Data-Driven Sample Average Approximation with Covariate Information

Rohit Kannan

Wisconsin Institute for Discovery
University of Wisconsin-Madison

November 10, 2020

Joint work with Güzin Bayraksan and Jim Luedtke

Funding: MACSER project (DOE)

Traditional data-driven stochastic programming

- Traditional SP: minimize expected cost

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

Traditional data-driven stochastic programming

- Traditional SP: minimize expected cost

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

- Data-driven SP: given (i.i.d.) samples $\{y^i\}_{i=1}^n$ of Y, construct Sample Average Approximation (SAA)

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

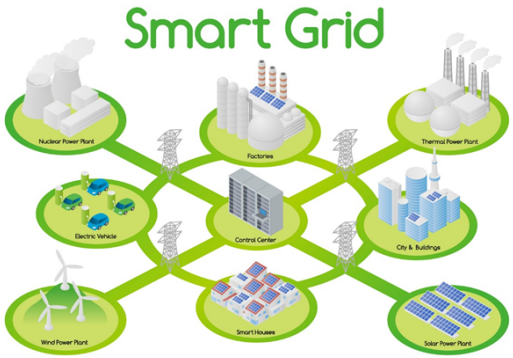
- SAA theory: optimal value and solutions converge as $n \to \infty$, error is $O_p(n^{-1/2})$ (Shapiro et al., 2009)

Stochastic programming with covariate information

Enter Machine Learning

- Assume we have historical data of form $\mathcal{D}_n := \{(y^i, x^i)\}_{i=1}^n$ (parameters and *covariates*)
  - Covariates are also referred to as *features* or *side information*
- When making decision z, we observe a *new* covariate x, which we can use to predict y (with error)
- How to integrate learning (predicting Y given $X = x$) with optimization?

Example applications



Production/Inventory Planning
(Ban et al., 2018; Bertsimas and Kallus, 2020)

Y: Product demands

X: Seasonality; Web search results

z: Production and Inventory decisions

Power Grid Scheduling
(Donti et al., 2017)

Y: Load; Renewable energy outputs

X: Weather observations; Time/Season

z: Generator scheduling decisions

Problem setup

Given

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

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 $X = x$

Want to solve

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

Assume

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

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

- Known function class $\mathcal{F}$ such that $f^* \in \mathcal{F}$

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, 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, 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

Empirical Residuals-based Sample Average Approximation

Approach suggested in Kim and Mehrotra (2015), Sen and Deng (2018); analyzed in Ban et al. (2018) for a specific application

1. Estimate f^* using your favorite ML method $\Rightarrow \hat{f}_n$, and 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 $X = x$

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

Empirical Residuals-based Sample Average Approximation

Approach suggested in Kim and Mehrotra (2015), Sen and Deng (2018); analyzed in Ban et al. (2018) for a specific application

1. Estimate f^* using your favorite ML method $\Rightarrow \hat{f}_n$, and 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
- Surprisingly, no general convergence analysis

Toward convergence theory: Definitions

Notation:

- $v^*(x)$ = optimal value of true problem
- $S^\kappa(x)$ = set of κ-optimal solutions of true problem
- $\hat{z}_n^{ER}(x) \in \hat{S}_n^{ER}(x)$ = set of optimal solutions to (ER-SAA)

Toward convergence theory: Definitions

Notation:

- $v^*(x)$ = optimal value of true problem
- $S^\kappa(x)$ = set of κ-optimal solutions of true problem
- $\hat{z}_n^{ER}(x) \in \hat{S}_n^{ER}(x)$ = set of optimal solutions to (ER-SAA)

Asymptotic optimality: the out-of-sample "cost" of data-driven solutions approaches the minimum cost of the true problem as the number of data samples increases

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

Toward convergence theory: Definitions

Notation:

- $v^*(x)$ = optimal value of true problem
- $S^\kappa(x)$ = set of κ-optimal solutions of true problem
- $\hat{z}_n^{ER}(x) \in \hat{S}_n^{ER}(x)$ = set of optimal solutions to (ER-SAA)

Asymptotic optimality: the out-of-sample "cost" of data-driven solutions approaches the minimum cost of the true problem as the number of data samples increases

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

Assume for this talk: Two-stage stochastic LP setting

$$\min_{z \in \mathcal{Z}} c_z^\mathsf{T} z + \mathbb{E}_Y[Q(z, Y)],$$

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

Asymptotic optimality of ER-SAA solutions

Assumption: The regression procedure satisfies

- Pointwise error consistency: $\hat{f}_n(x) \xrightarrow{p} f^*(x)$
- 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: Under the above assumptions, the ER-SAA solution $\hat{z}_n^{ER}(x)$ is asymptotically optimal for a.e. x

Finite sample guarantees for ER-SAA solutions

sub-Gaussian errors ε, $\kappa > 0$: optimality gap, $\delta \in (0, 1)$: reliability level

Estimate sample size n required for $\mathbb{P}\left\{\hat{S}_n^{ER}(x) \subseteq S^{\kappa}(x)\right\} \geq 1 - \delta$,
i.e., "optimal solutions of (ER-SAA) are nearly optimal to the true problem with probability $\geq 1 - \delta$"

Finite sample guarantees for ER-SAA solutions

sub-Gaussian errors ε, $\kappa > 0$: optimality gap, $\delta \in (0,1)$: reliability level

Estimate sample size n required for $\mathbb{P}\left\{\hat{S}_n^{ER}(x) \subseteq S^{\kappa}(x)\right\} \geq 1-\delta$, i.e., "optimal solutions of (ER-SAA) are nearly optimal to the true problem with probability $\geq 1-\delta$"

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

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

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

Finite sample guarantees for ER-SAA solutions

sub-Gaussian errors ε, $\kappa > 0$: optimality gap, $\delta \in (0,1)$: reliability level

Estimate sample size n required for $\mathbb{P}\left\{\hat{S}_n^{ER}(x) \subseteq S^{\kappa}(x)\right\} \geq 1-\delta$, i.e., "optimal solutions of (ER-SAA) are nearly optimal to the true problem with probability $\geq 1-\delta$"

- 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

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

Resource allocation model (Luedtke, 2014)

Two-stage resource allocation LP model

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

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

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

Resource allocation model (Luedtke, 2014)

Two-stage resource allocation LP model

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

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

  where $\varepsilon_j \sim \mathcal{N}(0, \sigma_j^2)$, $p \in \{0.5, 1, 2\}$, $\dim(X) \in \{10, 100\}$
- Fit linear model with OLS regression (even when $p \neq 1$)

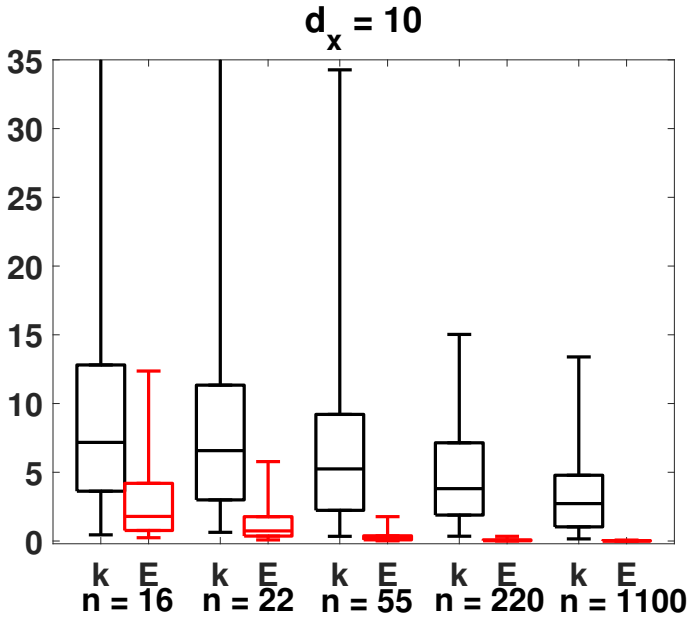
$$Y_j = \alpha_j + \sum_{l=1}^{\dim(X)} \beta_{jl} X_l + \eta_j, \quad \forall j \in \{1, \cdots, 30\},$$

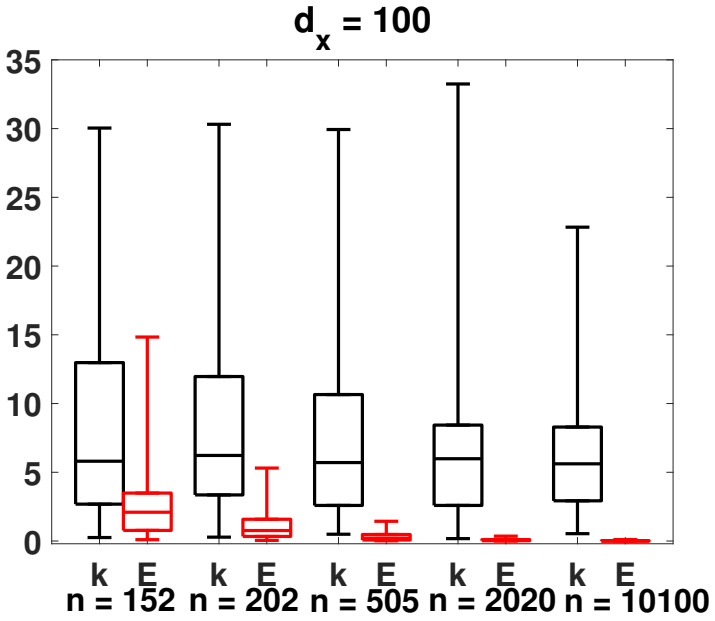
  where η_j are zero-mean errors
- Estimate optimality gap of ER-SAA solutions $\hat{z}_n^{ER}(x)$

Results with correct model class ($p = 1$)

Red (E): ER-SAA + OLS

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





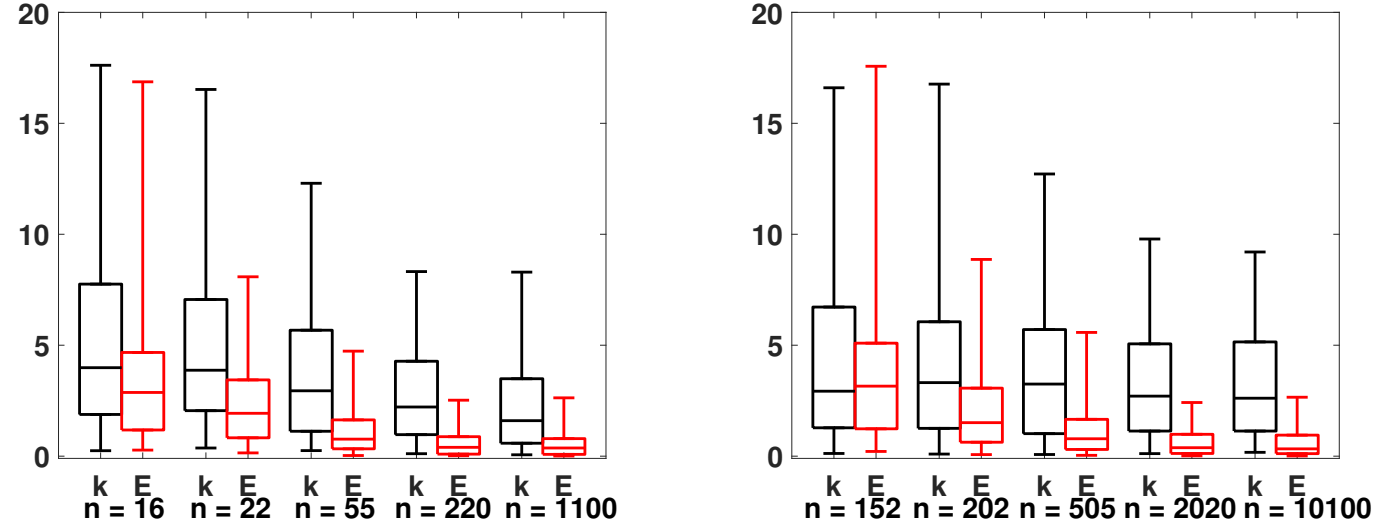
Boxes: 25, 50, and 75 percentiles of upper confidence bounds;
Whiskers: 2 and 98 percentiles

Results with misspecified model class ($p \neq 1$)

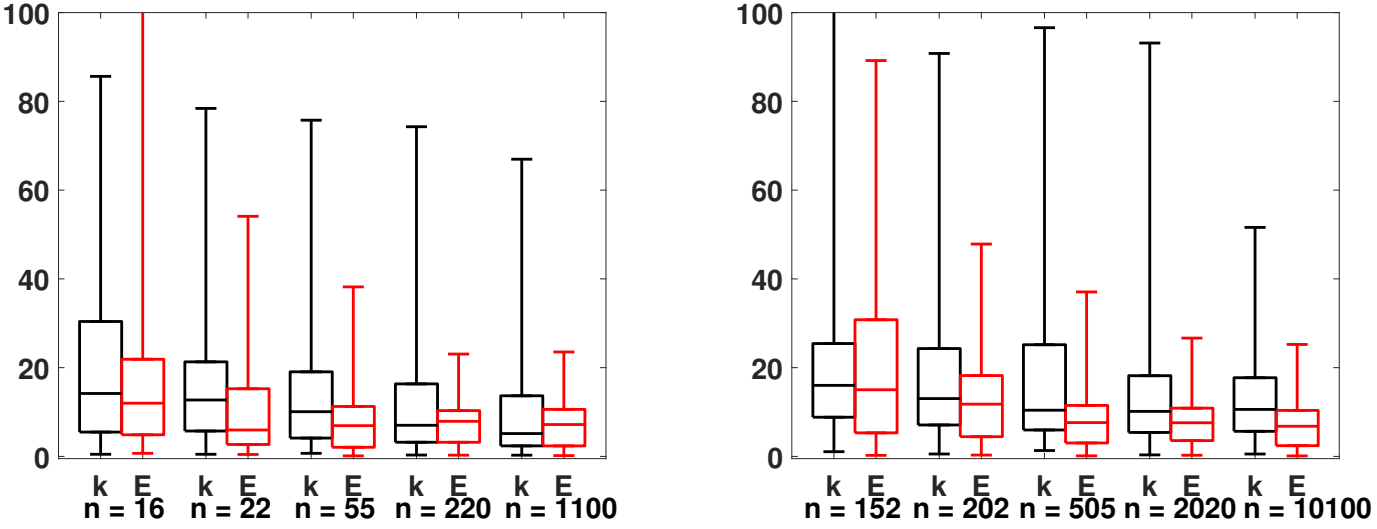
Red (E): ER-SAA + OLS, Black (k): Reweighted SAA with kNN

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

$\mathbf{p = 0.5}$



$\mathbf{p = 2}$



Concluding remarks

Empirical residuals SAA: 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

Preprint on Optimization Online:

`http://www.optimization-online.org/DB_FILE/2020/07/7932.pdf`

Includes

- two new data-driven SAA formulations
- rate of convergence results
- additional computational experiments + source code

Questions? rohit.kannan@wisc.edu

References

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.

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

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.

K. Kim and S. Mehrotra. A two-stage stochastic integer programming approach to integrated staffing and scheduling with application to nurse management. *Operations Research*, 63(6):1431–1451, 2015.

J. Luedtke. A branch-and-cut decomposition algorithm for solving chance-constrained mathematical programs with finite support. *Mathematical Programming*, 146(1-2): 219–244, 2014.

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.

A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.