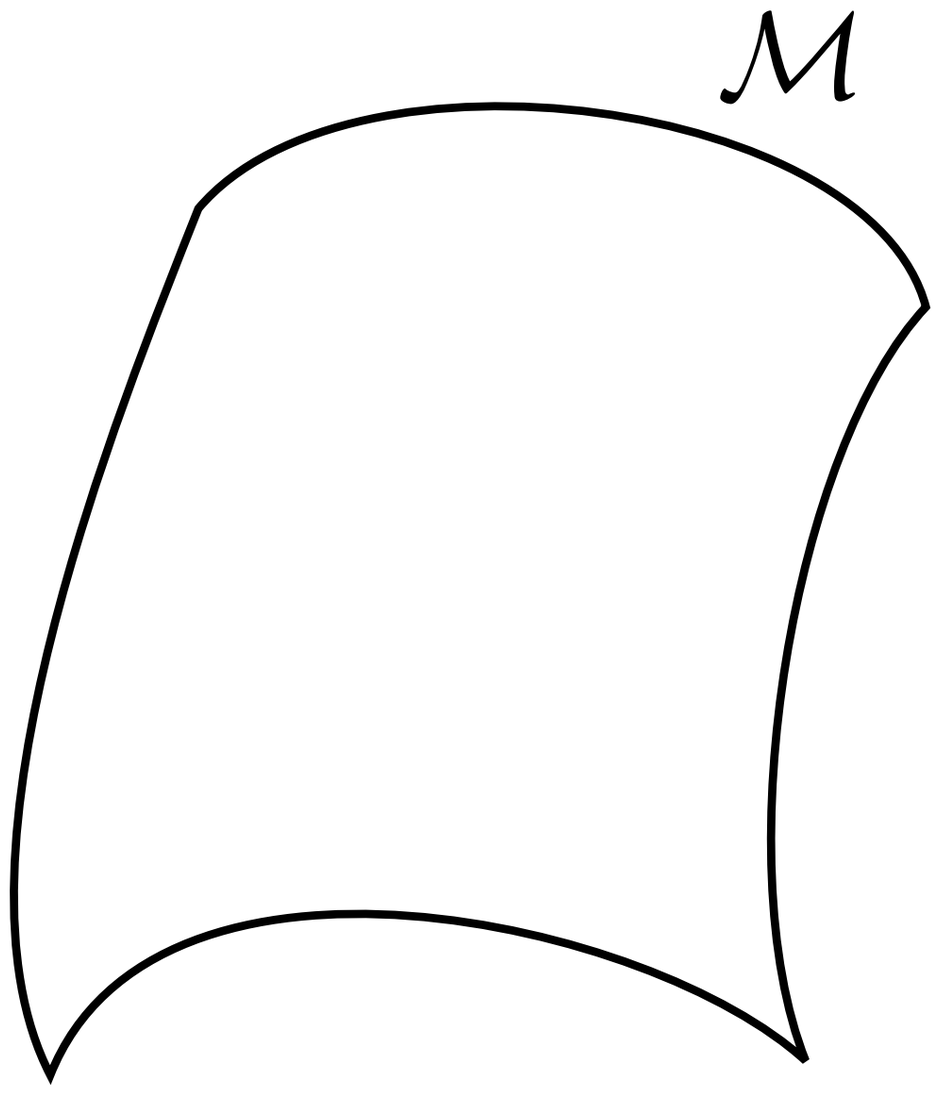
with b, Σ Lipschitz, and Σ uniformly elliptic on $\mathcal{M}$. Let q the density of the stationary measure of X_t. Let $\delta > 0$ be small enough, and $\tau > 0$. By collecting $p > (\tau^2 + d)\delta^{-4}$ paths of length $O(\delta)$ from each of $O(\delta^{-d})$ initial conditions sampled from μ_0, ATLAS returns a stochastic process $\hat{X}_t$ which has, with probability at least $1 - 2e^{-\tau^2}$, the following properties. $\hat{X}_t$ has a stationary measure, with density $\hat{q}$, such that

$$TV(dq, \Phi_{\#}^{-1}(d\hat{q})) \leq C\delta \log(1/\delta)\,.$$

Here Φ is the map from $\mathcal{M}$ to the collection of approximate tangent spaces. It is invertible for δ small enough.

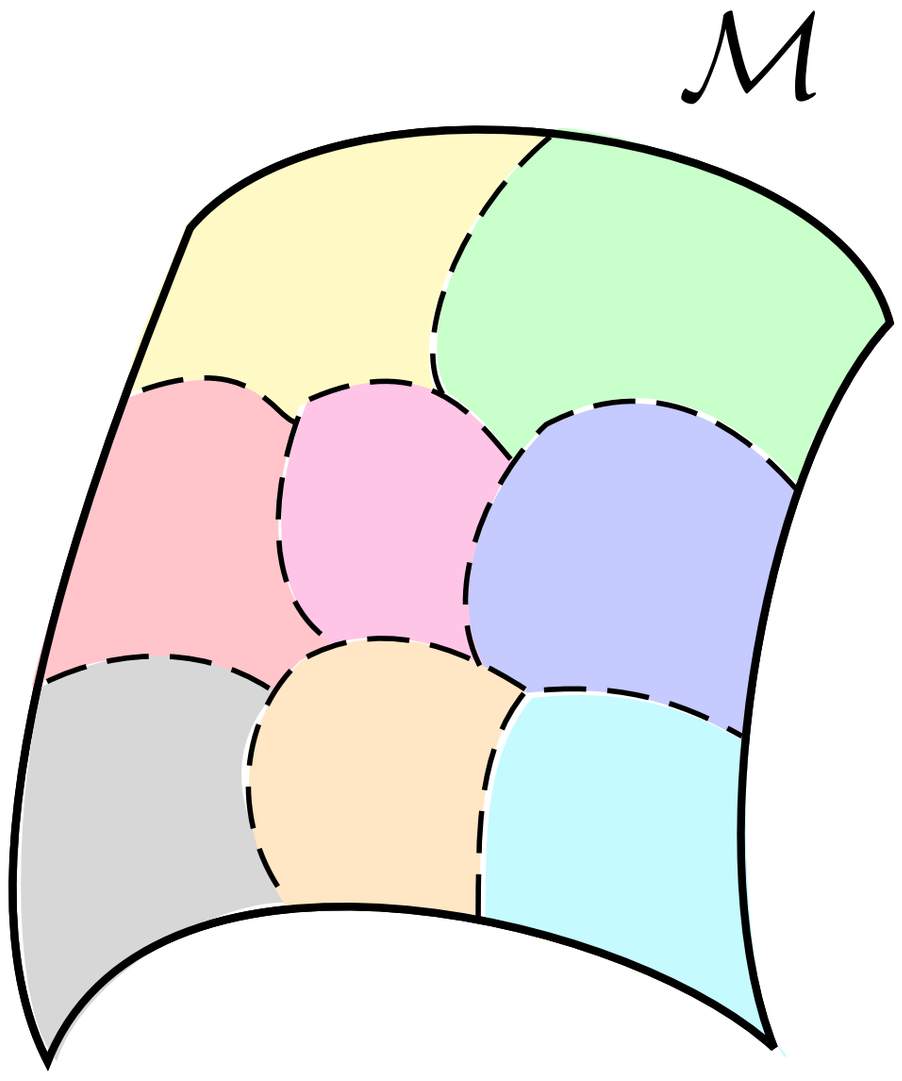
This result fits within the ideas that "short time accuracy implies long time accuracy" when averaging occurs and there is an underlying large-scale smoothness [J. Mattingly, A. Stuart, M. Tretyakov, E. Vanden-Eijnden, ...]

Sketch of construction



Sketch of construction

. Divide configuration space using a δ-net; $\mathcal{M}$ unknown, we use only samples.

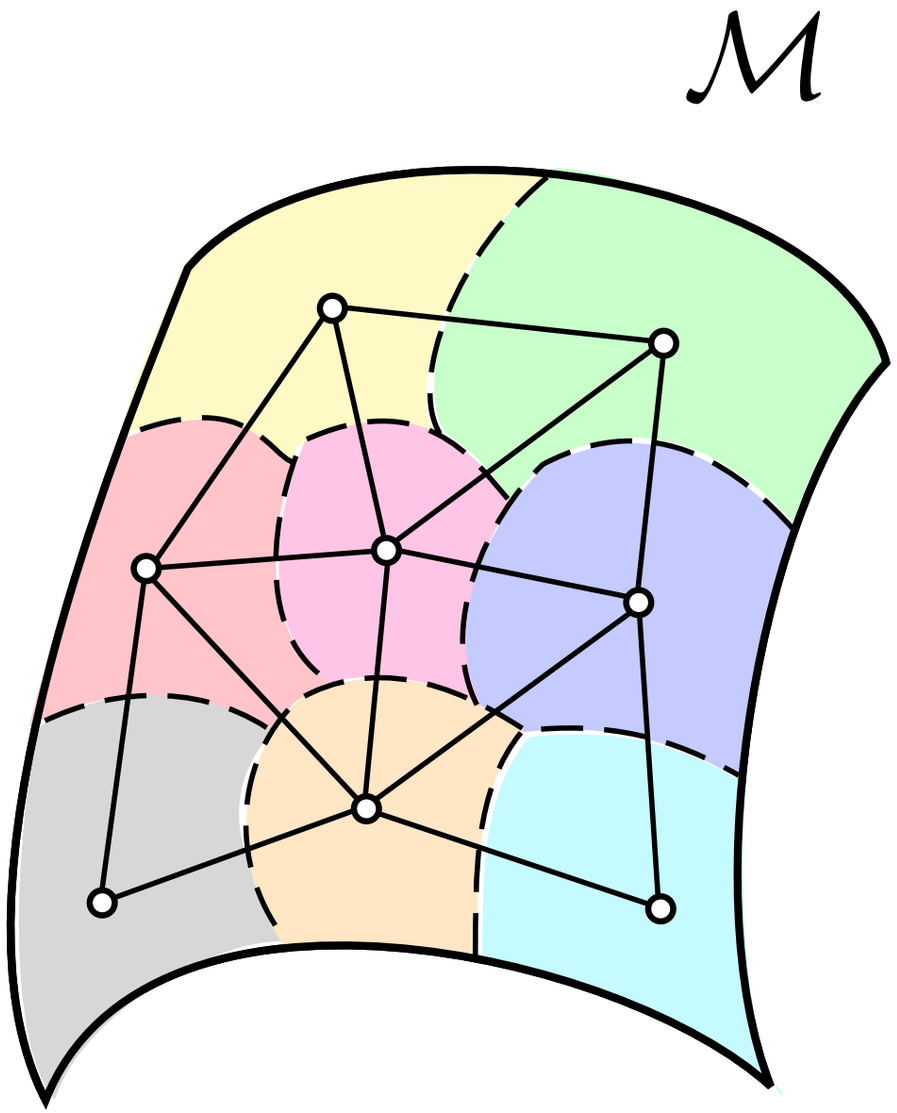


Γ a δ-net if $x \neq y \in \Gamma \implies d(x,y) > \frac{\delta}{2}$ and for every $x \in \mathcal{M}$ there is $y \in \Gamma$ with $d(x,y) < \delta$

Use cover trees to construct in online fashion in $O(C^d Dn \log n)$

Sketch of construction

. Divide configuration space using a δ-net; $\mathcal{M}$ unknown, we use only samples.



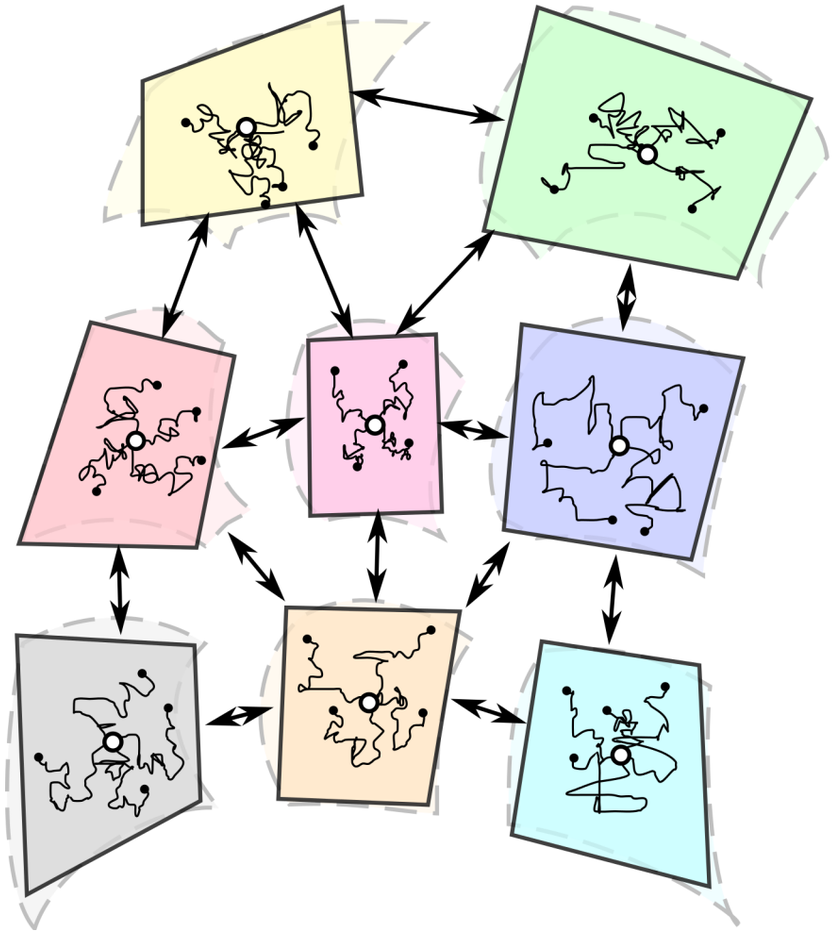
Connect each $y_k \in \Gamma$ to its neighbors.
This connectivity will be used to transition between local reduced simulators.
Also $O(C^d n \log n)$, and in parallel.

Sketch of construction

. Divide configuration space using a δ-net; $\mathcal{M}$ unknown, we use only samples.

. Construct local Euclidean charts in each piece of partition

Use Multi-Dimensional Scaling
to obtain maps Φ_k from
neighborhood of $y_k \in \Gamma$
to $C_k \subset \mathbb{R}^d$
Completely local calculation,
Can use landmarks to speed up.



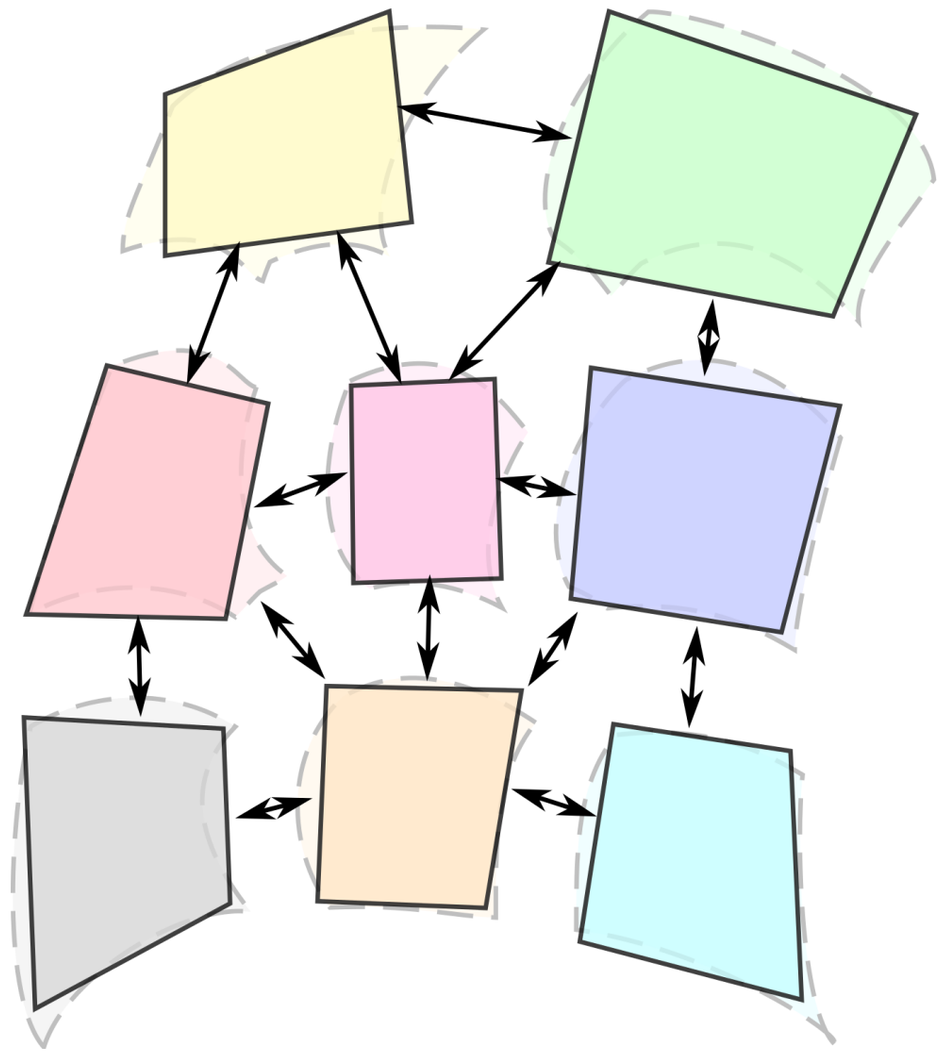
Sketch of construction

. Divide configuration space using a δ-net; $\mathcal{M}$ unknown, we use only samples.

. Construct local Euclidean charts in each piece of partition

. Construct connections between charts

A transition map between C_k and $C_{k'}$ is learned whenever $k \sim k'$.
We use linear maps $S_{k,k'} : \mathbb{R}^d \to \mathbb{R}^d$.



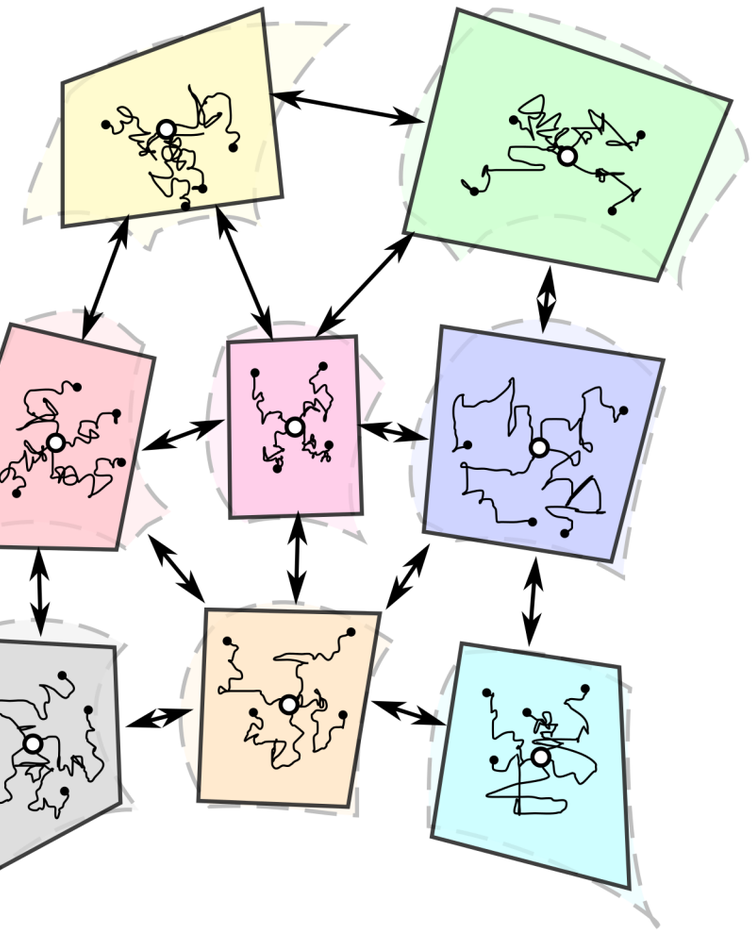
Sketch of construction

. Divide configuration space using a δ-net; $\mathcal{M}$ unknown, we use only samples.

. Construct local Euclidean charts in each piece of partition

. Construct connections between charts

. Learn simulators on charts



In each chart we fit a constant coefficient Itô diffusion:
$d\overline{X}_t = \bar{b}dt + \overline{\sigma}dB_t$
We estimate $\bar{b}$ and $\overline{\sigma}$ by running
p paths of length $O(\delta)$.
Turns out we need $p = O(d\delta^{-4})$ in order to obtain accuracy δ.

Sketch of construction

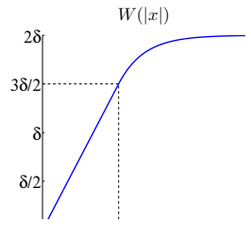
. Divide configuration space using a δ-net; $\mathcal{M}$ unknown, we use only samples.

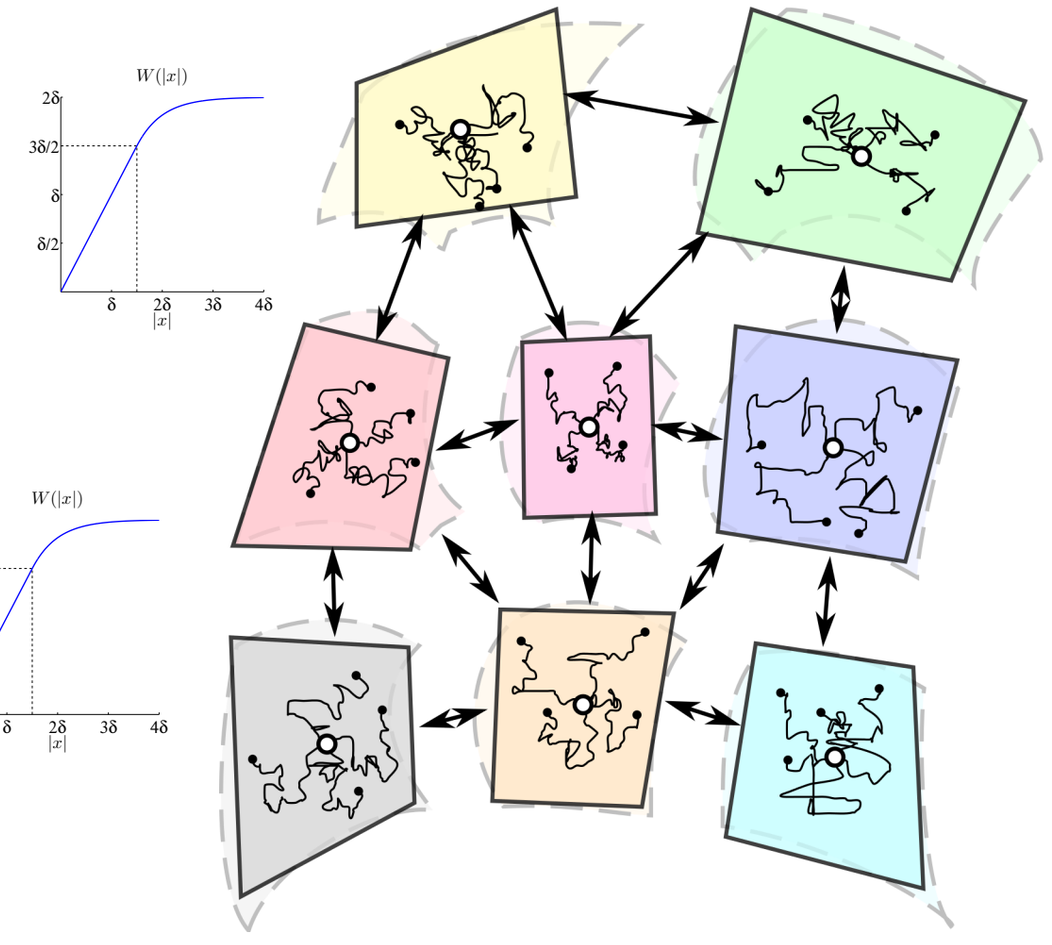
. Construct local Euclidean charts in each piece of partition

. Construct connections between charts

. Learn simulators on charts

. Glue simulators





ATLAS time step:

$$x \leftarrow x + \overline{b_k}\Delta t + \overline{\sigma}_k \Delta B$$
$$x \leftarrow W(x)$$
$$k' = \text{argmin}_{l \sim k} ||x - \Phi_k(y_l)||$$
$$x \rightarrow T_{k,k'}(x)$$

Large time guarantees

Theorem [M. Crosskey, MM]. Let $\mathcal{M}$ be a closed compact manifold, and suppose X_t is a stochastic process on $\mathcal{M}$ satisfying

$$dX_t = b(X_t)dt + \Sigma(X_t)dB_t$$

with b, Σ Lipschitz, and Σ uniformly elliptic on $\mathcal{M}$. Let q the density of the stationary measure of X_t. Let $\delta > 0$ be small enough, and $\tau > 0$. By collecting $p > (\tau^2 + d)\delta^{-4}$ paths of length $O(\delta)$ from each of $O(\delta^{-d})$ initial conditions sampled from μ_0, ATLAS returns a stochastic process $\hat{X}_t$ which has, with probability at least $1 - 2e^{-\tau^2}$, the following properties. $\hat{X}_t$ has a stationary measure, with density $\hat{q}$, such that

$$TV(dq, \Phi_{\#}^{-1}(d\hat{q})) \leq C\delta \log(1/\delta)\,.$$

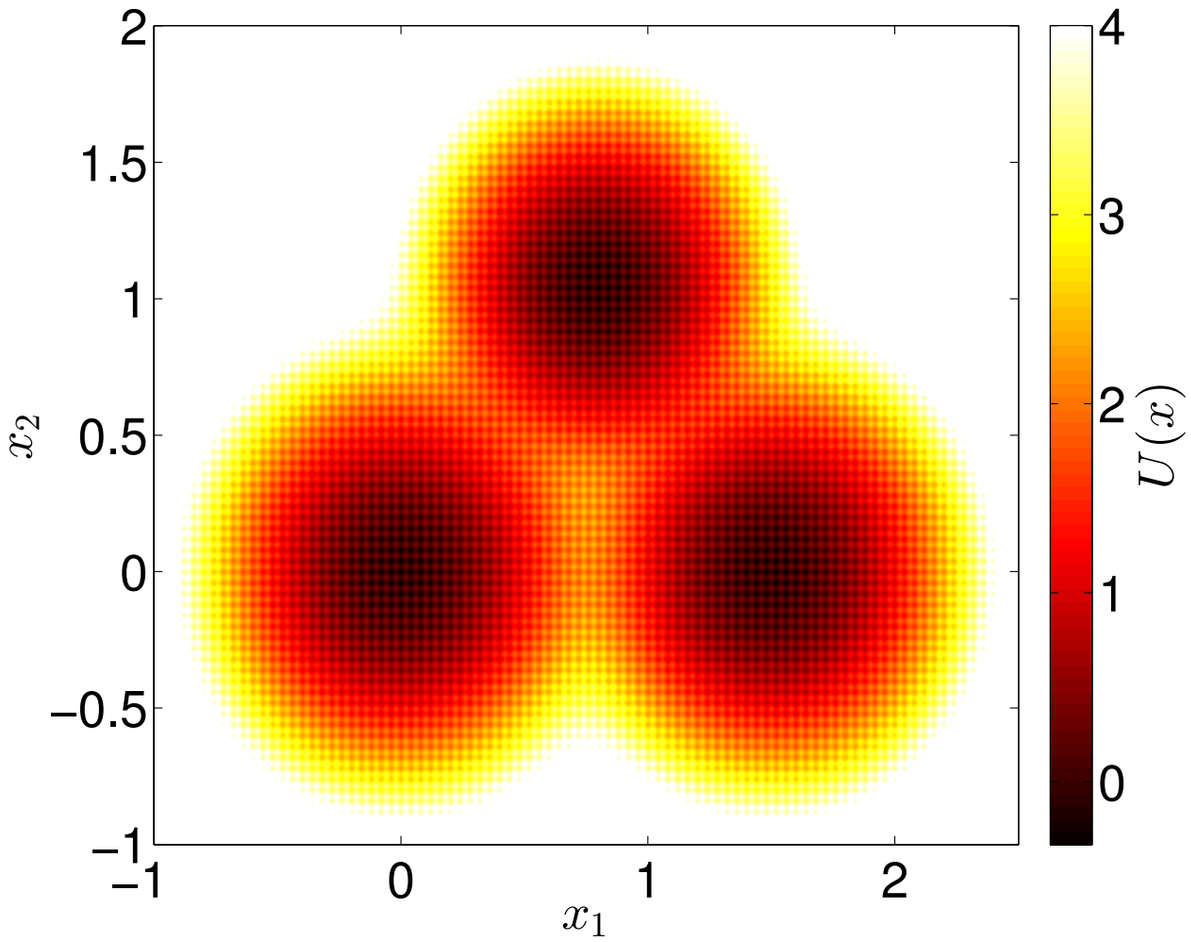
Here Φ is the map from $\mathcal{M}$ to the collection of approximate tangent spaces. It is invertible for δ small enough.

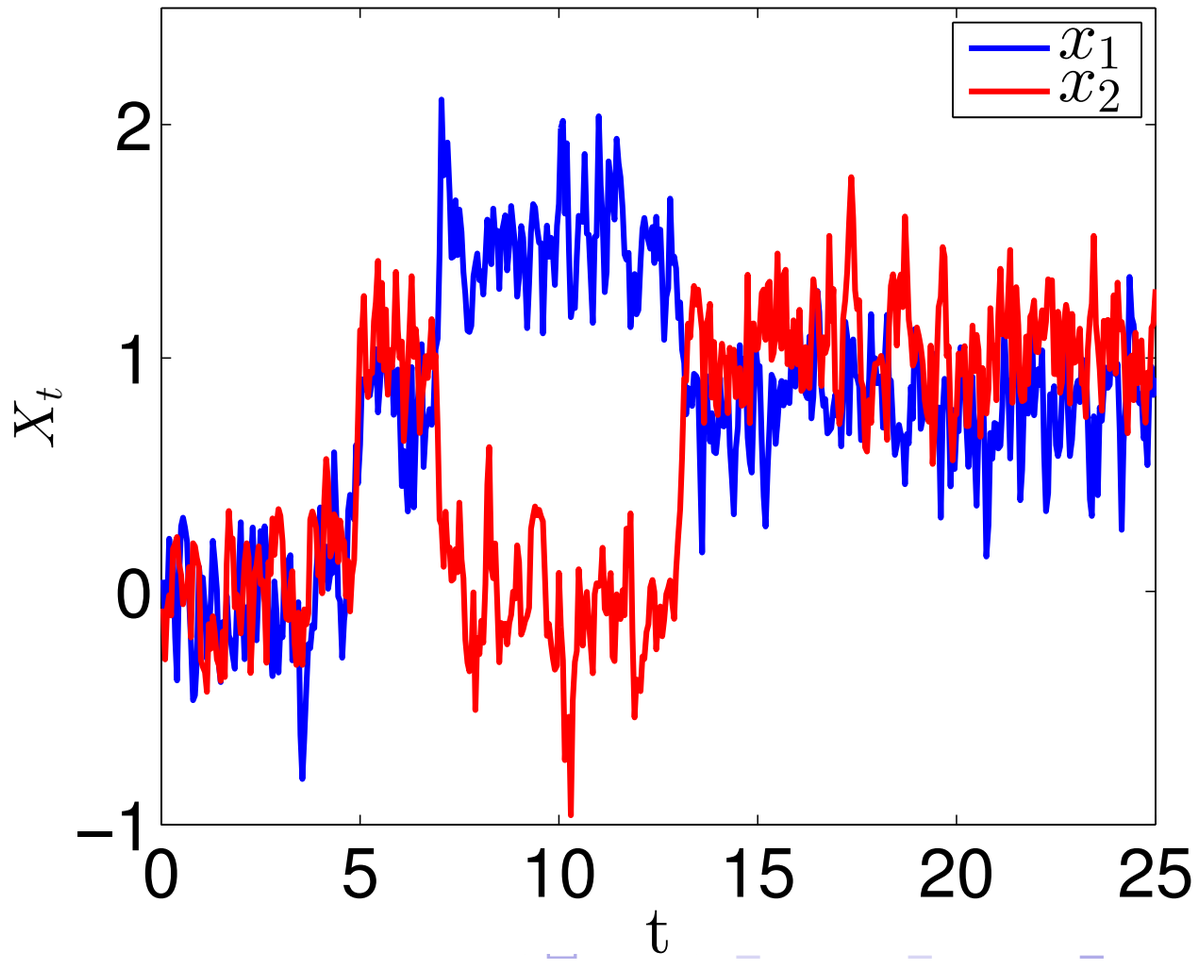
This result fits within the ideas that "short time accuracy implies long time accuracy" when averaging occurs and there is an underlying large-scale smoothness [J. Mattingly, A. Stuart, M. Tretyakov, E. Vanden-Eijnden, ...]

Examples: 2-D

Brownian motion in a potential well

$$dX_t = -\nabla U(X_t) + dB_t$$



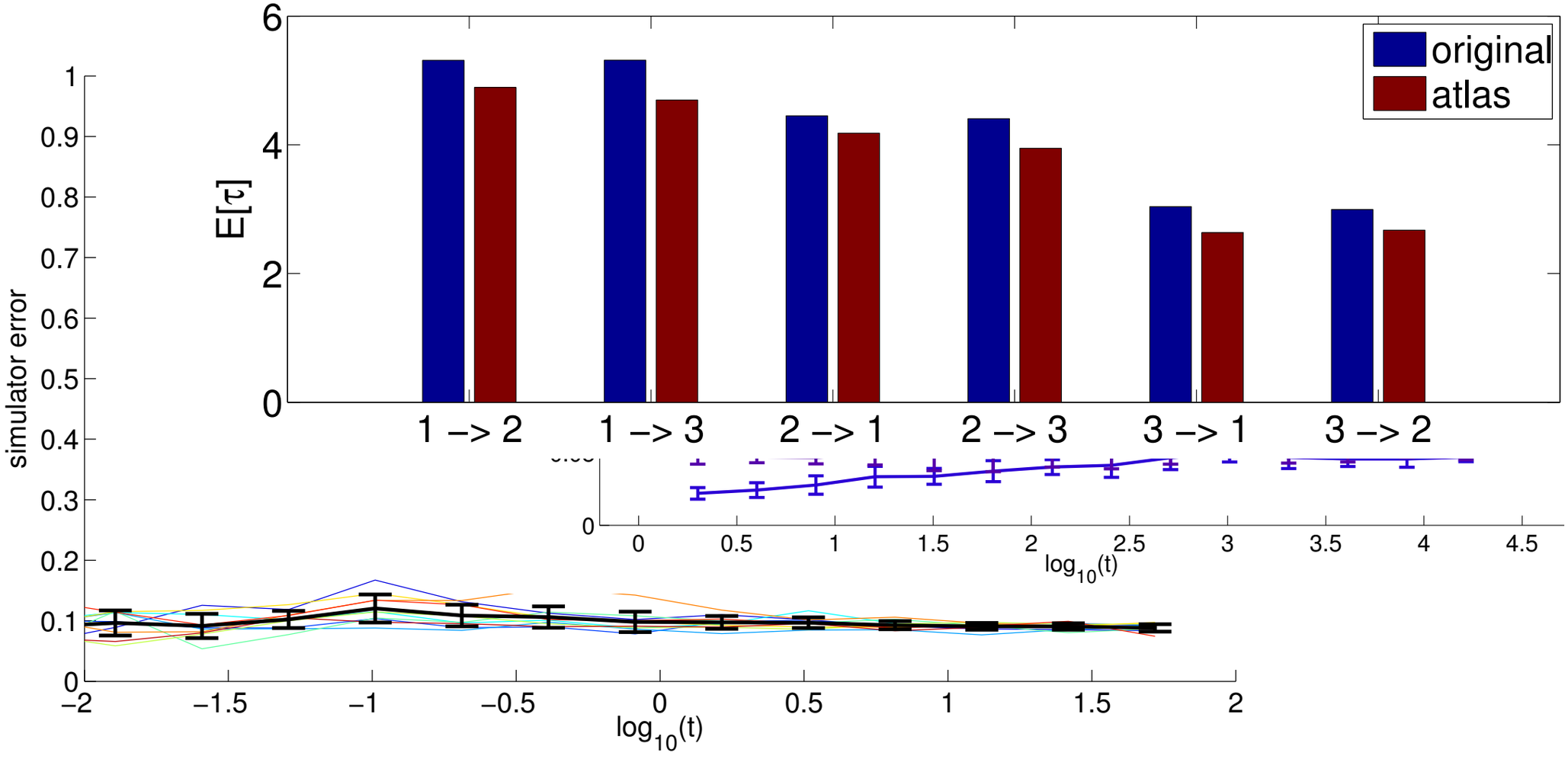


Examples: 2-D

Measure of error:

$$||p_t(x,\cdot) - \hat{p}_t(x,\cdot)||_{L^1(\mathcal{M})}\,,$$

where $p_t(x,\cdot)$ is the probability of being at $\cdot$ starting from x according to $\mathcal{S}$ and similarly for $\hat{p}$.
Look at the above, binned according to a partition associated with the net, averaged of x, as a function of t, aver all timescales.



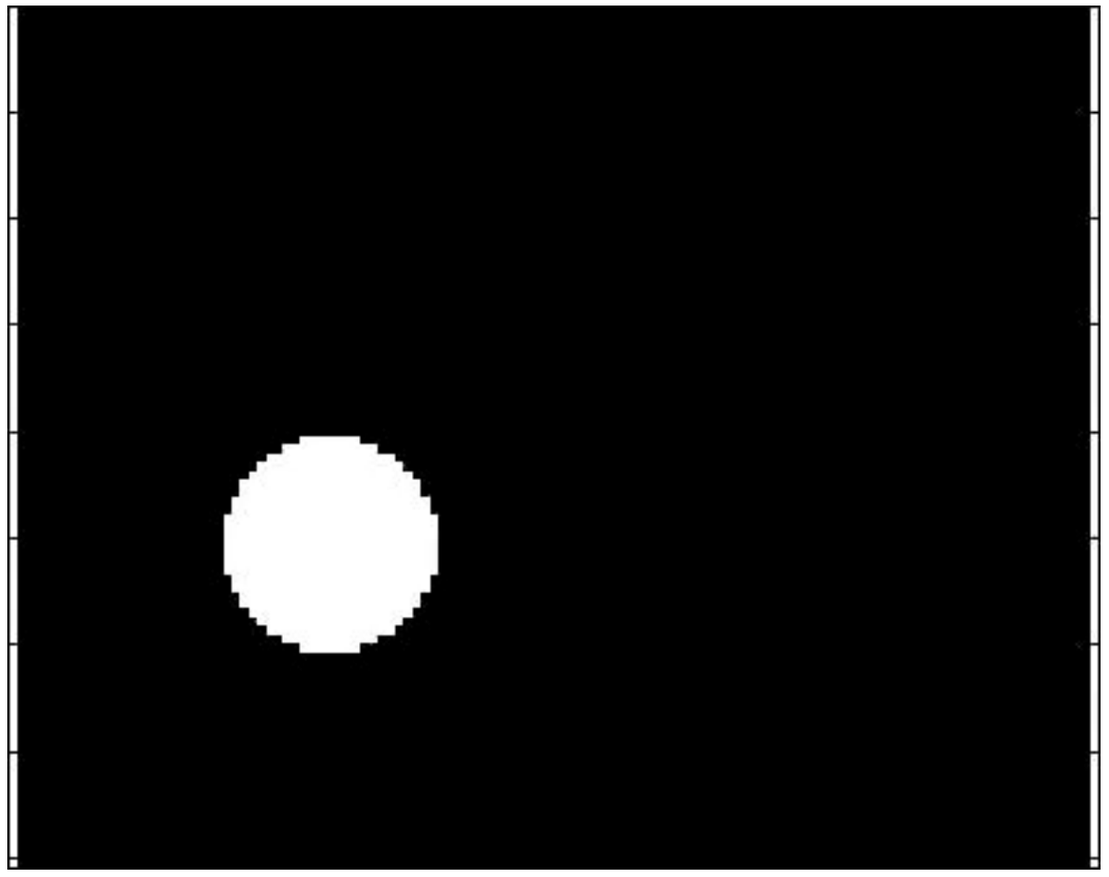
Examples: 12,500-D

Brownian motion in a potential well

$$dX_t = -\nabla U(X_t) + dB_t$$

obtained by mapping the 2-d rough potential to a (2-d) manifold in $\mathbb{R}^{12,500}$, endowed with L^1 distance, where each point is an image of a circle with center at the location corresponding to the 2-d example.

- $\delta = 0.2$
- $t_0 = 4 \cdot 10^{-2} = 800$ steps
- 230 charts
- $p = 2 \cdot 10^3$ samples per chart

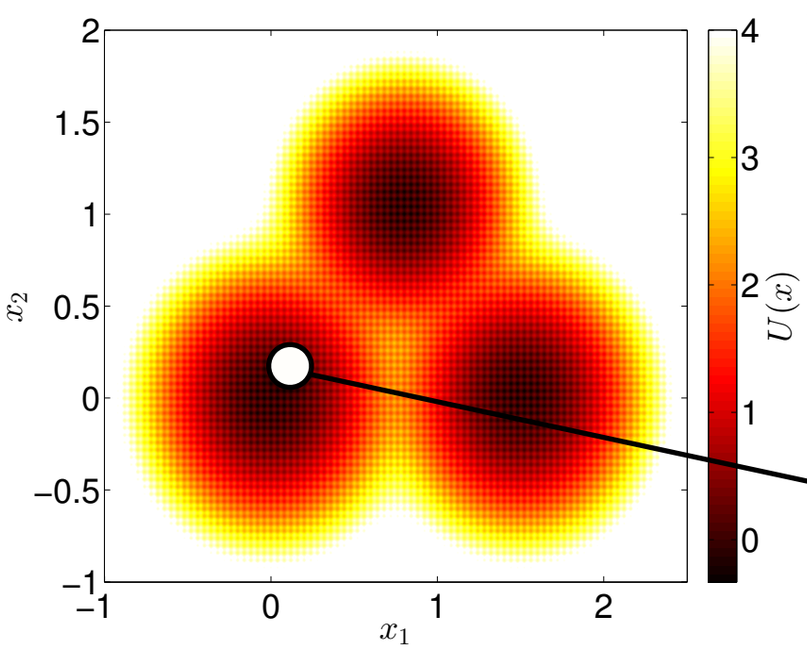


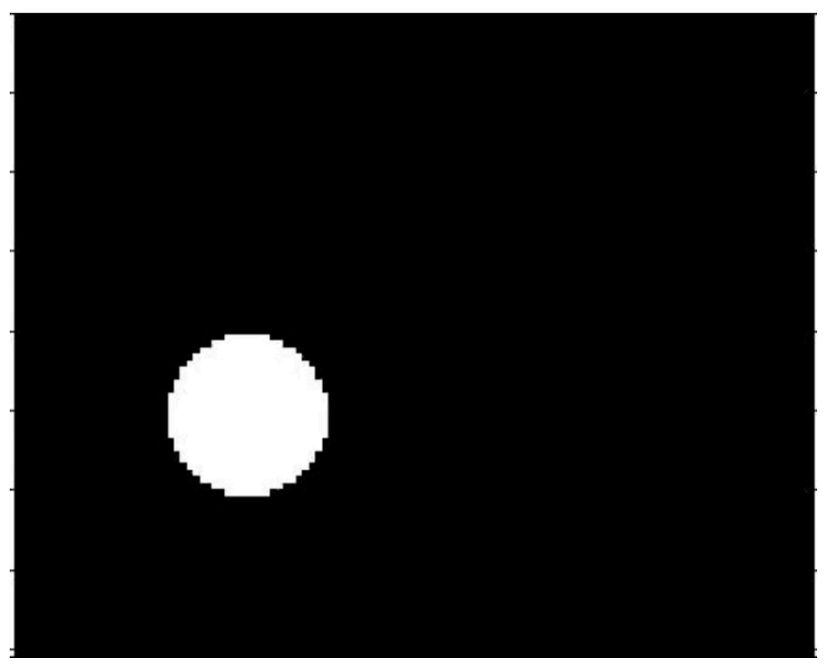
Examples: 12,500-D

Brownian motion in a potential well

$$dX_t = -\nabla U(X_t) + dB_t$$

obtained by mapping the 2-d rough potential to a (2-d) manifold in $\mathbb{R}^{12,500}$, endowed with L^1 distance, where each point is an image of a circle with center at the location corresponding to the 2-d example.



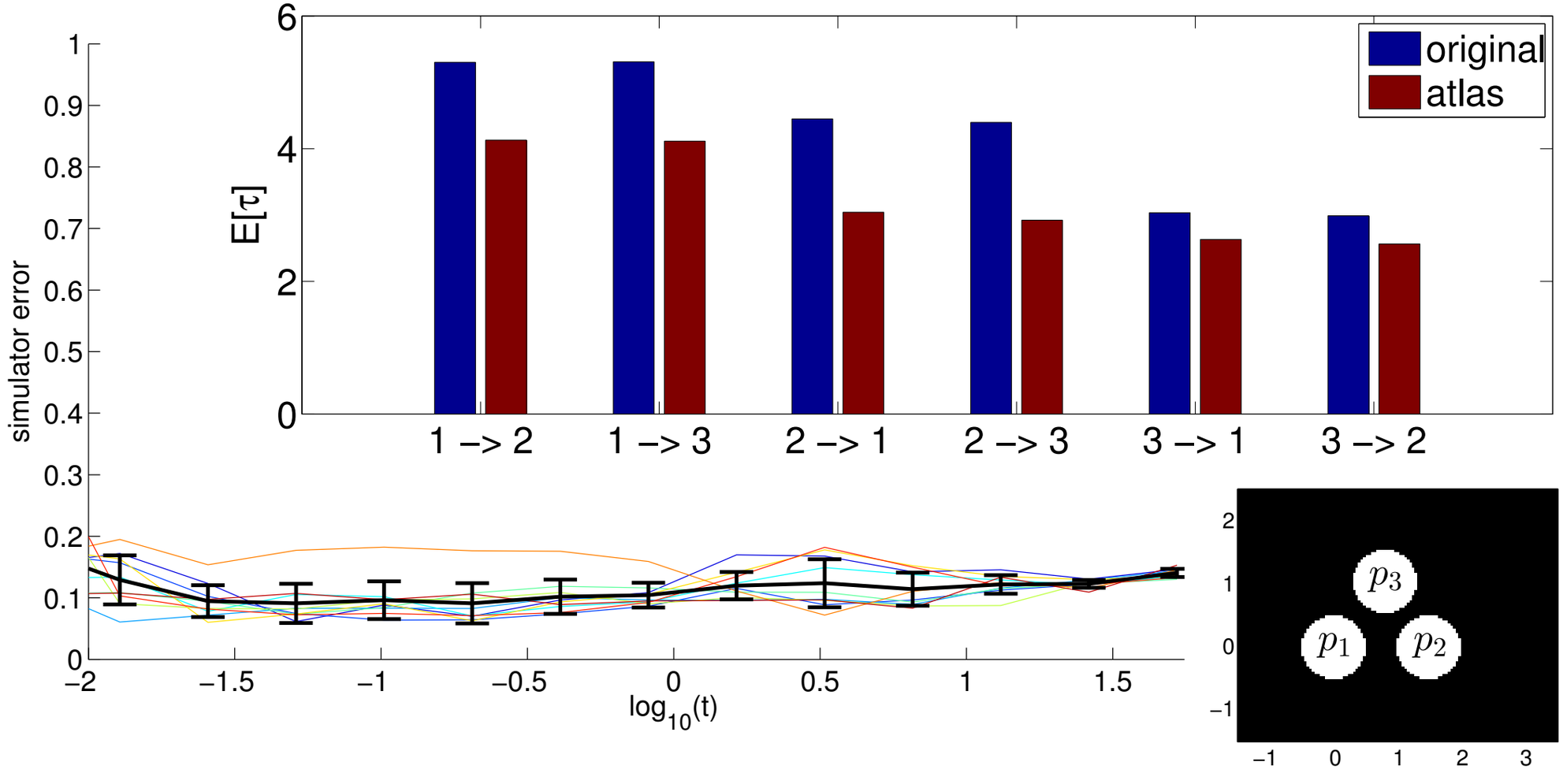


Examples: 12,500-D

Measure of error:

$$||p_t(x,\cdot) - \hat{p}_t(x,\cdot)||_{L^1(\mathcal{M})}\,,$$

where $p_t(x,\cdot)$ is the probability of being at $\cdot$ starting from x according to $\mathcal{S}$ and similarly for $\hat{p}$.
Look at the above, binned according to a partition associated with the net, averaged of x, as a function of t, aver all timescales.



Examples: 82-D, chaotic

Noise may arise from ensembles of deterministic chaotic processes. Multiscale ODE with a scale ϵ:

$$\begin{cases} \dot{X}_t^\epsilon = \epsilon f(X_t^\epsilon) + g(Y_t), & X_0^\epsilon = x \\ \dot{Y}_t = h(Y_t) & Y_0 = y \end{cases}$$

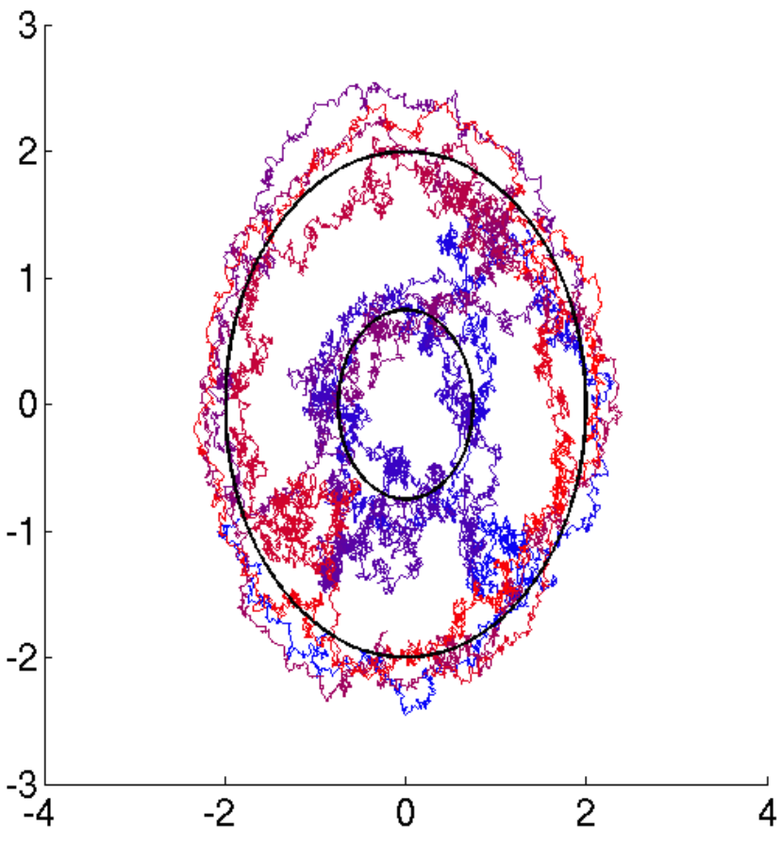
If the dynamics for Y_t alone admits an invariant measure μ, and $\mathbb{E}_\mu[g] = O(\epsilon)$, then the above behaves like the SDE $dX_s = b(X_s)ds + \sigma(X_s)dB_s$, on the timescale $s = \epsilon t$ in the limit $\epsilon \to 0$. For fixed ϵ, difficult to simulate directly due to the timescale separation.

We choose Y_t=Lorentz '96 with 80 dimensions, and $X_t \in \mathbb{R}^2$ so that there are two limit cycles consisting of two concentric circles.

. $\epsilon = 0.1$

. $\delta = 0.18 = 240$ steps

. $t_0 = \frac{1}{4}\epsilon^{-1}$



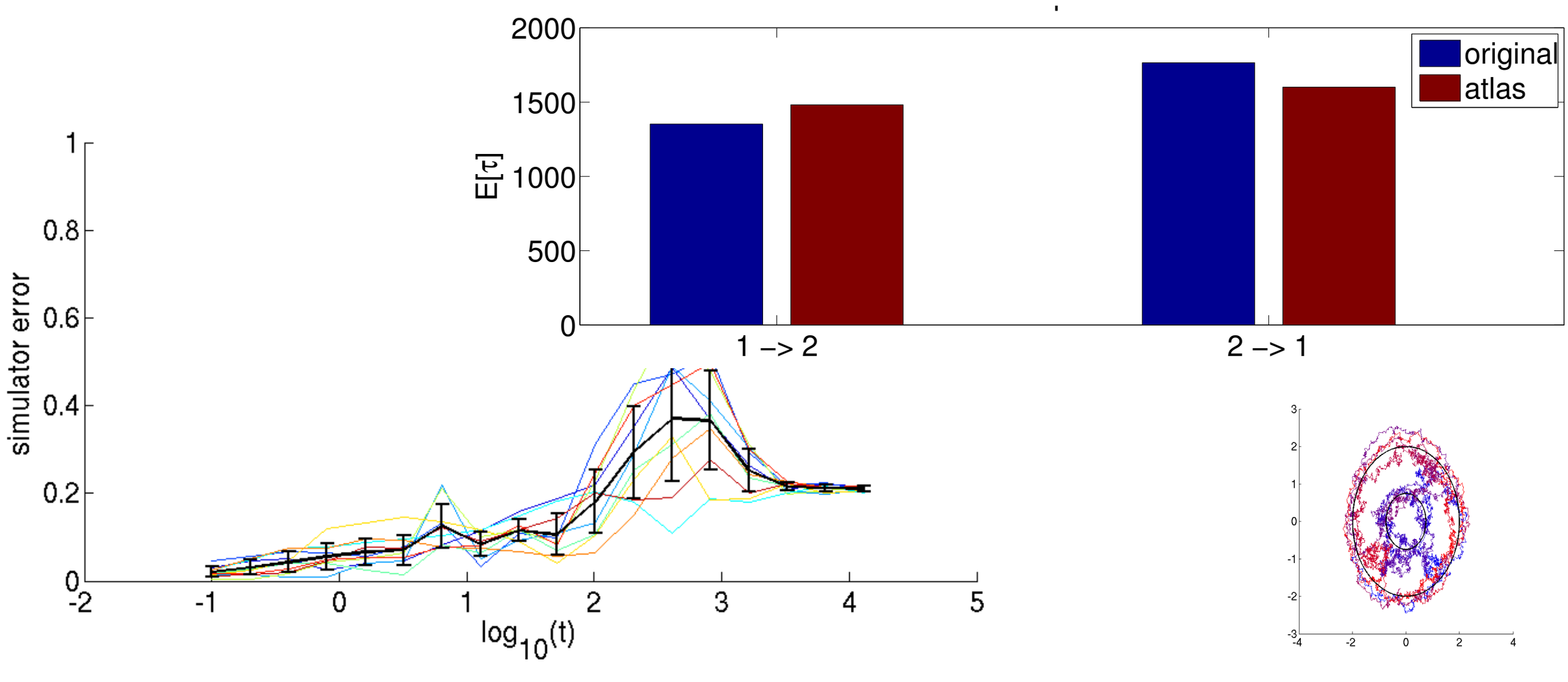
Examples: 82-D, chaotic

Measure of error:

$$||p_t(x,\cdot) - \hat{p}_t(x,\cdot)||_{L^1(\mathcal{M})}\,,$$

where $p_t(x,\cdot)$ is the probability of being at $\cdot$ starting from x according to $\mathcal{S}$ and similarly for $\hat{p}$.

Look at the above, binned according to a partition associated with the net, averaged of x, as a function of t, aver all timescales.



Conclusions

ATLAS

- decouples the ability of sampling in any way from interesting regions from the ability of getting a reduced model
- learns locally and in parallel model reduction by sampling short paths
- learns maps to stitch together the local models
- reduced simulator has correct large-time statistics (guaranteed for stationary distribution)

Extensions:

- *Multiscale*: choose scale and dimension adaptively.
- Fully online mode with *exploration*
- *Molecular Dynamics*
- Generalize theory to *other large-time functionals* [transition rates and beyond], and other notions of closeness in p
- *Hypoellipticity* (e.g. second order Langevin); *no homogeneous systems, MCMC*

This is a NUMERICAL SCHEME. It exploits structure in the equations that is not exploited by any other numerical scheme.

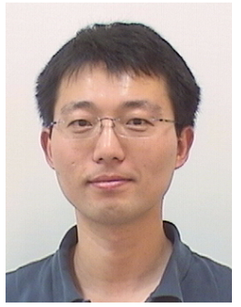
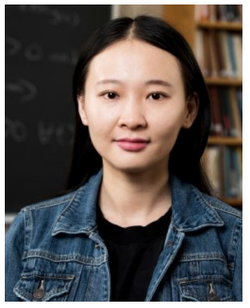
Two stories

Learning interaction kernels of agent-based systems. Given trajectories of a system of interacting agents, that may exhibit emergent behavior (e.g. flocking), can we learn interaction kernels, in a flexible non-parametric fashion, without being cursed by the high dimension of the state space?

Model reduction for stochastic processes (diffusions and Lengevin dynamics) on manifolds. Given the ability of sampling inital conditions and short paths, learn a stochastic process on a manifold (with both the process and the manifold being unknown) that approximates the original not only at short time scales, but also at long time scales. Combination of manifold learning and learning of SDE's.

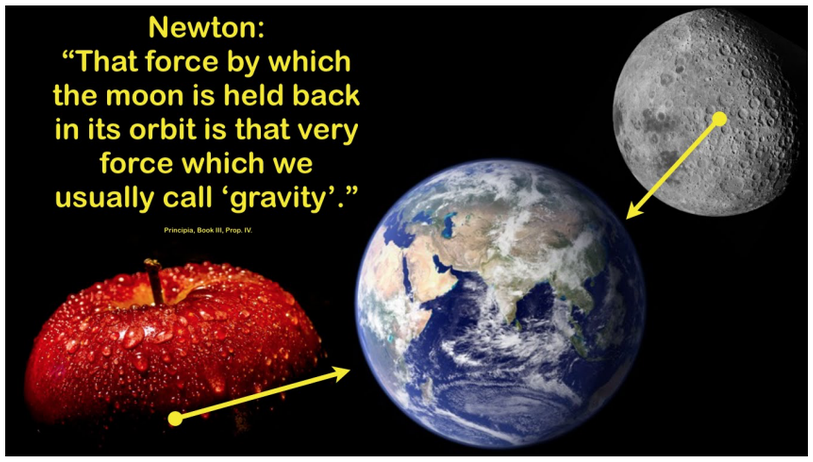
From https://www.youtube.com/watch?v=bb9ZTbYGRdc

Learning of Interaction Rules for agent-based systems


Given trajectories of dynamical system of interacting agents, learn the interaction rules. Applications: biological systems, particle systems.
Further goals: hypothesis testing for agent-based systems; transfer learning; agents on networks; collaborative and competitive games; learning dictionary for complex dynamical systems.

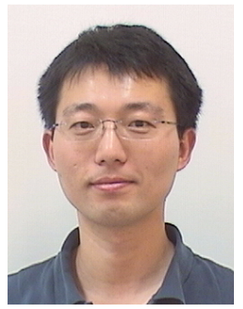
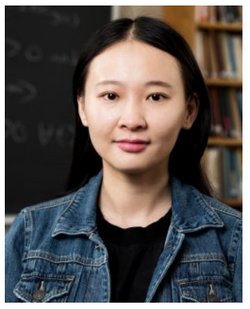
Lots of recent interest
in ML for learning
ODE's and PDE's
e.g. M. Fornasier, N. Kutz,
Y. Kevrekidis, R. Ward...



From https://www.youtube.com/watch?v=gJhn7WmXWVY

From https://www.youtube.com/watch?v=bb9ZTbYGRdc

Learning of Interaction Rules for agent-based systems

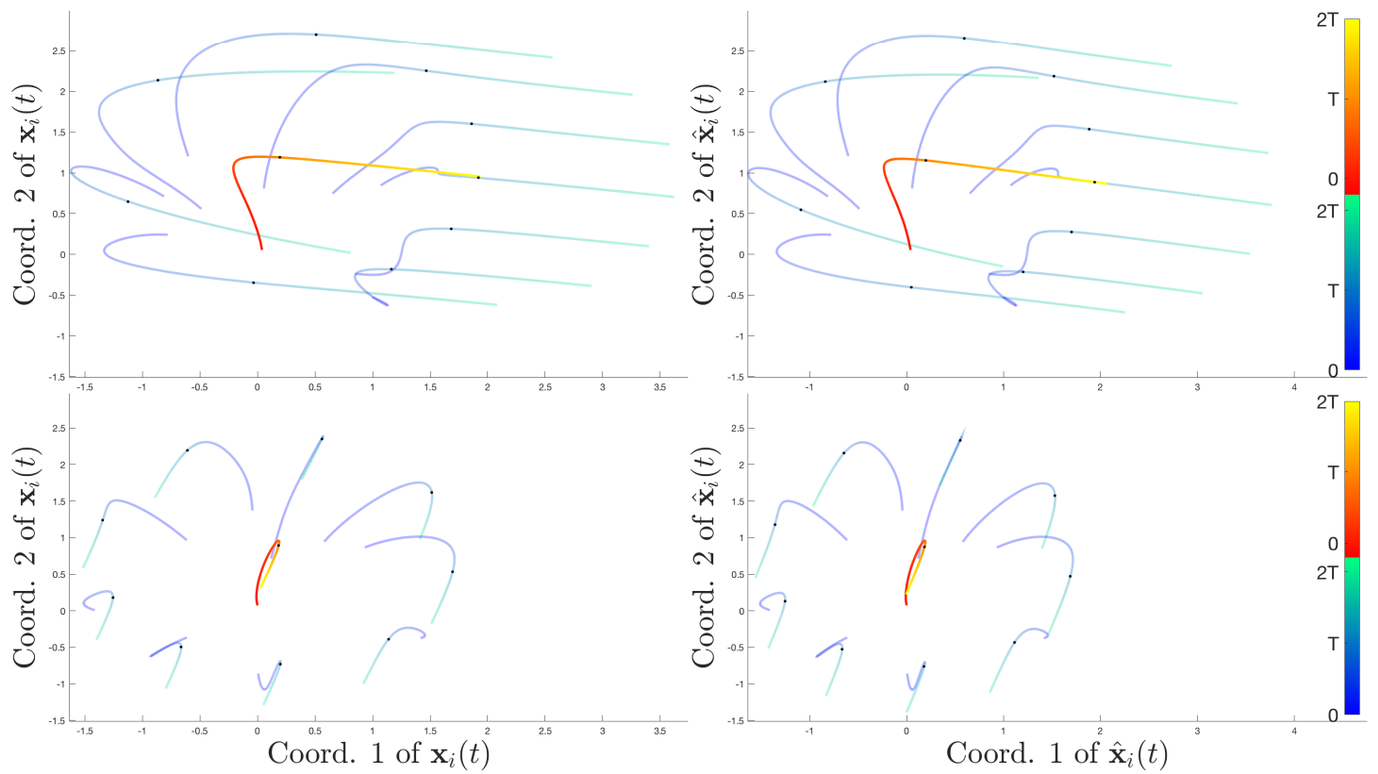

Given observations of the positions of agents $\{\mathbf{x}_i\}_{i=1}^N$ at different times $\{t_l\}_{l=1}^L$ and/or for different initial conditions $\{\mathbf{x}^{(m)}(0)\}_{m=1}^M$, evolving for example according to

$$\dot{\mathbf{x}}_i = \sum_{i'} \phi(||\mathbf{x}_i - \mathbf{x}_{i'}||)(\mathbf{x}_{i'} - \mathbf{x}_i)$$

we want to learn ϕ. Different limits: $N \to +\infty$ (mean-field limit, joint work with M. Fornasier and M. Bongini), $M \to +\infty$ (joint current work with F. Lu, M. Zhong and S. Tang).

Interesting extensions to:
- higher-order systems,
- stochastic systems,
- agents of different types,
- varying environment.

...



Second-order prey-predator model.
Left: true trajectories; Right: trajectories with learned interactions.

The Mean-field limit

Rewriting

$$\dot{\mathbf{x}}_i = \sum_{i'} \phi(||\mathbf{x}_i - \mathbf{x}_{i'}||)(\mathbf{x}_{i'} - \mathbf{x}_i) = \frac{1}{N} \sum_{i'} \frac{\Phi'(||\mathbf{x}_i - \mathbf{x}_{i'}||)}{||\mathbf{x}_i - \mathbf{x}_{i'}||}(\mathbf{x}_i - \mathbf{x}_{i'})$$

we see this is the gradient flow of the energy $\mathcal{J}_N(\mathbf{X}) = \frac{1}{2N} \sum_{i,i'=1}^{N} \Phi(||\mathbf{x}_i - \mathbf{x}_{i'}||)$.

Considering the measure $\mu^N(t) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\mathbf{x}_i(t)}$, we may let $N \to +\infty$ to obtain (under suitable regularity assumptions on Φ) the mean field equations

$$\partial_t \mu(t) = -\nabla \cdot \left(\left(-\frac{\Phi'(||\cdot||)}{||\cdot||} * \mu(t) \right) \mu(t) \right) , \quad \mu(0) = \mu_0 \, .$$

This is also a gradient flow for the energy $\mathcal{J}(\mu) = \int_{\mathbb{R}^{d \times d}} \Phi(||\mathbf{x} - \mathbf{y}||) d\mu(\mathbf{x}) d\mu(\mathbf{y})$ on the space of probability measures with Wasserstein distance.
This was studied in *Inferring Interaction Rules from Observations of Evolutive Systems I: The Variational Approach*, M. Bongini, M. Fornasier, M. Hansen, and MM, M3S, 2017

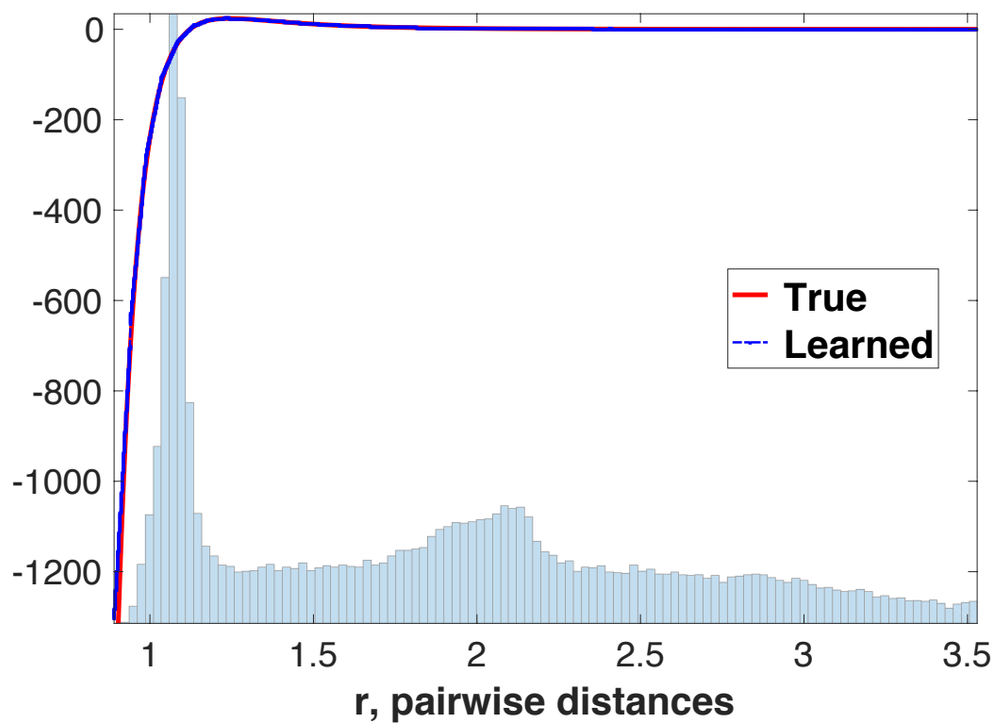
Learning the Interaction Kernel

Observations: $\{(\mathbf{x}_i, \dot{\mathbf{x}}_i)^{(m)}(t_l)\}_{i=1,l=1,m=1}^{N,L,M}$, where $\mathbf{x}^{(m)}(0) \sim \mu_0$ for some μ_0 on $\mathbb{R}^d$. Note that each state of the system is in $\mathbb{R}^{dN}$.
All we want however is the one-dimensional interaction kernel ϕ in the equations

$$\dot{\mathbf{x}}_{i'}(t) = \sum_{i'} \phi(\underbrace{||\mathbf{x}_{i'}(t) - \mathbf{x}_i(t)||}_{r_{ii'}(t)})(\mathbf{x}_{i'}(t) - \mathbf{x}_i(t))\,.$$

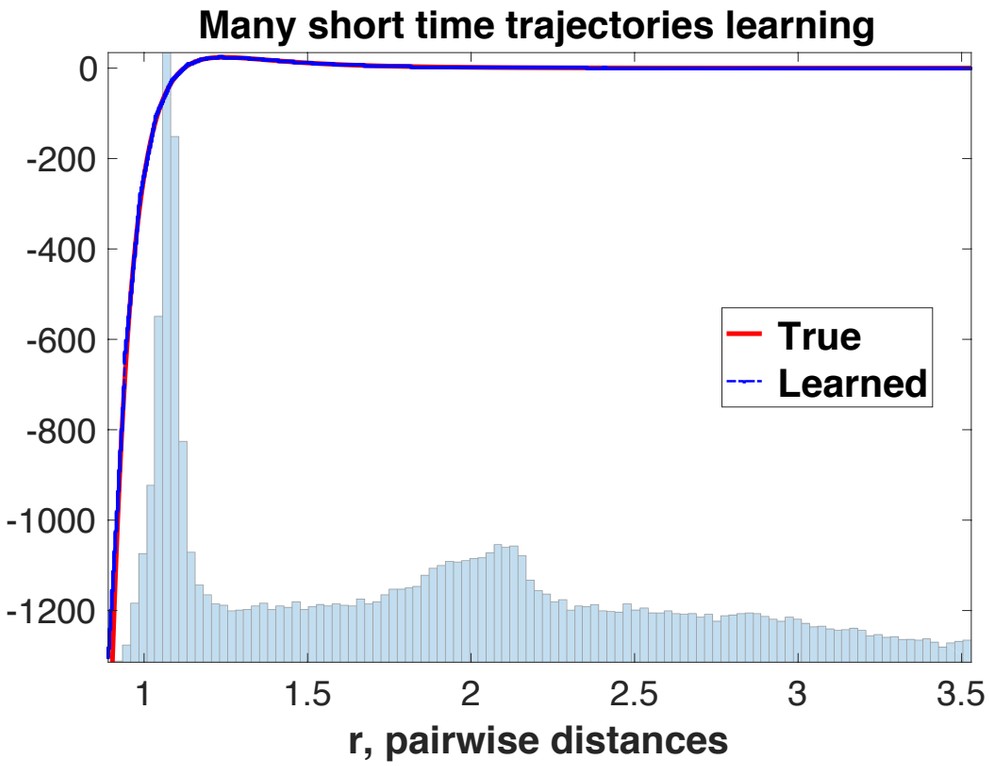
Fix the time scale $[0, T]$. We introduce the measure on $\mathbb{R}_+$ defined by
$\rho_T^L(r) := \frac{1}{\binom{N}{2}L} \sum_{l=1}^{L} \mathbb{E}_{X(0)\sim\mu_0} \left[\sum_{i,i'=1,i<i'}^{N} \delta_{r_{ii'}(t_l)}(r)\right].$

Example. The Lennard Jones force is the derivative of the potential
$V_{LJ}(r) = 4\epsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right)$.
Right figure: In blue the LJ ϕ,
superimposed to an empirical estimate of ρ_T^L,
for a system of $N = 7$ agents, and L, T small.

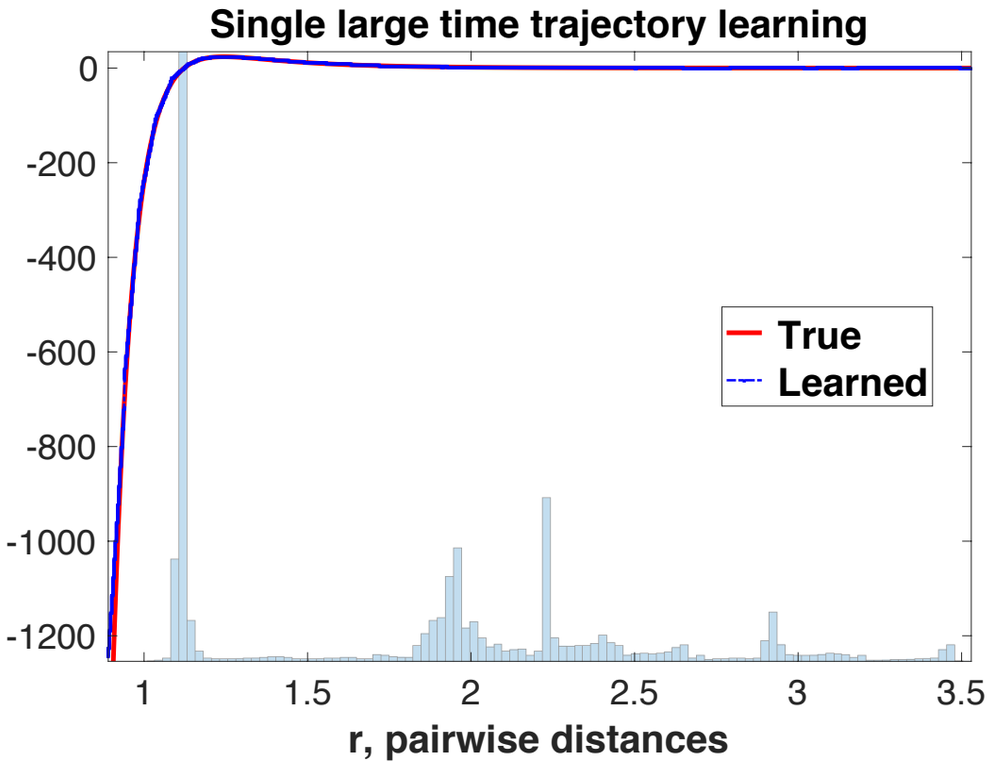


Example: L-J kernel and ρ_L^T

Example. The Lennard Jones force is the derivative of the potential
$V_{LJ}(r) = 4\epsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right)$.
Right figure: In blue the LJ ϕ,
superimposed to an empirical estimate of ρ_T^L,
for a system of $N = 7$ agents, and L, T small.



Example (cont'd). The measure ρ_T^L does depend on L and T.
With the same system as above,
we consider here L and T large.
ρ_T^L is much more concentrated,
due to the system approaching equilibrium..



The estimator

Observations: $\{(\mathbf{x}_i, \dot{\mathbf{x}}_i)(t_l)\}_{I=1,l=1}^{N,L}$, for M different IC's, from

$$\dot{\mathbf{x}}_{i'}(t) = \sum_{i'} \phi(||\mathbf{x}_{i'}(t) - \mathbf{x}_i(t)||)(\mathbf{x}_{i'}(t) - \mathbf{x}_i(t)) =: \mathbf{f}_\phi(\mathbf{x}_i(t))\,.$$

When the t_l's are equi-spaced in time, we consider the estimator $\widehat{\varphi}_{L,M,\mathcal{H}}$ minimizing over some set of functions $\mathcal{H}$ the empirical error functional

$$\mathcal{E}_{L,M}(\varphi) := \frac{1}{LMN} \sum_{l,m,i=1}^{L,M,N} \left\| \dot{\mathbf{x}}_i^{(m)}(t_l) - \mathbf{f}_\varphi(\mathbf{x}_i^{(m)}(t_l)) \right\|^2,$$

over $\varphi \in \mathcal{H}$, a simple hypothesis space of functions on $\mathbb{R}_+$, with dimension n which will be chosen dependent on M:

$$\hat{\phi}_{L,M,\mathcal{H}} := \arg\min_{\varphi \in \mathcal{H}} \mathcal{E}_{L,M}(\varphi)\,.$$

For $\mathcal{H}$ linear subspace, this is a least squares problem (Gauss, Legendre). Coercivity constant is related to smallest singular value of matrix in the LS problem.

Coercivity condition

$$\mathcal{E}_{L,M}(\varphi) := \frac{1}{LMN} \sum_{l,m,i=1}^{L,M,N} \left\| \dot{\mathbf{x}}_i^{(m)}(t_l) - \mathbf{f}_\varphi(\mathbf{x}_i^{(m)}(t_l)) \right\|^2,$$

$$\hat{\phi}_{L,M,\mathcal{H}} := \arg\min_{\varphi \in \mathcal{H}} \mathcal{E}_{L,M}(\varphi)\,.$$

We shall assume that the unknown interaction kernel ϕ is in the admissible class $\mathcal{K}_{R,S} := \{\varphi \in C^1(\mathbb{R}_+) :\ \mathrm{supp}\varphi \subset [0,R], \sup_{r\in[0,R]} |\varphi(r)| + |\varphi'(r)| \le S\}$.
Coercivity condition: $\forall \varphi\ :\ \varphi(\cdot)\cdot \in L^2(\rho_T^L)$, for $0 < c_L \le (N-1)/N^2$

$$c_L \|\varphi(\cdot)\cdot\|^2_{L^2(\rho_T^L)} \le \frac{1}{NL} \sum_{l,i=1}^{L,N} \mathbb{E} \left\| \frac{1}{N} \sum_{i'=1}^{N} \varphi(r_{ii'}(t_l)) \mathbf{r}_{ii'}(t_l) \right\|^2.$$

Lemma. Coercivity $\implies$ unique minimizer of $\lim_{M\to+\infty} \mathcal{E}_{L,M}(\varphi)$ over $\varphi \in \mathcal{H}$

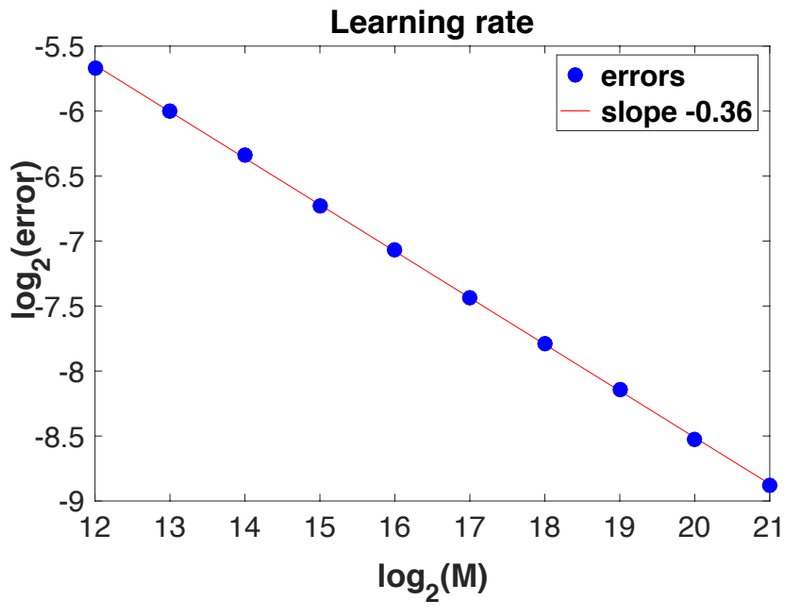
Main Theorem

Theorem. Let $\{\mathcal{H}_n\}_n$ be a sequence of subspaces of $L^\infty[0,R]$, with $\dim(\mathcal{H}_n) \leq c_0 n$ and $\inf_{\varphi\in\mathcal{H}_n} \|\varphi(\cdot)-\phi(\cdot)\|_{L^\infty([0,R])} \leq c_1 n^{-s}$, for some constants $c_0, c_1, s > 0$. It exists, for example, if ϕ is s-Hölder regular. Choose $n_* = (M/\log M)^{\frac{1}{2s+1}}$: then for some $C = C(c_0, c_1, c_L, R, S)$

$$\mathbb{E}[\|\widehat{\phi}_{L,M,\mathcal{H}_{n_*}}(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho_L^T)}] \leq C\left(\frac{\log M}{M}\right)^{\frac{s}{2s+1}}.$$

- The good: Rate in M is optimal, in fact even optimal in the case of regression, where we would be given $(r_m, \phi(r_m))_{m=1}^M$.
- The bad: no dependency on L.

Example. The Lennard Jones kernel is *not* admissible, yet since particles rarely get very close to each other, we obtain a convergence rate close to optimal.



Errors on trajectories

Proposition. Assume $\widehat{\phi}(\|\cdot\|)\cdot \in \mathrm{Lip}(\mathbb{R}^d)$, with Lipschitz constant C_{Lip}. Let $\widehat{\mathbf{X}}(t)$ and $\mathbf{X}(t)$ be the solutions of systems with kernels $\widehat{\phi}$ and ϕ respectively, started from the same initial condition. Then for each trajectory

$$\sup_{t\in[0,T]} \|\widehat{\mathbf{X}}(t) - \mathbf{X}(t)\|^2 \leq 2Te^{8T^2C_{\mathrm{Lip}}^2}\int_0^T \left\|\dot{\mathbf{X}}(t) - \mathbf{f}_{\hat{\phi}}(\mathbf{X}(t))\right\|^2 dt\,,$$

and on average w.r.t. the distribution μ_0 of initial conditions:

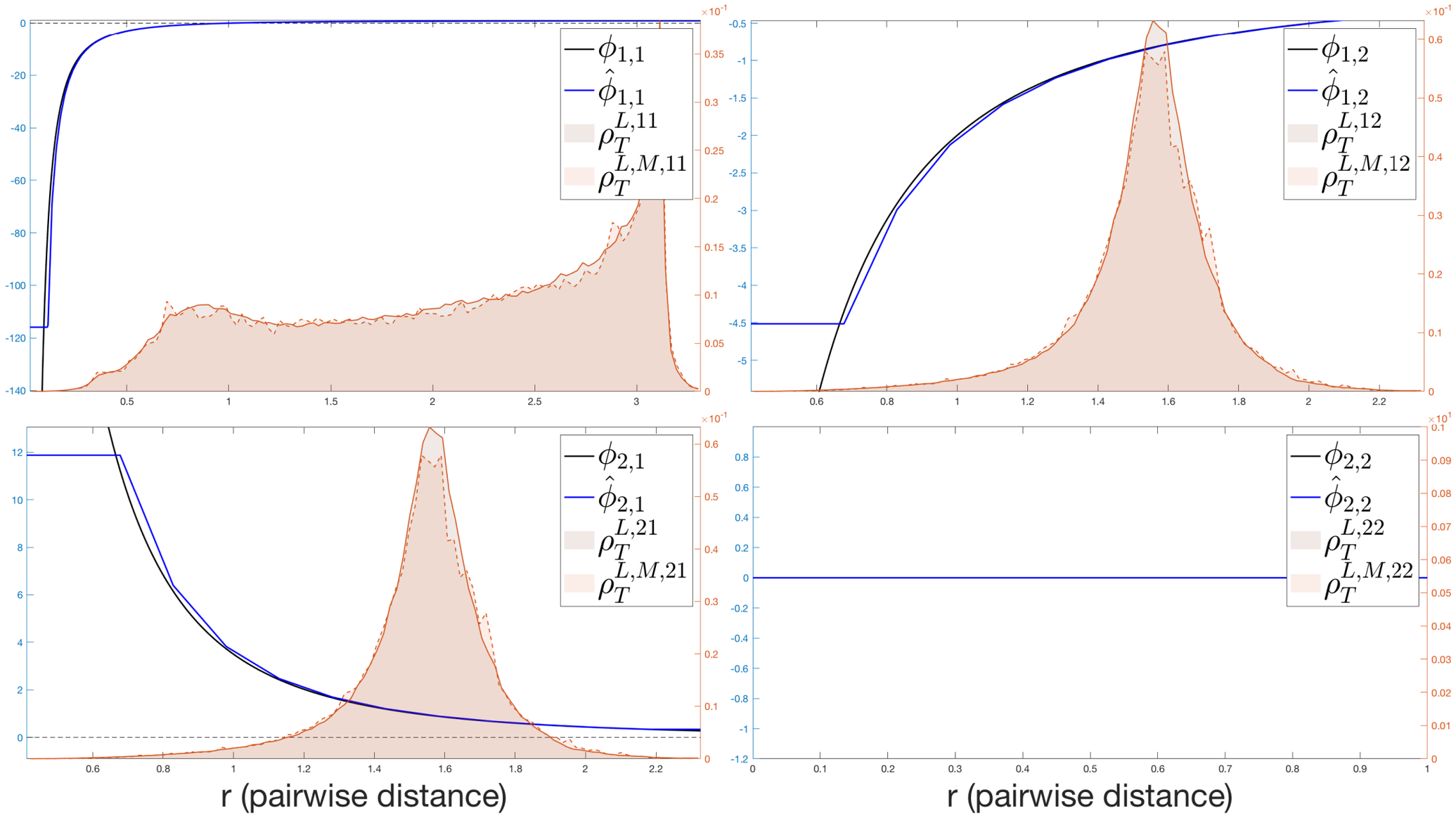
$$\mathbb{E}_{\mu_0}[\sup_{t\in[0,T]} \|\widehat{\mathbf{X}}(t) - \mathbf{X}(t)\|] \leq C(T, C_{\mathrm{Lip}})\sqrt{N}\|\hat{\phi}(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho_T)}\,,$$

where $C(T, C_{\mathrm{Lip}})$ is a constant depending on T and C_{Lip}.

Examples: multi-type agents

We may extend to first order agent systems with multiple types of agents, with different interaction kernels for each directed pair of interactions.

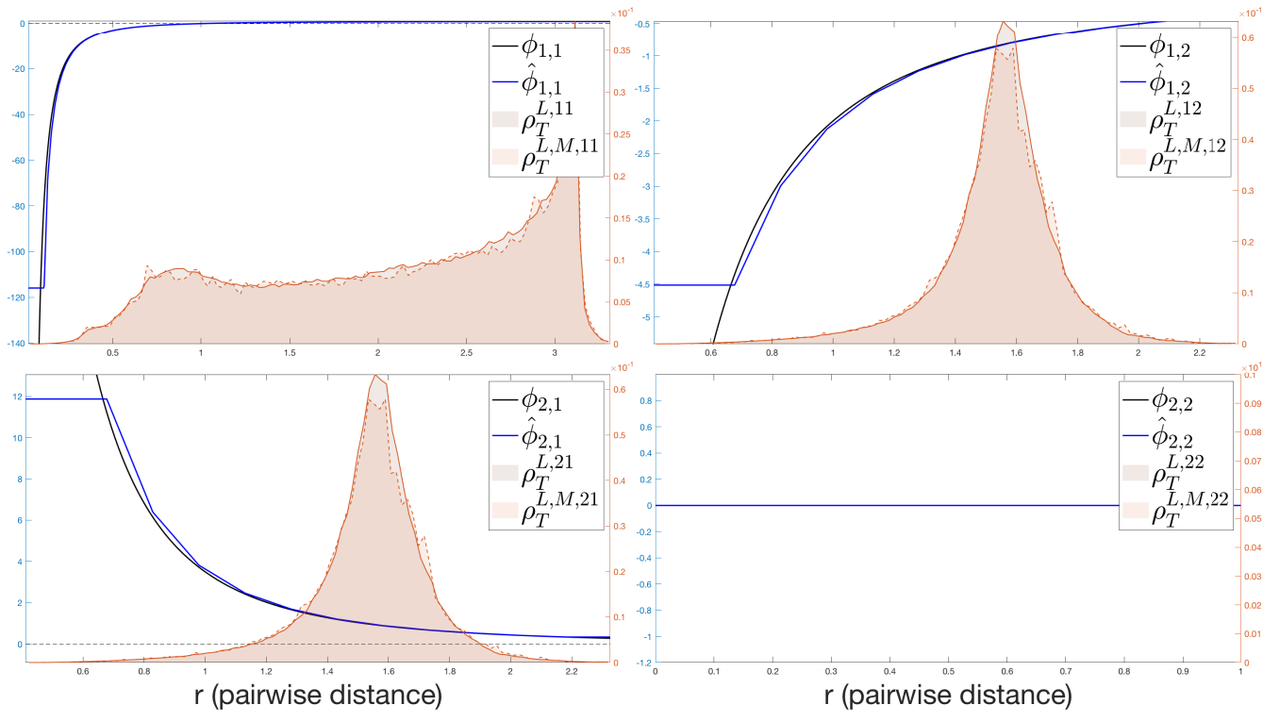
$$\dot{\boldsymbol{x}}_i(t) = \sum_{i'=1}^{N} \frac{\kappa_{k_{i'}}}{N_{k_{i'}}} \phi_{k_i k_{i'}}(r_{ii'}(t)) \boldsymbol{r}_{ii'}(t)$$

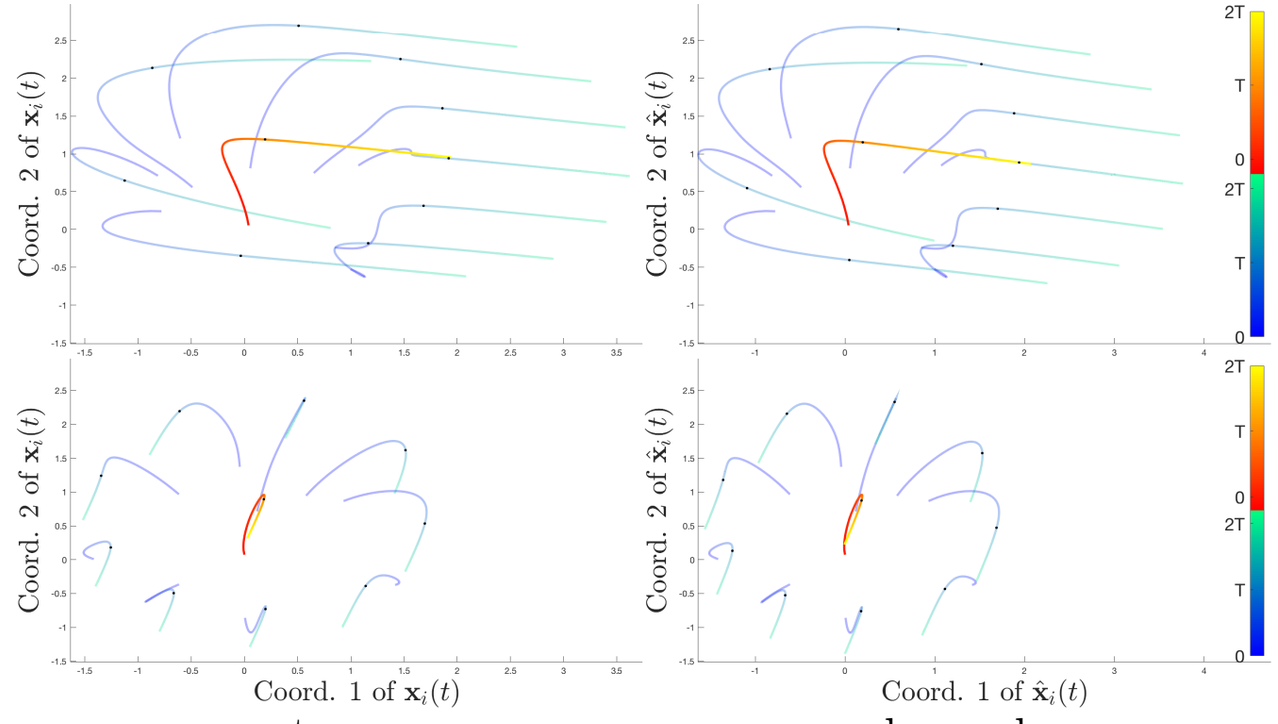


Examples: multi-type agents

We may extend to first order agent systems with multiple types of agents, with different interaction kernels for each directed pair of interactions.

$$\dot{\boldsymbol{x}}_i(t) = \sum_{i'=1}^{N} \frac{\kappa_{k_{i'}}}{N_{k_{i'}}} \phi_{k_i k_{i'}}(r_{ii'}(t)) \boldsymbol{r}_{ii'}(t)$$





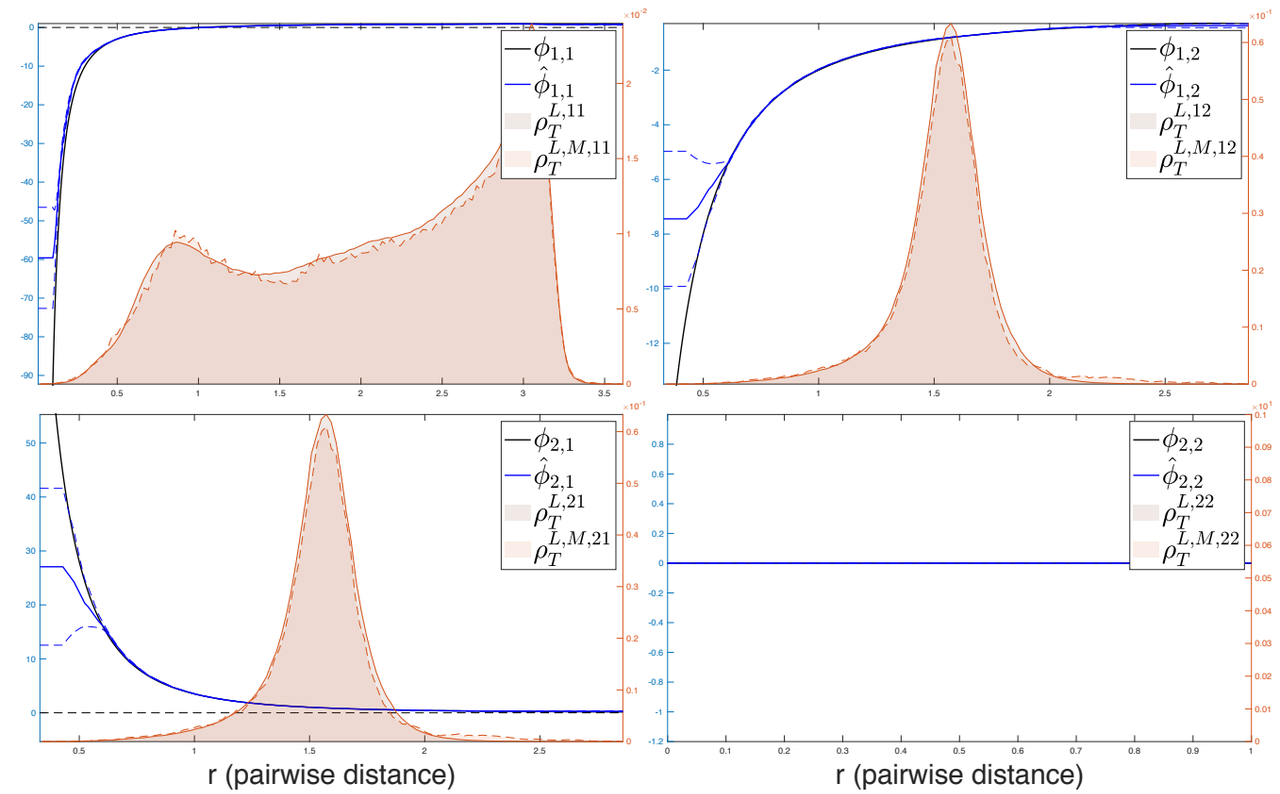
true learned

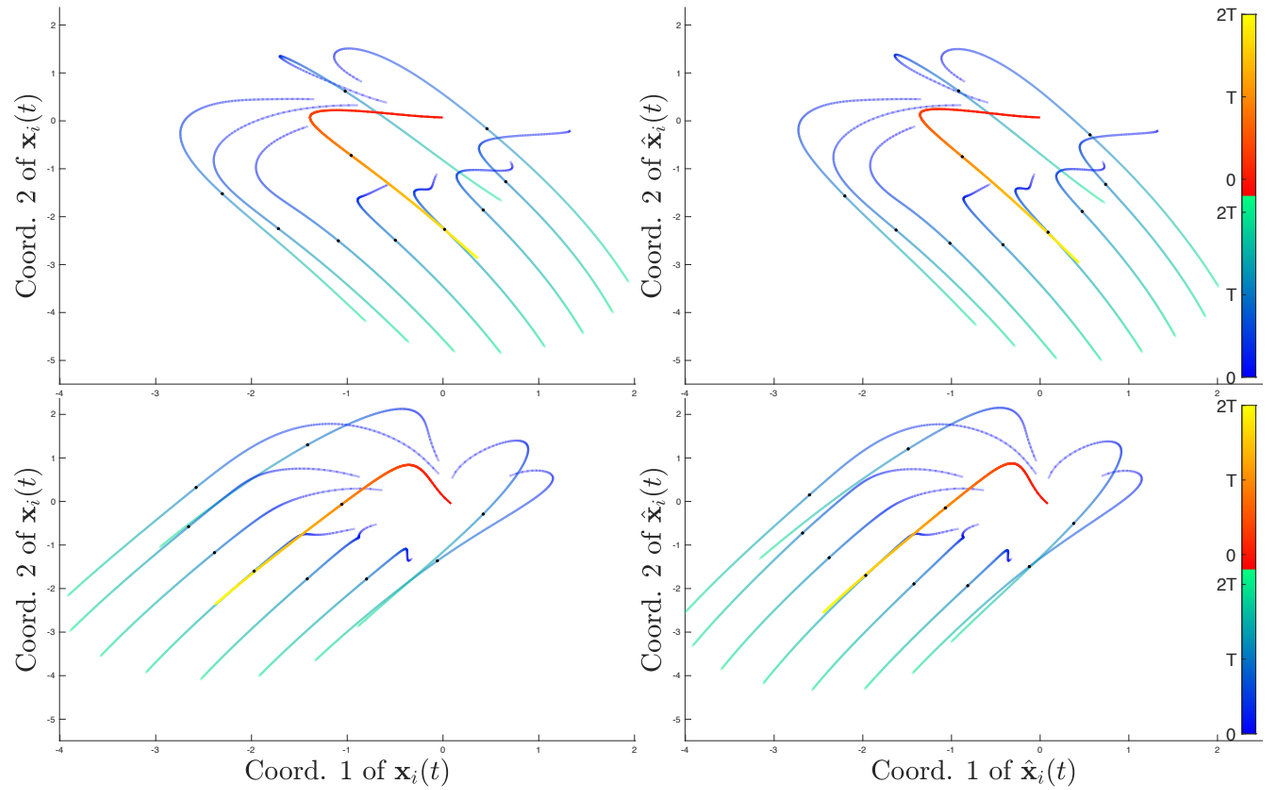
Example 1st order Prey-Predator system. Left: the interaction kernels and ρ_L^T's. Right: trajectories of the true system (left col.) and learned system (right col.) with an initial condition from training data (top) and a new one (bottom).

Examples: multi-type agents + noise

We may extend to first order agent systems with multiple types of agents, with different interaction kernels for each directed pair of interactions.

$$\dot{\boldsymbol{x}}_i(t) = \sum_{i'=1}^{N} \frac{\kappa_{k_{i'}}}{N_{k_{i'}}} \phi_{k_i k_{i'}}(r_{ii'}(t)) \boldsymbol{r}_{ii'}(t)$$





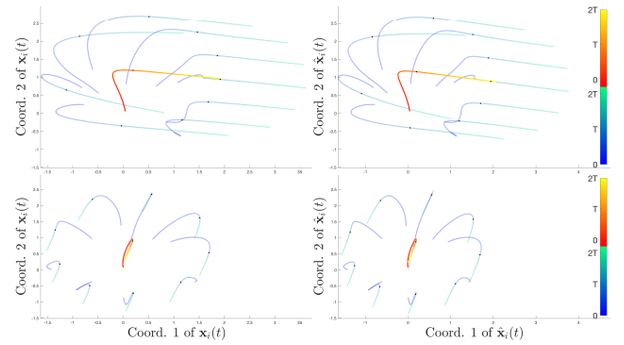
Example 1st order Prey-Predator system + noise: multiplicative noise $\sim \frac{1}{10}\text{Unif}[-\frac{1}{2}, \frac{1}{2}]$ is added to observed positions and velocities.

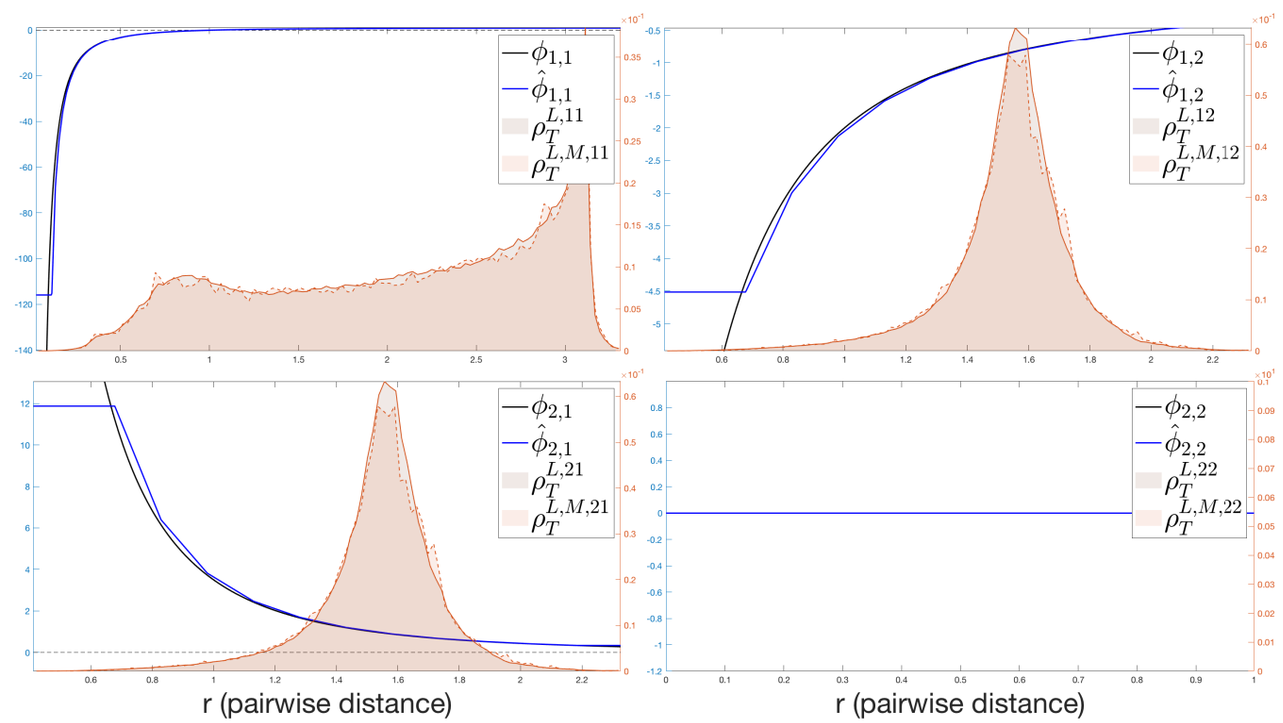
Examples: multi-type agents + scaling N

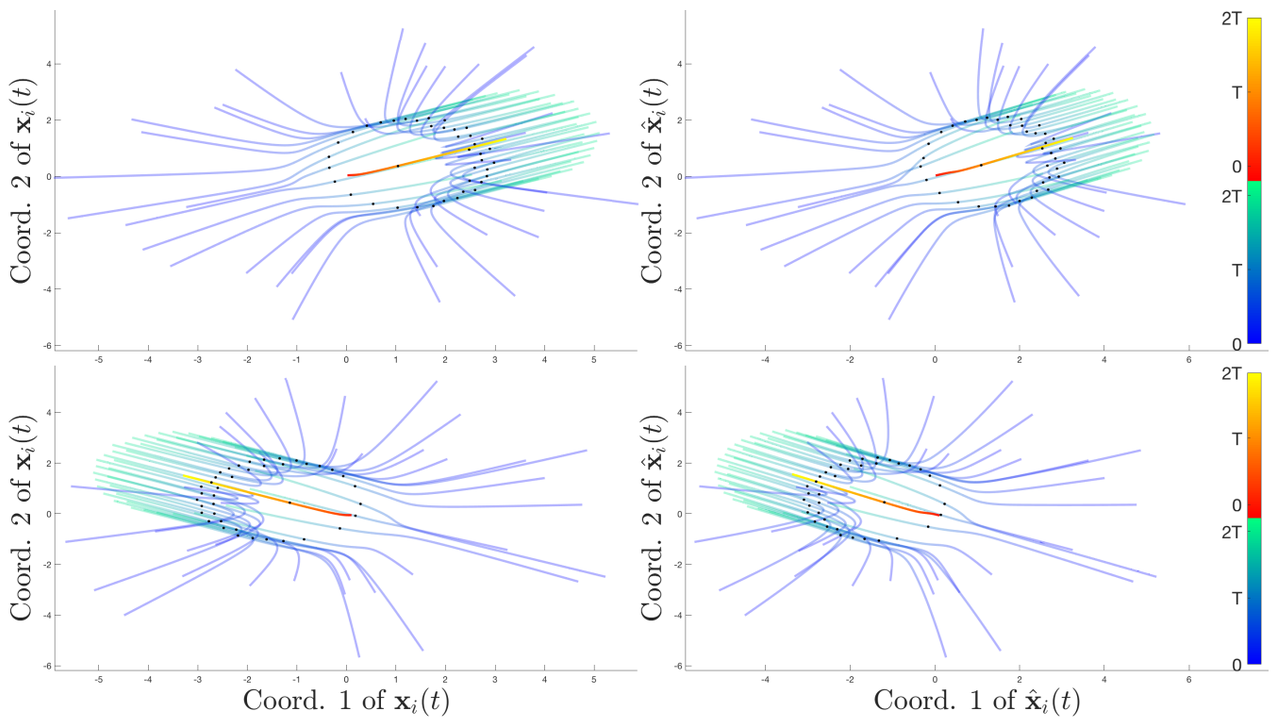
We may extend to first order agent systems with multiple types of agents, with different interaction kernels for each directed pair of interactions.

$$\dot{\boldsymbol{x}}_i(t) = \sum_{i'=1}^{N} \frac{\kappa_{k_{i'}}}{N_{k_{i'}}} \phi_{k_i k_{i'}}(r_{ii'}(t)) \boldsymbol{r}_{ii'}(t)$$

$N \mapsto 4N$



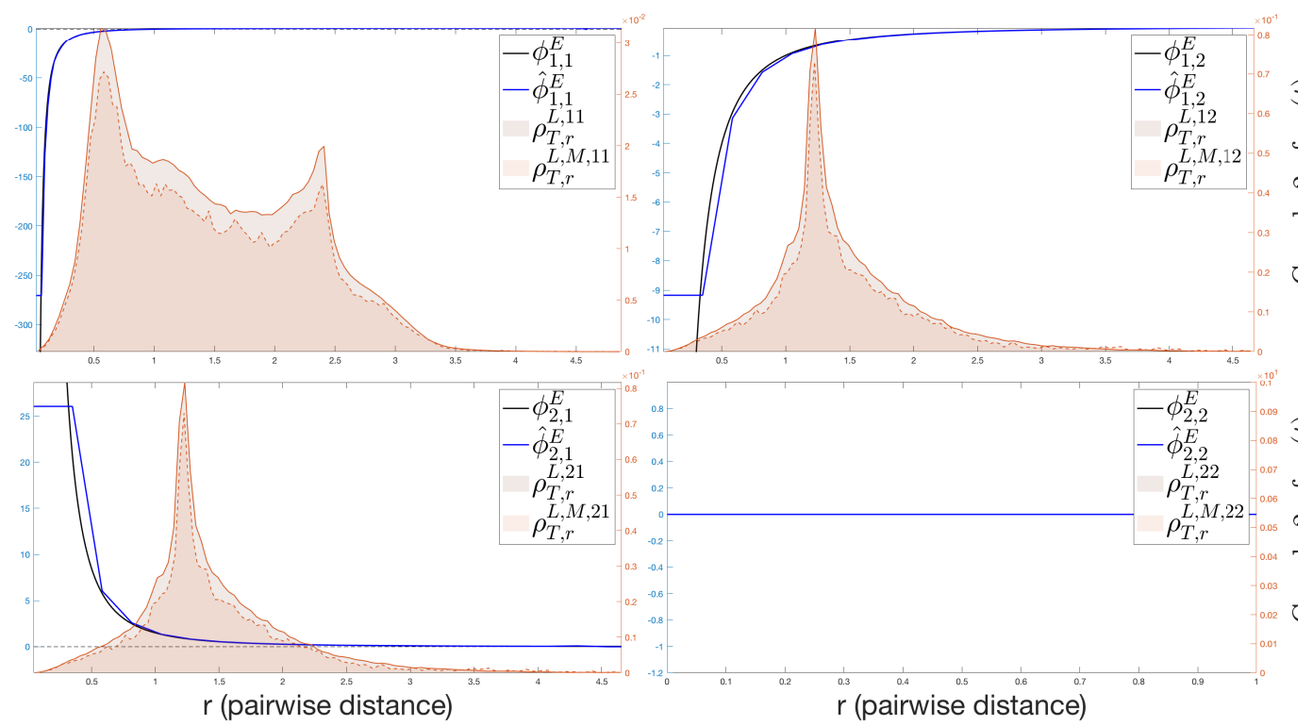


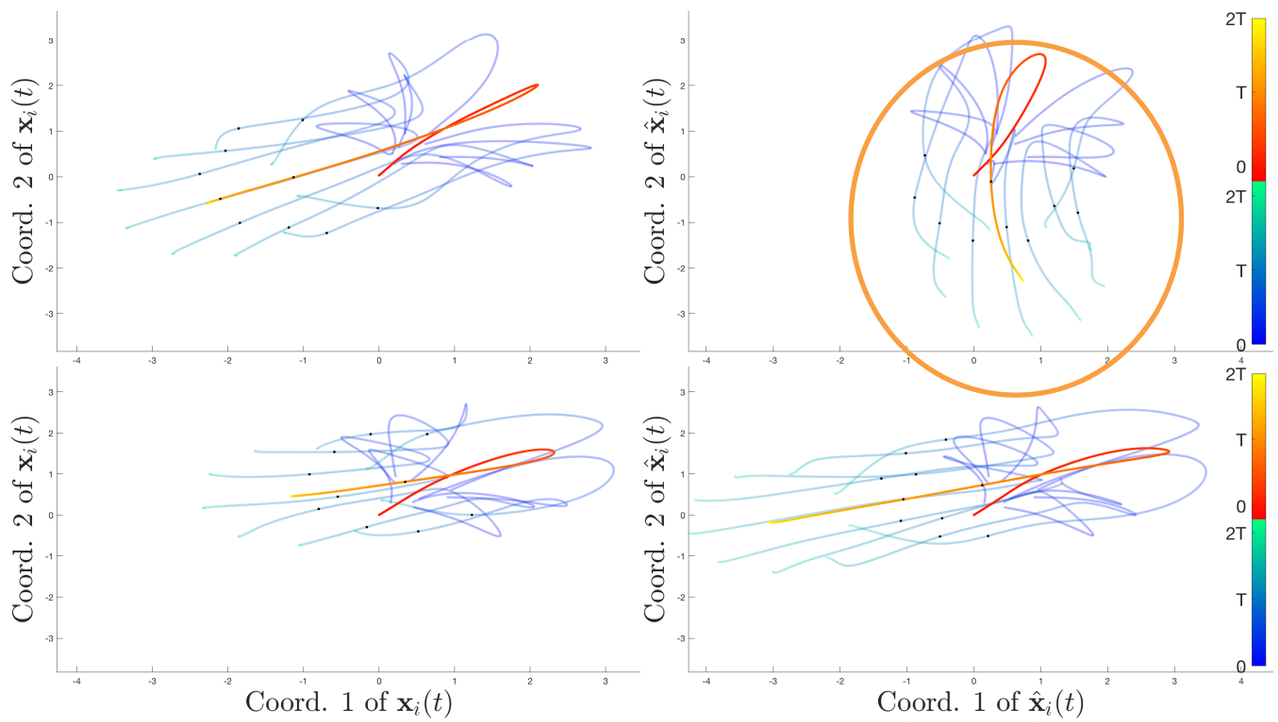


Example 1st order Prey-Predator system. Left: the interaction kernels and ρ_L^T's. Right: trajectories of the true system (left col.) and learned system (right col.) with an initial condition from training data (top) and a new one (bottom).

Examples: 2nd order systems

$$\begin{cases} m_i \ddot{\boldsymbol{x}}_i = F_i^{\boldsymbol{v}}(\dot{\boldsymbol{x}}_i, \xi_i) + \sum_{i'=1}^{N} \frac{\kappa^{\boldsymbol{v}}_{k_{i'}}}{N_{k_{i'}}} \left(\phi^E_{k_i k_{i'}}(r_{ii'}) \boldsymbol{r}_{ii'} + \phi^A_{k_i k_{i'}}(r_{ii'}) \dot{\boldsymbol{r}}_{ii'} \right) \\ \dot{\xi}_i = F_i^{\xi}(\xi_i) + \sum_{i'=1}^{N} \frac{\kappa^{\xi}_{k_{i'}}}{N_{k_{i'}}} \phi^{\xi}_{k_i k_{i'}}(r_{ii'}) \xi_{ii'} \end{cases}$$



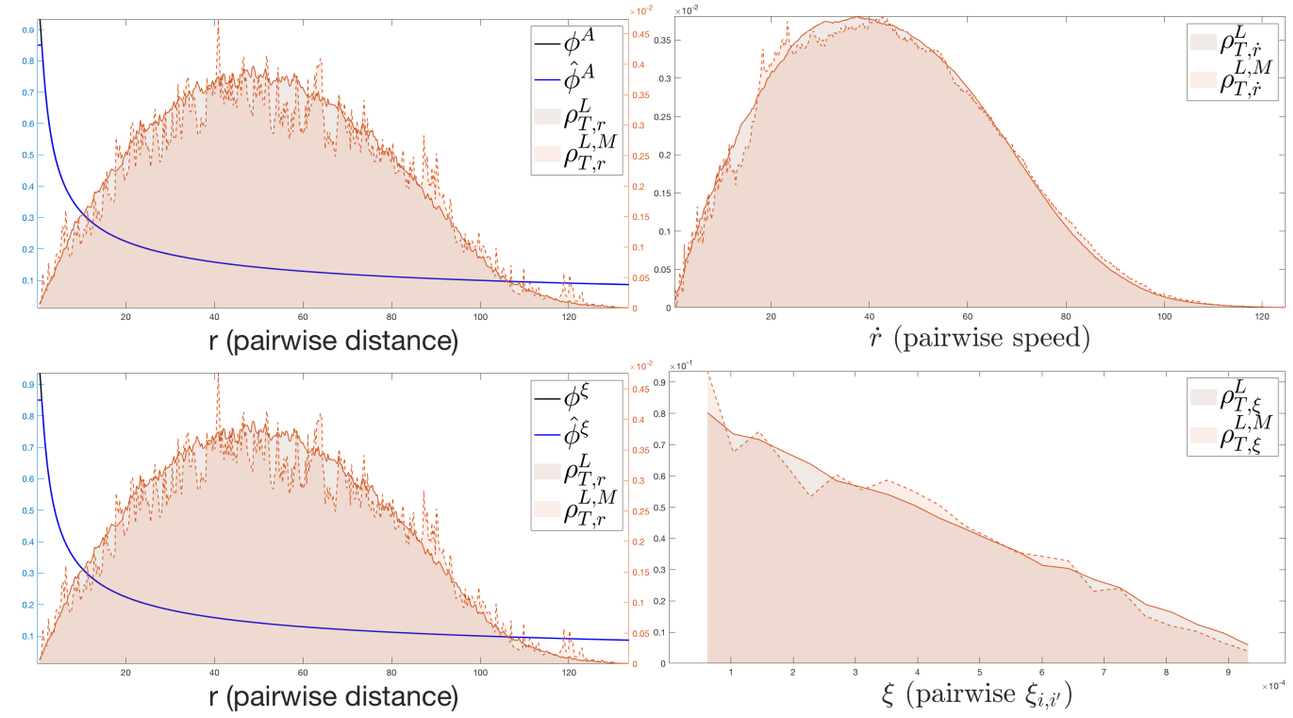


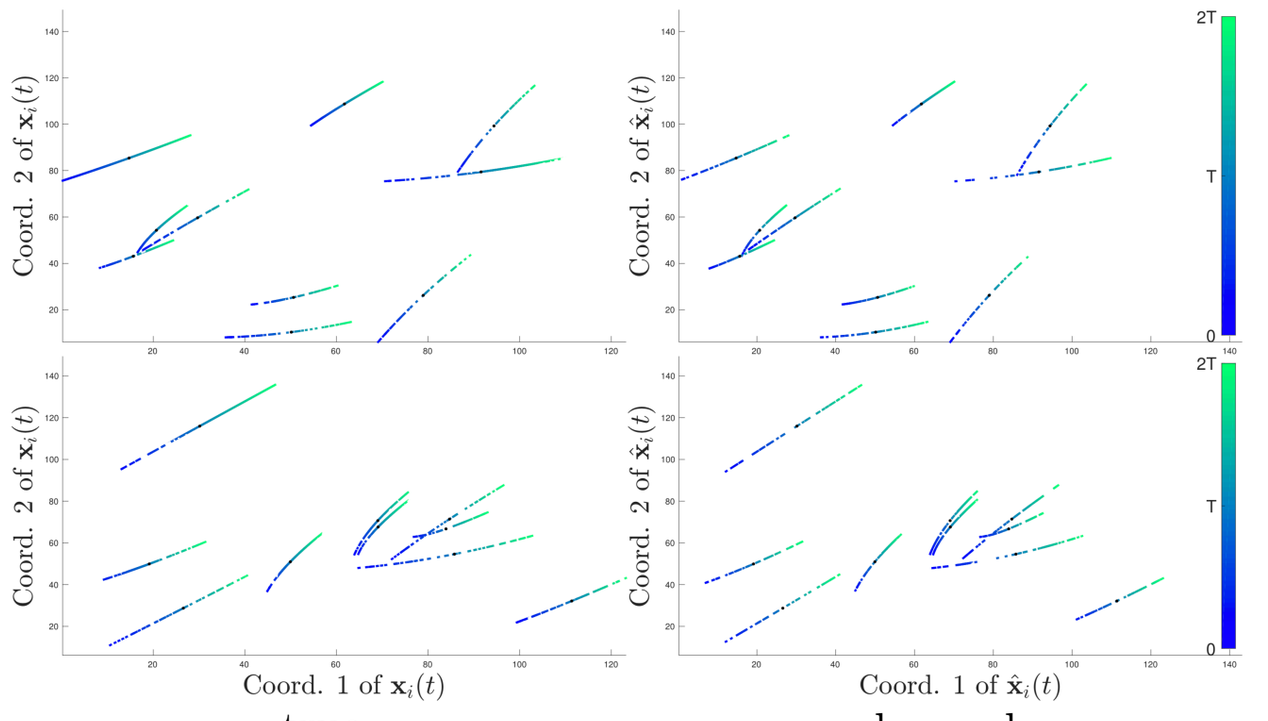
true learned

Example 2nd order Prey-Predator system. Left: the interaction kernels and ρ_L^T's. Right: trajectories of the true system (left col.) and learned system (right col.) with an initial condition from training data (top) and a new one (bottom).

Example with environment (phototaxis)

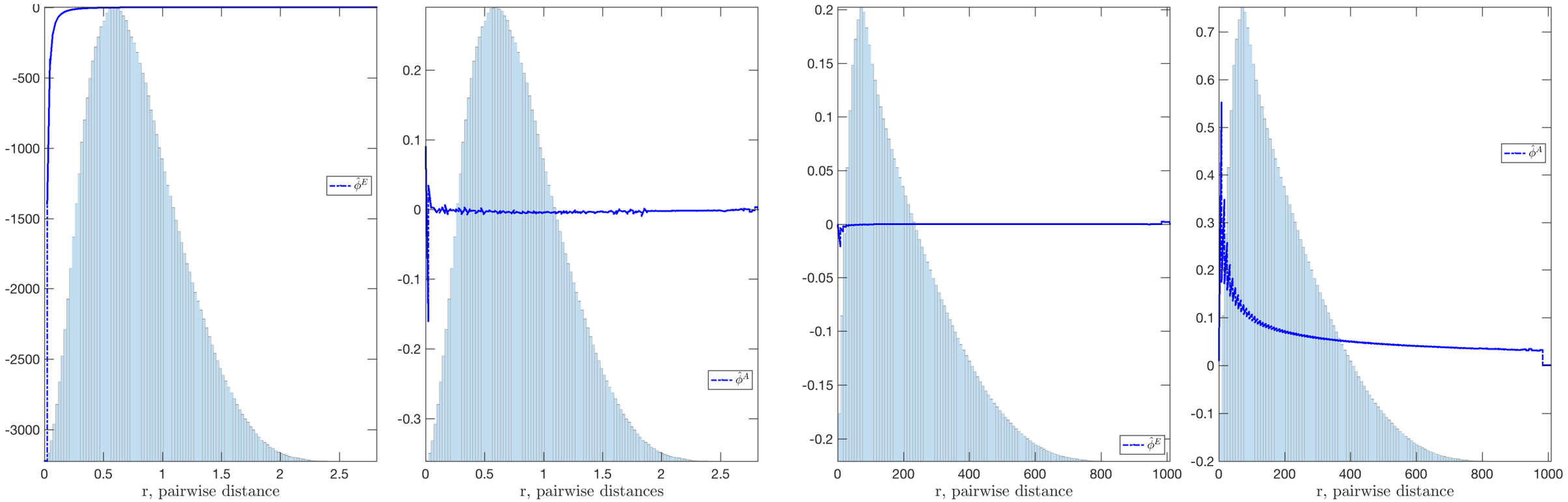
$$\begin{cases} m_i \ddot{\boldsymbol{x}}_i = F_i^{\boldsymbol{v}}(\dot{\boldsymbol{x}}_i, \xi_i) + \displaystyle\sum_{i'=1}^{N} \frac{\kappa^{\boldsymbol{v}}_{k_{i'}}}{N_{k_{i'}}} \left(\phi^E_{k_i k_{i'}}(r_{ii'}) \boldsymbol{r}_{ii'} + \phi^A_{k_i k_{i'}}(r_{ii'}) \dot{\boldsymbol{r}}_{ii'} \right) \\ \dot{\xi}_i = F_i^{\xi}(\xi_i) + \displaystyle\sum_{i'=1}^{N} \frac{\kappa^{\xi}_{k_{i'}}}{N_{k_{i'}}} \phi^{\xi}_{k_i k_{i'}}(r_{ii'}) \xi_{ii'} \end{cases}$$





Example 2nd order Phototaxis model, which includes an environment modeling light, interacting with the agents. Left: the interaction kernels and ρ_L^T's. Right: trajectories of the true system (left col.) and learned system (right col.) with an initial condition from training data (top) and a new one (bottom).

Testing hypotheses for agent systems



Example We want to test if a 2nd order system is driven by energy or alignment interactions. Left: we learn a general model (with both types of interaction) on a system with only energy interaction terms: we obtain $\hat{\phi}^A$ is ≈ 0. Right: learning on a system with only alignment term yields $\hat{\phi}^E \approx 0$.

Example We want to test if a system is governed by 1st or 2nd order interactions. We are able to tell the difference reliably, by testing the predictions of the learned models on trajectories.

True	Learned as 1^{st} order	Learned as 2^{nd} order
1^{st} order	**0.039 $\pm$ 0.16**	28 ± 21
2^{nd} order	3.1 ± 0.99	**0.58 $\pm$ 0.89**

Conclusions

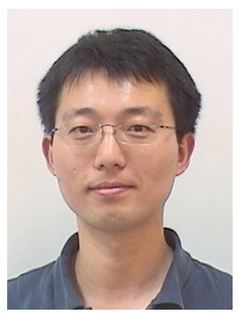
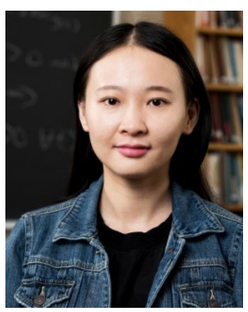
- Learning agent-based type system may be performed efficiently, nonparametrically, at least in special cases, notwithstanding the high-dimensional state space.

- Important generalizations: 1st- and 2nd-order, multi-type; more general interaction kernels.

- Hypothesis testing; transfer learning; dictionary learning for dynamical systems.

- Many open problems. E.g.: quantifying information needed for learning; stochasticity; hidden variables; general interaction kernels; ...

- Many applications: biological systems, particle systems, learning forces in molecular systems, ...

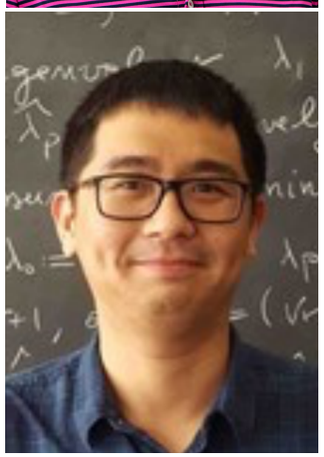




Inference of interaction laws in systems of agents from trajectory data , F. Lu, S. Tang, M. Zhong, MM

Wenjing Liao (now Asst. Prof. GAtech). Multiscale approximation to manifolds, dictionary learning; regression on manifolds.

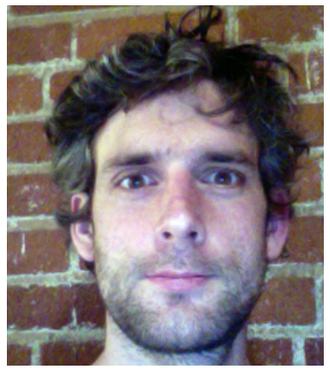
Stefano Vigogna: regression on manifolds, learning high-dimensional functions with known and unknown structure.

James Murphy (soon Asst. Prof. Tufts U.). Hyper-spectral imaging, clustering, medical data analysis.

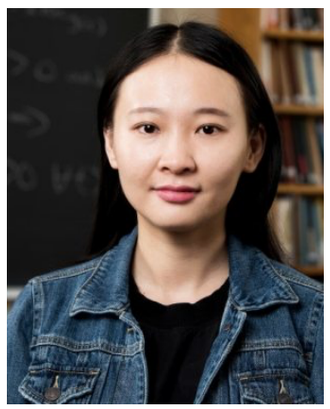


Ming Zhong. Learning of agent-based dynamical systems.

Paul Escande. Multiscale compression of non homogeneous blurring kernels for imaging; learning of maps; multi-modal data; shape analysis.



Sam Gerber (now at tech company). Multiscale optimal transportation; visualization of data.



Sui Tang. Learning of agent-based dynamical systems, in Euclidean space and on graphs. Signal processing, phase retrieval.

+ **ECG** data (modeling and prediction)
+ **Cardiac MRI** data for arrhythmia and sudden cardiac death risk assessment
+ **Learning in metric spaces**, for example for images and multi-model data
+ **Zero-shot learning** for detection of novel image classes

THANK YOU!

www.math.jhu.edu/~mauro