Minimax rates for heterogeneous effect estimation

Edward Kennedy

Department of Statistics & Data Science
Carnegie Mellon University

U Melbourne, 1 May 2023

Punchline

There have been many many proposals in recent years for flexible estimation of *heterogeneous causal effects*

- but crucial theoretical gaps remain, especially when effects have nontrivial structure (e.g., smoothness/sparsity)
- can current methods be improved? what is the best possible error one could achieve?

The goals of this work are: *flexible estimators + minimax rates*

1. more flexible estimators, with better error guarantees
   - (Kennedy, 2020)
2. resolve open question of minimax optimality
   - (Kennedy, Balakrishnan, Robins, & Wasserman, 2022)

Heterogeneous Causal Effects

Treatments/policies are often studied at population level

- i.e., the average outcome if *all versus none* were treated

However this can obscure important heterogeneity

- effect may be zero on average - but in theory could be benefitting some and harming others

Why should we care about heterogeneity?

- improve understanding of variation, help inform policy & optimize treatment decisions
- critical across medicine & social sciences

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 DECEMBER 31, 2020 VOL. 383 NO. 27

Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine

Fernando P. Polack, M.D., Stephen J. Thomas, M.D., Nicholas Kitchin, M.D., Judith Absalon, M.D., Alejandra Gurtman, M.D., Stephen Lockhart, D.M., John L. Perez, M.D., Gonzalo Pérez Marc, M.D., Edson D. Moreira, M.D., Cristiano Zerbini, M.D., Ruth Bailey, B.Sc., Kena A. Swanson, Ph.D., Satrajit Roychoudhury, Ph.D., Kenneth Koury, Ph.D., Ping Li, Ph.D., Warren V. Kalina, Ph.D., David Cooper, Ph.D., Robert W. Frenck, Jr., M.D., Laura L. Hammitt, M.D., Özlem Türeci, M.D., Haylene Nell, M.D., Axel Schaefer, M.D., Serhat Ünal, M.D., Dina B. Tresnan, D.V.M., Ph.D., Susan Mather, M.D., Philip R. Dormitzer, M.D., Ph.D., Uğur Şahin, M.D., Kathrin U. Jansen, Ph.D., and William C. Gruber, M.D., for the C4591001 Clinical Trial Group*

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 FEBRUARY 4, 2021 VOL. 384 NO. 5

Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine

L.R. Baden, H.M. El Sahly, B. Essink, K. Kotloff, S. Frey, R. Novak, D. Diemert, S.A. Spector, N. Rouphael, C.B. Creech, J. McGettigan, S. Khetan, N. Segall, J. Solis, A. Brosz, C. Fierro, H. Schwartz, K. Neuzil, L. Corey, P. Gilbert, H. Janes, D. Follmann, M. Marovich, J. Mascola, L. Polakowski, J. Ledgerwood, B.S. Graham, H. Bennett, R. Pajon, C. Knightly, B. Leav, W. Deng, H. Zhou, S. Han, M. Ivarsson, J. Miller, and T. Zaks, for the COVE Study Group*

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Safety and Efficacy of Single-Dose Ad26.COV2.S Vaccine against Covid-19

J. Sadoff, G. Gray, A. Vandebosch, V. Cárdenas, G. Shukarev, B. Grinsztejn, P.A. Goepfert, C. Truyers, H. Fennema, B. Spiessens, K. Offergeld, G. Scheper, K.L. Taylor, M.L. Robb, J. Treanor, D.H. Barouch, J. Stoddard, M.F. Ryser, M.A. Marovich, K.M. Neuzil, L. Corey, N. Cauwenberghs, T. Tanner, K. Hardt, J. Ruiz-Guiñazú, M. Le Gars, H. Schuitemaker, J. Van Hoof, F. Struyf, and M. Douoguih, for the ENSEMBLE Study Group*

Subgroup	Placebo (N=14,073)	mRNA-1273 (N=14,134)	Vaccine Efficacy (95% CI)
	no. of events/total no.		
All patients	185/14,073	11/14,134	94.1 (89.3–96.8)
Age			
≥18 to <65 yr	156/10,521	7/10,551	95.6 (90.6–97.9)
≥65 yr	29/3552	4/3583	86.4 (61.4–95.2)
Age, risk for severe Covid-19			
18 to <65 yr, not at risk	121/8403	5/8396	95.9 (90.0–98.3)
18 to <65 yr, at risk	35/2118	2/2155	94.4 (76.9–98.7)
≥65 yr	29/3552	4/3583	86.4 (61.4–95.2)
Sex			
Male	87/7462	4/7366	95.4 (87.4–98.3)
Female	98/6611	7/6768	93.1 (85.2–96.8)
At risk for severe Covid-19			
Yes	43/3167	4/3206	90.9 (74.7–96.7)
No	142/10,906	7/10,928	95.1 (89.6–97.7)
Race and ethnic group			
White	144/8916	10/9023	93.2 (87.1–96.4)
Communities of color	41/5132	1/5088	97.5 (82.2–99.7)

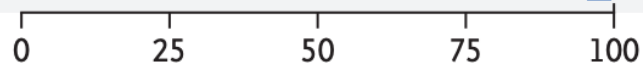
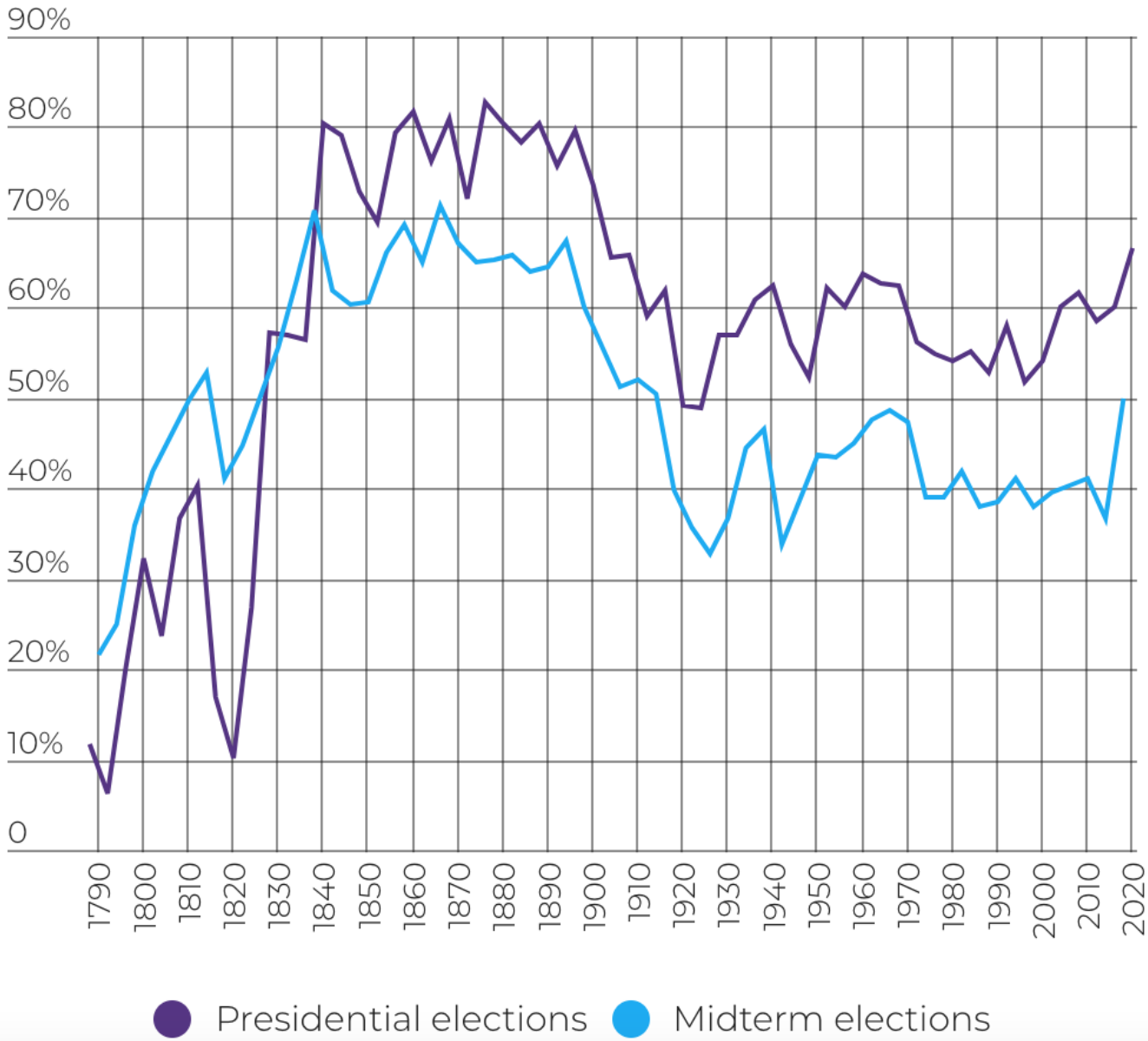


Figure 4. Vaccine Efficacy of mRNA-1273 to Prevent Covid-19 in Subgroups.

Table 3. Vaccine Efficacy against Covid-19 with Onset at Least 14 Days and at Least 28 Days after Administration of Vacc Population).*

Variable	≥14 Days after Administration†				
	Ad26.COV2.S		Placebo		Vaccine Efficacy (95% CI)
	no.	*person-yr*	*no.*	*person-yr*	%
Worldwide					
No. of participants	19,514		19,544		
Moderate to severe–critical Covid-19	173	3113.9	509	3089.1	66.3 (59.9 to 71.8)
Severe–critical Covid-19	19	3124.7	80	3121.0	76.3 (57.9 to 87.5)
United States					
No. of participants	9,119		9,086		
Moderate to severe–critical Covid-19	51	1414.0	196	1391.3	74.4 (65.0 to 81.6)
Severe–critical Covid-19	4	1417.2	18	1404.8	78.0 (33.1 to 94.6)
Brazil					
No. of participants	3,370		3,355		
Moderate to severe–critical Covid-19	39	555.7	114	548.8	66.2 (51.0 to 77.1)
Severe–critical Covid-19	2	558.9	11	556.8	81.9 (17.0 to 98.1)
South Africa					
No. of participants	2,473		2,496		
Moderate to severe–critical Covid-19	43	377.6	90	379.2	52.0 (30.3 to 67.4)
Severe–critical Covid-19	8	380.2	30	382.9	73.1 (40.0 to 89.4)

Voter turnout rates, 1789 - 2020



Get out the vote

Voters are older, wealthier, and whiter than non-voters

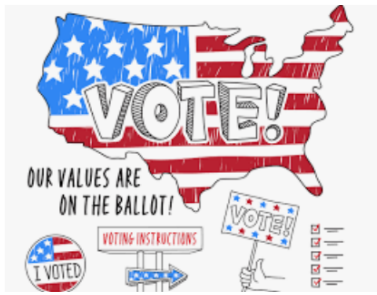


TABLE 6 Effects of Mobilization Strategies Listed from Most Effective to Least Effective

Mobilization Strategy	Effect
Face-to-Face Canvass	8%
Average Volunteer Phone Calls	3%
Text Messaging	3%
Street Signs in New York City	3%
Leaflets	1.2%
Direct Mail	0.6%
Average Commercial Phone Calls	0.55%
Robo Calls	none
E-mail	none

Note: The data in this table come from Nickerson (2007b), Green and Gerber (2004), and Panagopoulos (2009).

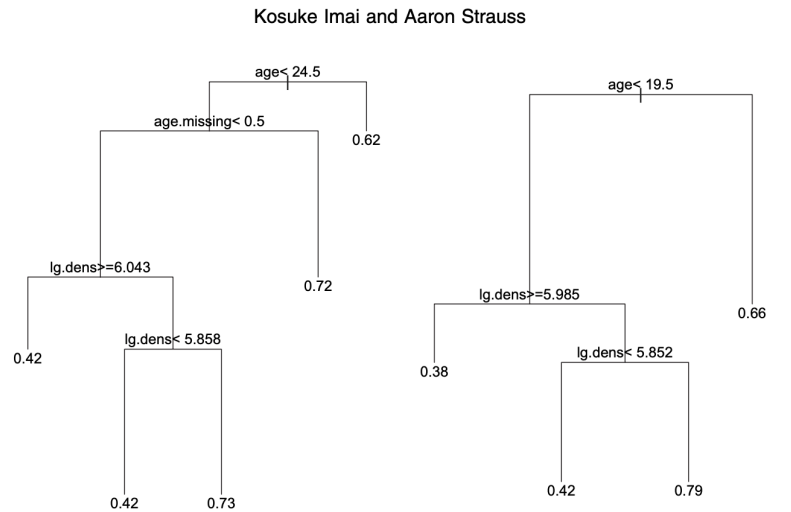


Fig. 2 Final classification trees for the control group (left panel) and treatment group (right

Setup

Consider the classic causal inference data structure:

- covariates $X \in \mathbb{R}^d$, treatment $A \in \{0, 1\}$, outcome $Y \in \mathbb{R}$

e.g., in GOTV example:

- $X =$ city, party affiliation, voting history, age, family size, race
- $A =$ whether contacted by canvasser
- $Y =$ whether subject voted in local election or not

Let Y^a denote counterfactual outcome under treatment $A = a$

- e.g., $Y^1 =$ whether would've voted if contacted, $Y^0 =$ if not

Targets & identification

The average treatment effect (ATE) is

$$\mathbb{E}(Y^1 - Y^0)$$

i.e., the mean outcome if *all versus none* were treated

Heterogeneous effect estimation $\implies$ conditional ATE (CATE)

$$\tau(x) = \mathbb{E}(Y^1 - Y^0 \mid X = x)$$

Under standard *no unmeasured confounding* assumptions we have

$$\mathbb{E}(Y^a \mid X = x) = \mathbb{E}(Y \mid X = x, A = a) \equiv \mu_a(x)$$

and so

$$\tau(x) = \mu_1(x) - \mu_0(x)$$

Key Idea

How to estimate $\tau(x) = \mu_1(x) - \mu_0(x)$? Simplest just plugs in

$$\widehat{\tau}(x) = \widehat{\mu}_1(x) - \widehat{\mu}_0(x)$$

However this can be highly suboptimal (MSE too large)

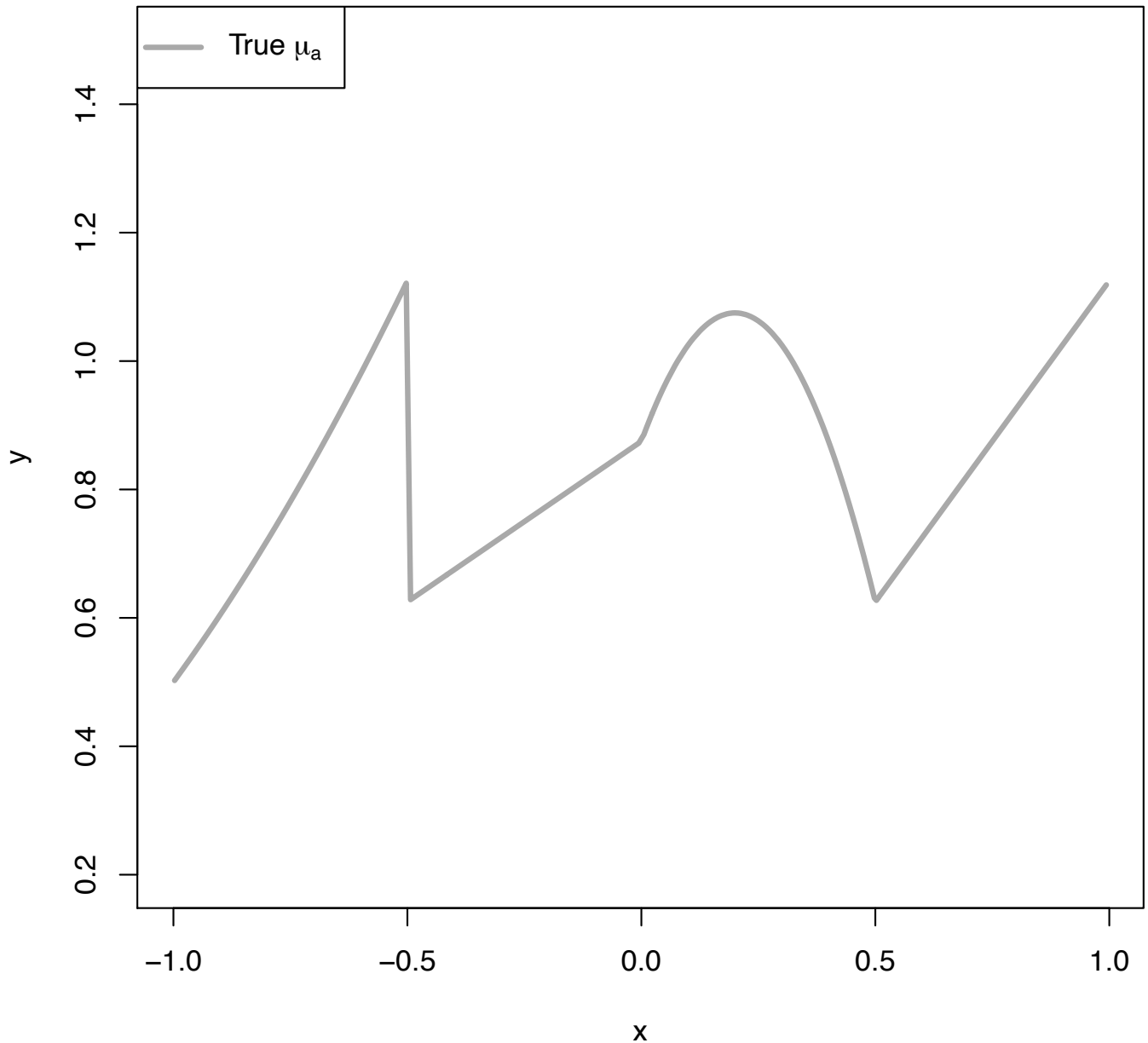
Key idea: complexity of $\tau(x)$ can be very different from $\mu_a(x)$

- μ_a is a natural outcome process that may be very complex
- CATE $\tau(x)$ could very well be constant or even null!

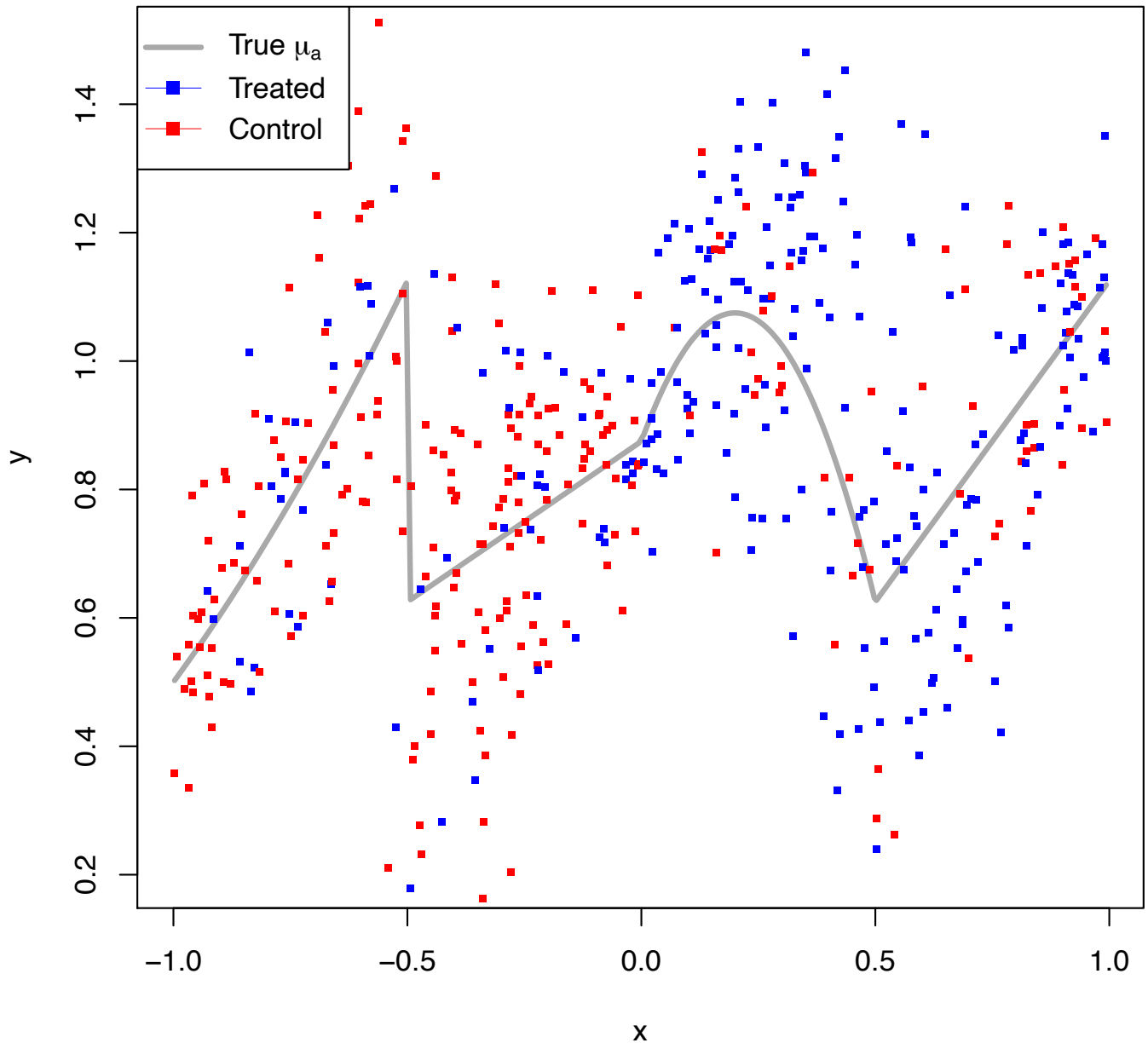
In general $\tau(x)$ has to be at least as smooth/sparse as μ_a

- but could be *much more* smooth/sparse

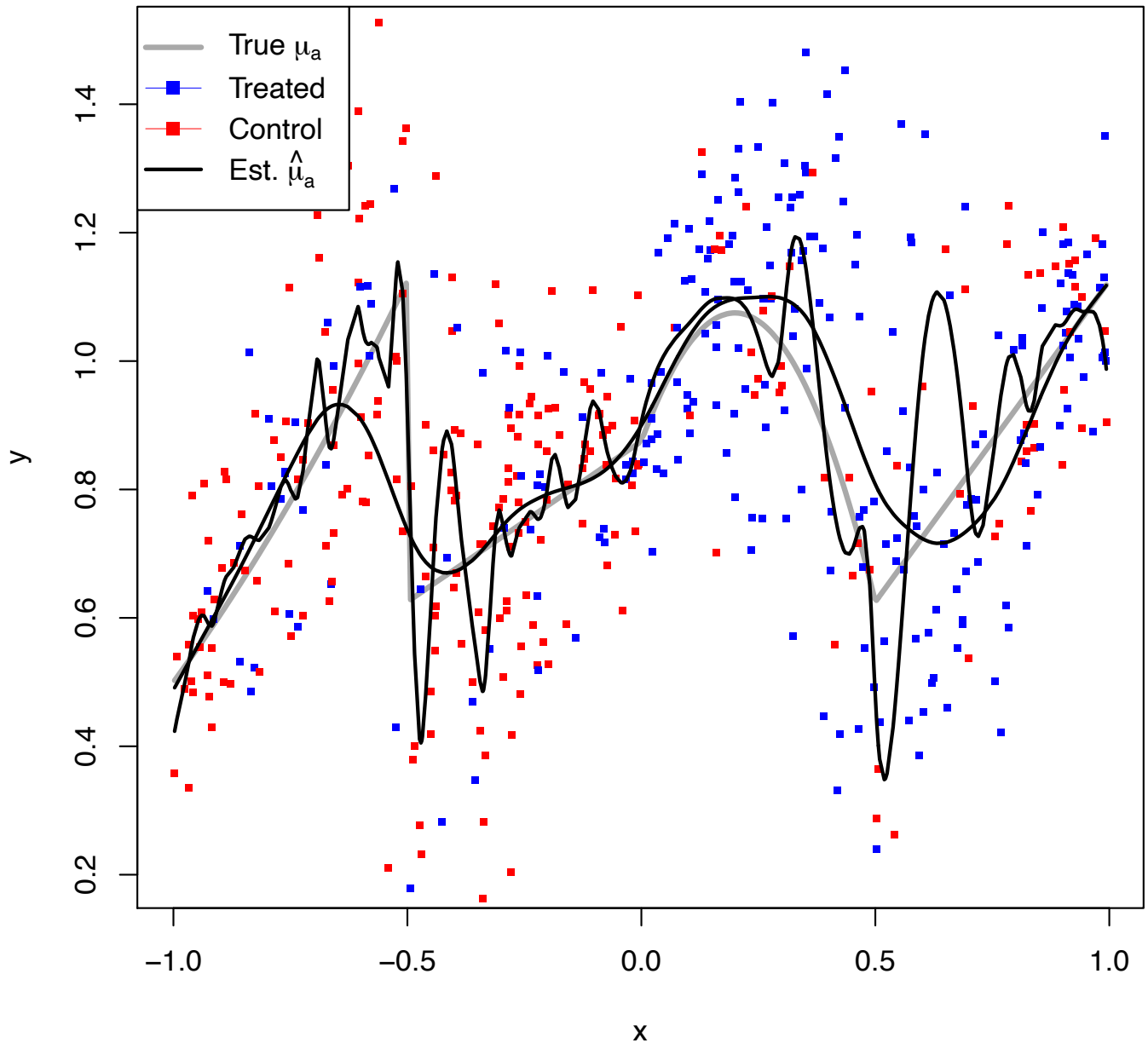
Toy Ex



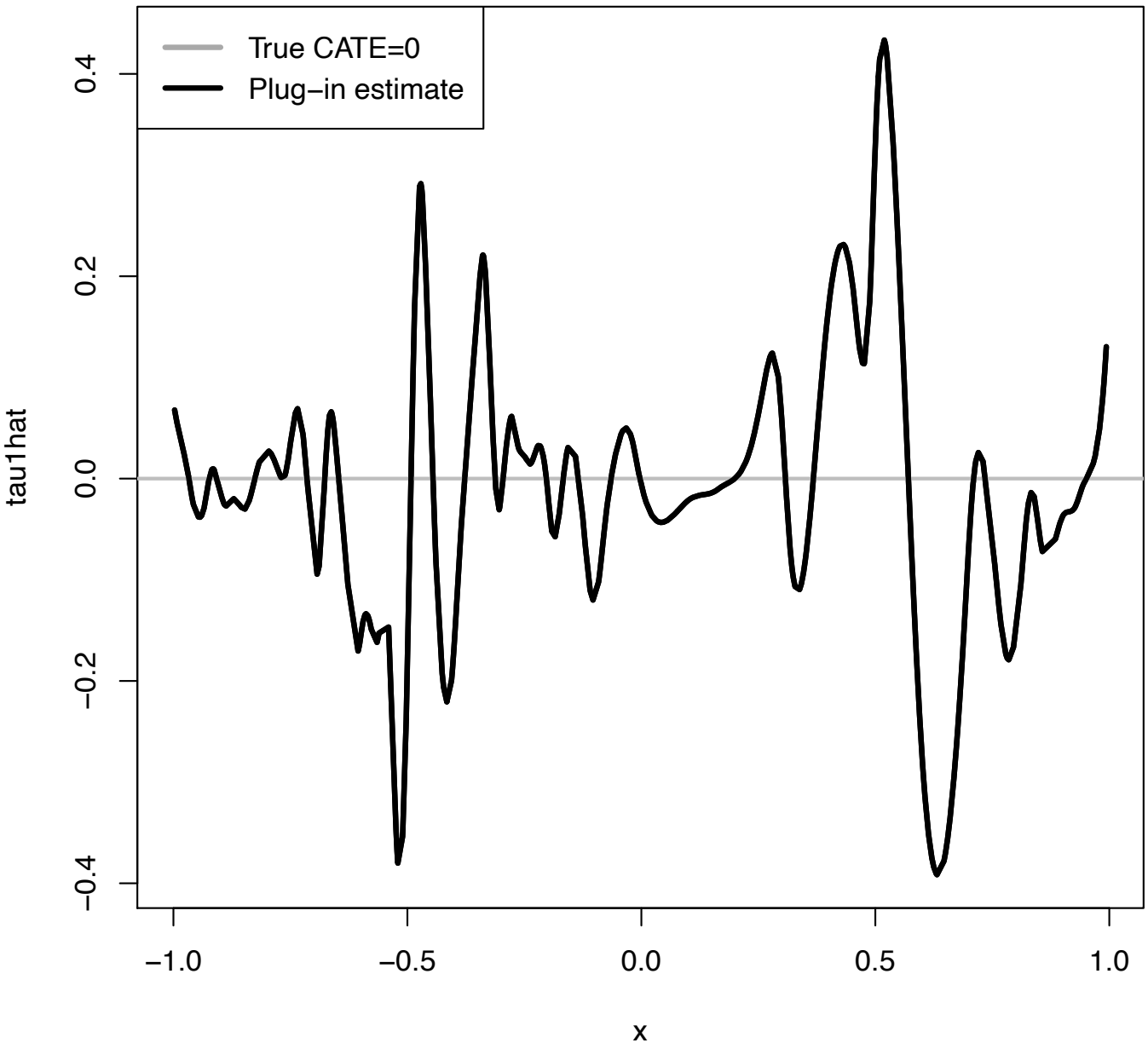
Toy Ex



Toy Ex


True μ_a
Treated
Control
Est. $\hat{\mu}_a$
y
x
0.2
0.4
0.6
0.8
1.0
1.2
1.4
−1.0
−0.5
0.0
0.5
1.0

Toy Ex


True CATE=0
Plug−in estimate
tau1hat
0.4
0.2
0.0
−0.2
−0.4
−1.0
−0.5
0.0
0.5
1.0
x

Benchmarking optimality

So plug-in will generally be deficient/suboptimal

- what's the best performance we can hope for?

First simple approach: suppose we regressed $(Y^1 - Y^0)$ on X

- this is an oracle estimator - *it knows potential outcomes*
- call this oracle $\widetilde{\tau}^*(x)$

Then if $\tau(x)$ is s-sparse, we might hope for

$$\text{RMSE}\left\{\widehat{\tau}(x)\right\} \sim \text{RMSE}\left\{\widetilde{\tau}^*(x)\right\} \sim \sqrt{\frac{s \log d}{n}}$$

Or if $\tau(x)$ is s-smooth, we might hope for

$$\text{RMSE}\left\{\widehat{\tau}(x)\right\} \sim \text{RMSE}\left\{\widetilde{\tau}^*(x)\right\} \sim n^{-\frac{1}{2+d/s}}$$

What happens with ATE?

How can we exploit any potential simplicity in τ, like oracle?

- can take intuition from ATE case (CATE $\approx$ ATE in small bin)

For estimating ATEs, a lot is known (but not all!)

- doubly robust (i.e., *semiparametric / targeted / double ML*) methods can be efficient in large nonparametric models

DR intuition: correct bias of plug-in by estimating & incorporating propensity scores $\pi(x) = \mathbb{P}(A = 1 \mid X = x)$

DR Est = Avg(*estimated pseudo-outcome*)
= Avg(regression prediction + IPW weighted residuals)

$$\widehat{\psi}_{ate} = \mathbb{P}_n \left[\widehat{\mu}_1(X) - \widehat{\mu}_0(X) + \frac{A - \widehat{\pi}(X)}{\widehat{\pi}(X)\{1 - \widehat{\pi}(X)\}} \left\{ Y - \widehat{\mu}_A(X) \right\} \right]$$

DR-Learner

Intuition: for ATEs we average, for CATEs let's regress

- we call this "DR-Learner"
- first proposed by van der Laan (2005, 2013), recently rediscovered, but little in the way of general analysis

Algorithm (DR-Learner)

Step 1. *Nuisance training:*

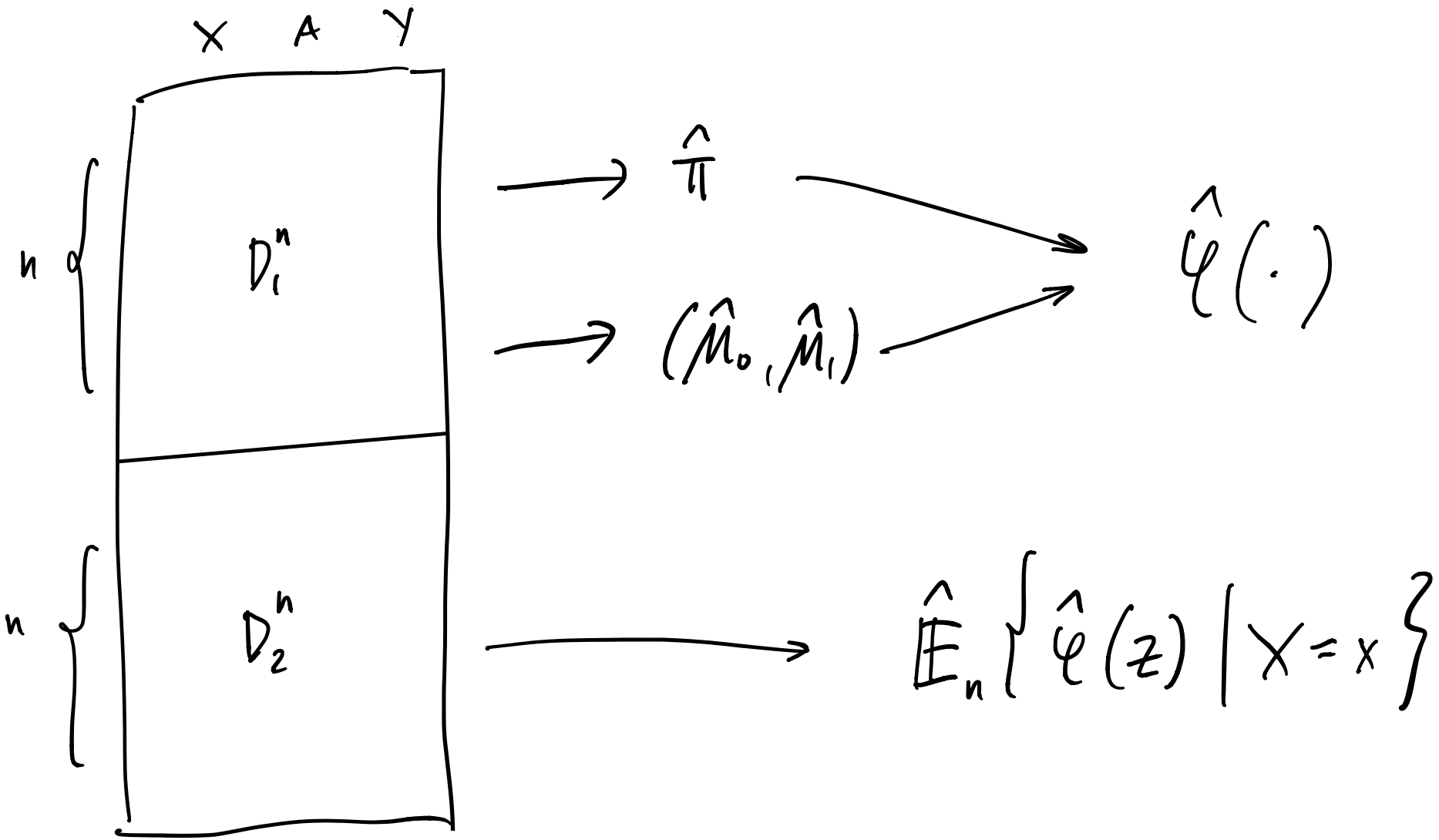
Construct estimates $(\widehat{\pi}, \widehat{\mu}_0, \widehat{\mu}_1)$ *of* (π, μ_0, μ_1) *using* D_1^n.

Step 2. *Pseudo-outcome regression: Construct the pseudo-outcome*

$$\widehat{\varphi}(Z) = \widehat{\mu}_1(X) - \widehat{\mu}_0(X) + \frac{A - \widehat{\pi}(X)}{\widehat{\pi}(X)\{1 - \widehat{\pi}(X)\}} \Big\{ Y - \widehat{\mu}_A(X) \Big\}$$

and regress it on covariates X *in the test sample* D_2^n*, yielding*

$$\widehat{\tau}_{dr}(x) = \widehat{\mathbb{E}}_n\{\widehat{\varphi}(Z) \mid X = x\}.$$


X A Y
n
D_1^n
n
D_2^n
$\hat{\pi}$
$(\hat{\mu}_0, \hat{\mu}_1)$
$\hat{\varphi}(\cdot)$
$\hat{\mathbb{E}}_n\{\hat{\varphi}(Z) \mid X = x\}$

Some DR-Learner history

First proposed by van der Laan (2005, 2013)

- but did not give specific error bounds

Kernel & series versions rederived in recent years

- *Lee et al (2017), Semenova & Chernozhukov (2017), Zimmert & Lechner (2019), Fan et al (2019)*
- but: tailored to particular 2nd-stage methods
- also: used restrictive assumptions on nuisance estimators and/or didn't allow simpler CATE

Foster & Syrgkanis (2019) studied ERM version

- but: error bounds were not doubly robust, also global & loose relative to oracle

DR-Learner

In particular, these previous analyses obtain rates like:

$$\text{RMSE}\left(\widehat{\tau}\right) \sim \text{RMSE}\left(\widetilde{\tau}^*\right) + \sqrt{k_n}\left\{\text{RMSE}\left(\widehat{\pi}\right) \times \text{RMSE}\left(\widehat{\mu}_a\right)\right\}$$

with $k_n \to \infty$, or

$$\text{RMSE}\left(\widehat{\tau}\right) \sim \text{RMSE}\left(\widetilde{\tau}^*\right) + \text{RMSE}\left(\widehat{\pi}\right)^2 + \text{RMSE}\left(\widehat{\mu}_a\right)^2$$

Q: What can we say about the DR-Learner if we are agnostic about what regression tools we use in the 1st & 2nd stage?

- Can these conditions for oracle optimality be improved?

DR-Learner error bound

Theorem (DR-Learner Master Theorem)

Assume $\widehat{\mathbb{E}}_n$ *is stable and* $\widehat{\varphi}$ *is consistent.*

Denote oracle estimator by $\widetilde{\tau}(x) = \widehat{\mathbb{E}}_n(Y^1 - Y^0 \mid X)$, *with risk* $R^*(x) = MSE\Big(\widetilde{\tau}(x)\Big) = \mathbb{E}[\{\widetilde{\tau}(x) - \tau(x)\}^2]$.

Then

$$\widehat{\tau}_{dr}(x) \;=\; \widetilde{\tau}(x) + O_{\mathbb{P}}\left(\widehat{\mathbb{E}}_n\{\widehat{b}(X) \mid X = x\}\right) + o_{\mathbb{P}}\left(\sqrt{R^*(x)}\right).$$

for "doubly robust" bias term

$$\widehat{b}(x) = \sum_{a=0}^{1} \frac{\Big\{\widehat{\pi}(x) - \pi(x)\Big\}\Big\{\widehat{\mu}_a(x) - \mu_a(x)\Big\}}{a\widehat{\pi}(x) + (1-a)(1-\widehat{\pi}(x))}$$

Stability

Definition

The estimator $\widehat{\mathbb{E}}_n$ is *stable* (with respect to distance d) if

$$\frac{\widehat{\mathbb{E}}_n\{\widehat{f}(Z) \mid X = x\} - \widehat{\mathbb{E}}_n\{f(Z) \mid X = x\} - \widehat{\mathbb{E}}_n\{\widehat{b}(X) \mid X = x\}}{RMSE^*} \xrightarrow{p} 0$$

whenever $d(\widehat{f}, f) \xrightarrow{p} 0$, for

- $\widehat{b}(x) = \mathbb{E}\{\widehat{f}(Z) - f(Z) \mid D^n, X = x\}$ the conditional bias of $\widehat{f}$,
- $RMSE^{*2} = \mathbb{E}\left(\left[\widehat{\mathbb{E}}_n\{f(Z) \mid X = x\} - \mathbb{E}\{f(Z) \mid X = x\}\right]^2\right)$.

Theorem 1 of Kennedy (2020): Linear smoothers are stable with respect to a weighted $L_2(\mathbb{P})$ norm.

Discussion: DR-Learner error bound

This is a nearly model-free, method-agnostic error bound

- shows DR-Learner error can't deviate from oracle by more than product of nuisance errors
- essentially a CATE analog of DR results for ATEs
- *allows faster rates for CATE even w/slower nuisance rates*
- gives smaller risk vs. previous method-specific results

This result is **very general**, allowing generic methods/assumptions

- now we specialize to classic Hölder s-smooth functions
- i.e., all derivatives up to $s-1$ bounded, & highest continuous

DR-Learner error bounds: Smoothness

Corollary (DR-Learner Under Smoothness)

Suppose assumptions of DR-Learner Theorem hold, and that:

1. π *is* α*-smooth, and estimated with MSE* $n^{-1/\left(2+\frac{d}{\alpha}\right)}$.
2. μ_a *are* β*-smooth, and estimated with MSE* $n^{-1/\left(2+\frac{d}{\beta}\right)}$.

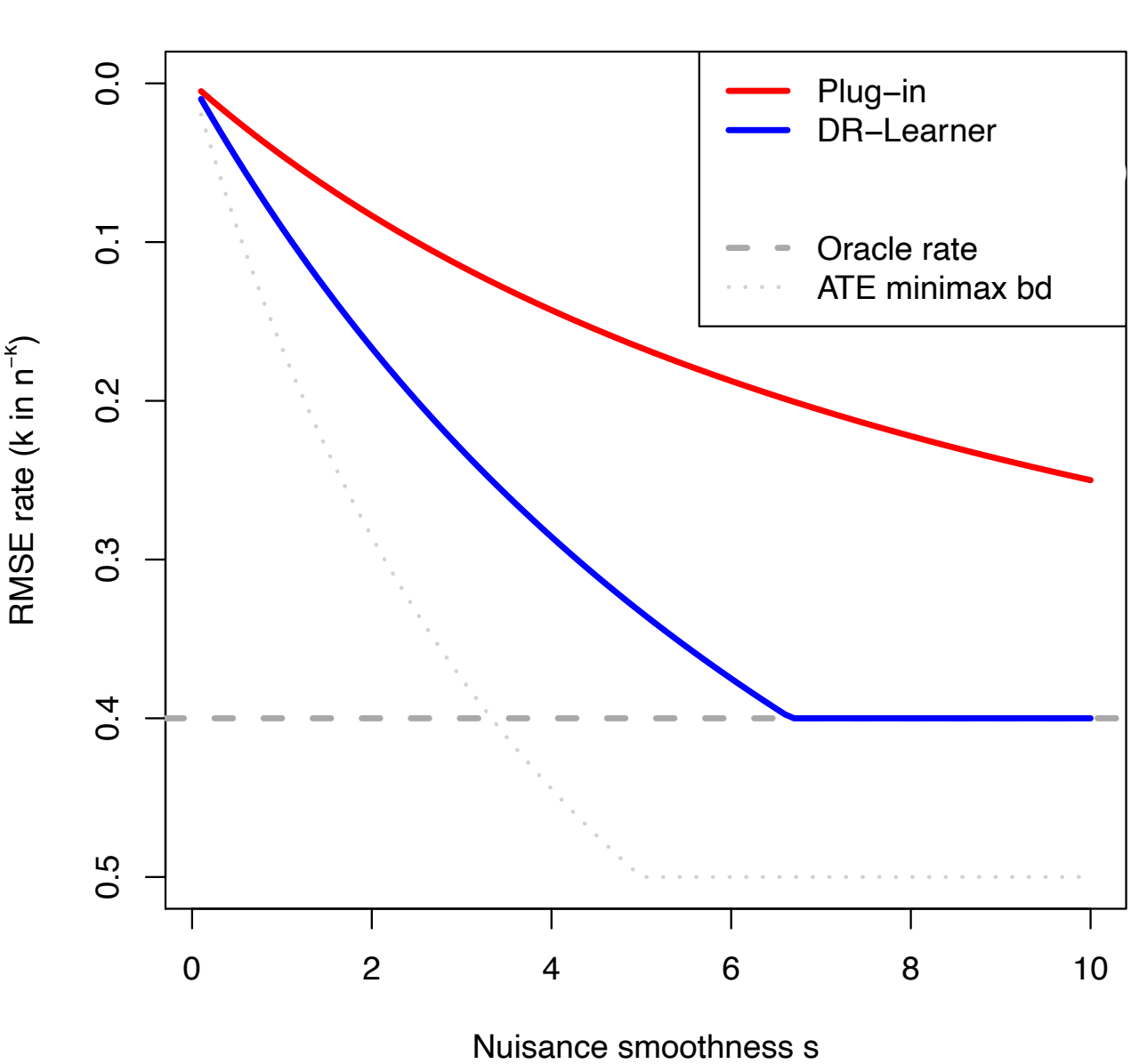
If CATE τ *is* γ*-smooth and* $\widehat{\mathbb{E}}_n$ *is minimax optimal, then*

$$\widehat{\tau}_{dr}(x) - \tau(x) \;=\; O_{\mathbb{P}}\left(n^{-1/\left(2+\frac{d}{\gamma}\right)} + n^{-1/\left(2+\frac{d}{\alpha}\right)} n^{-1/\left(2+\frac{d}{\beta}\right)} \right)$$

and thus the DR-Learner achieves oracle rate if

$$\sqrt{\alpha\beta} \geq \frac{d/2}{\sqrt{1+\frac{d}{\gamma}\left(1+\frac{d}{2\bar{s}}\right)}}$$

for $\bar{s}$ *the harmonic mean of* α *and* β.


Dimension d=20, CATE smoothness γ=2d
Plug−in
DR−Learner
Oracle rate
ATE minimax bd
RMSE rate (k in n^−k)
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
Nuisance smoothness s

Discussion

Previous result shows DR-Learner can adapt to CATE smoothness

- even when propensity score & regressions are less smooth
- gives sufficient conditions for oracle rate $n^{-\gamma/(2\gamma+d)}$

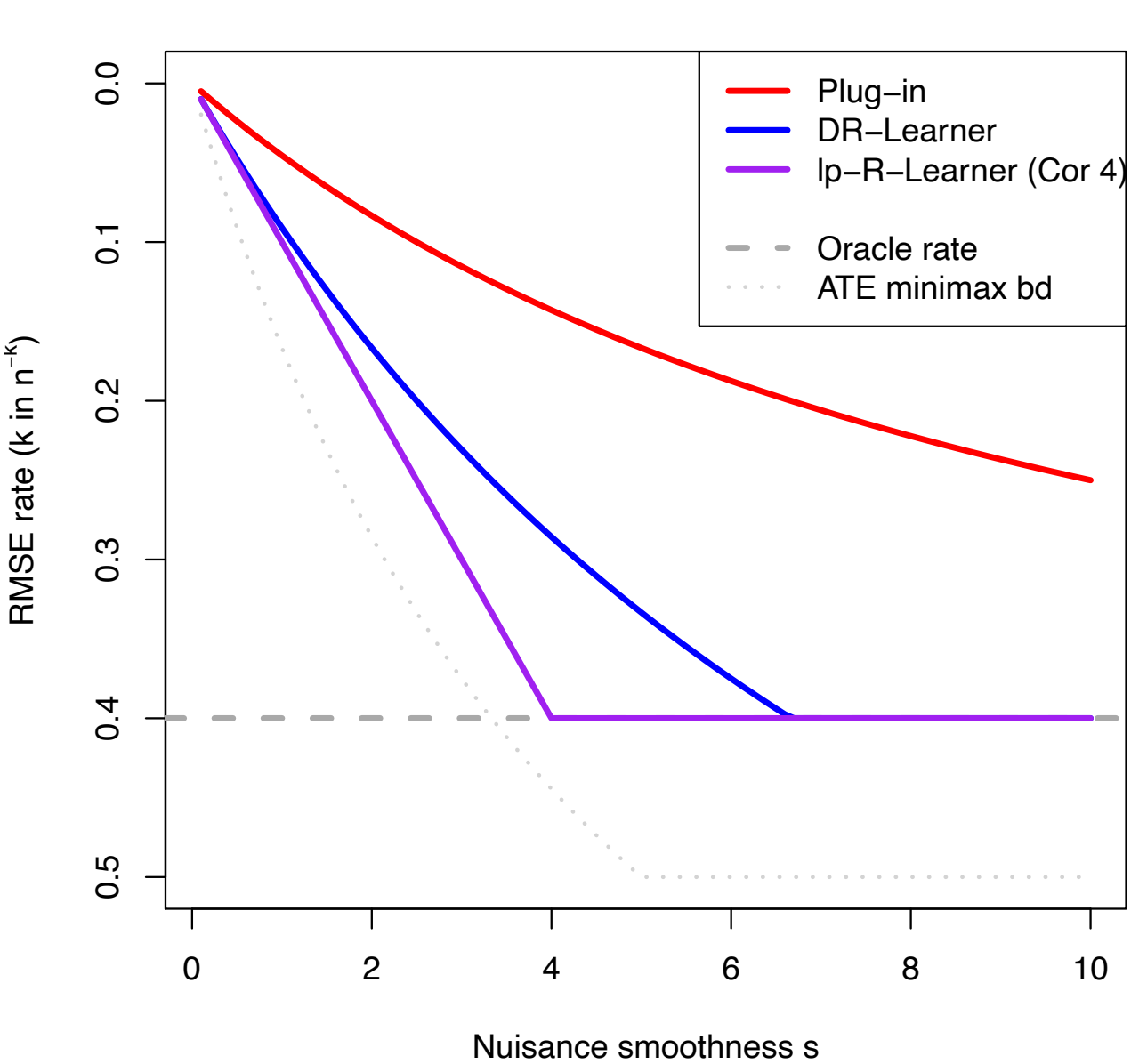
Analogous condition for DR estimator of ATE: $\sqrt{\alpha\beta} \geq d/2$

- as CATE gets more smooth, these conditions align

Term dividing $d/2$ is *"lowered bar"* for optimal estimation due to oracle rate being slower than root-n

- arose in dose-response but missed in recent CATE papers

See paper for result for generic regression w/estimated outcomes


Dimension d=20, CATE smoothness γ=2d
Plug−in
DR−Learner
lp−R−Learner (Cor 4)
Oracle rate
ATE minimax bd
RMSE rate (k in n^−k)
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
Nuisance smoothness s

Minimax optimality

Q: Are these MSE rates optimal? Can they be improved?

A natural way to characterize optimality is via the minimax rate

$$R_n = \inf_{\widehat{\tau}} \sup_{P \in \mathcal{P}} \mathbb{E}_P |\widehat{\tau}(x) - \tau_P(x)|$$

i.e., the best possible (worst-case) error, *across all estimators*

Minimax rates are well-understood in many problems:

- smooth nonparametric regression: $n^{-1/\left(2+\frac{d}{s}\right)}$
- smooth functional estimation: $\max\{n^{-1/\left(1+\frac{d}{4s}\right)}, 1/\sqrt{n}\}$
- sparse linear regression: $\sqrt{s \log d / n}$
- density estimation w/measurement error: $(\log n)^{-s}$

Minimax optimality

Minimax rates have crucial implications, practical & theoretical

- gives benchmark for best possible performance
- precisely illustrates fundamental limits / statistical difficulty

Main idea in deriving lower bounds:

- construct distributions so similar they're indistinguishable
- but for which parameter is maximally separated

$\Longrightarrow$ then *no estimator* can have error smaller than separation

For nonlinear functionals, *mixture distributions* are required

Minimax optimality

Three ingredients in deriving minimax lower bound:

1. pair of mixture distributions
2. distance between their n-fold products (ideally small)
3. separation of parameter (ideally large)

Lemma (Tsybakov)

Let $P_\lambda, Q_\lambda \in \mathcal{P}$ and let ϖ be a prior distribution over λ. If

$$H^2\left(\int P_\lambda^n \, d\varpi(\lambda), \int Q_\lambda^n \, d\varpi(\lambda)\right) \leq \alpha < 2$$

and $|\psi(P_\lambda) - \psi(Q_\lambda)| \geq s > 0$, then

$$R_n = \inf_{\widehat{\psi}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left|\widehat{\psi} - \psi(P)\right| \geq \frac{|s|}{4}\left\{1 - \sqrt{\alpha\left(1 - \frac{\alpha}{4}\right)}\right\}.$$

Construction

Main idea:

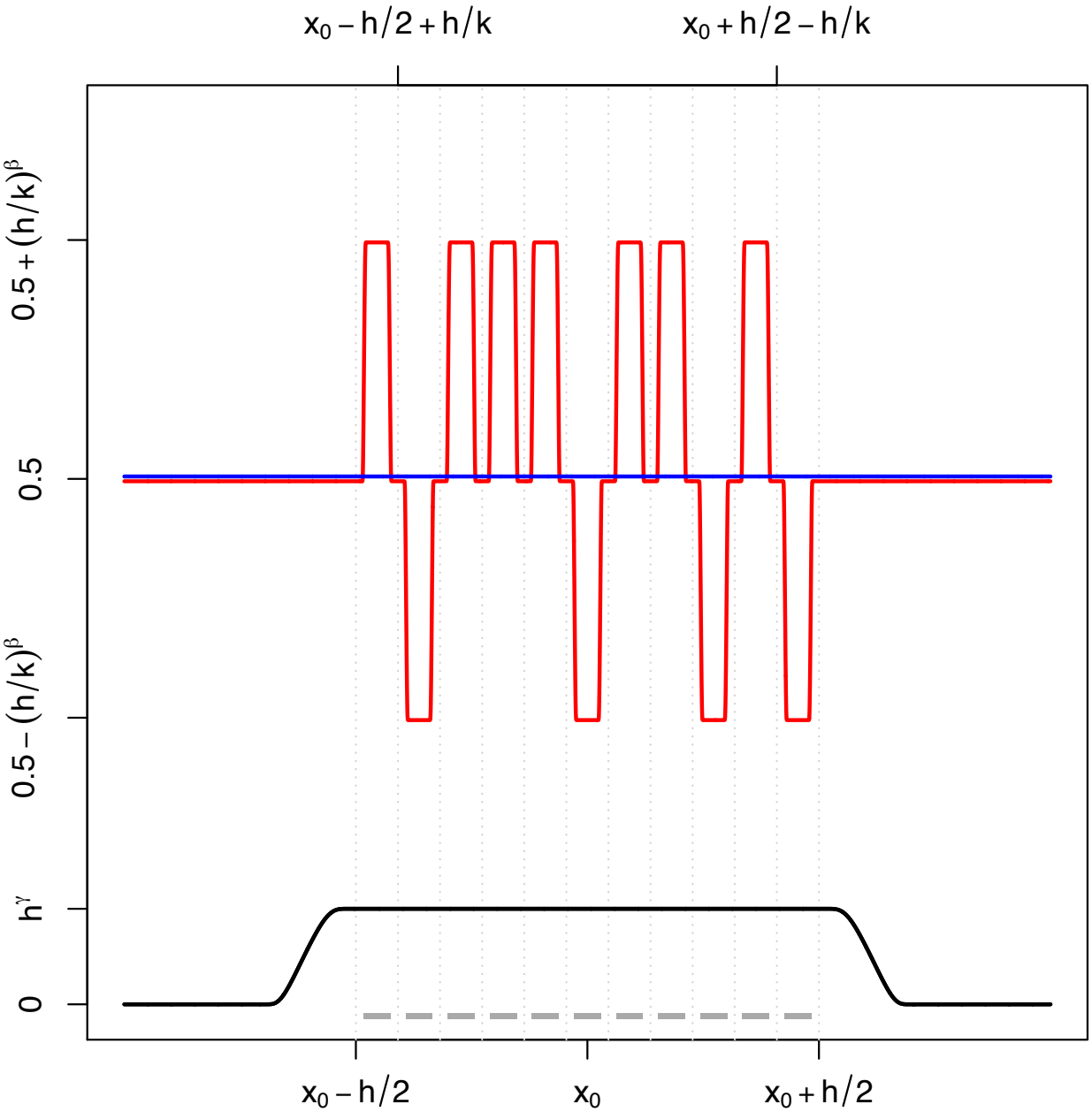
- perturb CATE with a flat-top bump at x_0
- perturb PS π and regression μ_0, but only locally near x_0
- only get observations at flat parts of bumps

Also: less smooth nuisance also perturbed under both P_λ and Q_λ

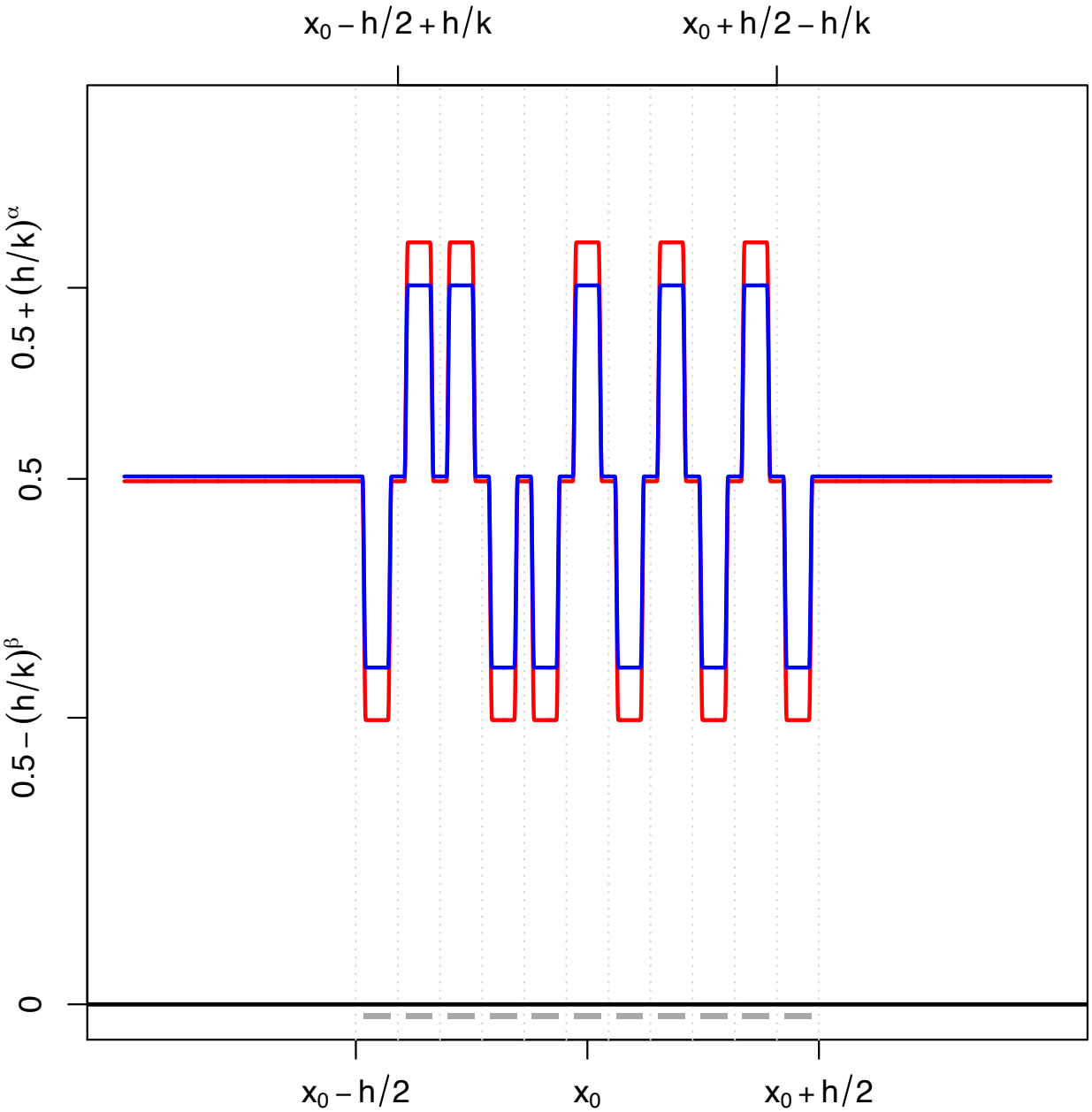
This is like a combination of lower bound constructions for

- nonparametric regression (cf. Tsybakov 2009)
- functional estimation (Birge & Massart '95, Robins et al. '09)

Null P_λ



Alt. Q_λ



Hellinger distance

In general, the distance between mixtures can be complicated

- we give a local adaptation of a nice lemma from Robins et al. (2009) to relate to simple posteriors over 1 observation

Proposition

Let $s \equiv \frac{\alpha+\beta}{2}$. *Under some conditions we have*

$$H^2\left(\int P_\lambda^n \, d\varpi(\lambda), \int Q_\lambda^n \, d\varpi(\lambda)\right) \lesssim \left(\frac{n^2 h^d}{k/h^d}\right)\left(k/h^d\right)^{-4s/d}.$$

Now we choose $h^\gamma \sim (h/k^{1/d})^{2s}$ and $k \sim n^{(d/2s - d/\gamma)/(1+d/2\gamma+d/4s)}$ to ensure the Hellinger distance is bounded (e.g., less than one)

Overall minimax rate

Since the separation in the CATE is h^γ, this implies the following minimax lower bound:

Theorem

Let $\mathcal{P}$ denote the model where

- *$f(x)$ satisfies some conditions (see paper),*
- *$\pi(x)$ is α-smooth,*
- *$\mu_0(x)$ is β-smooth,*
- *$\tau(x)$ is γ-smooth,*

Then for $s \equiv (\alpha+\beta)/2$ the minimax rate is lower bounded as

$$\inf_{\widehat{\psi}} \sup_{P\in\mathcal{P}} \mathbb{E}_P|\widehat{\tau}(x_0) - \tau_P(x_0)| \gtrsim \begin{cases} n^{-1/\left(1+\frac{d}{2\gamma}+\frac{d}{4s}\right)} & \text{if } s < \frac{d/4}{1+d/2\gamma} \\ n^{-1/\left(2+\frac{d}{\gamma}\right)} & \text{otherwise} \end{cases}$$

Discussion

This result verifies Kennedy (2020) conjecture that $s \geq \frac{d/4}{1+d/2\gamma}$ is necessary for achieving the oracle rate

Crucially, shows how CATE is hybrid regression/functional

- smooth nonparametric regression rates scale with $d/2\gamma$
- functional estimation rates with $d/4s$

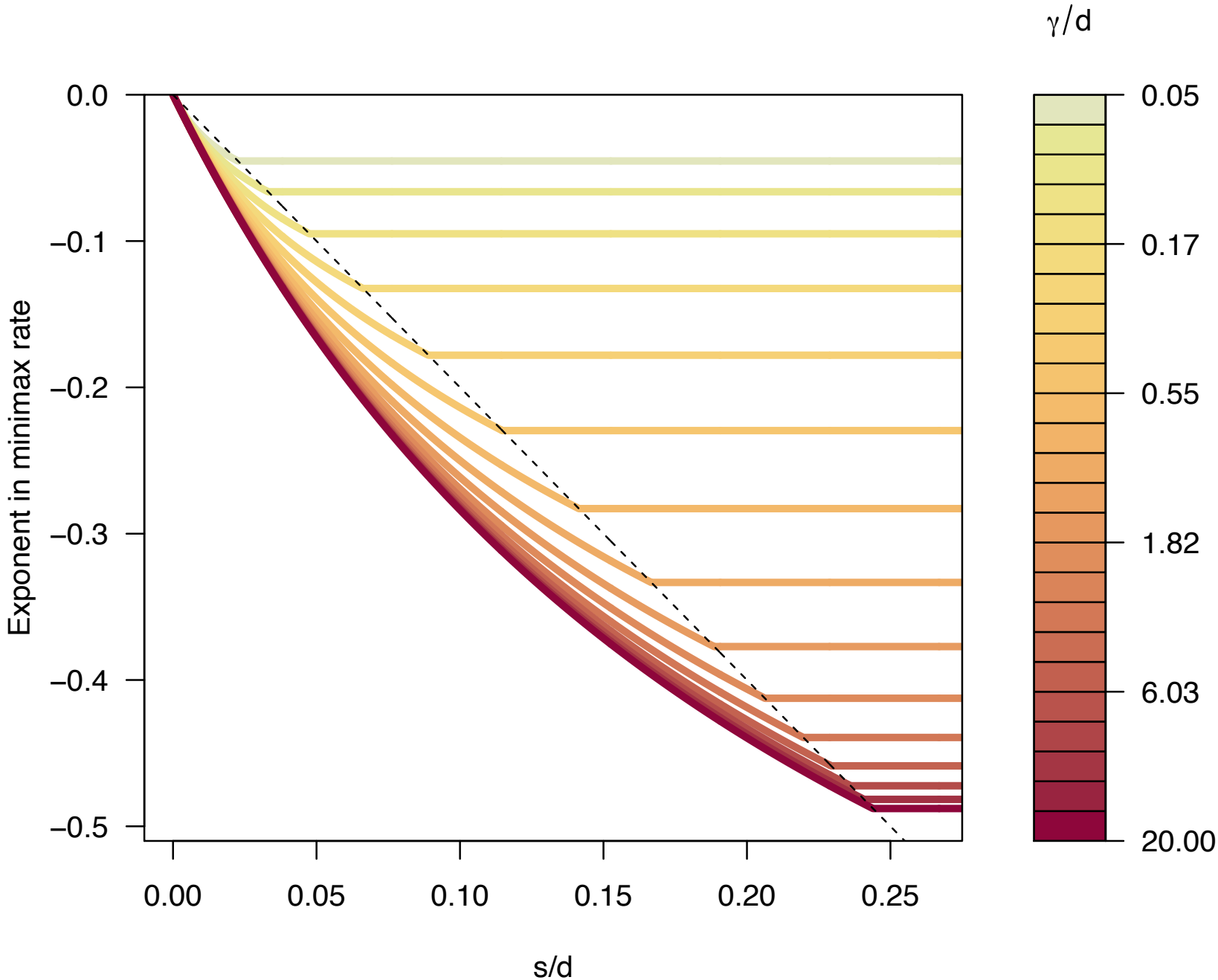
We showed the minimax rate for the CATE is

$$n^{-1/\left(1+\frac{d}{2\gamma}+\frac{d}{4s}\right)}$$

$\rightarrow$ which scales with the sum! $d/2\gamma + d/4s$

We expect similar phenomena for other hybrid creatures

- dose-response curve, counterfactual density, etc.


γ/d
0.05
0.17
0.55
1.82
6.03
20.00
Exponent in minimax rate
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
0.00
0.05
0.10
0.15
0.20
0.25
s/d

Attainability

In the paper we construct a new estimator that attains the bound

- *under some conditions on covariate density estimation*

Estimator targets a $\pi(1-\pi)$-weighted local polynomial projection of the CATE

$$\arg\min_{\beta} \mathbb{E}\left[\pi(x)\{1-\pi(x)\}K_h(x)\left\{\tau(x)-\beta^{\mathrm{T}}\rho_h(x)\right\}^2\right]$$

using a localized form of higher-order influence function methods (Robins et al., 2008, 2017, etc.)

- estimator is localized U-statistic with localized basis kernel
- tuning parameters: h controls localization, k basis terms
- estimator pretends CATE is a polynomial locally near x_0

Intuition

Consider estimation of CATE in semiparametric model with $\tau(x) = \tau$ constant. Classic efficient estimator $\widehat{\tau}$ solves:

$$0 = \mathbb{P}_n \left[\left\{ A - \widehat{\pi}(X) \right\} \left\{ Y - \widehat{\mu}_0(X) - \widehat{\tau} A \right\} \right] \equiv \mathbb{P}_n \{ \widehat{\varphi}(Z; \widehat{\tau}) \}$$

$\rightarrow$ **Robinson's double residual regression**

Improved U-statistic-based approach: $\widehat{\tau}$ solves

$$0 = \mathbb{P}_n \{ \widehat{\varphi}(Z; \widehat{\tau}) \} - \mathbb{U}_n \left\{ (A_1 - \widehat{\pi}_1) b(X_1)^{\mathrm{T}} \Omega^{-1} b(X_2) (Y_2 - \widehat{\mu}_{02} - \widehat{\tau} A_2) \right\}$$

$\rightarrow$ **Robins et al. higher-order influence functions**

Our method is essentially a local polynomial version of this

Estimator #2: Undersmoothed R-Learner

The original "R-Learner" was proposed by Robinson (1988) for efficiently estimating a constant/parametric CATE

Simplest form: for regression function $\eta(x) = \mathbb{E}(Y \mid X = x)$, do

linear regression of (outcome residuals) on (treatment residuals)

$$\text{lm}\Big(\{Y - \widehat{\eta}(X)\} \sim \{A - \widehat{\pi}(X)\}\Big)$$

Intuition: The slope estimates a weighted treatment effect, with weights largest at Xs where we see both treated & controls

Idea: instead of linear regression, do nonparametric regression of residuals on residuals (interacted with covariates)

- Robins ('08), Nie & Wager ('17), Chernozhukov et al. ('17)
- Kennedy ('20): undersmoothed local polynomial

More motivation & intuition

The R-learner from Kennedy (2020) can be viewed as estimating the locally weighted projection parameter

$$\tau_h(x_0) = \rho_h(x_0)^{\mathrm{T}}\theta,$$

for coefficients $\theta = Q^{-1}R$ with

$$Q = \int \rho(x)K_h(x)\pi(x)\{1-\pi(x)\}\rho(x)^{\mathrm{T}}\ d\mathbb{P}(x)$$

$$R = \int \rho(x)K_h(x)\pi(x)\{1-\pi(x)\}\tau(x)\ d\mathbb{P}(x),$$

i.e., the $K_h(x)\pi(x)(1-\pi(x))$-weighted *least squares projection* of the CATE $\tau(x)$ on the Legendre series $\rho(x)$.

Estimator #3: Higher-order R-learner

R-learner can be viewed as doubly robust-style estimator of θ

- suggests improving via higher-order influence functions (Robins et al., 2008, 2017, etc.)

Thus our proposed estimator is $\widehat{\tau}(x_0) = \rho(x_0)^{\mathrm{T}} \widehat{Q}^{-1} \widehat{R}$ for

$$\widehat{Q} = \mathbb{U}_n \left[\rho(X_1) K_h(X_1) \left\{ \widehat{\varphi}_{a1}(Z_1) + \widehat{\varphi}_{a2}(Z_1, Z_2) K_h(X_2) \right\} \rho(X_1)^{\mathrm{T}} \right]$$
$$\widehat{R} = \mathbb{U}_n \left[\rho(X_1) K_h(X_1) \left\{ \widehat{\varphi}_{y1}(Z_1) + \widehat{\varphi}_{y2}(Z_1, Z_2) K_h(X_2) \right\} \right]$$

- uses U-statistic terms to further correct bias
- kind of like doing extra undersmoothed regressions, but without undersmoothing estimators $\widehat{\pi}, \widehat{\mu}_0$ directly

Some distinctions

However our estimator is not just a "standard" higher order estimator of the projection parameter

- it has some extra localization, which wouldn't arise if you only cared about projection parameter

Extra localization:

- U-statistic term localized wrt both observations

$$\widehat{R} = \mathbb{U}_n \left[\rho(X_1) K_h(X_1) \left\{ \widehat{\varphi}_{y1}(Z_1) + \widehat{\varphi}_{y2}(Z_1, Z_2) K_h(X_2) \right\} \right]$$

- basis terms $b_h(x)$ in $\widehat{\varphi}(Z_1, Z_2)$ are localized:
  $\rightarrow$ like taking functions only near x_0, stretching them out, and approximating stretched function, to get *smaller bias*

Bias & variance

Proposition

Under regularity conditions,

$$\mathbb{E}\,|\widehat{\tau}(x_0) - \tau_h(x_0)| \;\lesssim\; \left(h/k^{1/d}\right)^{2s} + \|\widehat{\pi} - \pi\|_{F^*}\|\widehat{\mu}_0 - \mu_0\|_{F^*}\|\widehat{\Omega}^{-1} - \Omega^{-1}\| + \sqrt{\frac{1}{nh^d}}\left\{1 + \sqrt{\frac{k}{nh^d}}\left(1 + \|\widehat{\Omega}^{-1} - \Omega^{-1}\|\right)\right\}$$

Further, if the covariate density is estimated accurately enough,

$$\mathbb{E}\,|\widehat{\tau}(x_0) - \tau(x_0)| \;\lesssim\; h^{\gamma} + \left(h/k^{1/d}\right)^{2s} + \frac{\sqrt{k}}{nh^d}$$

Overall rate

$h^\gamma + \left(h/k^{1/d}\right)^{2s}$ is the bias

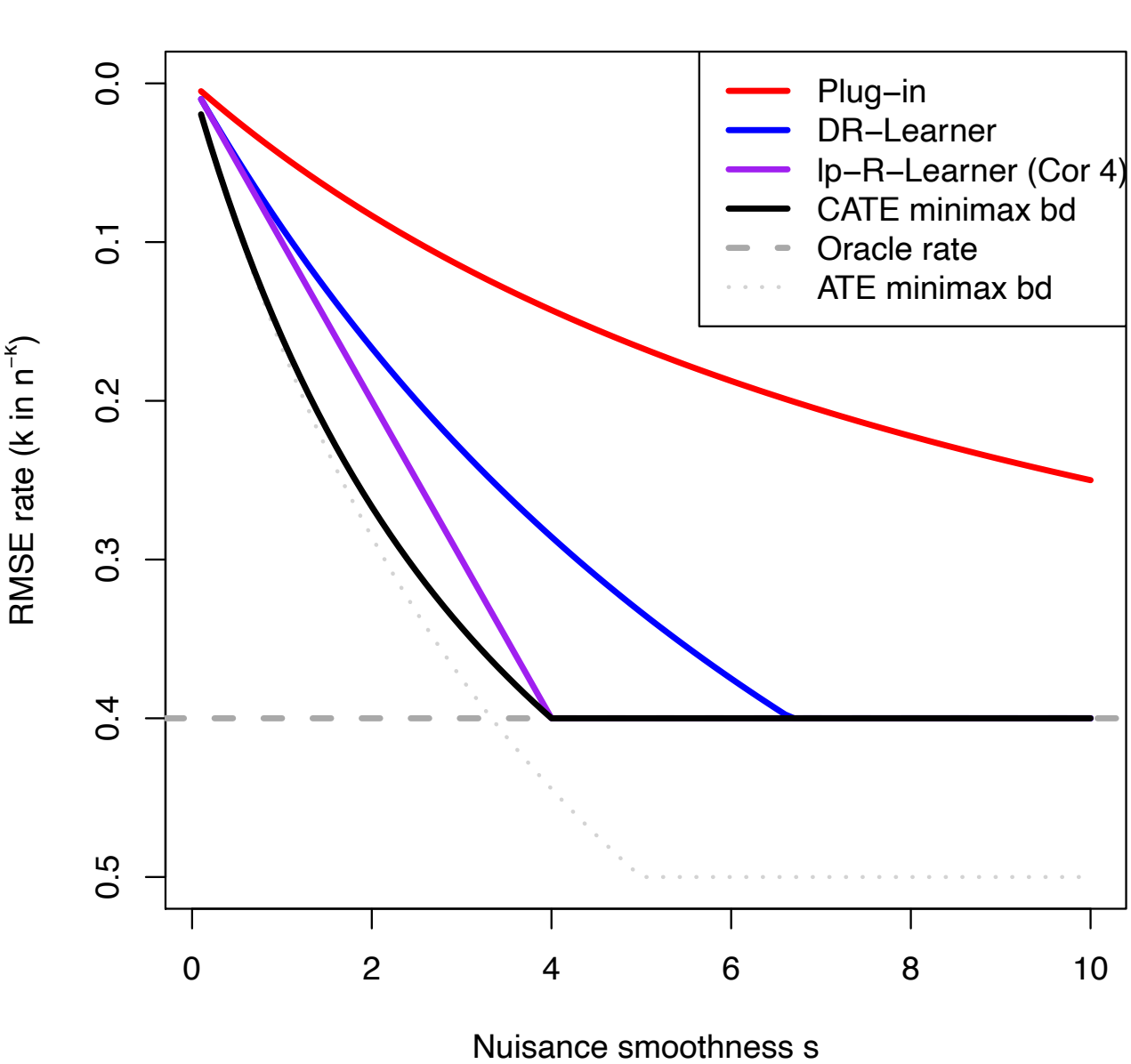
- intuition: h^γ is bias even if we observed potential outcomes
- second part is usual squared nuisance bias $k^{-2s/d}$, shrunk by h^{2s} since only care about small window

$k/(nh^d)^2$ is the variance

- intuition: U-statistic in nh^d observations, with kernel depending on k-dimensional basis

Balancing gives the minimax rate!

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P |\widehat{\tau}(x_0) - \tau_P(x_0)| \lesssim \begin{cases} n^{-1/\left(1+\frac{d}{2\gamma}+\frac{d}{4s}\right)} & \text{if } s < \frac{d/4}{1+d/2\gamma} \\ n^{-1/\left(2+\frac{d}{\gamma}\right)} & \text{otherwise} \end{cases}$$


Dimension d=20, CATE smoothness γ=2d
Plug−in
DR−Learner
lp−R−Learner (Cor 4)
CATE minimax bd
Oracle rate
ATE minimax bd
RMSE rate (k in n^−k)
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
Nuisance smoothness s

Alternative model

Assuming $\eta(X) = \mathbb{E}(Y \mid X)$ is β-smooth instead of $\mu_0(X)$:

Theorem

Let $\mathcal{P}$ denote the model where

- *$f(x)$ satisfies some conditions (see paper),*
- *$\pi(x)$ is α-smooth,*
- *$\eta(x) = \mathbb{E}(Y \mid X = x)$ is β-smooth,*
- *$\tau(x)$ is γ-smooth,*

Then for $s \equiv (\alpha + \beta)/2$ the minimax rate is lower bounded as

$$\inf_{\widehat{\psi}} \sup_{P \in \mathcal{P}} \mathbb{E}_P |\widehat{\tau}(x_0) - \tau_P(x_0)| \gtrsim \begin{cases} n^{-1/\left(1 + \frac{d}{2\gamma} + \frac{d}{4(s \wedge \alpha)}\right)} & \text{if } s \wedge \alpha < \frac{d/4}{1 + d/2\gamma} \\ n^{-1/\left(2 + \frac{d}{\gamma}\right)} & \text{otherwise} \end{cases}$$

Summary

Lots of CATE estimators developed recently

- but many not very well-understood
- previously unclear how to benchmark - *what is optimal?*

Our contributions:

1. more flexible estimators, with stronger guarantees
2. resolution of minimax optimality story

Lots of unanswered questions & future work

- role of covariate density, other function classes
- lots to do wrt theory, methods, & application!

Kennedy (2020): arxiv.org/abs/2004.14497

Newer one w/ Siva & Jamie & Larry: arxiv.org/abs/2203.00837

Feel free to email with any questions:
edward@stat.cmu.edu

Thank you!

Smooth functions

In the following we illustrate result with *s*-smooth functions

- these are functions in the Hölder class $\mathcal{H}(s)$
- intuitively: smooth fns close to $\lfloor s \rfloor$-order Taylor approxs

Formally $\mathcal{H}(s)$ contains $\lfloor s \rfloor$-times continuously differentiable functions w/bdd partial derivatives, and for which

$$|D^m f(x) - D^m f(x')| \lesssim \|x - x'\|^{s - \lfloor s \rfloor}$$

for all x, x' and $m = (m_1, ..., m_d)$ such that $\sum_j m_j = \lfloor s \rfloor$, where $D^m = \frac{\partial^{\lfloor s \rfloor}}{\partial_{x_1}^{m_1} ... \partial_{x_d}^{m_d}}$ is the multivariate partial derivative operator

Similar results can be obtained in any model w/known MSE rates

Oracle inequality for regression w/estimated outcomes

Underlying the DR-Learner error bound is a general oracle inequality for regression with estimated/imputed outcomes

This setup arises in a wide variety of problems

- V-specific CATE for $V \subset X$
- regression with censored/missing outcomes (Fan & Gijbels 1994, Rubin & van der Laan 2006, Wang et al. 2010)
- dose-response curve estimation (Kennedy et al. 2017), heterogeneous effects of cts treatment
- conditional IV effects
- partially linear IV (Ai & Chen 2003, Newey & Powell 2003)

Improving the DR-Learner?

Now we have general conditions for DR-Learner to be optimal

- optimal when product of nuisance MSEs is smaller order

What if this condition fails? Any hope at oracle rates?

We pursue bias reduction using undersmoothing

- classic trick in parameter estimation
- idea: estimate nuisances with too little bias, too large variance

Since DR-learner error bound involves product of MSEs, undersmoothing won't (immediately) help

- also some additional complications involving estimating $1/\pi$
- $\implies$ considered local polynomial adaptation of R-Learner

Estimator #2: Undersmoothed R-Learner

The original "R-Learner" was proposed by Robinson (1988) for efficiently estimating a constant/parametric CATE

Simplest form: for regression function $\mu(x) = \mathbb{E}(Y \mid X = x)$, do linear regression of (outcome residuals) on (treatment residuals)

$$\text{lm}\Big(\{Y - \widehat{\mu}(X)\} \sim \{A - \widehat{\pi}(X)\}\Big)$$

Intuition: The slope estimates a weighted treatment effect, with weights largest at Xs where we see both treated & controls

Idea: instead of linear regression, do nonparametric regression of residuals on residuals (interacted with covariates)

- Robins ('08), Nie & Wager ('17), Chernozhukov et al. ('17)

lp-R-Learner error bounds

Theorem (lp-R-Learner Error Bound)

Assume:

1. *Estimator $\widehat{\mu}$ & observations Z are bdd, & X has bdd density.*
2. *PS estimates satisfy $\epsilon \leq \widehat{\pi}_j \leq 1 - \epsilon$ for some $\epsilon > 0$.*
3. *Eigenvalue condition on design matrices $\widehat{Q}$, $\widetilde{Q}$ (see paper).*
4. *$(\widehat{\pi}_j, \widehat{\mu})$ satisfy Condition NE, w/bias bds holding $\forall x' \in B_h(x)$*

Then, undersmoothing $(\widehat{\pi}, \widehat{\mu})$ and if CATE $\tau(x)$ is γ-smooth,

$$\widehat{\tau}(x) - \tau(x) = O_{\mathbb{P}}\left(n^{-\gamma/(2\gamma+d)} + n^{-2s/d}\right)$$

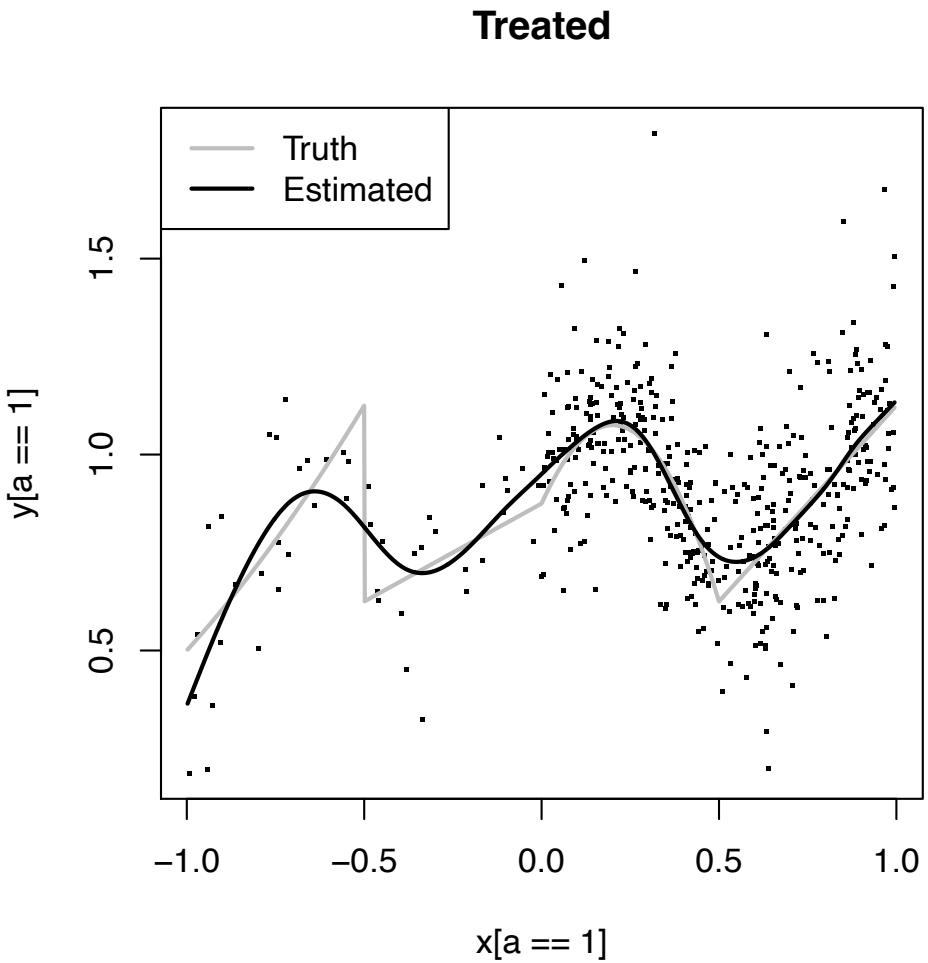
where $s = \frac{\alpha+\beta}{2}$ is avg smoothness of propensity score & regression.

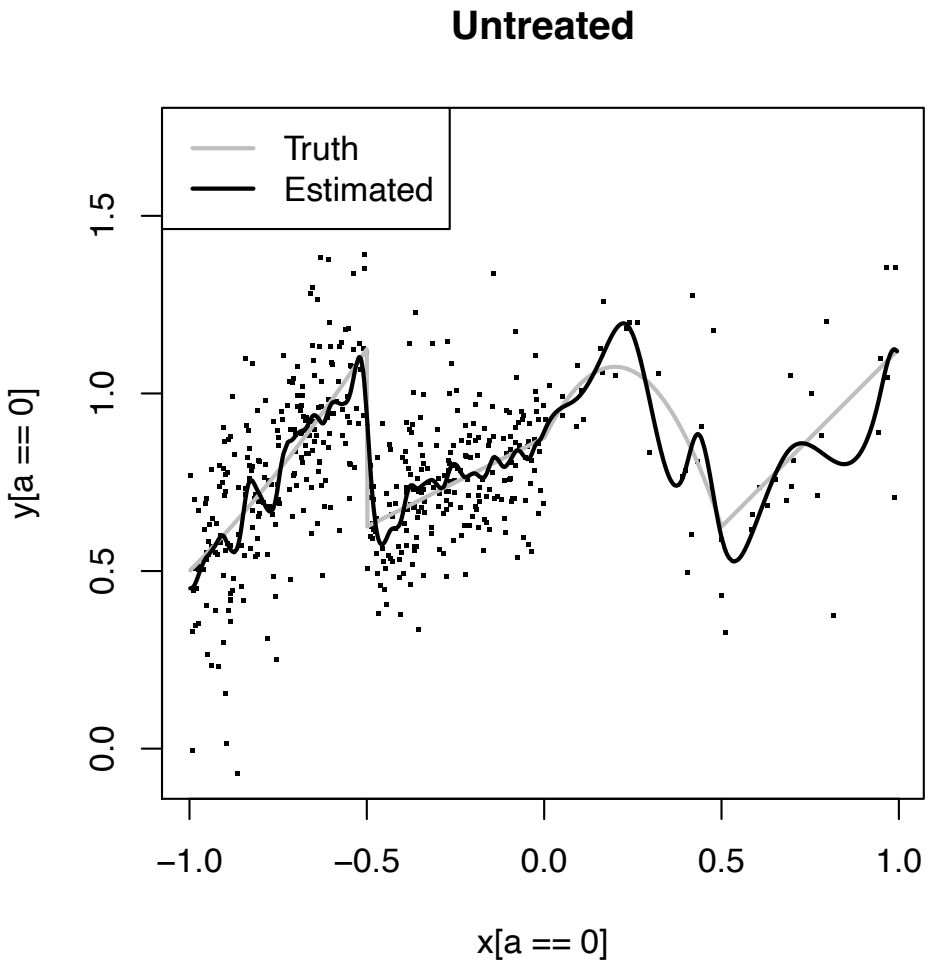
Simulations: Polynomial model

Now we study polynomial model from earlier example

$$X \sim \text{Unif}[-1, 1] \qquad \pi(x) = 0.5 + 0.4 \times \text{sign}(x)$$

$\mu_1 = \mu_0$ equal to piecewise polynomial from Gyorfi et al. (2002)


Treated
Truth
Estimated
y[a == 1]
x[a == 1]
0.5
1.0
1.5
−1.0
−0.5
0.0
0.5
1.0


Untreated
Truth
Estimated
y[a == 0]
x[a == 0]
0.0
0.5
1.0
1.5
−1.0
−0.5
0.0
0.5
1.0

Simulations: Polynomial model

Now we study polynomial model from earlier example

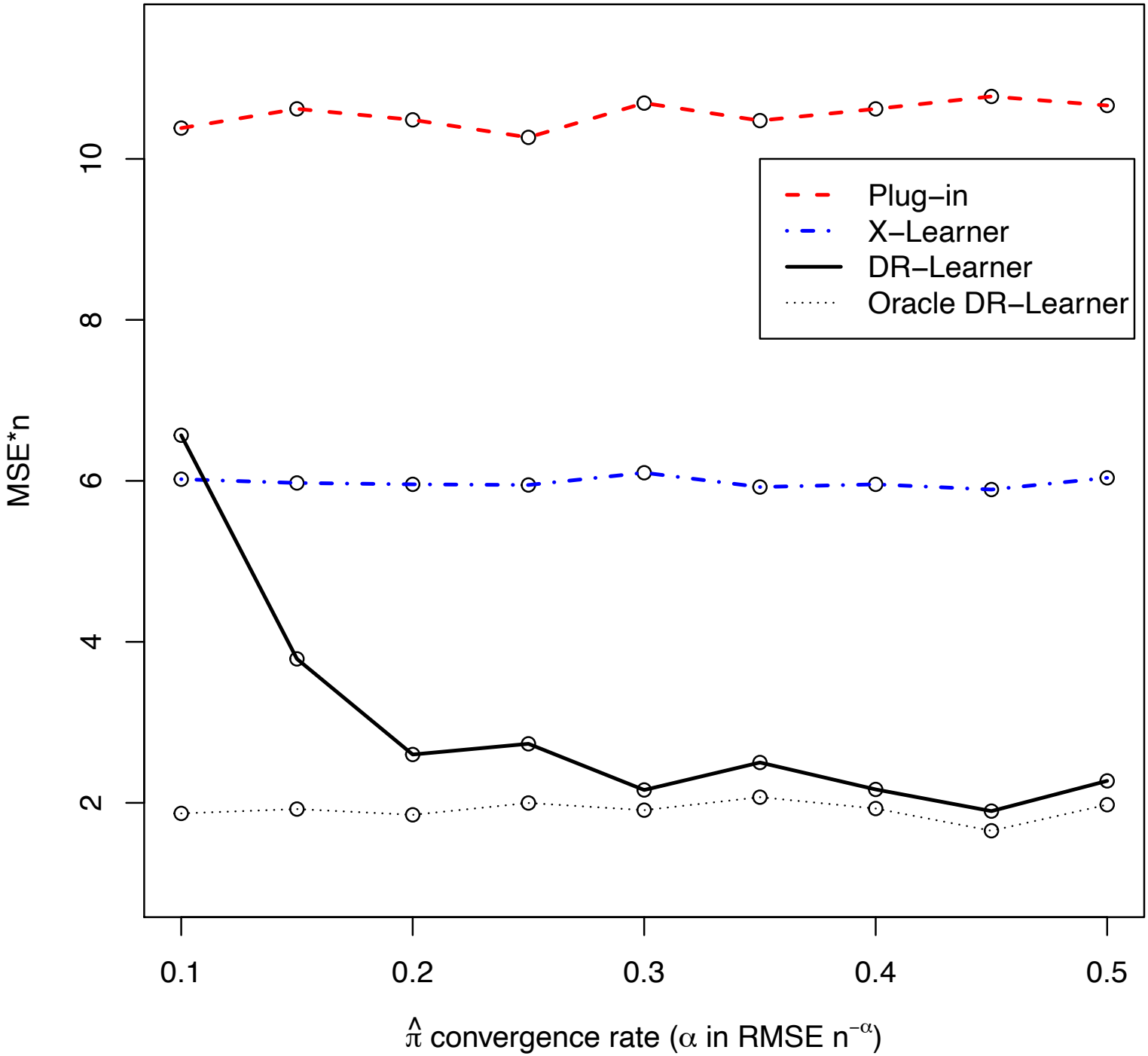
$$X \sim \text{Unif}[-1, 1] \qquad \pi(x) = 0.5 + 0.4 \times \text{sign}(x)$$

$\mu_1 = \mu_0$ equal to piecewise polynomial from Gyorfi et al. (2002)

Outcome & 2nd-stage regressions fit via *smoothing.spline* in R

PS fit as $\widehat{\pi} = \text{expit}\{\text{logit}(\pi) + \epsilon_n\}$ where $\epsilon_n \sim N(n^{-\alpha}, n^{-2\alpha})$

- allows for precise control of RMSE$(\widehat{\pi}) \sim n^{-\alpha}$


MSE*n
2
4
6
8
10
0.1
0.2
0.3
0.4
0.5
$\hat{\pi}$ convergence rate (α in RMSE $n^{-\alpha}$)
Plug−in
X−Learner
DR−Learner
Oracle DR−Learner

Illustration

Green et al. (2003): effects of canvassing on voter turnout

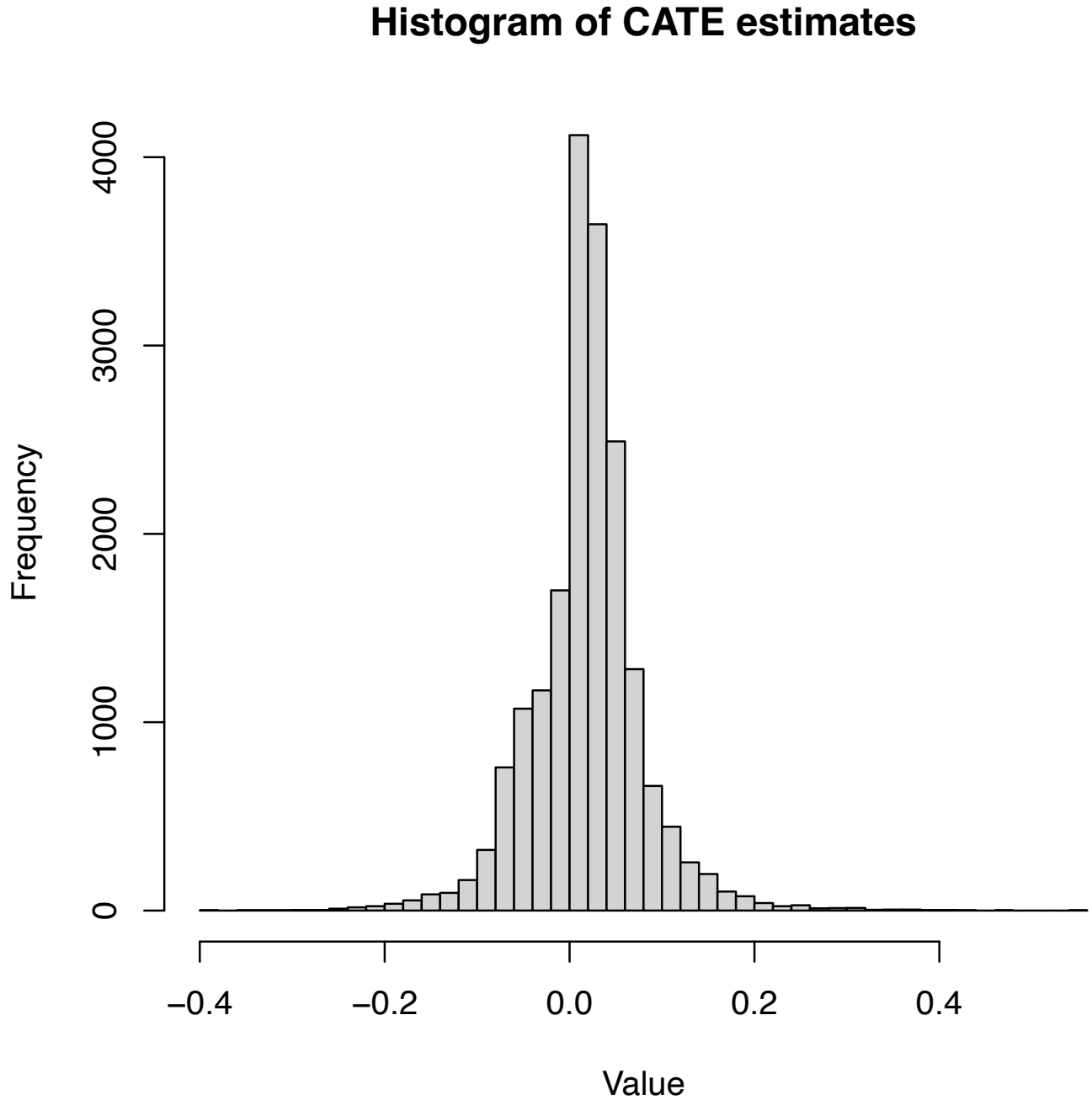
- $n \approx$19k registered voters across 6 large cities
- encouraged half to vote in local elections w/F2F contact

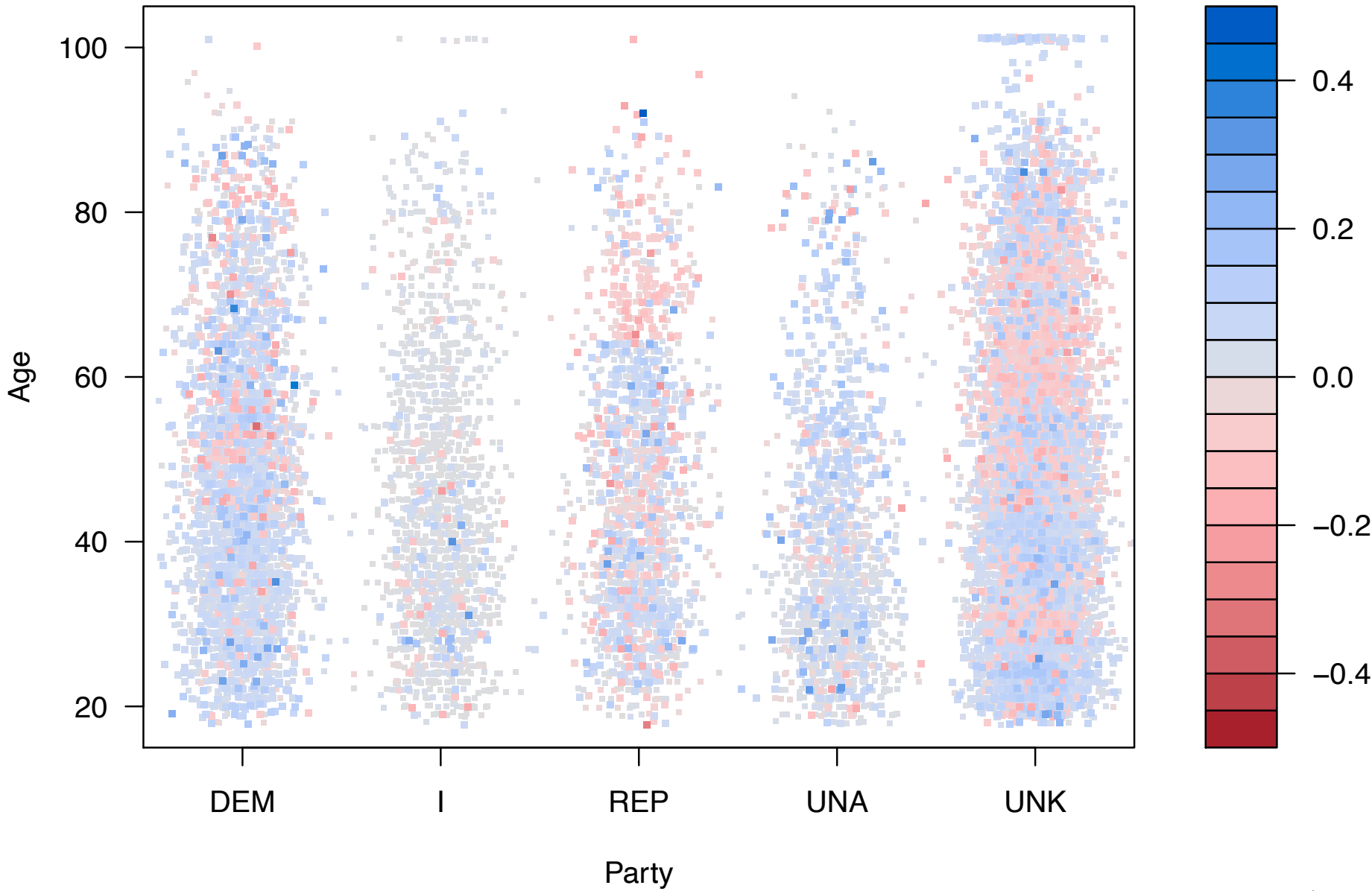
Data:

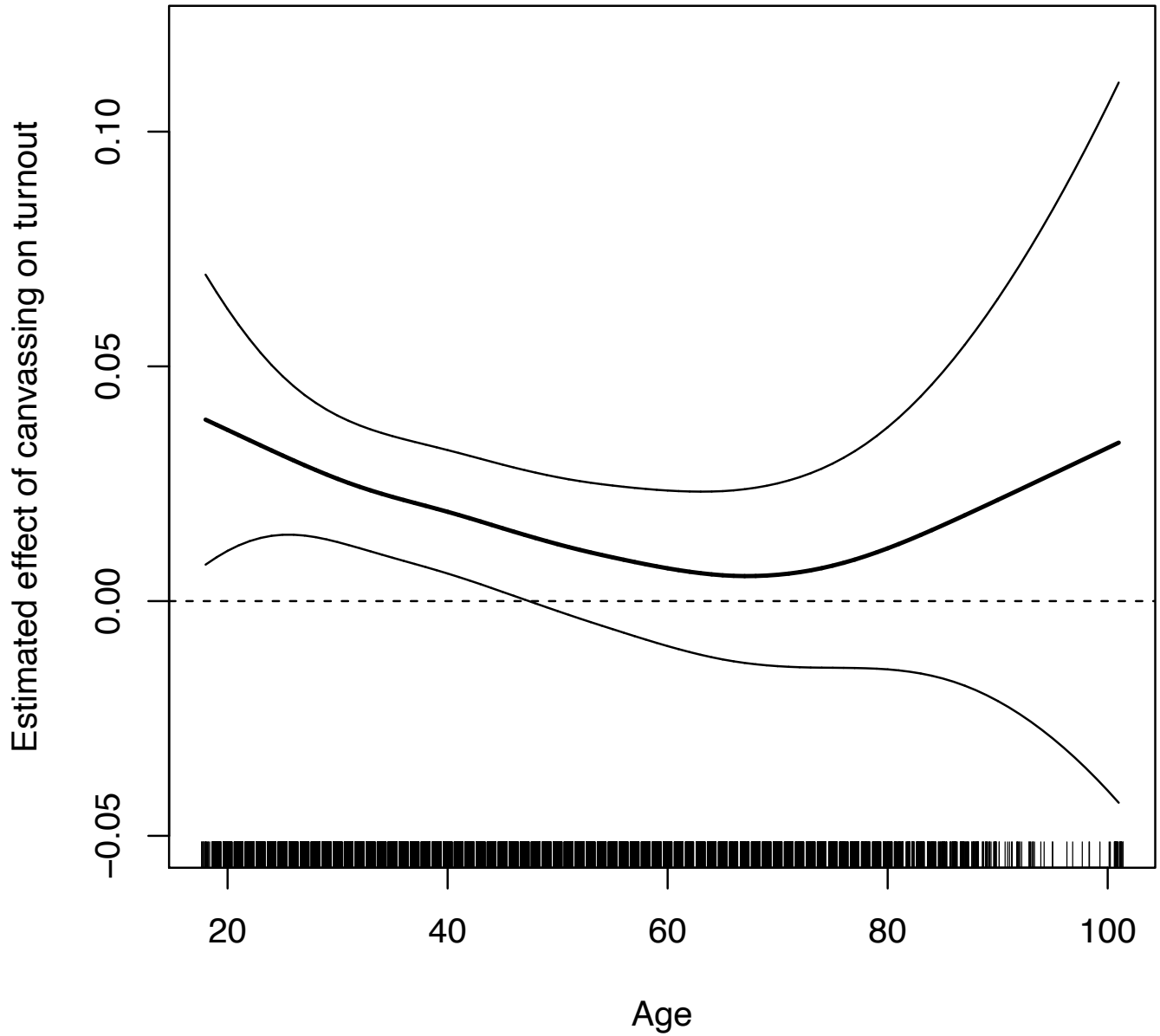
- $X =$ city, party affiliation, voting history, age, family size, race
- $A =$ whether (randomly) assigned to in-person encouragement
- $Y =$ whether subject voted in local election or not

Used proposed DR-Learner, with $K = 2$ folds and random forests

- Also estimated $\mathbb{E}(Y^1 - Y^0 \mid X_1)$ for $X_1 =$ age w/GAM


Histogram of CATE estimates
Frequency
0
1000
2000
3000
4000
−0.4
−0.2
0.0
0.2
0.4
Value


100
80
60
40
20
Age
DEM
I
REP
UNA
UNK
Party
0.4
0.2
0.0
−0.2
−0.4


Estimated effect of canvassing on turnout
0.10
0.05
0.00
−0.05
20
40
60
80
100
Age

Condition (NE, Nuisance Estimators)

Nuisance estimators $(\widehat{\pi}_a, \widehat{\pi}_b, \widehat{\mu})$ *are (a) linear smoothers*

$$\widehat{\pi}_j(x) = \sum_{i \in D_{1j}^n} w_{i\alpha}(x; X_{1j}^n) A_i \quad \textit{and} \quad \widehat{\mu}(x) = \sum_{i \in D_{1b}^n} w_{i\beta}(x; X_{1b}^n) Y_i$$

with weights $w_{i\cdot}(x; X_{1\cdot}^n)$ *depending on parameter* k, *(b) satisfying*

$$\left(\sum_{i=1}^n w_{i\alpha}(x; X_{1j}^n)^2 \right) \vee \left(\sum_{i=1}^n w_{i\beta}(x; X_{1j}^n)^2 \right) \lesssim \frac{k}{n}$$

and (c) yielding pointwise conditional bias bounds

$$\left| \mathbb{E}\{\widehat{\pi}_j(x) \mid X_{1j}^n\} - \pi(x) \right| \lesssim k^{-\frac{\alpha}{d}} \ \& \ \left| \mathbb{E}\{\widehat{\mu}(x) \mid X_{1b}^n\} - \mu(x) \right| \lesssim k^{-\frac{\beta}{d}}$$

Conditions NE(a–c) are standard in nonparametrics

For NE(a), many popular estimators are linear smoothers

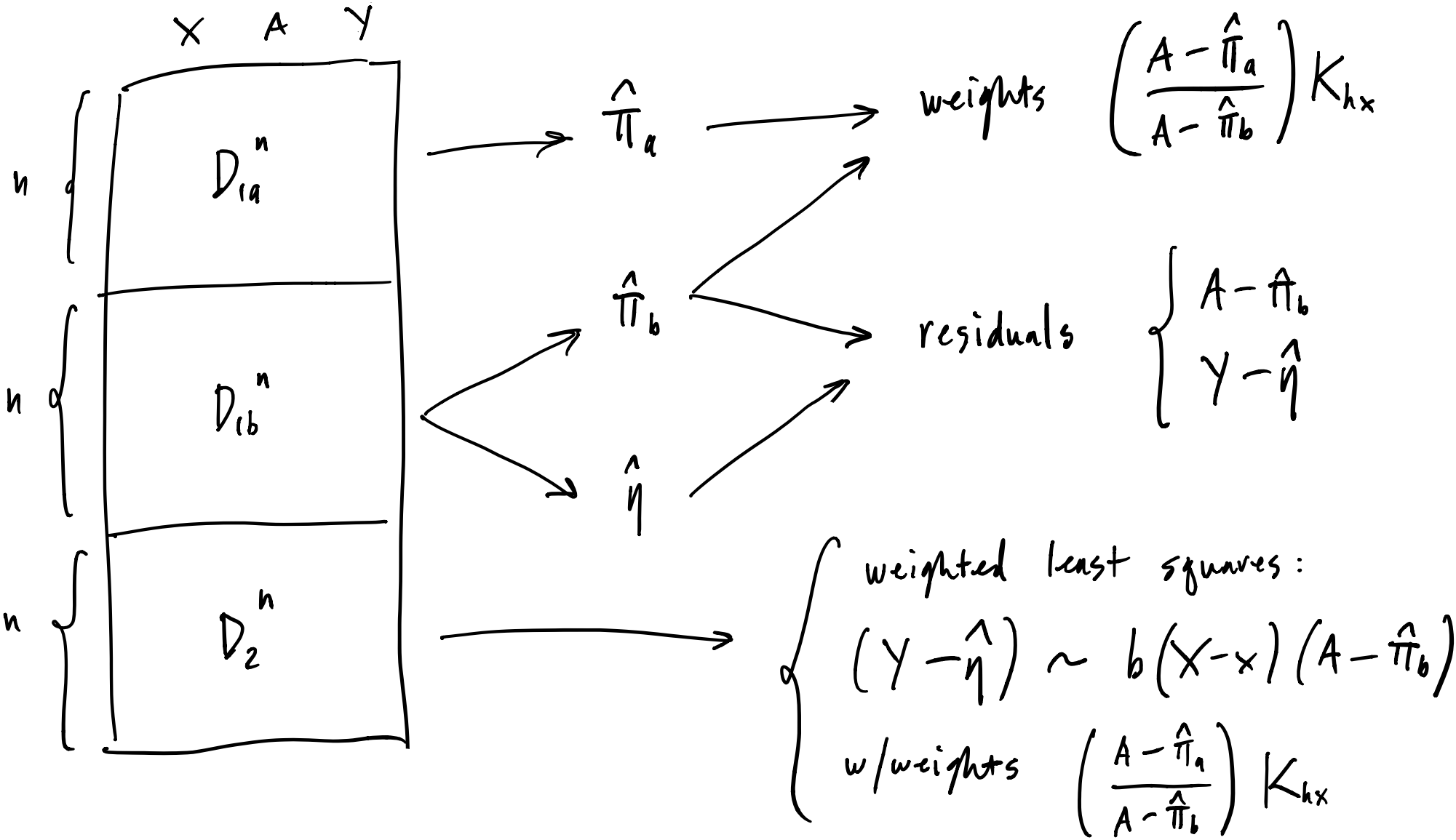
- but greedy RFs & locally adaptive methods generally excluded

NE(b) holds for kernel/local polynomial estimators with $h \sim k^{-d}$, and for many series estimators (with $k = \#$ basis terms)

- Fourier, splines, CDV wavelets, local polynomial partitioning

NE(c) holds for series and local polynomial estimators, for example, when underlying regression is appropriately smooth

- e.g., if PS π is α-smooth and regression μ is β-smooth


X A Y
n
D_{1a}^n
n
D_{1b}^n
n
D_2^n
$\hat{\pi}_a$
$\hat{\pi}_b$
$\hat{\eta}$
weights $\left(\frac{A-\hat{\pi}_a}{A-\hat{\pi}_b}\right) K_{hx}$
residuals $\begin{cases} A-\hat{\pi}_b \\ Y-\hat{\eta} \end{cases}$
weighted least squares:
$(Y-\hat{\eta}) \sim b(X-x)(A-\hat{\pi}_b)$
w/weights $\left(\frac{A-\hat{\pi}_a}{A-\hat{\pi}_b}\right) K_{hx}$

lp-R-Learner error bounds

Theorem (lp-R-Learner Error Bound)

Assume:

1. *Estimator* $\widehat{\mu}$ *&* *observations* Z *are bdd, &* X *has bdd density.*
2. *PS estimates satisfy* $\epsilon \leq \widehat{\pi}_j \leq 1 - \epsilon$ *for some* $\epsilon > 0$.
3. *Eigenvalue condition on design matrices* $\widehat{Q}$, $\widetilde{Q}$ *(see paper).*
4. $(\widehat{\pi}_j, \widehat{\mu})$ *satisfy Condition NE, w/bias bds holding* $\forall x' \in B_h(x)$

Then, if $\frac{k/n}{\sqrt{nh^d}} \to 0$ *and the CATE* $\tau(x)$ *is* γ*-smooth,*

$$\widehat{\tau}_r(x) - \tau(x) = O_{\mathbb{P}}\left(h^\gamma + k^{-2s/d} + k^{-2\alpha/d} + \frac{1}{\sqrt{nh^d}}\left(1 + \frac{k}{n}\right)\right)$$

where $s = \frac{\alpha+\beta}{2}$ *is avg smoothness of propensity score & regression.*

Error bound discussion

The first three terms $h^\gamma + k^{-2s/d} + k^{-2\alpha/d}$ are the bias

- h^γ is the bias of an oracle with access to (π, μ)
- the other two terms are from nuisance estimation: $k^{-2s/d}$ is product of $\widehat{\pi}$ and $\widehat{\mu}$ bias, while $k^{-2\alpha/d}$ is squared $\widehat{\pi}$ bias

Heuristic: lp-R-Learner uses least squares, so like product of "$(X^\mathrm{T} X)^{-1}$" (involving $\widehat{\pi}_a$ & $\widehat{\pi}_b$) with "$X^\mathrm{T} Y$" (involving $\widehat{\pi}_a$ & $\widehat{\mu}$)

The next two terms $\frac{1}{\sqrt{nh^d}}\left(1 + \frac{k}{n}\right)$ are the variance

- $(nh^d)^{-1/2}$ is the variance of an oracle with access to (π, μ)
- second is the product of nuisance SDs with oracle variance
- standard setup would require $k \log k / n \to 0$ for Condition NE to hold, making nuisance variance asymptotically negligible

Error bound discussion

Result shows that an undersmoothed lp-R-Learner can achieve oracle rate under *weaker conditions* than DR-Learner bound

- $s \geq \frac{d/4}{1+d/2\gamma} \implies$ up to $1/2$ the smoothness
- this is lower bar than $s \geq d/4$ condition for $\sqrt{n}$-rates for ATE
- note interesting interaction with γ:
  $\gamma \to \infty \implies s \geq d/4$, and $\gamma \to 0 \implies s \geq 0$

New paper: we prove this condition is minimal in a minimax sense!

Note when oracle rate is not achieved, the rate $n^{-2s/d}$ is slower than usual functional minimax rate $n^{-4s/(4s+d)}$

- e.g., when $s = d/8$ it is $n^{-1/4}$ versus $n^{-1/3}$

Estimator #3: Higher-order R-learner

The proposed estimator is then defined as

$$\widehat{\tau}(x_0) = \rho(x_0)^{\mathrm{T}} \widehat{Q}^{-1} \widehat{R} \tag{1}$$

where

$$\widehat{Q} = \mathbb{U}_n \left[\rho(X_1) K_h(X_1) \left\{ \widehat{\varphi}_{a1}(Z_1) + \widehat{\varphi}_{a2}(Z_1, Z_2) K_h(X_2) \right\} \rho(X_1)^{\mathrm{T}} \right]$$

$$\widehat{R} = \mathbb{U}_n \left[\rho(X_1) K_h(X_1) \left\{ \widehat{\varphi}_{y1}(Z_1) + \widehat{\varphi}_{y2}(Z_1, Z_2) K_h(X_2) \right\} \right]$$

and

$$\widehat{\varphi}_{a1}(Z) = A\{A - \widehat{\pi}(X)\}$$

$$\widehat{\varphi}_{y1}(Z) = \{Y - \widehat{\mu}(X)\}\{A - \widehat{\pi}(X)\}$$

$$\widehat{\varphi}_{a2}(Z_1, Z_2) = -\{A_1 - \widehat{\pi}(X_1)\} b_h(X_1)^{\mathrm{T}} \widehat{\Omega}^{-1} b_h(X_2) A_2$$

$$\widehat{\varphi}_{y2}(Z_1, Z_2) = -\{A_1 - \widehat{\pi}(X_1)\} b_h(X_1)^{\mathrm{T}} \widehat{\Omega}^{-1} b_h(X_2) \{Y_2 - \widehat{\mu}(X_2)\}$$

$$b_h(x) = h^{-d} b\{1/2 + (x - x_0)/h\} \mathbb{1}(\|x - x_0\| \leq h)$$

$$\widehat{\Omega} = \int b_h(x) K_h(x) b_h(x)^{\mathrm{T}} \, d\widehat{F}(x)$$