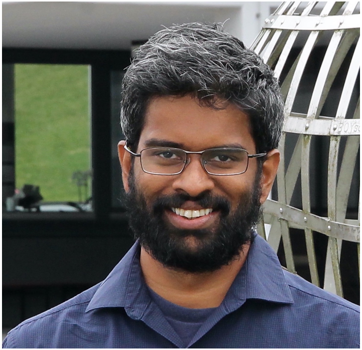
Fundamental limits of structure-agnostic functional estimation

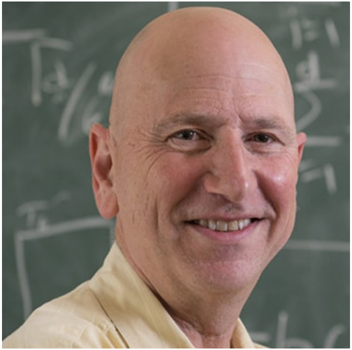
Edward Kennedy

Siva Balakrishnan & Larry Wasserman

Department of Statistics & Data Science
Carnegie Mellon University

ACIC, 24 May 2023





Punchline

Optimality & fundamental statistical limits in causal inference

- much is unknown, many open problems
- e.g., what's best possible performance of effect estimator?

To shed some light on this, in this work we give:

- new model & framework for black-box functional estimation
- new minimax rates for functionals/parameters in Gaussian sequence model, density functionals, & causal inference

Causal inference & functional estimation

After identification, many causal problems equate to statistical functional/parameter estimation

E.g., denote covariates X, treatment A, outcome Y, and

$$\pi(x) = \mathbb{P}(A = 1 \mid X = x), \quad \mu_1(x) = \mathbb{E}(Y \mid X = x, A = 1)$$

then under *consistency* / *positivity* / *no unmeasured confounding*:

$$\mathbb{E}(Y^1) = \mathbb{E}\Big\{\mu_1(X)\Big\} = \mathbb{E}\left\{\frac{AY}{\pi(X)}\right\}$$

Goal is <u>not</u> to estimate whole distribution P, or even (π, μ_1), well

- instead, we want accurate estimates of causal parameter
- similar to other functional estimation settings *outside causal*

Expected conditional covariance

Here we focus on the expected conditional covariance parameter

$$\psi = \mathbb{E}\{\text{cov}(A, Y \mid X)\} = \mathbb{E}\Big\{AY - \pi(X)\mu(X)\Big\}$$

for $\mu(x) = \mathbb{E}(Y \mid X = x)$, which arises in many diverse settings:

- *constant effect estimators under misspecification*
- *overlap weights / weighted effects* (Crump et al. 2006)
- *independence testing* (Shah & Peters 2020)
- *causal influence* (Diaz 2022)
- *marginal incremental effects* (Zhou & Opacic 2022)

Some estimators

A plug-in estimator:

$$\widehat{\psi}_{pi} = \mathbb{P}_n\Big\{AY - \widehat{\pi}(X)\widehat{\mu}(X)\Big\}$$

A doubly-robust / first-order estimator (e.g., Robinson 1988):

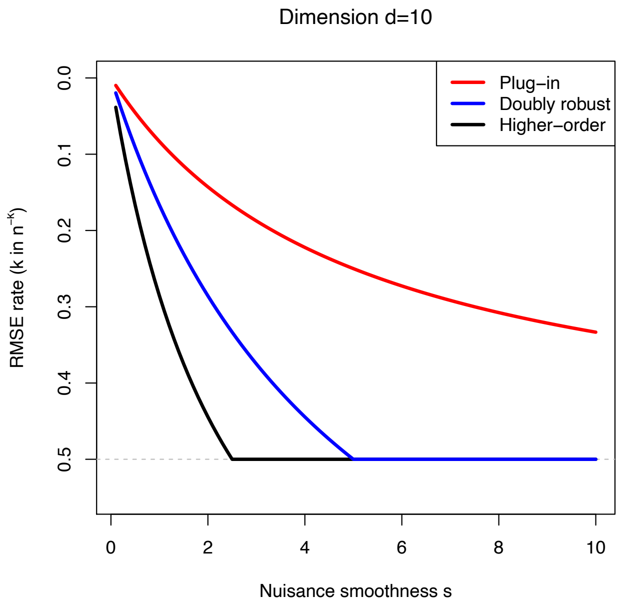
$$\widehat{\psi}_{dr} = \mathbb{P}_n\left[\Big\{A - \widehat{\pi}(X)\Big\}\Big\{Y - \widehat{\mu}(X)\Big\}\right]$$

A higher-order estimator (Robins et al. 2008):

$$\widehat{\psi}_{hi} = \widehat{\psi}_{dr} - \frac{1}{n(n-1)}\sum_{i \neq j}\Big\{A_i - \widehat{\pi}(X_i)\Big\}K_h(X_i, X_j)\Big\{Y_j - \widehat{\mu}(X_j)\Big\}$$

How should we compare these & similar estimators?

- one option: Holder smoothness classes ($\approx$ s bdd derivatives)


Dimension d=10
Plug−in
Doubly robust
Higher−order
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

Black-box / structure-agnostic viewpoint

Regardless of smoothness, the doubly robust estimator satisfies

$$\mathbb{E}|\widehat{\psi}_{dr} - \psi| \lesssim \frac{1}{\sqrt{n}} + \|\widehat{\pi} - \pi\|\|\widehat{\mu} - \mu\|$$

and this error can be small under sparsity, bdd variation, etc.

This motivates black-box approach we often see in practice:

- *throw kitchen sink* at estimating (π, μ)
- put into plug-in/DR estimator, hoping rates *"fast enough"*

But this approach is sub-optimal in smoothness classes

- need more complicated higher-order estimators
- "structure-agnostic" guarantees not so beneficial here

Q: Can we formalize black-box model? What is optimal there?

Minimax optimality

A natural way to characterize optimality is via the minimax rate

$$R_n = \inf_{\widehat{\psi}} \sup_{P \in \mathcal{P}} \mathbb{E}_P |\widehat{\psi} - \psi_P|$$

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

Minimax rates have crucial implications, practical & theoretical

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

Minimax rates are well-understood in many problems:

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

A new minimax framework

We propose a new **black-box model** for minimax analysis

- we only assume pilot propensity $\widehat{\pi}$ and regression $\widehat{\mu}$ estimators are accurate in an $L_2(P)$ sense, nothing else

Our model is:

$$\mathcal{P}(r_n, s_n) = \Big\{ \text{all distributions } P : \|\widehat{\pi} - \pi\| \lesssim r_n, \ \|\widehat{\mu} - \mu\| \lesssim s_n \Big\}$$

(along with some boundedness conditions)

We do not assume (r_n, s_n) are known to the statistician

- so estimators in this model will be adaptive to (r_n, s_n)

Now the formal question is

$$\inf_{\widehat{\psi}} \sup_{P \in \mathcal{P}(r_n, s_n)} \mathbb{E}_P |\widehat{\psi} - \psi_P| \asymp ???$$

A new minimax framework

Some notable distinctions vs. usual (e.g., smooth/sparse) models:

We impose structure implicitly via accuracy in pilot estimators

- assumption strength depends on the accuracy (r_n, s_n)

Following popular practice, we take conditional perspective

- half sample to estimate nuisances, *rest to estimate functional*
- we treat pilot estimates $(\widehat{\pi}, \widehat{\mu})$ as fixed
- Bickel & Ritov (1988), Robins et al (2008), Chernuzhukov et al (2018), Foster & Syrgkanis (2019), etc.

Local minimax flavor

- can think of this as a local minimax problem, localized around $(\widehat{\pi}, \widehat{\mu})$, rather than around true parameter (π, μ)

The main result

Theorem

Let $\mathcal{P}(r_n, s_n)$ denote the model where

$$\|\widehat{\pi} - \pi\| \lesssim r_n \quad \text{and} \quad \|\widehat{\mu} - \mu\| \lesssim s_n.$$

Then the minimax rate is

$$\inf_{\widehat{\psi}} \sup_{P \in \mathcal{P}(r_n, s_n)} \mathbb{E}_P |\widehat{\psi} - \psi_P| \asymp \frac{1}{\sqrt{n}} + r_n \times s_n$$

(see paper for similar sequence model / density functional results).

$\rightarrow$ Here doubly robust estimator can't be meaningfully improved!

Minimax lower bound

Intuition for minimax 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

Three ingredients in deriving minimax lower bound:

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

Construction

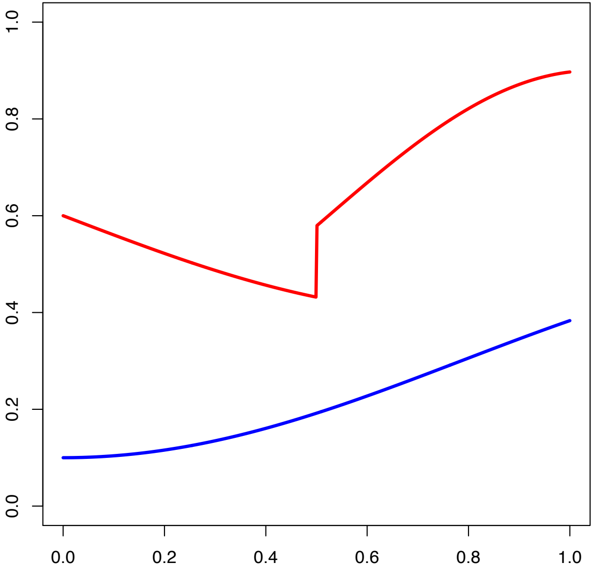
Intuition: perturbed nuisances need not be smooth

- can make them essentially impossible to estimate
- then only information comes from pilot estimates

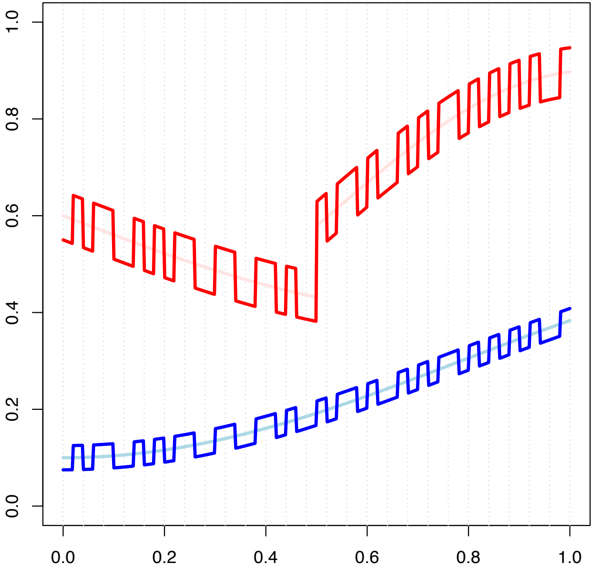
Pair of distributions:

- under null P, take (π, μ) to be given estimates $(\widehat{\pi}, \widehat{\mu})$
- under alternative Q_λ, add k bumps w/random direction λ, and height approx. equal to r_n and s_n (for π, μ, resp.)

Null P



Alt. Q_λ



Construction

Intuition: perturbed nuisances need not be smooth

- can make them essentially impossible to estimate
- then only information comes from pilot estimates

Pair of distributions:

- under null P, take (π, μ) to be given estimates $(\widehat{\pi}, \widehat{\mu})$
- under alternative Q_λ, add k bumps w/random direction λ, and height approx. equal to r_n and s_n (for π, μ, resp.)

Functional separation:

$$\psi(P) = \int \widehat{\pi}\widehat{\mu} \ , \quad \psi(Q_\lambda) - \psi(P) \ \gtrsim \ r_n \times s_n$$

Hellinger distance: $H^2 \lesssim \frac{n^2}{k} \left(r_n^4 + s_n^4 \right)$

Some implications

- There's a strong sense in which popular DR/TMLE/DML -style estimators are optimal, from black-box perspective
  - even when nuisances estimated at slower than $n^{-1/4}$ rates
- rate benefits from *higher-order estimators* will necessarily require more assumptions
- "doubly robust inference" methods, which yield root-n rates as long as either nuisance is converging at $n^{-1/4}$, are necessarily using more assumptions (sparse glm, smoothness)

Summary

Still a long way to go understanding optimality in causal inference

Our contributions here:

1. new black-box framework, giving complementary perspective
2. new structure-agnostic minimax rates for functional estimation in sequence model, density/causal parameters

Lots of unanswered questions & future work:

- other functionals, classes of functionals; adaptivity; other models; and more

On arxiv now! $\rightarrow$ arxiv.org/abs/2305.041167

The Fundamental Limits of Structure-Agnostic Functional Estimation

Sivaraman Balakrishnan†, Edward Kennedy† and Larry Wasserman†

Department of Statistics and Data Science†

Carnegie Mellon University,
Pittsburgh, PA 15213.

{siva,edward,larry}@stat.cmu.edu

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

Thank you!