Semi-Supervised Deep Sobolev Regression: Estimation, Variable Selection, and Beyond

Chenguang Duan

Wuhan University
Joint work with Zhao Ding, Yuling Jiao, and Jerry Zhijian Yang

July 11, 2025

Nonparametric regression model

$$Y_i = f_0(X_i) + \xi_i$$

- $X_1, \ldots, X_n \sim^{\text{i.i.d.}} \mu$.
- $\xi_1, \ldots, \xi_n \sim^{\text{i.i.d.}} N(0, \sigma^2 I_d)$.
- $\xi_{1:n}$ is independent of $X_{1:n}$.

The goal of the regression is to find the unknown function f_0 using labeled data $\{(X_i, Y_i)\}_{i=1}^n$.

Nonparametric regression model

$$Y_i = f_0(X_i) + \xi_i$$

- $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mu$.
- $\xi_1, \ldots, \xi_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2 I_d)$.
- $\xi_{1:n}$ is independent of $X_{1:n}$.

The goal of the regression is to find the unknown function f_0 using labeled data $\{(X_i, Y_i)\}_{i=1}^n$.

Least-squares regression

$$f_0 = \underset{f \text{ measurable}}{\arg\min} \; L(f) := \mathbb{E}_{(X,Y)}\left[(Y - f(X))^2\right]$$

Nonparametric regression model

$$Y_i = f_0(X_i) + \xi_i$$

- $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mu$.
- $\xi_1, \ldots, \xi_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2 I_d)$.
- $\xi_{1:n}$ is independent of $X_{1:n}$.

The goal of the regression is to find the unknown function f_0 using labeled data $\{(X_i, Y_i)\}_{i=1}^n$.

Least-squares regression

$$f_0 = \underset{f \text{ measurable}}{\arg\min} \; L(f) := \mathbb{E}_{(X,Y)}\left[(Y - f(X))^2\right]$$

- The expectation is intractable.
- Minimization over the measurable function class is intractable.

Nonparametric regression model

$$Y_i = f_0(X_i) + \xi_i$$

- $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mu$.
- $\xi_1, \ldots, \xi_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2 I_d)$.
- $\xi_{1:n}$ is independent of $X_{1:n}$.

The goal of the regression is to find the unknown function f_0 using labeled data $\{(X_i, Y_i)\}_{i=1}^n$.

Least-squares regression

$$f_0 = \underset{f \text{ measurable}}{\arg\min} \; L(f) := \mathbb{E}_{(X,Y)}\left[(Y - f(X))^2\right]$$

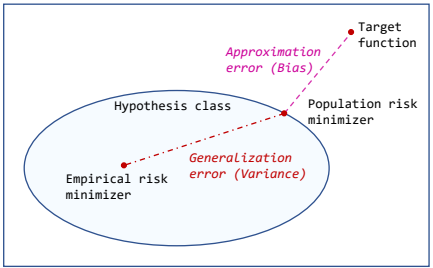
- The expectation is intractable. Monte Carlo Approximation
- Minimization over the measurable function class is intractable. Parameterization

$$\widehat{f}_n \in \arg\min_{f \in \mathcal{F}} \widehat{L}_n(f) := \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2$$

$$\widehat{f}_n \in \arg\min_{f \in \mathcal{F}} \widehat{L}_n(f) := \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2$$

Error decomposition

$$\mathbb{E}_S \left[\|\widehat{f}_n - f_0\|_2^2 \right] \leq \underbrace{\inf_{f \in \mathcal{F}} \|f - f_0\|_{L^2(\Omega)}}_{\text{Approximation error}} + \underbrace{\mathbb{E}_S \left[\sup_{f \in \mathcal{F}} (L(f) - \widehat{L}_n(f)) \right]}_{\text{Generalization error}}$$



Bias-Variance trade-off

Choose the hypothesis class as a **deep neural network** class.

A neural network $\phi : \mathbb{R}^d \to \mathbb{R}$ is a function defined by

$$f(x) = T_L(\rho(T_{L-1}(\cdots \rho(T_0(x)) \cdots))), \tag{DNN}$$

where the activation function ρ is applied component-wisely and $T_\ell(x) = A_\ell x + b_\ell$ is an affine transformation with $A_\ell \in \mathbb{R}^{N_{\ell+1} \times N_\ell}$ and $b_\ell \in \mathbb{R}^{N_\ell}$ for $\ell = 0, \ldots, L$. Then the ρ-activated deep neural network class $N_\rho(L, S, B)$ is defined as

$$N_\rho(L, S, B) = \left\{ f(x) \text{ has the form of (DNN)} : \sum_{0 \le \ell \le L+1} (\|A_\ell\|_0 + \|b_\ell\|_2) \le S,\ \|f\|_{L^\infty(\mathbb{R}^d)} \le B \right\},$$

where L is called the depth of neural networks, S represents the number of non-zero parameters, and B is the uniform bound of neural networks.

Approximation error of DNN

- Approximation in L^p-norm
  (Dmitry Yarotsky 2017, Zuowei Shen et al. 2019, Johannes Schmidt-Hieber 2020, Jianfeng Lu et al. 2020, Yuling Jiao et al. 2023)
- Approximation in Sobolev norms
  (Ingo Gühring et al. 2021, **Chenguang Duan et al. 2022**)
- Approximation with Lipschitz constraint
  (Jian Huang et al. 2022, Yuling Jiao et al. 2023, **Zhao Ding et al. 2024**)

Generalization error of DNN

Empirical process + Sample complexity of DNN

- (Local) Rademacher complexity (Peter L. Bartlett et al. 2002, 2005)
- Size-independent Rademacher complexity of DNN (Noah Golowich et al. 2018)
- VC-dimention bound for DNN (Martin Anthony et al. 1999, Peter L. Bartlett et al. 1998, 2019)

Suppose that $f_0 \in \mathcal{H}^s(\Omega)$, then the minimax optimal rate of en estimator $\widehat{f}_n$ is given as:

$$\mathbb{E}_S\left[\|\widehat{f}_n - f_0\|^2_{L^2(\Omega)}\right] = \mathcal{O}\left(n^{-\frac{2s}{d+2s}}\right).$$

Suppose that $f_0 \in \mathcal{H}^s(\Omega)$, then the minimax optimal rate of en estimator $\widehat{f}_n$ is given as:

$$\mathbb{E}_S\left[\|\widehat{f}_n - f_0\|^2_{L^2(\Omega)}\right] = \mathcal{O}\left(n^{-\frac{2s}{d+2s}}\right).$$

The convergence in L^2-norm can **NOT** imply the convergence of gradient.

- For example,

$$\widehat{f}_n(x) = \frac{\sin(nx)\log n}{n} \xrightarrow{L^p} 0, \quad \text{for each } 1 \le p \le \infty.$$

  However, $\widehat{f}_n'(x) = \cos(nx)\log n \to +\infty$.

Suppose that $f_0 \in \mathcal{H}^s(\Omega)$, then the minimax optimal rate of en estimator $\widehat{f}_n$ is given as:

$$\mathbb{E}_S\left[\|\widehat{f}_n - f_0\|^2_{L^2(\Omega)}\right] = \mathcal{O}\left(n^{-\frac{2s}{d+2s}}\right).$$

The convergence in L^2-norm can **NOT** imply the convergence of gradient.

- For example,

$$\widehat{f}_n(x) = \frac{\sin(nx)\log n}{n} \xrightarrow{L^p} 0, \quad \text{for each } 1 \le p \le \infty.$$

However, $\widehat{f}_n'(x) = \cos(nx)\log n \to +\infty$.

How can we simultaneously estimate both the regression function f_0 and its gradient ∇f_0?

Applications in inverse problems, nonparametric variable selection, generative learning ...

Least-squares regression with gradient penalty

$$L^{\lambda}(f) := \underbrace{\mathbb{E}_{(X,Y)}\left[(Y - f(X))^2\right]}_{\text{least-squares}} + \underbrace{\lambda |f|^2_{H^1(\Omega)}}_{\text{gradient penalty}} = \|f - f_0\|^2_{L^2(\Omega)} + \sigma^2 + \lambda |f|_{H^1(\Omega)}.$$

Least-squares regression with gradient penalty

$$L^\lambda(f) := \underbrace{\mathbb{E}_{(X,Y)}\left[(Y - f(X))^2\right]}_{\text{least-squares}} + \underbrace{\lambda |f|^2_{H^1(\Omega)}}_{\text{gradient penalty}} = \|f - f_0\|^2_{L^2(\Omega)} + \sigma^2 + \lambda |f|_{H^1(\Omega)}.$$

Variational problem

Find $f^\lambda \in H^1(\Omega)$, such that $\delta L^\lambda(f^\lambda, v) = 0$ for each $v \in H^1(\Omega)$, that is,

$$(f^\lambda - f_0, v)_{L^2(\Omega)} + \lambda(\nabla(f^\lambda - f_0), \nabla v)_{L^2(\Omega)} = -\lambda(\nabla f_0, \nabla v)_{L^2(\Omega)}.$$

Least-squares regression with gradient penalty

$$L^\lambda(f) := \underbrace{\mathbb{E}_{(X,Y)}\left[(Y - f(X))^2\right]}_{\text{least-squares}} + \underbrace{\lambda |f|^2_{H^1(\Omega)}}_{\text{gradient penalty}} = \|f - f_0\|^2_{L^2(\Omega)} + \sigma^2 + \lambda |f|_{H^1(\Omega)}.$$

Variational problem

Find $f^\lambda \in H^1(\Omega)$, such that $\delta L^\lambda(f^\lambda, v) = 0$ for each $v \in H^1(\Omega)$, that is,

$$(f^\lambda - f_0, v)_{L^2(\Omega)} + \lambda(\nabla(f^\lambda - f_0), \nabla v)_{L^2(\Omega)} = -\lambda(\nabla f_0, \nabla v)_{L^2(\Omega)}.$$

Notice that f^λ is the solution to the elliptic equation

$$\begin{cases} -\lambda \Delta f + f = f_0, & \text{in } \Omega, \\ \nabla f \cdot \mathbf{n} = 0, & \text{on } \partial\Omega. \end{cases}$$

Substituting $v = f^\lambda - f_0$ into $\delta L^\lambda(f^\lambda, v) = 0$ implies

$$\overbrace{\|f^\lambda - f_0\|^2_{L^2(\Omega)}}^{\text{interior } L^2 \text{ error}} + \lambda \overbrace{|f^\lambda - f_0|^2_{H^1(\Omega)}}^{\text{interior gradient error}}$$
$$= \lambda(\Delta f_0, f_0 - f^\lambda)_{L^2(\Omega)} \leq \lambda \|\Delta f_0\|_{L^2(\Omega)} \|f^\lambda - f_0\|_{L^2(\Omega)}.$$

Substituting $v = f^\lambda - f_0$ into $\delta L^\lambda(f^\lambda, v) = 0$ implies

$$\overbrace{\|f^\lambda - f_0\|^2_{L^2(\Omega)}}^{\text{interior } L^2 \text{ error}} + \lambda \overbrace{|f^\lambda - f_0|^2_{H^1(\Omega)}}^{\text{interior gradient error}}$$
$$= \lambda(\Delta f_0, f_0 - f^\lambda)_{L^2(\Omega)} \leq \lambda \|\Delta f_0\|_{L^2(\Omega)} \|f^\lambda - f_0\|_{L^2(\Omega)}.$$

- Interior L^2 error:

$$\|f^\lambda - f_0\|^2_{L^2(\Omega)} \leq \lambda^2 \|\Delta f_0\|^2_{L^2(\Omega)}.$$

Substituting $v = f^\lambda - f_0$ into $\delta L^\lambda(f^\lambda, v) = 0$ implies

$$\overbrace{\|f^\lambda - f_0\|^2_{L^2(\Omega)}}^{\text{interior } L^2 \text{ error}} + \lambda \overbrace{|f^\lambda - f_0|^2_{H^1(\Omega)}}^{\text{interior gradient error}}$$
$$= \lambda(\Delta f_0, f_0 - f^\lambda)_{L^2(\Omega)} \leq \lambda \|\Delta f_0\|_{L^2(\Omega)} \|f^\lambda - f_0\|_{L^2(\Omega)}.$$

- Interior L^2 error:

$$\|f^\lambda - f_0\|^2_{L^2(\Omega)} \leq \lambda^2 \|\Delta f_0\|^2_{L^2(\Omega)}.$$

- Interior gradient error:

$$|f^\lambda - f_0|^2_{H^1(\Omega)} \leq \lambda \|\Delta f_0\|^2_{L^2(\Omega)}.$$

Substituting $v = f^\lambda - f_0$ into $\delta L^\lambda(f^\lambda, v) = 0$ implies

$$\overbrace{\|f^\lambda - f_0\|^2_{L^2(\Omega)}}^{\text{interior } L^2 \text{ error}} + \lambda \overbrace{|f^\lambda - f_0|^2_{H^1(\Omega)}}^{\text{interior gradient error}}$$
$$= \lambda(\Delta f_0, f_0 - f^\lambda)_{L^2(\Omega)} \leq \lambda \|\Delta f_0\|_{L^2(\Omega)} \|f^\lambda - f_0\|_{L^2(\Omega)}.$$

- Interior L^2 error:

$$\|f^\lambda - f_0\|^2_{L^2(\Omega)} \leq \lambda^2 \|\Delta f_0\|^2_{L^2(\Omega)}.$$

- Interior gradient error:

$$|f^\lambda - f_0|^2_{H^1(\Omega)} \leq \lambda \|\Delta f_0\|^2_{L^2(\Omega)}.$$

The Sobolev-penalized regressor f^λ is an estimator of f_0 in both L^2-norm and H^1-semi-norm.

Sobolev-Penalized Risk

$$L^{\lambda}(f) := \mathbb{E}_{(X,Y)}\left[(Y - f(X))^2\right] + \lambda |f|^2_{H^1(\Omega)}$$

- The expectation is intractable. — Monte Carlo Approximation
- Minimization over the measurable function class is intractable. — Parameterization

Sobolev-Penalized Risk

$$L^{\lambda}(f) := \mathbb{E}_{(X,Y)}\left[(Y - f(X))^2\right] + \lambda |f|^2_{H^1(\Omega)}$$

- The expectation is intractable. — Monte Carlo Approximation
- Minimization over the measurable function class is intractable. — Parameterization

---

Empirical Sobolev-Penalized Risk Minimization

$$\widehat{f}^{\lambda}_{n,m} \in \underset{f \in \mathrm{conv}(\mathcal{F})}{\arg\min}\ \widehat{L}^{\lambda}_{n,m}(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 + \frac{\lambda}{m}\sum_{j=1}^{m}\|\nabla f(Z_j)\|_2^2.$$

- Labeled data: $(X_i, Y_i) \sim^{\text{i.i.d.}} P$
- Unlabeled data: $Z_j \sim^{\text{i.i.d.}} \mathrm{unif}(\Omega)$

Sobolev-Penalized Risk

$$L^{\lambda}(f) := \mathbb{E}_{(X,Y)}\left[(Y - f(X))^2\right] + \lambda |f|^2_{H^1(\Omega)}$$

- The expectation is intractable. — Monte Carlo Approximation
- Minimization over the measurable function class is intractable. — Parameterization

---

Empirical Sobolev-Penalized Risk Minimization

$$\widehat{f}^{\lambda}_{n,m} \in \underset{f \in \mathrm{conv}(\mathcal{F})}{\arg\min}\ \widehat{L}^{\lambda}_{n,m}(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 + \frac{\lambda}{m}\sum_{j=1}^{m}\|\nabla f(Z_j)\|_2^2.$$

- Labeled data: $(X_i, Y_i) \sim^{\text{i.i.d.}} P$
- Unlabeled data: $Z_j \sim^{\text{i.i.d.}} \mathrm{unif}(\Omega)$

Semi-supervised framework

- **A1. Sub-Gaussian noise.** The noise ξ is sub-Gaussian with mean 0 and finite variance proxy σ^2.

- **A1. Sub-Gaussian noise.** The noise ξ is sub-Gaussian with mean 0 and finite variance proxy σ^2.
- **A2. Bounded hypothesis.** There exists an absolute positive constant B_0, such that $\sup_{x\in\Omega} |f_0(x)| \leq B_0$. Further, functions in hypothesis class $\mathcal{F}$ are also bounded, that is, $\sup_{x\in\Omega} |f(x)| \leq B_0$.

- **A1. Sub-Gaussian noise.** The noise ξ is sub-Gaussian with mean 0 and finite variance proxy σ^2.
- **A2. Bounded hypothesis.** There exists an absolute positive constant B_0, such that $\sup_{x\in\Omega} |f_0(x)| \leq B_0$. Further, functions in hypothesis class $\mathcal{F}$ are also bounded, that is, $\sup_{x\in\Omega} |f(x)| \leq B_0$.
- **A3. Bounded derivatives of hypothesis.** There exists positive constants $\{B_{1,k}\}_{k=1}^d$, such that $\sup_{x\in\Omega} |D_k f_0(x)| \leq B_{1,k}$ for $1 \leq k \leq d$. Further, the first-order partial derivatives of functions in hypothesis class $\mathcal{F}$ are also bounded, i.e., $\sup_{x\in\Omega} |D_k f(x)| \leq B_{1,k}$ for each $1 \leq k \leq d$ and $f \in \mathcal{F}$. Denote by $B_1^2 := \sum_{k=1}^d B_{1,k}^2$.

- **A1. Sub-Gaussian noise.** The noise ξ is sub-Gaussian with mean 0 and finite variance proxy σ^2.
- **A2. Bounded hypothesis.** There exists an absolute positive constant B_0, such that $\sup_{x\in\Omega} |f_0(x)| \le B_0$. Further, functions in hypothesis class $\mathcal{F}$ are also bounded, that is, $\sup_{x\in\Omega} |f(x)| \le B_0$.
- **A3. Bounded derivatives of hypothesis.** There exists positive constants $\{B_{1,k}\}_{k=1}^d$, such that $\sup_{x\in\Omega} |D_k f_0(x)| \le B_{1,k}$ for $1 \le k \le d$. Further, the first-order partial derivatives of functions in hypothesis class $\mathcal{F}$ are also bounded, i.e., $\sup_{x\in\Omega} |D_k f(x)| \le B_{1,k}$ for each $1 \le k \le d$ and $f \in \mathcal{F}$. Denote by $B_1^2 := \sum_{k=1}^d B_{1,k}^2$.
- **A4. Regularity of regression function.** The regression function satisfies $\Delta f_0 \in L^2(\Omega)$ and $\nabla f_0 \cdot \mathbf{n} = 0$ a.e. on $\partial\Omega$, where $\boldsymbol{n}$ is the unit normal to the boundary.

Oracle inequality

Under **A1** to **A4**. Suppose $n \geq \log N(B_0\delta, \mathcal{F}, L^2(\mathcal{D}))$ and $m \geq \max_k \log N(B_{1,k}\delta, D_k\mathcal{F}, L^2(\mathcal{S}))$. Then

$$\mathbb{E}\big[\|\widehat{f}^{\lambda}_{n,m} - f_0\|^2_{L^2(\Omega)}\big] \lesssim \beta\lambda^2 + \varepsilon_{\text{app}}(\mathcal{F}, \lambda) + \varepsilon_{\text{gen}}(\mathcal{F}, n) + \varepsilon^{\text{reg}}_{\text{gen}}(\nabla\mathcal{F}, m),$$
$$\mathbb{E}\big[\|\nabla(\widehat{f}^{\lambda}_{n,m} - f_0)\|^2_{L^2(\Omega)}\big] \lesssim \beta\lambda + \lambda^{-1}\varepsilon_{\text{app}}(\mathcal{F}, \lambda) + \lambda^{-1}\varepsilon_{\text{gen}}(\mathcal{F}, n) + \lambda^{-1}\varepsilon^{\text{reg}}_{\text{gen}}(\nabla\mathcal{F}, m),$$

where $\beta = \|\Delta f_0\|^2_{L^2(\Omega)} + B_1^2$. The approximation error $\varepsilon_{\text{app}}(\mathcal{F}, \lambda)$, the generalization errors $\varepsilon_{\text{gen}}(\mathcal{F}, n)$ and $\varepsilon^{\text{reg}}_{\text{gen}}(\nabla\mathcal{F}, m)$ are defined, respectively, as

$$\varepsilon_{\text{app}}(\mathcal{F}, \lambda) = \inf_{f\in\mathcal{F}} \Big\{ \|f - f_0\|^2_{L^2(\Omega)} + \lambda\|\nabla(f - f_0)\|^2_{L^2(\Omega)} \Big\},$$
$$\varepsilon_{\text{gen}}(\mathcal{F}, n) = (B_0^2 + \sigma^2)(\log n) \inf_{\delta>0} \Big\{ \Big(\frac{2\log N(B_0\delta, \mathcal{F}, L^2(\mathcal{D}))}{n} \Big)^{\frac{1}{2}} + \delta \Big\},$$
$$\varepsilon^{\text{reg}}_{\text{gen}}(\nabla\mathcal{F}, m) = B_1^2 \inf_{\delta>0} \Big\{ \max_{1\leq k\leq d} \frac{\log N(B_{1,k}\delta, D_k\mathcal{F}, L^2(\mathcal{S}))}{m} + \delta \Big\}.$$

Approximation error

Let $\Omega \subseteq K \subseteq \mathbb{R}^d$ be two bounded domain. Set the hypothesis class as a deep ReQU neural network $\mathcal{F} = \mathcal{N}(L, W, S)$ with $L = \mathcal{O}(\log N)$ and $S = \mathcal{O}(N^d)$. Then for each $\phi \in C^s(K)$ with $s \in \mathbb{N}_{\geq 2}$, there exists $f \in \mathcal{F}$ such that

$$\|f - \phi\|_{L^2(\Omega)} \leq CN^{-s}\|\phi\|_{C^s(K)},$$
$$\|\nabla(f - \phi)\|_{L^2(\Omega)} \leq CN^{-(s-1)}\|\phi\|_{C^s(K)},$$

where C is a constant independent of N.

Approximation error

Let $\Omega \subseteq K \subseteq \mathbb{R}^d$ be two bounded domain. Set the hypothesis class as a deep ReQU neural network $\mathcal{F} = \mathcal{N}(L, W, S)$ with $L = \mathcal{O}(\log N)$ and $S = \mathcal{O}(N^d)$. Then for each $\phi \in C^s(K)$ with $s \in \mathbb{N}_{\geq 2}$, there exists $f \in \mathcal{F}$ such that

$$\|f - \phi\|_{L^2(\Omega)} \leq CN^{-s}\|\phi\|_{C^s(K)},$$
$$\|\nabla(f - \phi)\|_{L^2(\Omega)} \leq CN^{-(s-1)}\|\phi\|_{C^s(K)},$$

where C is a constant independent of N.

Generalization error

Suppose the activation function is piecewise-polynomial. Let $\mathcal{D} = \{X_i\}_{i=1}^n$ and $\mathcal{S} = \{Z_j\}_{j=1}^m$. Then

$$\log N(\delta, N(L, S, B), L^2(\mathcal{D})) \lesssim LS \log(S) \log\left(\frac{nB}{\delta}\right),$$
$$\log N(\delta, D_k N(L, S, B), L^2(\mathcal{S})) \lesssim L^2 S \log(S) \log\left(\frac{mB}{\delta}\right).$$

Under **A1** to **A4**. Let $\Omega \subseteq K \subseteq \mathbb{R}^d$ be two bounded domain. Assume that $f_0 \in C^s(K)$ with $s \in \mathbb{N}_{\geq 2}$. Set the hypothesis class as a deep ReQU neural network class $\mathcal{F} = N(L, W, S)$ with $L = \mathcal{O}(\log n)$ and $S = \mathcal{O}(n^{\frac{d}{d+4s}})$. Let $\lambda = \mathcal{O}(n^{-\frac{s}{d+4s}} \log^2 n)$. Then

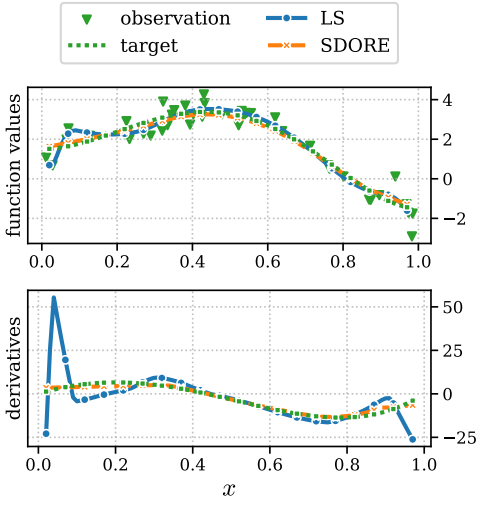
$$\mathbb{E}\left[\|\widehat{f}^{\lambda}_{n,m} - f_0\|^2_{L^2(\Omega)}\right] \leq \mathcal{O}\left(n^{-\frac{2s}{d+4s}} \log^4 n\right) + \mathcal{O}\left(n^{\frac{d}{d+4s}} \log^4 nm^{-1}\right),$$
$$\mathbb{E}\left[\|\nabla(\widehat{f}^{\lambda}_{n,m} - f_0)\|^2_{L^2(\Omega)}\right] \leq \mathcal{O}\left(n^{-\frac{s}{d+4s}} \log^2 n\right) + \mathcal{O}\left(n^{\frac{d+s}{d+4s}} \log^2 nm^{-1}\right).$$

Under **A1** to **A4**. Let $\Omega \subseteq K \subseteq \mathbb{R}^d$ be two bounded domain. Assume that $f_0 \in C^s(K)$ with $s \in \mathbb{N}_{\geq 2}$. Set the hypothesis class as a deep ReQU neural network class $\mathcal{F} = N(L, W, S)$ with $L = \mathcal{O}(\log n)$ and $S = \mathcal{O}(n^{\frac{d}{d+4s}})$. Let $\lambda = \mathcal{O}(n^{-\frac{s}{d+4s}} \log^2 n)$. Then

$$\mathbb{E}\left[\|\widehat{f}^{\lambda}_{n,m} - f_0\|^2_{L^2(\Omega)}\right] \leq \mathcal{O}\left(n^{-\frac{2s}{d+4s}} \log^4 n\right) + \mathcal{O}\left(n^{\frac{d}{d+4s}} \log^4 nm^{-1}\right),$$
$$\mathbb{E}\left[\|\nabla(\widehat{f}^{\lambda}_{n,m} - f_0)\|^2_{L^2(\Omega)}\right] \leq \mathcal{O}\left(n^{-\frac{s}{d+4s}} \log^2 n\right) + \mathcal{O}\left(n^{\frac{d+s}{d+4s}} \log^2 nm^{-1}\right).$$

Simultaneously estimate both the regression function f_0 and its gradient ∇f_0.

$$f_0(x) = 1 + 36x^2 - 59x^3 + 21x^5 + 0.5\cos(\pi x)$$



- **A5. Sparsity structure**
  There exists $f_0^* : \mathbb{R}^{d^*} \to \mathbb{R}$ $(1 \leq d^* < d)$ such that for each $x := (x_1, \ldots, x_d) \in \mathbb{R}^d$,

$$f_0(x_1, \ldots, x_d) = f_0^*(x_{j_1}, \ldots, x_{j_{d^*}}), \quad \{j_1, \ldots, j_{d^*}\} \subseteq [d].$$

- **A5. Sparsity structure**
  There exists $f_0^* : \mathbb{R}^{d^*} \to \mathbb{R}$ $(1 \leq d^* < d)$ such that for each $x := (x_1, \ldots, x_d) \in \mathbb{R}^d$,

$$f_0(x_1, \ldots, x_d) = f_0^*(x_{j_1}, \ldots, x_{j_{d^*}}), \quad \{j_1, \ldots, j_{d^*}\} \subseteq [d].$$

The derivatives can indicate whether a variable is relevant to the output.

- **A5. Sparsity structure**
  There exists $f_0^* : \mathbb{R}^{d^*} \to \mathbb{R}$ $(1 \le d^* < d)$ such that for each $x := (x_1, \ldots, x_d) \in \mathbb{R}^d$,

$$f_0(x_1, \ldots, x_d) = f_0^*(x_{j_1}, \ldots, x_{j_{d^*}}), \quad \{j_1, \ldots, j_{d^*}\} \subseteq [d].$$

The derivatives can indicate whether a variable is relevant to the output.

Relevant variable

- A variable $k \in [d]$ is irrelevant for the function f with respect to Lebesgue measure on Ω, if

$$D_k f(X) = 0 \quad \text{almost surely,}$$

and relevant otherwise. The set of relevant variables is defined as

$$\mathcal{I}(f) = \{k \in [d] : \|D_k f\|_{L^2(\Omega)} > 0\}.$$

	Is f_0 sparse?	Convergence rate
$\mathbb{E}[\|\widehat{f}^{\lambda}_{n,m} - f_0\|^2_{L^2(\Omega)}]$	✗	$\widetilde{\mathcal{O}}(n^{-\frac{2s}{d+4s}})$
$\mathbb{E}[\|\widehat{f}^{\lambda}_{n,m} - f_0\|^2_{L^2(\Omega)}]$	✓	$\widetilde{\mathcal{O}}(n^{-\frac{2s}{d^*+4s}})$
$\mathbb{E}[\|\nabla(\widehat{f}^{\lambda}_{n,m} - f_0)\|^2_{L^2(\Omega)}]$	✗	$\widetilde{\mathcal{O}}(n^{-\frac{s}{d+4s}})$
$\mathbb{E}[\|\nabla(\widehat{f}^{\lambda}_{n,m} - f_0)\|^2_{L^2(\Omega)}]$	✓	$\widetilde{\mathcal{O}}(n^{-\frac{2s}{d^*+4s}})$

	Is f_0 sparse?	Convergence rate
$\mathbb{E}[\|\widehat{f}^{\lambda}_{n,m} - f_0\|^2_{L^2(\Omega)}]$	✗	$\widetilde{\mathcal{O}}(n^{-\frac{2s}{d+4s}})$
$\mathbb{E}[\|\widehat{f}^{\lambda}_{n,m} - f_0\|^2_{L^2(\Omega)}]$	✓	$\widetilde{\mathcal{O}}(n^{-\frac{2s}{d^*+4s}})$
$\mathbb{E}[\|\nabla(\widehat{f}^{\lambda}_{n,m} - f_0)\|^2_{L^2(\Omega)}]$	✗	$\widetilde{\mathcal{O}}(n^{-\frac{s}{d+4s}})$
$\mathbb{E}[\|\nabla(\widehat{f}^{\lambda}_{n,m} - f_0)\|^2_{L^2(\Omega)}]$	✓	$\widetilde{\mathcal{O}}(n^{-\frac{2s}{d^*+4s}})$

Selection consistency

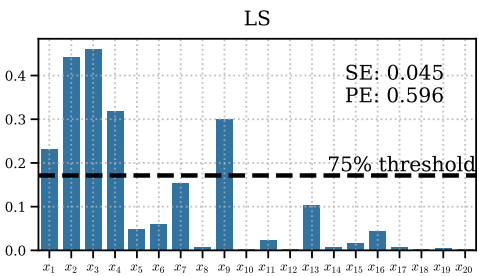
Under **A1** to **A5**. It follows that

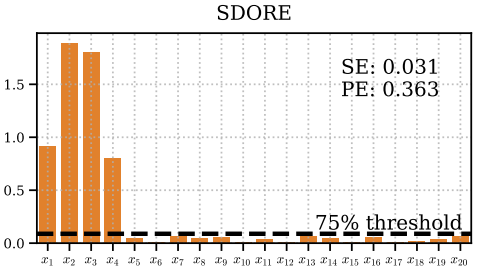
$$\lim_{n\to\infty} \Pr\left\{\mathcal{I}(f_0) = \mathcal{I}(\widehat{f}^{\lambda}_{n,m})\right\} = 1,$$

where $\lambda = \mathcal{O}(n^{-\frac{s}{d^*+4s}} \log^2 n)$, and m is sufficiently large.

$$f_0(x_1, \ldots, x_{20}) = \sum_{i=1}^{3} \sum_{j=i+1}^{4} x_i x_j.$$

sparse structure





Elliptic equation with unknown source

$$\begin{cases} -\nabla \cdot (a(x)\nabla u(x)) + c(x)u = f(x), & \text{in } \Omega, \\ \nabla u \cdot \mathbf{n} = 0, & \text{on } \partial\Omega. \end{cases}$$

Measurement model

$$Y = \mathcal{S}(f^\dagger)(X) + \xi.$$

- $u^\dagger = \mathcal{S}(f^\dagger)$ is the solution to the elliptic equation.
- $X \sim \mathrm{unif}(\Omega)$, and $\xi \sim \mathrm{subG}(\sigma^2)$ is the random noise independent of X.

Recovering Procedure

- Sobolev-penalized regression using interior measurements

$$\widehat{u}^{\lambda}_{n,m} \in \underset{u \in \mathrm{conv}(\mathcal{U})}{\arg\min}\ \widehat{L}^{\lambda}_{n,m}(u) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - u(X_i))^2 + \frac{\lambda}{m}\sum_{j=1}^{m}\|\nabla u(Z_j)\|_2^2.$$

  - Interior position-measurement pairs: $\{(X_i, Y_i)\}_{i=1}^{n}$. expensive
  - Positions variables $Z_1, \ldots, Z_m \overset{\text{i.i.d.}}{\sim} \mathrm{unif}(\Omega)$ very cheap

Recovering Procedure

- Sobolev-penalized regression using interior measurements

$$\widehat{u}_{n,m}^{\lambda} \in \underset{u \in \mathrm{conv}(\mathcal{U})}{\arg\min}\ \widehat{L}_{n,m}^{\lambda}(u) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - u(X_i))^2 + \frac{\lambda}{m}\sum_{j=1}^{m}\|\nabla u(Z_j)\|_2^2.$$

  - Interior position-measurement pairs: $\{(X_i, Y_i)\}_{i=1}^{n}$. expensive
  - Positions variables $Z_1, \ldots, Z_m \sim^{\text{i.i.d.}} \mathrm{unif}(\Omega)$ very cheap

- Recovering unknown source using gradient estimator
  Find $\widehat{f}_{n,m}^{\lambda}$, such that for each $v \in H^1(\Omega)$,

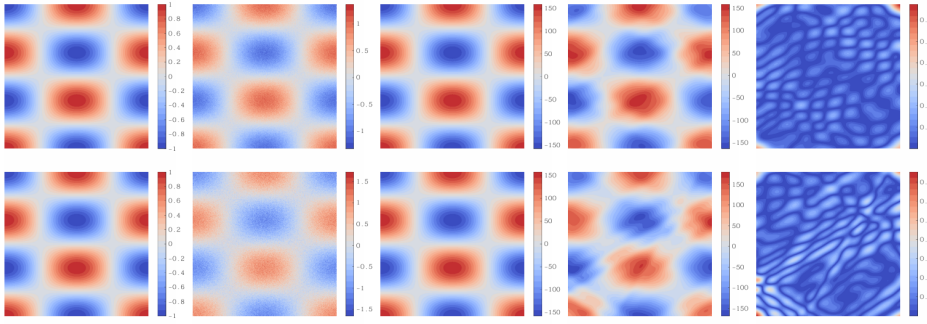
$$(a(x)\nabla \widehat{u}_{n,m}^{\lambda}, \nabla v)_{L^2(\Omega)} + (c(x)\widehat{u}_{n,m}^{\lambda}, v)_{L^2(\Omega)} = (\widehat{f}_{n,m}^{\lambda}, v)_{L^2(\Omega)}.$$

  Since $\widehat{u}_{n,m}^{\lambda} \in H^2(\Omega)$, we find

$$\widehat{f}_{n,m}^{\lambda} = -\nabla \cdot (a(x)\nabla \widehat{u}_{n,m}^{\lambda}) + c(x)\nabla \widehat{u}_{n,m}^{\lambda} \in L^2(\Omega).$$

Convergence rate in weak norm

$$\mathbb{E}\left[\|\widehat{f}^{\lambda}_{n,m} - f^{\dagger}\|_{(H^1(\Omega))^*}\right] \lesssim \mathcal{O}\left(n^{-\frac{s}{2(d+4s)}}\right).$$



(a) Clear data (b) Noisy data (c) Exact source (d) Recovery (e) Point-wise error

- Simultaneously estimations of both the regression function and its gradient.
- Nonasymptotic convergence rate.
- Applications in nonparametric variable selection and inverse problems.

Reference: Zhao Ding, Chenguang Duan, Yuling Jiao, and Jerry Zhijian Yang. Semi-Supervised Deep Sobolev Regression: Estimation and Variable Selection by ReQU Neural Network. *IEEE Transactions on Information Theory*, 2025.

Thanks for your attention!

Homepage:
`https://chenguangduan.github.io/`
Google Scholar:
`https://scholar.google.com/citations?user=RpmGgyMAAAAJ`
Email: `cgduan.math@gmail.com`