Causal Machine Learning – Fall 2023

# Week 6: Deep Nets & Two Step Semiparametrics

**Max H. Farrell & Sanjog Misra**

## Topics to cover

1. Convergence Rates for Deep Nets
2. Two step semiparametric inference
    - Why do we care about all this rate of convergence stuff?

# Nonparametrics – Last class

- Fitting a linear model in each of $J$ bins:

$$\left|\hat{f}(x) - f(x)\right| = O_p\left(\sqrt{\frac{J}{n}} + J^{-2}\right)$$

- Connected lines or not, same result
- MSE optimal $J \asymp n^{1/5}$

$$\Rightarrow \text{RMSE} = n^{-\frac{2}{5}} = n^{-\frac{2}{2(2\times2+1)}} = n^{-\frac{\text{smoothness}}{2\times\text{smoothness}+\text{dim}}}$$

  b/c fitting lines needs the 2nd derivative.

- ATE optimal $J$? Difficult or unknown
- In general:

$$\text{Var} = \frac{1}{\text{effective sample size}} = \frac{\#\text{ params}}{n}$$

$$\text{Bias} = (\#\text{ params})^{-(\text{smoothness})}$$

# Nonparametrics – Deep Nets

Main result of Farrell, Liang, Misra (2021, *Econometrica*)

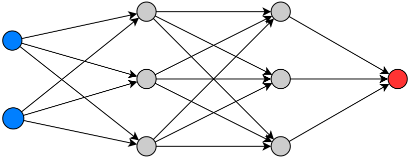
$$\left|\hat{f}_{\text{DNN}}(x) - f(x)\right| = O_p\left(\sqrt{\frac{W \times L \log(W) \log(n)}{n}} + \epsilon_n\right)$$

- $W$ = number of parameters
- $L$ = Depth
- $\epsilon_n$ = bias, which depends on the architecture

Rate is not as fast

- The variance part is not just $\frac{\#\text{ params}}{n}$
- Extra $L$ and $\log()$ terms

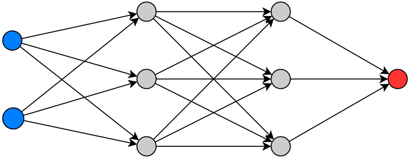
# Nonparametrics – Deep Nets



Number of parameters $W$:

$$
\begin{aligned}
W &= (d+1)H_1 + \sum_{l=2}^{L}(H_{l-1}+1)H_l + (H_L+1) \\
&= (d+1)H + (L-1)(H^2+H) + H + 1 \\
&\asymp LH^2
\end{aligned}
$$

# Nonparametrics – Deep Nets



Approximation depends on how complex the deep net can be:

$$\begin{aligned}\epsilon_n &\leq (WL\log(W))^{-\mathrm{smoothness}/2\times\mathrm{dim}} \\ &\asymp \left(H^2L^2\log(H^2L)\right)^{-\mathrm{smoothness}/2\times\mathrm{dim}}\end{aligned}$$

# Nonparametrics – Deep Nets

Putting the variance and bias together to get the best rate:

$$H \asymp n^{-\frac{\mathrm{dim}}{2\times(\mathrm{smoothness}+\mathrm{dim})}} \log^2(n) \qquad \text{and} \qquad L \asymp \log(n)$$

$$\Rightarrow \quad \left|\hat{f}_{\mathrm{DNN}}(x) - f(x)\right| = O_p\left(n^{-\frac{\mathrm{smoothness}}{2(\mathrm{smoothness}+\mathrm{dim})}} \log^8(n)\right)$$

- Not as fast as before, but fast enough for inference later
- Same features as usual
  - Smoother functions are easier to approximate
  - Curse of dimensionality
- Other research shows that DNNs can adapt to certain low dimensional structures if they are present
  - even if you do not know that in advance.
  - E.g. additive model has dim=1: $f(x_1, x_2, \ldots, x_d) = f_1(x_1) + f_2(x_2) + \cdots + f_d(x_d)$

# Semiparametric Two Step Estimation & Inference

- Basically the main goal of the class
- Semiparametric: Inference target is finite dimensional, first stage is nonparametric/ML
- Key ideas:
  1. First step correction
  2. Influence function based estimator & double robustness
  3. Sample splitting & cross fitting
  - $2 + 3 = \text{DML}$

Today we will stick with half the ATE

- Parameter of interest is $\mu = \mathbb{E}[Y(1)]$
- Identification $\mu = \mathbb{E}[\mathbb{E}[Y \mid T = 1, X]] = \mathbb{E}[\mu_1(X)]$
- Nonparametric first step: $\hat{\mu}_1(x)$

# Semiparametric Two Step Estimation & Inference

Remember in week 3 we did **parametric** two step estimation:

$$\widehat{\mathbb{E}[Y(1)]} = \frac{1}{n}\sum_{i=1}^{n} x_i' \hat{\beta}_1$$

The big conclusion was to show that the first stage *estimation* had an impact on the second-stage *inference*.

$$\sqrt{n}\left(\widehat{\mathbb{E}[Y(1)]} - \mathbb{E}[Y(1)]\right)$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\underbrace{x_i'\beta_1 - \mathbb{E}[X\beta_1]}_{\text{Plug in part}} + \underbrace{\mathbb{E}[X']\mathbb{E}[TXX']^{-1}t_i x_i \varepsilon_i}_{\text{First step correction}}\Big\} + o_p(1)$$

$$\to_d \mathcal{N}\big(0, \mathbb{V}[\phi(z_i)]\big)$$

# Semiparametric Two Step Estimation & Inference

But now our first stage is more complicated

$$\widehat{\mathbb{E}[Y(1)]} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_1(x_i)$$

What impact does this have on second stage inference?

- For inference/testing/CI, we need a Normal approximation from a CLT
- CLT applies to sample averages
- $\Rightarrow$ We need to do the same calculation that we did in week 3:

$$\sqrt{n}\left(\widehat{\mathbb{E}[Y(1)]} - \mathbb{E}[Y(1)]\right) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_1(x_i) - \mathbb{E}[\mu_1(X)]\right) \overset{?}{=} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \textbf{?}$$