When doubly robust estimators collapse

Betsy Ogburn
Johns Hopkins University

arxiv.org/abs/2304.14545

David Bruns-Smith
UC Berkeley

Oliver Dukes
Ghent University

Avi Feller
UC Berkeley

Thank you to the Simons Institute for Theoretical Computer Science

We will connect three threads of research to elucidate a family of popular double machine learning estimators:

1. Linear weighting is equivalent to linear outcome regression
2. A parametric outcome model can be doubly robust
3. Theoretically, undersmoothed nonparametric regression can be efficient
   But practically, undersmoothing is a “dark art” (Bruce Hansen)

⟶ Many popular “double machine learning” estimators collapse and can be analyzed as undersmoothed outcome regression

Goals

- To demystify Automatic Debased Machine Learning (AutoDML), a class of double machine learning that has exploded in popularity
- To see how and why some outcome models are doubly robust
- To provide practical guidance for hyperparameter selection in kernel ridge regression

Outline

- Estimating a missing (counterfactual) mean
- Balancing weights and augmented balancing weights (aka AutoDML)
- New results show that augmented linear balancing weights collapse to a linear outcome model
- Implications for kernel ridge and double kernel ridge in practice

Review:
Estimating a counterfactual mean

Fundamental problem of causal inference

Goal: estimate $ATE = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$

Challenge: need to estimate unobserved counterfactual means

$$\mathbb{E}[Y(1) \mid T = 0] \text{ and } \mathbb{E}[Y(0) \mid T = 1]$$

Estimating $\mathbb{E}[Y(1)]$

3 identifying functionals → 3 types of estimators:

- Weighting (aka Riesz representer)

$$\mathbb{E}\left[\frac{T}{e(X)}Y\right], e(X) = P(T = 1 \mid X)$$

- Outcome regression

$$\mathbb{E}[\mathbb{E}[Y \mid T = 1, X]]$$

- Doubly robust

$$\mathbb{E}\left[\mathbb{E}[Y \mid T = 1, X] + \frac{T}{e(X)}\{Y - \mathbb{E}[Y \mid T = 1, X]\}\right]$$

When everything is linear…

In this talk, $\mathbb{E}[Y \mid T, X]$ and $\frac{T}{e(X)}$ are both assumed to be linear in X

(or in a possibly infinite basis expansion of X:

RKHS, HAL, infinite-width neural networks, “honest” random forests)

When everything is linear...

$$w(X) = \frac{T}{e(X)} = X\theta \qquad \Longrightarrow \quad \hat{\mathbb{E}}[Y(1)] = X\hat{\theta}Y, \text{ where } \hat{\theta} = (X^T X)^{-1} X^T$$

$$\mathbb{E}[Y \,|\, T = 1, X] = X\beta \qquad \Longrightarrow \quad \hat{\mathbb{E}}[Y(1)] = \hat{\beta}Y \text{ where } \hat{\beta} = X(X^T X)^{-1} X^T$$

OLS is doubly robust!

$$\hat{\beta}_{\text{ols}}$$

$$X(X^T X)^{-1} X^T Y$$

$$\hat{w}$$

[Robins et al., 2007]

OLS is doubly robust!

$$\underbrace{X(X^TX)^{-1}X^TY}_{\hat{w}}$$

$\hat{\beta}_{\text{ols}}$

$$\rightarrow \mathbb{E}[Y(1)] \text{ if}$$

1. $\hat{w} \rightarrow \frac{1}{e(X)}$
2. $X\hat{\beta}_{ols} \rightarrow \mathbb{E}[Y \mid T = 1, X]$

[Robins et al., 2007]

OLS is doubly robust!

$$\underbrace{X\hat{\beta} + \hat{w}(Y - X\hat{\beta})}_{\text{Doubly robust}} = \underbrace{X\hat{\beta}}_{\text{Outcome regression}} = \underbrace{\hat{w}Y}_{\text{Weighting}}$$

[Robins et al., 2007]

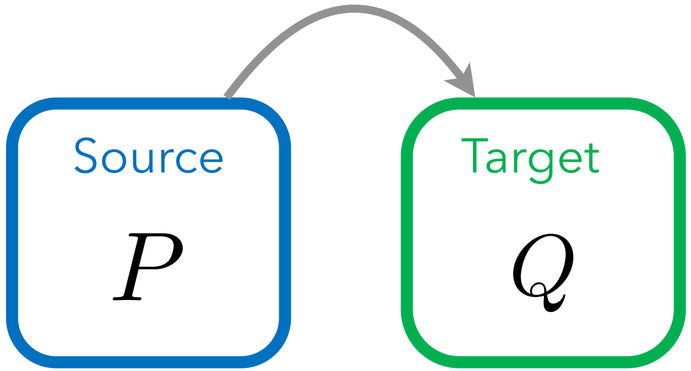
Foreshadowing

We generalize the “OLS is doubly robust” result to any nonparametric outcome and weight models that are linear in a basis expansion of X

Review:

Estimating a missing mean

Estimating the mean of a (partially) unobserved outcome



$\mathbb{E}_Q[Y]$

Y observed

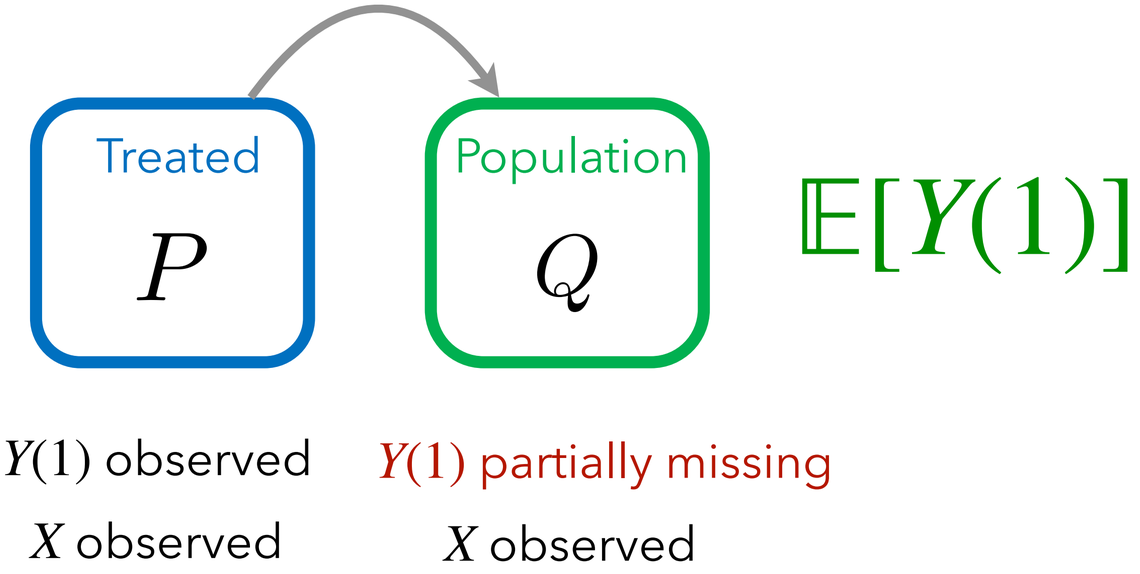
X observed

Y missing

X observed

Weighting estimator: $\mathbb{E}_P\left[w(X)Y\right]$ with $w(X) = \frac{dQ}{dP}(X)$

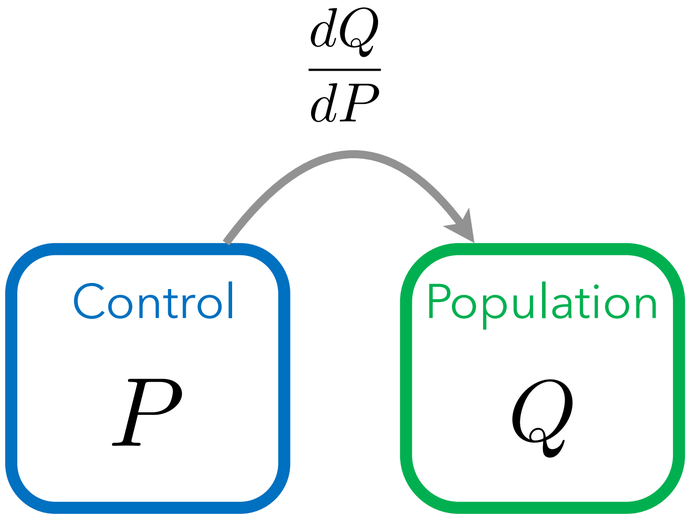
Estimating the mean of a (partially) unobserved outcome



Weighting estimator: $\mathbb{E}_P\left[w(X)Y\right]$ with $w(X) = \dfrac{1}{e(X)}$

Estimating the weights

Traditional IPW: $\hat{e}(X) \approx \mathbb{P}(\text{treat} \mid X) \rightarrow$ plug in $\dfrac{1}{\hat{e}(X)}$



Balancing weights: $\boxed{\hat{w}(X) \approx \dfrac{dQ}{dP}(X)}$

- Estimate via constrained optimization / moment equations
  [Ben-Michael, Feller, et al., 2021]
- Proposed many times in different areas:

  Survey calibration [Deville and Särndal, 1992]

  Synthetic control [Abadie and Gardeazabal, 2003]

  Direct density ratio estimation [Kanamori, Hido, Sugiyama, 2009]

  Balancing weights [Hainmuller 2012, Zubizarreta 2015]

  Automatic estimator of Riesz representer [Chernozhukov et al, 2022]

Review:
Balancing Weights

Review: Ben-Michael, Feller, et al. [2021]

Automatic Debiased Machine Learning (AutoDML)

Debiased machine learning of global and local parameters using regularized Riesz representers V Chernozhukov, WK Newey, R Singh The Econometrics Journal 25 (3), 576--601	311 *	2022
Automatic Debiased Machine Learning of Causal and Structural Effects V Chernozhukov, WK Newey, R Singh Econometrica 90 (3), 967-1027	86 *	2022
Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests V Chernozhukov, W Newey, VM Quintas-Martínez, V Syrgkanis International Conference on Machine Learning, 3901-3914	11	2022

Generalized Regression Estimators (GREG)

Calibration estimators in survey sampling JC Deville, CE Särndal Journal of the American statistical Association 87 (418), 376-382	2577	1992

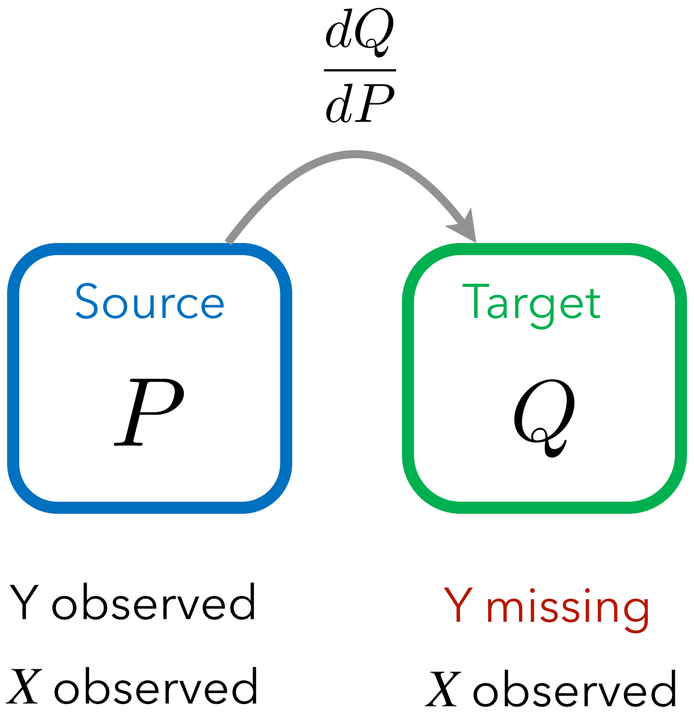
Direct Density Ratio Estimation

Direct importance estimation with model selection and its application to covariate shift adaptation M Sugiyama, S Nakajima, H Kashima, P Buenau, M Kawanabe Advances in neural information processing systems 20	922	2007
Density ratio estimation in machine learning M Sugiyama, T Suzuki, T Kanamori Cambridge University Press	541	2012
A least-squares approach to direct importance estimation T Kanamori, S Hido, M Sugiyama The Journal of Machine Learning Research 10, 1391-1445	537	2009

Balancing Weights

Approximate residual balancing: Debiased inference of average treatment effects in high dimensions S Athey, GW Imbens, S Wager Journal of the Royal Statistical Society Series B 80 (4), 597-623	470 *	2018
Stable weights that balance covariates for estimation with incomplete outcome data JR Zubizarreta Journal of the American Statistical Association 110 (511), 910-922	407	2015
Augmented minimax linear estimation DA Hirshberg, S Wager The Annals of Statistics 49 (6), 3206-3227	95 *	2021

Background: Balancing weights



$$\mathbb{E}_P\left[\frac{dQ}{dP}(X)Y\right] = \mathbb{E}_Q[Y]$$

$w(X)$

Background: Balancing weights

Population Balance Property

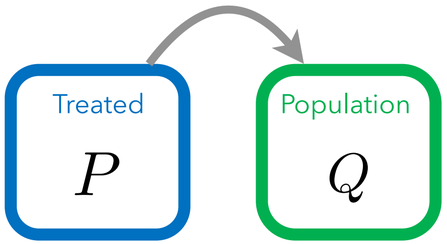
$$\forall f(), \mathbb{E}_p[w(X)f(X)] - \mathbb{E}_q[f(X)] = 0$$

$$\frac{dQ}{dP}(X)$$

unique weights that balance all $f(X)$

(Why $f(X)$? Think of $\mathbb{E}_p[Y|X]$)

Background: Balancing weights



$\mathbb{E}_Q[Y(1)]$

Population Balance Property

$$\mathbb{E}_p[w(X)f(X)] - \mathbb{E}_q[f(X)] = 0$$

$$\frac{1}{e(X)}$$

unique weights that *nonparametrically* balance $\mathbb{E}[Y \mid X]$

Background: Balancing weights

$$\forall f(),\ \mathbb{E}_p[w(X)f(X)] - \mathbb{E}_q[f(X)] = 0$$

$$\Longleftrightarrow$$

$\frac{dQ}{dP}(X)$ is the unique solution to

$$\sup_f \{\mathbb{E}_P[w(X)f(X)] - \mathbb{E}_Q[f(X)]\} = 0$$

imbalance

Background: Balancing weights

$$\sup_{f} \{\mathbb{E}_P[w(X)f(X)] - \mathbb{E}_Q[f(X)]\} = 0$$

$$\sup_{f \in \mathcal{F}} \{\mathbb{E}_P[w(X)f(X)] - \mathbb{E}_Q[f(X)]\} = 0$$

restrict our attention to function class $\mathcal{F}$

Background: Balancing weights

$$\sup_{f \in \mathcal{F}} \{\mathbb{E}_P[w(X)f(X)] - \mathbb{E}_Q[f(X)]\} = 0$$

$$\frac{dQ}{dP}(X)$$

No longer unique weights to satisfy this property!

Background: Balancing weights

$$\sup_{f \in \mathcal{F}} \{\mathbb{E}_P[w(X)f(X)] - \mathbb{E}_Q[f(X)]\} = 0$$

$\mathbb{E}[Y|X]$ in $\mathcal{F} \implies$

any weights that satisfy this equation can be used to identify and estimate $\mathbb{E}_Q[Y]$

Background: Balancing weights

$$\sup_{f \in \mathcal{F}} \{\mathbb{E}_P[w(X)f(X)] - \mathbb{E}_Q[f(X)]\} \leq \delta$$

$\mathbb{E}[Y|X]$ in $\mathcal{F} \implies$

any weights that satisfy this inequality have small bias for estimating $\mathbb{E}_Q[Y]$

Background: Linear balancing weights

Consider the class of linear functionals

$$\mathcal{F} = \{f(x) = x\theta\}$$

$$\sup_{f \in \mathcal{F}} \left\{ \mathbb{E}_p[w(X) f(X)] - \mathbb{E}_q[f(X)] \right\} \longrightarrow \left\| w(X) X_p - \bar{X}_q \right\|_*$$

Population imbalance

Finite sample linear imbalance in * norm (e.g., ℓ_2 imbalance)

Background: Linear balancing weights

Minimum variance

$$\min_{w \in \mathbb{R}^n} \|w\|_2^2$$

$$\text{s.t.} \left\| wX_p - \bar{X}_q \right\|_* \leq \delta$$

- hyperparameter
- selected in finite samples with CV
- $\to 0$ with *n*

Background: Linear balancing weights

Balancing Weights:

$$\min_{w \in \mathbb{R}^n} \|w\|_2^2$$

$$\text{s.t. } \left\| w X_p - \bar{X}_q \right\|_* \leq \delta$$

Equivalent Problems
(Primal / Dual)

Automatic Estimate of Riesz Representer:

$$\min_{\theta \in \mathbb{R}^d} \left\{ \theta^T (X_p^T X_p) \theta - 2\theta^T \bar{X}_q + \delta' \|\theta\| \right\}$$

[Chernozhukov et al, 2022]

Background: Linear balancing weights

Balancing Weights:

$$\min_{w \in \mathbb{R}^n} \|w\|_2^2$$

$$\text{s.t. } \left\| w X_p - \bar{X}_q \right\|_* \leq \delta$$

Equivalent Problems
(Primal / Dual)

Automatic Estimate of Riesz Representer:

$$\min_{\theta \in \mathbb{R}^d} \left\{ \theta^T (X_p^T X_p) \theta - 2\theta^T \bar{X}_q + \delta' \|\theta\| \right\}$$

[Chernozhukov et al, 2022]

$$\hat{w}(X) := X\hat{\theta}$$

Augmented balancing weights, aka AutoDML

(Regularized) outcome model

(Regularized) linear balancing weights

$$\hat{\mathbb{E}}[X_q \hat{\beta}_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}(X_p)(y_p - X_p \hat{\beta}_{\text{reg}})]$$

Generalized Regression Estimator (GREG) [Deville and Särndal, 1992]

De-biased Lasso [Javanmard and Montanari, 2014]

Approximate Residual Balancing [Athey, Imbens, and Wager, 2018]

Augmented Minimax Linear Estimation [Hirshberg and Wager, 2021]

Augmented Synthetic Control Method [Ben-Michael, Feller, and Rothstein, 2021]

Automatic Debiased Machine Learning (AutoDML) [Chernozhukov et al, 2022]

Regularized outcome model optimizes MSE

$$\hat{\mathbb{E}}[X_q \hat{\beta}_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}(X_p)(y_p - X_p \hat{\beta}_{\text{reg}})]$$

Augmentation term: corrects bias,
undersmoothes relative to outcome model

Regularized outcome model optimizes MSE

$$\hat{\mathbb{E}}[X_q \hat{\beta}_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}(X_p)(y_p - X_p \hat{\beta}_{\text{reg}})]$$

Augmentation term: corrects bias, undersmoothes relative to outcome model

Double machine learning

- Bias of DR estimator behaves as product of $(\hat{\beta} - \beta)(\hat{w} - w)$
- Can get $\sqrt{n}$ rates even if $\hat{\beta}$ and $\hat{w}$ converge slowly

Terminology break

"Ridge"

"ℓ_2-penalized outcome model"

"ℓ_2-penalized weight model"

dual norm

"ℓ_2 balancing"

"Lasso"

"ℓ_1-penalized outcome model"

"ℓ_1-penalized weight model"

dual norm

"ℓ_∞ balancing"

A little more notation…

Outcome regularization hyperparameter

λ

Weight regularization hyperparameter

δ

Maximum allowed imbalance

$$\|wX_p - \bar{X}_q\|_* \leq \delta$$

Exact Balancing Weights

$\delta = 0$ $\quad \hat{w}_{\text{exact}} X_p = \bar{X}_q$

Clean up covariates

Center Covariates

$$\bar{X}_p = 0 \qquad \bar{y}_p = 0$$

$$\Delta_q := \bar{X}_q - \bar{X}_p = \bar{X}_q$$

Covariate Shift

≡ add intercept to weight and outcome models

Diagonal Covariance

$$X_p^T X_p = \mathrm{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_d^2)$$

discuss general case in paper

New Results

(Regularized) outcome model

$$\hat{\mathbb{E}}[X_q \hat{\beta}_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}(X_p)(y_p - X_p \hat{\beta}_{\text{reg}})]$$

(Regularized) outcome model

(Regularized) linear balancing weights

$$\hat{w}(X) := X\hat{\theta}$$

$$\hat{\mathbb{E}}[X_q\hat{\beta}_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}(X_p)(y_p - X_p\hat{\beta}_{\text{reg}})]$$

(Regularized) outcome model

(Regularized) linear balancing weights

$$\hat{w}(X) := X\hat{\theta}$$

$$\hat{\mathbb{E}}[X_q\hat{\beta}_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}(X_p)(y_p - X_p\hat{\beta}_{\text{reg}})]$$

$$= \hat{\mathbb{E}}[X_q\hat{\beta}_{\text{aug}}]$$

(Regularized) outcome model

(Regularized) linear balancing weights

$$\hat{w}(X) := X\hat{\theta}$$

$$\hat{\mathbb{E}}[X_q\hat{\beta}_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}(X_p)(y_p - X_p\hat{\beta}_{\text{reg}})]$$

$$= \hat{\mathbb{E}}[X_q\hat{\beta}_{\text{aug}}]$$

Equivalent to undersmoothed outcome model

(Regularized) outcome model

(Regularized) linear balancing weights

$$\hat{w}(X) := X\hat{\theta}$$

$$\hat{\mathbb{E}}[X_q\hat{\beta}_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}(X_p)(y_p - X_p\hat{\beta}_{\text{reg}})]$$

X can be infinite dimensional

$$= \hat{\mathbb{E}}[X_q\hat{\beta}_{\text{aug}}]$$

Extend to nonlinear weights

Equivalent to undersmoothed outcome model

(Regularized) outcome model

(Regularized) linear balancing weights

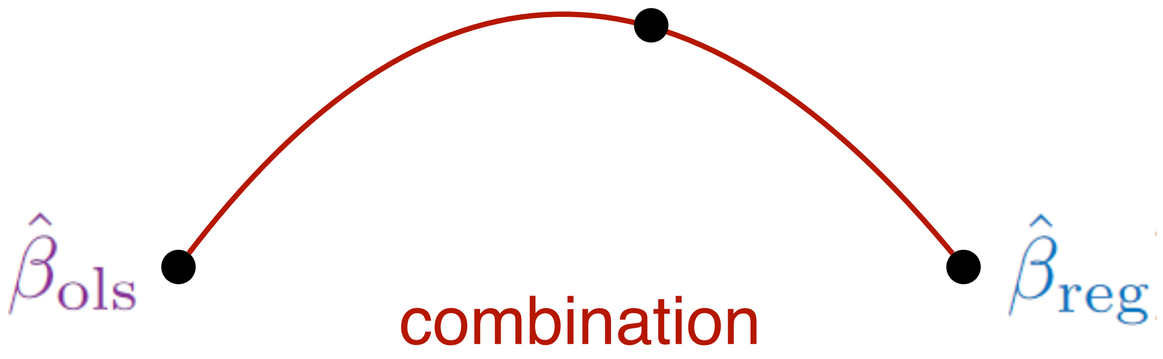
$$\hat{w}(X) := X\hat{\theta}$$

$$\hat{\mathbb{E}}[X_q\hat{\beta}_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}(X_p)(y_p - X_p\hat{\beta}_{\text{reg}})]$$

Can collapse to OLS in practice (e.g., LaLonde)

$$= \hat{\mathbb{E}}[X_q\hat{\beta}_{\text{aug}}]$$

Characterize for different model choices



Recall…

$$y_p^T \hat{w}_{\text{exact}} = \bar{X}_q \hat{\beta}_{\text{ols}}$$

[Fuller, 2002; Kline, 2011; Chattopadhay and Zubizarreta, 2021]

Recall…

$$y_p^T \hat{w}_{\text{exact}} = \bar{X}_q \hat{\beta}_{\text{ols}}$$

$$= \bar{X}_q (X_p^T X_p)^{-1} X_p^T y_p$$

Recall…

$$y_p^T \hat{w}_{\text{exact}} = \bar{X}_q \hat{\beta}_{\text{ols}}$$

$$= \bar{X}_q \overbrace{(X_p^T X_p)^{-1} X_p^T y_p}^{\hat{\beta}_{\text{ols}}}$$

Recall…

$$y_p^T \hat{w}_{\text{exact}} = \bar{X}_q \hat{\beta}_{\text{ols}}$$

$$= \bar{X}_q \underbrace{(X_p^T X_p)^{-1} X_p^T}_{\hat{w}_{\text{exact}}} y_p$$

$\hat{\beta}_{\text{ols}}$

Warmup: Linear balancing Weights ↔ Outcome Modeling

Old equivalence

$$y_p^T \hat{w}_{\text{exact}} = \bar{X}_q \hat{\beta}_{\text{ols}}$$

Warmup: Linear balancing Weights ↔ Outcome Modeling

Old equivalence

$$y_p^T \hat{w}_{\text{exact}} = \bar{X}_q \hat{\beta}_{\text{ols}}$$

Generalization

$$y_p^T \hat{w}_{\delta} = \hat{X}_q \hat{\beta}_{\text{ols}}$$

Warmup: Linear balancing Weights ↔ Outcome Modeling

Old equivalence

$$y_p^T \hat{w}_{\text{exact}} = \bar{X}_q \hat{\beta}_{\text{ols}}$$

Any linear balancing weights

$$\hat{w}_\delta := X_p \hat{\theta}_\delta$$

Generalization

$$y_p^T \hat{w}_\delta = \hat{X}_q \hat{\beta}_{\text{ols}}$$

Warmup: Balancing and OLS

Old equivalence

$$y_p^T \hat{w}_{\text{exact}} = \bar{X}_q \hat{\beta}_{\text{ols}}$$

Any linear balancing weights

$$\hat{w}_\delta := X_p \hat{\theta}_\delta$$

Reweighted covariates

$$\hat{X}_q := \hat{w}_\delta{}^T X_p$$

Generalization

$$y_p^T \hat{w}_\delta = \hat{X}_q \hat{\beta}_{\text{ols}}$$

Augmented Balancing Weights as Linear Regression

Apply last result: $= \hat{X}_q \hat{\beta}_{OLS}$

$$\hat{\mathbb{E}}[X_q \hat{\beta}_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}(X_p)(y_p - X_p \hat{\beta}_{\text{reg}})]$$

Augmented Balancing Weights as Linear Regression

Apply last result: $= \hat{X}_q \hat{\beta}_{OLS}$

$$\hat{\mathbb{E}}[X_q \hat{\beta}_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}(X_p)(y_p - X_p \hat{\beta}_{\text{reg}})] = \hat{\mathbb{E}}[X_q \hat{\beta}_{\text{aug}}]$$

Augmented Balancing Weights as Linear Regression

Apply last result: $= \hat{X}_q \hat{\beta}_{OLS}$

$$\hat{\mathbb{E}}[X_q \hat{\beta}_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}(X_p)(y_p - X_p \hat{\beta}_{\text{reg}})] = \hat{\mathbb{E}}[X_q \hat{\beta}_{\text{aug}}]$$

$$\hat{\beta}_{\text{aug}} = a_\delta \circ \hat{\beta}_{\text{ols}} + (1 - a_\delta) \circ \hat{\beta}_{\text{reg}} \qquad a_\delta \in [0, 1]^d$$

Augmented Balancing Weights as Linear Regression

$$\hat{\mathbb{E}}[X_q \hat{\beta}_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}(X_p)(y_p - X_p \hat{\beta}_{\text{reg}})] = \hat{\mathbb{E}}[X_q \hat{\beta}_{\text{aug}}]$$

$$\hat{\beta}_{\text{aug}} = a_\delta \circ \hat{\beta}_{\text{ols}} + (1 - a_\delta) \circ \hat{\beta}_{\text{reg}} \qquad a_\delta \in [0, 1]^d$$

$\delta \to \infty$

$a_\delta \to 0$

Augmented Balancing Weights as Linear Regression

$$\hat{\mathbb{E}}[X_q \hat{\beta}_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}(X_p)(y_p - X_p \hat{\beta}_{\text{reg}})] = \hat{\mathbb{E}}[X_q \hat{\beta}_{\text{aug}}]$$

$$\hat{\beta}_{\text{aug}} = a_\delta \circ \hat{\beta}_{\text{ols}} + (1 - a_\delta) \circ \hat{\beta}_{\text{reg}} \qquad a_\delta \in [0, 1]^d$$

$\delta \to 0$

$a_\delta \to 1$

Augmented Balancing Weights as Linear Regression

$$\hat{\mathbb{E}}[X_q\hat{\beta}_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}(X_p)(y_p - X_p\hat{\beta}_{\text{reg}})] = \hat{\mathbb{E}}[X_q\hat{\beta}_{\text{aug}}]$$

$$\hat{\beta}_{\text{aug}} = a_\delta \circ \hat{\beta}_{\text{ols}} + (1 - a_\delta) \circ \hat{\beta}_{\text{reg}} \qquad a_\delta \in [0, 1]^d$$

$\delta \to 0$ $\qquad$ $a_\delta \to 1$

AutoDML procedure can collapse to OLS in practice!

This happens frequently when cross-validating the Riesz loss

Implications

$$\hat{\mathbb{E}}[X_q\hat{\beta}_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}(X_p)(y_p - X_p\hat{\beta}_{\text{reg}})] = \hat{\mathbb{E}}[X_q\hat{\beta}_{\text{aug}}]$$

If we could estimate $\mathbb{E}[X_q\hat{\beta}_{aug}]$ directly, it would be doubly robust

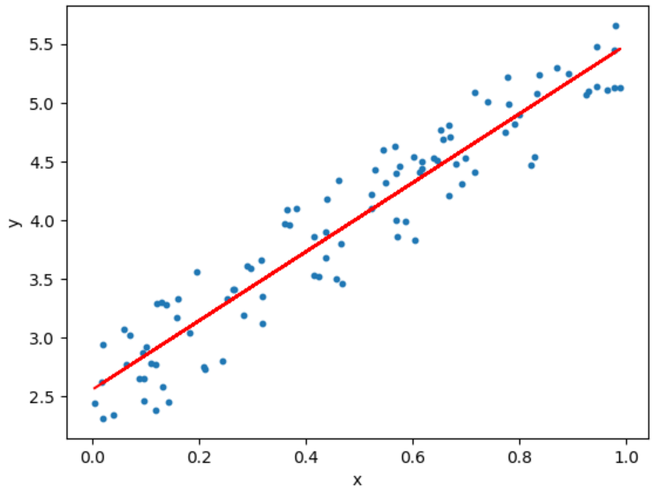
$$\mathbb{E}[X_q\hat{\beta}_{aug}] \rightarrow \mathbb{E}_Q[Y] \text{ if}$$

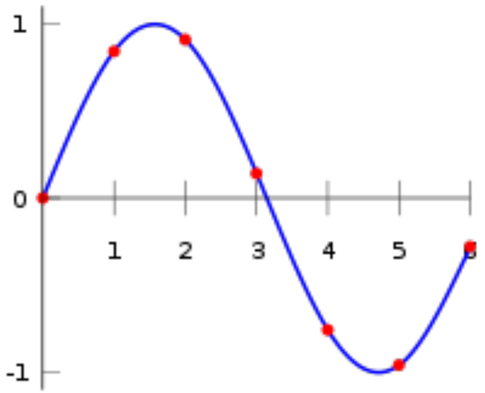
1. $\hat{w}(X_p) \rightarrow \frac{dQ}{dP}(X)$
2. $X_q\hat{\beta}_{reg} \rightarrow \mathbb{E}[Y|X]$

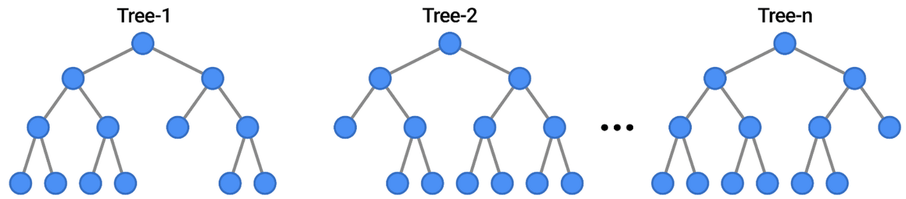
Numeric results hold for a broad class

Linear in any features

- Includes ridge, lasso, kernel ridge, PCR, honest random forests
- Includes d > n or infinite-dimensional settings → use minimum-norm interpolant for OLS
- Last layer embedding from pre-trained LLM







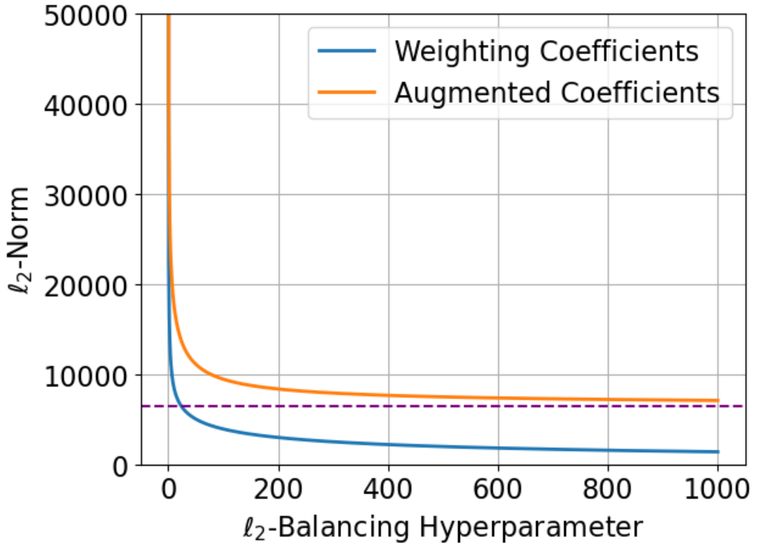
Riesz Representer: Extend to general linear functionals

“Undersmoothed” linear regression: special cases

Ridge outcome + ridge weights (ℓ_2 balancing)

[e.g., Singh, 2021; variation of Ben-Michael et al., 2021]

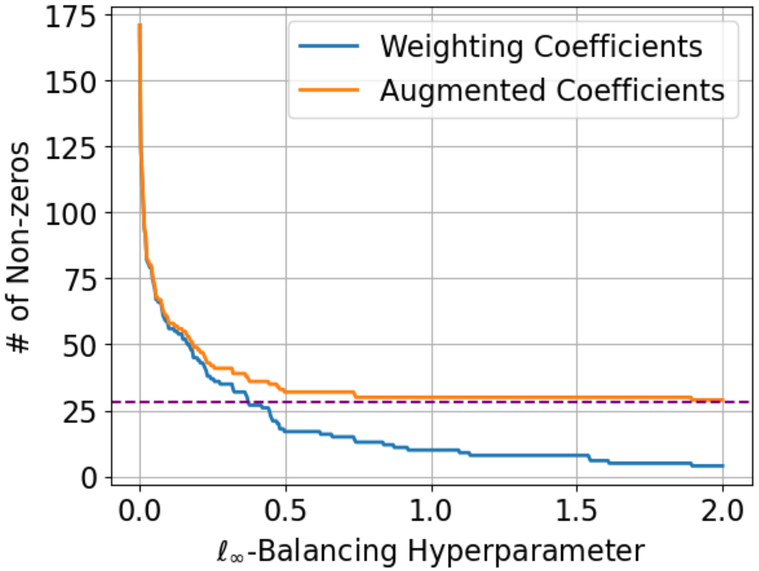
- $\hat{\beta}_{\text{aug}} = \hat{\beta}_{\text{ridge}} \rightarrow$ single, **undersmoothed** ridge regression
- *Asymptotics:* optimally undersmoothed kernel ridge



Lasso outcome + lasso weights (ℓ_∞ balancing)

[e.g., Chernozhukov et al., 2022; variation of Athey et al., 2018]

- $\hat{\beta}_{\text{aug}}$: “double selection” [Belloni et al., 2014]
- Undersmoothing in # of included covariates (ℓ_0 “norm”)



Double (Kernel) Ridge

(ℓ_2 balancing + ridge regression)

Ridge + Ridge: Summary

$$\hat{\mathbb{E}}[X_q \hat{\beta}_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}(X_p)(y_p - X_p \hat{\beta}_{\text{reg}})] = \hat{\mathbb{E}}[X_q \hat{\beta}_{\text{aug}}]$$

$$\hat{\beta}_{\text{aug}} = a_\delta \circ \hat{\beta}_{\text{ols}} + (1 - a_\delta) \circ \hat{\beta}_{\text{reg}} \qquad a_\delta \in [0, 1]^d$$

Ridge + Ridge: Summary

$$\hat{\mathbb{E}}[X_q \hat{\beta}_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}(X_p)(y_p - X_p \hat{\beta}_{\text{reg}})] = \hat{\mathbb{E}}[X_q \hat{\beta}_{\text{aug}}]$$

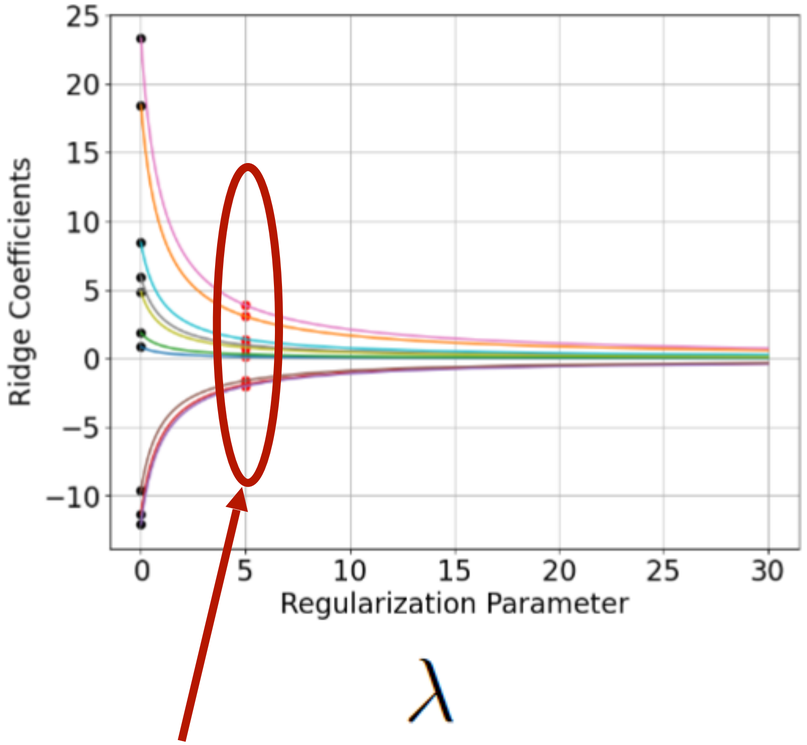
$$\hat{\beta}_{\text{aug}} = a_\delta \circ \hat{\beta}_{\text{ols}} + (1 - a_\delta) \circ \hat{\beta}_{\text{reg}} \qquad a_\delta \in [0, 1]^d$$

$\hat{w}(X)$ + $\hat{\beta}_{\text{reg}}$ ⇒ $\hat{\beta}_{\text{aug}}$ undersmoothed ridge

ridge ridge

Ridge + Ridge

$$\hat{\beta}^{\lambda}_{\text{ridge}}$$



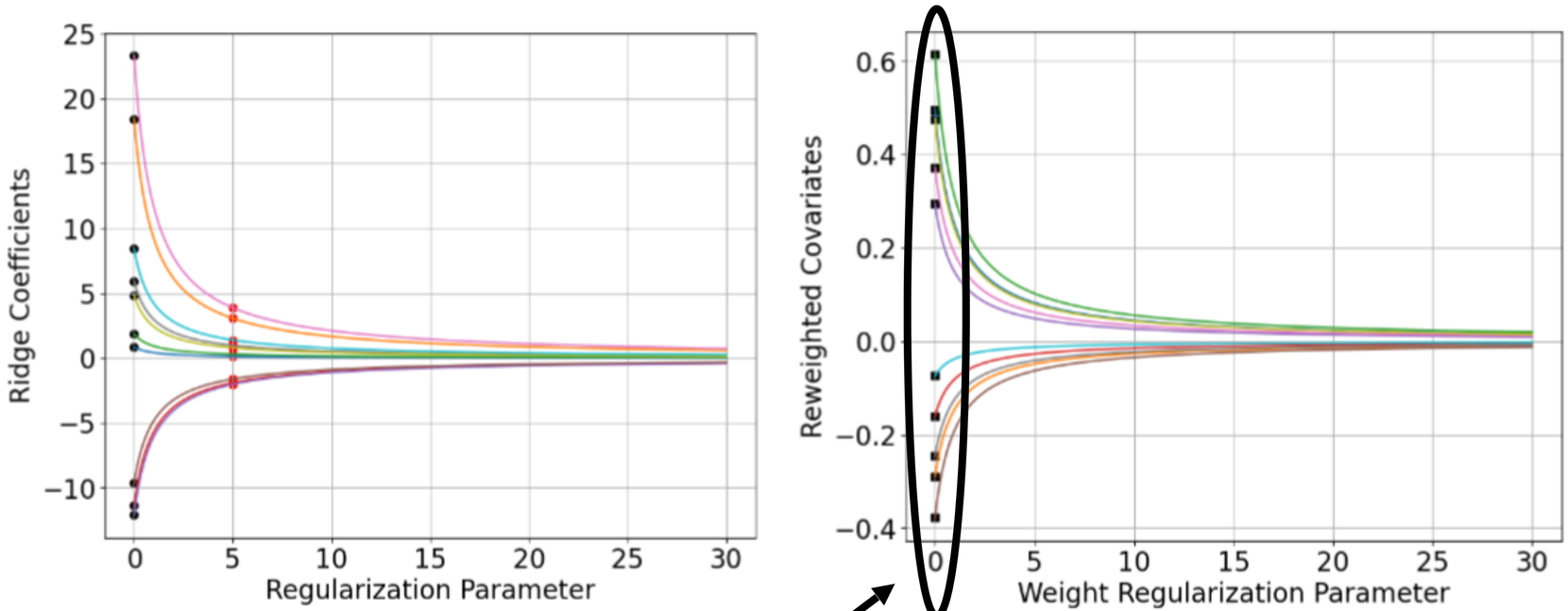
λ

Cross-validation

Ridge + Ridge

$\hat{\beta}^{\lambda}_{\text{ridge}}$

$\hat{X}_q := \hat{w}_\delta{}^T X_p$



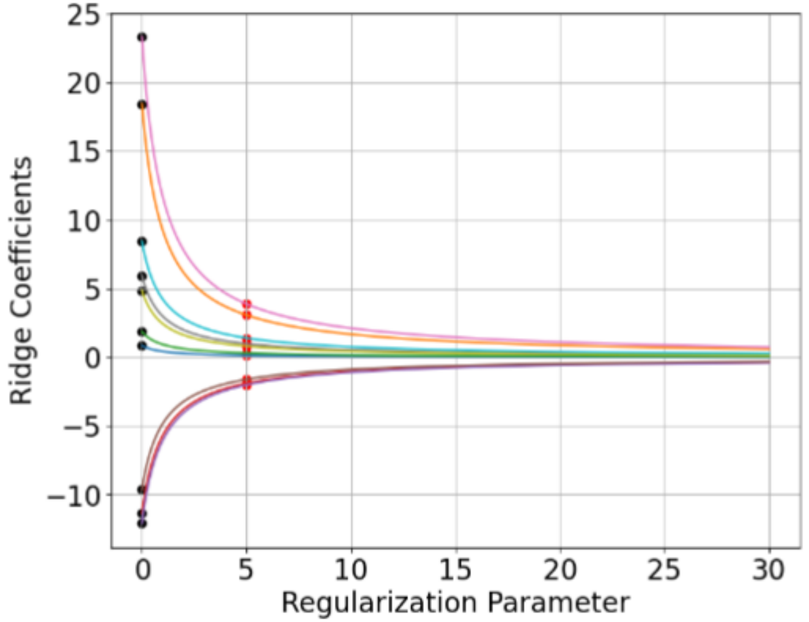
λ

δ

$\bar{X}_q$

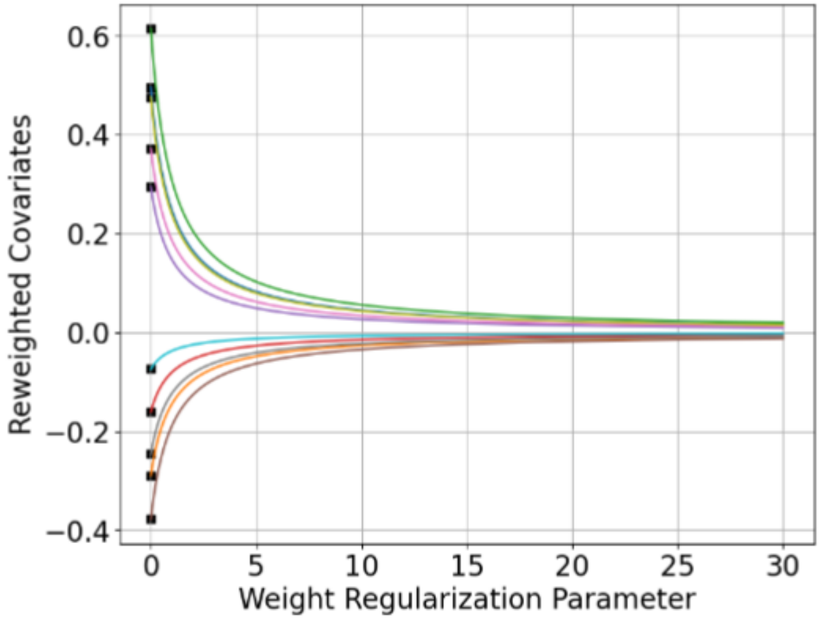
Ridge + Ridge

$\hat{\beta}^{\lambda}_{\text{ridge}}$



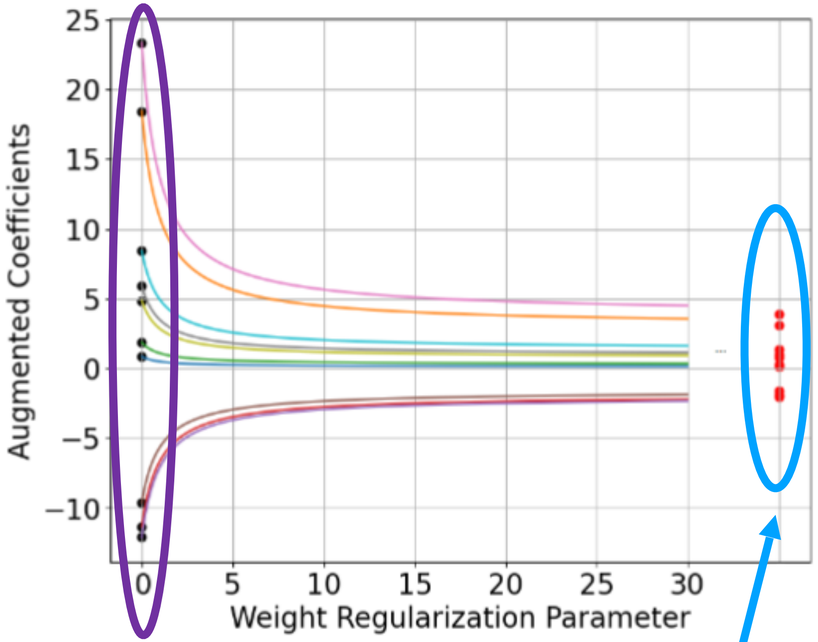
λ

$\hat{X}_q := \hat{w}_\delta{}^T X_p$



δ

$\hat{\beta}_{\text{aug}}$



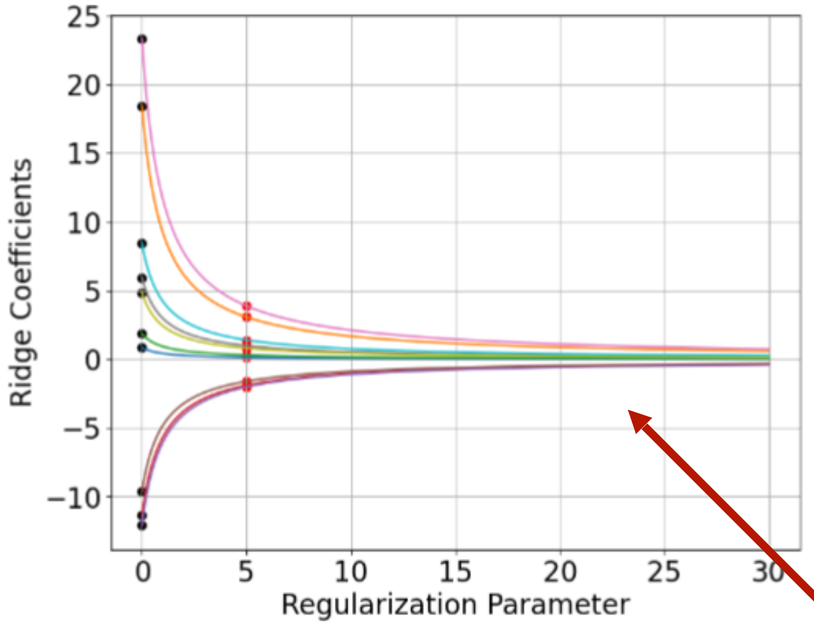
δ

$\hat{\beta}_{\text{ols}}$

$\hat{\beta}^{\lambda_{\text{cv}}}_{\text{ridge}}$

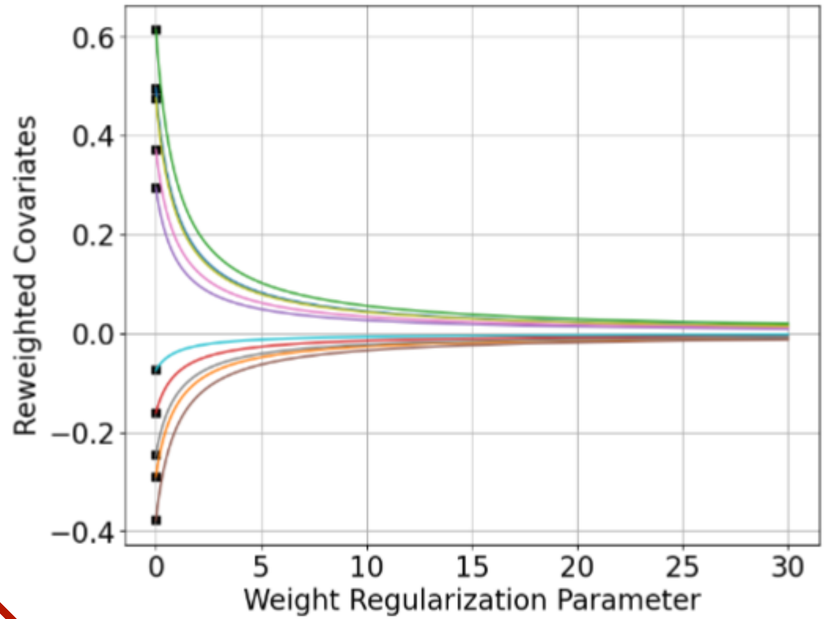
Ridge + Ridge

$\hat{\beta}^{\lambda}_{\text{ridge}}$



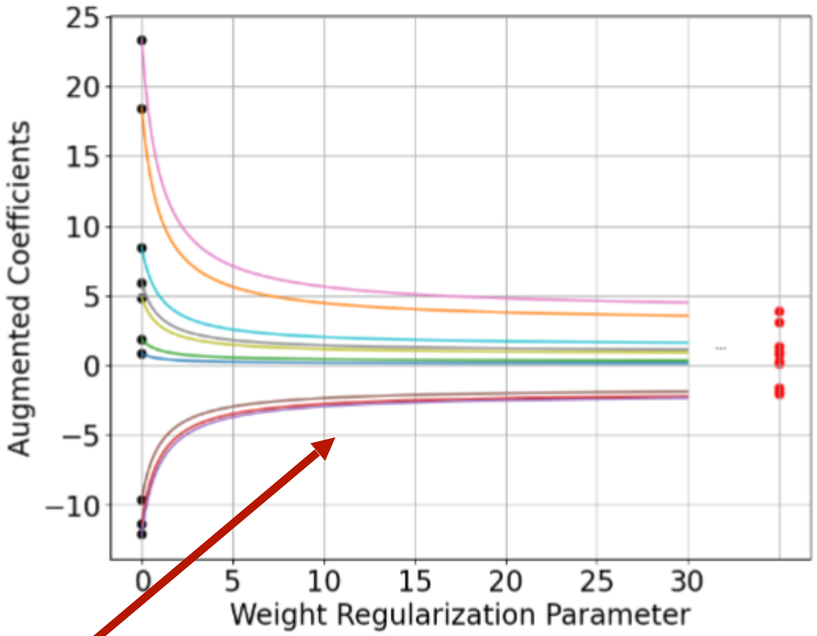
λ

$\hat{X}_q := \hat{w}_\delta{}^T X_p$



δ

$\hat{\beta}_{\text{aug}}$



δ

Look the same, except different asymptotes

Ridge + Ridge = Ridge!

$$\hat{\mathbb{E}}[X_q \hat{\beta}_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}(X_p)(y_p - X_p \hat{\beta}_{\text{reg}})] = \hat{\mathbb{E}}[X_q \hat{\beta}_{\text{aug}}]$$

Undersmoothing!

λ δ $\gamma_j := \dfrac{\delta\lambda}{\sigma_j^2 + \lambda + \delta}$

$$\hat{\beta}^{\lambda}_{\text{ridge},j} = \left(\frac{\sigma_j^2}{\sigma_j^2 + \lambda}\right) \hat{\beta}_{\text{ols},j} \quad + \quad \hat{w}^{\delta}_{\ell_2} \quad \Rightarrow \quad \hat{\beta}_{\text{aug},j} = \left(\frac{\sigma_j^2}{\sigma_j^2 + \gamma_j}\right) \hat{\beta}_{\text{ols},j}$$

Ridge Regression

Ridge Weights

(Generalized)
Ridge Regression

Undersmoothed ridge regression is doubly robust

$$\hat{\mathbb{E}}[X_q\hat{\beta}_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}(X_p)(y_p - X_p\hat{\beta}_{\text{reg}})] = \hat{\mathbb{E}}[X_q\hat{\beta}_{\text{aug}}]$$

$$\mathbb{E}[X_q\hat{\beta}_{aug}] \rightarrow \mathbb{E}_Q[Y] \text{ if}$$

1. $\hat{w}(X_p) \rightarrow \frac{dQ}{dP}(X)$
2. $X_q\hat{\beta}_{reg} \rightarrow \mathbb{E}[Y|X]$

The “dark art” of undersmoothing

Regularized regression

- Typically tuning parameter chosen by minimizing MSE in cross validation
- In high dimensions does not achieve $\sqrt{n}$ rate of convergence

Undersmoothing, aka overfitting

[Newey 1994; Newey et al, 1998; Artefaie et al. 2023; lots of recent work]

- Prioritizes bias over variance; not MSE optimal
- Optimal undersmoothing can achieve $\sqrt{n}$ rates!
- Often, achieving optimal undersmoothing in practice is a **dark art** (-Bruce Hansen)

….. until now! (for kernel ridge)

Kernel ridge + Kernel ridge: Undersmoothing + Asymptotics

Debiased Kernel Methods

Rahul Singh
MIT Economics
rahul.singh@mit.edu

Kernel-based off-policy estimation without overlap: Instance optimality beyond semiparametric efficiency

Wenlong Mou[◇] Peng Ding[†] Martin J. Wainwright[◇,†,‡] Peter L. Bartlett[◇,†,⋆]

Kernel-based covariate functional balancing for observational studies FREE

Raymond K W Wong, Kwun Chuen Gary Chan

Biometrika, Volume 105, Issue 1, March 2018, Pages 199–213, https://doi.org/10.1093/biomet/asx069
Published: 08 December 2017 **Article history** ▾

Minimax Linear Estimation of the Retargeted Mean

David A. Hirshberg* Arian Maleki[†] José R. Zubizarreta[‡]

March 1, 2021

Existing results

Augmented balancing weights
"Double kernel ridge"
[Wong and Chan, 2018; Singh, 2021]

$\hat{\beta}_{\text{reg}}$ $\lambda \asymp n^{-1/2}$

$\hat{w}(X)$ $\delta \asymp n^{-1/2}$

Optimally undersmoothed
Single kernel ridge
[Hirshberg et al., 2019; Mou et al., 2023]

$\gamma \asymp n^{-1}$

Optimally undersmoothed
ℓ_2 balancing
[Wong and Chan, 2018]

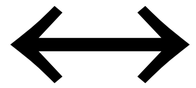
$\gamma \asymp n^{-1}$

New results

$\hat{\beta}_{\text{reg}}$ $\lambda \asymp n^{-1/2}$

$\hat{w}(X)$ $\delta \asymp n^{-1/2}$

equivalent
in finite samples



$\hat{\beta}_{\text{aug}}$ $\gamma \asymp n^{-1}$

New results

$\hat{\beta}_{\text{reg}}$ $\quad \lambda \asymp n^{-1/2}$

$\hat{w}(X)$ $\quad \delta \asymp n^{-1/2}$

equivalent
in finite samples

$\leftrightarrow$

$\hat{\beta}_{\text{aug}}$ $\quad \gamma \asymp n^{-1}$

$\Longrightarrow$ New practical guidance:

Cross-validate $\hat{\beta}_{\text{reg}}$ to select λ

$\delta = \lambda$

$\gamma = \lambda^2$

New results

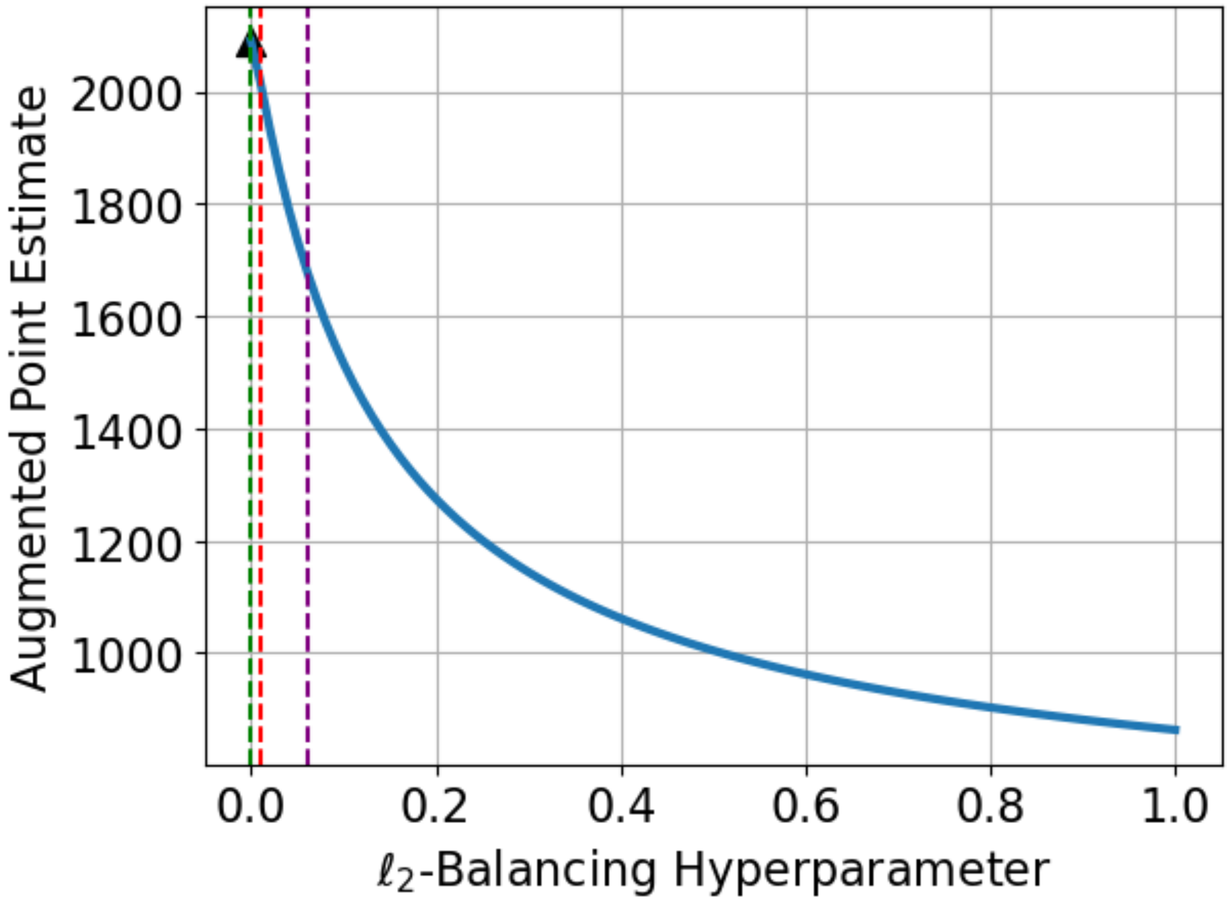
In simulations we find that cross validating the Riesz loss often collapses to OLS, while this method performs surprisingly well!

$\Longrightarrow$ New practical guidance:

Cross-validate $\hat{\beta}_{\text{reg}}$ to select λ $\rightarrow$ $\delta = \lambda$

$\rightarrow$ $\gamma = \lambda^2$

Lalonde



The point estimate for the augmented estimator across the weighting hyperparameter δ; the black triangle corresponds to the OLS point estimate, the green dotted line corresponds to cross-validated balance, the red dotted line corresponds to cross-validated Riesz loss, and the purple dotted line corresponds to the ridge outcome hyperparameter.

New results

New hyperparameter regimes for infinite dimensional RKHSs

$$\gamma \in (n^{-2}, n^{-2/3})$$

(depending on smoothness and effective dimension)

New results

Finite sample bias and variance of double kernel ridge:

$$B_q^2(\lambda,\delta) = \beta_0^T(\hat{\Sigma}+\Gamma_{\lambda,\delta})^{-1}\Gamma_{\lambda,\delta}\mathbb{E}[\Phi_q]^T\mathbb{E}[\Phi_q]\Gamma_{\lambda,\delta}(\hat{\Sigma}+\Gamma_{\lambda,\delta})^{-1}\beta_0$$

$$V_q(\lambda,\delta) = \frac{\sigma^2}{n} tr\left[\hat{\Sigma}(\hat{\Sigma}+\Gamma_{\lambda,\delta})^{-1}\mathbb{E}[\Phi_q]^T\mathbb{E}[\Phi_q](\hat{\Sigma}+\Gamma_{\lambda,\delta})^{-1}\right].$$

Dobriban and Wager, 2018

Ridge + Ridge: Summary

$$\hat{\mathbb{E}}[X_q \hat{\beta}_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}(X_p)(y_p - X_p \hat{\beta}_{\text{reg}})] = \hat{\mathbb{E}}[X_q \hat{\beta}_{\text{aug}}]$$

$$\hat{\beta}_{\text{aug}} = a \circ \hat{\beta}_{\text{ols}} + (1 - a) \circ \hat{\beta}_{\text{reg}}$$

$\hat{w}(X)$ ridge + $\hat{\beta}_{\text{reg}}$ ridge ⇒ $\hat{\beta}_{\text{aug}}$ undersmoothed ridge

Wrapping Up

Augmented balancing weights as linear regression

Our paper: Under linearity, equivalent to single ("undersmoothed") regression

- Augmenting with balancing weights → shifts model toward OLS, sometimes all the way
- Ridge outcome + ridge weighting → single, **undersmoothed** ridge regression
- Lasso outcome + lasso weighting → **double selection**
- Optimally undersmoothed kernel ridge regression
- Non-linear links, general linear functionals

Many extensions in progress:

- Practical recommendations; hyperparameter tuning; inference …

Some follow-up projects:

- **Synthetic controls**: simplex constraints, comparison with Local Projections
- **Multicalibration**: optimality vs robustness in algorithmic fairness
- **Proximal causal inference**: extending results to Fredholm integral equations

Thank you!