Introduction to Nonparametric Learning of Kernels in Operators

Fei Lu

Department of Mathematics, Johns Hopkins University

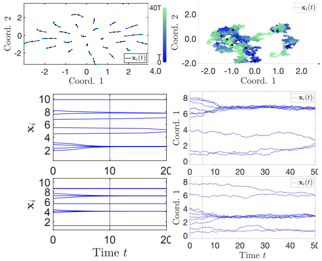
Plan:
Lecture 1. Overview and a review of classical learning theory
Lecture 2. Learning interaction kernels in interacting particle systems
Lecture 3. Coercivity condition and minimax rate of convergence
Lecture 4. Learning interaction kernels in mean-field equations
Lecture 5. Data adaptive RKHS Tikhonov regularization
Lecture 6. Small noise analysis of RKHS regularizations

Lecture 2. Learning kernels in interacting particle systems

Learning interaction kernel $K_{\phi}(x-y) = \phi(|x-y|)\frac{x-y}{|x-y|}$

$$dX_t^i = \frac{1}{N}\sum_{j=1}^{N} K_{\phi}(X_t^j - X_t^i)dt + \sqrt{2\nu}dB_t^i, \quad 1 \le i \le N, X_t^i \in \mathbb{R}^d$$

$$\Leftrightarrow \dot{\mathbf{X}}_t = R_{\phi}(\mathbf{X}_t) + \sqrt{2\nu}\dot{\mathbf{B}}_t, \quad \mathbf{X}_t \in \mathbb{R}^{N\times d}$$



Finite N:

- Data: M trajectories of particles $\{\mathbf{X}_{t_1:t_L}^{(m)}\}_{m=1}^{M}$
- ODEs/SDEs: Opinion Dynamics, Lennard-Jones, Prey-Predator; 1st/2nd order
- Statistical learning

Goal: algorithm, identifiability, convergence

1. Review of learning theory in Lecture 1
2. Computation: loss function and regression
3. Main results: theory and numerical tests
4. The coercivity condition (with open questions)

Outline

1. Review of learning theory in Lecture 1
2. Computation: loss function and regression
3. Main results: theory and numerical tests
4. The coercivity condition (with open questions)

Review of Lecture 1:
Theory for learning kernels in operators

Variational approach: $\mathcal{H}_n := \mathrm{span}\{\phi_i\}_{i=1}^n$, $\phi = \sum_{i=1}^n c_i\phi_i$,

$$\widehat{\phi}_{n,M} = \arg\min_{\phi\in\mathcal{H}_n} \mathcal{E}_M(\phi)$$

$$\mathcal{E}_M(\phi) = \frac{1}{TM}\sum_{m=1}^M \int_0^T \left|\dot{\mathbf{X}}_t^{(m)} - R_\phi(\mathbf{X}_t^{(m)})\right|^2 dt = c^\top A_{n,M}c - 2c^\top b_{n,M} + Const.$$

$$\nabla\mathcal{E}_M = 0 \Rightarrow \quad \widehat{c} = A_{n,M}^{-1} b_{n,M} \Rightarrow \quad \widehat{\phi}_{n,M} = \sum_i \widehat{c}_i\phi_i$$

- How to choose $\mathcal{H}_n := \mathrm{span}\{\phi_i\}_{i=1}^n$?
- $A_{n,M}^{-1}$ exists? $A_{n,M}^{-1} b_{n,M}$ stable?
- Convergence of $\phi_{n,M}$?
- Identifiability of ϕ_{true}? ...

New elements:

Exploration measure: $\rho_T \sim \{|X_t^i - X_t^j|\}_{i,j,t}^{(m)}$

$$\langle\langle\phi,\phi\rangle\rangle := \frac{1}{T}\int_0^T \mathbb{E}[\langle R_\phi(\mathbf{X}_t), R_\phi(\mathbf{X}_t)\rangle]dt$$

Coercivity condition: $\langle\langle\phi,\phi\rangle\rangle \geq c_{\mathcal{H}}\|\phi\|_{L^2_{\rho_T}}^2$

From the loss function, find its minimizer in $\mathcal{H}$

Exploration measure ρ for the variable of ϕ

- Empirical distribution $\rho_M \xrightarrow{M\to\infty} \rho$ (LLN)
- Intrinsic to the dynamics: initial distribution $\mu_0^{\otimes N}$ and kernel
- Function space L^2_ρ
  - $\mathcal{H} =$ piecewise polynomials $\subset L^2_\rho$
  - singular kernels $\subset L^2_\rho$

Coercivity condition $\langle\langle \phi, \phi \rangle\rangle \geq c_{\mathcal{H}} \|\phi\|^2_{L^2_{\rho_T}}$, $\forall \phi \in \mathcal{H}$

- The loss function is uniformly convex
- Ensures identifiability and error bounds
- Ensures $(A_{n,M}(i,j)) \approx (\langle\langle \phi_i, \phi_j \rangle\rangle)$ well-conditioned w.h.p. in regression

Learning theory: classical v.s. new

Classical learning theory

Given: Data $\{(X_m, Y_m)\}_{m=1}^M \sim (X, Y)$

Goal: find f s.t. $Y = f(X)$

$$\mathcal{E}(f) = \mathbb{E}[|Y - f(X)|^2] = \|f - f_{true}\|^2_{L^2(\rho_X)}$$

$$\approx \mathcal{E}_M(f) = \frac{1}{M}\sum_{m=1}^M |Y_m - f(X_m)|^2$$

Learning kernel

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

Goal: find ϕ s.t. $\dot{\mathbf{X}}_t = R_\phi(\mathbf{X}_t)$

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

$$\approx \mathcal{E}_M(\phi) = \frac{1}{M}\sum_{m=1}^M |\dot{\mathbf{X}}^{(m)} - R_\phi(\mathbf{X}^{(m)})|^2$$

Learning theory: classical v.s. new

Classical learning theory

Given: Data$\{(X_m, Y_m)\}_{m=1}^M \sim (X, Y)$

Goal: find f s.t. $Y = f(X)$

$$\mathcal{E}(f) = \mathbb{E}[|Y - f(X)|^2] = \|f - f_{true}\|^2_{L^2(\rho_X)}$$

$$\approx \mathcal{E}_M(f) = \frac{1}{M}\sum_{m=1}^M |Y_m - f(X_m)|^2$$

- Function space: $L^2(\rho_X)$.
- Identifiability: $\mathbb{E}[Y|X=x] = \underset{f\in L^2(\rho_X)}{\arg\min}\, \mathcal{E}(f)$.

Learning kernel

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

Goal: find ϕ s.t. $\dot{\mathbf{X}}_t = R_\phi(\mathbf{X}_t)$

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

$$\approx \mathcal{E}_M(\phi) = \frac{1}{M}\sum_{m=1}^M |\dot{\mathbf{X}}^{(m)} - R_\phi(\mathbf{X}^{(m)})|^2$$

1 Review of learning theory in Lecture 1

Learning theory: classical v.s. new

Classical learning theory

Given: Data$\{(X_m, Y_m)\}_{m=1}^M \sim (X, Y)$

Goal: find f s.t. $Y = f(X)$

$$\mathcal{E}(f) = \mathbb{E}[|Y - f(X)|^2] = \|f - f_{true}\|^2_{L^2(\rho_X)}$$

$$\approx \mathcal{E}_M(f) = \frac{1}{M}\sum_{m=1}^M |Y_m - f(X_m)|^2$$

- Function space: $L^2(\rho_X)$.
- Identifiability: $\mathbb{E}[Y|X = x] = \arg\min_{f \in L^2(\rho_X)} \mathcal{E}(f)$.
- $A_{n,M} \approx \mathbb{E}[\phi_i(X)\phi_j(X)] = I_n$ by setting $\{\phi_i\}$ ONB in $L^2(\rho_X)$.
- **Error bounds for** $\widehat{f}_{n_M}$ (asymptotic/non-asymptotic)

Learning kernel

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

Goal: find ϕ s.t. $\dot{\mathbf{X}}_t = R_\phi(\mathbf{X}_t)$

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

$$\approx \mathcal{E}_M(\phi) = \frac{1}{M}\sum_{m=1}^M |\dot{\mathbf{X}}^{(m)} - R_\phi(\mathbf{X}^{(m)})|^2$$

- Function space: $L^2(\rho)$.
- Identifiability: $\arg\min_{\phi \in L^2_\rho} \mathcal{E}(\phi)$??

1 Review of learning theory in Lecture 1

Learning theory: classical v.s. new

Classical learning theory

Given: Data $\{(X_m, Y_m)\}_{m=1}^M \sim (X, Y)$

Goal: find f s.t. $Y = f(X)$

$$\mathcal{E}(f) = \mathbb{E}[|Y - f(X)|^2] = \|f - f_{true}\|^2_{L^2(\rho_X)}$$

$$\approx \mathcal{E}_M(f) = \frac{1}{M}\sum_{m=1}^M |Y_m - f(X_m)|^2$$

- Function space: $L^2(\rho_X)$.
- Identifiability: $\mathbb{E}[Y|X = x] = \arg\min_{f \in L^2(\rho_X)} \mathcal{E}(f)$.
- $A_{n,M} \approx \mathbb{E}[\phi_i(X)\phi_j(X)] = I_n$ by setting $\{\phi_i\}$ ONB in $L^2(\rho_X)$.
- **Error bounds for $\widehat{f}_{n_M}$** (asymptotic/non-asymptotic)

Learning kernel

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

Goal: find ϕ s.t. $\dot{\mathbf{X}}_t = R_\phi(\mathbf{X}_t)$

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

$$\approx \mathcal{E}_M(\phi) = \frac{1}{M}\sum_{m=1}^M |\dot{\mathbf{X}}^{(m)} - R_\phi(\mathbf{X}^{(m)})|^2$$

- Function space: $L^2(\rho)$.
- Identifiability: $\arg\min_{\phi \in L^2_\rho} \mathcal{E}(\phi)$??
- $A_{n,M} \approx \mathbb{E}[R_{\phi_i}(\mathbf{X})R_{\phi_j}(\mathbf{X})] \approx I_n$ by setting $\{\phi_i\}$ ONB in L^2_ρ??.

1 Review of learning theory in Lecture 1

Learning theory: classical v.s. new

Classical learning theory

Given: Data $\{(X_m, Y_m)\}_{m=1}^M \sim (X, Y)$

Goal: find f s.t. $Y = f(X)$

$$\mathcal{E}(f) = \mathbb{E}[|Y - f(X)|^2] = \|f - f_{true}\|^2_{L^2(\rho_X)}$$

$$\approx \mathcal{E}_M(f) = \frac{1}{M}\sum_{m=1}^M |Y_m - f(X_m)|^2$$

- Function space: $L^2(\rho_X)$.
- Identifiability: $\mathbb{E}[Y|X = x] = \underset{f \in L^2(\rho_X)}{\arg\min}\ \mathcal{E}(f)$.
- $A_{n,M} \approx \mathbb{E}[\phi_i(X)\phi_j(X)] = I_n$ by setting $\{\phi_i\}$ ONB in $L^2(\rho_X)$.
- **Error bounds for** $\widehat{f}_{n_M}$ (asymptotic/non-asymptotic)

Learning kernel

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

Goal: find ϕ s.t. $\dot{\mathbf{X}}_t = R_\phi(\mathbf{X}_t)$

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

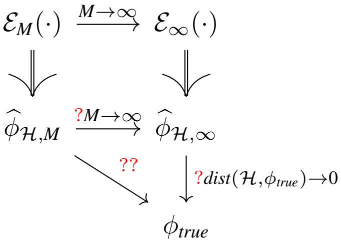
$$\approx \mathcal{E}_M(\phi) = \frac{1}{M}\sum_{m=1}^M |\dot{\mathbf{X}}^{(m)} - R_\phi(\mathbf{X}^{(m)})|^2$$

- Function space: $L^2(\rho)$.
- Identifiability: $\underset{\phi \in L^2_\rho}{\arg\min}\ \mathcal{E}(\phi)$??
- $A_{n,M} \approx \mathbb{E}[R_{\phi_i}(\mathbf{X})R_{\phi_j}(\mathbf{X})] \approx I_n$ by setting $\{\phi_i\}$ ONB in L^2_ρ??.
- **Error bounds for** $\widehat{\phi}_{n_M}$ **?**

1 Review of learning theory in Lecture 1

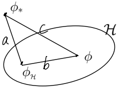
Convergence/Error bounds

- For $\mathcal{E}_M \to \mathcal{E}_\infty$; LLN/CLT, concentration inequalities; (Uniform in $\mathcal{H}$)
- Pass them to estimator:



$$\mathcal{E}_\infty(\phi) - \mathcal{E}_\infty(\phi_{\mathcal{H}}) \geq c_{\mathcal{H}}\|\phi - \phi_{\mathcal{H}}\|^2$$

$c^2 - a^2 > b^2$



$$\langle\langle \phi, \psi \rangle\rangle := \frac{1}{T}\int_0^T \mathbb{E}[\langle R_\phi(\mathbf{X}_t), R_\psi(\mathbf{X}_t)\rangle]dt$$

$$\mathcal{E}_\infty(\phi) = \frac{1}{T}\int_0^T \mathbb{E}[|\dot{\mathbf{X}}_t - R_\phi(\mathbf{X}_t)|^2]dt = \langle\langle \phi - \phi_{true}, \phi - \phi_{true} \rangle\rangle$$

Coercivity condition: $\langle\langle \phi, \phi \rangle\rangle \geq c_{\mathcal{H}}\|\phi\|^2, \quad \forall \phi \in \mathcal{H}.$

Outline

Computation: loss function and regression

Variational approach: $\mathcal{H}_n := \mathrm{span}\{\phi_i\}_{i=1}^n$, $\phi = \sum_{i=1}^n c_i\phi_i$,

$$\widehat{\phi}_{n,M} = \underset{\phi\in\mathcal{H}_n}{\arg\min}\, \mathcal{E}_M(\phi) = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{T}\int_0^T \left|\dot{\mathbf{X}}_t^{(m)} - R_{\phi}(\mathbf{X}_t^{(m)})\right|^2 dt = c^\top A_{n,M} c - 2c^\top b_{n,M} + Const.$$

$$\nabla\mathcal{E}_M = 0 \Rightarrow \quad \widehat{c} = A_{n,M}^{-1} b_{n,M} \Rightarrow \quad \widehat{\phi}_{n,M} = \sum_i \widehat{c}_i\phi_i$$

Computation: loss function and regression

Variational approach: $\mathcal{H}_n := \mathrm{span}\{\phi_i\}_{i=1}^n$, $\phi = \sum_{i=1}^n c_i\phi_i$,

$$\widehat{\phi}_{n,M} = \underset{\phi\in\mathcal{H}_n}{\arg\min}\ \mathcal{E}_M(\phi) = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{T}\int_0^T \left|\dot{\mathbf{X}}_t^{(m)} - R_\phi(\mathbf{X}_t^{(m)})\right|^2 dt = c^\top A_{n,M} c - 2c^\top b_{n,M} + Const.$$

$$\nabla\mathcal{E}_M = 0 \Rightarrow \quad \widehat{c} = A_{n,M}^{-1} b_{n,M} \Rightarrow \quad \widehat{\phi}_{n,M} = \sum_i \widehat{c}_i\phi_i$$

- Loss function: key in a variational/learning approach
- Regression: (where fundamental questions arise)
  - How to choose $\mathcal{H}_n := \mathrm{span}\{\phi_i\}_{i=1}^n$?
  - $A_{n,M}^{-1}$ exists? $A_{n,M}^{-1}b_{n,M}$ stable? LSE, regularization?
  - Exploration measure ρ
  - Function space
- Convergence of $\phi_{n,M}$ as M increases?

Loss function

Given data $\{\mathbf{X}^{(m)}_{t_0:t_L}\}_{m=1}^{M}$, to recover ϕ in

$$dX^i_t = \frac{1}{N}\sum_{j=1}^{N} K_\phi(X^j_t - X^i_t)dt + \sqrt{2\nu}dB^i_t, \quad 1 \le i \le N, X^i_t \in \mathbb{R}^d$$

$$\Leftrightarrow d\mathbf{X}_t = R_\phi(\mathbf{X}_t)dt + \sqrt{2\nu}d\mathbf{B}_t, \quad \mathbf{X}_t \in \mathbb{R}^{N\times d}$$

Loss function $\mathcal{E}_M(\phi) = \frac{1}{M}\sum_{m=1}^{M} \mathcal{E}(\phi, \mathbf{X}^{(m)}_{[0,T]})$

- Deterministic ($\nu = 0$): $\mathcal{E}(\phi, \mathbf{X}_{[0,T]}) = \frac{1}{T}\int_0^T \|\dot{\mathbf{X}}_t - R_\phi(\mathbf{X}_t)\|^2 dt$

Loss function

Given data $\{\mathbf{X}^{(m)}_{t_0:t_L}\}_{m=1}^{M}$, to recover ϕ in

$$dX_t^i = \frac{1}{N}\sum_{j=1}^{N} K_\phi(X_t^j - X_t^i)dt + \sqrt{2\nu}dB_t^i, \quad 1 \le i \le N, X_t^i \in \mathbb{R}^d$$

$$\Leftrightarrow d\mathbf{X}_t = R_\phi(\mathbf{X}_t)dt + \sqrt{2\nu}d\mathbf{B}_t, \quad \mathbf{X}_t \in \mathbb{R}^{N\times d}$$

Loss function $\mathcal{E}_M(\phi) = \frac{1}{M}\sum_{m=1}^{M}\mathcal{E}(\phi, \mathbf{X}^{(m)}_{[0,T]})$

- Deterministic ($\nu = 0$): $\mathcal{E}(\phi, \mathbf{X}_{[0,T]}) = \frac{1}{T}\int_0^T \|\dot{\mathbf{X}}_t - R_\phi(\mathbf{X}_t)\|^2 dt$
- Stochastic: $\mathcal{E}(\phi, \mathbf{X}_{[0,T]}) = \frac{1}{T}\int_0^T -2\langle d\mathbf{X}_t, R_\phi(\mathbf{X}_t)\rangle_{\mathbb{R}^{N\times d}} + \|R_\phi(\mathbf{X}_t)\|^2 dt$
  - $-$ log-likelihood ratio of the path $\mathbf{X}_{[0,T]}$
  - Discrete-approximation with $d\mathbf{X}_{t_l} \approx \mathbf{X}_{t_{l+1}} - \mathbf{X}_{t_l}$
    $=$ the $-$log-likelihood of Euler-Maruyama
    $\mathbf{X}_{t_{l+1}} - \mathbf{X}_{t_l} \approx R_\phi(\mathbf{X}_{t_l})\Delta t + \sqrt{2\nu\Delta t}\mathbf{W}_l$
- $\mathcal{E}_M(\phi)$ is quadratic in $\phi \longrightarrow$ Regression

Nonparametric Regression

Quadratic loss: $\mathcal{H}_n := \mathrm{span}\{\phi_i\}_{i=1}^n$, $\phi = \sum_{i=1}^n c_i\phi_i$,

$$\widehat{\phi}_{n,M} = \arg\min_{\phi\in\mathcal{H}_n} \mathcal{E}_M(\phi), \quad \nabla\mathcal{E}_M = 0 \Rightarrow \boxed{\widehat{c} = A_{n,M}^{-1} b_{n,M}} \Rightarrow \quad \widehat{\phi}_{n,M} = \sum_i \widehat{c}_i \phi_i$$

$$\mathcal{E}_M(\phi) = \frac{1}{M}\sum_{m=1}^M \frac{1}{T}\int_0^T \left\| \dot{\mathbf{X}}_t^{(m)} - R_\phi(\mathbf{X}_t^{(m)}) \right\|^2 dt = c^\top A_{n,M} c - 2c^\top b_{n,M} + Const.,$$

$$A_{n,M}(i,j) = \frac{1}{MT}\sum_{m=1}^M \int_0^T \langle R_{\phi_i}(\mathbf{X}_t^{(m)}), R_{\phi_j}(\mathbf{X}_t^{(m)})\rangle dt$$

$$b_{n,M}(i) = \frac{1}{MT}\sum_{m=1}^M \int_0^T \langle R_{\phi_i}(\mathbf{X}_t^{(m)}), d\mathbf{X}_t^{(m)})\rangle$$

- How to choose $\mathcal{H}_n := \mathrm{span}\{\phi_i\}_{i=1}^n$?
- $A_{n,M}^{-1}$ exists? $A_{n,M}^{-1} b_{n,M}$ stable? LSE, regularization?

Nonparametric Regression

Quadratic loss: $\mathcal{H}_n := \mathrm{span}\{\phi_i\}_{i=1}^n$, $\phi = \sum_{i=1}^n c_i \phi_i$,

$$\widehat{\phi}_{n,M} = \arg\min_{\phi \in \mathcal{H}_n} \mathcal{E}_M(\phi), \quad \nabla \mathcal{E}_M = 0 \Rightarrow \boxed{\widehat{c} = A_{n,M}^{-1} b_{n,M}} \Rightarrow \quad \widehat{\phi}_{n,M} = \sum_i \widehat{c}_i \phi_i$$

$$\mathcal{E}_M(\phi) = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{T} \int_0^T \left\| \dot{\mathbf{X}}_t^{(m)} - R_\phi(\mathbf{X}_t^{(m)}) \right\|^2 dt = c^\top A_{n,M} c - 2c^\top b_{n,M} + Const.,$$

$$A_{n,M}(i,j) = \frac{1}{MT} \sum_{m=1}^{M} \int_0^T \langle R_{\phi_i}(\mathbf{X}_t^{(m)}), R_{\phi_j}(\mathbf{X}_t^{(m)}) \rangle dt$$

$$b_{n,M}(i) = \frac{1}{MT} \sum_{m=1}^{M} \int_0^T \langle R_{\phi_i}(\mathbf{X}_t^{(m)}), d\mathbf{X}_t^{(m)} \rangle$$

- How to choose $\mathcal{H}_n := \mathrm{span}\{\phi_i\}_{i=1}^n$?
- $A_{n,M}^{-1}$ exists? $A_{n,M}^{-1} b_{n,M}$ stable? LSE, regularization?
  - Exploration measure ρ: ϕ_i supported in $\mathrm{supp}(\rho)$ (ONB in L^2_ρ)
  - Local basis (B-spline), global basis (polynomials, RKHS)
  - LSE OK for singular $A_{n,M}$ (NO need of regularization if CC)

Outline

1. Review of learning theory in Lecture 1
2. Computation: loss function and regression
3. Main results: theory and numerical tests
4. The coercivity condition (with open questions)

Results for deterministic systems

Consistency of estimator

Theorem ([LZTM19])

Assume the coercivity condition. Let $\{\mathcal{H}_n\}$ be a sequence of ↑ compact convex subsets of $C([0,R])$ such that $\inf_{\psi\in\mathcal{H}_n} \|\psi - \phi_{true}\|_\infty \to 0$ as $n \to \infty$. Then

$$\lim_{n\to\infty} \lim_{M\to\infty} \|\widehat{\phi}_{M,\mathcal{H}_n} - \phi_{true}\|_{L^2_{\rho_T}} = 0, \text{ almost surely.}$$

[LZTM19]: L., M Zhong, S Tang, and M Maggioni. Nonparametric inference of interaction laws in systems of agents from trajectory data. Proc. Natl. Acad. Sci. USA. 116 (29) 14424–14433. 2019

Rate of convergence: upper bound

Theorem ([LZTM19])

Let $\{\mathcal{H}_n\}$ be a seq. of compact convex subspaces of $C[0,R]$ s.t.

$$\dim(\mathcal{H}_n) \leq c_0 n, \text{ and } \inf_{\psi \in \mathcal{H}_n} \|\psi - \phi_{true}\|_\infty \leq c_1 n^{-s}.$$

Assume the coercivity condition. Choose $n_* = (M/\log M)^{\frac{1}{2s+1}}$: then

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

- The 2nd condition is about regularity: $\phi \in C^s$
- Choice of $\dim(\mathcal{H}_n)$: adaptive to s and M

Prediction

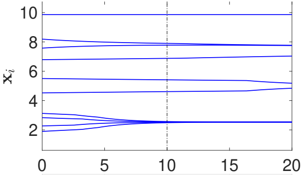
Theorem ([LZTM19])

Denote by $\widehat{\mathbf{X}}(t)$ and $\mathbf{X}(t)$ the solutions of the systems with kernels $\widehat{\phi}$ and ϕ_{true} respectively, starting from the same initial conditions that are drawn i.i.d from μ_0. Then we have

$$\mathbb{E}_{\mu_0}[\sup_{t\in[0,T]} \|\widehat{\mathbf{X}}(t) - \mathbf{X}(t)\|^2] \lesssim e^{CT}\sqrt{N}\|\widehat{\phi} - \phi_{true}\|^2_{L^2_{\rho_T}},$$

- Follows from Grownwall's inequality

Example: Opinion Dynamics



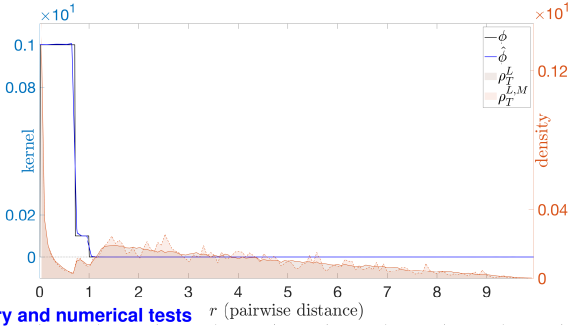
$N = 10, \mathbf{x}_i \in \mathbb{R}.$

$M = 250, \mu_0 = Unif[0, 10]^{10}$

$\mathcal{T} = [0, 10], 200$ discrete instances

$\mathcal{H} =$ piecewise constant functions

The estimated kernels:



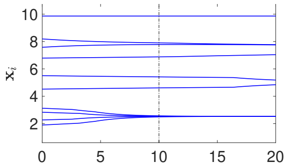
Example: Opinion Dynamics

$N = 10, \mathbf{x}_i \in \mathbb{R}.$

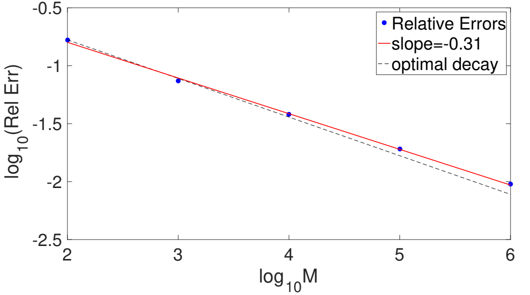
$M = 250, \mu_0 = Unif[0, 10]^{10}$

$\mathcal{T} = [0, 10]$, 200 discrete instances

$\mathcal{H} =$ piecewise constant functions



The rate of convergence:



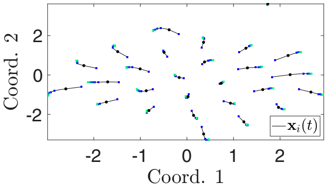
Examples: Lennard-Jones Dynamics

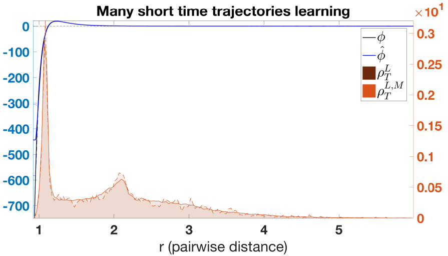
The Lennard-Jones potential

$$V_{LJ}(r) = 4\epsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right) \Rightarrow \quad \phi(r)r = V'_{LJ}(r)$$

$$\dot{x}_i(t) = \frac{1}{N}\sum_{j=1, j\neq i}^{N} \phi(|x_i - x_j|)(x_j - x_i)$$

- piecewise linear estimator; Gaussian initial conditions.

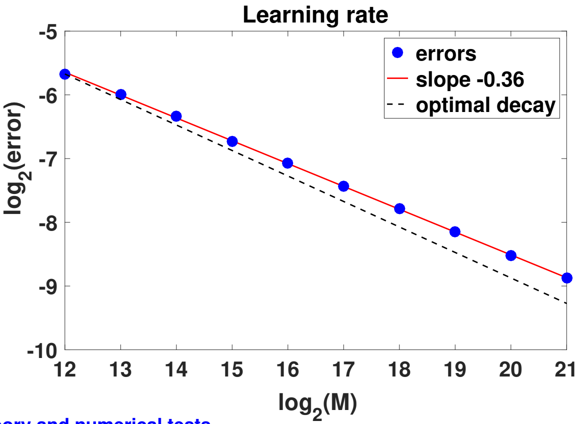




Optimal rate

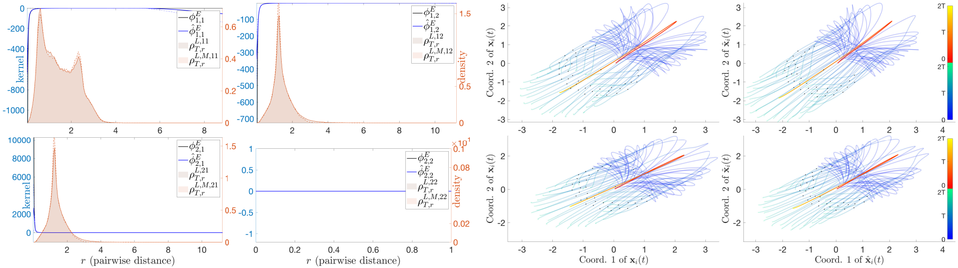
$$V_{LJ}(r) = 4\epsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right) \Rightarrow \quad \phi(r)r = V'_{LJ}(r)$$

- V_{LJ} is highly singular, yet we get close to optimal rate (-0.4).



Example: 2nd-order Prey-Predator system

$$\begin{cases} \ddot{\boldsymbol{X}}^m = \mathcal{F}^{\boldsymbol{v}}(\dot{\boldsymbol{X}}^m, \Xi^m) + \mathbf{f}_{\boldsymbol{\phi}^E}(\boldsymbol{X}^m) + \mathbf{f}_{\boldsymbol{\phi}^A}(\boldsymbol{X}^m, \dot{\boldsymbol{X}}^m) \\ \dot{\Xi}^m = \mathcal{F}^{\xi}(\Xi^m) + \mathbf{f}_{\boldsymbol{\phi}^\xi}(\boldsymbol{X}^m, \Xi^m), \end{cases}$$

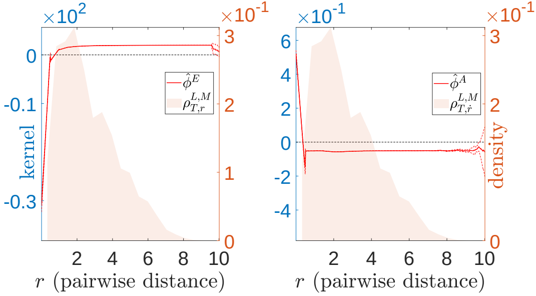


Example: model selection

- **Order selection**

	Learned as 1^{st} order	Learned as 2^{nd} order
1^{st} order system	$\mathbf{0.01 \pm 0.002}$	1.6 ± 1.1
2^{nd} order system	1.7 ± 0.3	$\mathbf{0.2 \pm 0.06}$

- **Interaction type selection**



Stochastic systems: similar results

Theorem ([LMT21foc])

For any ϵ, with a high probability ($> 1-\delta$, $M \geq M_{\delta,\epsilon}$)

$$\|\widehat{\phi}_{L,T,M,\mathcal{H}} - \phi\|^2_{L^2_\rho} \leq \|\widehat{\phi}_{T,\infty,\mathcal{H}} - \phi\|^2_{L^2_\rho} + C\left(\epsilon n/M + \Delta t\right),$$

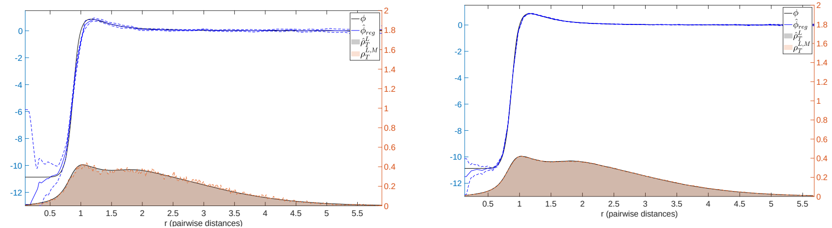
With a high probability and in expectation:

$$\|\widehat{\phi}_{L,T,M,\mathcal{H}} - \phi\|^2_{L^2_\rho} \lesssim c_{\mathcal{H}}^{-2}\left(\frac{\log M}{M}\right)^{\frac{2s}{2s+1}}$$

- Discrete data: Euler-Maruyama ≈ the likelihood ratio
- Concentration for Martingales/unbounded r.v.
- Two types of arguments
  - Learning theory (applicable for generic regression)
  - Regression $(A_{n,M}, b_{n,M})$

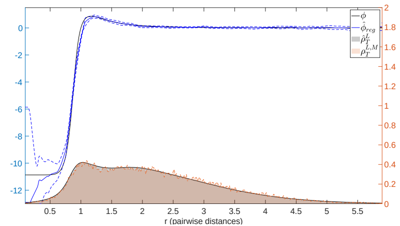
Lennard-Jones kernel estimators:

M=32 M=1024

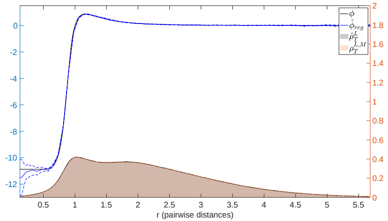


Lennard-Jones kernel estimators:

M=32

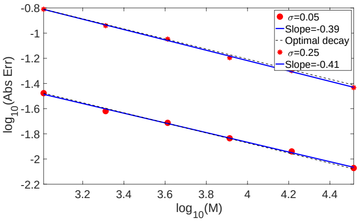


M=1024

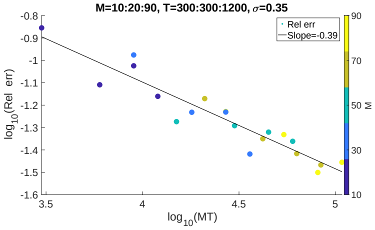


Close to optimal rates of convergence:

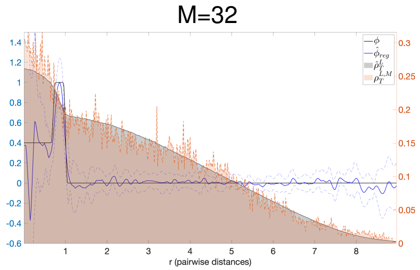
Rate in M

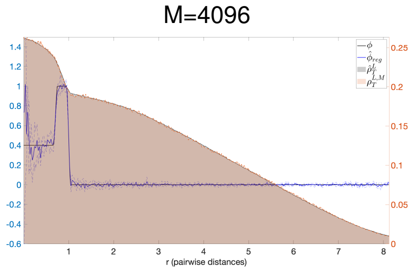


Rate in $M \times T$



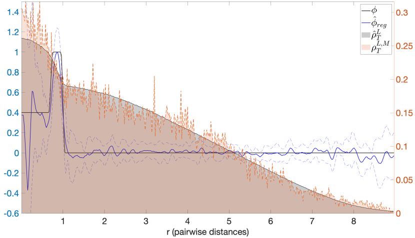
Opinion dynamics kernel estimators:



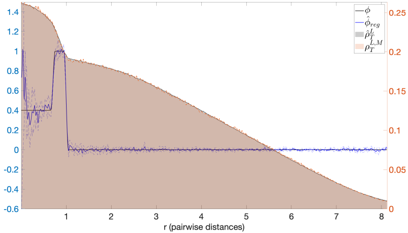


Opinion dynamics kernel estimators:

M=32

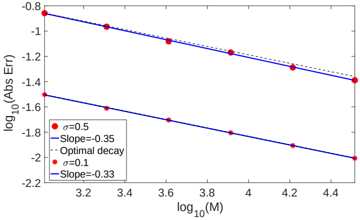


M=4096

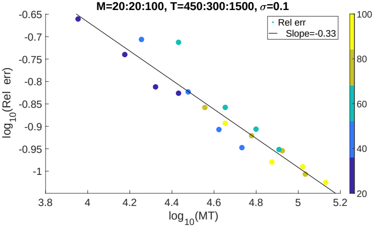


Close to optimal rates of convergence:

Rate in M



Rate in $M \times T$



Outline

The coercivity condition

$\langle\langle \phi, \psi \rangle\rangle = \frac{1}{TN} \int_0^T \mathbb{E}[\langle R_\phi(\mathbf{X}_t), R_\psi(\mathbf{X}_t)\rangle] dt$

$$\langle\langle \phi, \phi \rangle\rangle \geq c_{\mathcal{H}}^T \|\phi\|_{L^2_\rho}^2, \ \forall \phi \in \mathcal{H}.$$

When does it hold?

- Partial results for $\mathcal{H} = L^2_\rho$ in [LMT21jmlr, LLMTZ21spa,LL21]
  - $t = 0$ with Gaussian initial distribution [LZMT19], $c_{\mathcal{H}} = \frac{1}{N-2}$
  - stochastic: Gaussian process or $r^{2\beta}$ stationary $N = 3$ [LLMTZ21spa]
  - stochastic: $r^{2\beta}$ nonlinear stationary $N > 3$ [LL20]
- Open: non-stationary? A compact $\mathcal{H} \subset C(\text{supp}(\rho_T))$?
- No coercivity on $L^2_{\rho_T}$ when $N \to \infty$ since $c_{\mathcal{H}} \to 0$

The coercivity condition

$\langle\langle \phi, \psi \rangle\rangle = \frac{1}{TN} \int_0^T \mathbb{E}[\langle R_\phi(\mathbf{X}_t), R_\psi(\mathbf{X}_t) \rangle] dt$

$$\langle\langle \phi, \phi \rangle\rangle \geq c_{\mathcal{H}}^T \|\phi\|_{L^2_\rho}^2, \ \forall \phi \in \mathcal{H}.$$

Recall:

$$R_\phi(\mathbf{X}_t)_i = \frac{1}{N} \sum_{j=1}^N \phi(|\mathbf{r}_t^{i,j}|) \frac{\mathbf{r}_t^{i,j}}{|\mathbf{r}_t^{i,j}|}, \quad \mathbf{r}_t^{i,j} = X_t^j - X_t^i, \quad r_t^{ij} = |\mathbf{r}_t^{ij}|$$

$$\begin{aligned} \langle R_\phi(\mathbf{X}_t), R_\psi(\mathbf{X}_t) \rangle &= \langle \frac{1}{N} \sum_{j=1}^N \phi(r_t^{ij}) \frac{\mathbf{r}_t^{ij}}{r_t^{ij}}, \frac{1}{N} \sum_{j=1}^N \psi(r_t^{ij}) \frac{\mathbf{r}_t^{ij}}{r_t^{ij}} \rangle \\ &= \sum_{i=1}^N \frac{1}{N^2} \sum_{j,k=1}^N \phi(r_t^{ij}) \psi(r_t^{ik}) \frac{\langle \mathbf{r}_t^{ij}, \mathbf{r}_t^{ik} \rangle}{r_t^{ik} r_t^{ij}} \end{aligned}$$

Exchangeability implies

$$\frac{1}{N}\mathbb{E}[\langle R_\phi(\mathbf{X}_t), R_\phi(\mathbf{X}_t)\rangle] = \sum_{i=1}^{N} \frac{1}{N^3} \sum_{\substack{j,k=1,\\ j\neq i, k\neq i}}^{N} \underbrace{\mathbb{E}[\phi(|\mathbf{r}_t^{ji}|)\phi(|\mathbf{r}_t^{ki}|)\frac{\langle \mathbf{r}_t^{ji}, \mathbf{r}_t^{ki}\rangle}{|\mathbf{r}_t^{ji}||\mathbf{r}_t^{ki}|}]}_{I_{ijk}}$$

$$= \frac{(N-1)(N-2)I_{123} + (N-1)I_{122}}{N^2},$$

- $I_{ijk} = I_{123}$ for $\{(i,j,k), j \neq i, k \neq i, j \neq k\}$, $\Rightarrow N(N-1)(N-2)$ copies of I_{123};
- $I_{ijk} = I_{122}$ for $\{(i,j,k), j = k \neq i\}$, $\Rightarrow N(N-1)$ copies of I_{122}.

Note that $\frac{1}{T}\int_0^T I_{122}dt = \frac{1}{T}\int_0^T \mathbb{E}[\phi(r_t^{12})^2]dt = \|\phi\|^2_{L^2_{\rho_T}}$. Denote

$$\langle \phi, L_G\phi\rangle := \frac{1}{T}\int_0^T I_{123}(t)dt = \frac{1}{T}\int_0^T \mathbb{E}[\phi(|\mathbf{r}_t^{12}|)\phi(|\mathbf{r}_t^{13}|)\frac{\langle \mathbf{r}_t^{12}, \mathbf{r}_t^{13}\rangle}{|\mathbf{r}_t^{12}||\mathbf{r}_t^{13}|}]dt.$$

$$\langle\langle \phi, \phi\rangle\rangle = \frac{1}{T}\int_0^T \frac{(N-1)(N-2)I_{123} + (N-1)I_{122}}{N^2}dt$$

$$= \frac{(N-1)(N-2)}{N^2}\langle \phi, L_G\phi\rangle + \frac{N-1}{N^2}\|\phi\|^2_{L^2_{\rho_T}}$$

Coercivity condition in operator form:

$$\langle\langle \phi, \phi \rangle\rangle \geq c_{\mathcal{H}} \|\phi\|^2_{L^2_{\rho_T}} \Leftrightarrow \frac{(N-1)(N-2)}{N^2} \inf_{\phi \in \mathcal{H}, \|\phi\|_{L^2_{\rho_T}}=1} \langle \phi, L_G \phi \rangle + \frac{N-1}{N^2} \geq c_{\mathcal{H}}$$

- $\langle\langle \phi, \phi \rangle\rangle = \langle \mathcal{L}_{\overline{G}} \phi, \phi \rangle$ with $\mathcal{L}_{\overline{G}} = \frac{(N-1)(N-2)}{N^2} L_G + \frac{N-1}{N^2} I$
- A sufficient condition for $c_{\mathcal{H}} = \frac{N-1}{N^2}$ with $\mathcal{H} = L^2_{\rho_T}$:

$$\langle \phi, L_G \phi \rangle := \frac{1}{T} \int_0^T \mathbb{E}[\phi(|\mathbf{r}_t^{12}|)\phi(|\mathbf{r}_t^{13}|) \frac{\langle \mathbf{r}_t^{12}, \mathbf{r}_t^{13} \rangle}{|\mathbf{r}_t^{12}||\mathbf{r}_t^{13}|}] dt \geq 0, \forall \phi \in L^2_{\rho_T}$$

$$\frac{1}{T}\int_0^T \mathbb{E}[\phi(|\mathbf{r}_t^{12}|)\phi(|\mathbf{r}_t^{13}|)\frac{\langle \mathbf{r}_t^{12}, \mathbf{r}_t^{13}\rangle}{|\mathbf{r}_t^{12}||\mathbf{r}_t^{13}|}]dt \geq 0, \forall \phi \in L^2_{\rho_T}$$

$t = 0$, Gaussian:

Theorem (Lemma 3.2, LZTM19)

Let (X, Y, Z) be exchangeable mean zero Gaussian in $\mathbb{R}^d$ with $\mathrm{cov}(X) - \mathrm{cov}(X, Y) = \lambda I_d$ for $\lambda > 0$. Then, the coercivity condition holds true with $c_{\mathcal{H}} = \frac{N-1}{N^2}$ on L^2_ρ with $\rho \propto r^{d-1}e^{-r^2/3}$:

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

- $LHS = \int\int \phi(r)\phi(s)G(r,s)drds$ and show G is positive definite.
- Open: general distribution?

Theorem (LLMTZ21spa,LL20)

Stochastic system. Then, CC holds with $c_{\mathcal{H}} = \frac{N-1}{N^2}$ *on* $\mathcal{H} = L^2_\rho$*:*

- *Linear* $\phi(r) = \theta r$*; IC: non-degenerate exchangeable Gaussian.*
- *Nonlinear stationary:* $\Phi(r) = ar^{2\beta} + \Phi_0(r)$*, with* $a > 0$*,* $\beta \in [\frac{1}{2}, 1]$*,* $\Phi_0 \in C^2$ *s.t.* $f(u, v) = \Phi(|u - v|)$ *is negative definite, and* $\lim_{r\to\infty} \Phi(r) = +\infty$*.*

Major idea:

- Write it in the form with an integral operator with kernel G_T

$$\frac{1}{T}\int_0^T \mathbb{E}[\phi(|\mathbf{r}_t^{12}|)\phi(|\mathbf{r}_t^{13}|)\frac{\langle \mathbf{r}_t^{12}, \mathbf{r}_t^{13}\rangle}{|\mathbf{r}_t^{12}||\mathbf{r}_t^{13}|}]dt = \int \phi(r)\phi(s)G_T(r, s)\rho(r)\rho(s)drds$$

- Gaussian: G_T is strictly positive definite using Müntz type theorems (i.e., $\{r^{2k}\}_{k\geq 0}$ is dense in L^2_ρ)
- Nonlinear: a “comparison to Gaussian kernels” technique
- Open: Non-stationary? Relax the symmetry?

A new idea: conditional independence.

At IC with iid components: by independence

$$\mathbb{E}[\phi(r_{12})\frac{\mathbf{r}_{12}}{r_{12}}|X_1] = \mathbb{E}[\phi(r_{13})\frac{\mathbf{r}_{13}}{r_{13}}|X_1]$$

Then,

$$\begin{aligned} &\mathbb{E}[\phi(r_{12})\psi(r_{13})\frac{\langle \mathbf{r}_{12}, \mathbf{r}_{13}\rangle}{r_{12}r_{13}}] \\ =&\mathbb{E}[\langle \mathbb{E}[\phi(r_{12})\frac{\mathbf{r}_{12}}{r_{12}}|X_1], \mathbb{E}[\phi(r_{13})\frac{\mathbf{r}_{13}}{r_{13}}|X_1]\rangle] \\ =&\mathbb{E}[\left\|\mathbb{E}[\phi(r_{12})\frac{\mathbf{r}_{12}}{r_{12}}|X_1]\right\|^2] \geq 0 \end{aligned}$$

Question: can we extend it to the stochastic process?

- the EM generated process: yes
- continuous-time process?

Summary

Learn the interaction kernel in IPS, $N < \infty$

- Multiple trajectories M
- New from classical learning theory: coercivity condition
  - $\mathcal{E}_\infty(\phi) - \mathcal{E}_\infty(\phi_{\mathcal{H}}) \geq c_{\mathcal{H}} \|\phi - \phi_{\mathcal{H}}\|_2^2$
  - Well-posed inversion
  - Open questions: important for nonlocal dependence (later)
- Nonparametric regression: rate of convergence in M

Summary

Learn the interaction kernel in IPS, $N < \infty$

- Multiple trajectories M
- New from classical learning theory: coercivity condition
  - $\mathcal{E}_\infty(\phi) - \mathcal{E}_\infty(\phi_\mathcal{H}) \geq c_\mathcal{H} \|\phi - \phi_\mathcal{H}\|_2^2$
  - Well-posed inversion
  - Open questions: important for nonlocal dependence (later)
- Nonparametric regression: rate of convergence in M

Next questions:

- Is the rate minimax?
- what if $N \to \infty$?
  Data: particle ensemble/microscopic density,

$$u_N(x,t) = \frac{1}{N} \sum_{i=1}^{N} \delta_{X_t^i}(x) \to u(x,t)$$

  an inverse problem for the mean-field equations.