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
What is the law of interaction?

\[ \dot{X}_i^t = \frac{1}{N} \sum_{j=1, j \neq i}^{N} m_j K_\phi (X_j^t - X_i^t), \]

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

- 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 $\phi$

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

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

**Large N ($\gg 1$)**
- 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

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

- Data: $M$ trajectories of particles $\{X_{t_1:t_L}^{(m)}\}_{m=1}^{M}$
- Goal: estimate $\phi$
Finitely many particles

\[ R_{\phi}(X_t) = \dot{X}_t - \sqrt{2 \nu} \dot{B}_t \text{ & Data } \{ X^{(m)}_{t_i:t} \}_{m=1}^M \]

- Loss function (or log-likelihood for SDEs):
  \[
  \hat{\phi}_{n,M} = \arg\min_{\phi \in \mathcal{H}_n} \mathcal{E}_M(\phi) = \frac{1}{M} \sum_{m=1}^M \int_0^T |\dot{X}_m^t - R_\phi(X_m^t)|^2 dt
  \]

- Nonparametric Regression: \( \mathcal{H}_n = \text{span}\{\phi_i\}_{i=1}^n \), \( \phi = \sum_i c_i \phi_i \)
  \[
  \mathcal{E}_M(\phi) = c^T A c - 2 b^T c \quad \Rightarrow \quad \hat{\phi}_{n,M} = \sum_{1 \leq i \leq n} \hat{c}_i \phi_i, \quad \hat{c} = A^{-1} b
  \]
Finitely many particles

\[ R_\phi(X_t) = \dot{X}_t - \sqrt{2\nu}B_t \quad \& \text{Data } \{X_{t_i}^{(m)}\}_{m=1}^{M} \]

- Loss function (or log-likelihood for SDEs):
  \[ \hat{\phi}_{n,M} = \arg\min_{\phi \in \mathcal{H}_n} \mathcal{E}_M(\phi) = \frac{1}{M} \sum_{m=1}^{M} \int_{0}^{T} |\dot{X}_t^{m} - R_\phi(X_t^{m})|^2 dt \]

- Nonparametric Regression: \( \mathcal{H}_n = \text{span}\{\phi_i\}_{i=1}^{n}, \phi = \sum_i c_i \phi_i \)
  \[ \mathcal{E}_M(\phi) = c^\top Ac - 2b^\top c \quad \Rightarrow \quad \hat{\phi}_{n,M} = \sum_{1 \leq i \leq n} \hat{c}_i \phi_i, \quad \hat{c} = A^{-1}b \]

- Choice of \( \mathcal{H}_n \) ? function space?
- Identifiability/Well-posedness?
- Convergence and rate?
Classical learning in a nutshell

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 \( \psi = \sum_i c_i \phi_i \in \mathcal{H}_n = \text{span}\{ \phi_i \}_{i=1}^{n} \):
  \[ \mathcal{E}_M(\psi) = c^\top A c - 2 b^\top c \Rightarrow \hat{\phi}_{n,M} = \sum_{1 \leq i \leq n} \hat{c}_i \phi_i, \quad \hat{c} = A^{-1} b \]

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

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

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

Goal: Estimate \(\phi\) s.t. \(\dot{X}_t \approx R_\phi(X_t) = \left[\frac{1}{N} \sum_{j=1}^N K_\phi(X^j_t, X^i_t)\right]\)

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

- Choice of \(\mathcal{H}_n\): similar
  Function space: \(L^2(\rho)\), exploration measure \(\rho \sim |X^i - X^j|\)

- Identifiability: unique minimizer \(\arg\min_{\phi \in L^2_\rho} \mathcal{E}(\phi)??\)

\[A \approx \left(\mathbb{E}[R_{\phi_i}(X)R_{\phi_j}(X)]\right)_{i,j} \geq c_{\mathcal{H}} I_n \Leftarrow \text{Coercivity condition} \downarrow\]

- Convergence rate: \(\checkmark\)
Theorem (Convergence with minimax rate \[LZTM19,LMT21,LMT22\])

Let \( \{H_n\} \) compact convex in \( L^\infty \) with \( \text{dist}(\phi_{true}, H_n) \sim n^{-s} \). Assume the coercivity condition on \( \bigcup_n H_n \). Set \( n^* = (M/\log M)^{1/(2s+1)} \). Then

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

- \( \dim(H_n) \) 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 $\mathcal{H}$

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

- 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 $\mathcal{H} \subset C(\text{supp}(\rho))$?
- No coercivity on $L^2_\rho$ when $N \to \infty$ since $c_\mathcal{H} \to 0$
Part 2: Infinitely many particles

Inverse problem for mean-field PDEs

Goal: Identify $\phi$ 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

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

Candidates:

- **Discrepancy**: \( \mathcal{E}(\phi) = \| \partial_t u - \nu \Delta u - \nabla \cdot (u(K_\phi \ast u)) \|^2 \)
  - discrete data \( \rightarrow \) 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** ↓
A probabilistic loss functional

\[ \mathcal{E}(\phi) := \frac{1}{T} \int_{0}^{T} \int_{\mathbb{R}^d} \left[ |K_\phi \ast u|^2 u - 2\nu u(\nabla \cdot K_\phi \ast u) + 2\partial_t u(\Phi \ast u) \right] dx \, dt \]

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

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

- 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 \( \phi = \sum_{i=1}^{n} c_i \phi_i \in \mathcal{H}_n \):

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

- Choice of \( \mathcal{H}_n \) & function space of learning?
  - Exploration measure \( \rho \leftarrow |X_t - X'_t| \)
- 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

\[ \mathcal{E}(\phi) = \langle L_G^{-1} \phi, \phi \rangle_{L_2^2} - 2 \langle \phi^D, \phi \rangle + \text{const.} \]

\[ \nabla \mathcal{E}(\phi) = L_G \phi - \phi^D = 0 \quad \Rightarrow \quad \hat{\phi} = L_G^{-1} \phi^D \]

- **Identifiability:** \( A^{-1} b \leftrightarrow L_G^{-1} \phi^D \)
  - \( L_G^{-1} \): positive compact operator

- Coercivity condition on \( \mathcal{H} \) (not \( L_2^2 \rho \))

\[ c_{\mathcal{H}} = \inf_{\phi \in \mathcal{H}, \|\phi\|_{L_2^2} = 1} \langle L_G^{-1} \phi, \phi \rangle > 0 \]
Convergence rate

**Theorem (Numerical error bound [Lang-Lu20])**

Let $\mathcal{H}_n = \text{span}\{\phi_i\}_{i=1}^n$ 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:

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

- $\Delta x^\alpha$ 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 $\phi$ discontinuous,
  - low rate when $\phi$ 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:

\[ dX_t^i = \frac{1}{N} \sum_{j=1}^{N} K_\phi(X_t^j, X_t^i)dt + \sqrt{2\nu} dB_t^i \]

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

Classical learning
\[ \{(x_i, \phi(x_i) + \epsilon_i)\} \]
Nonlocal dependence
Values are undetermined from data

Learning kernel
\[ \{(u_k, R_\phi[u_k] + \eta_k)\} \]

Inversion
\[ \hat{\phi} = I^{-1}\phi^D \]
\[ \hat{\phi} = (I + \lambda Q)^{-1}\phi^D \]
\[ \hat{\phi} = L_G^{-1}\phi^D \]
\[ \hat{\phi} = (L_G + \lambda L_G^{-1})^{-1}\phi^D \]

- Coercivity condition (with it ✔ without it ☐)
- Space-aware Regularization
- Convergence (minimax rate)