Experimental Evaluation of Causal Machine Learning

Kosuke Imai

Harvard University

Department of Statistics and the Florence Center for Data Science
University of Florence

Joint work with Michael Lingzhi Li (Harvard Business School)

Motivation

- Two revolutions over the past 20 years:
  1. causal inference
  2. machine learning

- Causal machine learning
  1. individualized treatment rules
  2. heterogeneous treatment effects

- Experimental evaluation of causal machine learning (ML)
  - ML algorithms do not necessarily work well in practice
  - uncertainty quantification is important and yet difficult
  - evaluate causal ML before putting it in practice

Evaluating Individualized Treatment Rules

- Individualized treatment rules (ITRs)
  - designed to increase efficiency of policies or treatments
  - personalized medicine, micro-targeting in business/politics

- Existing literature:
  1. estimation of heterogeneous treatment effects
  2. active development of optimal ITRs
  3. extensive use of ML algorithms

- Goal: use a randomized experiment to *evaluate generic ITRs*
  1. use a separate experiment to evaluate ITRs developed with other data
  2. use the same experiment to construct and evaluate ITRs
- Imai and Li. "Experimental Evaluation of Individualized Treatment Rules." *Journal of the American Statistical Association*, Forthcoming.

Key Contributions

1. Neyman's repeated sampling framework
   - random treatment assignment, random sampling
   - no modeling assumption or asymptotic approximation
   - extend analysis to cross-fitting: random splitting

2. Evaluation measures
   - shortcomings of existing metrics
   - incorporating a budget constraint
   - overall evaluation metric for general ITRs

Evaluation without a Budget Constraint

- Setup
  - Binary treatment: $T_i \in \{0, 1\}$
  - Pre-treatment covariates: $\mathsf{X} \in \mathcal{X}$
  - No interference: $Y_i(T_1 = t_1, T_2 = t_2, \ldots, T_n = t_n) = Y_i(T_i = t_i)$
  - Random sampling of units:

$$(Y_i(1), Y_i(0), \mathsf{X}_i) \overset{\text{i.i.d.}}{\sim} \mathcal{P}$$

  - Completely randomized treatment assignment:

$$\Pr(T_i = 1 \mid Y_i(1), Y_i(0), \mathsf{X}_i) = \frac{n_1}{n} \quad \text{where} \quad n_1 = \sum_{i=1}^{n} T_i$$

- Fixed (for now) ITR:

$$f : \mathcal{X} \longrightarrow \{0, 1\}$$

  - based on any ML algorithm or even a heuristic rule
  - sample splitting for experimental data, separate observational data

Neyman's Inference for the Standard Metric

- Standard metric (Population Average "Value" or PAV):

$$\lambda_f \;=\; \mathbb{E}\{Y_i(f(\mathrm{X}_i))\}$$

- A natural estimator:

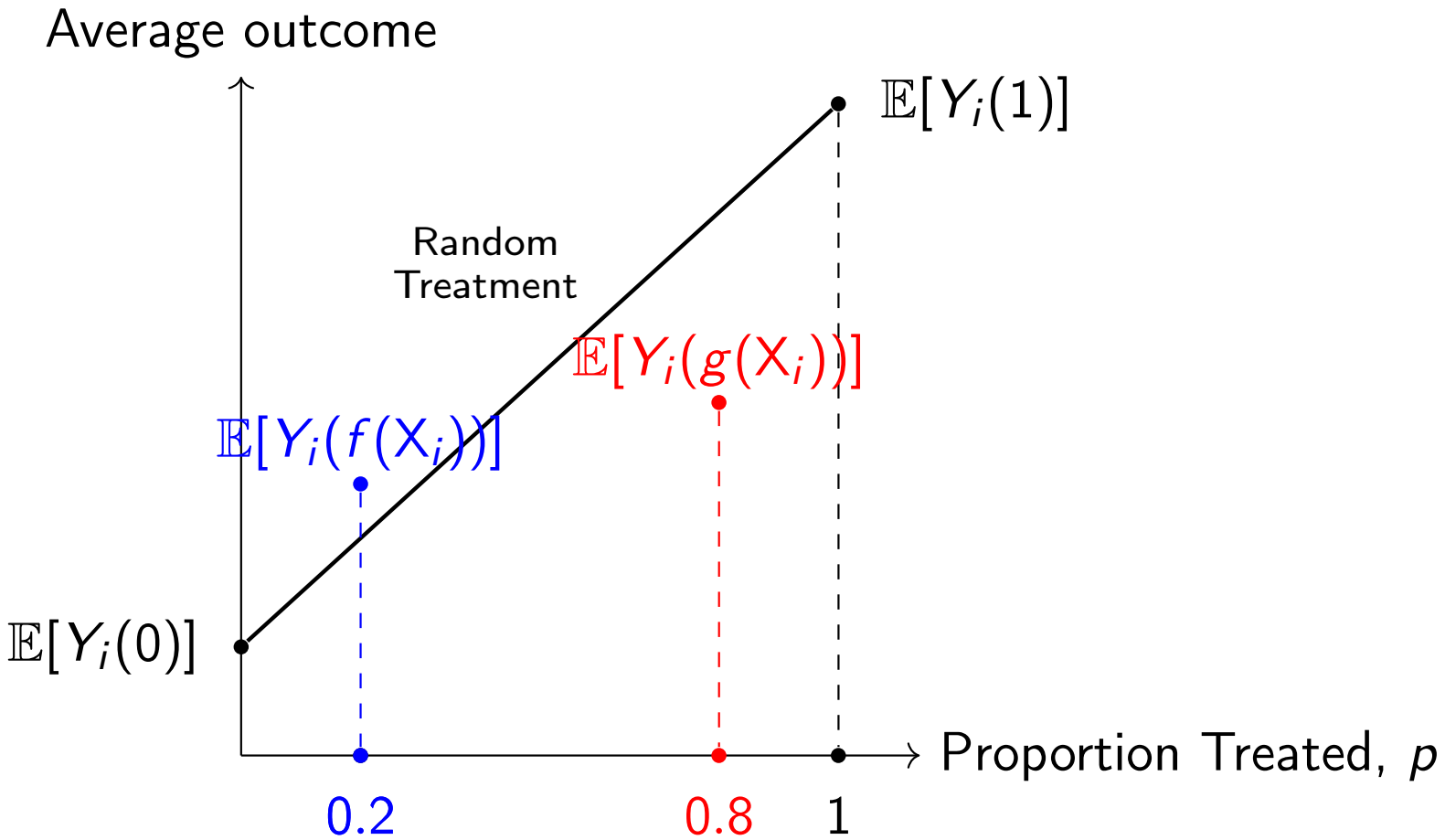
$$\hat{\lambda}_f(\mathcal{Z}) \;=\; \frac{1}{n_1} \underbrace{\sum_{i=1}^{n} Y_i T_i f(\mathrm{X}_i)}_{\substack{\text{treated units who should} \\ \text{be treated}}} + \frac{1}{n_0} \underbrace{\sum_{i=1}^{n} Y_i(1-T_i)(1-f(\mathrm{X}_i))}_{\substack{\text{untreated units who should} \\ \text{not be treated}}},$$

  where $\mathcal{Z} = \{\mathrm{X}_i, T_i, Y_i\}_{i=1}^n$
- Unbiasedness: $\mathbb{E}\{\hat{\lambda}_f(\mathcal{Z})\} = \lambda_f$
- Usual variance:

$$\mathbb{V}\{\hat{\lambda}_f(\mathcal{Z})\} \;=\; \frac{\mathbb{E}(S_{f1}^2)}{n_1} + \frac{\mathbb{E}(S_{f0}^2)}{n_0},$$

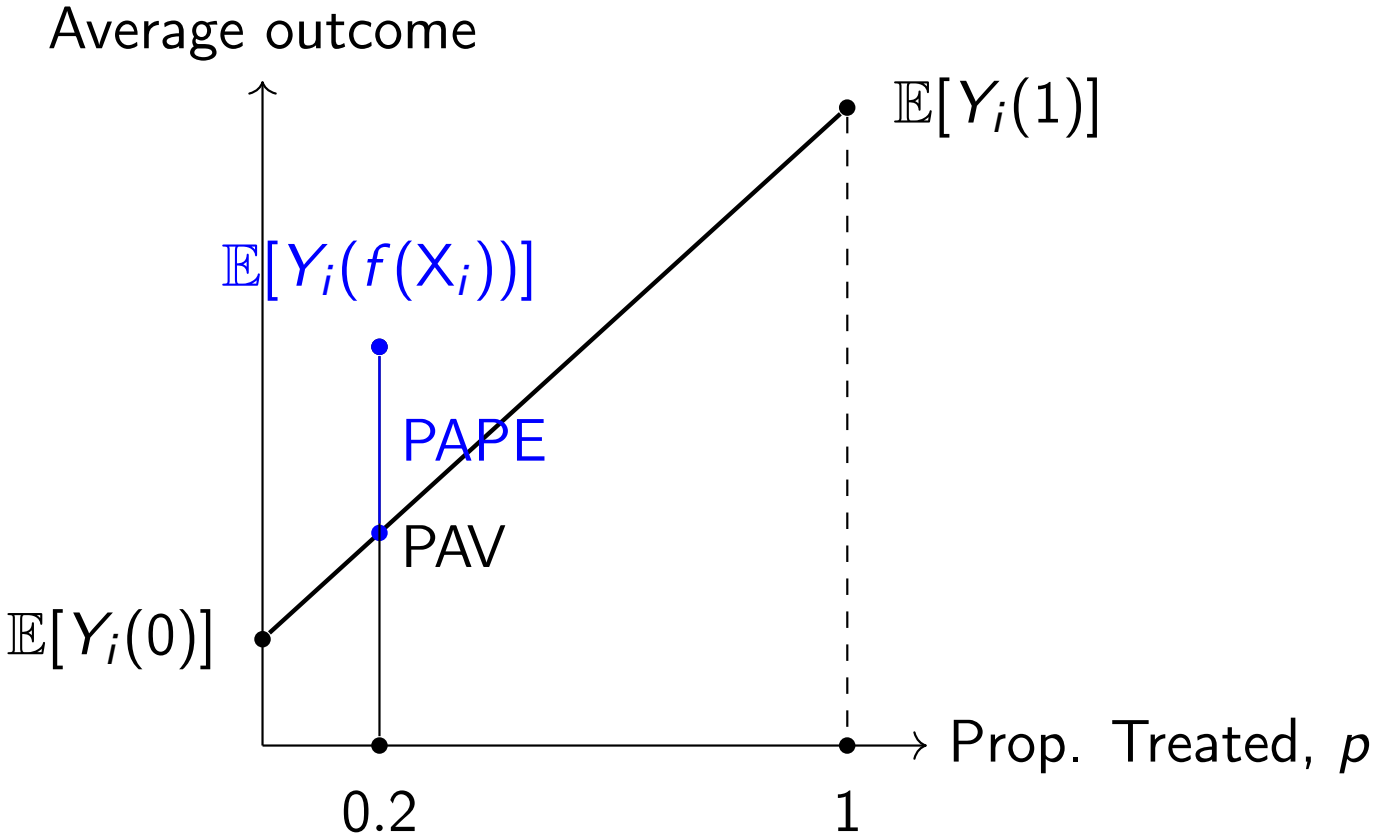
  where $S_{ft}^2 = \sum_{i=1}^n (Y_{fi}(t) - \overline{Y_f(t)})^2/(n-1)$,
  $Y_{fi}(t) = 1\{f(\mathrm{X}_i) = t\}Y_i(t)$, and $\overline{Y_f(t)} = \sum_{i=1}^n Y_{fi}(t)/n$ for $t = \{0, 1\}$

A Problem of Comparing ITRs Using the PAV



- $\lambda_f < \lambda_g$: but g is performing worse than the random (i.e., non-individualized) treatment rule whereas f is not
- Need to account for the proportion treated

Accounting for the Proportion of Treated Units



- Population Average Prescriptive Effect (PAPE):

$$\tau_f \;=\; \mathbb{E}\{Y_i(f(X_i)) - p_f\, Y_i(1) - (1 - p_f)\, Y_i(0)\}$$

where $p_f = \Pr(f(X_i) = 1)$ is the proportion treated under f

Estimating the Population Average Prescriptive Effect

- An unbiased estimator of PAPE τ_f:

$$\hat{\tau}_f(\mathcal{Z}) = \frac{n}{n-1}\Bigg[\underbrace{\frac{1}{n_1}\sum_{i=1}^{n} Y_i T_i f(\mathrm{X}_i) + \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-T_i)(1-f(\mathrm{X}_i))}_{\text{PAV of ITR}} \underbrace{- \frac{\hat{p}_f}{n_1}\sum_{i=1}^{n} Y_i T_i - \frac{1-\hat{p}_f}{n_0}\sum_{i=1}^{n} Y_i(1-T_i)\Bigg]}_{\substack{\text{PAV of random treatment rule with} \\ \text{the same treated proportion}}}$$

  where $\hat{p}_f = \sum_{i=1}^{n} f(\mathrm{X}_i)/n$

- We also derive its variance, and propose its consistent estimator
- Not invariant to additive transformation: $Y_i + c$
- Solution: centering $\mathbb{E}(Y_i(1) + Y_i(0)) = 0 \rightsquigarrow$ minimum variance

Estimating and Evaluating ITRs via Cross-Fitting

- Estimate and evaluate an ITR using the same experimental data
- How should we account for both estimation uncertainty and evaluation uncertainty under the Neyman's framework?
- Setup:
  - Learning algorithm
$$F : \mathcal{Z} \longrightarrow \mathcal{F}$$
  - K-fold cross-fitting: $\mathcal{Z} = \{\mathcal{Z}_1, \mathcal{Z}_2, \ldots, \mathcal{Z}_K\}$
$$\hat{f}_{-k} \;=\; F(\mathcal{Z}_1, \mathcal{Z}_2, \ldots, \mathcal{Z}_{k-1}, \mathcal{Z}_{k+1}, \ldots, \mathcal{Z}_K)$$
  - Evaluation metric estimators:
$$\hat{\lambda}_F \;=\; \frac{1}{K}\sum_{k=1}^{K} \hat{\lambda}_{\hat{f}_{-k}}(\mathcal{Z}_k), \quad \hat{\tau}_F \;=\; \frac{1}{K}\sum_{k=1}^{K} \hat{\tau}_{\hat{f}_{-k}}(\mathcal{Z}_k)$$
- Uncertainty over both evaluation data and all random sets of training data (of a fixed size) as well as treatment assignment

Causal Estimands under Cross-fitting

- Population Average Value (PAV)
  - Generalized ITR averaging over the random sampling of training data $\mathcal{Z}^{tr}$ (due to random splitting)

$$\bar{f}_F(\mathsf{x}) \;=\; \mathbb{E}\{\hat{f}_{\mathcal{Z}^{tr}}(\mathsf{x}) \mid \mathsf{X}_i = \mathsf{x}\} \;=\; \Pr(\hat{f}_{\mathcal{Z}^{tr}}(\mathsf{x}) = 1 \mid \mathsf{X}_i = \mathsf{x})$$

  - Estimand

$$\lambda_F \;=\; \mathbb{E}\left\{\bar{f}_F(\mathsf{X}_i)\, Y_i(1) + (1 - \bar{f}_F(\mathsf{X}_i))\, Y_i(0)\right\}$$

- Population Average Prescriptive Effect (PAPE)
  - Proportion treated

$$p_F \;=\; \mathbb{E}\{\bar{f}_F(\mathsf{X}_i)\}.$$

  - Estimand

$$\tau_F \;=\; \mathbb{E}\{\lambda_F - p_F\, Y_i(1) - (1 - p_F)\, Y_i(0)\}.$$

Inference under Cross-Fitting

- Under Neyman's framework, the cross-fitting estimators are unbiased, i.e., $\mathbb{E}(\hat{\lambda}_F) = \lambda_F$ and $\mathbb{E}(\hat{\tau}_F) = \tau_F$
- The variance of the PAV estimator

$$\mathbb{V}(\hat{\lambda}_F) = \underbrace{\frac{\mathbb{E}(S^2_{\hat{f}1})}{m_1} + \frac{\mathbb{E}(S^2_{\hat{f}0})}{m_0}}_{\text{evaluation uncertainty}} + \underbrace{\mathbb{E}\left\{\mathrm{Cov}(\hat{f}_{\mathcal{Z}^{tr}}(\mathbf{X}_i), \hat{f}_{\mathcal{Z}^{tr}}(\mathbf{X}_j) \mid \mathbf{X}_i, \mathbf{X}_j)\tau_i\tau_j\right\}}_{\text{estimation uncertainty}}$$
$$- \underbrace{\frac{K-1}{K}\mathbb{E}(S^2_F)}_{\substack{\textit{efficiency gain due} \\ \textit{to cross-fitting}}}$$

  for $i \neq j$ where m_t is the size of the training set with $T_i = t$, $\tau_i = Y_i(1) - Y_i(0)$, $S^2_F = \sum_{k=1}^{K}\left\{\hat{\lambda}_{\hat{f}_{-k}}(\mathcal{Z}_k) - \overline{\hat{\lambda}_{\hat{f}_{-k}}(\mathcal{Z}_k)}\right\}^2/(K-1)$
- Analogous results for the PAPE τ_F

Evaluation with a Budget Constraint

- Policy makers often face a binding budget constraint p
- Scoring rule:

$$s : \mathcal{X} \longrightarrow \mathcal{S} \quad \text{where} \quad \mathcal{S} \subset \mathbb{R}$$

- Example: CATE $s(\mathrm{x}) = \mathbb{E}(Y_i(1) - Y_i(0) \mid \mathrm{X}_i = \mathrm{x})$
- (Fixed) ITR with a budget constraint:

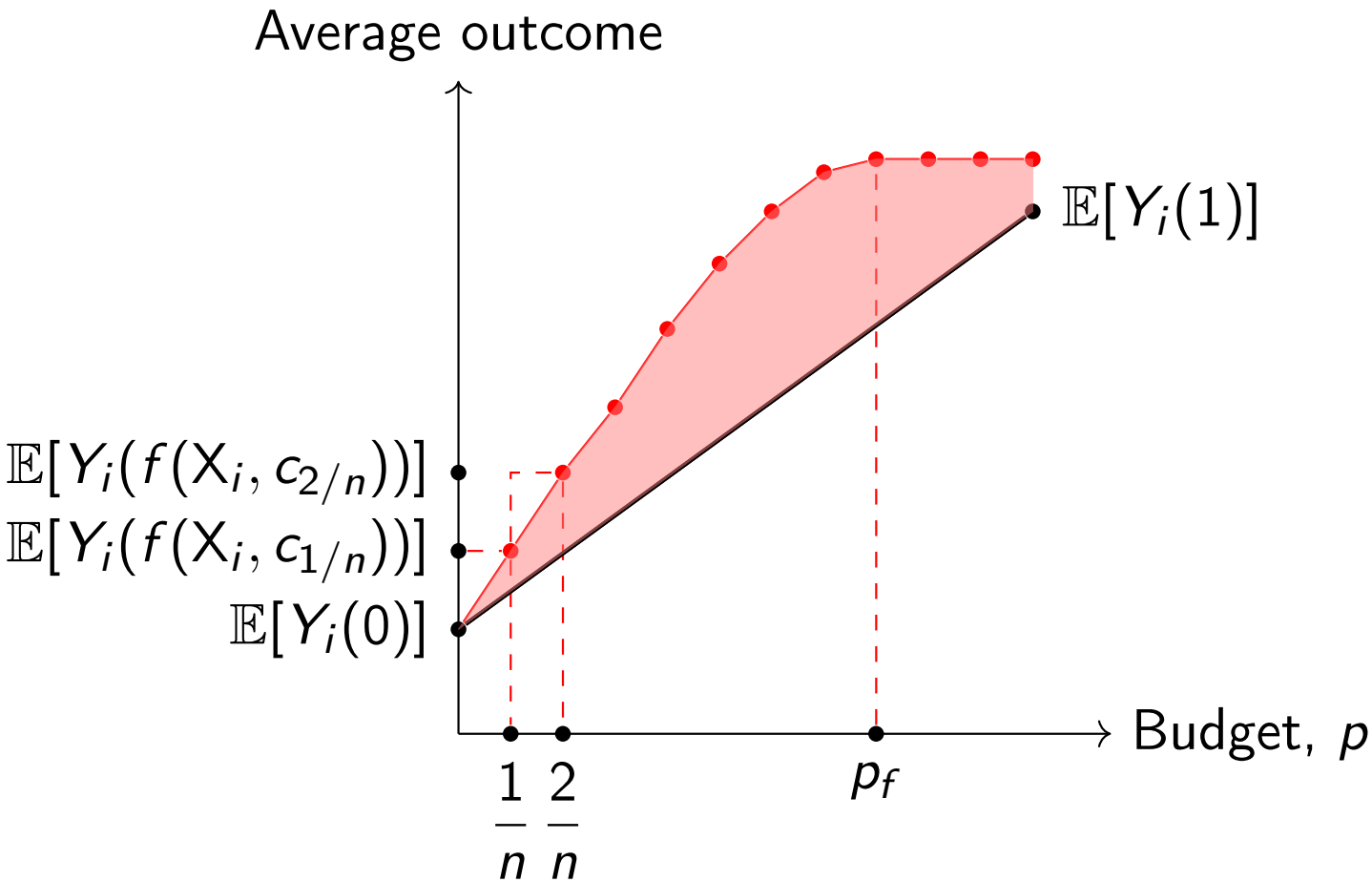
$$f(\mathrm{X}_i, c) \;=\; 1\{s(\mathrm{X}_i) > c\},$$

  where $c_p(f) \;=\; \inf\{c \in \mathbb{R} : \Pr(f(\mathrm{X}_i, c) = 1) \leq p\}$
- PAPE under a budget constraint

$$\tau_{fp} \;=\; \mathbb{E}\{\, Y_i(f(\mathrm{X}_i, c_p(f))) - pY_i(1) - (1-p)\, Y_i(0)\}.$$

- We derive the bias (and its finite sample bound) and variance under the Neyman's framework
- Extensions: cross-fitting, diff. in PAPE between two ITRs

The Area Under Prescriptive Effect Curve (AUPEC)



- Measure of performance across different budget constraints
- We show how to do inference with and without cross-fitting
- Normalized AUPEC = average percentage gain using an ITR over the randomized treatment rule across a range of budget contraints

Simulations

- Atlantic Causal Inference Conference data analysis challenge
- Data generating process
  - 8 covariates from the Infant Health and Development Program (originally, 58 covariates and 4,302 observations)
  - population distribution = original empirical distribution
  - Model

$$Y_i(t) = \mu(X_i) + \tau(X_i)t + \sigma(X_i)\epsilon_i,$$

where $t = 0, 1$, $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, and

$$\begin{aligned}
\mu(x) &= -\sin(\Phi(\pi(x))) + x_{43}, \\
\pi(x) &= 1/[1 + \exp\{3(x_1 + x_{43} + 0.3(x_{10} - 1)) - 1\}], \\
\tau(x) &= \xi(x_3 x_{24} + (x_{14} - 1) - (x_{15} - 1)), \\
\sigma(x) &= 0.25\sqrt{\mathbb{V}(\mu(x) + \pi(x)\tau(x))}.
\end{aligned}$$

- Two scenarios: large vs. small treatment effects $\xi \in \{2, 1/3\}$
- Sample sizes: $n \in \{100, 500, 2,000\}$

Results I: Fixed ITR

- No budget constraint, 20% constraint
- f: Bayesian Additive Regression Tree (BART)
- g: Causal Forest
- h: LASSO

		$n = 100$			$n = 500$			$n = 2000$		
Estimator	truth	cov.	bias	s.d.	cov.	bias	s.d.	cov.	bias	s.d.
Small effect										
$\hat{\tau}_f$	0.066	94.3	0.005	0.124	96.2	0.001	0.053	95.1	0.001	0.026
$\hat{\tau}_f(c_{0.2})$	0.051	93.2	−0.002	0.109	94.4	0.001	0.046	95.2	0.002	0.021
$\widehat{\Gamma}_f$	0.053	95.3	0.001	0.106	95.1	0.001	0.045	94.8	−0.001	0.024
$\widehat{\Delta}_{0.2}(f,g)$	−0.022	94.0	0.006	0.122	95.4	0.002	0.051	96.0	0.000	0.026
$\widehat{\Delta}_{0.2}(f,h)$	−0.014	93.9	−0.001	0.131	94.9	−0.000	0.060	95.3	−0.000	0.030
Large effect										
$\hat{\tau}_f$	0.430	94.7	−0.000	0.163	95.7	0.000	0.064	94.4	-0.000	0.031
$\hat{\tau}_f(c_{0.2})$	0.356	94.7	0.004	0.159	95.7	0.002	0.072	95.8	0.000	0.035
$\widehat{\Gamma}_f$	0.363	94.3	−0.005	0.130	94.9	0.003	0.058	95.7	0.000	0.029
$\widehat{\Delta}_{0.2}(f,g)$	−0.000	96.9	0.008	0.151	97.9	−0.002	0.073	98.0	−0.000	0.026
$\widehat{\Delta}_{0.2}(f,h)$	0.000	94.7	−0.004	0.140	97.7	−0.001	0.065	96.6	0.000	0.033

Results II: Estimated ITR

- 5-fold cross fitting
- F: LASSO
- std. dev. for $n = 500$ is roughly half of the fixed $n = 100$ case

	$n = 100$			$n = 500$			$n = 2000$		
Estimator	cov.	bias	s.d.	cov.	bias	s.d.	cov.	bias	s.d.
Small effect									
$\hat{\lambda}_F$	96.4	0.001	0.216	96.7	0.002	0.100	97.2	0.002	0.046
$\hat{\tau}_F$	94.6	−0.002	0.130	95.5	−0.002	0.052	94.4	−0.000	0.027
$\hat{\tau}_F(c_{0.2})$	95.4	−0.003	0.120	95.4	−0.002	0.043	96.8	0.001	0.029
$\widehat{\Gamma}_F$	98.2	0.002	0.117	96.8	−0.001	0.048	95.9	0.001	0.001
Large effect									
$\hat{\lambda}_H$	96.9	−0.007	0.261	96.5	−0.003	0.125	97.3	0.001	0.062
$\hat{\tau}_F$	93.6	−0.000	0.171	93.0	0.000	0.093	95.3	0.001	0.041
$\hat{\tau}_F(c_{0.2})$	94.8	−0.002	0.170	96.2	−0.005	0.075	95.8	0.001	0.037
$\widehat{\Gamma}_F$	98.5	0.001	0.126	98.9	0.005	0.053	99.0	0.001	0.026

Application to the STAR Experiment

- Experiment involving 7,000 students across 79 schools
- Randomized treatments (kindergarden):
  1. $T_i = 1$: small class (13–17 students)
  2. $T_i = 0$: regular class (22–25)
  3. regular class with aid
- Outcome: SAT scores
- Literature on heterogeneous treatments in labor economics
- 10 covariates
  - 4 demographics: gender, race, birth month, birth year
  - 6 school characteristics: urban/rural, enrollment size, grade range, number of students on free lunch, percentage white, number of students on school buses
- Sample size: $n = 1,911$, 5-fold cross-fitting
- Average Treatment Effects:
  - SAT reading: 6.78 (s.e.=1.71)
  - SAT math: 5.78 (s.e.=1.80)

Results I: ITR Performance

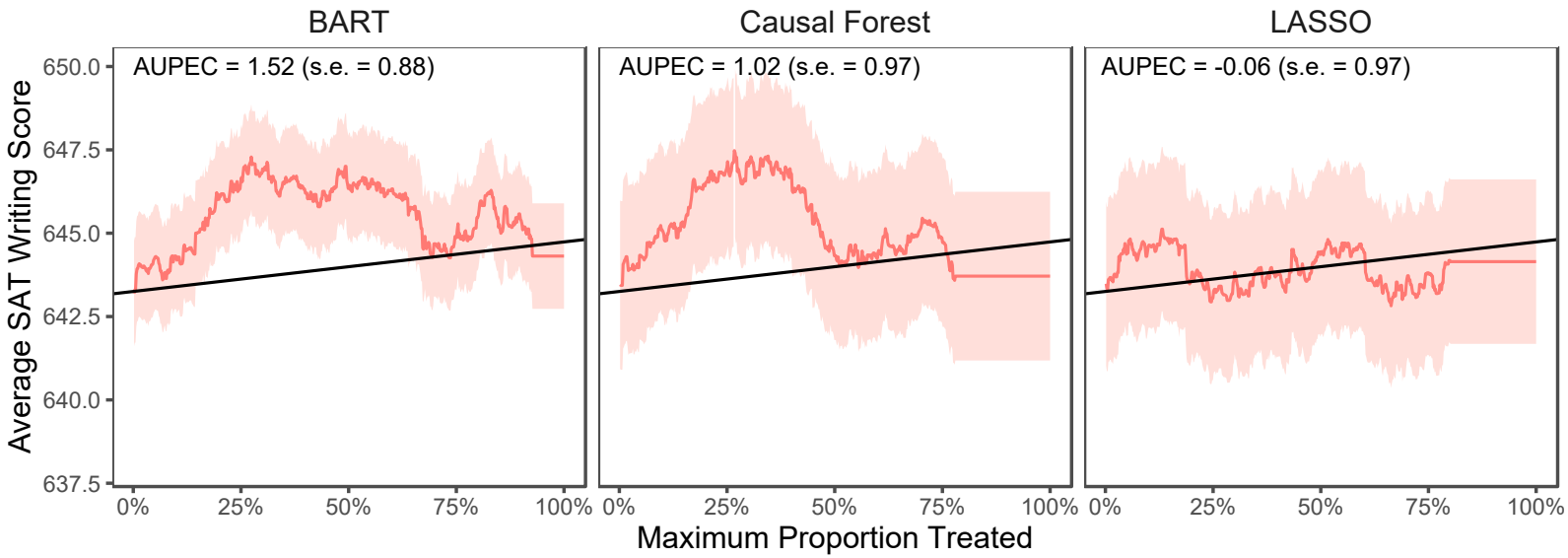
	BART			Causal Forest			LASSO		
	est.	s.e.	treated	est.	s.e.	treated	est.	s.e.	treated
Fixed ITR									
No budget constraint									
Reading	0	0	100%	−0.38	1.14	84.3%	−0.41	1.10	84.4%
Math	0.52	1.09	86.7	0.09	1.18	80.3	1.73	1.25	78.7
Writing	−0.32	0.72	92.7	−0.70	1.18	78.0	−0.30	1.26	80.0
Budget constraint									
Reading	−0.89	1.30	20	0.66	1.23	20	−1.17	1.18	20
Math	0.70	1.25	20	2.57	1.29	20	1.25	1.32	20
Writing	2.60	1.17	20	2.98	1.18	20	0.28	1.19	20
Estimated ITR									
No budget constraint									
Reading	0.19	0.37	99.3%	0.31	0.77	86.6%	0.32	0.53	87.6%
Math	0.92	0.75	84.7	2.29	0.80	79.1	1.52	1.60	75.2
Writing	1.12	0.86	88.0	1.43	0.71	67.4	0.05	1.37	74.8
Budget constraint									
Reading	1.55	1.05	20	0.40	0.69	20	−0.15	1.41	20
Math	2.28	1.15	20	1.84	0.73	20	1.50	1.48	20
Writing	2.31	0.66	20	1.90	0.64	20	−0.47	1.34	20

Results II: Comparison between ML Algorithms

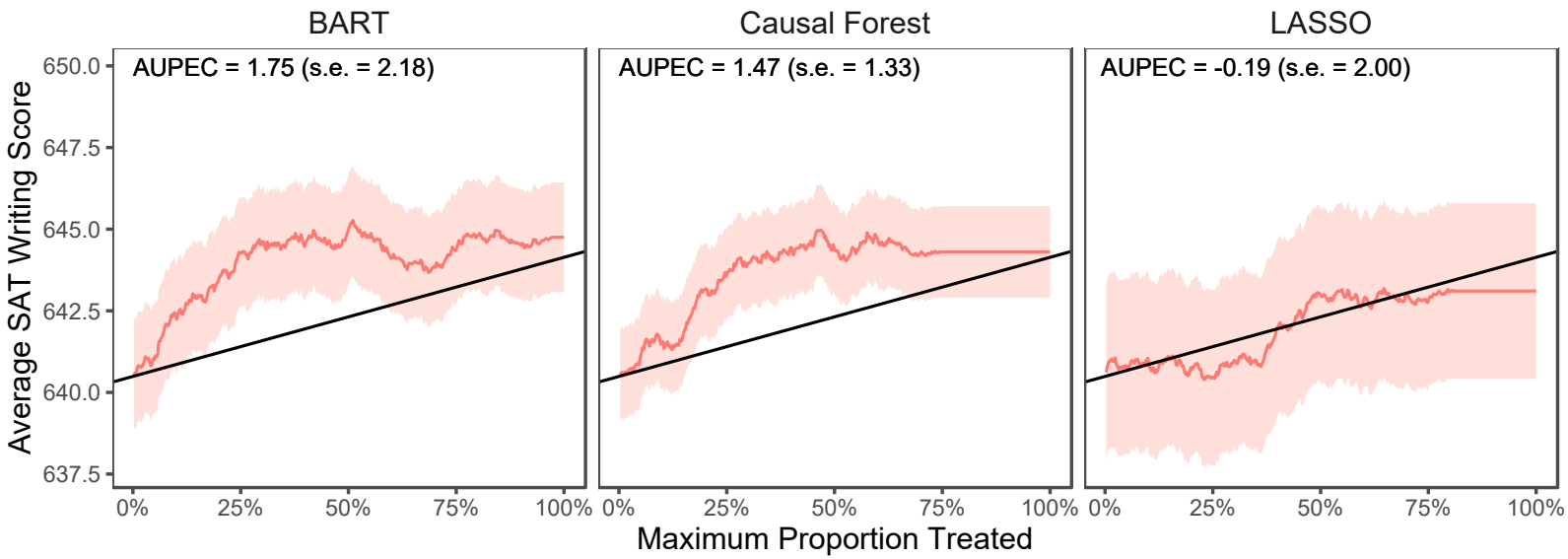
	Causal Forest vs. BART		Causal Forest vs. LASSO		BART vs. LASSO	
	est.	95% CI	est.	95% CI	est.	95% CI
Fixed ITR						
Math	1.55	[−0.35, 3.45]	1.83	[−0.50, 4.16]	0.28	[−2.39, 2.95]
Reading	1.86	[−0.79, 4.51]	1.31	[−1.49, 4.11]	−0.55	[−4.02, 2.92]
Writing	0.38	[−1.66, 2.42]	2.69	[−0.27, 5.65]	2.32	[−0.53, 5.15]
Estimated ITR						
Reading	−1.15	[−3.99, 1.69]	0.55	[−1.05, 2.15]	1.70	[−0.90, 4.30]
Math	−0.43	[−2.57, 3.43]	0.34	[−1.32, 2.00]	0.77	[−1.99, 3.53]
Writing	−0.41	[−1.63, 0.80]	2.37	[0.76, 3.98]	2.79	[1.32, 4.26]

Results III: AUPEC

Fixed ITR



Estimated ITR



Evaluation of Heterogeneous Treatment Effects

- Another popular use of ML in causal inference
- Estimation of heterogeneous treatment effects: random forest, BART, Lasso, etc.

- How can we make valid inference for heterogeneous treatment effects discovered via a generic ML algorithm?
  - cannot assume ML algorithms converge uniformly
  - avoid computationally intensive method (e.g., repeated cross-fitting)
  - use Neyman's repeated sampling framework for inference

- Imai and Li. "Statistical Inference for Heterogeneous Treatment Effects Discovered by Generic Machine Learning in Randomized Experiments." `https://arxiv.org/pdf/2203.14511.pdf`

Setup and Causal Quantities of Interest

- Conditional Average Treatment Effect (CATE):

$$\tau(\mathrm{x}) \;=\; \mathbb{E}(Y_i(1) - Y_i(0) \mid \mathrm{X}_i = \mathrm{x})$$

- CATE estimation based on ML algorithm

$$s : \mathcal{X} \longrightarrow \mathcal{S} \subset \mathbb{R}$$

- Sorted Group Average Treatment Effect (GATES; Chernozhukov et al. 2019)

$$\tau_k \;:=\; \mathbb{E}(Y_i(1) - Y_i(0) \mid c_{k-1}(s) \leq s(\mathrm{X}_i) < c_k(s))$$

for $k = 1, 2, \ldots, K$ where c_k represents the cutoff between the $(k-1)$th and kth groups

GATES Estimation as ITR Evaluation

- A natural GATES estimator

$$\hat{\tau}_k \;=\; \frac{K}{n_1}\sum_{i=1}^{n} Y_i T_i \hat{f}_k(\mathrm{X}_i) - \frac{K}{n_0}\sum_{i=1}^{n} Y_i(1-T_i)\hat{f}_k(\mathrm{X}_i),$$

  where $\hat{f}_k(\mathrm{X}_i) \;=\; 1\{s(\mathrm{X}_i) \geq \hat{c}_k(s)\} - 1\{s(\mathrm{X}_i) \geq \hat{c}_{k-1}(s)\}$

- Rewrite this as the PAPE:

$$\hat{\tau}_k \;=\; K\Bigg\{\underbrace{\frac{1}{n_1}\sum_{i=1}^{n} Y_i T_i \hat{f}_k(\mathrm{X}_i) + \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-T_i)(1-\hat{f}_k(\mathrm{X}_i))}_{\text{estimated PAV}} - \underbrace{\frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-T_i)}_{\text{no one gets treated}}\Bigg\}$$

- We can use our previous results!
- Inference for GATES under cross-fitting

Two Nonparametric Tests of Heterogeneity

1. Treatment effect heterogeneity:
   - Null hypothesis
   $$H_0 : \hat{\boldsymbol{\tau}} = (\hat{\tau}_1 - \hat{\tau}, \cdots, \hat{\tau}_K - \hat{\tau})^\top$$
   - Reference distribution
   $$\hat{\boldsymbol{\tau}}^\top \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\tau}} \xrightarrow{d} \chi^2_K$$

2. Rank-consistent treatment effect heterogeneity:
   - Null hypothesis
   $$H_0^* : \tau_1 \leq \tau_2 \leq \cdots \leq \tau_K$$
   - Reference distribution
   $$\left(\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}^*(\hat{\boldsymbol{\tau}})\right)^\top \boldsymbol{\Sigma}^{-1} \left(\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}^*(\hat{\boldsymbol{\tau}})\right) \xrightarrow{d} \bar{\chi}^2_K$$

   where
   $$\boldsymbol{\mu}^*(\boldsymbol{x}) = \underset{\boldsymbol{\mu}}{\operatorname{argmin}} \|\boldsymbol{\mu} - \boldsymbol{x}\|_2^2 \quad \text{subject to } \mu_1 \leq \mu_2 \leq \cdots \leq \mu_K,$$

   with $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_K)^\top$ and $\boldsymbol{x} \in \mathbb{R}^K$

A Simulation Study

- 2016 ACIC competition (Dorie *et al.*, 2019)
- Sample size $n = 4,802$ and 58 covariates, taken from a real study
- We generate data sets using their data generating process

- Sample size: $n = 100, 500$, and $2,500$
- Number of groups: $K = 5$
- Sample splitting: trained on the original ACIC data
- Cross-fitting: 5-fold
- ML algorithms: BART, Causal Forest, and Lasso
- Finite sample properties (sample splitting and cross-fitting)
  1. GATES estimation
  2. Nonparametric tests (treatment effect homogeneity $\rightsquigarrow$ false; rank-consistency $\rightsquigarrow$ true)

Cross-Fitting Case: GATES

	$n = 100$				$n = 500$				$n = 2500$			
Estimator	truth	bias	s.d.	coverage	truth	bias	s.d.	coverage	truth	bias	s.d.	coverage
Causal Forest												
$\hat{\tau}_1$	3.976	−0.053	2.971	94.0%	2.900	−0.007	1.572	95.6%	2.210	−0.007	0.594	97.7%
$\hat{\tau}_2$	4.173	−0.061	2.584	95.9	4.112	−0.038	1.075	98.2	4.057	0.011	0.541	98.6
$\hat{\tau}_3$	4.286	−0.012	2.560	96.7	4.510	−0.054	1.058	97.7	4.545	0.019	0.465	98.1
$\hat{\tau}_4$	4.400	−0.119	2.865	97.4	4.799	0.066	1.149	97.9	4.951	−0.009	0.509	98.6
$\hat{\tau}_5$	4.569	0.140	3.447	94.1	5.086	0.001	1.620	96.0	5.643	−0.006	0.620	98.3
LASSO												
$\hat{\tau}_1$	4.191	−0.125	3.196	97.6%	4.017	−0.025	1.488	96.0%	3.752	−0.004	0.669	96.0%
$\hat{\tau}_2$	4.205	0.036	2.281	97.5	4.137	−0.069	1.027	97.9	4.028	−0.019	0.590	98.9
$\hat{\tau}_3$	4.268	−0.126	2.354	96.6	4.291	−0.019	1.000	97.9	4.323	0.037	0.488	97.5
$\hat{\tau}_4$	4.334	−0.003	2.536	96.8	4.430	0.035	1.174	96.8	4.571	0.033	0.642	97.2
$\hat{\tau}_5$	4.406	0.111	3.615	96.2	4.530	0.047	1.811	95.0	4.732	0.022	0.697	95.3

Cross-Fitting Case: Nonparametric Tests

	$n = 100$		$n = 500$		$n = 2500$	
	rejection rate	median p-value	rejection rate	median p-value	rejection rate	median p-value
Causal Forest						
Homogeneous Treatment Effects	1.4%	0.790	4.6%	0.712	51.4%	0.041
Consistent Treatment Effects	1.4%	0.702	0.8%	0.845	0.0%	0.976
LASSO						
Homogeneous Treatment Effects	0.6%	0.880	1.8%	0.850	9.0%	0.664
Consistent Treatment Effects	1.0%	0.722	0.6%	0.769	0.2%	0.889

Empirical Application

- National Supported Work Demonstration Program (LaLonde 1986)
- Temporary employment program to help disadvantaged workers by giving them a guaranteed job for 9 to 18 months

- Data
  - sample size: $n_1 = 297$ and $n_0 = 425$
  - outcome: annualized earnings in 1978 (36 months after the program)
  - 7 pre-treatment covariates: demographics and prior earnings

- Setup
  - ML algorithms: Causal Forest, BART, and LASSO
  - Sample-splitting: 2/3 of the data as training data
  - Cross-fitting: 3 folds
  - 5 fold cross-validation for tuning parameters

GATES Estimates (in 1,000 US Dollars)

	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_3$	$\hat{\tau}_4$	$\hat{\tau}_5$
Sample-splitting					
Causal Forest	3.40	0.13	−0.85	−1.91	7.21
	[−1.29, 3.40]	[−5.37, 5.63]	[−5.22, 3.52]	[−5.16, 1.34]	[1.22, 13.19]
BART	2.90	−0.73	−0.02	3.25	2.57
	[−2.25, 8.06]	[−5.05, 3.58]	[−3.47, 3.43]	[−1.53, 8.03]	[−3.82, 8.97]
LASSO	1.86	2.62	−2.07	1.39	4.17
	[−3.59, 7.30]	[−1.69, 6.93]	[−5.39, 1.26]	[−2.95, 5.73]	[−2.30, 10.65]
Cross-fitting					
Causal Forest	−3.72	1.05	5.32	−2.64	4.55
	[−6.52, −0.93]	[−2.28, 4.37]	[2.63, 8.01]	[−5.07, −0.22]	[1.14, 7.96]
BART	0.40	−0.15	−0.40	2.52	2.19
	[−3.79, 4.59]	[−2.54, 2.23]	[−3.37, 2.56]	[−0.99, 6.03]	[−0.73, 5.11]
LASSO	0.65	0.45	−2.88	1.32	5.02
	[−3.65, 4.94]	[−3.28, 4.18]	[−5.38, −0.38]	[−1.83, 4.48]	[−0.14, 10.18]

Nonparametric Tests

	Causal Forest		**BART**		**LASSO**	
	stat	p-value	stat	p-value	stat	p-value
Sample-splitting						
Homogeneous Treatment Effects	9.78	0.082	2.76	0.737	5.26	0.362
Rank-consistent Treatment Effects	3.07	0.323	1.13	0.657	3.14	0.302
Cross-fitting						
Homogeneous Treatment Effects	30.29	0.000	2.32	0.803	10.79	0.056
Rank-consistent Treatment Effects	0.06	0.691	0.04	0.885	0.45	0.711

Concluding Remarks

- Causal machine learning is everywhere
  - estimation of heterogeneous treatment effects (HTEs)
  - development of individualized treatment rules (ITRs)
- Inference about HTEs and ITRs has been largely model-based
  - We show how to experimentally evaluate HTEs and ITRs
  - No modeling assumption or asymptotic approximation is required
  - Complex machine learning algorithms can be used
  - Applicable to cross-fitting estimators
  - Simulations: good small sample performance
- Ongoing extension: dynamic ITRs
- Open source software: evalITR: Evaluating Individualized Treatment Rules at CRAN `https://CRAN.R-project.org/package=evalITR`