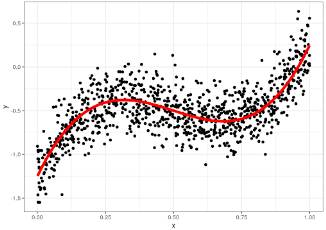
Towards a Unified Analysis of Neural Networks in Nonparametric Instrumental Variable Regression: Optimization and Generalization

Zonghao Chen Atsushi Nitanda Arthur Gretton Taiji Suzuki

October 30, 2025

Background

- Nonparametric regression is used everywhere in statistics.



- Regression fails when ...
  - there exist unobserved confounding

Background: Causal inference and Instrumental Variable

- 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.

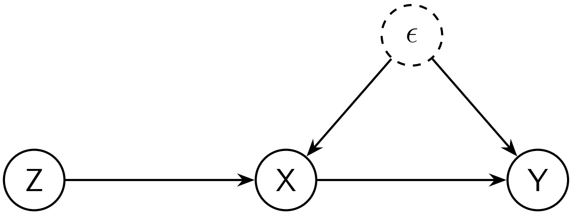
$$Y = h_\circ(X) + \epsilon, \qquad \text{where } \epsilon \not\perp X.$$

- Instrumental variable Z that affects Y only through X: price of the cigarette.

$$\epsilon \perp Z$$

- Conditioning both sides on Z

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



Background: NPIV versus NPR

- Nonparametric regression (NPR): $Y = h_\circ(X) + \epsilon$ with $\epsilon \perp X$.
  - Conditioning both sides on X:
$$\mathbb{E}[Y \mid X] = h_\circ(X). \tag{NPR}$$
  - Target $h_\circ = \arg\min_h \mathbb{E}_{YX}[(Y - h(X))^2]$.
  - Least squares estimator
$$\theta^* = \arg\min_\theta \frac{1}{n}\sum_{i=1}^{n}(h_\theta(\mathbf{x}_i) - \mathbf{y}_i)^2, \quad \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n \sim P_{XY}.$$
  - h_θ is a neural network parameterized by θ.
- Nonparametric instrumental variable regression (NPIV): $Y = h_\circ(X) + \epsilon$ with $\epsilon \not\perp X$ but $\epsilon \perp Z$.
  - $\mathbb{E}[Y \mid X] = h_\circ(X) + \mathbb{E}[\epsilon \mid X] \neq h_\circ(X)$.
  - Conditioning both sides on Z:
$$\mathbb{E}[Y \mid Z] = \mathbb{E}[h_\circ(X) \mid Z]. \tag{NPIV}$$

Problem: How to estimate $h_\circ$?

Background: NPIV and 2SLS

- Define $T : L^2(P_X) \to L^2(P_Z)$ as the unknown conditional expectation operator defined by $(Tf)(Z) = \mathbb{E}[f(X) \mid Z]$.

$$\mathbb{E}[Y \mid Z] = \mathbb{E}[h_\circ(X) \mid Z] =: (Th_\circ)(Z). \tag{NPIV}$$

- Target $h_\circ = \arg\min_h \mathbb{E}_{YZ}[(Y - (Th)(Z))^2]$.
  - T is unknown yet it is a conditional expectation and hence can be learned via regression.
- Two-stage least squares (2SLS)

$$\begin{aligned} \theta_z^*(\theta_x) &= \arg\min_{\theta_z} \frac{1}{m} \sum_{i=1}^{m} (h_{\theta_z}(\mathbf{z}_i) - h_{\theta_x}(\mathbf{x}_i))^2 , \quad \{\mathbf{x}_i, \mathbf{z}_i\}_{i=1}^m \sim P_{XZ} \\ \theta_x^* &= \arg\min_{\theta_x} \frac{1}{n} \sum_{i=1}^{n} (h_{\theta_z^*(\theta_x)}(\mathbf{z}_i) - \mathbf{y}_i)^2, \quad \{\mathbf{z}_i, \mathbf{y}_i\}_{i=1}^n \sim P_{ZY}. \end{aligned} \tag{2SLS}$$

- $h_{\theta_z^*(\theta_x)}(\mathbf{z}_i) \approx \mathbb{E}[h_{\theta_x}(X) \mid Z = \mathbf{z}_i] = (Th_{\theta_x})(\mathbf{z}_i)$.

Background: Offline reinforcement learning

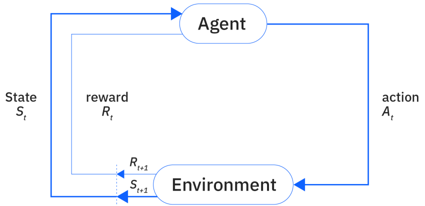
- s: state, a: action, r: reward, γ: discount factor.
- $Q(s, a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_0 = a]$ denotes the expected long-term return when taking action a in state s.
- Bellman equation:

$$\mathbb{E}[r \mid s, a] = Q(s, a) - \gamma \mathbb{E}[Q(s', a') \mid s, a].$$

- The Bellman equation shars the same structure as NPIV:

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

- Correspondence: $Y \to r$, $Z \to (s, a)$, $X \to (s', a')$.



Target: Optimization and Generalization

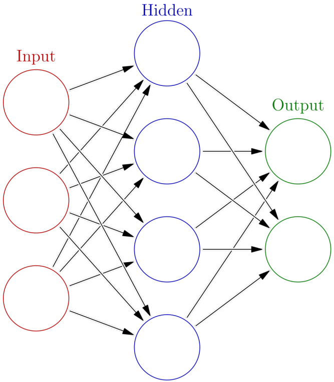
- Two-stage least squares (2SLS)
  - Both h_{θ_x} and h_{θ_z} are neural networks.

$$
\begin{aligned}
&\text{Stage I:} \quad \theta_z^*(\theta_x) = \arg\min_{\theta_z} \frac{1}{m} \sum_{i=1}^{m} \left(h_{\theta_z}(\mathbf{z}_i) - h_{\theta_x}(\mathbf{x}_i)\right)^2, \quad \{\mathbf{x}_i, \mathbf{z}_i\}_{i=1}^m \sim P_{XZ} \\
&\text{Stage II:} \quad \theta_x^* = \arg\min_{\theta_x} \frac{1}{n} \sum_{i=1}^{n} \left(h_{\theta_z^*(\theta_x)}(\mathbf{z}_i) - \mathbf{y}_i\right)^2, \quad \{\mathbf{z}_i, \mathbf{y}_i\}_{i=1}^n \sim P_{ZY}.
\end{aligned}
\tag{2SLS}
$$

Bilevel optimization theory: Does gradient based algorithm can actually find the global optimum θ_z^*, θ_x^*? If it does, what is the iteration complexity?

Statistical theory: Given the global optimal θ_z^*, θ_x^*, is $h_{\theta_x^*}$ a consistent estimator of $h_\circ$? If it does, what is the sample complexity?

Mean-field neural networks



Background: Mean-field two-layer neural networks

- Consider neural networks with a single hidden layer of size N:
  - $\mathscr{X} = [x^{(1)}, \dots, x^{(N)}] \in (\mathbb{R}^{d_x})^N$ are the network parameters and $\mathbf{x}$ is the network input

$$h(\mathbf{x}, \mathscr{X}) = \frac{1}{N} \sum_{i=1}^{N} \Psi(\mathbf{x}, x^{(i)})$$

- Here, $\Psi(\mathbf{x}, x) = w_2 a(w_1^\top \mathbf{x} + b)$ with parameters $x = (w_1, w_2, b)$ and a being an activation function.
- As the empirical distribution $\frac{1}{N} \sum_{i=1}^{N} \delta_{x^{(i)}} \to \mu$ as $N \to \infty$:

$$h_\mu(\mathbf{x}) = \int \Psi(\mathbf{x}, x) \mathrm{d}\mu(x) = \mathbb{E}_{X \sim \mu}[\Psi(\mathbf{x}, X)].$$

  - So called "mean-field neural network".

Question: What is the purpose of considering the mean-field limit?

Background: Mean-field perspective of two-layer neural networks

- Consider the squared loss with ℓ_2-norm regularization

$$F(\mu) := \frac{1}{2}\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\rho}\left[(\mathbb{E}_{X\sim\mu}[\Psi(\mathbf{x},X)] - \mathbf{y})^2\right] + \frac{\zeta}{2}\mathbb{E}_{X\sim\mu}[\|X\|^2],$$

- ρ is a data distribution, e.g. $\rho = \frac{1}{n}\sum_{i=1}^{n}\delta_{(\mathbf{x}_i,\mathbf{y}_i)}$.
- F is linear convex in μ: for any probability measures $\mu, \nu \in \mathcal{P}$,

$$F(\vartheta\mu + (1-\vartheta)\nu) \leq \vartheta F(\mu) + (1-\vartheta)F(\nu), \quad \forall\vartheta \in (0,1).$$

  - Optimization problem in $\mathcal{P}$: Wasserstein gradient flow!

Definition (Wasserstein gradient)

The first variation δG of $G : \mathcal{P} \to \mathbb{R}$ at $\mu \in \mathcal{P}$ is defined as a functional $\mathcal{P} \times \mathbb{R}^d \to \mathbb{R}$ that satisfies $\lim_{\epsilon\to 0}\epsilon^{-1}G(\epsilon\nu + (1-\epsilon)\mu) = \int \delta G(\mu)(x)\mathrm{d}(\nu - \mu)$ for any $\nu \in \mathcal{P}(\mathbb{R}^d)$. The Wasserstein gradient $\boldsymbol{\nabla} G$ of $G : \mathcal{P} \to \mathbb{R}$ at $\mu \in \mathcal{P}$ is defined as a functional $\mathcal{P} \times \mathbb{R}^d \to \mathbb{R}^d$ that satisfies $\boldsymbol{\nabla} G(\mu)(\cdot) = \nabla\delta G(\mu)(\cdot)$.

Background: Mean-field perspective of two-layer neural networks

- Wasserstein gradient of F at $\mu \in \mathcal{P}$ evaluated at $x \in \mathbb{R}^d$.

$$\boldsymbol{\nabla} F(\mu)(x) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\rho}\left[(\mathbb{E}_{X\sim\mu}[\Psi(\mathbf{x}, X)] - \mathbf{y})\nabla\Psi(\mathbf{x}, x)\right] + \zeta x.$$

- Wasserstein gradient of F at $\mu_{\mathscr{X}} = \frac{1}{N}\sum_{i=1}\delta_{x^{(i)}}$ evaluated at $x^{(i)} \in \mathbb{R}^d$.

$$\boldsymbol{\nabla} F(\mu_{\mathscr{X}})(x^{(i)}) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\rho}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\Psi(\mathbf{x}, x^{(i)}) - \mathbf{y}\right)\nabla\Psi(\mathbf{x}, x^{(i)})\right] + \zeta x^{(i)}.$$

- Euclidean gradient of the loss
$L(\mathscr{X}) := \frac{1}{2}\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\rho}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\Psi(\mathbf{x}, x^{(i)}) - \mathbf{y}\right)^2\right] + \frac{\zeta}{2}\frac{1}{N}\sum_{i=1}^{N}[\|x^{(i)}\|^2],.$

$$\nabla_{x^{(i)}} L(\mathscr{X}) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\rho}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\Psi(\mathbf{x}, x^{(i)}) - \mathbf{y}\right)\frac{\nabla\Psi(\mathbf{x}, x^{(i)})}{N}\right] + \frac{\zeta}{N}x^{(i)}$$

  - The Wasserstein gradient descent $x_{s+1}^{(i)} = x_s^{(i)} - \gamma\boldsymbol{\nabla} F(\mu_{\mathscr{X},s})(x_s^{(i)})$ coincides with the Euclidean gradient descent $x_{s+1}^{(i)} = x_s^{(i)} - \gamma\nabla_{x^{(i)}} L(\mathscr{X}_s)$ with a rescaled learning rate!

Background: Mean-field perspective of two-layer neural networks

- At iteration $s \in \{0, \dots, S\}$ and for any $i \in \{1, \dots, N\}$:

$$x_{s+1}^{(i)} = x_s^{(i)} - \gamma \nabla F(\mu_{\mathscr{X}})(x_s^{(i)}) + \sqrt{2\sigma\gamma}\, \xi_s^{(i)}.$$

  - $\{\xi_s^{(i)}\}_{i=1}^N$ are N i.i.d samples from d dimensional unit Gaussian.
- Define an entropic regularized objective $\mathscr{F}(\mu) = F(\mu) + \sigma \mathrm{Ent}(\mu)$.
  - $\mathrm{Ent}(\mu) = \int \log \mu(x)\mu(x)\mathrm{d}x$.
  - Noisy gradient descent is Wasserstein gradient descent of $\mathscr{F}$.
- The global optimum $\mu^* := \arg\min_\mu \mathscr{F}(\mu)$.

Question: Does noisy gradient descent can actually find the global optimum μ^*? If it does, what is the iteration complexity?

Assumption (Bounded and smooth neural networks)

There exists a universal positive constant R such that $\sup_{x \in \mathbb{R}^{d_x}, \mathbf{x} \in \mathcal{X}} |\Psi_{\mathbf{x}}(x)| \leq R$ and $\sup_{x \in \mathbb{R}^{d_x}, \mathbf{x} \in \mathcal{X}} |\nabla_x \Psi_{\mathbf{x}}(x)| \leq R$.

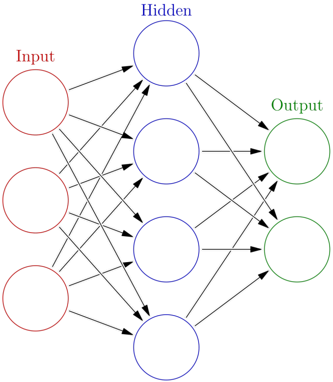
Background: Mean-field perspective of two-layer neural networks

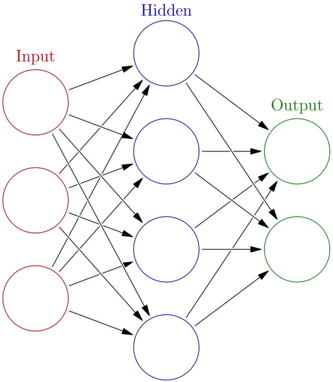
- Define $h_*(\mathbf{x}) = \int \Psi(\mathbf{x}, x)\mathrm{d}\mu_*(x)$ the output of the optimal mean-field neural network.
- Define $\hat{h}_S(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^N \Psi(\mathbf{x}, x_S^{(i)})$ the output of a trained neural network at time S.
- For any input $\mathbf{x} \in \mathcal{X}$,

$$\mathbb{E}\left[\left(\hat{h}_S(\mathbf{x}) - h_*(\mathbf{x})\right)^2\right] \leq \underbrace{\mathcal{O}(N^{-1})}_{\text{finite particle error}} + \underbrace{\mathcal{O}\left(\frac{\gamma^2 + \gamma\sigma d}{C_{\mathrm{LSI}}\sigma}\right)}_{\text{time discretization error}} + \underbrace{\mathcal{O}(\exp(-\gamma C_{\mathrm{LSI}}\sigma S))}_{\text{optimization error}}.$$

  - Expectation is taken over the randomness in initialization and noise at each iteration.
  - $C_{\mathrm{LSI}} = \Theta(\sigma^{-1}\exp(-\zeta^{-1}\sigma^{-1}\sqrt{d}))$ describes the 'difficulty' of learning μ_*.
    - It reflects the curse of dimensionality .

Mean-field neural networks in 2SLS





Mean-field perspective of 2SLS

- Two-stage least squares (2SLS)

$$
\begin{aligned}
\text{Stage I:} \quad & \mathscr{Z}^*(\mathscr{X}) = \underset{\mathscr{Z}\in(\mathbb{R}^{d_z})^{N_x}}{\arg\min} \ \frac{1}{2}\mathbb{E}_\rho\left[(h(\mathbf{z},\mathscr{Z}) - h(\mathbf{x},\mathscr{X}))^2\right], \\
\text{Stage II:} \quad & \mathscr{X}^* = \underset{\mathscr{X}\in(\mathbb{R}^{d_x})^{N_z}}{\arg\min} \ \frac{1}{2}\mathbb{E}_\rho\left[(h(\mathbf{z},\mathscr{Z}^*(\mathscr{X})) - \mathbf{y})^2\right].
\end{aligned} \tag{1}
$$

- $h(\mathbf{x},\mathscr{X}) = \frac{1}{N_x}\sum_{i=1}^{N_x}\Psi_{\mathbf{x}}(x^{(i)})$ where $\mathscr{X} = [x^{(1)},\ldots,x^{(N_x)}] \in (\mathbb{R}^{d_x})^{N_x}$ are the network parameters and $\mathbf{x}$ is the network input.
- $h(\mathbf{z},\mathscr{Z}) = \frac{1}{N_z}\sum_{i=1}^{N_z}\Psi_{\mathbf{z}}(z^{(i)})$ where $\mathscr{Z} = [z^{(1)},\ldots,z^{(N_z)}] \in (\mathbb{R}^{d_z})^{N_z}$ are the network parameters and $\mathbf{z}$ is the network input.
- ρ is the data distribution over $(\mathbf{x},\mathbf{z},\mathbf{y})$.
- A shorthand notation: $\Psi_{\mathbf{x}}(x^{(i)}) = \Psi(\mathbf{x}, x^{(i)})$ and $\Psi_{\mathbf{z}}(z^{(i)}) = \Psi(\mathbf{z}, z^{(i)})$.

Mean-field perspective of 2SLS

- Mean field neural networks $\int_{\mathbb{R}^{d_x}} \Psi_{\mathbf{x}}(x)\mathrm{d}\mu_x(x)$ and $\int_{\mathbb{R}^{d_z}} \Psi_{\mathbf{z}}(z)\mathrm{d}\mu_z(z)$ where μ_x, μ_z are the mean-field limit of the hidden layer.
- ℓ_2 and entropic regularizations for both stages:

$$\text{Stage I:} \quad \mu_z^*(\mu_x) = \underset{\mu_z \in \mathcal{P}(\mathbb{R}^{d_z})}{\arg\min} \frac{1}{2}\mathbb{E}_\rho[(\textstyle\int \Psi_{\mathbf{z}}\mathrm{d}\mu_z - \int \Psi_{\mathbf{x}}\mathrm{d}\mu_x)^2] + \frac{\zeta_1}{2}\mathbb{E}_{\mu_z}[\|z\|^2] + \sigma_1 \mathrm{Ent}(\mu_z),$$

$$\text{Stage II:} \quad \mu_x^* = \underset{\mu_x \in \mathcal{P}(\mathbb{R}^{d_x})}{\arg\min} \frac{1}{2}\mathbb{E}_\rho[(\textstyle\int \Psi_{\mathbf{z}}\mathrm{d}\mu_z^*(\mu_x) - \mathbf{y})^2] + \frac{\zeta_2}{2}\mathbb{E}_{\mu_x}[\|x\|^2] + \sigma_2 \mathrm{Ent}(\mu_x). \tag{Bi-MFLD}$$

- A bilevel optimization problem over $\mathcal{P}(\mathbb{R}^{d_x})$ and $\mathcal{P}(\mathbb{R}^{d_z})$.
  - Popular methods like explicit gradient (autodiff) and implicit gradient (high-order gradient) do not work.
  - For fixed μ_x, Stage I $\mu_z^*(\mu_x)$ can be solved via standard mean field Langevin dynamics.

Mean-field perspective of 2SLS

- Some notation:

$$F_1(\mu_x, \mu_z) = \frac{1}{2}\mathbb{E}_\rho[(\int \Psi_{\mathbf{z}} \,\mathrm{d}\mu_z - \int \Psi_{\mathbf{x}} \,\mathrm{d}\mu_x)^2] + \frac{\zeta_1}{2}\mathbb{E}_{\mu_z}[\|z\|^2]$$
$$F_2(\mu_x, \mu_z) = \frac{1}{2}\mathbb{E}_\rho[(\int \Psi_{\mathbf{z}} \,\mathrm{d}\mu_z - \mathbf{y})^2] + \frac{\zeta_2}{2}\mathbb{E}_{\mu_x}[\|x\|^2].$$

- $\mathscr{F}_1(\mu_x, \mu_z) = F_1(\mu_x, \mu_z) + \sigma_1 \mathrm{Ent}(\mu_z)$ and $\mathscr{F}_2(\mu_x, \mu_z) = F_2(\mu_x, \mu_z) + \sigma_2 \mathrm{Ent}(\mu_x)$.
- Bilevel optimization problem is

Stage I: $\mu_z^*(\mu_x) = \underset{\mu_z \in \mathcal{P}(\mathbb{R}^{d_z})}{\arg\min}\ \mathscr{F}_1(\mu_x, \mu_z)$, Stage II: $\mu_x^* = \underset{\mu_x \in \mathcal{P}(\mathbb{R}^{d_x})}{\arg\min}\ \mathscr{F}_2(\mu_x, \mu_z^*(\mu_x))$.

Observations:

1. The partial Wasserstein gradients $\mu_x \mapsto F_1(\mu_x, \mu_z)$ and $\mu_x \mapsto F_2(\mu_x, \mu_z)$; $\mu_z \mapsto F_1(\mu_x, \mu_z)$ and $\mu_z \mapsto F_1(\mu_x, \mu_z)$ are simple.
2. The nested Wasserstein gradient of $\mu_x \mapsto F_2(\mu_x, \mu_z^*(\mu_x))$ is nasty.

Mean-field perspective of 2SLS: Lagrangian formulation

- The bilevel optimization problem

$$\mu_z^*(\mu_x) = \underset{\mu_z \in \mathcal{P}(\mathbb{R}^{d_z})}{\arg\min}\ \mathscr{F}_1(\mu_x, \mu_z), \quad \mu_x^* = \underset{\mu_x \in \mathcal{P}(\mathbb{R}^{d_x})}{\arg\min}\ \mathscr{F}_2(\mu_x, \mu_z^*(\mu_x)). \qquad \text{(Bilevel)}$$

- A constrained optimization problem
  - Stage I problem re-casted as a constraint.

$$\min_{\mu_x, \mu_z} \mathscr{F}_2(\mu_x, \mu_z), \quad \mathscr{F}_1(\mu_x, \mu_z) - \mathscr{F}_1(\mu_x, \mu_z^*(\mu_x)) \leq \varepsilon. \qquad (\varepsilon\text{-constrained})$$

- A Lagrangian optimization problem

$$(\mu_{x,\lambda}^*, \mu_{z,\lambda}^*) = \arg\min_{\mu_x, \mu_z} \mathscr{L}_\lambda(\mu_x, \mu_z) := \mathscr{F}_2(\mu_x, \mu_z) + \lambda \left(\mathscr{F}_1(\mu_x, \mu_z) - \mathscr{F}_1(\mu_x, \mu_z^*(\mu_x))\right). \qquad (\lambda\text{-penalty})$$

  - When $\lambda = +\infty$, it recovers the bilevel optimization problem.
  - When $\lambda < +\infty$, one needs to take into account an additional approximation error.

Mean-field perspective of 2SLS: Lagrangian formulation

- Main challenge:

$$
\begin{aligned}
(\mu^*_{x,\lambda}, \mu^*_{z,\lambda}) &= \arg\min_{\mu_x,\mu_z} \mathscr{L}_\lambda(\mu_x, \mu_z) \\
&= \arg\min_{\mu_x,\mu_z} \mathscr{F}_2(\mu_x, \mu_z) + \lambda \left(\mathscr{F}_1(\mu_x, \mu_z) - \mathscr{F}_1(\mu_x, \mu^*_z(\mu_x))\right)
\end{aligned}
$$

Proposition 1 (Wasserstein gradient of $\mathscr{L}_\lambda$)

Let $\mu^*_z(\mu_x) = \arg\min_{\mu_z} \mathscr{F}_1(\mu_x, \mu_z)$ be the solution to the stage I optimization. Then,

$$
\begin{aligned}
\boldsymbol{\nabla}_1 \mathscr{L}_\lambda(\mu_x, \mu_z) &= \boldsymbol{\nabla}_1 \mathscr{F}_2(\mu_x, \mu_z) + \lambda \boldsymbol{\nabla}_1 \mathscr{F}_1(\mu_x, \mu_z) - \lambda \boldsymbol{\nabla}_1 \mathscr{F}_1(\mu_x, \mu^*_z(\mu_x)), \\
\boldsymbol{\nabla}_2 \mathscr{L}_\lambda(\mu_x, \mu_z) &= \boldsymbol{\nabla}_2 \mathscr{F}_2(\mu_x, \mu_z) + \lambda \boldsymbol{\nabla}_2 \mathscr{F}_1(\mu_x, \mu_z).
\end{aligned}
$$

$\boldsymbol{\nabla}_1$ *(resp. $\boldsymbol{\nabla}_2$) denotes the Wasserstein gradient with the first (resp. second) argument.*

- The Wasserstein gradient of the mapping $\mu_x \mapsto \mathscr{F}_1(\mu_x, \mu^*_z(\mu_x))$ only involves the partial derivative with the first argument (envelope theorem).
- We avoid the nasty Wasserstein gradient of $\mu_x \mapsto \mathscr{F}_2(\mu_x, \mu^*_z(\mu_x))$.

Mean-field perspective of 2SLS: Lagrangian formulation

- Convexity of $\mathscr{L}_\lambda(\mu_x, \mu_z) = \mathscr{F}_2(\mu_x, \mu_z) + \lambda\left(\mathscr{F}_1(\mu_x, \mu_z) - \mathscr{F}_1(\mu_x, \mu_z^*(\mu_x))\right)$.

Observations:

1. The partial mapping $\mu_z \mapsto \mathscr{L}_\lambda(\mu_x, \mu_z)$ is convex, for any fixed $\mu_x \in \mathcal{P}_2(\mathbb{R}^{d_x})$.
2. The partial mapping $\mu_x \mapsto \mathscr{L}_\lambda(\mu_x, \mu_z)$ is not convex, for any fixed $\mu_z \in \mathcal{P}_2(\mathbb{R}^{d_x})$.
3. The joint mapping $(\mu_x, \mu_z) \mapsto \mathscr{L}_\lambda(\mu_x, \mu_z)$ is not convex.

Question: How to exploit this partial convexity $\mu_z \mapsto \mathscr{L}_\lambda(\mu_x, \mu_z)$?

- Innerloop: $\mu_z^*(\mu_x) = \arg\min_{\mu_z} \mathscr{F}_1(\mu_x, \mu_z)$, $\tilde{\mu}_z^*(\mu_x) = \arg\min_{\mu_z} \mathscr{L}_\lambda(\mu_x, \mu_z)$.
- Outerloop: $\mu_{x,\lambda}^* = \arg\min_{\mu_x} \mathscr{L}_\lambda\left(\mu_x,\, \tilde{\mu}_z^*(\mu_x),\, \mu_z^*(\mu_x)\right)$.
- Noisy gradient descent!

Mean-field perspective of 2SLS: Lagrangian formulation

- Inner-loop algorithm

$$\mu_z^*(\mu_x) = \arg\min_{\mu_z} \mathscr{F}_1(\mu_x, \mu_z) = \arg\min_{\mu_z} F_1(\mu_x, \mu_z) + \sigma_1 \mathrm{Ent}(\mu_z),$$

$$\tilde{\mu}_z^*(\mu_x) = \arg\min_{\mu_z} \mathscr{L}_\lambda(\mu_x, \mu_z) = \arg\min_{\mu_z} F_2(\mu_x, \mu_z) + \lambda F_1(\mu_x, \mu_z) + \lambda\sigma_1 \mathrm{Ent}(\mu_z).$$

- Fast convergence due to partial convexity of $\mu_z \mapsto F_1(\mu_x, \mu_z)$ and $\mu_z \mapsto F_2(\mu_x, \mu_z) + \lambda F_1(\mu_x, \mu_z)$ for fixed μ_x.

Algorithm $\textsc{InnerLoop}(\mu_x,\ T,\ \alpha,\ \beta,\ \lambda,\ \sigma_1)$

1: Initialize $\mu_{\mathscr{Z},0} = \frac{1}{N_z}\sum_{j=1}^{N_z} \delta_{z_0^{(j)}}$ and $\tilde{\mu}_{\mathscr{Z},0} = \frac{1}{N_z}\sum_{j=1}^{N_z} \delta_{\tilde{z}_0^{(j)}}$.
2: **for** $t = 0, \ldots, T$ **do**
3: **for** $i = 1, \ldots, N_z$ **do**
4: $z_{t+1}^{(i)} = z_t^{(i)} - \alpha \boldsymbol{\nabla}_2 F_1(\mu_x, \mu_{\mathscr{Z},t})(z_t^{(i)}) + \sqrt{2\alpha\sigma_1}\, \xi_{z,t}^{(i)}$.
5: $\tilde{z}_{t+1}^{(i)} = \tilde{z}_t^{(i)} - \beta \nabla_2 F_2(\mu_x, \tilde{\mu}_{\mathscr{Z},t})(\tilde{z}_t^{(i)}) - \beta\lambda \boldsymbol{\nabla}_2 F_1(\mu_x, \tilde{\mu}_{\mathscr{Z},t})(\tilde{z}_t^{(i)}) + \sqrt{2\beta\lambda\sigma_1}\, \tilde{\xi}_{z,t}^{(i)}$.
6: **end for**
7: **end for**

Mean-field perspective of 2SLS: Lagrangian formulation

Assumption 1 (Bounded and smooth neural networks)

There exists a universal positive constant R *such that* $\sup_{x \in \mathbb{R}^{d_x}, \mathbf{x} \in \mathcal{X}} |\Psi_{\mathbf{x}}(x)| \leq R$ *and* $\sup_{z \in \mathbb{R}^{d_z}, \mathbf{z} \in \mathcal{Z}} |\Psi_{\mathbf{z}}(z)| \leq R$*. Also,* $\sup_{x \in \mathbb{R}^{d_x}, \mathbf{x} \in \mathcal{X}} |\nabla_x \Psi_{\mathbf{x}}(x)| \leq R$ *and* $\sup_{z \in \mathbb{R}^{d_z}, \mathbf{z} \in \mathcal{Z}} |\nabla_z \Psi_{\mathbf{z}}(z)| \leq R$.

- It works for two-layer neural networks with tanh/ReLU plus smooth output clipping.

Assumption 2 (Bounded target)

There exists a universal constant M *such that the target random variable* $|Y| \leq M$ *and* $|h_\circ(X)| \leq M$ *almost surely.*

- Boundedness of $|Y|$ can be relaxed to sub-Gaussian residual $Y - (Th_\circ)(Z)$.

Mean-field perspective of 2SLS: Lagrangian formulation

Proposition 2 (Inner-loop convergence towards $\mu_z^*(\mu_x)$ and $\tilde{\mu}_z^*(\mu_x)$)

Suppose Assumption 1 and 2 hold. Given a fixed $\mu_x \in \mathcal{P}_2(\mathbb{R}^{d_x})$. Let $\mathscr{Z} = \{z^{(i)}\}_{i=1}^{N_z}$ and $\tilde{\mathscr{Z}} = \{\tilde{z}^{(i)}\}_{i=1}^{N_z}$ be the output of the inner-loop algorithm $\textsc{InnerLoop}(\mu_x, T, \alpha, \beta, \lambda, \sigma_1)$. *Denote $\mu_z^{(N_z)}$ and $\tilde{\mu}_z^{(N_z)}$ as the joint distribution of these N_z particles $\mathscr{Z}$. Suppose the step size satisfy $\alpha \leq \frac{1}{\zeta_1}$ and $\beta \leq \frac{1}{\lambda\zeta_2}$. For any $T > 0$,*

$$\frac{\sigma_1}{N_z}\mathrm{KL}\left(\mu_z^{(N_z)}, (\mu_z^*(\mu_x))^{\otimes N_z}\right) \leq \frac{R^2}{N_z} + \frac{\alpha^2 + \alpha\sigma_1 d_z}{C_{\mathrm{LSI},z}\sigma_1} + \mathcal{O}(\exp(-C_{\mathrm{LSI},z}\sigma_1\alpha T))$$

$$\frac{\sigma_2}{N_z}\mathrm{KL}\left(\tilde{\mu}_z^{(N_z)}, (\tilde{\mu}_z^*(\mu_x))^{\otimes N_z}\right) \leq \frac{R^2}{N_z} + \frac{\beta^2 + \beta\sigma_1 d_z}{C_{\mathrm{LSI},z}\sigma_1} + \mathcal{O}(\exp(-C_{\mathrm{LSI},z}\sigma_1\beta T)).$$

- $C_{\mathrm{LSI},z} = \Theta(\frac{\zeta_1}{\sigma_1}\exp(-\frac{R^2}{\zeta_1\sigma_1}\sqrt{d_z/\pi}))$.
- Direct application of mean-field results.

- Outer-loop algorithm

$$\mu^*_{x,\lambda} = \arg\min_{\mu_x} \mathscr{L}_\lambda(\mu_x, \tilde{\mu}^*_z(\mu_x), \mu^*_z(\mu_x)) = \arg\min_{\mu_x} L_\lambda(\mu_x, \tilde{\mu}^*_z(\mu_x), \mu^*_z(\mu_x)) + \sigma_2 \mathrm{Ent}(\mu_x)$$

- Its Wasserstein gradient $\boldsymbol{\nabla} L_\lambda(\mu_x, \tilde{\mu}^*_z(\mu_x))(x)$ equals (via envelope theorem)

$$\boldsymbol{\nabla}_1 F_2(\mu_x, \tilde{\mu}^*_z(\mu_x))(x) + \lambda(\boldsymbol{\nabla}_1 \mathscr{F}_1(\mu_x, \tilde{\mu}^*_z(\mu_x))(x) - \boldsymbol{\nabla}_1 \mathscr{F}_1(\mu_x, \mu^*_z(\mu_x))(x))$$

Algorithm OUTERLOOP: F^2BMLD (Fully-first order Bilevel MFLD)

1: Initialize $\mu_{\mathscr{X},0} = \frac{1}{N_x} \sum_{j=1}^{N_x} \delta_{x_0^{(j)}}$.
2: **for** $s = 0, \dots, S$ **do**
3: $\tilde{\mu}_{\mathscr{Z},s}, \mu_{\mathscr{Z},s} \leftarrow$ INNERLOOP$(\mu_{\mathscr{X},s})$.
4: **for** $i = 1, \dots, N_x$ **do**
5:
$$x^{(i)}_{s+1} = x^{(i)}_s - \gamma\Big(\boldsymbol{\nabla}_1 F_2(\mu_{\mathscr{X},s}, \tilde{\mu}_{\mathscr{Z},s})(x^{(i)}_s) + \lambda(\boldsymbol{\nabla}_1 \mathscr{F}_1(\mu_{\mathscr{X},s}, \tilde{\mu}_{\mathscr{Z},s})(x^{(i)}_s) - \boldsymbol{\nabla}_1 \mathscr{F}_1(\mu_{\mathscr{X},s}, \mu_{\mathscr{Z},s})(x^{(i)}_s))\Big) + \sqrt{2\gamma\sigma_2}\,\xi^{(i)}_{x,s}.$$
6: **end for**
7: **end for**

Algorithm ✓ Theory ?

Mean-field perspective of 2SLS: Lagrangian formulation

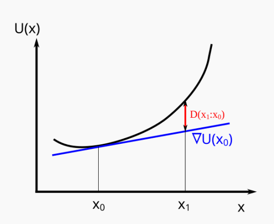
Question: How to prove convergence $\mu_{\mathscr{X},S} = \frac{1}{N_x}\sum_{i=1}^{N_x} \delta_{x_S^{(i)}} \to \mu_{x,\lambda}^*$?

- The key is convexity!
- $\mu_x \mapsto L_\lambda(\mu_x, \tilde{\mu}_z^*(\mu_x), \mu_z^*(\mu_x))$ is only weakly convex.

Lemma (Lower-bound on the Bregman divergence of L_λ)

Suppose Assumption 1 holds. Then, we have $B_{L_\lambda}(\mu_x, \mu_x') \geq -\frac{R^3\lambda}{4\sigma_1}\mathrm{TV}^2(\mu_x, \mu_x')$.

- L_λ is more convex as σ_1 increases yet less convex as λ increases.



Mean-field perspective of 2SLS: Lagrangian formulation

Theorem 3 (Convergence bound)

Suppose Assumption 1 and 2 hold. Let $\mathfrak{c} > 0$ and assume that $\sigma_1\sigma_2\mathfrak{c} \geq 4R^3\lambda$. Suppose the step size $\gamma \leq \zeta_2^{-1}$. Given a fixed $\lambda > 0$, for any number of iterations $S \in \mathbb{N}^+$, we have

$$\mathcal{H}(S) \lesssim \exp\left(-\sigma_2 C_{\mathrm{LSI},x} S\gamma\right) + \frac{\lambda R^2}{\sigma_1 N_x} + \frac{\lambda^2\left(\sqrt{\frac{\mathfrak{KL}}{N_z}}+\sqrt{\frac{\tilde{\mathfrak{KL}}}{N_z}}\right)}{\sigma_2 C_{\mathrm{LSI},x}} + \frac{\lambda^2\left(\gamma^2+\gamma\sigma_2 d_x\right)}{\sigma_2 C_{\mathrm{LSI},x}} + \mathfrak{c}^2 C_{\mathrm{LSI},x}.$$

- $\mathfrak{KL}$ and $\tilde{\mathfrak{KL}}$ represents the inner-loop optimization error.
- $\mathfrak{c}$ is a slack parameter arising from the weak convexity of L_λ.
- Define $h_{*,\lambda}(\mathbf{x}) = \int \Psi_{\mathbf{x}}(x)\, \mathrm{d}\mu^*_{x,\lambda}(x)$ the global optimum mean field network. Define $\hat{h}_S(\mathbf{x}) = \frac{1}{N_x}\sum_{i=1}^{N_x} \Psi_{\mathbf{x}}(x_S^{(i)})$ where $\{x_S^{(i)}\}_{i=1}^{N_x}$ are the output of F^2*BMLD*.

$$\forall \mathbf{x} \in \mathcal{X}, \quad \mathbb{E}\left[\left(\hat{h}_S(\mathbf{x}) - h_{*,\lambda}(\mathbf{x})\right)^2\right] \leq \sqrt{\sigma_2^{-1}\mathcal{H}(S) + \frac{\lambda\mathfrak{c}}{\sigma_1\sigma_2} + \frac{\lambda}{N_x\sigma_1\sigma_2} + \frac{1}{N_x}}.$$

- The optimization bound wants small λ.

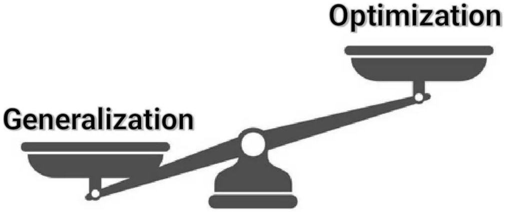
Mean-field perspective of 2SLS: Generalization

Optimization theory:

We have proved that F^2BMLD can indeed find the global optimum solution $h_{*,\lambda}$.

Statistical theory:

How well does $h_{*,\lambda}$ generalize towards $h_\circ$ when given finite i.i.d samples over $(\mathbf{x}, \mathbf{z}, \mathbf{y})$?



Mean-field perspective of 2SLS: Generalization

- Given m i.i.d samples $\{\mathbf{z}_i, \mathbf{x}_i\}_{i=1}^m \sim P_{ZX}$ in stage I and n i.i.d samples $\{\mathbf{z}_i, \mathbf{y}_i\}_{i=1}^n \sim P_{ZY}$ in stage II:

$$\mathscr{F}_1(\mu_x, \mu_z) = \sum_{i=1}^{m} \frac{1}{2m} \left(\int \Psi(\mathbf{z}_i, z)\mathrm{d}\mu_z - \int \Psi(\mathbf{x}_i, x)\mathrm{d}\mu_x\right)^2 + \frac{\zeta_1}{2}\mathbb{E}_{\mu_z}[\|z\|^2] + \sigma_1 \mathrm{Ent}(\mu_z),$$

$$\mathscr{F}_2(\mu_x, \mu_z) = \sum_{i=1}^{n} \frac{1}{2n} \left(\int \Psi(\mathbf{z}_i, z)\mathrm{d}\mu_z^*(\mu_x) - \mathbf{y}_i\right)^2 + \frac{\zeta_2}{2}\mathbb{E}_{\mu_x}[\|x\|^2] + \sigma_2 \mathrm{Ent}(\mu_x).$$

- Recall that $T : L^2(P_X) \to L^2(P_Z)$ defined as $T : f \mapsto \mathbb{E}[f(X) \mid Z]$ and NPIV:

$$\mathbb{E}[Y \mid Z] = \mathbb{E}[h_\circ(X) \mid Z]. \qquad \text{(NPIV)}$$

Mean-field perspective of 2SLS: Generalization

Assumption 3 (Stage II well-specifiedness)

$h_\circ$ belongs to a KL restricted Barron space $\mathcal{B}_{M_x} := \{\int \Psi(\cdot, x)\mathrm{d}\mu_x(x) \mid \mathrm{KL}(\mu_x, \nu_x) \le M_x\}$, where $\nu_x = \mathcal{N}(0, \zeta_2\sigma_2^{-1}\mathrm{Id}_{d_x})$. That is, there exists a measure $\mu_x^\circ \in \mathcal{B}_{M_x}$ such that $h_\circ(\cdot) = \int \Psi(\cdot, x)\mathrm{d}\mu_x^\circ$.

Assumption 4 (Stage I well-specifiedness)

The conditional expectation $T[\int \Psi(\cdot, x)\mathrm{d}\mu_x(x)](\mathbf{z}) = \int \mathbb{E}[\Psi(X, x) \mid Z = \mathbf{z}]\,\mathrm{d}\mu_x(x)$ belongs to a KL restricted Barron space $\mathcal{B}_{M_z} := \{\int \Psi(\cdot, z)\mathrm{d}\mu_z(z) \mid \mathrm{KL}(\mu_z, \nu_z) \le M_z\}$, where $\nu_z = \mathcal{N}(0, \zeta_1\sigma_1^{-1}\mathrm{Id}_{d_z})$. That is, there exists a measure $\mu_z^\circ(\mu_x) \in \mathcal{B}_{M_z}$ such that $T[\int \Psi(\cdot, x)\mathrm{d}\mu_x(x)](\mathbf{z}) = \int \Psi(\cdot, z)\mathrm{d}\mu_z^\circ(\mu_x)$.

- M_x, M_z are universal constants that control the size of the Barron spaces.

Mean-field perspective of 2SLS: Generalization

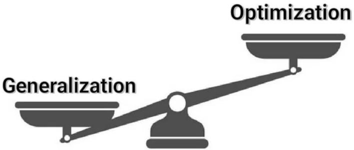
Theorem 4 (Generalization bound)

Suppose Assumption 1,2,3,4 hold. For $\lambda > 0$, let $\mu^*_{x,\lambda}$ be the optimal solution to the Lagrangian problem and $h_{*,\lambda}(\mathbf{x}) = \int \Psi(\mathbf{x}, x)\, \mathrm{d}\mu^*_{x,\lambda}(x)$ be its associated mean field neural network. Then, with $P_{XZY}^{\otimes(m+n)}$ probability at least $1 - 8\delta$,

$$\mathbb{E}_{P_Z}\left[\Big((Th_{*,\lambda})(Z) - (Th_\circ)(Z)\Big)^2\right] \lesssim \sigma_2 M_x + \sigma_1 M_z + \frac{R^2(R+M)^2}{\sigma_1 \lambda} + \sqrt{\frac{M_z + \frac{1}{\sigma_1} + \log(\delta^{-1})}{m}} + \sqrt{\frac{M_x + \frac{1}{\sigma_2} + \log(\delta^{-1})}{n}}.$$

- The generalization bound wants large λ so the Lagrangian problem is more faithful to the original bilevel optimization problem.
- $\mathcal{O}(m^{-\frac{1}{2}})$ and $\mathcal{O}(n^{-\frac{1}{2}})$ arise from Rademacher complexity bound.
    - Two-stage regression so we need both $m, n \to \infty$.

Mean-field perspective of 2SLS: Optimization and Generalization



- Trade-off on $\lambda, \sigma_1, \sigma_2$ in terms of optimization and generalization.
- Optimization bound: $\mathbb{E}\left[\left(\hat{h}_S(\mathbf{x}) - h_{*,\lambda}(\mathbf{x})\right)^2\right] = \mathcal{O}(\lambda^2 + \sigma_1^{-1} + \sigma_2^{-1})$.
- Generalization bound: $\mathbb{E}_{P_Z}\left[\left((Th_{*,\lambda})(Z) - (Th_\circ)(Z)\right)^2\right] = \mathcal{O}(\lambda^{-1} + \sigma_1 + \sigma_2)$.
- Unfortunately, there does not exist a pair of $\lambda, \sigma_1, \sigma_2$ such that both errors vanish.

Experiments: Offline RL on Cartpole

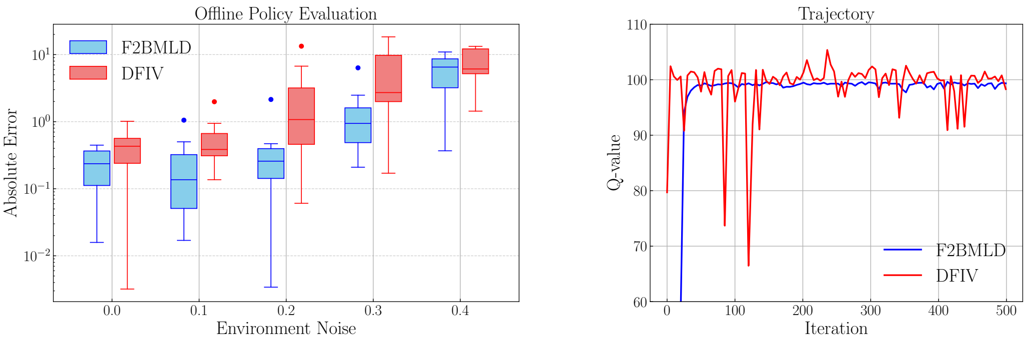


Figure: **Left:** Comparison of DFIV and F2BMLD in terms of target policy value. **Right:** Comparison of DFIV and F2BMLD training trajectories.

- λ is selected from a set $\{0.1, 1.0, 10.0\}$.
- More stable trajectory because of fully-first order gradient in optimization.

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