Statistical learning and inverse problems from interacting particle systems

Fei Lu

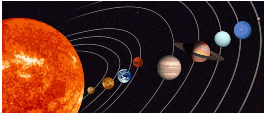
Department of Mathematics, Johns Hopkins University

June 15, 2023
Tianyuan Mathematical Center Conference:
Random Dynamical System
Wuhan University

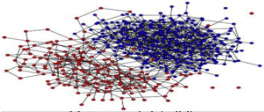
JOHNS HOPKINS UNIVERSITY

NSF

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?

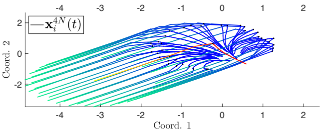
[a] (1) Cucker+Smale: On the mathematics of emergence. 2007. (2) Vicsek+Zafeiris: Collective motion. 2012. (3) Motsch+Tadmor: Heterophilious Dynamics Enhances Consensus. 2014 ...

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:

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

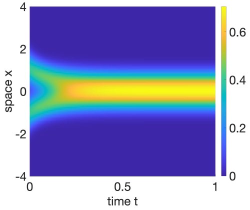


Large N ($>>$ 1)

- 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

Part 1: Finitely many particles

Statistical learning from M sample trajectories

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

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

Finitely many particles

$$R_{\phi}(\boldsymbol{X}_t) = \dot{\boldsymbol{X}}_t - \sqrt{2\nu}\dot{\boldsymbol{B}}_t \quad \text{\& Data } \{\boldsymbol{X}^{(m)}_{t_1:t_L}\}_{m=1}^{M}$$

- Loss function (or log-likelihood for SDEs):

$$\hat{\phi}_{n,M} = \arg\min_{\phi\in\mathcal{H}_n} \mathcal{E}_M(\phi) = \frac{1}{M}\sum_{m=1}^{M}\int_0^T |\dot{\boldsymbol{X}}_t^m - R_{\phi}(\boldsymbol{X}_t^m)|^2 dt$$

- Nonparametric Regression: $\mathcal{H}_n = \mathrm{span}\{\phi_i\}_{i=1}^n$, $\phi = \sum_i c_i\phi_i$

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

Finitely many particles

$$R_{\phi}(\boldsymbol{X}_t) = \dot{\boldsymbol{X}}_t - \sqrt{2\nu}\dot{\boldsymbol{B}}_t \quad \& \text{ Data } \{\boldsymbol{X}^{(m)}_{t_1:t_L}\}_{m=1}^M$$

- Loss function (or log-likelihood for SDEs):

$$\hat{\phi}_{n,M} = \arg\min_{\phi\in\mathcal{H}_n} \mathcal{E}_M(\phi) = \frac{1}{M}\sum_{m=1}^{M}\int_0^T |\dot{\boldsymbol{X}}_t^m - R_{\phi}(\boldsymbol{X}_t^m)|^2 dt$$

- Nonparametric Regression: $\mathcal{H}_n = \mathrm{span}\{\phi_i\}_{i=1}^n$, $\phi = \sum_i c_i\phi_i$

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

  - Choice of $\mathcal{H}_n$? function space?
  - Identifiability/Well-posedness?
  - Convergence and rate?

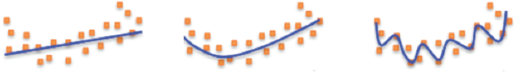
Classical learning in a nutshell

$$\text{Data}\{(x_m, y_m)\}_{m=1}^M \sim (X, Y) \Rightarrow \text{find } \phi \text{ s.t. } Y = \phi(X)$$

- Loss function: $\hat{\phi}_{n,M} = \arg\min_{\phi\in\mathcal{H}_n} \mathcal{E}_M(\phi) = \frac{1}{M}\sum_{m=1}^M |Y_m - \phi(X_m)|^2$.
- Regression: with $\psi = \sum_i c_i\phi_i \in \mathcal{H}_n = \mathrm{span}\{\phi_i\}_{i=1}^n$:

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

- ▸ Choice of $\mathcal{H}_n \subset C^s$ in $L^2(\rho_X)$: $n_* = (M/\log M)^{\frac{1}{2s+d}}$



- ▸ Well-posed: $\phi_{optimal} = \mathbb{E}[Y|X=x] = \arg\min_{\phi\in L^2(\rho_X)} \mathcal{E}(\phi)$
- ▸ Minimax rate $\mathbb{E}[\|\widehat{\phi}_{n_*,M} - \phi_{optimal}\|^2_{L^2(\rho_X)}] \approx \left(\frac{\log M}{M}\right)^{\frac{s}{2s+d}}$

Learning kernel

$$\text{Given: Data}\{\boldsymbol{X}_{[0,T]}^{(m)}\}_{m=1}^{M}$$

$$\text{Goal: Estimate } \phi \text{ s.t. } \dot{\boldsymbol{X}}_t \approx R_\phi(\boldsymbol{X}_t) = [\frac{1}{N}\sum_{j=1}^{N} K_\phi(X_t^j, X_t^i)]$$

$$\mathcal{E}(\phi) = \mathbb{E}|\dot{\boldsymbol{X}} - R_\phi(\boldsymbol{X})|^2 \neq \|\phi - \phi_{true}\|_{L^2(\rho)}^2$$

- Choice of $\mathcal{H}_n$: similar
  Function space: $L^2(\rho)$, exploration measure $\rho \sim |X^i - X^j|$
- Identifiability: unique minimizer $\arg\min_{\phi \in L^2_\rho} \mathcal{E}(\phi)$??
  $A \approx \left(\mathbb{E}[R_{\phi_i}(\boldsymbol{X})R_{\phi_j}(\boldsymbol{X})]\right)_{i,j}\ ?\geq?\, c_{\mathcal{H}} I_n \Leftarrow$ Coercivity condition $\downarrow$
- Convergence rate: $\checkmark$

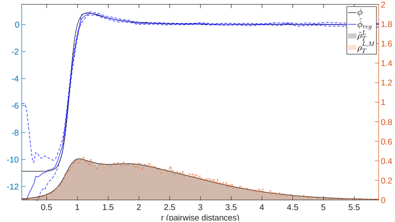
Theorem (Convergence with minimax rate [LZTM19,LMT21,LMT22])

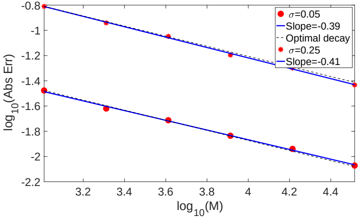
Let $\{\mathcal{H}_n\}$ compact convex in L^∞ with $\mathrm{dist}(\phi_{true}, \mathcal{H}_n) \sim n^{-s}$. Assume the coercivity condition on $\cup_n \mathcal{H}_n$. Set $n_* = (M/\log M)^{\frac{1}{2s+1}}$. Then

$$\mathbb{E}_{\mu_0}[\|\widehat{\phi}_{n_*,M} - \phi_{true}\|_{L^2_\rho}] \leq C \left(\frac{\log M}{M}\right)^{\frac{s}{2s+1}}.$$

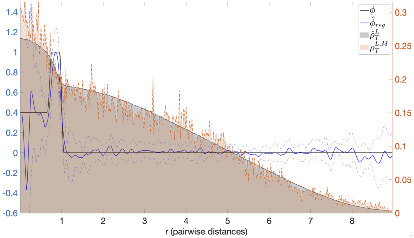
- $\dim(\mathcal{H}_n)$ adaptive to s ($\phi_{true} \in C^s$) and M
- Concentration inequalities for r.v. or martingale
- Ongoing: lower bound

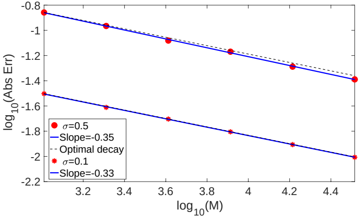
Lennard-Jones kernel estimators:





Opinion dynamics kernel estimators:





Coercivity condition on $\mathcal{H}$

$$\frac{1}{T}\int_0^T \mathbb{E}[R_\phi(\boldsymbol{X}_t)R_\phi(\boldsymbol{X}_t)]dt \geq c_{\mathcal{H}}\|\phi\|^2_{L^2_\rho}, \quad \forall \phi \in \mathcal{H}$$

- Partial results: $c_{\mathcal{H}} = \frac{1}{N-2}$ for $\mathcal{H} = L^2_\rho$
  - Gaussian or $\Phi(r) = r^{2\beta}$ stationary process [LLMTZ21spa,LL20]
  - Harmonic analysis: strictly positive definite integral kernel

$$\mathbb{E}[\phi(|X-Y|)\phi(|X-Z|)\frac{\langle X-Y, X-Z\rangle}{|X-Y||X-Z|}] \geq 0, \forall \phi \in L^2_\rho$$

- Open: non-stationary? A compact $\mathcal{H} \subset C(\mathrm{supp}(\rho))$?
- No coercivity on L^2_ρ when $N \to \infty$ since $c_{\mathcal{H}} \to 0$

Part 2: Infinitely many particles

Inverse problem for mean-field PDEs

Goal: Identify ϕ from discrete data $\{u(x_m, t_l)\}_{m,l=1}^{M,L}$ of

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

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

Loss functional

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

Candidates:

- Discrepancy: $\mathcal{E}(\phi) = \|\partial_t u - \nu \Delta u - \nabla.(u(K_\phi * u))\|^2$
  - discrete data $\to$ error in derivative approx.
  - denoising+smoothing [Kang+Liao etc22]
- Wasserstein-2: $\mathcal{E}(\phi) = W_2(u^\phi, u)$

  costly: requires many PDE simulations in optimization
- Weak SINDY [Bortz etc21,22]: parametric
- A probabilistic loss functional $\downarrow$

A probabilistic loss functional

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

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

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

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

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

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

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

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

Identifiability

$$\mathcal{E}(\phi) = \langle L_{\overline{G}}\phi, \phi\rangle_{L^2_\rho} - 2\langle \phi^D, \phi\rangle + const.$$

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

- Identifiability: $A^{-1}b \leftrightarrow L_{\overline{G}}^{-1}\phi^D$
  - $L_{\overline{G}}$: positive compact operator

- Coercivity condition on $\mathcal{H}$ (not L^2_ρ)

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

Convergence rate

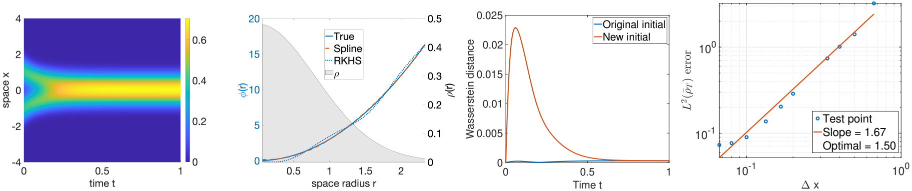
Theorem (Numerical error bound [Lang-Lu20])

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

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

- 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

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

Summary and future directions

Nonparametric/Variational learning of interaction kernels

- Finite N (ODEs/SDEs): statistical learning
- $N = \infty$ (Mean-field PDEs): inverse problem

Learning kernels in operators:

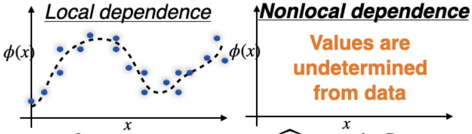
- Identifiability: a coercivity condition
- Algorithms with performance guarantees

Learning kernel in operators:

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

$$\partial_t u = \nu\Delta u + \nabla\cdot[u(K_\phi * u)] \qquad \Leftrightarrow R_\phi[u(\cdot,t)] = f(\cdot,t)$$

	Classical learning	**Learning kernel**
	$\{(x_i, \phi(x_i) + \epsilon_i)\}$	$\{(u_k, R_\phi[u_k] + \eta_k)\}$



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

- Coercivity condition (with it ✓ without it $\Downarrow$)
- Space-aware Regularization
- Convergence (minimax rate)