Nonparametric Instrumental Variable Regression with Observed Covariates

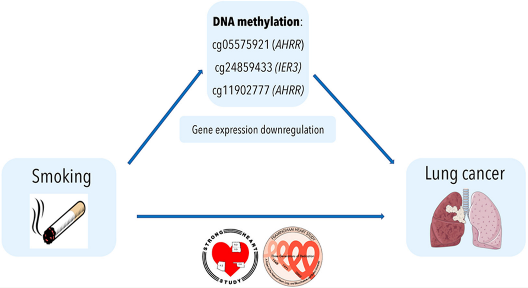
Zikai Shen* Zonghao Chen* Dimitri Meunier Ingo Steinwart
Arthur Gretton† Zhu Li†

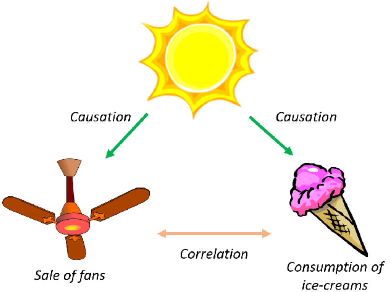
December 15, 2025

Submitted to *Annals of Statistics*

Background

- Nonparametric regression is used everywhere in statistics.
  - Kernel based estimators, (deep) neural networks, nearest neighbours, etc.
- Regression fails when ....
  - there exist confounding that affects both the input and the output.





How to prove a direct relation under (unobserved) confounding?

Background: Instrumental Variable

- In this talk, we only consider additive confounding ($\epsilon \not\perp X$).

$$Y = f_*(X) + \epsilon, \quad \mathbb{E}[\epsilon \mid X] \neq \mathbb{E}[\epsilon] = 0.$$

  - f_* is the target of interest.
  - Dose response curve, causal parameter, potential outcome, structural function, etc.
  - This shall be contrasted with standard regression setting ($\epsilon \perp X$).
- Regression always outputs a biased estimate $\mathbb{E}[Y \mid X] = f_*(X) + \mathbb{E}[\epsilon \mid X] \neq f_*(X)$.
  - This bias remains regardless of the estimators: kernel, neural networks, wavelets, and the number of observations.

Question: How to find f_* with unobserved confounder ϵ?

Background: Instrumental Variable

Definition (Instrumental Variables (Informal))

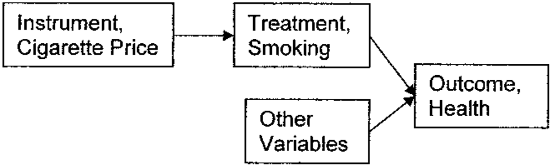
Instrumental variables Z affect Y only through X and is independent of ϵ.

- A valid instrumental variable in the smoking example could be 'price of cigarette'.

$$Y = f_*(X) + \epsilon, \quad \mathbb{E}[\epsilon \mid X] \neq 0, \quad \mathbb{E}[\epsilon \mid Z] = 0.$$

- Conditioning both sides on Z

$$\mathbb{E}[Y \mid Z] = \mathbb{E}[f_*(X) \mid Z]. \tag{NPIV}$$



Background: Instrumental Variable

- P_Z, P_X, P_Y denote the marginal distributions of P.
- $\mathbb{E}[Y \mid Z] = \mathbb{E}[f_*(X) \mid Z]$ is in fact an ill-posed inverse problem

$$\begin{aligned} Y &= \mathbb{E}[f_*(X) \mid Z] + Y - \mathbb{E}[Y \mid Z] \\ &= (Tf_*)(Z) + \upsilon. \end{aligned}$$

  - where T is a conditional expectation operator.

$$T : L^2(P_X) \to L^2(P_Z), \quad (Tf)(\mathbf{z}) = \mathbb{E}[f(X) \mid Z = \mathbf{z}].$$

  - υ satisfies $\mathbb{E}[\upsilon \mid Z] = 0$.
- Ill-posedness: T is compact and infinite-dimensional so T^{-1} is unbounded.
  - Intuition: compactness is a result of the smoothing effect from 'convolution':

$$\mathbb{E}[f(X) \mid Z = \mathbf{z}] = \int f(\mathbf{x}) p(\mathbf{x} \mid \mathbf{z}) \mathrm{d}\mathbf{x}.$$

Background: Instrumental Variable with Observed Covariates

- In practice, one has access to observed covariates (confounders) O.
  - For instance, one's occupation.

$$Y = f_*(X, O) + \epsilon, \quad \mathbb{E}[\epsilon \mid Z, O] = 0.$$

- An ill-posed inverse problem

$$Y = (Tf_*)(Z, O) + \upsilon, \quad \mathbb{E}[\upsilon \mid Z, O] = 0,$$

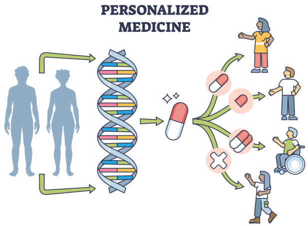
  - where T is a conditional expectation operator.

$$T : L^2(P_{XO}) \to L^2(P_{ZO}), \quad (Tf)(\mathbf{z}, \mathbf{o}) = \mathbb{E}[f(X, O) \mid Z = \mathbf{z}, O = \mathbf{o}].$$

- One has access to n i.i.d samples $\{\mathbf{x}_i, y_i, \mathbf{z}_i, \mathbf{o}_i\}_{i=1}^n$ from P which is the joint data distribution over (X, Y, Z, O).
- We call this problem nonparametric instrumental variable with observed covariates (NPIV-O).

NPIV-O

- The observed covariates O brings two advantages
  - Practitioners adjust for as many observed covariates as possible in reality.
    - Occupation, income, age, disease history, etc.
  - Personalized causal effect estimation by conditioning on $O = \mathbf{o}$.
    - The effect of smoking on lung cancer for manual laborers (heterogeneous treatment effect).



Problem: How to estimate f_* given $\{\mathbf{x}_i, y_i, \mathbf{z}_i, \mathbf{o}_i\}_{i=1}^n \sim P$?

NPIV-O

- The observed covariates O brings two challenges for its theoretical analysis
  - a) The anisotropic smoothness of $f_* : \mathcal{X} \times \mathcal{O} \to \mathbb{R}$.
  - b) The partial identity of T, which makes T no longer compact!

$$\mathfrak{G}_B = \left\{ g \in L^2(P_{XO}) \,\middle|\, \exists g_2 \in L^2(P_O) \text{ such that } \forall \mathbf{x} \in \mathcal{X}, \mathbf{o} \in \mathcal{O},\ g(\mathbf{x}, \mathbf{o}) = g_2(\mathbf{o}) \right\}.$$

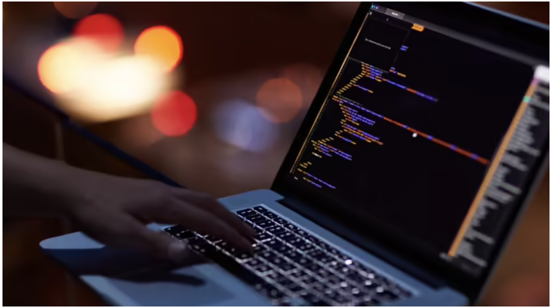
- $T^* T|_{\mathfrak{G}_B}$ is an identity operator.

$$\begin{aligned} \forall g \in \mathfrak{G}_B : \ (Tg)(\mathbf{z}, \mathbf{o}) &= \mathbb{E}[g(X, O) \mid Z = \mathbf{z}, O = \mathbf{o}] \\ &= \mathbb{E}[g_2(O) \mid Z = \mathbf{z}, O = \mathbf{o}] \\ &= g_2(\mathbf{o}) = g(\mathbf{x}, \mathbf{o}). \end{aligned}$$

- Previous analysis of NPIV relies heavily on the compactness of T!

NPIV-O

- The challenge arises in the theoretical analysis.
- But first... Let us see the algorithms.



Algorithm: Kernel 2SLS

- Suppose $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric positive definite function.
- There exists a unique reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ associated with k such that 1) $k(\mathbf{x}, \cdot) \in \mathcal{H}$. 2) the reproducing property $\langle f, k(\mathbf{x}, \cdot)\rangle_{\mathcal{H}} = f(\mathbf{x})$.
  - When $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}'$, $\mathcal{H}$ is the space of linear functions.
- $k(\mathbf{x}, \cdot) =: \phi(\mathbf{x}) \in \mathcal{H}$ is a nonlinear 'infinite' dimensional feature map.
- Tensor product kernels: $\mathfrak{K}([\mathbf{z}, \mathbf{x}], [\mathbf{z}', \mathbf{x}']) = k_{\mathcal{Z}}(\mathbf{z}, \mathbf{z}') \cdot k_{\mathcal{X}}(\mathbf{x}, \mathbf{x}')$.
- The tensor product RKHS $\mathcal{H} = \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Z}}$ with feature map

$$\phi([\mathbf{z}, \mathbf{x}]) = \phi_{\mathcal{Z}}(\mathbf{z}) \otimes \phi_{\mathcal{X}}(\mathbf{x}).$$

Algorithm: Kernel 2SLS

- NPIV-O: Learn f_* from

$$\mathbb{E}[Y \mid Z, O] = \mathbb{E}[f_*(X) \mid Z, O] = (Tf_*)(Z, O).$$

- Target $f_* = \arg\min_f \mathbb{E}_{YZO}[(Y - (Tf)(Z, O))^2]$.
  - T is unknown yet it is a conditional expectation and hence can be learned via regression.
- Two-stage least squares (2SLS)

Stage I: learn T. Stage II: learn f_*.

- The domains are $\mathcal{O} = [0,1]^{d_o}$, $\mathcal{X} = [0,1]^{d_x}$, $\mathcal{Z} = [0,1]^{d_z}$.
- We introduce four kernels $k_{\mathcal{X}}, k_{\mathcal{Z}}, k_{\mathcal{O},1}, k_{\mathcal{O},2}$ with associated $\mathcal{H}_{\mathcal{X}}, \mathcal{H}_{\mathcal{Z}}, \mathcal{H}_{\mathcal{O},1}, \mathcal{H}_{\mathcal{O},2}$.
  - The reason we need two RKHSs on $\mathcal{O}$ will be clear later on.

Algorithm: Kernel 2SLS (Stage I)

- T is a conditional expectation operator.

$$T : L^2(P_{XO}) \to L^2(P_{ZO}), \quad (Tf)(\mathbf{z}, \mathbf{o}) = \mathbb{E}[f(X, O) \mid Z = \mathbf{z}, O = \mathbf{o}].$$

- Kernel conditional mean embedding:

$$F_* : \mathcal{Z} \times \mathcal{O} \to \mathcal{H}_{\mathcal{X}}, \quad F_*(\mathbf{z}, \mathbf{o}) = \mathbb{E}[\phi_{\mathcal{X}}(X) \mid Z = \mathbf{z}, O = \mathbf{o}] \in \mathcal{H}_{\mathcal{X}}.$$

- F_* is a RKHS analogue of conditional expectation operator T.
- To see why, $\forall f \in \mathcal{H}_{\mathcal{O},2} \otimes \mathcal{H}_{\mathcal{X}}$, we have

$$\begin{aligned}
&\langle f, \phi_{\mathcal{O},2}(\mathbf{o}) \otimes F_*(\mathbf{z}, \mathbf{o}) \rangle_{\mathcal{H}_{\mathcal{O},2} \otimes \mathcal{H}_{\mathcal{X}}} \\
&= \langle f, \phi_{\mathcal{O},2}(\mathbf{o}) \otimes \mathbb{E}[\phi_{\mathcal{X}}(X) \mid Z = \mathbf{z}, O = \mathbf{o}] \rangle_{\mathcal{H}_{\mathcal{O},2} \otimes \mathcal{H}_{\mathcal{X}}} \\
&= \mathbb{E}[\langle f, \phi_{\mathcal{O},2}(\mathbf{o}) \otimes \phi_{\mathcal{X}}(X) \rangle_{\mathcal{H}_{\mathcal{O},2} \otimes \mathcal{H}_{\mathcal{X}}} \mid Z = \mathbf{z}, O = \mathbf{o}] \qquad \text{(Linearity of } \mathbb{E}) \\
&= \mathbb{E}[f(X, O) \mid Z = \mathbf{z}, O = \mathbf{o}] \qquad \text{(Reproducing property)} \\
&= (Tf)(\mathbf{z}, \mathbf{o}).
\end{aligned}$$

Algorithm: Kernel 2SLS (Stage I)

Stage I: We learn $F_* = \mathbb{E}[\phi_{\mathcal{X}}(X) \mid Z, O]$ with ridge regression.

- Stage I: Learn F_* with $\{(\tilde{\mathbf{z}}_i, \tilde{\mathbf{o}}_i, \tilde{\mathbf{x}}_i)\}_{i=1}^{\tilde{n}}$.

$$\hat{F}_\xi := \arg\min_{F \in \mathcal{G}} \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \|\phi_{\mathcal{X}}(\tilde{\mathbf{x}}_i) - F(\tilde{\mathbf{z}}_i, \tilde{\mathbf{o}}_i)\|_{\mathcal{H}_X}^2 + \xi \|F\|_{\mathcal{G}}^2,$$

- $\mathcal{G}$ is a vector-valued RKHS which contain mappings from $\mathcal{Z} \times \mathcal{O} \to \mathcal{H}_{\mathcal{X}}$.
  - $\mathcal{G}$ is isometrically isomorphic to the space $S_2(\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{O},1}, \mathcal{H}_{\mathcal{X}})$ of Hilbert-Schmidt operators from $\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{O},1}$ to $\mathcal{H}_{\mathcal{X}}$.
- ξ is a regularization parameter, and $\hat{F}_\xi$ admits a closed-form expression.

Algorithm: Kernel 2SLS (Stage II)

Stage II: We learn f_* with ridge regression.

- Stage II: Learn f_* with $\{(\mathbf{z}_i, \mathbf{o}_i, y_i)\}_{i=1}^n$.

$$\hat{f}_\lambda := \inf_{f \in \mathcal{H}_\mathcal{X} \otimes \mathcal{H}_{\mathcal{O},2}} \lambda \|f\|^2_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_{\mathcal{O},2}} + \frac{1}{n} \sum_{i=1}^n \left(y_i - \left\langle f, \hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i) \otimes \phi_{O,2}(\mathbf{o}_i) \right\rangle_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_{\mathcal{O},2}} \right)^2.$$

- Recall that $\langle f, \phi_{\mathcal{O},2}(\mathbf{o}) \otimes F_*(\mathbf{z}, \mathbf{o}) \rangle_{\mathcal{H}_{\mathcal{O},2} \otimes \mathcal{H}_\mathcal{X}} = (Tf)(\mathbf{z}, \mathbf{o})$.
- λ is a regularization parameter, and $\hat{f}_\lambda$ admits a closed-form expression.

Algorithm ✓ Theory ?

Theory: Learning risk

- The learning risk is

$$\|\hat{f}_\lambda - f_*\|_{L^2(P_{XO})}.$$

Learning rate: How fast $\|\hat{f}_\lambda - f_*\|_{L^2(P_{XO})} \to 0$ as $n \to \infty$?

- Many papers in NPIV only prove learning rate of

$$\|T\hat{f}_\lambda - Tf_*\|_{L^2(P_{ZO})},$$

which is a weaker metric.

- T is a bounded operator.

$$\begin{aligned}
\|Tf\|^2_{L^2(P_{ZO})} &= \mathbb{E}_{ZO}\left[(\mathbb{E}[f(X,O) \mid Z,O])^2\right] \\
&\le \mathbb{E}_{Z,O}\left[\mathbb{E}\left[f(X,O)^2 \mid Z,O\right]\right] \quad \text{(Jensen inequality)} \\
&= \|f\|^2_{L^2(P_{XO})}.
\end{aligned}$$

Assumptions: Smoothness

- Stage I target F_* : $F_*(\mathbf{z},\mathbf{o}) = \mathbb{E}[\phi_{\mathcal{X}}(X) \mid Z=\mathbf{z}, O=\mathbf{o}] = \int \phi_{\mathcal{X}}(\mathbf{x})p(\mathbf{x}\mid \mathbf{z},\mathbf{o})\mathrm{d}\mathbf{x}$.
  - The regularity of F_* is completely decided by the conditional distribution $P_{X|Z,O}$.

Assumption (Conditional distribution)

Let $m_o, m_z \in \mathbb{N}^+$. The map $(\mathbf{z},\mathbf{o}) \mapsto p(\mathbf{x}\mid \mathbf{z},\mathbf{o})$ satisfies:

$$\rho := \max_{|\boldsymbol{\alpha}|\le m_z} \max_{|\boldsymbol{\beta}|\le m_o} \sup_{\mathbf{x}\in\mathcal{X},\mathbf{z}\in\mathcal{Z},\mathbf{o}\in\mathcal{O}} |\partial_{\mathbf{z}}^{\boldsymbol{\alpha}}\partial_{\mathbf{o}}^{\boldsymbol{\beta}} p(\mathbf{x}\mid\mathbf{z},\mathbf{o})| < \infty$$

- Stage II target $f_* : \mathcal{X}\times\mathcal{O}\to\mathbb{R}$.

Assumption (Anisotropic Besov space target)

$f_* \in B^{s_x,s_o}_{2,\infty}(\mathcal{X}\times\mathcal{O}) \cap L^\infty(\mathcal{X}\times\mathcal{O})$.

- Can be extended to allow more anisotropic smoothness within X and O.
- The regularity of f_* and F_* on O might be different: justifies two kernels $k_{\mathcal{O},1}, k_{\mathcal{O},2}$.

Assumption: Partial smoothing effect of T

Assumption (Completeness)

For all functions $f \in L^2(P_{XO})$, $(Tf)(Z, O) = \mathbb{E}[f(X, O) \mid Z, O] = 0$ implies that $f(X, O) = 0$ almost surely.

- Guarantees the uniqueness of f_*.
- For non-asymptotic convergence, we need stronger assumptions on T.

Definition (Partial Fourier transform)

For a function $f : \mathcal{X} \times \mathcal{O} \to \mathbb{R}$ such that $f(\cdot, \mathbf{o}) \in L^1(\mathbb{R}^{d_x})$ for any $\mathbf{o} \in \mathcal{O}$, we define its partial Fourier transform as

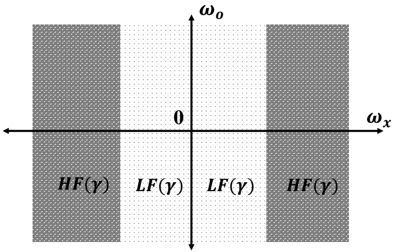
$$\mathcal{F}_x[f](\boldsymbol{\omega}_x, \mathbf{o}) = \int_{\mathbb{R}^{d_x}} f(\mathbf{x}, \mathbf{o}) \exp(-i\langle \mathbf{x}, \boldsymbol{\omega}_x \rangle) \, \mathrm{d}\mathbf{x}.$$

Assumption: Partial smoothing effect of T

For any scalar $\gamma \in (0, 1)$, we define the following two sets of functions:

$$\mathrm{LF}(\gamma) := \left\{ f : \mathbb{R}^{d_x + d_o} \to \mathbb{R} \;\middle|\; \forall \mathbf{o} \in \mathcal{O},\ \operatorname{supp}\big(\mathcal{F}_{\mathbf{x}}[f(\cdot, \mathbf{o})]\big) \subseteq \{\boldsymbol{\omega}_x \in \mathbb{R}^{d_x} : \|\boldsymbol{\omega}_x\|_2 \leq \gamma^{-1}\} \right\}.$$

$$\mathrm{HF}(\gamma) := \left\{ f : \mathbb{R}^{d_x + d_o} \to \mathbb{R} \;\middle|\; \forall \mathbf{o} \in \mathcal{O},\ \operatorname{supp}\big(\mathcal{F}_{\mathbf{x}}[f(\cdot, \mathbf{o})]\big) \subseteq \{\boldsymbol{\omega}_x \in \mathbb{R}^{d_x} : \|\boldsymbol{\omega}_x\|_2 \geq \gamma^{-1}\} \right\}.$$



We consider partial Fourier transform due to the partial identity structure of T.

Assumption: Partial smoothing effect of T

Assumption (Fourier measure of partial contractivity of T)

$\exists c_1 > 0$ *and* $\exists \eta_1 \in [0, \infty)$, *such that* $\forall \gamma \in (0,1)$ *and* $\forall f \in \mathrm{HF}(\gamma) \cap L^\infty(P_{XO})$*:*

$$\|Tf\|_{L^2(P_{ZO})} \leq c_1 \gamma^{d_x \eta_1} \|f\|_{L^2(P_{XO})}.$$

Assumption (Fourier measure of partial ill-posedness of T)

$\exists c_0 > 0$ *and* $\exists \eta_0 \in [0, \infty)$, *such that* $\forall \gamma \in (0,1)$ *and* $\forall f \in \mathrm{LF}(\gamma) \cap L^\infty(P_{XO})$*:*

$$\|Tf\|_{L^2(P_{ZO})} \geq c_0 \gamma^{d_x \eta_0} \|f\|_{L^2(P_{XO})}.$$

- η_1 quantifies the *partial smoothing* effect of T on a function f's high-frequency components with respect to X.
- η_0 quantifies the *partial anti-smoothing* effect of T on a function f's low-frequency components with respect to X.
- We construct an example of T when both assumptions hold. These assumptions are hard to verify in practice.

Assumption: Connection to RKHS

Question: What is the connection of Fourier smoothing and RKHS?

- An Gaussian RKHS with lengthscale γ_x can be defined through Fourier transform:

$$\mathcal{H}_{\mathcal{X},\gamma_x} = \left\{ f : \mathbb{R}^{d_x} \to \mathbb{R} \left| \int_{\mathbb{R}^{d_x}} \left| \mathcal{F}_{\mathbf{x}}[f]\left(\boldsymbol{\omega}_x\right) \right|^2 \exp\left(\gamma_x^2 \left\|\boldsymbol{\omega}_x\right\|_2^2\right) \mathrm{d}\boldsymbol{\omega}_x < \infty \right. \right\},$$

- For $f \in \mathcal{H}_{X,\gamma_x}$, the bulk of its Fourier spectrum $\mathcal{F}_{\mathbf{x}}[f](\boldsymbol{\omega}_x)$ would belong to the ball $\{\boldsymbol{\omega}_x : \|\boldsymbol{\omega}_x\|_2 \leq \gamma_x^{-1}\}$ with remaining spectrum decaying exponentially as $\boldsymbol{\omega}_x \to \infty$.
- We employ Gaussian kernels in stage II:

$$k_{\gamma_x}(\mathbf{x}, \mathbf{x}') = \exp\left(-\sum_{j=1}^{d_x} \frac{(x_j - x_j')^2}{\gamma_x^2}\right), \quad k_{\gamma_o}(\mathbf{o}, \mathbf{o}') = \exp\left(-\sum_{j=1}^{d_o} \frac{(o_j - o_j')^2}{\gamma_o^2}\right).$$

- The kernel lengthscales γ_x, γ_o are tuned adaptive to the anisotropic smoothness of f_*.

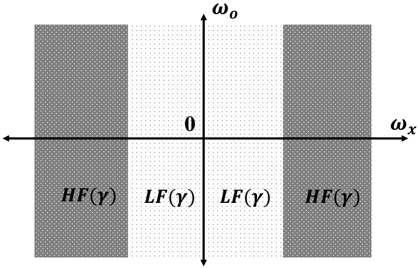
Challenges of NPIV-O revisited

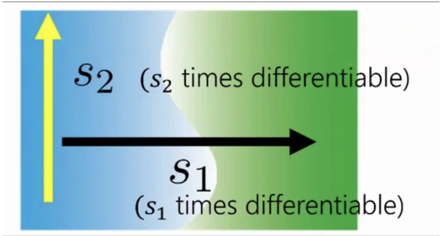
Challenge One: Partial identity of T.

Solution One: Fourier measure of partial smoothing and ill-posedness of T

Challenge Two: Anisotropic smoothness of f_*.

Solution Two: Gaussian kernel lengthscales adaptive to s_x, s_o.





Assumptions: Data generating distribution

Assumption (Upper and lower bounded marginal densities)

The joint probability measures P_{ZO} and P_{XO} admit probability density functions p_{ZO} and p_{XO}. There exists a universal constant $a > 0$ such that $a^{-1} \geq p_{ZO}(\mathbf{z}, \mathbf{o}) \geq a$ for all $(\mathbf{z}, \mathbf{o}) \in [0,1]^{d_z+d_o}$ and $a^{-1} \geq p_{XO}(\mathbf{x}, \mathbf{o}) \geq a$ for all $(\mathbf{x}, \mathbf{o}) \in [0,1]^{d_x+d_o}$.

- Standard assumptions for Besov spaces.

Assumption (Subgaussian noise)

$\forall(\mathbf{z}, \mathbf{o}) \in \mathcal{Z} \times \mathcal{O}$, the residual $\upsilon := Y - (Tf_*)(Z, O)$ is σ-subgaussian conditioned on $Z = \mathbf{z}, O = \mathbf{o}$.

- Standard assumptions for high probability upper bound.

Theory: Upper bounds

1. Suppose all assumptions hold.
2. We choose stage I kernels $k_{\mathcal{O},1}$ and $k_{\mathcal{Z}}$ are Matérn kernels whose associated RKHS $\mathcal{H}_O$ and $\mathcal{H}_Z$ are equivalent to $W_2^{m_o}(\mathcal{O})$ and $W_2^{m_z}(\mathcal{Z})$. Define $d^\dagger = (d_z m_z^{-1}) \vee (d_o m_o^{-1})$.
3. Suppose stage II kernels k_{X,γ_x} and k_{O,γ_o} are Gaussian kernels with lengthscales

$$\gamma_x = n^{-\frac{\frac{1}{d_x}}{1+2(\frac{s_x}{d_x}+\eta_1)+\frac{d_o}{s_o}(\frac{s_x}{d_x}+\eta_1)}}, \quad \gamma_o = n^{-\frac{\frac{1}{s_o}(\frac{s_x}{d_x}+\eta_1)}{1+2(\frac{s_x}{d_x}+\eta_1)+\frac{d_o}{s_o}(\frac{s_x}{d_x}+\eta_1)}}.$$

4. Stage I regularization $\xi = \tilde{n}^{-\frac{1}{1+d^\dagger}}$ and stage II regularization $\lambda = n^{-1}$.
5. Stage I sample size satisfies $\tilde{n} \gtrsim n^{2+d^\dagger}$

Then, we have with high probability,

$$\left\|\hat{f}_\lambda - f_*\right\|_{L^2(P_{XO})} \lesssim n^{-\frac{\frac{s_x}{d_x}+\eta_1-\eta_0}{1+2(\frac{s_x}{d_x}+\eta_1)+\frac{d_o}{s_o}(\frac{s_x}{d_x}+\eta_1)}} \cdot (\log n)^{\frac{d_x+d_o+1+d_x\eta_0}{2}}.$$

Theory: Upper bounds

$$\text{Upper Bound:} \quad \tilde{\mathcal{O}}_P\left(n^{-\frac{\frac{s_x}{d_x}+\eta_1-\eta_0}{1+2(\frac{s_x}{d_x}+\eta_1)+\frac{d_o}{s_o}\frac{s_x}{d_x}+\frac{d_o}{s_o}\eta_1}}\right).$$

- We take $\eta_1 = \eta_0 = \eta$ such that we have a precise characterization of the partial smoothing effect of T.
- Our derived upper rate interpolates between the known optimal L^2-rates of
  - NPR ($\eta_0 = \eta_1 = 0$), the upper bound simplifies to $\tilde{\mathcal{O}}_P(n^{-\frac{1}{2\tilde{s}+1}})$ with $\tilde{s} = (d_o/s_o + d_x/s_x)^{-1}$ being the intrinsic smoothness.
  - NPIV ($d_o = 0$ and $\eta_0 = \eta_1 = \eta > 0$), our upper bound simplifies to $\tilde{\mathcal{O}}_P(n^{-\frac{s_x}{d_x+2(s_x+\eta d_x)}})$.
- We consider sufficiently large stage I sample size $\tilde{n}$, a regime regime where we can study rate-optimality with respect to n.

Proof sketch

- Recall $\hat{f}_\lambda$ and define $\bar{f}_\lambda$:

$$\bar{f}_\lambda := \underset{f \in \mathcal{H}_{\gamma_X,\gamma_O}}{\arg\min}\, \lambda \|f\|^2_{\mathcal{H}_{\gamma_X,\gamma_O}} + \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \langle f, F_*(\mathbf{z}_i,\mathbf{o}_i) \otimes \phi_{O,\gamma_o}(\mathbf{o}_i)\rangle_{\mathcal{H}_{\gamma_X,\gamma_O}}\right)^2.$$

$$\hat{f}_\lambda := \underset{f \in \mathcal{H}_{\gamma_X,\gamma_O}}{\arg\min}\, \lambda \|f\|^2_{\mathcal{H}_{\gamma_X,\gamma_O}} + \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left\langle f, \hat{F}_\xi(\mathbf{z}_i,\mathbf{o}_i) \otimes \phi_{O,\gamma_o}(\mathbf{o}_i)\right\rangle_{\mathcal{H}_{\gamma_X,\gamma_O}}\right)^2.$$

- The proof can be divided into three steps:
  - Step 1: Bound $\|T\hat{f}_\lambda - T\bar{f}_\lambda\|_{L^2(P_{ZO})}$. This is the stage I error.
  - Step 2: Bound $\|T\bar{f}_\lambda - Tf_*\|_{L^2(P_{ZO})}$. This is the stage II error.
  - Combine the above to bound $\|T\hat{f}_\lambda - Tf_*\|_{L^2(P_{ZO})}$.
  - Step 3: Apply partial Fourier condition to obtain the bound $\|\hat{f}_\lambda - f_*\|_{L^2(P_{XO})}$.

Proof sketch: Step 1 and Step 2

- Stage I error: $\|T\hat{f}_\lambda - T\bar{f}_\lambda\|_{L^2(P_{ZO})}$ can be translated to a bound that involves $\|\hat{F}_\xi - F_*\|_{L^2}$ and $\|\hat{F}_\xi - F_*\|_{\mathcal{G}}$.
  - Optimal rate of conditional mean embedding.
- Stage II error: $\|T\bar{f}_\lambda - Tf_*\|_{L^2(P_{ZO})}$ can be translated to a generalization error of kernel ridge regression with a new RKHS $\mathcal{H}_{FO}$.

$$\mathcal{H}_{FO} = \left\{ f : \mathcal{Z} \times \mathcal{O} \to \mathbb{R} \mid \exists w \in \mathcal{H}_{\gamma_x,\gamma_o}, f(\mathbf{z},\mathbf{o}) \equiv \langle w, F_*(\mathbf{z},\mathbf{o}) \otimes \phi_{\gamma_o}(\mathbf{o}) \rangle_{\mathcal{H}_{\gamma_x,\gamma_o}} \right\}$$

- Denote

$$f_\lambda := \underset{f \in \mathcal{H}_{\gamma_x,\gamma_o}}{\arg\min} \lambda \|f\|^2_{\mathcal{H}_{\gamma_x,\gamma_o}} + \|T(f - f_*)\|^2_{L^2(P_{ZO})} = \underset{h \in \mathcal{H}_{FO}}{\arg\min} \lambda \|h\|^2_{\mathcal{H}_{FO}} + \|h - Tf_*\|^2_{L^2(P_{XO})}$$

  - Estimation error $\|T\bar{f}_\lambda - Tf_\lambda\|_{L^2(P_{ZO})}$. We use Fischer and Steinwart [2020].
  - Approximation error $\|Tf_\lambda - Tf_*\|_{L^2(P_{ZO})}$. We use Hang and Steinwart [2021].

Theory: Lower bounds

For all learning methods $D \mapsto \hat{f}_D$ ($D = (\mathbf{z}_i, \mathbf{x}_i, \mathbf{o}_i, y_i)_{i=1}^n$), $\forall \tau > 0$, and sufficiently large $n \geq 1$, there exists a distribution P over (Z, X, O, Y) inducing a NPIV-O model

$$Y = f_*(X, O) + \epsilon, \quad \mathbb{E}[\epsilon | Z, O] = 0 \ ,$$

such that all assumptions in the upper bound are satisfied, and with high probability,

$$\left\| \hat{f}_D - f_* \right\|_{L^2(P_{XO})} \gtrsim n^{-\frac{\frac{s_X}{d_X}}{1+2(\frac{s_X}{d_X}+\eta_1)+\frac{d_O}{s_O}\frac{s_X}{d_X}}} (\log n)^{-d_X}.$$

Lower Bounds

$$\text{Lower Bound:} \quad \tilde{\mathcal{O}}_P\left(n^{-\frac{\frac{s_X}{d_X}}{1+2\left(\frac{s_X}{d_X}+\eta\right)+\frac{d_O}{s_O}\frac{s_X}{d_X}}}\right), \quad \text{Upper Bound:} \quad \tilde{\mathcal{O}}_P\left(n^{-\frac{\frac{s_X}{d_X}}{1+2\left(\frac{s_X}{d_X}+\eta\right)+\frac{d_O}{s_O}\frac{s_X}{d_X}+\frac{d_O}{s_O}\eta}}\right).$$

- Both interpolate between the known optimal L^2-rates for NPIV ($d_o = 0$) and anisotropic kernel ridge regression ($\eta = 0$).
- There exists a gap $\frac{d_o}{s_o}\eta$ between the upper and lower bounds.
- We hypothesize that the gap is an inherent limitation of the kernel 2SLS algorithm.
  - For upper bound, we set $\gamma_X^{s_x+d_x\eta} = \gamma_O^{s_o}$. For lower bound, we set $\gamma_X^{s_x} = \gamma_O^{s_o}$.

Conclusions

- Theoretical challenges posed by the presence of observed covariates.
  - **Challenge 1: Anisotropic smoothness.**
    Solution: adaptive kernel lengthscales.
  - **Challenge 2: Partial smoothing by the operator T.**
    Solution: a novel Fourier-based characterization of the partial smoothing effect of T.
- We prove an upper bound for kernel 2SLS and the first minimax lower bound.
- We identify a gap between our bounds which we posit is fundamental to kernel 2SLS.

About Me

- Zonghao Chen
- 4th year PhD Student at University College London (UCL)
  - Foundational AI Centre
  - Gatsby Computational Neuroscience Unit
- Graduated from Tsinghua University in 2022
  - Department of EE
- Kernel (nonparametric) methods, causal inference, statistical learning theory

S. Fischer and I. Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21(205):1–38, 2020.

H. Hang and I. Steinwart. Optimal learning with anisotropic gaussian svms. *Applied and Computational Harmonic Analysis*, 55:337–367, 2021.