Data-driven stochastic optimization with covariate information

Rohit Kannan

Wisconsin Institute for Discovery
University of Wisconsin-Madison

September 29, 2020

Funding sources: MACSER project (DOE), BP

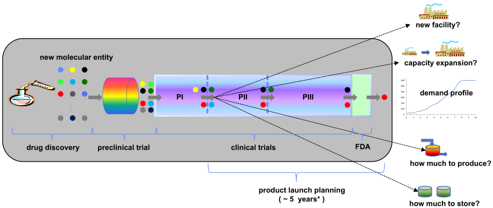
Academic background

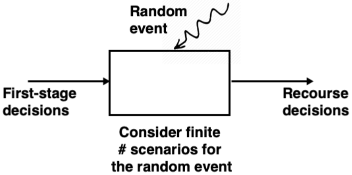
- Bachelors in Chemical Engineering from IIT Madras
  - Measure of Granger causality for nonlinear multivariate processes.
- Ph.D. in Chemical Engineering from MIT
  - Algorithms, analysis, and software for the global optimization of two-stage stochastic programs (SPs)
- Postdoctoral Associate at UW-Madison (since Jan. 2018)
  - Algorithms for data-driven stochastic optimization with application to power systems

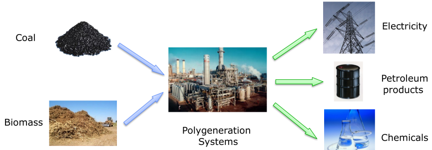
Outline

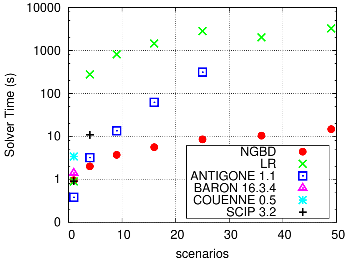
1. Doctoral research highlights
   - Global optimization of SPs
   - Viability of B&B algorithms
2. Postdoctoral research highlights
3. Data-driven SP with covariate information
4. Empirical Residuals SAA
5. Computational experiments
6. Extensions

Global optimization of two-stage stochastic programs





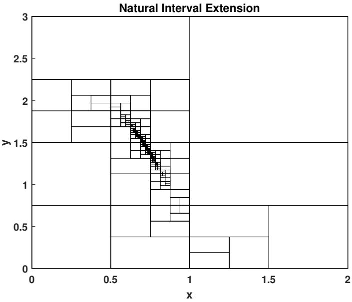


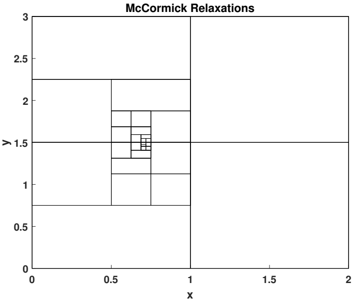


- Even solving nominal problem is hard
- Complexity of generic methods grows exponentially with number of scenarios
- Developed first fully-decomposable algorithm with provable convergence

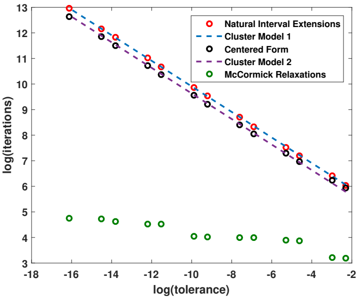
NGBD & LR: decomposition methods
Rest: State-of-the-art solvers

When are B&B algorithms viable?





- Branch-and-bound (B&B) algorithms may face “cluster problem” depending on bounding method used
- Built general theory to understand requirements on bounding method to avoid this problem

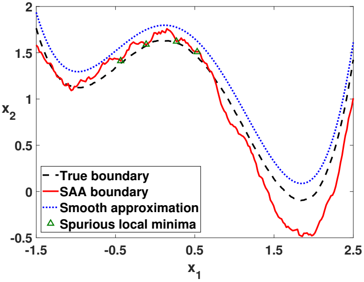


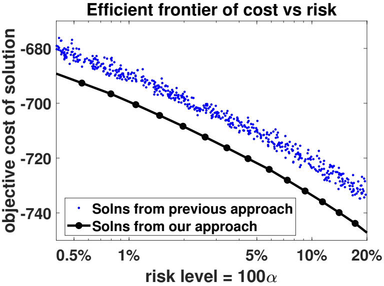
Outline

Optimization with reliability constraints

$$\begin{aligned} \min \ & \text{System cost} \\ \text{s.t.} \ & \text{Deterministic constraints} \\ & \text{Prob(constraints with uncertain parameters hold)} \geq 1 - \alpha \end{aligned}$$

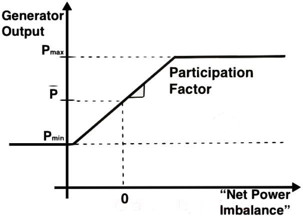
- Reliability constraints can be nonsmooth, nonconvex, and even hard to evaluate with high accuracy!
- Previous approaches are either suboptimal, or do not scale
- Designed a stochastic subgradient method for approximating the efficient frontier of cost versus risk



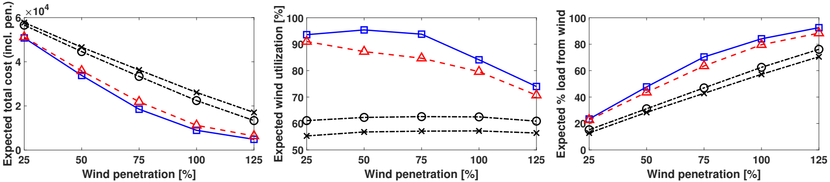


Better integration of renewables

- Generators balance variability in loads and renewables by activating reserves using piecewise-affine control policy
  - Captures behavior of generators when they reach their limits
  - Less conservative solutions than affine policy that forces generator outputs to lie within limits with high probability



- Tailored decomposition method for DC-OPF



□: our approach. Δ: penalty approach. ∘ and ×: affine policy + chance constraints

Joint work with Jim Luedtke and Line Roald

Integrated learning & optimization under uncertainty
Focus of this presentation

Example application: power generation scheduling (unit commitment)

- Uncertain parameters: renewables outputs, loads
- Covariates: current and past weather observations, current time (hour/day of week/month)

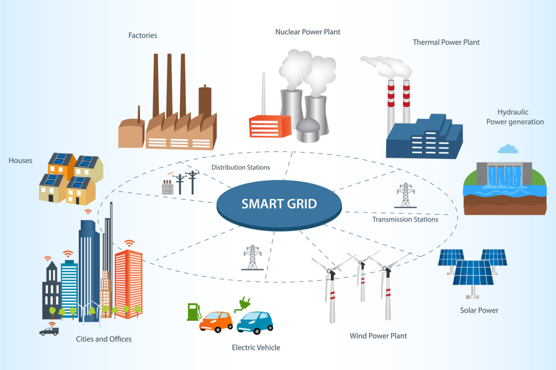


Image credit: IEEE Innovation at Work

Outline

Optimization under uncertainty

General (ill-posed) optimization with uncertain parameters Y:

$$\min_{z \in \mathcal{Z}} c(z, Y)$$

- $\mathcal{Z}$ is the feasible region (assume known)
- Y is a vector of uncertain parameters $\Rightarrow$ Problem not well defined

Optimization under uncertainty

General (ill-posed) optimization with uncertain parameters Y:

$$\min_{z \in \mathcal{Z}} c(z, Y)$$

- $\mathcal{Z}$ is the feasible region (assume known)
- Y is a vector of uncertain parameters $\Rightarrow$ Problem not well defined

Popular modeling approaches:

1. Stochastic: assuming distribution of Y known, minimize average cost

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

2. Robust: assuming support of Y known, minimize worst-case cost

$$\min_{z \in \mathcal{Z}} \max_{y \in \mathcal{Y}} c(z, y)$$

Example: Resource allocation model (Luedtke [2014])

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

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

- Y_j: uncertain demand of customer type j
- z_i: quantity of resource i (allocate before observing demands)
- 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

Data-driven stochastic programming

Traditional SP paradigm

- Minimize expected cost

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

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

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

- Sample average approximation (SAA) theory: optimal value and solutions converge as $n \to \infty$, error is $O_p(n^{-1/2})$

Data-driven stochastic programming

Traditional SP paradigm

- Minimize expected cost

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

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

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

- Sample average approximation (SAA) theory: optimal value and solutions converge as $n \to \infty$, error is $O_p(n^{-1/2})$

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*)
- 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) with optimization?

Traditional integrated learning and optimization

Separate learning and optimization steps

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

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

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

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

- Modular approach
- Can expect to work well if (and likely only if) prediction is accurate

Improved integrated learning and optimization

Approach 1: Modify the learning step

- Change loss function in ML training step to reflect use of prediction in optimization model
- E.g., Kao et al. [2009], Donti et al. [2017], Elmachtoub and Grigas [2017]
- Results in a challenging training problem
- Less modular then traditional approach

Improved integrated learning and optimization

Approach 1: Modify the learning step

- Change loss function in ML training step to reflect use of prediction in optimization model
- E.g., Kao et al. [2009], Donti et al. [2017], Elmachtoub and Grigas [2017]
- Results in a challenging training problem
- Less modular then traditional approach

Approach 2 (this work): Modify the optimization step

- Change optimization model to reflect uncertainty in prediction
- Bertsimas and Kallus [2019], Kim and Mehrotra [2015], Sen and Deng [2018], Ban et al. [2018]

Improved integrated learning and optimization

Approach 1: Modify the learning step

- Change loss function in ML training step to reflect use of prediction in optimization model
- E.g., Kao et al. [2009], Donti et al. [2017], Elmachtoub and Grigas [2017]
- Results in a challenging training problem
- Less modular then traditional approach

Approach 2 (this work): Modify the optimization step

- Change optimization model to reflect uncertainty in prediction
- Bertsimas and Kallus [2019], Kim and Mehrotra [2015], Sen and Deng [2018], Ban et al. [2018]

Approach 3: Direct solution learning

- Attempt to directly learn a mapping from x to a solution z (Bertsimas and Kallus [2019], Ban and Rudin [2018])
- Handling constraints and large dimensions of z is challenging

Problem setup

Given

- Joint observations $\mathcal{D}_n := \{(y^i, x^i)\}_{i=1}^n$, assumed to be drawn from joint random variables (Y, X)
- New random covariate observation $X = x$

Want to solve

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

Minimize expected cost given the observed covariate x

Outline

Problem setup

Given data $\mathcal{D}_n := \{(y^i, x^i)\}_{i=1}^n$, new 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 $f^* \in \mathcal{F}$

Joint work with Jim Luedtke and Güzin Bayraksan

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 model $\Rightarrow \hat{f}_n$, and compute *empirical residuals* $\hat{\varepsilon}^i_n := y^i - \hat{f}_n(x^i)$, $i \in [n]$
2. Use $\{\hat{f}_n(x) + \hat{\varepsilon}^i_n\}_{i=1}^n$ as proxy for samples of Y given $X = x$

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

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 model $\Rightarrow \hat{f}_n$, and compute *empirical residuals* $\hat{\varepsilon}^i_n := y^i - \hat{f}_n(x^i)$, $i \in [n]$
2. Use $\{\hat{f}_n(x) + \hat{\varepsilon}^i_n\}_{i=1}^n$ as proxy for samples of Y given $X = x$

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

Modular like traditional approach: Can easily change prediction model in step 1

- Surprisingly, no general convergence analysis
- Improvements when sample size is small?

Small sample size variant

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

- 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 $\hat{\varepsilon}^i_n := y^i - \hat{f}_{-i}(x^i)$, $i \in [n]$
- Use $\{\hat{f}_n(x) + \hat{\varepsilon}^i_n\}_{i=1}^n$ (or $\{\hat{f}_{-i}(x) + \hat{\varepsilon}^i_n\}_{i=1}^n$) as proxy for samples of Y given $X = x$

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

Inspired by Jackknife methods [Barber et al., 2019]

Nonparametric reweighting-based SAA
Bertsimas and Kallus [2019]

- Solve the following reweighted SAA problem

$$\min_{z \in \mathcal{Z}} \sum_{i=1}^{n} w_n^i(x) c(z, y^i),$$

where $\{w_n^i(\cdot)\}_{i=1}^n$ are weight functions determined using $\mathcal{D}_n$

- Constant weights $\Rightarrow$ SAA that ignores covariate information
- Examples of weight functions
  - kNN-based: $w_n^{i,kNN}(x) = \frac{1}{k}\mathbb{I}[x^i \text{ is a kNN of } x]$
  - kernel-based: $w_n^{i,ker}(x) = \dfrac{\mathcal{K}\left(\frac{x^i - x}{h_n}\right)}{\sum_{j=1}^n \mathcal{K}\left(\frac{x^j - x}{h_n}\right)}$
  - others based on regression trees and random forests

Nonparametric reweighting-based SAA
Bertsimas and Kallus [2019]

- Solve the following reweighted SAA problem

$$\min_{z \in \mathcal{Z}} \sum_{i=1}^{n} w_n^i(x) c(z, y^i),$$

where $\{w_n^i(\cdot)\}_{i=1}^n$ are weight functions determined using $\mathcal{D}_n$

- Constant weights $\Rightarrow$ SAA that ignores covariate information
- Examples of weight functions
  - kNN-based: $w_n^{i,kNN}(x) = \frac{1}{k}\mathbb{I}[x^i \text{ is a kNN of } x]$
  - kernel-based: $w_n^{i,ker}(x) = \dfrac{\mathcal{K}\left(\frac{x^i - x}{h_n}\right)}{\sum_{j=1}^n \mathcal{K}\left(\frac{x^j - x}{h_n}\right)}$
  - others based on regression trees and random forests
- Advantages: minimal assumptions on f^* and $\mathcal{D}_n$
- Drawback: could be data-intensive when $\dim(X)$ is large

Toward convergence theory: Definitions

Notation:

- $v^*(x)$ = optimal value of true problem
- $S^\kappa(x)$ = set of κ-optimal solutions of true problem

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

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

Asymptotic optimality of ER-SAA solutions

Two-stage stochastic LP setting:

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

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

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

Asymptotic optimality of J-SAA solutions

Assumption: The regression procedure satisfies

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

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

Informal Theorem: Under the above assumptions, the J-SAA solution $\hat{z}_n^J(x)$ is asymptotically optimal for a.e. x

Rate of convergence of ER-SAA solutions

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

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

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

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

Informal Theorem: Under the above assumptions,

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

Finite sample guarantees for ER-SAA solutions

$\kappa > 0$: 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 approximation are nearly optimal to the true problem with probability $\geq 1 - \delta$"

Assumption: The errors ε are sub-Gaussian

Finite sample guarantees for ER-SAA solutions

$\kappa > 0$: 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 approximation are nearly optimal to the true problem with probability $\geq 1-\delta$"

Assumption: The errors ε are sub-Gaussian

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

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

Finite sample guarantees for ER-SAA solutions

$\kappa > 0$: 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 approximation are nearly optimal to the true problem with probability $\geq 1 - \delta$"

Assumption: The errors ε are sub-Gaussian

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

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

Finite sample guarantees for ER-SAA solutions

$\kappa > 0$: 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 approximation are nearly optimal to the true problem with probability $\geq 1 - \delta$"

Assumption: The errors ε are sub-Gaussian

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

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

Outline

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/Lasso regression (even when $p \neq 1$)

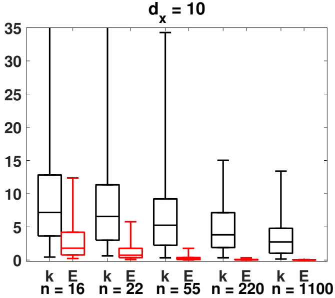
$$Y_j = \alpha_j + \sum_{l=1}^{\dim(X)} \beta_{jl} X_l + \eta_j, \quad \forall j \in \{1, \cdots, 30\},$$

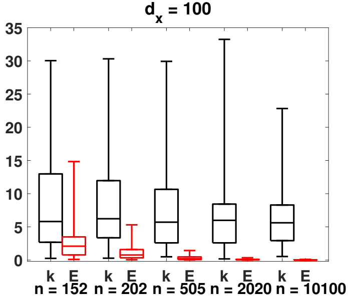
  where η_j are zero-mean errors

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

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

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

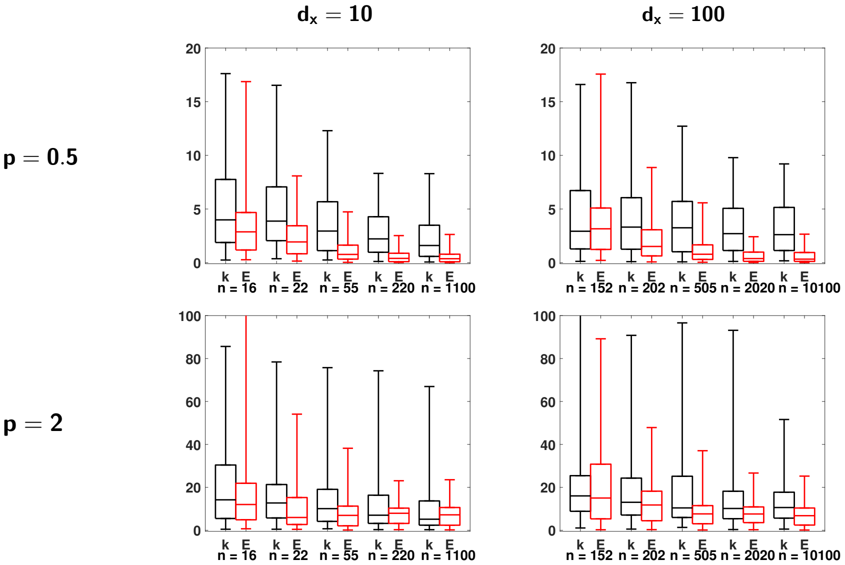




Boxes: 25, 50, and 75 percentiles of upper confidence bounds;
Whiskers: 2 and 98 percentiles

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

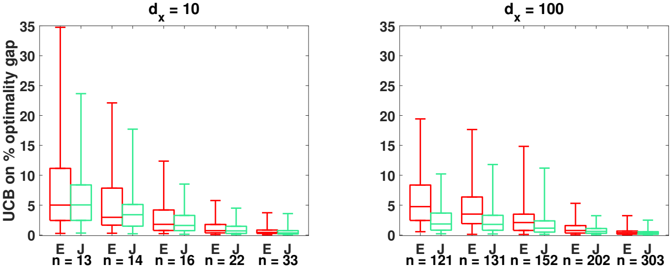
Red (E): ER-SAA + OLS, Black (k): Reweighted SAA with kNN



Advantage of the J-SAA formulation with limited data

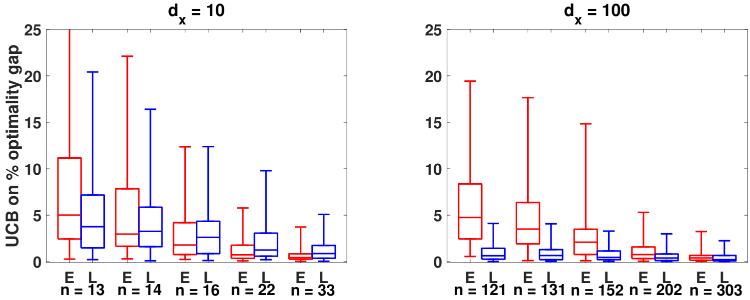
Red (E): ER-SAA + OLS, Green (J): J-SAA + OLS

p = 1



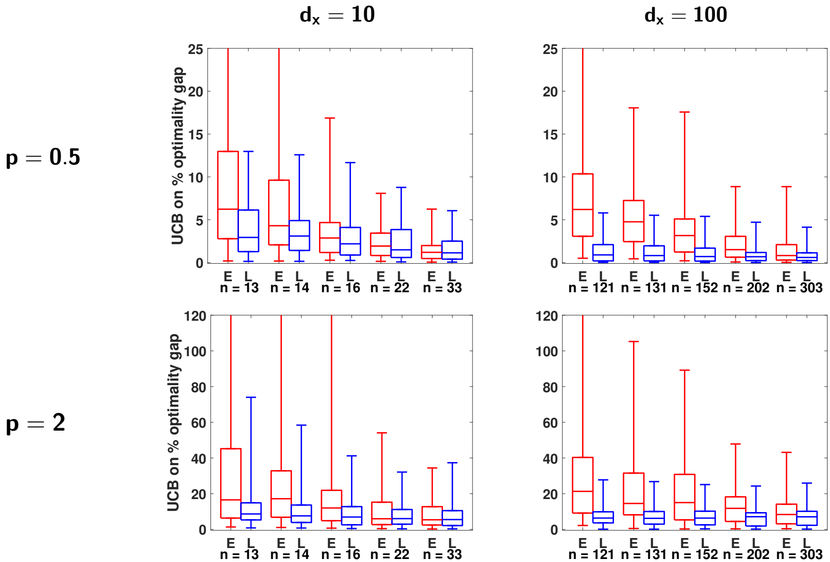
Modularity benefit: Bring on Lasso ($p = 1$)

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



Lasso results with misspecified model class ($p \neq 1$)

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



Outline

Distributionally robust optimization

- Alternative optimization model for small sample sizes n: Distributionally Robust Optimization (DRO)

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

$\hat{\mathcal{P}}_n(x)$ is a "confidence region" for distribution of Y given $X = x$ centered at the empirical residuals distribution $\hat{P}_n^{ER}(x)$

- Example: construct the ambiguity set $\hat{\mathcal{P}}_n(x)$ as

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

Joint work with Jim Luedtke and Güzin Bayraksan

Flavor of DRO results

Let $\hat{v}_n^{DRO}(x)$ and $\hat{z}_n^{DRO}(x)$ denote optimal value and solution of the p-Wasserstein DRO problem

Informal Theorem: Suppose the regression estimates $\hat{f}_n(\cdot)$ satisfy some finite sample guarantees. Then, for a suitable choice of the Wasserstein radius $\zeta_n(x)$:

- $\hat{z}_n^{DRO}(x)$ is asymptotically optimal for a.e. x, and
- the estimator $\hat{z}_n^{DRO}(\cdot)$ and the optimal value $\hat{v}_n^{DRO}(\cdot)$ satisfy the finite sample guarantee

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

Flavor of DRO results

Let $\hat{v}_n^{DRO}(x)$ and $\hat{z}_n^{DRO}(x)$ denote optimal value and solution of the p-Wasserstein DRO problem

Informal Theorem: Suppose the regression estimates $\hat{f}_n(\cdot)$ satisfy some finite sample guarantees. Then, for a suitable choice of the Wasserstein radius $\zeta_n(x)$:

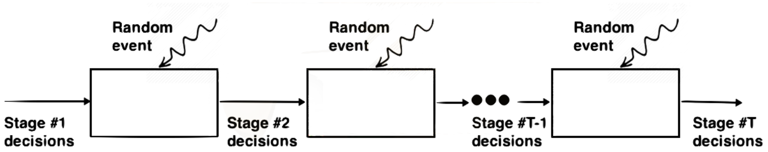
- $\hat{z}_n^{DRO}(x)$ is asymptotically optimal for a.e. x, and
- the estimator $\hat{z}_n^{DRO}(\cdot)$ and the optimal value $\hat{v}_n^{DRO}(\cdot)$ satisfy the finite sample guarantee

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

Challenge for practical use: choosing the DRO radius given $\mathcal{D}_n$

- Promising computational results using cross-validation

Multi-stage stochastic optimization



$$Q_t(x_{t-1}, \xi_{[t]}) := \min_{x_t \in X_t(x_{t-1}, \xi_t)} c_t^\top x_t + \mathbb{E}\left[Q_{t+1}(x_t, \xi_{[t+1]}) \mid \xi_{[t]} \right], \ \forall t \in [T],$$

$$X_t(x_{t-1}, \xi_t) := \left\{ x_t \in \mathbb{R}_+^{d_{x,t}} : B_t x_{t-1} + A_t x_t = h_t(\xi_t) \right\}, \ \forall t \in [T]$$

$\xi_{[t]} := (\xi_1, \cdots, \xi_t)$ and $\{\xi_t\}$ is a stochastic process satisfying

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

for i.i.d. errors $\{\varepsilon_t\}$

Joint work with Jim Luedtke and Nam Ho-Nguyen

Data-driven approximation

- Have $n+1$ historical observations $\mathcal{D}_n := \left\{\tilde{\xi}_{-n}, \tilde{\xi}_{-n+1}, \cdots, \tilde{\xi}_0\right\}$ of the stochastic process
- Estimate the function m_t^* by $\hat{m}_{t,n}$ using a regression method on $\mathcal{D}_n$. Solve for the empirical residuals $\{\hat{\varepsilon}_n^i\}_{i=1}^n$ from

$$\tilde{\xi}_{1-i} = \hat{m}_{1-i,n}(\tilde{\xi}_{-i}, \hat{\varepsilon}_n^i), \quad i \in [n]$$

- Empirical Residuals SAA:

$$\hat{Q}_{t,n}^{ER}(x_{t-1}, \xi_t) := \min_{x_t \in X_t(x_{t-1}, \xi_t)} c_t^\top x_t + \frac{1}{n} \sum_{i=1}^n \hat{Q}_{t+1,n}^{ER}(x_t, \hat{m}_{t+1,n}(\xi_t, \hat{\varepsilon}_n^i)), \ \forall t \in [T]$$

- For multistage stochastic LP, can solve with stochastic dual dynamic programming (SDDP)
- Different convergence analysis required since *same* empirical errors used in each time stage

Concluding remarks

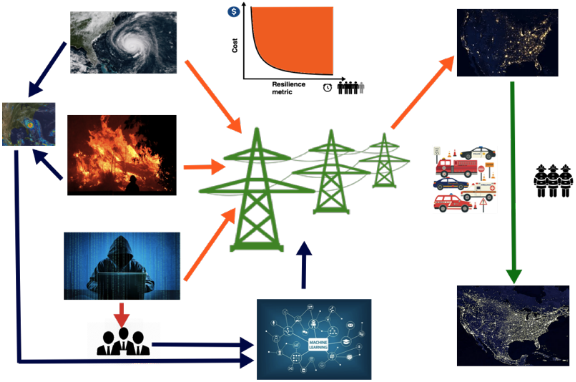
Empirical residuals SAA: A modular approach to using covariate information in optimization

- "Predict, then smart optimize" instead of "Smart predict, then optimize"
- 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

Extensions/future work

- Distributionally robust, multi-stage, stochastic constraints
- (Partially) remove assumption on independence of ϵ and X
- Lower bounds on required sample size?

Future research: resilient power grid



- Multi-objective and multi-stage optimization model
- Distributionally robust chance constraints, robust constraints
- Accurate and tractable power flow and restoration models
- Integrate machine learning and stochastic optimization models

References I

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

Gah-Yi Ban, Jérémie Gallien, and Adam 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.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *arXiv preprint arXiv:1905.02928*, pages 1–40, 2019.

Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *arXiv preprint arXiv:1402.5481*, pages 1–20, 2019.

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

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

Yi-hao Kao, Benjamin V Roy, and Xiang Yan. Directed regression. In *Advances in Neural Information Processing Systems*, pages 889–897, 2009.

Kibaek Kim and Sanjay 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.

References II

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

Suvrajeet Sen and Yunxiao 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.