Nonparametric Instrumental Variable Regression with Observed Covariates

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

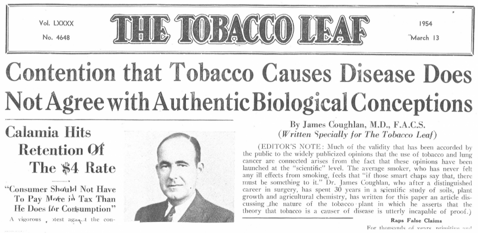
May 28, 2025

Submitted (soon) to *Annals of Statistics*

*, † Equal contribution in alphabetical order

Background: Causal Inference

- The causal effect of smoking X on the risk of lung cancer Y.
- Unobserved confounding ϵ that affects both X and Y: gene, occupation, childhood.
- It takes long for scientific community to agree that smoking increases risk of lung cancer.


Vol. LXXXX No. 4648 THE TOBACCO LEAF 1954 March 13

Contention that Tobacco Causes Disease Does Not Agree with Authentic Biological Conceptions

Calamia Hits Retention Of The $4 Rate

"Consumer Should Not Have To Pay More in Tax Than He Does for Consumption"

By James Coughlan, M.D., F.A.C.S.
(*Written Specially for The Tobacco Leaf*)

(EDITOR'S NOTE: Much of the validity that has been accorded by the public to the widely publicized opinions that the use of tobacco and lung cancer are connected arises from the fact that these opinions have been launched at the "scientific" level. The average smoker, who has never felt any ill effects from smoking, feels that "if those smart chaps say that, there must be something to it." Dr. James Coughlan, who after a distinguished career in surgery, has spent 30 years in a scientific study of soils, plant growth and agricultural chemistry, has written for this paper an article discussing the nature of the tobacco plant in which he asserts that the theory that tobacco is a causer of disease is utterly incapable of proof.)

Raps False Claims

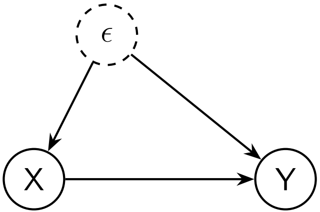
How to do causal inference with unobserved confounders ϵ?

Background: Instrumental Variable

- In this talk, we only consider *additive* confounding.

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

- f_* is the target of interest.
    - Dose response curve, causal parameter, potential outcome, structural function, etc.
- Regression only recovers the conditional mean $\mathbb{E}[Y \mid X] \neq f_*(X)$ and always outputs a *biased* estimate.



Background: Instrumental Variable

- Instrumental variables Z affect Y only through X and is independent of ϵ.
    - A valid instrumental variable could be 'price of cigarette'.

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

- Conditioning on both sides

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

- P is the joint data distribution over Z, X, Y. P_Z, P_X, P_Y denote the marginals.
- In fact an ill-posed inverse problem

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

- where T is a conditional expectation operator.

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

- Ill-posedness: T is compact so T^{-1} is unbounded.

Background: Instrumental Variable

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

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

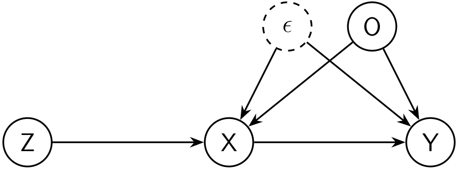
- An ill-posed inverse problem

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

- where T is a conditional expectation operator.

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

- We focus on *nonparametric* instrumental variable with observed covariates (NPIV-O).



Setting: NPIV-O

- The observed covariates O brings two advantages
  - Practitioners adjust for as many observed covariates as possible.
    - 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.
- 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 structural dichotomy of T between a compact operator and an identity operator.

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

- $T^*T|_{\mathfrak{G}_A}$ (T^*T restricted to $\mathfrak{G}_A$) is compact; $T^*T|_{\mathfrak{G}_B}$ is an identity operator.
  - Partial smoothing effect of T.
- Existing work on NPIV relies on the compactness of T.
  - Stratification on O is a simple yet statistically inefficient fix.

Contributions

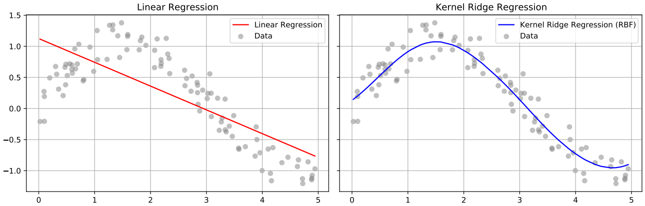
- We adapt an existing algorithm kernel 2SLS to observed covariates.
- For challenge a), we tune kernel lengthscales so that the algorithm adapts to the anisotropic smoothness of f_*.
- For challenge b), we introduce a novel Fourier measure of *partial* smoothing effect of T.
- We prove upper learning rates for kernel 2SLS and the first minimax lower learning rates for NPIV-O.
- Our analysis can be applied to an emerging field of proximal causal inference.

$$\mathbb{E}[Y \mid Z, X] = (Tf_*)(W, X), \text{ with } T : L^2(P_{ZX}) \to L^2(P_{WX}).$$

  - W, Z are proxy variables.

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

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

- Ill-posed inverse problem with $T : L^2(P_{XO}) \to L^2(P_{ZO})$:

$$Y = (Tf_*)(Z, O) + \upsilon, \quad (Tf)(\mathbf{z}, \mathbf{o}) = \mathbb{E}[f(X, O) \mid Z = \mathbf{z}, O = \mathbf{o}].$$

- 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

Main challenge: Unknown conditional expectation operator T!

- Conditional mean embedding:

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

  - It is a Hilbert-space valued integral (Bochner integral).
- $\forall f \in \mathcal{H}_{\mathcal{O},2} \otimes \mathcal{H}_{\mathcal{X}}$, we have

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

- F_* is a kernel analogue of conditional expectation operator T.

Algorithm: Kernel 2SLS

- 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 $\mathcal{Z} \times \mathcal{O} \rightarrow \mathcal{H}_{\mathcal{X}}$.
- $\mathcal{G}$ 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.
- $\hat{F}_\xi$ admits a closed-form expression.

Algorithm: Kernel 2SLS

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

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

- λ is a regularization parameter.
- $\hat{f}_\lambda$ admits a closed-form expression.
- We employ Gaussian kernels

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

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

Theory: Learning risk

- The learning risk is

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

- 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.

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

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:

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

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

Assumption (Anisotropic Besov space target)

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

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

Assumption: Partial smoothing effect of T

Assumption (Completeness)

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

- Identification: $\mathcal{N}(T)^{\perp} = \{0\}$.
- For non-asymptotic convergence, we need stronger assumptions that characterize the degree of smoothing of 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:

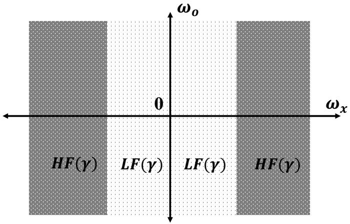
$$\mathrm{LF}(\gamma) := \left\{ f : \mathbb{R}^{d_x+d_o} \to \mathbb{R} \;\middle|\; \forall \mathbf{o} \in \mathcal{O},\ f(\cdot,\mathbf{o}) \in L^1(\mathbb{R}^{d_x}), \right.$$

$$\left. \operatorname{supp}\big(\mathcal{F}_{\mathbf{x}}[f(\cdot,\mathbf{o})]\big) \subseteq \left\{ \boldsymbol{\omega}_x \in \mathbb{R}^{d_x} : \|\boldsymbol{\omega}_x\|_2 \leq \gamma^{-1} \right\} \right\}.$$

$$\mathrm{HF}(\gamma) := \left\{ f : \mathbb{R}^{d_x+d_o} \to \mathbb{R} \;\middle|\; \forall \mathbf{o} \in \mathcal{O},\ f(\cdot,\mathbf{o}) \in L^1(\mathbb{R}^{d_x}), \right.$$

$$\left. \operatorname{supp}\big(\mathcal{F}_{\mathbf{x}}[f(\cdot,\mathbf{o})]\big) \subseteq \left\{ \boldsymbol{\omega}_x \in \mathbb{R}^{d_x} : \|\boldsymbol{\omega}_x\|_2 \geq \gamma^{-1} \right\} \right\}.$$

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* $\forall f \in \mathrm{HF}(\gamma) \cap L^\infty(P_{XO})$:

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

- It quantifies the *partial smoothing* effect of T on a function f's high-frequency components with respect to X.

Assumption: Partial smoothing effect of T

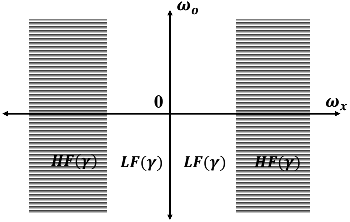
Assumption (Fourier measure of partial ill-posedness of T)

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

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

- It captures the *partial anti-smoothing* behaviour of T on a function f's low frequency components with respect to X.
- We set the constants $c_0 = c_1 = 1$ for simplicity.
- $\eta_0 \geq \eta_1$.
- These assumptions hold for Fourier series (not Fourier transforms!) when $\{e_n(\mathbf{x}) = \exp(i2n\pi\mathbf{x})\}_{n \geq 1}$ are the eigenbasis for T^*T and the eigenvalues of T^*T decay polynomially.
- These assumptions are hard to verify in practice.
    - Link conditions, sieve measure of ill-posedness, etc.

Assumption: Connection to RKHS



- An Gaussian RKHS can be defined through Fourier transform:

$$\mathcal{H}_{\mathcal{X},\gamma_x} = \left\{ f : \mathbb{R}^{d_x} \to \mathbb{R} \,\middle|\, \int_{\mathbb{R}^{d_x}} \left| \mathcal{F}_{\mathbf{x}}[f]\left(\boldsymbol{\omega}_x\right) \right|^2 \exp\left(\frac{1}{4} \gamma_x^2 \left\| \boldsymbol{\omega}_x \right\|_2^2 \right) \mathrm{d}\boldsymbol{\omega}_x < \infty \right\},$$

- 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 formulate our Assumptions with Fourier transforms for its generality.

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. Suppose stage I kernels $k_{\mathcal{O}}$ and $k_{\mathcal{Z}}$ are Matérn kernels whose associated RKHS $\mathcal{H}_O$ and $\mathcal{H}_Z$ are norm 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 $\lambda = n^{-1}$.

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 for NPIV without observed covariates and anisotropic kernel ridge regression.
    - When $\eta_0 = \eta_1 = 0$, the upper bound simplifies to $\tilde{\mathcal{O}}_P(n^{-\frac{1}{2\tilde{s}+1}})$ with $\tilde{s} = (d_o/s_o + d_x/s_x)^{-1}$ beings the intrinsic smoothness, which matches the known optimal rate in NPR.
    - When $d_o = 0$ and $\eta_0 = \eta_1 > 0$, our upper bound simplifies to $\tilde{\mathcal{O}}_P(n^{-\frac{s_x}{d_x+2(s_x+\eta d_x)}})$, which matches the known optimal rate in NPIV.

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

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

- The lower bounds also interpolate between the known optimal L^2-rates for NPIV without observed covariates and anisotropic kernel ridge regression, same as the upper bounds.
- There exists a gap between the upper and lower bounds even when $\eta_1 = \eta_0$.
- We hypothesize that the gap is an inherent limitation of the kernel 2SLS algorithm.

Conclusions

- Presence of observed covariates pose additional theoretical challenges in NPIV-O.
  - a) Anisotropic smoothness
  - b) Partial smoothing effect of T
- To tackle challenge a), we modify the existing kernel 2SLS to observed covariates with adaptive kernel lengthscales.
- To tackle challenge b), we propose a novel Fourier measure of partial smoothing effect.
- We prove an upper bound for kernel 2SLS and the first minimax lower bound for NPIV-O.
- We identify a gap between our bounds which we posit is fundamental to kernel 2SLS.

More About Me

- 3rd year PhD Student at UCL, AI Centre and Gatsby Unit
- Graduated from THUEE in 2022
- Kernel (nonparametric) methods, causal inference, statistical learning theory
- Visiting RIKEN AIP now (Summer 2025)