Residuals-Based Distributionally Robust Optimization with Covariate Information

arXiv:2012.01088

Rohit Kannan

Center for Nonlinear Studies Postdoctoral Fellow
Los Alamos National Laboratory

Robust Optimization Webinar on April 8, 2021

Joint work with Güzin Bayraksan and Jim Luedtke

Funding: DOE (MACSER Project)

Outline

Traditional Data-Driven Stochastic Programming

- Traditional SP: minimize expected system cost assuming feasible region $\mathcal{Z}$ and distribution of Y known

$$\min_{z \in \mathcal{Z}} \mathbb{E}_Y[c(z, Y)]$$

Traditional Data-Driven Stochastic Programming

- Traditional SP: minimize expected system cost assuming feasible region $\mathcal{Z}$ and distribution of Y known

$$\min_{z \in \mathcal{Z}} \mathbb{E}_Y[c(z, Y)]$$

- Data-driven SP: have access to samples $\{y^i\}_{i=1}^n$ of Y

$$\min_{z \in \mathcal{Z}} \mathbb{E}_Y[c(z, Y)] \approx \min_{z \in \mathcal{Z}} \frac{1}{n} \sum_{i=1}^{n} c(z, y^i) \qquad \text{(SAA)}$$

- Sample Average Approximation theory: as sample size $n \to \infty$, optimal value and solutions of (SAA) converge to those of true SP at rate $O_p(n^{-1/2})$

Example: Mean-Risk Portfolio Optimization

$$\min_{z \in \mathcal{Z}} \ \mathbb{E}_Y[-Y^\mathsf{T} z] + \rho \, \mathrm{CVaR}_\beta(-Y^\mathsf{T} z),$$

where $\mathcal{Z} := \left\{ z \in \mathbb{R}^{d_z}_+ : \sum_i z_i = 1 \right\}$.

- z_i: fraction of capital invested in asset i
- Y_i: uncertain net return of asset i
- $\mathrm{CVaR}_\beta \approx$ average of the $100(1-\beta)\%$ worst return outcomes
- $\rho \geq 0$ and $\beta \in [0, 1)$: risk parameters (e.g., $\rho = 10$, $\beta = 0.8$)

Stochastic Programming with Covariate Information

- Use covariates X to inform distribution of random vector Y
  - Covariates also called *features* or *side information*
- When making decision z, we observe a *new* covariate $X = x$
- Goal: solve the conditional SP

$$\min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z, Y) \mid X = x\right]$$

Example Application: Power Grid Scheduling

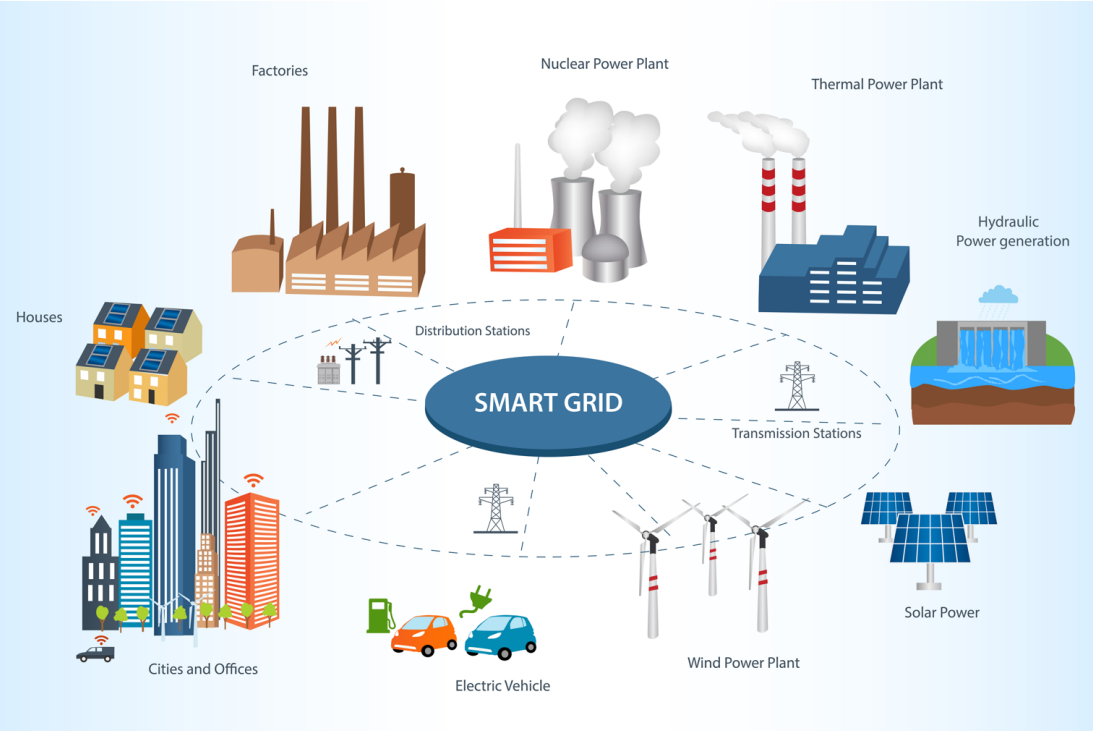


Image credit: IEEE Innovation at Work

- Decisions z: Generator schedules
- Uncertain Parameters Y: Load, Renewable energy outputs
- Covariates X: Weather observations, Time of day/Season

Example Application: Production Planning

Image credit: AIDIAONE

- Decisions z: Production and Inventory levels
- Uncertain Parameters Y: Product demands
- Covariates X: Seasonality, Web search results

Stochastic Programming with Covariate Information

Enter Machine Learning

- When making decision z, we observe a *new* covariate $X = x$
- Goal: solve the conditional SP

$$\min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z, Y) \mid X = x\right]$$

Stochastic Programming with Covariate Information

Enter Machine Learning

- When making decision z, we observe a *new* covariate $X = x$
- Goal: solve the conditional SP

$$\min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z, Y) \mid X = x\right]$$

- Assume we have uncertain parameter and covariate data pairs (not necessarily i.i.d.)

$$\mathcal{D}_n := \{(y^i, x^i)\}_{i=1}^n$$

- How to construct data-driven approximation to conditional SP?

Stochastic Programming with Covariate Information

Enter Machine Learning

- When making decision z, we observe a *new* covariate $X = x$
- Goal: solve the conditional SP

$$\min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z, Y) \mid X = x\right]$$

- Assume we have uncertain parameter and covariate data pairs (not necessarily i.i.d.)

$$\mathcal{D}_n := \{(y^i, x^i)\}_{i=1}^n$$

- How to construct data-driven approximation to conditional SP?
  1. Learn: predict Y given $X = x$
  2. Optimize: integrate learning into optimization (with errors)

Traditional Integrated Learning and Optimization

1. Use data to train your favorite ML prediction model:

$$\hat{f}_n(\cdot) \in \arg\min_{f(\cdot)\in\mathcal{F}} \sum_{i=1}^{n} \ell(f(x^i), y^i) + \rho(f)$$

2. Given observed covariate $X = x$, use point prediction within deterministic optimization model

$$\min_{z\in\mathcal{Z}} c(z, \hat{f}_n(x))$$

Traditional Integrated Learning and Optimization

1. Use data to train your favorite ML prediction model:

$$\hat{f}_n(\cdot) \in \arg\min_{f(\cdot)\in\mathcal{F}} \sum_{i=1}^{n} \ell(f(x^i), y^i) + \rho(f)$$

2. Given observed covariate $X = x$, use point prediction within deterministic optimization model

$$\min_{z\in\mathcal{Z}} c(z, \hat{f}_n(x))$$

- Modular: separate learning and optimization steps
- Expect to work well if (and likely only if) prediction is accurate

Improved Integrated Learning and Optimization

Approach 1: Modify the learning step[1]

- Change loss function in ML training step to reflect use of prediction within optimization model
- More challenging training problem + less modular

[1] Kao et al. (2009); Donti et al. (2017); Elmachtoub and Grigas (2017); . . .

[2] Ban et al. (2018); Bertsimas and Kallus (2020); Sen and Deng (2018); Bertsimas et al. (2019); Esteban-Pérez and Morales (2020); . . .

[3] Bertsimas and Kallus (2020); Ban and Rudin (2018); Bazier-Matte and Delage (2020); . . .

Improved Integrated Learning and Optimization

Approach 1: Modify the learning step[1]

- Change loss function in ML training step to reflect use of prediction within optimization model
- More challenging training problem + less modular

Approach 2 (this work): Modify the optimization step[2]

- Change optimization model to reflect uncertainty in prediction

[1] Kao et al. (2009); Donti et al. (2017); Elmachtoub and Grigas (2017); . . .

[2] Ban et al. (2018); Bertsimas and Kallus (2020); Sen and Deng (2018); Bertsimas et al. (2019); Esteban-Pérez and Morales (2020); . . .

[3] Bertsimas and Kallus (2020); Ban and Rudin (2018); Bazier-Matte and Delage (2020); . . .

Improved Integrated Learning and Optimization

Approach 1: Modify the learning step[1]

- Change loss function in ML training step to reflect use of prediction within optimization model
- More challenging training problem + less modular

Approach 2 (this work): Modify the optimization step[2]

- Change optimization model to reflect uncertainty in prediction

Approach 3: Direct solution learning[3]

- Attempt to directly learn a mapping from x to a solution z
- Handling constraints and large dimensions of z is challenging

[1] Kao et al. (2009); Donti et al. (2017); Elmachtoub and Grigas (2017); . . .

[2] Ban et al. (2018); Bertsimas and Kallus (2020); Sen and Deng (2018); Bertsimas et al. (2019); Esteban-Pérez and Morales (2020); . . .

[3] Bertsimas and Kallus (2020); Ban and Rudin (2018); Bazier-Matte and Delage (2020); . . .

Outline

Problem Setup

Given

- Joint observations $\mathcal{D}_n := \{(y^i, x^i)\}_{i=1}^n$ of random vectors Y, X
- New random covariate observation $X = x$ (current context)

Want to solve

$$v^*(x) := \min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z, Y) \mid X = x\right]$$

Problem Setup

Given

- Joint observations $\mathcal{D}_n := \{(y^i, x^i)\}_{i=1}^n$ of random vectors Y, X
- New random covariate observation $X = x$ (current context)

Want to solve

$$v^*(x) := \min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z, Y) \mid X = x\right]$$

Assume (for now)

- True model: $Y = f^*(X) + \varepsilon$ with X and ε *independent*

$$\implies v^*(x) \equiv \min_{z \in \mathcal{Z}} \mathbb{E}_\varepsilon[c(z, f^*(x) + \varepsilon)]$$

- We know a function class $\mathcal{F}$ such that $f^* \in \mathcal{F}$

Empirical Residuals-based Sample Average Approximation

Approach suggested in Sen and Deng (2018); analyzed for a specific application in Ban et al. (2018)

1. Estimate f^* using your favorite ML method $\Rightarrow \hat{f}_n$

K., Bayraksan, and Luedtke. Data-driven sample average approximation with covariate information. Submitted

Empirical Residuals-based Sample Average Approximation

Approach suggested in Sen and Deng (2018); analyzed for a specific application in Ban et al. (2018)

1. Estimate f^* using your favorite ML method $\Rightarrow \hat{f}_n$

   Compute *empirical residuals* $\hat{\varepsilon}_n^i := y^i - \hat{f}_n(x^i)$, $i \in [n]$

K., Bayraksan, and Luedtke. Data-driven sample average approximation with covariate information. Submitted

Empirical Residuals-based Sample Average Approximation

Approach suggested in Sen and Deng (2018); analyzed for a specific application in Ban et al. (2018)

1. Estimate f^* using your favorite ML method $\Rightarrow \hat{f}_n$

   Compute *empirical residuals* $\hat{\varepsilon}_n^i := y^i - \hat{f}_n(x^i)$, $i \in [n]$

2. Use $\{\hat{f}_n(x) + \hat{\varepsilon}_n^i\}_{i=1}^n$ as proxy for samples of Y given $X = x$

$$\hat{z}_n^{ER}(x) \in \arg\min_{z \in \mathcal{Z}} \frac{1}{n} \sum_{i=1}^{n} c(z, \hat{f}_n(x) + \hat{\varepsilon}_n^i) \qquad \text{(ER-SAA)}$$

K., Bayraksan, and Luedtke. Data-driven sample average approximation with covariate information. Submitted

Empirical Residuals-based Sample Average Approximation

Approach suggested in Sen and Deng (2018); analyzed for a specific application in Ban et al. (2018)

1. Estimate f^* using your favorite ML method $\Rightarrow \hat{f}_n$

   Compute *empirical residuals* $\hat{\varepsilon}_n^i := y^i - \hat{f}_n(x^i)$, $i \in [n]$

2. Use $\{\hat{f}_n(x) + \hat{\varepsilon}_n^i\}_{i=1}^n$ as proxy for samples of Y given $X = x$

$$\hat{z}_n^{ER}(x) \in \arg\min_{z \in \mathcal{Z}} \frac{1}{n} \sum_{i=1}^{n} c(z, \hat{f}_n(x) + \hat{\varepsilon}_n^i) \qquad \text{(ER-SAA)}$$

- Modular like traditional approach
- Our contribution: general convergence analysis

K., Bayraksan, and Luedtke. Data-driven sample average approximation with covariate information. Submitted

Empirical Residuals-based Sample Average Approximation

Approach suggested in Sen and Deng (2018); analyzed for a specific application in Ban et al. (2018)

1. Estimate f^* using your favorite ML method $\Rightarrow \hat{f}_n$

   Compute *empirical residuals* $\hat{\varepsilon}_n^i := y^i - \hat{f}_n(x^i)$, $i \in [n]$

2. Use $\{\hat{f}_n(x) + \hat{\varepsilon}_n^i\}_{i=1}^n$ as proxy for samples of Y given $X = x$

$$\hat{z}_n^{ER}(x) \in \arg\min_{z \in \mathcal{Z}} \frac{1}{n} \sum_{i=1}^{n} c(z, \hat{f}_n(x) + \hat{\varepsilon}_n^i) \qquad \text{(ER-SAA)}$$

- Modular like traditional approach
- Our contribution: general convergence analysis
- Improvements when sample size is small?

K., Bayraksan, and Luedtke. Data-driven sample average approximation with covariate information. Submitted

New Small Sample Variant of ER-SAA

Mitigate effects of overfitting by using *leave-one-out residuals*

1. Estimate f^* separately with each data point i left out (leave-one-out regression) $\Rightarrow \hat{f}_{-i}(\cdot)$ for $i \in [n]$

K., Bayraksan, and Luedtke. Data-driven sample average approximation with covariate information. Submitted

New Small Sample Variant of ER-SAA

Mitigate effects of overfitting by using *leave-one-out residuals*

1. Estimate f^* separately with each data point i left out (leave-one-out regression) $\Rightarrow \hat{f}_{-i}(\cdot)$ for $i \in [n]$

   Compute leave-one-out residuals $\hat{\varepsilon}_n^i := y^i - \hat{f}_{-i}(x^i)$, $i \in [n]$

K., Bayraksan, and Luedtke. Data-driven sample average approximation with covariate information. Submitted

New Small Sample Variant of ER-SAA

Mitigate effects of overfitting by using *leave-one-out residuals*

1. Estimate f^* separately with each data point i left out (leave-one-out regression) $\Rightarrow \hat{f}_{-i}(\cdot)$ for $i \in [n]$

   Compute leave-one-out residuals $\hat{\varepsilon}_n^i := y^i - \hat{f}_{-i}(x^i)$, $i \in [n]$

2. Use $\{\hat{f}_n(x) + \hat{\varepsilon}_n^i\}_{i=1}^n$ or $\{\hat{f}_{-i}(x) + \hat{\varepsilon}_n^i\}_{i=1}^n$ as proxy for samples of Y given $X = x$

$$\hat{z}_n^J(x) \in \arg\min_{z \in \mathcal{Z}} \frac{1}{n} \sum_{i=1}^n c(z, \hat{f}_n(x) + \hat{\varepsilon}_n^i) \qquad \text{(J-SAA)}$$

Inspired by Jackknife methods (Barber et al., 2019)

K., Bayraksan, and Luedtke. Data-driven sample average approximation with covariate information. Submitted

New Small Sample Variant of ER-SAA

Mitigate effects of overfitting by using *leave-one-out residuals*

1. Estimate f^* separately with each data point i left out (leave-one-out regression) $\Rightarrow \hat{f}_{-i}(\cdot)$ for $i \in [n]$

   Compute leave-one-out residuals $\hat{\varepsilon}_n^i := y^i - \hat{f}_{-i}(x^i)$, $i \in [n]$

2. Use $\{\hat{f}_n(x) + \hat{\varepsilon}_n^i\}_{i=1}^n$ or $\{\hat{f}_{-i}(x) + \hat{\varepsilon}_n^i\}_{i=1}^n$ as proxy for samples of Y given $X = x$

$$\hat{z}_n^J(x) \in \arg\min_{z \in \mathcal{Z}} \frac{1}{n} \sum_{i=1}^n c(z, \hat{f}_n(x) + \hat{\varepsilon}_n^i) \qquad \text{(J-SAA)}$$

Inspired by Jackknife methods (Barber et al., 2019)

This talk: DRO formulation around (ER-SAA) as alternative to (J-SAA)

K., Bayraksan, and Luedtke. Data-driven sample average approximation with covariate information. Submitted

A Sampling of ER-SAA Theory

- $v^*(x) = \min_{z \in \mathcal{Z}} \mathbb{E}_\varepsilon[c(z, f^*(x) + \varepsilon)]$
  $= \text{optimal value of true conditional SP}$
- $\hat{z}_n^{ER}(x) = \text{ER-SAA solution}$

Asymptotic optimality: the out-of-sample cost of data-driven solutions approaches the optimal value of the true conditional SP as the sample size increases

$$\mathbb{E}_\varepsilon\left[c(\hat{z}_n^{ER}(x), f^*(x) + \varepsilon)\right] \xrightarrow{p} v^*(x)$$

See http://www.optimization-online.org/DB_HTML/2020/07/7932.html for more theory + numerical experiments

A Sampling of ER-SAA Theory

- $v^*(x) = \min_{z \in \mathcal{Z}} \mathbb{E}_\varepsilon[c(z, f^*(x) + \varepsilon)]$
  $= \text{optimal value of true conditional SP}$
- $\hat{z}_n^{ER}(x) = \text{ER-SAA solution}$

Asymptotic optimality: the out-of-sample cost of data-driven solutions approaches the optimal value of the true conditional SP as the sample size increases

$$\mathbb{E}_\varepsilon\left[c(\hat{z}_n^{ER}(x), f^*(x) + \varepsilon)\right] \xrightarrow{p} v^*(x)$$

Setting: two-stage stochastic Mixed-Integer Linear Programs

$$\min_{z \in \mathcal{Z}} c_z^\mathsf{T} z + \mathbb{E}\left[Q(z, Y) \mid X = x\right],$$

$$\text{where} \quad Q(z, Y) := \min_{v \in \mathbb{R}_+^{d_v}} \left\{q_v^\mathsf{T} v : Wv = h(Y) - T(Y)z\right\}$$

See http://www.optimization-online.org/DB_HTML/2020/07/7932.html for more theory + numerical experiments

Rate of Convergence of ER-SAA Solutions

Assumption: There is a constant $r \in (0, 1]$ such that the regression procedure satisfies

- Pointwise error rate: $\|f^*(x) - \hat{f}_n(x)\|^2 = O_p\left(n^{-r}\right)$
- Mean-squared estimation error rate:

$$\frac{1}{n}\sum_{i=1}^{n}\|f^*(x^i) - \hat{f}_n(x^i)\|^2 = O_p(n^{-r})$$

K., Bayraksan, and Luedtke. Data-driven sample average approximation with covariate information. Submitted

Rate of Convergence of ER-SAA Solutions

Assumption: There is a constant $r \in (0, 1]$ such that the regression procedure satisfies

- Pointwise error rate: $\|f^*(x) - \hat{f}_n(x)\|^2 = O_p\left(n^{-r}\right)$
- Mean-squared estimation error rate:

$$\frac{1}{n}\sum_{i=1}^{n}\|f^*(x^i) - \hat{f}_n(x^i)\|^2 = O_p(n^{-r})$$

- OLS regression, Lasso satisfy assumption with $r = 1$
- CART, RF regression satisfy assumption with $r = \frac{O(1)}{\dim(X)}$

K., Bayraksan, and Luedtke. Data-driven sample average approximation with covariate information. Submitted

Rate of Convergence of ER-SAA Solutions

Assumption: There is a constant $r \in (0, 1]$ such that the regression procedure satisfies

- Pointwise error rate: $\|f^*(x) - \hat{f}_n(x)\|^2 = O_p\left(n^{-r}\right)$
- Mean-squared estimation error rate:

$$\frac{1}{n}\sum_{i=1}^{n}\|f^*(x^i) - \hat{f}_n(x^i)\|^2 = O_p(n^{-r})$$

- OLS regression, Lasso satisfy assumption with $r = 1$
- CART, RF regression satisfy assumption with $r = \frac{O(1)}{\dim(X)}$

Informal Theorem (Rate of Convergence)

Under the above assumptions, ER-SAA solution $\hat{z}_n^{ER}(x)$ satisfies

$$\mathbb{E}_{\varepsilon}\left[c(\hat{z}_n^{ER}(x), f^*(x) + \varepsilon)\right] = v^*(x) + O_p(n^{-r/2})$$

K., Bayraksan, and Luedtke. Data-driven sample average approximation with covariate information. Submitted

Outline

Distributionally robust optimization (DRO)

- Minimize worst-case expected cost over a set of distributions

$$\hat{z}_n^{DRO}(x) \in \arg\min_{z \in \mathcal{Z}} \max_{Q \in \hat{\mathcal{P}}_n(x)} \mathbb{E}_{Y \sim Q}[c(z, Y)]$$

$\hat{\mathcal{P}}_n(x) =$ "confidence region" for distribution of Y given $X = x$

K., Bayraksan, and Luedtke. Residuals-based DRO with covariate information. Submitted

Distributionally robust optimization (DRO)

- Minimize worst-case expected cost over a set of distributions

$$\hat{z}_n^{DRO}(x) \in \arg\min_{z \in \mathcal{Z}} \max_{Q \in \hat{\mathcal{P}}_n(x)} \mathbb{E}_{Y \sim Q}[c(z, Y)]$$

$\hat{\mathcal{P}}_n(x) =$ "confidence region" for distribution of Y given $X = x$

- If $\hat{\mathcal{P}}_n(x)$ only comprises the ER-SAA distribution

$$\hat{P}_n^{ER}(x) := \frac{1}{n} \sum_{i=1}^{n} \delta_{\hat{f}_n(x) + \hat{\varepsilon}_n^i},$$

then recover the ER-SAA solution

- Motivation: DRO regularizes small sample ER-SAA, yielding solutions with better out-of-sample performance

K., Bayraksan, and Luedtke. Residuals-based DRO with covariate information. Submitted

Empirical Residuals-based DRO (ER-DRO)

Given ambiguity set $\hat{\mathcal{P}}_n(x)$ centered at $\hat{P}_n^{ER}(x)$, solve

$$\hat{z}_n^{DRO}(x) \in \arg\min_{z \in \mathcal{Z}} \sup_{Q \in \hat{\mathcal{P}}_n(x)} \mathbb{E}_{Y \sim Q}[c(z, Y)]$$

Examples of ambiguity sets $\hat{\mathcal{P}}_n(x)$:

- Wasserstein ambiguity sets of order $p \in [1, +\infty)$:

$$\hat{\mathcal{P}}_n(x) := \Big\{\text{distributions } Q \text{ such that the } p\text{-Wasserstein distance between } Q \text{ and } \hat{P}_n^{ER}(x) \leq \zeta_n(x)\Big\}$$

- Other ambiguity sets based on phi-divergences, sample robust optimization, ...

K., Bayraksan, and Luedtke. Residuals-based DRO with covariate information. Submitted

Towards Convergence Theory for Wasserstein ER-DRO

Assumption: For any risk level $\alpha \in (0,1)$, there exists a constant $\kappa_{p,n}(\alpha, x) > 0$ such that the regression procedure satisfies

$$\mathbb{P}\{\|f^*(x) - \hat{f}_n(x)\|^p > \kappa_{p,n}^p(\alpha, x)\} \leq \alpha, \quad \text{and}$$

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}\|f^*(x^i) - \hat{f}_n(x^i)\|^p > \kappa_{p,n}^p(\alpha, x)\right\} \leq \alpha.$$

Towards Convergence Theory for Wasserstein ER-DRO

Assumption: For any risk level $\alpha \in (0,1)$, there exists a constant $\kappa_{p,n}(\alpha, x) > 0$ such that the regression procedure satisfies

$$\mathbb{P}\{\|f^*(x) - \hat{f}_n(x)\|^p > \kappa_{p,n}^p(\alpha, x)\} \leq \alpha, \quad \text{and}$$

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}\|f^*(x^i) - \hat{f}_n(x^i)\|^p > \kappa_{p,n}^p(\alpha, x)\right\} \leq \alpha.$$

Example: holds for Wasserstein order $p = 2$ and

- OLS, Lasso with $\kappa_{2,n}^2(\alpha, x) = O(n^{-1}\log(\alpha^{-1}))$
- CART, RF with $\kappa_{2,n}^2(\alpha, x) = O(n^{-1}\log(\alpha^{-1}))^{O(1)/d_x}$

Towards Convergence Theory for Wasserstein ER-DRO

Assumption: For any risk level $\alpha \in (0,1)$, there exists a constant $\kappa_{p,n}(\alpha, x) > 0$ such that the regression procedure satisfies

$$\mathbb{P}\{\|f^*(x) - \hat{f}_n(x)\|^p > \kappa_{p,n}^p(\alpha, x)\} \leq \alpha, \quad \text{and}$$

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}\|f^*(x^i) - \hat{f}_n(x^i)\|^p > \kappa_{p,n}^p(\alpha, x)\right\} \leq \alpha.$$

Towards Convergence Theory for Wasserstein ER-DRO

Assumption: For any risk level $\alpha \in (0,1)$, there exists a constant $\kappa_{p,n}(\alpha, x) > 0$ such that the regression procedure satisfies

$$\mathbb{P}\{\|f^*(x) - \hat{f}_n(x)\|^p > \kappa_{p,n}^p(\alpha, x)\} \leq \alpha, \quad \text{and}$$

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}\|f^*(x^i) - \hat{f}_n(x^i)\|^p > \kappa_{p,n}^p(\alpha, x)\right\} \leq \alpha.$$

Given covariate realization x and risk level $\alpha \in (0,1)$, use

$$\zeta_n(\alpha, x) := 2\kappa_{p,n}\left(\tfrac{\alpha}{4}, x\right) + \bar{\kappa}_{p,n}\left(\tfrac{\alpha}{2}\right)$$

as the radius of the Wasserstein ambiguity set, where

$\bar{\kappa}_{p,n}\left(\frac{\alpha}{2}\right) =$ traditional Wasserstein DRO radius that is used if we know f^* (Kuhn et al., 2019)

Towards Convergence Theory for Wasserstein ER-DRO

Assumption: For any risk level $\alpha \in (0,1)$, there exists a constant $\kappa_{p,n}(\alpha, x) > 0$ such that the regression procedure satisfies

$$\mathbb{P}\{\|f^*(x) - \hat{f}_n(x)\|^p > \kappa_{p,n}^p(\alpha, x)\} \leq \alpha, \quad \text{and}$$

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}\|f^*(x^i) - \hat{f}_n(x^i)\|^p > \kappa_{p,n}^p(\alpha, x)\right\} \leq \alpha.$$

Given covariate realization x and risk level $\alpha \in (0,1)$, use

$$\zeta_n(\alpha, x) := 2\kappa_{p,n}\left(\tfrac{\alpha}{4}, x\right) + \bar{\kappa}_{p,n}\left(\tfrac{\alpha}{2}\right)$$

as the radius of the Wasserstein ambiguity set, where

$\bar{\kappa}_{p,n}\left(\frac{\alpha}{2}\right) =$ traditional Wasserstein DRO radius that is used if we know f^* (Kuhn et al., 2019)

Radius guarantees that $\mathbb{P}\{d_W(\hat{P}_n^{ER}(x), P_{Y|X=x}) > \zeta_n(\alpha, x)\} \leq \alpha$

Flavor of Wasserstein ER-DRO Results

Informal Theorem (Finite Sample Certificate Guarantee)

For the above choice of the Wasserstein radius $\zeta_n(\alpha, x)$, the solution $\hat{z}_n^{DRO}(x)$ and the optimal value $\hat{v}_n^{DRO}(x)$ satisfy

$$\mathbb{P}\left\{\mathbb{E}_\varepsilon\left[c(\hat{z}_n^{DRO}(x), f^*(x) + \varepsilon)\right] \leq \hat{v}_n^{DRO}(x)\right\} \geq 1 - \alpha$$

Flavor of Wasserstein ER-DRO Results

Informal Theorem (Finite Sample Certificate Guarantee)

For the above choice of the Wasserstein radius $\zeta_n(\alpha, x)$, the solution $\hat{z}_n^{DRO}(x)$ and the optimal value $\hat{v}_n^{DRO}(x)$ satisfy

$$\mathbb{P}\left\{\mathbb{E}_\varepsilon\left[c(\hat{z}_n^{DRO}(x), f^*(x)+\varepsilon)\right] \leq \hat{v}_n^{DRO}(x)\right\} \geq 1-\alpha$$

Informal Theorem (Rate of Convergence)

Suppose there is a sequence of risk levels $\{\alpha_n\} \subset (0,1)$ such that $\sum_n \alpha_n < +\infty$ and the radius satisfies $\lim_{n\to\infty} \zeta_n(\alpha_n, x) = 0$. Then the sequence $\{\hat{z}_n^{DRO}(x)\}$ of solutions satisfies

$$\mathbb{E}_\varepsilon\left[c(\hat{z}_n^{DRO}(x), f^*(x)+\varepsilon)\right] = v^*(x) + O_p(\zeta_n(\alpha_n, x))$$

Choosing the Wasserstein Radius in Practice

- Theoretical Wasserstein radius: involves unknown constants and is typically conservative
- Use cross-validation to specify the radius $\zeta_n(x)$
  - Approach 1: Ignore covariate information altogether while choosing ζ_n
  - Approach 2: Use the data $\mathcal{D}_n$ to choose ζ_n independently of the covariate realization $X = x$
  - Approach 3: Use both the data $\mathcal{D}_n$ and the covariate realization $X = x$ to choose the radius $\zeta_n(x)$
- Approach 3 is more data intensive than Approaches 1 & 2

Numerical Study: Mean-Risk Portfolio Optimization

- Consider instance with 10 assets
- Uncertain returns Y generated according to

$$Y_j = \nu_j^* + \sum_{l=1}^{3} \mu_{jl}^* (X_l)^{\theta} + \bar{\varepsilon}_j + \omega, \quad \forall j \in \{1, \ldots, 10\},$$

where $\bar{\varepsilon}_j \sim \mathcal{N}(0, 0.02j)$, $\omega \sim \mathcal{N}(0, 0.02)$, $\theta \in \{0.5, 1, 2\}$, $\dim(X) \in \{10, 100\}$

Numerical Study: Mean-Risk Portfolio Optimization

- Consider instance with 10 assets
- Uncertain returns Y generated according to

$$Y_j = \nu_j^* + \sum_{l=1}^{3} \mu_{jl}^* (X_l)^{\theta} + \bar{\varepsilon}_j + \omega, \quad \forall j \in \{1, \ldots, 10\},$$

where $\bar{\varepsilon}_j \sim \mathcal{N}(0, 0.02j)$, $\omega \sim \mathcal{N}(0, 0.02)$, $\theta \in \{0.5, 1, 2\}$, $\dim(X) \in \{10, 100\}$

- Fit linear model with OLS/Lasso regression (even when $\theta \neq 1$)

$$Y_j = \nu_j + \sum_{l=1}^{\dim(X)} \mu_{jl} X_l + \eta_j, \quad \forall j \in \{1, \ldots, 10\},$$

where η_j are zero-mean errors

- Estimate optimality gap of solutions $\hat{z}_n^{ER}(x)$ and $\hat{z}_n^{DRO}(x)$

Results with OLS and Correct Model Class ($\theta = 1$)

I^*: Ideal Wasserstein radius (only for benchmarking)

1 & **2**: Wasserstein radius specified using Approaches 1 & 2

E: ER-SAA + OLS

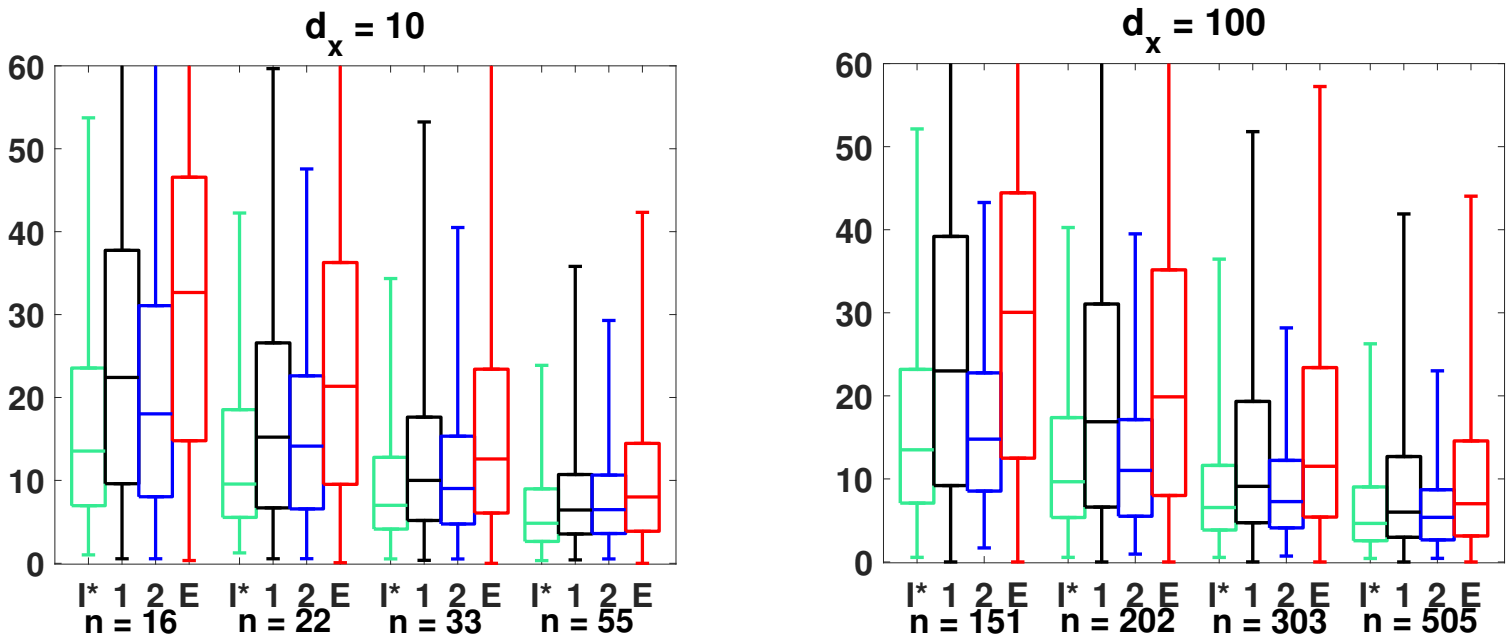
Results with OLS and Correct Model Class ($\theta = 1$)

I^*: Ideal Wasserstein radius (only for benchmarking)

1 & **2**: Wasserstein radius specified using Approaches 1 & 2

E: ER-SAA + OLS

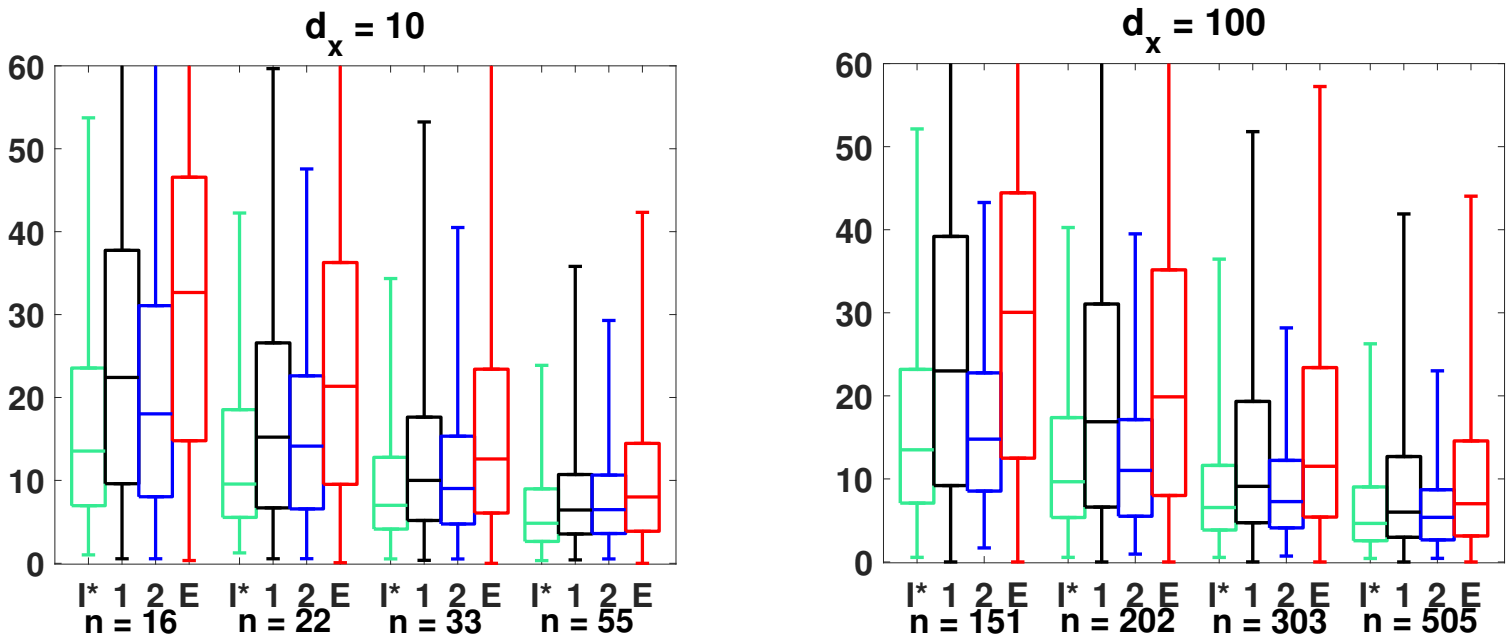
Lower y-axis value $\implies$ closer to optimal



Results with OLS and Correct Model Class ($\theta = 1$)

I^*: Ideal Wasserstein radius (only for benchmarking)
1 & **2**: Wasserstein radius specified using Approaches 1 & 2
E: ER-SAA + OLS

Lower y-axis value $\implies$ closer to optimal

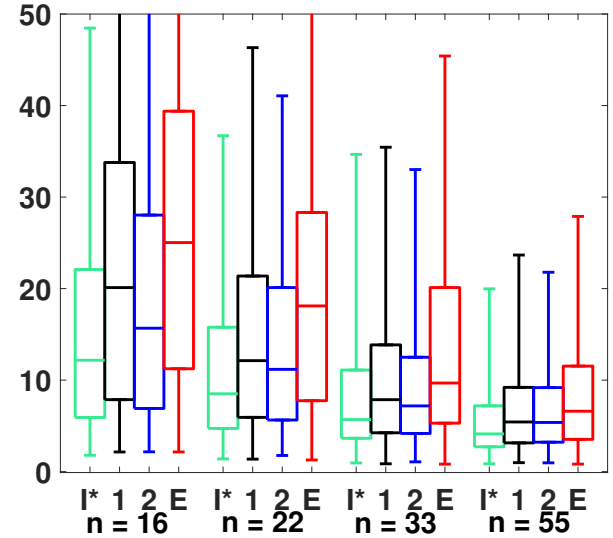


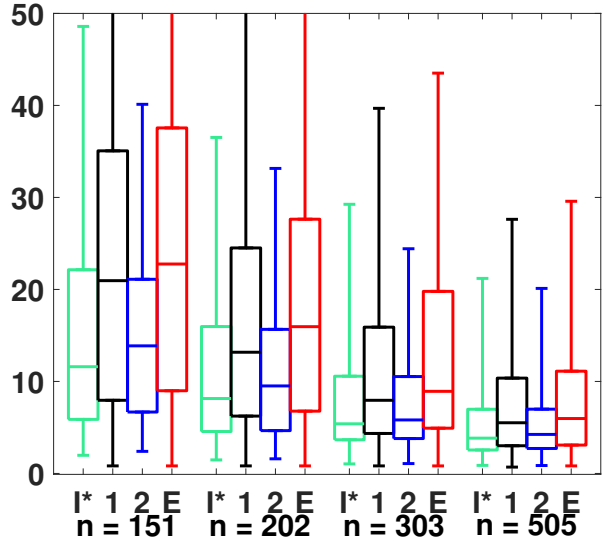
Boxes: 25, 50, and 75 percentiles of upper confidence bounds
Whiskers: 2 and 98 percentiles
Sample sizes: $\{1.5,\ 2,\ 3,\ 5\} \times (\dim(X) + 1)$

Results with OLS and Misspecified Model Class ($\theta \neq 1$)

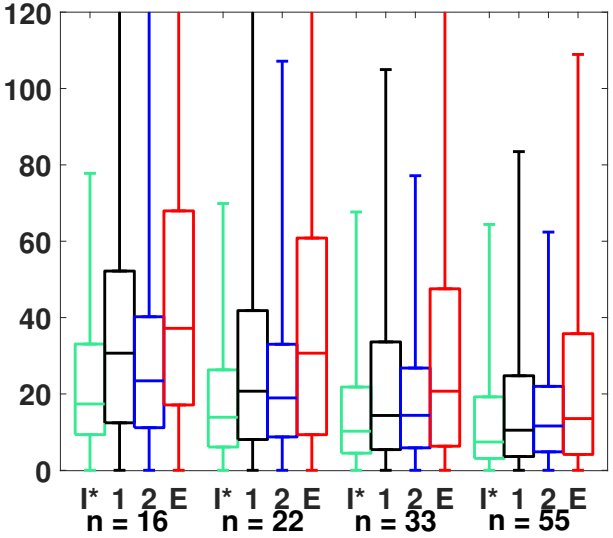
$\mathbf{d_x = 10}$ $\mathbf{d_x = 100}$

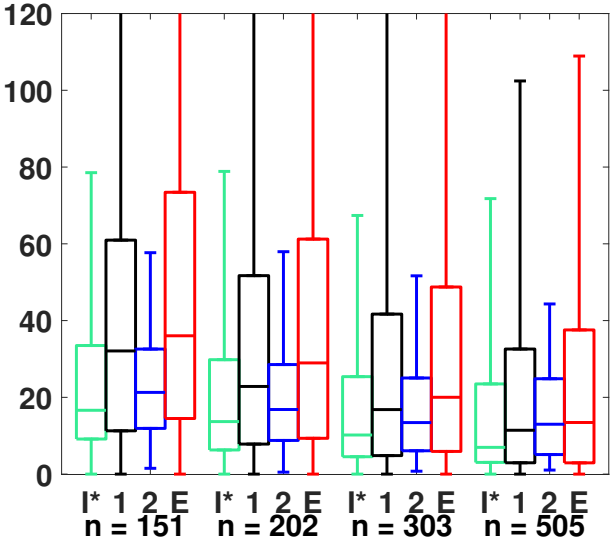
$\boldsymbol{\theta = 0.5}$





$\boldsymbol{\theta = 2}$





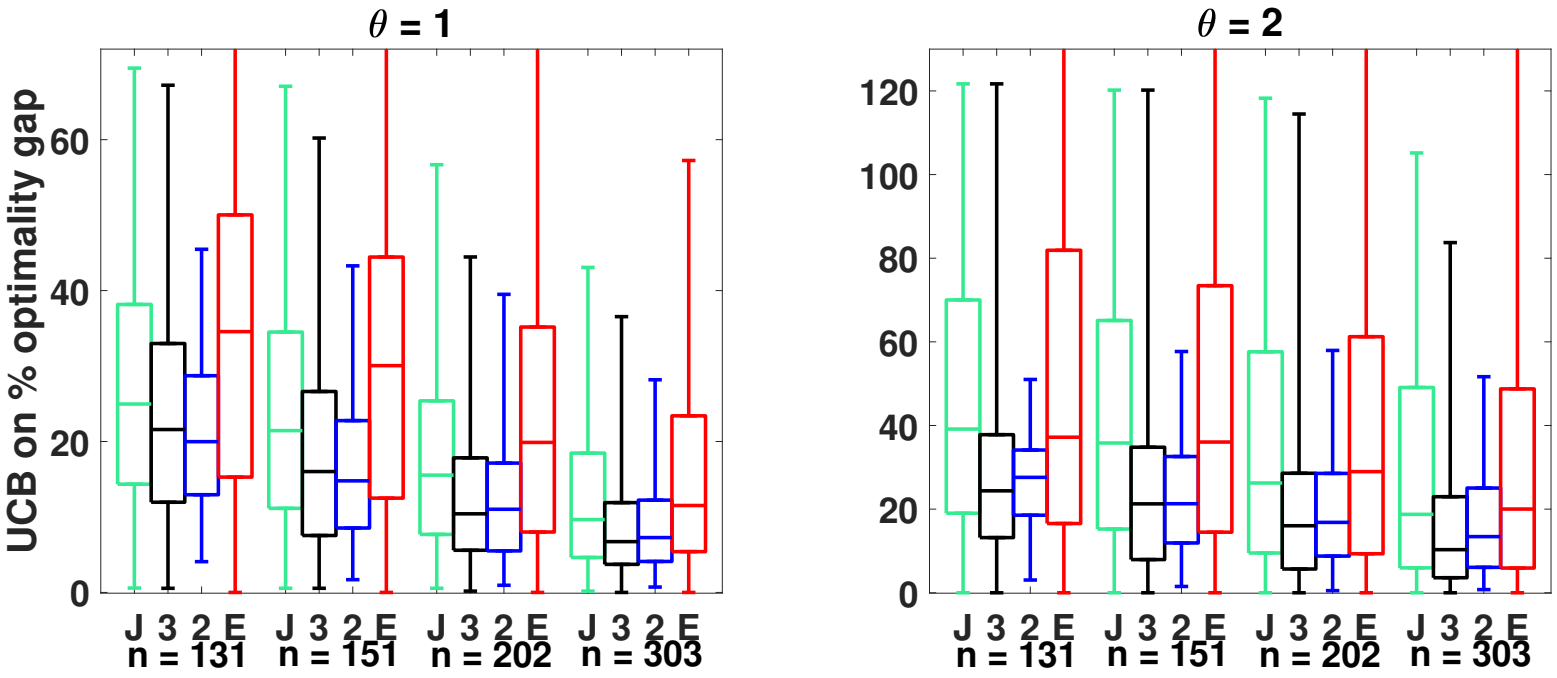
Comparison with J-SAA for $d_x = 100$

J: J-SAA + OLS

3 & **2**: Wasserstein radius specified using Approaches 3 & 2

E: ER-SAA + OLS

Lower y-axis value $\implies$ closer to optimal



Boxes: 25, 50, and 75 percentiles of upper confidence bounds

Whiskers: 2 and 98 percentiles

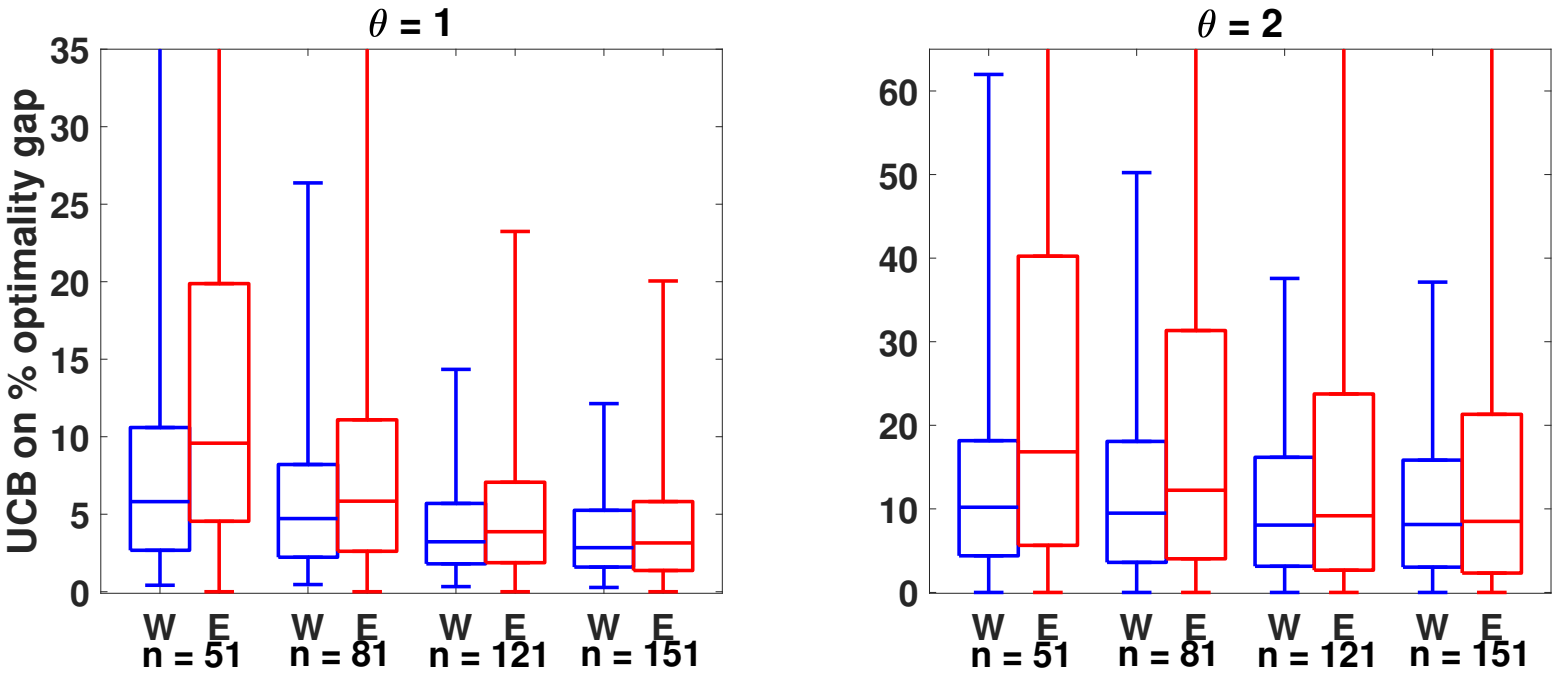
Sample sizes: $\{1.3,\ 1.5,\ 2,\ 3\} \times (\dim(X) + 1)$

Modularity Benefit for $d_x = 100$: Bring on Lasso

W: Wasserstein radius for ER-DRO + Lasso using Approach 2

E: ER-SAA + Lasso

Lower y-axis value $\implies$ closer to optimal



Boxes: 25, 50, and 75 percentiles of upper confidence bounds

Whiskers: 2 and 98 percentiles

Sample sizes: $\{0.5,\ 0.8,\ 1.2,\ 1.5\} \times (\dim(X) + 1)$

Outline

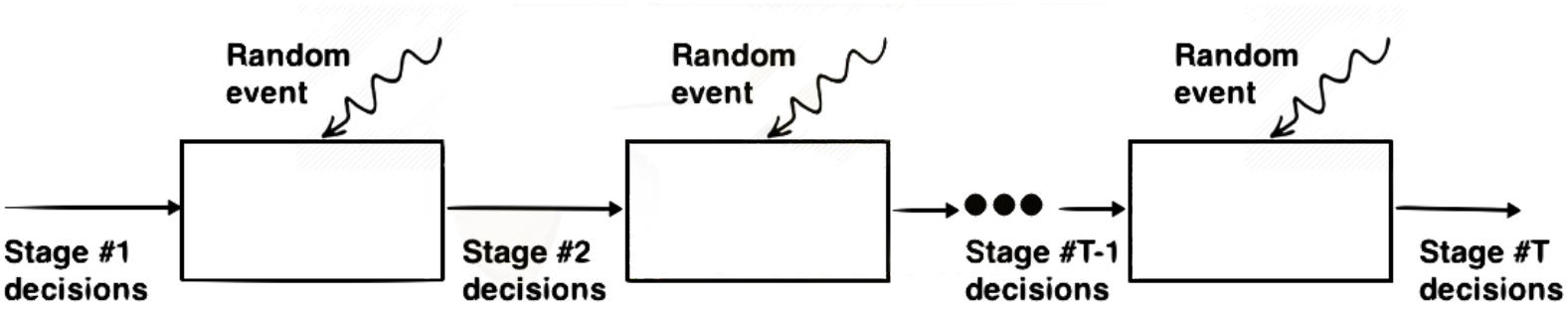
Handling Heteroscedastic Errors

arXiv:2101.03139

- Key assumption thus far: true model is $Y = f^*(X) + \varepsilon$ with errors ε independent of covariates X
- Assumption may be violated for some applications
  - Example: variability of product demands/wind generators can depend on seasonality/location
- Relaxed assumption: $Y = f^*(X) + Q^*(X)\varepsilon$ with X, ε indep.
  - Estimate f^* and $Q^* \implies$ estimate samples of ε
  - Theoretical results for ER-SAA and ER-DRO readily generalize

K., Bayraksan, and Luedtke. Heteroscedasticity-aware residuals-based contextual stochastic optimization

Multistage Stochastic Optimization



- Stochastic process $\{\xi_t\}$ and i.i.d. errors $\{\varepsilon_t\}$ satisfying

$$\xi_t = m_t^*(\xi_{t-1}) + \varepsilon_t, \quad \forall t \in \mathbb{Z}$$

- Given n historical observations of the stochastic process, estimate m_t^* by $\hat{m}_{t,n}$ and compute empirical residuals $\{\hat{\varepsilon}_n^i\}$
- Given ξ_{t-1}, use $\{\hat{m}_{t,n}(\xi_{t-1}) + \hat{\varepsilon}_n^i\}$ as scenarios for stage t
- Tailored convergence analysis required since *same empirical errors used* in each time stage

K., Ho-Nguyen, and Luedtke. Multistage stochastic optimization given time series data

Concluding Remarks

Empirical residuals formulations: A modular approach to using covariate information in optimization

- Converges under appropriate assumptions on prediction and optimization models
- Trade-off in choosing prediction model class: using a misspecified model can lead to better results with limited data
- Preprints on Optimization Online and arXiv

Concluding Remarks

Empirical residuals formulations: A modular approach to using covariate information in optimization

- Converges under appropriate assumptions on prediction and optimization models
- Trade-off in choosing prediction model class: using a misspecified model can lead to better results with limited data
- Preprints on Optimization Online and arXiv

Future research directions

- Formulations with stochastic constraints, discrete recourse decisions; robust multistage optimization
- Application to energy systems optimization

References I

G.-Y. Ban and C. Rudin. The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67(1):90–108, 2018.

G.-Y. Ban, J. Gallien, and A. J. Mersereau. Dynamic procurement of new products with covariate information: The residual tree method. Articles In Advance. *Manufacturing & Service Operations Management*, pages 1–18, 2018.

R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani. Predictive inference with the jackknife+. *arXiv preprint arXiv:1905.02928*, pages 1–40, 2019.

T. Bazier-Matte and E. Delage. Generalization bounds for regularized portfolio selection with market side information. *INFOR: Information Systems and Operational Research*, 58(2):374–401, 2020.

D. Bertsimas and N. Kallus. From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044, 2020.

D. Bertsimas, C. McCord, and B. Sturt. Dynamic optimization with side information. *arXiv preprint arXiv:1907.07307*, pages 1–37, 2019.

P. Donti, B. Amos, and J. Z. Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, pages 5484–5494, 2017.

A. N. Elmachtoub and P. Grigas. Smart "predict, then optimize". *arXiv preprint arXiv:1710.08005*, pages 1–38, 2017.

A. Esteban-Pérez and J. M. Morales. Distributionally robust stochastic programs with side information based on trimmings. *arXiv preprint arXiv:2009.10592*, 2020.

References II

Y.-h. Kao, B. V. Roy, and X. Yan. Directed regression. In *Advances in Neural Information Processing Systems*, pages 889–897, 2009.

D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS, 2019.

S. Sen and Y. Deng. Learning enabled optimization: Towards a fusion of statistical learning and stochastic programming. `http://www.optimization-online.org/DB_FILE/2017/03/5904.pdf`, 2018.

Asymptotic Optimality of ER-SAA Solutions

Assumption: The regression procedure satisfies

- Pointwise error consistency: $\hat{f}_n(x) \xrightarrow{p} f^*(x) \quad$ for a.e. x
- Mean-squared estimation error consistency:

$$\frac{1}{n}\sum_{i=1}^{n}\|f^*(x^i) - \hat{f}_n(x^i)\|^2 \xrightarrow{p} 0.$$

Informal Theorem (Asymptotic Optimality)

Under the above assumptions†, the ER-SAA solution $\hat{z}_n^{ER}(x)$ is asymptotically optimal for a.e. x, i.e.,

$$\mathbb{E}_\varepsilon\left[c(\hat{z}_n^{ER}(x), f^*(x) + \varepsilon)\right] \xrightarrow{p} v^*(x)$$

†Plus some mild standard assumptions on the true conditional SP

Rate of Convergence of ER-SAA Solutions

Assumption: There is a constant $r \in (0, 1]$ such that the regression procedure satisfies

- Pointwise error rate: $\|f^*(x) - \hat{f}_n(x)\|^2 = O_p\left(n^{-r}\right)$
- Mean-squared estimation error rate:

$$\frac{1}{n}\sum_{i=1}^{n}\|f^*(x^i) - \hat{f}_n(x^i)\|^2 = O_p(n^{-r})$$

- OLS regression, Lasso satisfy assumption with $r = 1$
- CART, RF regression satisfy assumption with $r = \frac{O(1)}{\dim(X)}$

Informal Theorem (Rate of Convergence)

Under the above assumptions, ER-SAA solution $\hat{z}_n^{ER}(x)$ satisfies

$$\mathbb{E}_{\varepsilon}\left[c(\hat{z}_n^{ER}(x), f^*(x) + \varepsilon)\right] = v^*(x) + O_p(n^{-r/2})$$

Finite Sample Guarantees for ER-SAA Solutions

Define

- $\hat{S}_n^{ER}(x) :=$ set of optimal solutions to ER-SAA
- $S^{\kappa}(x) :=$ set of κ-optimal solutions to the true conditional SP, i.e., points in $\mathcal{Z}$ with objective value $\leq v^*(x) + \kappa$

Assumption: The errors ε are sub-Gaussian (light tail distribution)

Given: target optimality gap $\kappa > 0$, unreliability level $\delta \in (0, 1)$

Goal: Estimate sample size n required for

$$\mathbb{P}\left\{\hat{S}_n^{ER}(x) \subseteq S^{\kappa}(x)\right\} \geq 1 - \delta,$$

i.e., with probability $\geq 1 - \delta$, optimal solutions of ER-SAA are κ-optimal to the true conditional SP

Finite Sample Guarantees for ER-SAA Solutions

Estimate sample size n required for $\mathbb{P}\left\{\hat{S}_n^{ER}(x) \subseteq S^{\kappa}(x)\right\} \geq 1-\delta$

- If f^* is linear and we use OLS regression, then require

$$n \geq \frac{O(1)}{\kappa^2}\left[d_z \log\left(\frac{O(1)}{\kappa}\right) + d_y \log\left(\frac{O(1)}{\delta}\right) + d_x d_y\right]$$

- If f^* is s-sparse linear and we use the Lasso, then require

$$n \geq \frac{O(1)}{\kappa^2}\left[d_z \log\left(\frac{O(1)}{\kappa}\right) + s d_y \log\left(\frac{O(1)}{\delta}\right) + s \log(d_x) d_y\right]$$

- If f^* is Lipschitz and we use kNN regression, then require

$$n \geq \frac{O(1) d_z}{\kappa^2} \log\left(\frac{O(1)}{\kappa}\right) + \left(\frac{O(1) d_y}{\kappa^2}\right)^{d_x}\left[d_x \log\left(\frac{O(1) d_x d_y}{\kappa^2}\right) + \log\left(\frac{O(1)}{\delta}\right)\right]$$

Numerical Study: Optimal Resource Allocation

$$\min_{z \geq 0} \; c^\mathsf{T} z + \mathbb{E}_Y[Q(z, Y)]$$

- z_i: quantity of resource $i \in \mathcal{I}$ (order before demands realized)
- Y_j: uncertain demand of customer type $j \in \mathcal{J}$

$$\begin{aligned} Q(z, Y) := \min_{w, v \geq 0} \quad & d^\mathsf{T} w \\ \text{s.t.} \quad & \sum_{j \in \mathcal{J}} v_{ij} \leq z_i, \quad \forall i \in \mathcal{I}, \\ & \sum_{i \in \mathcal{I}} \mu_{ij} v_{ij} + w_j \geq Y_j, \quad \forall j \in \mathcal{J}. \end{aligned}$$

- v_{ij}: amount of resource i allocated to customer type j
- w_j: amount of customer type j demand that is not met
- $\mu_{ij} \geq 0$: service rate of resource i for customer type j

Numerical Study: Optimal Resource Allocation

- Meet demands of 30 customer types for 20 resources
- Uncertain demands Y generated according to

$$Y_j = \alpha_j^* + \sum_{l=1}^{3} \beta_{jl}^* (X_l)^\theta + \varepsilon_j, \quad \forall j \in \{1, \cdots, 30\},$$

where $\varepsilon_j \sim \mathcal{N}(0, \sigma_j^2)$, $\theta \in \{0.5, 1, 2\}$, $\dim(X) \in \{10, 100\}$

- Fit linear model with OLS/Lasso regression (even when $\theta \neq 1$)

$$Y_j = \alpha_j + \sum_{l=1}^{\dim(X)} \beta_{jl} X_l + \eta_j, \quad \forall j \in \{1, \cdots, 30\},$$

where η_j are zero-mean errors

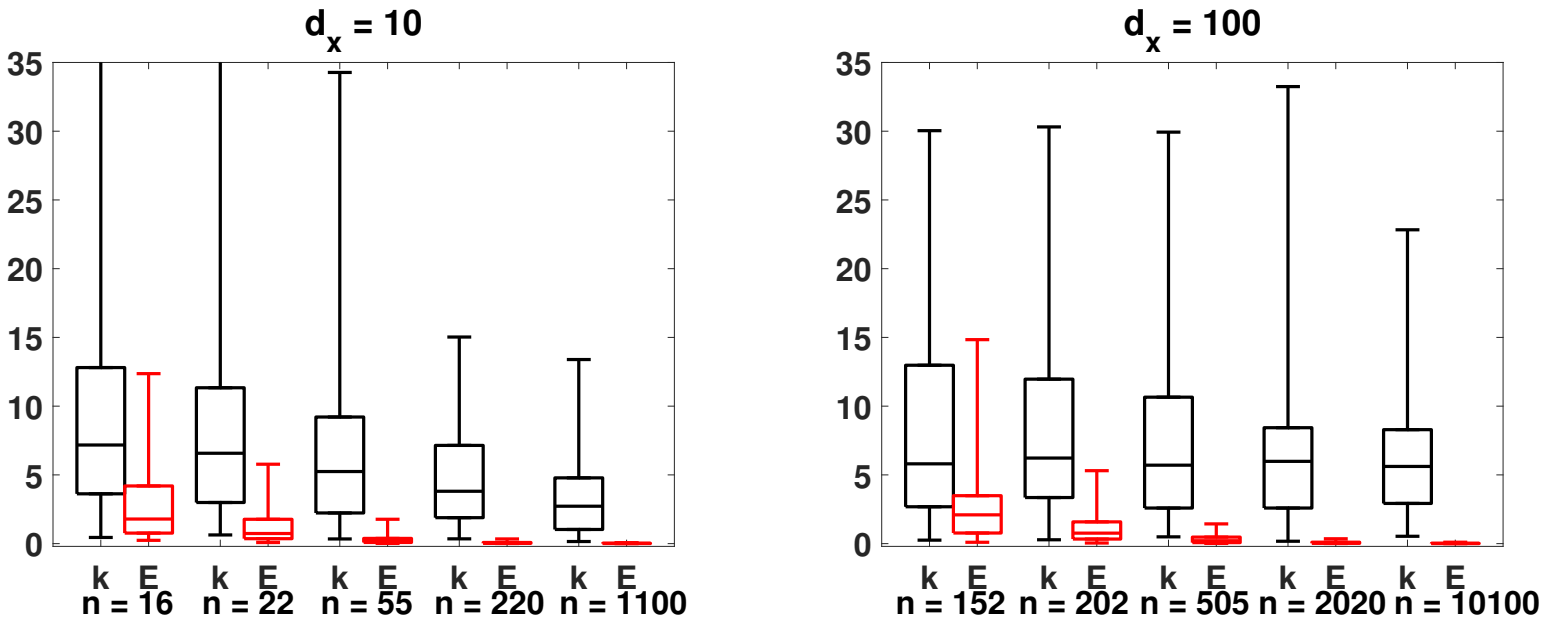
- Estimate optimality gap of solutions $\hat{z}_n^{ER}(x)$ and $\hat{z}_n^{J}(x)$

Results with Correct Model Class ($\theta = 1$)

Red (E): ER-SAA + OLS

Black (k): Reweighted SAA with kNN (Bertsimas and Kallus, 2020)

Lower y-axis value $\implies$ closer to optimal



Boxes: 25, 50, and 75 percentiles of upper confidence bounds

Whiskers: 2 and 98 percentiles

Sample sizes: $\{1.5,\ 2,\ 5,\ 20,\ 100\} \times (\dim(X) + 1)$

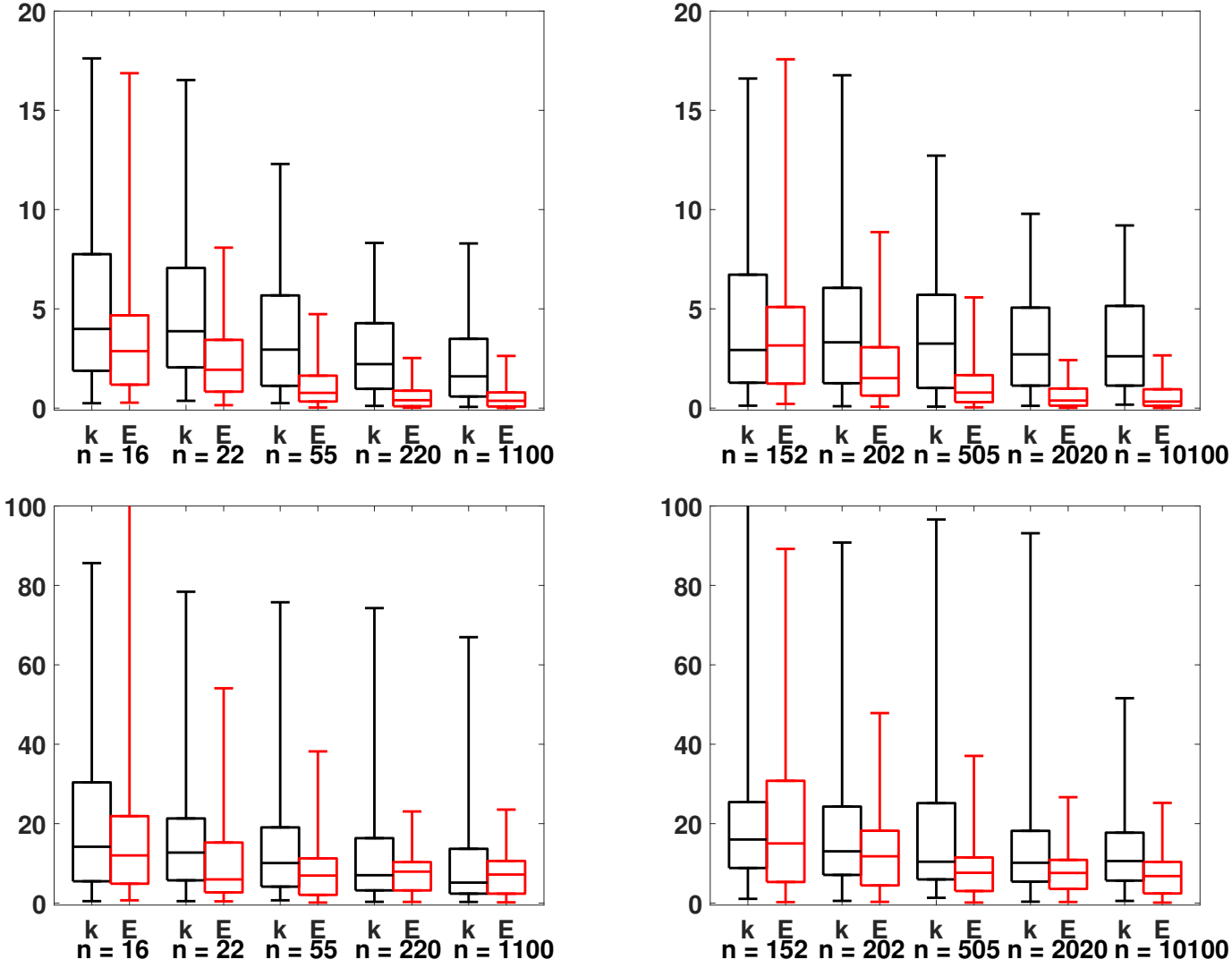
Results with Misspecified Model Class ($\theta \neq 1$)

Red (E): ER-SAA + OLS, Black (k): Reweighted SAA with kNN

$\mathbf{d_x = 10}$ $\mathbf{d_x = 100}$

$\boldsymbol{\theta = 0.5}$

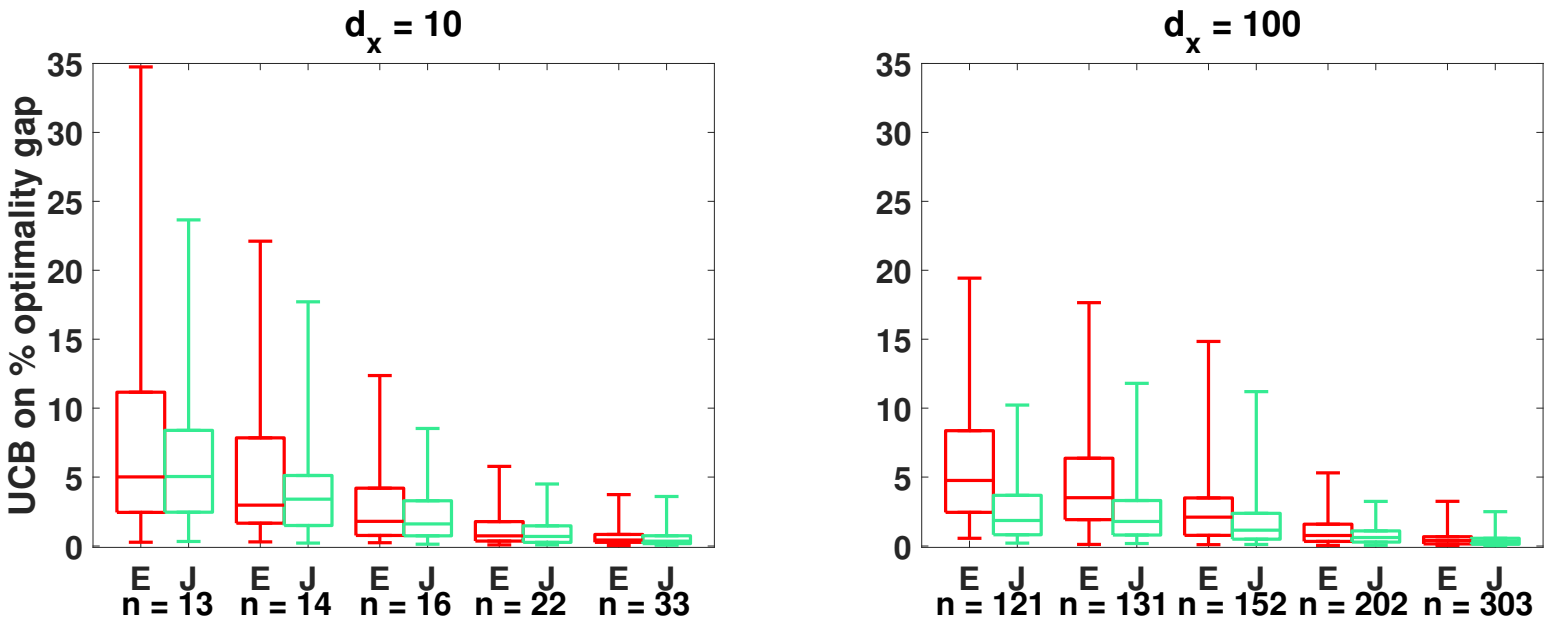
$\boldsymbol{\theta = 2}$



Advantage of J-SAA with Limited Data ($\theta = 1$)

Red (E): ER-SAA + OLS, Green (J): J-SAA + OLS

Lower y-axis value $\implies$ closer to optimal

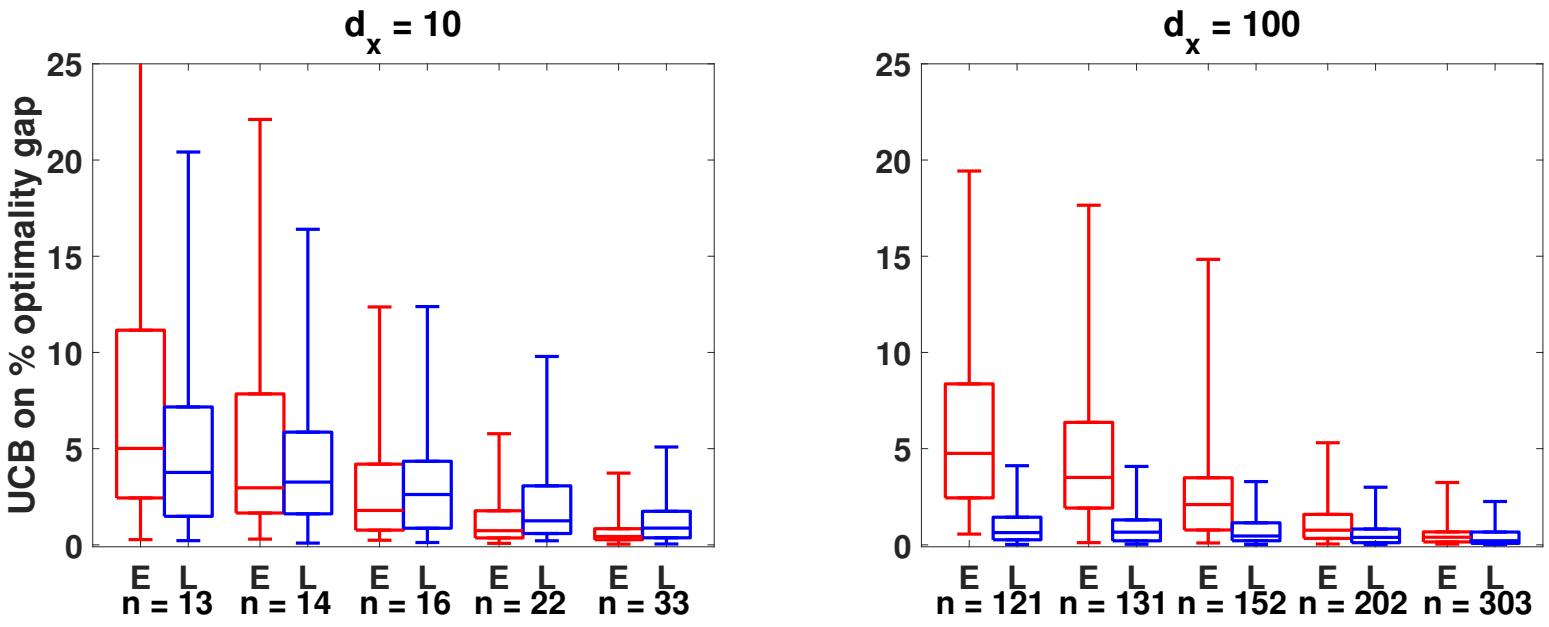


Boxes: 25, 50, and 75 percentiles of upper confidence bounds
Whiskers: 2 and 98 percentiles
Sample sizes: $\{1.1,\ 1.2,\ 1.5,\ 2,\ 3\} \times (\dim(X) + 1)$

Modularity Benefit: Bring on Lasso ($\theta = 1$)

Red (E): ER-SAA + OLS, Blue (L): ER-SAA + Lasso

Lower y-axis value $\implies$ closer to optimal



Boxes: 25, 50, and 75 percentiles of upper confidence bounds
Whiskers: 2 and 98 percentiles
Sample sizes: $\{1.1,\ 1.2,\ 1.5,\ 2,\ 3\} \times (\dim(X) + 1)$

Lasso Results with Misspecified Model Class ($\theta \neq 1$)

Red (E): ER-SAA + OLS, Blue (L): ER-SAA + Lasso

$\mathbf{d_x = 10}$ $\mathbf{d_x = 100}$

$\boldsymbol{\theta = 0.5}$

$\boldsymbol{\theta = 2}$