Data-Driven Multi-Stage Stochastic Optimization on Time Series

Jim Luedtke

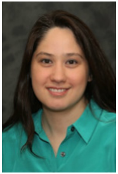
Department of Industrial and Systems Engineering
University of Wisconsin-Madison

May 21, 2022
Erice

Collaborators:

Rohit Kannan

Güzin Bayraksan

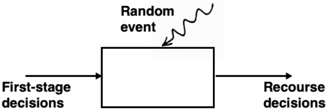


Nam Ho-Nguyen

Outline

1. Data-driven two-stage stochastic optimization
2. Distributionally Robust Extension
3. Multi-stage Stochastic Optimization

Two-Stage Stochastic Programming



- Traditional two-stage SP: minimize expected system cost *assuming* distribution of random vector Y known

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

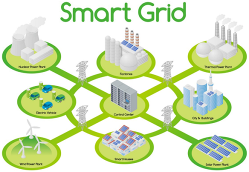
- Sample Average Approximation: given samples $\{y^i\}_{i=1}^n$ of Y

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

- SAA theory: optimal value and solutions converge as $n \to \infty$

Can we use covariates/features to better predict the random vector Y?

Stochastic Programming with Covariate Information



Power Grid Scheduling

Y: Load; Renewable energy outputs

X: Weather observations; Time/Season

z: Generator scheduling decisions

Production Planning/Scheduling

Y: Product demands; Prices

X: Seasonality; Web search results

z: Production and inventory decisions

- Given historical data on uncertain parameters and covariates

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

- When making decision z, we observe a *new* covariate $X = x$
- Goal: minimize expected cost given covariate observation x:

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

Stochastic Programming with Covariate Information

- Assume we have uncertain parameter and covariate data pairs

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

- When making decision z, we observe a *new* covariate $X = x$
- Goal: minimize expected cost given covariate observation x:

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

- Challenge: $\mathcal{D}_n$ may not include covariate observation $X = x$
- How to construct data-driven approximation to conditional SP?
  1. Learn: predict Y given $X = x$
  2. Optimize: integrate learning into optimization (with errors)
- Assume $Y = f^*(X) + Q^*(X)\varepsilon$ with X and ε *independent*

Separate Learning and Optimization

1. Use data to train our 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
- Does not yield asymptotically consistent solutions

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: 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]
[3] Bertsimas and Kallus [2020], Ban and Rudin [2018]

Empirical Residuals-based Sample Average Approximation

Approach (Sen and Deng [2018],Ban et al. [2018],Kannan et al. [2020a])

1. Use data to train our favorite ML prediction model $\Rightarrow \hat{f}_n, \hat{Q}_n$

$$\hat{f}_n(\cdot) \in \arg\min_{f(\cdot)\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\|y^i - f(x^i)\|^2$$

Compute *empirical residuals* $\hat{\varepsilon}^i_n := [\hat{Q}_n(x^i)]^{-1}\big(y^i - \hat{f}_n(x^i)\big),\ i \in [n]$

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

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

- Convergence conditions and rates: Kannan et al. [2020a]

Comparison: Nonparametric Reweighting-Based SAA

Bertsimas and Kallus [2020]

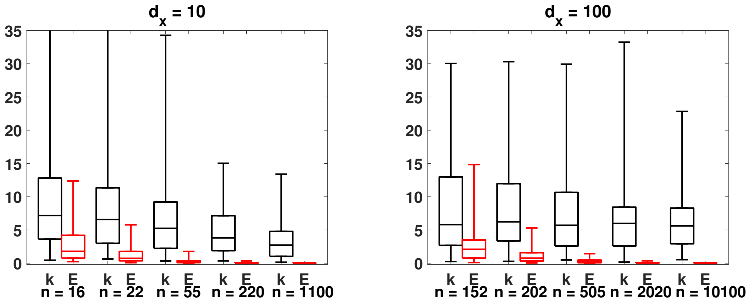
- Solve the following reweighted SAA problem

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

  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 $\dim(X)$ or $\dim(Y)$ is large

Results with Correct Model Class ($p = 1$)

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

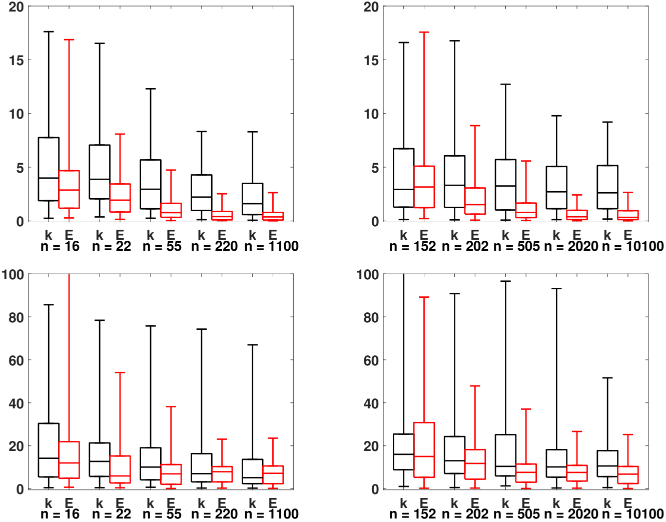


- 50 data replications × 20 covariates = 1000 experiments
- Boxes: 25 and 75 percentiles of upper confidence bounds; Whiskers: 2 and 98 percentiles

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

$\mathbf{p} = \mathbf{0.5}$

$\mathbf{p} = \mathbf{2}$



Outline

DRO Extension

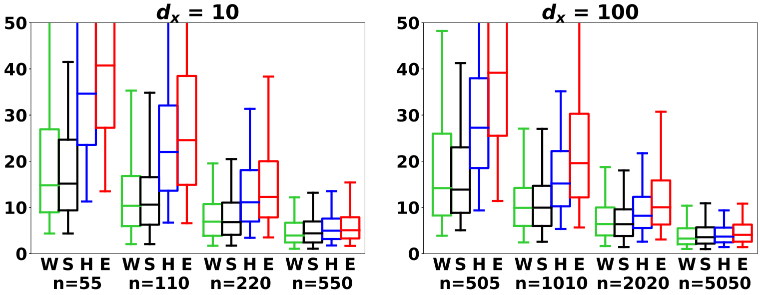
Summary of Kannan et al. [2020b]

- Solve the DRO model:

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

  where $\hat{\mathcal{P}}_n(x)$ is a data-driven ambiguity set for the distribution of Y given $X = x$ that is centered on residuals-based samples $\{\hat{f}_n(x) + \hat{Q}_n(x)\hat{\varepsilon}_n^i\}_{i=1}^n$
- Convergence shown for various forms of ambiguity set: Wasserstein/Monge, phi-divergence, sample-robust
- Key computational challenge: data-driven tuning of ambiguity set radius
  - Small data: best to use method that chooses radius indepedent of x
  - More data: obtain more consistent results by tuning radius to x

Sample Empirical DRO Results

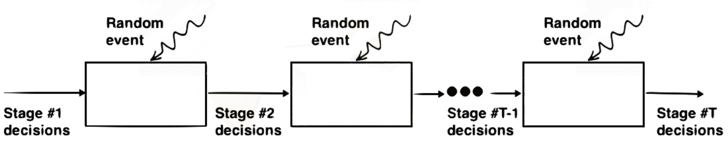


Boxes: 25 and 75 percentiles of upper confidence bounds;
Whiskers: 2 and 98 percentiles

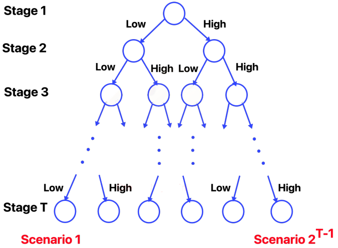
- W: Wassterstein/Monge
- S: Sample robust
- H: ϕ-divergence using Hellinger distance
- E: ER-SAA with no DRO

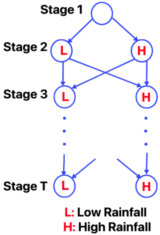
Outline

Multistage Stochastic Optimization



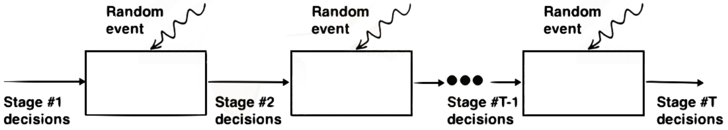
Complexity of multi-stage stochastic programs can grow significantly with the number of stages T!





Stochastic Dual Dynamic Programming (Pereira and Pinto [1991]): Exploit recombining scenario tree structure to limit number of value functions that need to be approximated.

Multistage Stochastic Optimization



- Decision Process: $z_1 \rightsquigarrow \xi_2 \rightsquigarrow z_2 \rightsquigarrow \cdots \xi_T \rightsquigarrow z_T$

At stage t, solve

$$\min_{z_t \in Z_t(z_{t-1},\xi_t)} \begin{matrix}\text{cost of decisions } z_t \\ \text{in current stage } t\end{matrix} + \begin{matrix}\text{expected cost of decisions } z_t \\ \text{in future stages given history } (\xi_1,\ldots,\xi_t)\end{matrix}$$

- Assume stationary time series model:
  $\xi_t = f^*(\xi_{t-1}) + Q^*(\xi_{t-1})\varepsilon_t$
- Goal: Given a *single historical trajectory* of $\{\xi_t\}$

$$\mathcal{D}_n := \{\tilde{\xi}_0, \tilde{\xi}_1, \cdots, \tilde{\xi}_n\}$$

estimate optimal first-stage decision z_1

Related Work and Goals

Bertsimas et al. [2021]:

- Assume given an *i.i.d. set of historical sample paths*
- Construct a robust optimization model with uncertainty sets built around sample paths
- Show asymptotic convergence to optimal solution as the *number of sample paths grows*
- Solve approximately using decision rule approximations

Other related work: Ban et al. [2018], Bertsimas and McCord [2019], Silva et al. [2021]

Our goals:

- Use *single historical sample path* (assuming time series model)
- Construct data-driven approximation that can be solved using Stochastic Dual Dynamic Programming (SDDP)
- Establish convergence as *size of sample path grows*

Problem Setup

- Given historical data from a *single trajectory* of $\{\xi_t\}$

$$\mathcal{D}_n := \{\tilde{\xi}^0, \tilde{\xi}^1, \cdots, \tilde{\xi}^n\}$$

- Want to solve

$$V_1(\xi_1) := \min_{z_1 \in Z_1(\xi_1)} f_1(z_1, \xi_1) + \mathbb{E}\left[V_2(z_1, \xi_2) \mid \xi_1\right],$$

where

$$V_t(z_{t-1}, \xi_{[t]}) := \min_{z_t \in Z_t(z_{t-1}, \xi_t)} \overbrace{f_t(z_t, \xi_t)}^{\text{stage } t \text{ cost}} + \overbrace{\mathbb{E}\left[V_{t+1}(z_t, \xi_{[t+1]}) \mid \xi_{[t]}\right]}^{\text{expected cost of future stages}}, \quad t \in [T-1],$$
$$V_T(z_{T-1}, \xi_{[T]}) := \min_{z_T \in Z_T(z_{T-1}, \xi_T)} f_T(z_T, \xi_T).$$

- Assume
  - True model: $\xi_t = f^*(\xi_{t-1}) + Q^*(\xi_{t-1})\varepsilon_t$ with i.i.d. errors $\{\varepsilon_t\}$
  - We know function classes $\mathcal{F}$, $\mathcal{Q}$ such that $f^* \in \mathcal{F}$, $Q^* \in \mathcal{Q}$

Empirical Residuals-based Sample Average Approximation

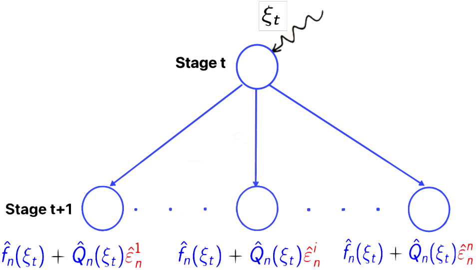
1. Estimate f^*, Q^* using our favorite ML method $\Rightarrow \hat{f}_n, \hat{Q}_n$

   Compute *empirical residuals*

$$\hat{\varepsilon}_n^i := [\hat{Q}_n(\tilde{\xi}^{i-1})]^{-1}\left(\tilde{\xi}^i - \hat{f}_n(\tilde{\xi}^{i-1})\right), \quad i \in [n]$$

2. Use $\{\hat{f}_n(\xi_t) + \hat{Q}_n(\xi_t)\hat{\varepsilon}_n^i\}_{i=1}^n$ as samples of ξ_{t+1} given ξ_t in SAA

Tailored convergence analysis required since same empirical errors $\hat{\varepsilon}_n^i$ used for all time stages



Convergence Theory

Assumptions on the multistage stochastic program:

Assumptions on the ML setup:

Asymptotic optimality

Convergence Theory

Assumptions on the multistage stochastic program:

- Can always take recourse decisions to keep system feasible
- The feasible region Z_t for each stage t is bounded

Assumptions on the ML setup:

Asymptotic optimality

Convergence Theory

Assumptions on the multistage stochastic program:

- Can always take recourse decisions to keep system feasible
- The feasible region Z_t for each stage t is bounded

Assumptions on the ML setup:

- The functions f^* and Q^* are Lipschitz continuous
- $\hat{f}_n \to f^*$ and $\hat{Q}_n \to Q^*$ uniformly on their domains

Asymptotic optimality

Convergence Theory

Assumptions on the multistage stochastic program:

- Can always take recourse decisions to keep system feasible
- The feasible region Z_t for each stage t is bounded

Assumptions on the ML setup:

- The functions f^* and Q^* are Lipschitz continuous
- $\hat{f}_n \to f^*$ and $\hat{Q}_n \to Q^*$ uniformly on their domains

Asymptotic optimality

Under above assumptions, as the historical sample size n increases, any first-stage ER-SAA solution converges to an optimal solution of the true multistage stochastic program

Convergence Theory

Result holds with these weaker assumptions on the ML setup:

- The functions f^*, $\hat{f}_n$, Q^*, and $\hat{Q}_n$ are Lipschitz continuous
- Mean-squared estimation error consistency:

$$\frac{1}{n}\sum_{i\in[n]}\|f^*(\tilde{\xi}^{i-1}) - \hat{f}_n(\tilde{\xi}^{i-1})\|^2 \xrightarrow{p} 0,$$

$$\frac{1}{n}\sum_{i\in[n]}\left\|\left[Q^*(\tilde{\xi}^{i-1})\right]^{-1} - \left[\hat{Q}_n(\tilde{\xi}^{i-1})\right]^{-1}\right\|^2 \xrightarrow{p} 0$$

- For each $t \in [T-1]$:

$$\mathbb{E}_{\varepsilon_t\sim P_n}\left[\|f^*(\xi_t) - \hat{f}_n(\xi_t)\| \middle| \xi_1\right] \xrightarrow{p} 0,$$

$$\mathbb{E}_{\varepsilon_t\sim P_n}\left[\|Q^*(\xi_t) - \hat{Q}_n(\xi_t)\| \middle| \xi_1\right] \xrightarrow{p} 0$$

$P_n := \frac{1}{n}\sum_{i\in[n]}\delta_{\tilde{\varepsilon}^i}$ is the true empirical distribution of errors

These assumptions can be readily verified for linear vector auto-regressive processes

Rates of Convergence

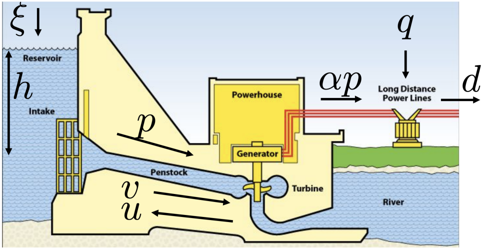
Assume

- The errors $\{\varepsilon_t\}$ obey a light-tailed distribution
- The true multistage stochastic program satisfies assumptions required for SAA convergence rate analysis (e.g., Shapiro et al. [2009])
- The regression estimates $\hat{f}_n$ and $\hat{Q}_n$ satisfy large deviation properties

Rates of convergence of regression estimates dictate rates of convergence of ER-SAA solutions

- For parametric time series models, rate of convergence of ER-SAA equals rate of convergence of classical SAA

Numerical Experiments: Hydrothermal Scheduling



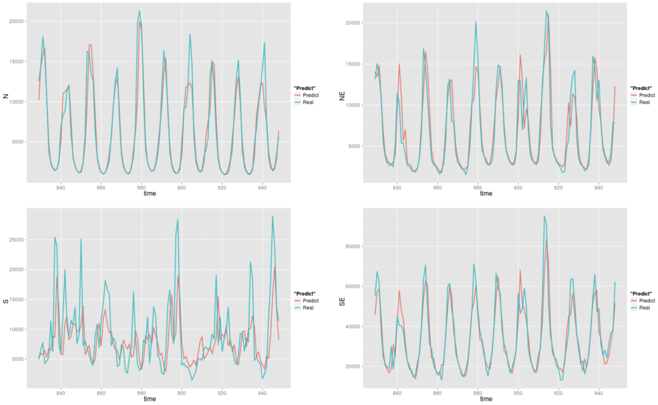
- Decisions z_t: Hydrothermal & natural gas generation, spillage
- Random vector ξ: Amount of rainfall

Numerical Experiments: Hydrothermal Scheduling

Assume true time series model for rainfall is of the form

$$\xi_t = (\alpha_t^* + \beta_t^* \xi_{t-1}) \exp(\varepsilon_t),$$

where $\alpha_t^* = \alpha_{t+12}^*, \quad \beta_t^* = \beta_{t+12}^*, \quad \varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \Sigma)$



Good fit to historical data over 8 decades! (Shapiro et al. [2012])

Numerical Experiments: Hydrothermal Scheduling

- Consider the Brazilian interconnected power system with four hydrothermal reservoirs
- Generate a sample trajectory of $\{\xi_t\}$ using time series model

$$\xi_t = (\alpha_t^* + \beta_t^* \xi_{t-1}) \exp(\varepsilon_t),$$

  where $\alpha_t^* = \alpha_{t+12}^*, \quad \beta_t^* = \beta_{t+12}^*, \quad \varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \Sigma)$

- Estimate coefficients $(\hat{\alpha}_t, \hat{\beta}_t)$ such that

$$\hat{\alpha}_t = \hat{\alpha}_{t+12}, \quad \hat{\beta}_t = \hat{\beta}_{t+12}$$

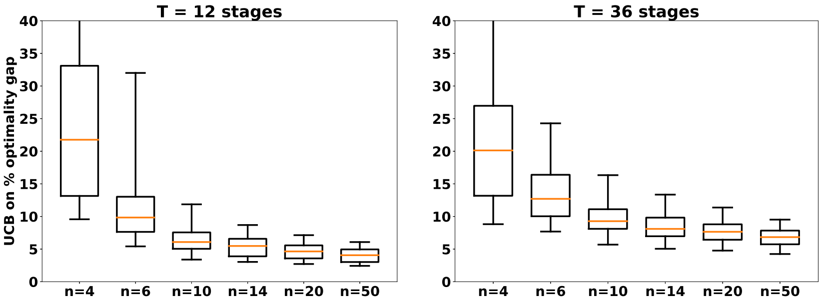
  Use these to estimate samples of the errors ε_t

- Solve the ER-SAA model using `SDDP.jl` Dowson and Kapelevich [2021].
  Estimate sub-optimality of ER-SAA solutions

Results When the Time Series Model is Correctly Specified

Estimate true heteroscedastic model: $\xi_t = (\alpha_t^* + \beta_t^* \xi_{t-1}) \exp(\varepsilon_t)$

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



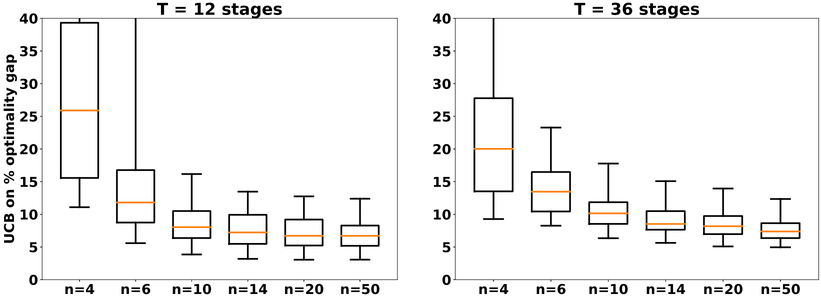
n: *years* of historical data (observations $= 12n$)

Boxes: 25, 50, and 75 percentiles of optimality gap estimates;
Whiskers: 5 and 95 percentiles

Results When the Time Series Model is Misspecified

Estimate seasonal additive error model: $\xi_t = \alpha_t^* + \beta_t^* \xi_{t-1} + \varepsilon_t$

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



n: number of historical samples *per month*

Boxes: 25, 50, and 75 percentiles of optimality gap estimates;
Whiskers: 5 and 95 percentiles

Concluding Remarks

ER-SAA: a modular approach to using covariate information in optimization under uncertainty

- DRO extension yields improved results in low data regime
- Multi-stage extension solvable using *Stochastic Dual Dynamic Programming*

Future research directions

- Robust multistage
- Discrete recourse decisions
- Possible for optimal value convergence rates to improve over prediction rate "limits"?

Questions? `jim.luedtke@wisc.edu`

Kannan, Bayraksan, and L. Data-Driven SAA With Covariate Information. Available on Optimization Online

Kannan, Bayraksan, and L. Residuals-Based DRO With Covariate Information. arXiv:2012.01088

Kannan, Ho-Nguyen, and L. Data-Driven Multistage Stochastic Optimization on Time Series. Working Paper

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.

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

Dimitris Bertsimas and Christopher McCord. From predictions to prescriptions in multistage optimization problems. *arXiv preprint arXiv:1904.11637*, pages 1–38, 2019.

Dimitris Bertsimas, Shimrit Shtern, and Bradley Sturt. A data-driven approach to multi-stage stochastic linear optimization. *Management Science (Forthcoming)*, 2021.

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.

Oscar Dowson and Lea Kapelevich. Sddp.jl: a julia package for stochastic dual dynamic programming. *INFORMS Journal on Computing*, 33(1):27–33, 2021.

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

References II

Rohit Kannan, Güzin Bayraksan, and James R Luedtke. Data-driven sample average approximation with covariate information. *Optimization Online*, 2020a.

Rohit Kannan, Güzin Bayraksan, and James R Luedtke. Residuals-based distributionally robust optimization with covariate information, 2020b. arXiv:2012.01088.

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

Mario VF Pereira and Leontina MVG Pinto. Multi-stage stochastic optimization applied to energy planning. *Mathematical programming*, 52(1):359–375, 1991.

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.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2009.

Alexander Shapiro, Wajdi Tekaya, J Paulo da Costa, and M Pereira Soares. Final report for technical cooperation between georgia institute of technology and ons–operador nacional do sistema elétrico. *Georgia Tech ISyE Report*, 2012.

Thuener Silva, Davi Valladão, and Tito Homem-de Mello. A data-driven approach for a class of stochastic dynamic optimization problems. *Computational Optimization and Applications*, 80(3):687–729, 2021.