Ridge Boosting is Both Robust and Efficient

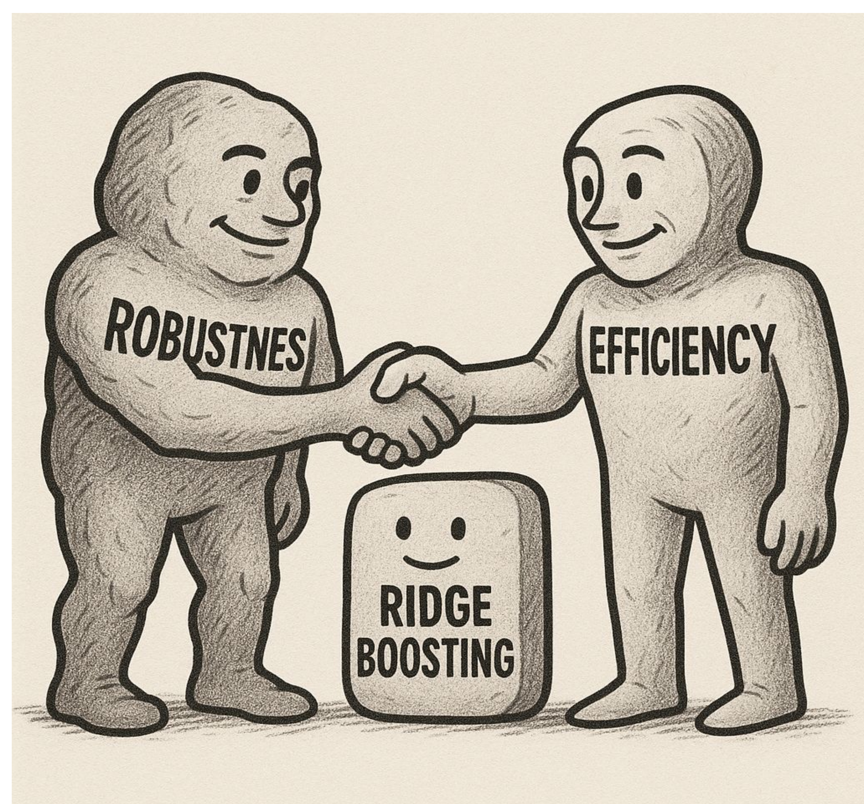
David Bruns-Smith (Stanford) Zhongming Xie (UC Berkeley)· Avi Feller (UC Berkeley)

Train: $X, Y \sim P$ → Covariate Shift → Test: $X, Y \sim Q$

Estimand $\mathbb{E}_Q[Y]$
- Causal Inference
- Subpopulations
- Offline RL

Robustness: Works well for any targets Q (e.g, Multi-accuracy).

Efficiency: Optimal for a particular Q (e.g, Double robustness).



Multi-accuracy (Developed in Fair ML, Uniformly low bias)

A predictor $\hat{f}_{\mathrm{M}}(X)$ is multiaccurate if:

$$\sup_{c \in \mathcal{C}} \left| \mathbb{E}_P[\ c(X)(Y - \hat{f}_{\mathrm{M}}(X))\ ] \right| \leq \alpha$$

$$\mathcal{C} := \left\{ c_1 = \frac{dQ_1}{dP},\ c_2 = \frac{dQ_2}{dP},\ c_3 = \frac{dQ_3}{dP} \right\}$$

→ **Universal Adaptability**

$$\mathbb{E}_{Q_1}[\hat{f}_{\mathrm{M}}(X)] \approx \mathbb{E}_{Q_1}[Y]$$
$$\mathbb{E}_{Q_2}[\hat{f}_{\mathrm{M}}(X)] \approx \mathbb{E}_{Q_2}[Y]$$
$$\mathbb{E}_{Q_3}[\hat{f}_{\mathrm{M}}(X)] \approx \mathbb{E}_{Q_3}[Y]$$

One way to get a multi-accurate estimator (ridge boosting):

Source Population: P — Fit a predictor → $\hat{f}(X)$ → Audit — Perform boosting with ridge → $\hat{f}_{\mathrm{M}}(X)$ → Target population Q — Plug in on target

This simple predictor is not just robust (Multi-accurate) but also optimal across many Q's.

$\hat{f}_{\mathrm{M}}(X)$ →

$$\hat{\mathbb{E}}[\hat{f}_{\mathrm{M}}(X_{q1})] = \hat{\mathbb{E}}[\hat{f}(X_{q1})] + \hat{\mathbb{E}}\left[\frac{d\hat{Q}_1}{dP}(X_p)(y_p - \hat{f}(X_p))\right]$$

$$\hat{\mathbb{E}}[\hat{f}_{\mathrm{M}}(X_{q2})] = \hat{\mathbb{E}}[\hat{f}(X_{q2})] + \hat{\mathbb{E}}\left[\frac{d\hat{Q}_2}{dP}(X_p)(y_p - \hat{f}(X_p))\right]$$

$$\hat{\mathbb{E}}[\hat{f}_{\mathrm{M}}(X_{q3})] = \hat{\mathbb{E}}[\hat{f}(X_{q3})] + \hat{\mathbb{E}}\left[\frac{d\hat{Q}_3}{dP}(X_p)(y_p - \hat{f}(X_p))\right]$$

We got this without the need of data from target population for training.

And we achieve both robustness and efficiency.

Auditing over a Hilbert space

Hilbert Space

$x \in \mathcal{H}$ $\quad \|x\|_{\mathcal{H}}^2 = x^T x$

$$\mathcal{C} = \{c(x) = \beta^T x : \|\beta\|_{\mathcal{H}} \leq B\}$$

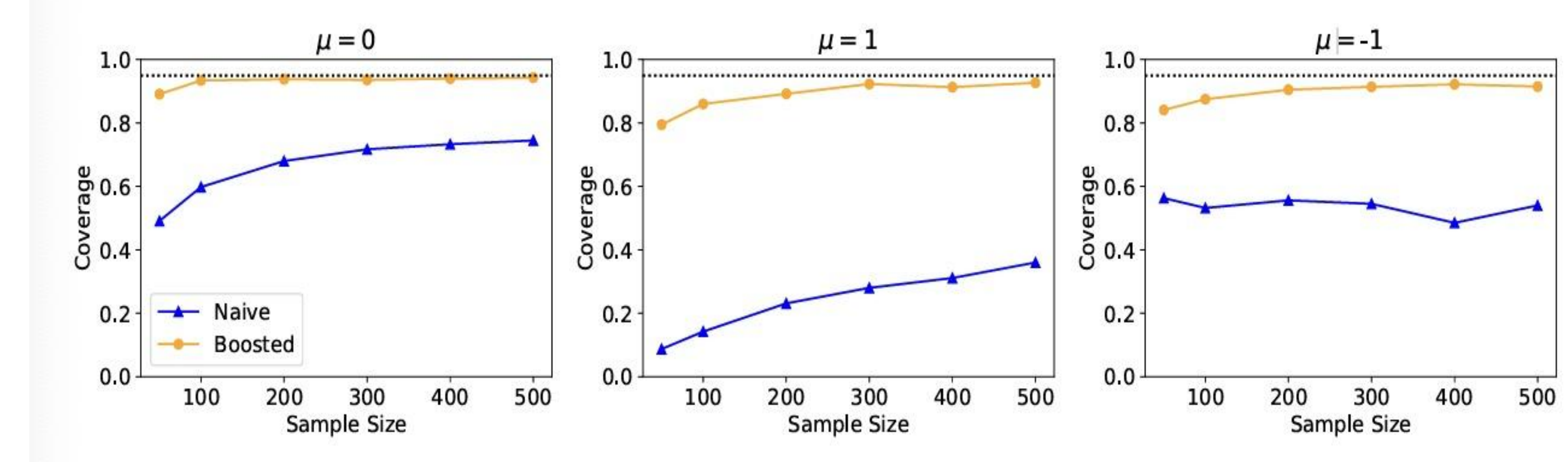
Ridge boosting

$$\min_{\beta \in \mathcal{H}} \{\|y_p - X_p\beta\|_2^2 + \lambda\|\beta\|_{\mathcal{H}}^2\}$$

Can be infinite dimensional

Could be random forests (with data splitting), last-layer embedding of LLM

This predictor is also doubly robust (achieve variance lower bound) across many Q's.



Why?: It implicitly estimate the density ratio (riesz representer)

We explicitly do ridge boosting

$$\hat{\beta} := \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \left\{ \|y_p - \hat{f}(X_p) - X_p\beta\|_2^2 + \lambda\|\beta\|_2^2 \right\}$$

$$\mathcal{C} = \{c \in \mathcal{H}, \|c\|_{\mathcal{H}} \leq B\}$$

And it implicitly estimate the density ratio for all Q:

$$\min_{w \in \mathcal{H}} \left\{ \mathbb{E}_P\left[\left(\frac{dQ}{dP}(X) - w(X)\right)^2\right] + \lambda\|w\|_{\mathcal{H}}^2 \right\}$$

See our paper for more results: