Inverse problems for mean-field equations of interacting particles

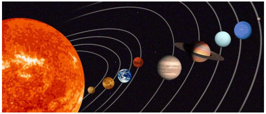
Fei Lu

Department of Mathematics, Johns Hopkins University

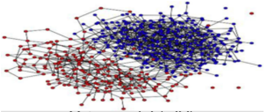
Joint with **Quanjun Lang, Qingci An, Yue Yu, Yvonne Ou**

July 15, 2023.
Institute of Computational Mathematics and
Scientific/Engineering Computing, CAS

What is the law of interaction ?



Popkin. Nature(2016)



Voter model (wiki)

$$\ddot{X}_t^i = \frac{1}{N}\sum_{j=1, j\neq i}^{N} m_j K_\phi(X_t^j - X_t^i),$$

$$K_\phi(x-y) = \nabla_x[\Phi(|x-y|)] = \phi(|x-y|)\frac{x-y}{|x-y|}.$$

- Newton's law of gravity $\phi(r) = \frac{c_1}{r^2}$
- Lennard-Jones potential: $\Phi(r) = \frac{c_1}{r^{12}} - \frac{c_2}{r^6}$.

- flocking birds, migrating cells?
- opinion dynamics ...? [a]

Infer the interaction kernel from data?

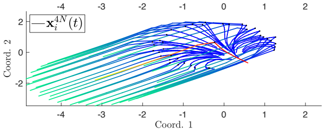
[a](1) Cucker+Smale: On the mathematics of emergence. 2007. (2) Vicsek+Zafeiris: Collective motion. 2012. (3) Motsch+Tadmor: Heterophilious Dynamics Enhances Consensus. 2014 ...

Learning the interaction kernel ϕ

$$dX_t^i = \frac{1}{N}\sum_{j=1}^{N} K_{\phi}(X_t^j - X_t^i)dt + \sqrt{2\nu}dB_t^i \quad \Leftrightarrow \dot{\boldsymbol{X}}_t = R_{\phi}(\boldsymbol{X}_t) + \sqrt{2\nu}\dot{\boldsymbol{B}}_t$$

Finite N: ("... 4 years ago ...")

- Data: M trajectories of particles $\{\boldsymbol{X}_{t_1:t_L}^{(m)}\}_{m=1}^{M}$
- Statistical learning

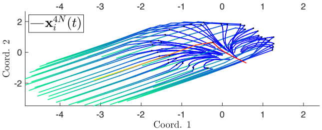


Learning the interaction kernel ϕ

$$dX_t^i = \frac{1}{N}\sum_{j=1}^{N} K_\phi(X_t^j - X_t^i)dt + \sqrt{2\nu}dB_t^i \quad \Leftrightarrow \dot{\boldsymbol{X}}_t = R_\phi(\boldsymbol{X}_t) + \sqrt{2\nu}\dot{\boldsymbol{B}}_t$$

Finite N: ("... 4 years ago ...")

- Data: M trajectories of particles $\{\boldsymbol{X}^{(m)}_{t_1:t_L}\}_{m=1}^M$
- Statistical learning

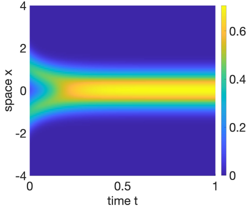


Large N ($>> 1$)

- Data: density of particles
  $\{u(x_m, t_l) \approx N^{-1}\sum_i \delta(X_{t_l}^i - x_m)\}_{m,l}$

$$\partial_t u = \nu\Delta u + \nabla\cdot[u(K_\phi * u)]$$

- Inverse problem for a PDE



Goal: algorithm, identifiability, convergence

Inverse problem for Mean-field PDE

Goal: Identify from data ϕ in

$$\partial_t u = \nu \Delta u + \nabla \cdot [u(K_\phi * u)], \quad x \in \mathbb{R}^d, t > 0,$$

where $K_\phi(x) = \nabla(\Phi(|x|)) = \phi(|x|)\frac{x}{|x|}$.

- Two types of data:
  - low-D: discrete data $\{u(x_m, t_l)\}_{m,l=1}^{M,L}$ with mesh $\{x_m\}$
  - high-D: particle samples $\{u_N(x, t_l) \approx M^{-1} \sum_{i=1}^{M} \delta(X_{t_l}^i - x)\}$
- Two types of equations: $\nu > 0$ or $\nu = 0$.

How? General & computationally efficient?

Inverse problem for Mean-field PDE

Goal: Identify from data ϕ in

$$\partial_t u = \nu \Delta u + \nabla \cdot [u(K_\phi * u)], \quad x \in \mathbb{R}^d, t > 0,$$

where $K_\phi(x) = \nabla(\Phi(|x|)) = \phi(|x|)\frac{x}{|x|}$.

- Two types of data:
  - low-D: discrete data $\{u(x_m, t_l)\}_{m,l=1}^{M,L}$ with mesh $\{x_m\}$
  - high-D: particle samples $\{u_N(x, t_l) \approx M^{-1} \sum_{i=1}^{M} \delta(X_{t_l}^i - x)\}$
- Two types of equations: $\nu > 0$ or $\nu = 0$.

How? General & computationally efficient?

- Variational /regression: loss functional
- Identifiability, Ill-posed: regularization

Loss functional

$$\partial_t u = \nu \Delta u + \nabla \cdot [u(K_\phi * u)]$$

Candidates:

- Discrepancy: $\mathcal{E}(\phi) = \|\partial_t u - \nu \Delta u - \nabla.(u(K_\phi * u))\|^2$
  - derivatives approx. from discrete data
  - Weak SINDY [Bortz etc21,22], denoising+smoothing [Kang+Liao etc22]
- Free energy: $\mathcal{E}(\phi) = C + |\int_{\mathbb{R}^d} u[(\Phi - \Phi_{true}) * u]dx|^2$

  limitted information from the 1st moment
- Wasserstein-2: $\mathcal{E}(\phi) = W_2(u^\phi, u)$

  costly: requires many PDE simulations in optimization
- **A probabilistic loss function** $\downarrow$
- A self-test loss function: simple, general

A probabilistic loss functional

$$\mathcal{E}(\phi) := \frac{1}{T}\int_0^T \int_{\mathbb{R}^d} \Big[\big| K_\phi * u \big|^2 u - 2\nu u (\nabla \cdot K_\phi * u) + 2\partial_t u (\Phi * u) \Big] \, dx \, dt$$

- $= -\mathbb{E}[$ log-likelihood $]$: McKean–Vlasov process

$$\begin{cases} d\overline{X}_t = - K_{\phi_{true}} * u(\overline{X}_t, t) dt + \sqrt{2\nu} dB_t, \\ \mathcal{L}(\overline{X}_t) = u(\cdot, t), \end{cases}$$

- Derivative free
- Suitable for high dimension: $Z_t = \overline{X}_t - \overline{X}_t'$

$$\mathcal{E}(\phi) = \frac{1}{T}\int_0^T \Big(\mathbb{E}|\mathbb{E}[K_\phi(Z_t)|\overline{X}_t]|^2 - 2\nu \mathbb{E}[\nabla \cdot K_\phi(Z_t)] + \partial_t \mathbb{E}\Phi(Z_t) \Big) \, dt$$

A self-test loss function

Weak form of the equation

$$
\begin{aligned}
\langle \partial_t u, v \rangle &= \nu \langle \Delta u, v \rangle + \langle \nabla \cdot [u(K_\phi * u)], v \rangle \\
&= \nu \langle u, \Delta v \rangle - \langle u(K_\phi * u), \nabla v \rangle, \quad \forall v \in C^\infty \ldots
\end{aligned}
$$

Take $v = \Phi * u$ s.t. $\nabla \Phi(|x|) = K_\phi(x) = \phi(|x|)\frac{x}{|x|}$,

$$
\langle \partial_t u, \Phi * u \rangle = \nu \langle u, \Delta \Phi * u \rangle - \langle u(K_\phi * u), K_\phi * u \rangle
$$

We regain the loss function

$$
\mathbb{E}(\phi) = \int_0^T [\langle \partial_t u, \Phi * u \rangle - \nu \langle u, \Delta \Phi * u \rangle + \langle u(K_\phi * u), K_\phi * u \rangle] dt
$$

- regardless of $\nu = 0$ or > 0
- Applicable to other PDEs: self-test (a better name?)

Nonparametric regression $\phi = \sum_{i=1}^{n} c_i\phi_i \in \mathcal{H}_n$:

$$\mathcal{E}_M(\phi) = c^\top Ac - 2b^\top c \quad \Rightarrow \quad \widehat{\phi}_{n,M} = \sum_{i=1}^{n} \widehat{c}_i\phi_i, \quad \widehat{c} = A^{-1}b$$

- Choice of $\mathcal{H}_n$ & function space of learning?
  - Exploration measure $\rho_T \leftarrow |\overline{X}_t - \overline{X}_t'|$
- Inverse problem well-posedness/ identifiability?
  - $\underset{\phi \in L^2(\rho)}{\arg\min}\ \mathcal{E}(\phi)$
- Convergence and rate? $\Delta x = M^{-1/d} \to 0$

Identifiability

$$\mathcal{E}(\phi) = \langle L_{\overline{G}}\phi, \phi\rangle - 2\langle \phi^D, \phi\rangle + const.$$
$$\nabla\mathcal{E}(\phi) = L_G\phi - \phi^D = 0 \quad \Rightarrow \widehat{\phi} = L_G^{-1}\phi^D$$

- Identifiability: $A^{-1}b \leftrightarrow L_{\overline{G}}^{-1}\phi^D$
  - $L_{\overline{G}}$: positive compact operator
  - Function space of identifiability (FSOI): $\overline{\mathrm{span}\{\psi_i\}_{\lambda_i>0}}$
- Coercivity condition on $\mathcal{H}$ (not $L^2(\rho)$)

$$c_{\mathcal{H}} = \inf_{\phi\in\mathcal{H}, \|\phi\|_{L^2(\rho_T)}=1} \langle L_{\overline{G}}\phi, \phi\rangle > 0$$

Convergence rate

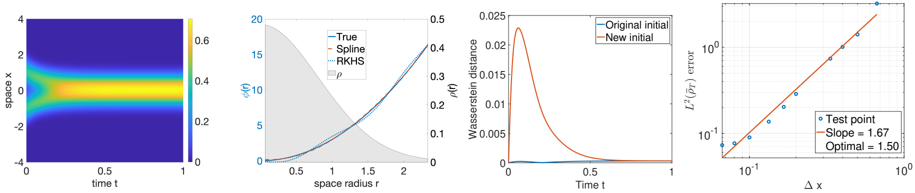
Theorem (Error bound [Lang-Lu22sisc])

Let $\mathcal{H}_n = \mathrm{span}\{\phi_i\}_{i=1}^n$ s.t. $\|\phi_{\mathcal{H}_n} - \phi\|_{L^2(\rho_T)} \lessapprox n^{-s}$. Assume the coercivity condition on $\cup\mathcal{H}_n$. Then, with $n \approx (\Delta x)^{-\alpha/(s+1)}$, we have:

$$\|\widehat{\phi}_{n,M} - \phi\|_{L^2(\rho_T)} \lessapprox (\Delta x)^{\alpha s/(s+1)}$$

- Δx^α comes from numerical integrator (e.g.,Riemann sum)
  - In statistical learning: $\alpha = 1/2$ (Monte Carlo, CLT)
- Trade-off: numerical error v.s. approximation error

Example: granular media $\phi(r) = 3r^2$



Data $u(x,t)$ Estimator Wasserstein-2 Rate

- Near optimal rate ($\phi \in W^{1,\infty}$)
- Other examples:
  - suboptimal when ϕ discontinuous,
  - low rate for singular ϕ

Learning kernels in operators

Learn the kernel ϕ:

$$R_\phi[u] = f$$

from data:

$$\mathcal{D} = \{(u_k, f_k)\}_{k=1}^N, \quad (u_k, f_k) \in \mathbb{X} \times \mathbb{Y}$$

- R_ϕ linear/nonlinear in u, but linear in ϕ
- Examples:
  - interaction kernel: $R_\phi[u] = \nabla \cdot [u(K_\phi * u)] = \partial_t u - \nu \Delta u$
  - Toeplitz/Hankel matrix
  - integral/nonlocal operators,...

Ill-posed inverse problem

$$\mathcal{E}(\phi) = \|R_\phi[u] - f\|_{\mathbb{Y}}^2 = \langle L_G\phi, \phi\rangle_{L^2(\rho)} - 2\langle \phi^D, \phi\rangle_{L^2(\rho)} + C$$
$$\nabla\mathcal{E}(\phi) = L_G\phi - \phi^D = 0 \quad \Rightarrow \widehat{\phi} = L_G^{-1}\phi^D$$

Ill-posed inverse problem

$$\mathcal{E}(\phi) = \|R_\phi[u] - f\|_{\mathbb{Y}}^2 = \langle L_G\phi, \phi\rangle_{L^2(\rho)} - 2\langle \phi^D, \phi\rangle_{L^2(\rho)} + C$$

$$\nabla\mathcal{E}(\phi) = L_G\phi - \phi^D = 0 \quad \Rightarrow \widehat{\phi} = L_G^{-1}\phi^D$$

Regularization

$$\mathcal{E}_\lambda(\phi) = \mathcal{E}(\phi) + \lambda\|\phi\|_Q^2 \rightarrow \widehat{\phi} = (L_G + \lambda Q)^{-1}\phi^D$$

- λ by the L-curve method [Hansen00]
- Regularization norm $\|\cdot\|_Q$? $Q = Id$, $Q = RKHS$? [many, Zhou13...]

Ill-posed inverse problem

$$\mathcal{E}(\phi) = \|R_\phi[u] - f\|_{\mathbb{Y}}^2 = \langle L_G \phi, \phi \rangle_{L^2(\rho)} - 2\langle \phi^D, \phi \rangle_{L^2(\rho)} + C$$

$$\nabla \mathcal{E}(\phi) = L_G \phi - \phi^D = 0 \quad \Rightarrow \widehat{\phi} = L_G^{-1} \phi^D$$

Regularization

$$\mathcal{E}_\lambda(\phi) = \mathcal{E}(\phi) + \lambda \|\phi\|_Q^2 \to \widehat{\phi} = (L_G + \lambda Q)^{-1} \phi^D$$

- λ by the L-curve method [Hansen00]
- Regularization norm $\|\cdot\|_Q$? $Q = Id$, $Q = RKHS$? [many, Zhou13...]

Data Adaptive RKHS Tikhonov Regularization [Lu+Lang+An22]

- norm of RKHS $H_G = L_G^{1/2} L^2(\rho) \leftrightarrow Q = L_G^{-1}$
- L_G is data dependent
- Computation: $\widehat{\phi} = (L_G + \lambda L_G^{-1})^{-1} \phi^D = (L_G^2 + \lambda I)^{-1} L_G \phi^D$

Regularization norms in computational practice:

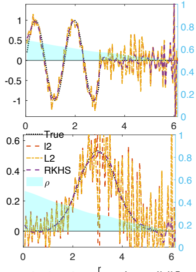
Table: Three regularizers using the norms of l^2, L^2_ρ and RKHS.

Regularizer name	$\mathbf{C}$	Regularized estimator
l2	$\mathbf{I}$	$\phi_\lambda^{l^2} = (\mathbf{A} + \lambda \mathbf{I})^{-1}\mathbf{b}$
L2	$\mathbf{B}$	$\phi_\lambda^{L^2} = (\mathbf{A} + \lambda \mathbf{B})^{-1}\mathbf{b}$
RKHS	$\mathbf{C}_{rkhs}$	$\phi_\lambda^{H_G} = (\mathbf{A} + \lambda \mathbf{C}_{rkhs})^{-1}\mathbf{b}$

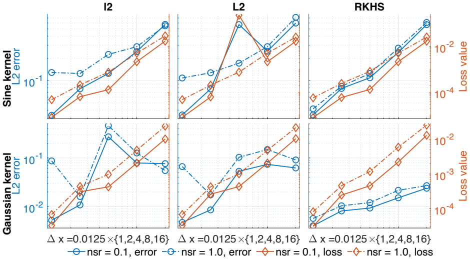
DARTR: Data Adaptive RKHS Tikhonov Regularization

$$R_{\phi}[u] = \nabla \cdot [u(K_{\phi} * u)] = f$$

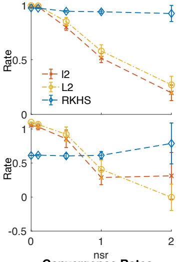
- Recover kernel from discrete noisy data
- Consistent convergence as mesh refines
- Recover nonlocal kernel in homogenization [Lu+An+Yue22]



Typical estimators, $\Delta x = 0.05$



Convergence of Estimators, nsr = 0.1 & 1



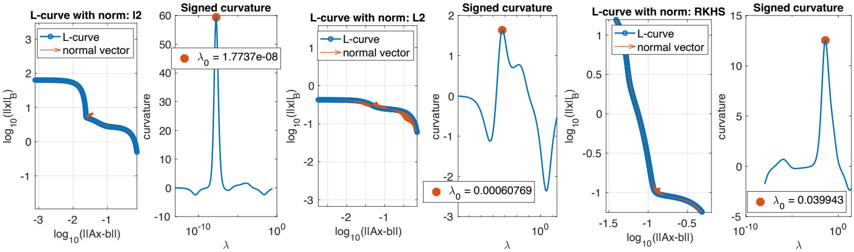
Convergence Rates

Why DARTR is better?
When / v.s. other norms?
Convergence rate?

Why DARTR is better?
When / v.s. other norms?
Convergence rate?

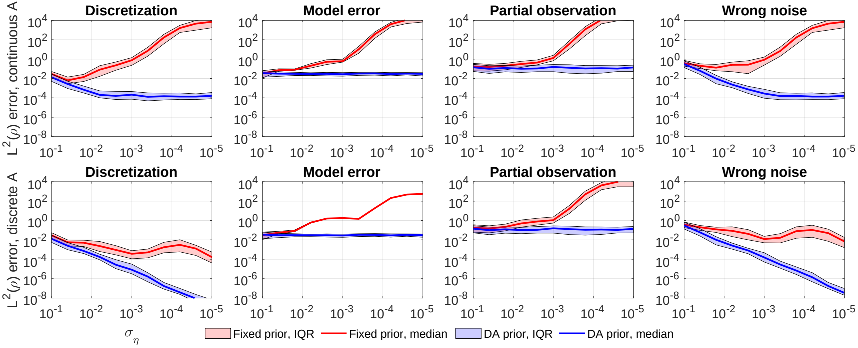
- Empirical: more robust L-curve
- Bayesian perspective: an adaptive prior [Chada+Lang+Lu+Xiong22]
- Fredholm equation: explicit RKHS [Lu+Ou23]
- Small noise analysis: fractional RKHSs [Lang+Lu23]
- Convergence rate: open, possible

More robust L-curve:



Bayesian: small noise limit of maximum of posterior

- $Q = I$: divergent estimator
- $Q = L_G^{-1}$: stable/convergent



DARTR for Fredholm equation

$$y(t) = \int_0^1 K(t,s)\phi(s)ds + \sigma \dot{W}(t),\ t \in \{t_i\}_{i=1}^m \subset [0,1].$$

$$G(s,s') := \int_0^1 K(t,s)K(t,s')\mu_m(dt), \quad \forall (s,s').$$

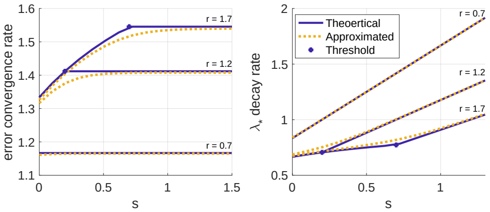
- RKHS with G as reproducing kernel: $H_G = L_G^{-1/2}(L^2_\rho)$
- G adaptive to data and the equation
- Nashed-Wahba74,..., Wahba77:
  - RKHS regularization, not G
  - Convergence of CV estimator

Small noise analysis for RKHS regularization

$$\widehat{\phi}^s_\lambda = (L_G + \lambda L_G^{-s})^{-1}\phi^D$$

- $s = 0$: L^2_ρ regularization
- $s > 0$: fractional RKHS ($s = 1$: RKHS)

$$\|\widehat{\phi}^s_\lambda - \phi_*\|^2_{L^2_\rho} = \sum_i (\lambda_i + \lambda\lambda_i^{-s})^{-2}(\sigma\lambda_i^{1/2}\xi_i - \lambda c_i)^2 + \sum_j d_j^2,$$



Surprise: over-smoothing OK in theory, but harder to select λ

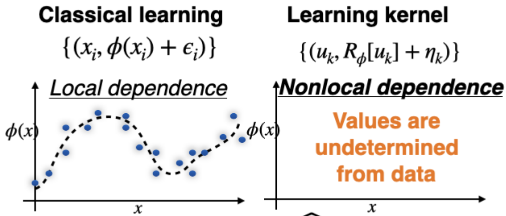
Summary and future directions

Inverse problems for mean-field PDE of interacting particles

- Construction of loss functions
- Nonparametric regression: identifiability
- Regularization: adaptive RKHSs
  Learning kernels in operators

Learning with nonlocal dependence:

statistical learning + inverse problem



Inversion $\widehat{\phi} = I^{-1}\phi^D$ $\qquad \widehat{\phi} = L_G^{-1}\phi^D$

Regularization $\widehat{\phi} = (I + \lambda Q)^{-1}\phi^D$ $\qquad \widehat{\phi} = (L_G + \lambda L_G^{-1})^{-1}\phi^D$

- Coercivity condition/ spectrum decay
- Convergence (minimax rate)
- High-D ϕ:
  - Iterative methods?
  - Regularization for NN in function space?