Data-driven Sample Average Approximation with Covariate Information

Güzin Bayraksan

Department of Integrated Systems Engineering
The Ohio State University

September 30, 2021

Joint work with Rohit Kannan and James R. Luedtke

CRM Optimization under Uncertainty Workshop, Montréal
Grateful to . . .

Rohit Kannan  
Los Alamos National Lab

Jim Luedtke  
University of Wisconsin-Madison

Gratefully acknowledge support of DOE Grant DE-AC02-06CH11357 (MACSER)
Introduction and Motivation

Empirical Residuals SAA
- Theoretical Analysis
- Computational Results

Distributionally Robust ER-SAA
- Theoretical Analysis
- Computational Results

Conclusions and Future Work
Traditional Stochastic Programs

Traditional Stochastic Programs consider

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

where

- $Y$ denotes the uncertain model parameters (a random vector),
- $\mathcal{Z}$ denotes the feasible region for decisions $z$. 

Suppose we have access to (i.i.d.) observations $\{y_i\}_{i=1}^n$ of $Y$.

Then, the Sample Average Approximation (SAA) is given by

$$\min_{z \in \mathcal{Z}} \frac{1}{n} \sum_{i=1}^{n} c(z, y_i).$$

SAA framework has well established theory: consistency, rates of convergence, ... (Shapiro et al., 2009) (Homem-de Mello and Bayraksan, 2014)
Traditional Stochastic Programs

- Traditional Stochastic Programs consider
  \[
  \min_{z \in \mathcal{Z}} \mathbb{E}[c(z, Y)],
  \]
  where
  - \(Y\) denotes the uncertain model parameters (a random vector),
  - \(\mathcal{Z}\) denotes the feasible region for decisions \(z\).

- Suppose we have access to (i.i.d.) observations \(\{y^i\}_{i=1}^n\) of \(Y\).
Traditional Stochastic Programs

- Traditional Stochastic Programs consider

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

where

- \(Y\) denotes the uncertain model parameters (a random vector),
- \(\mathcal{Z}\) denotes the feasible region for decisions \(z\).

- Suppose we have access to (i.i.d.) observations \(\{y^i\}_{i=1}^n\) of \(Y\).

- Then, the **Sample Average Approximation (SAA)** is given by

\[
\min_{z \in \mathcal{Z}} \frac{1}{n} \sum_{i=1}^n c(z, y^i).
\]

SAA framework has well established theory: consistency, rates of convergence, ... (Shapiro et al., 2009) (Homem-de Mello and Bayraksan, 2014)
Stochastic Programs with Covariate Information

Given a covariate realization $X = x$, update the Stochastic Program to

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

(Covariates) are also known as side information, or auxiliary variables, or features

( Bertsimas and Kallus, 2019)
Example Applications

Portfolio Optimization
(Bazier-Matte and Delage, 2020)

$Y$: Return of stocks
$X$: Market indicators; Company data
$z$: Portfolio decisions

Power Grid Scheduling
(Donti et al., 2017)

$Y$: Load; Renewal Energy Outputs
$X$: Weather forecast; Time/Day
$z$: Generator scheduling decisions
Suppose we have data of form (not necessarily i.i.d.)

\[ D_n := \{(y^i, x^i)\}_{i=1}^n \]

(uncertain parameters, and covariates)
How to Form SAA with Covariate Information?

- Suppose we have data of form (not necessarily i.i.d.)

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

(uncertain parameters, and covariates)

- When making decision \( z \), we observe a **new** covariate \( x \)
How to Form SAA with Covariate Information?

- Suppose we have data of form (not necessarily i.i.d.)
  \[ \mathcal{D}_n := \{(y^i, x^i)\}_{i=1}^n \]
  (uncertain parameters, and covariates)

- When making decision \( z \), we observe a new covariate \( x \)

- How can we form SAA with this data? Two Components:
  1. "Learn" — Predict \( Y \) given \( X = x \)
  2. "Optimize" — Integrate Learning into Optimization (with errors)
Integrated Learning and Optimization

Approach 1: Empirical Risk Minimization (ERM) [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
Introduction and Motivation

Integrated Learning and Optimization

Approach 1: Empirical Risk Minimization (ERM) [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

Approach 2: Modify the Optimization Step (this work)

- Change optimization model to reflect uncertainty in prediction
- Ban et al. (2018), Sen and Deng (2018), Bertsimas and Kallus (2019)
Integrated Learning and Optimization

Approach 1: Empirical Risk Minimization (ERM) [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

Approach 2: Modify the Optimization Step (this work)

- Change optimization model to reflect uncertainty in prediction
- Ban et al. (2018), Sen and Deng (2018), Bertsimas and Kallus (2019)

Approach 3: Modify the Learning Step

- Change loss function in training step to reflect use of prediction in optimization model (Donti et al. (2017), Elmachtoub and Grigas (2021))
- Results in a challenging training problem; Can be less modular but lower cost
Outline

1. Introduction and Motivation

2. Empirical Residuals SAA
   ▶ Theoretical Analysis
   ▶ Computational Results

3. Distributionally Robust ER-SAA
   ▶ Theoretical Analysis
   ▶ Computational Results

4. Conclusions and Future Work
“True” Relationship Between \( Y \) and \( X \)

- Assume

\[
Y|(X = x) = f^*(x) + \varepsilon,
\]

where

- \( f^*(x) := \mathbb{E}[Y \mid X = x] \) is the regression function
- \( f^* \) belongs to a known class of functions \( \mathcal{F} \) (can be infinite dimensional or depend on \( n \))
- \( \varepsilon \) is the associated regression error
- \( \varepsilon \) are independent of the covariates \( X \)
- \( \mathbb{E}[\varepsilon] = 0 \)
“True” Stochastic Program and SAA

- Under this structure, the “true” conditional stochastic program (SP) is equivalent to

\[
\min_{z \in \mathcal{Z}} \mathbb{E} [c(z, Y) \mid X = x]\tag{1}
\]
“True” Stochastic Program and SAA

Under this structure, the “true” conditional stochastic program (SP) is equivalent to

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

(1)
“True” Stochastic Program and SAA

- Under this structure, the “true” conditional stochastic program (SP) is equivalent to

\[
\min_{z \in \mathcal{Z}} \mathbb{E} \left[ c(z, f^*(x) + \varepsilon) \right]
\]  

(1)

- Given data \( \mathcal{D}_n := \{(y^i, x^i)\}_{i=1}^n \) (not necessarily i.i.d.) and errors

\[
\varepsilon^i := y^i - f^*(x^i), \quad \forall i \in \{1, \cdots, n\}
\]

- We can form **Full-Information SAA**

\[
\min_{z \in \mathcal{Z}} \frac{1}{n} \sum_{i=1}^n c(z, f^*(x) + \varepsilon^i)
\]

(2)
Under this structure, the “true” conditional stochastic program (SP) is equivalent to

$$\min_{z \in Z} \mathbb{E} \left[ c(z, f^*(x) + \varepsilon) \right]$$ (1)

Given data \( D_n := \{(y^i, x^i)\}_{i=1}^n \) (not necessarily i.i.d.) and errors

\[ \varepsilon^i := y^i - f^*(x^i), \quad \forall i \in \{1, \cdots, n\} \]

We can form Full-Information SAA

$$\min_{z \in Z} \frac{1}{n} \sum_{i=1}^{n} c(z, f^*(x) + \varepsilon^i)$$ (2)
Empirical Residuals-Based SAA

Approach suggested by Sen and Deng (2018) and Ban et al. (2018) is

1. **Estimate** $f^*$ using your favorite Statistical / Machine Learning (ML) model $\Rightarrow \hat{f}_n$

   *Example:*

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

   with some loss function $\ell : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \rightarrow \mathbb{R}_+$
2. **Compute empirical residuals**

\[
\hat{\varepsilon}_n^i := y^i - \hat{f}_n(x^i), \quad i = 1, \ldots, n
\]

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

\[
\min_{z \in Z} \frac{1}{n} \sum_{i=1}^n c(z, \hat{f}_n(x) + \hat{\varepsilon}_n^i)
\]  

(ER-SAA)
Empirical Residuals-Based SAA

2. Compute empirical residuals

\[ \hat{\varepsilon}_n^i := y^i - \hat{f}_n(x^i), \quad i = 1, \ldots, n \]

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

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

▶ General analysis

▶ Improvements when sample size is small?
Theoretical Analysis of ER-SAA
Sequence of Problems & Notation

- **“True” conditional stochastic program**
  \[
  \nu^*(x) := \min_{z \in \mathcal{Z}} \{ g(z; x) := \mathbb{E} [c(z, f^*(x) + \varepsilon)] \} \tag{3}
  \]

- **Full-Information SAA**
  \[
  \min_{z \in \mathcal{Z}} \left\{ g_n^*(z; x) := \frac{1}{n} \sum_{i=1}^{n} c(z, f^*(x) + \varepsilon^i) \right\} \tag{4}
  \]

- **Empirical Residuals-Based SAA (ER-SAA)**
  \[
  \hat{\nu}_n^{ER}(x) := \min_{z \in \mathcal{Z}} \left\{ \hat{g}_n^{ER}(z; x) := \frac{1}{n} \sum_{i=1}^{n} c\left(z, \hat{f}_n(x) + \hat{\varepsilon}_n^i \right) \right\} \tag{5}
  \]
Sequence of Problems & Notation

- **“True” conditional stochastic program**
  \[\nu^*(x) := \min_{z \in Z} \{g(z; x) := \mathbb{E}[c(z, f^*(x) + \epsilon)]\} \quad (3)\]

- **Full-Information SAA**
  \[
  \min_{z \in Z} \left\{ g_n^*(z; x) := \frac{1}{n} \sum_{i=1}^{n} c(z, f^*(x) + \epsilon^i) \right\} \quad (4)
  \]

- **Empirical Residuals-Based SAA (ER-SAA)**
  \[
  \hat{\nu}_n^{ER}(x) := \min_{z \in Z} \left\{ \hat{g}_n^{ER}(z; x) := \frac{1}{n} \sum_{i=1}^{n} c \left( z, \hat{f}_n(x) + \hat{\epsilon}_n^i \right) \right\} \quad (5)
  \]
Sequence of Problems & Notation

Notation regarding optimal solutions

- \( S^*(x) \) — set of optimal solutions to “true” problem
- \( \hat{S}^{ER}_n(x) \in \hat{S}^{ER}_n(x) \) — an optimal solution \( \in \) set of optimal solutions to (ER-SAA)

Assume \( S^*(x) \) and \( \hat{S}^{ER}_n(x) \) are nonempty for almost every (a.e.) \( x \in \mathcal{X} \)
Consistency and Asymptotic Optimality:
Assumption on Problem Structure

Assumption

For each \( z \in \mathcal{Z} \), the function \( c \) in problem (3) satisfies the Lipschitz condition

\[
|c(z, \bar{y}) - c(z, y)| \leq L(z) \|\bar{y} - y\|, \quad \forall y, \bar{y} \in \mathbb{R}^{d_y},
\]

with Lipschitz constant \( L \) satisfying \( \sup_{z \in \mathcal{Z}} L(z) < +\infty \).

Can be Satisfied by Two-Stage Stochastic Linear Programs

Note: This assumption can be relaxed to local Lipschitz continuity with additional conditions on the regression step.
Consistency and Asymptotic Optimality:

Assumption on Uniform Convergence of Full-Information SAA Objective Functions to True Problem Objective Function

Assumption

For a.e. \( x \in \mathcal{X} \), the sequence of sample average functions \( \{g_n^*(\cdot; x)\} \) defined in (4) converges in probability to the true function \( g(\cdot; x) \) defined in (3) uniformly on the set \( \mathcal{Z} \).

Follows under conditions stipulated in classical SAA analysis
Consistency and Asymptotic Optimality: 
Assumption on Learning Step

Assumption

The regression/learning procedure satisfies the following consistency properties:

1. **Pointwise error consistency**: \( \hat{f}_n(x) \xrightarrow{p} f^*(x) \) for a.e. \( x \in \mathcal{X} \),

2. **Mean-squared estimation error consistency**:
\[
\frac{1}{n} \sum_{i=1}^{n} \| f^*(x^i) - \hat{f}_n(x^i) \|^2 \xrightarrow{p} 0,
\]

where \( \xrightarrow{p} \) denotes convergence in probability.

Can hold under appropriate conditions for Ordinary Least Squares (OLS), k Nearest Neighbor (kNN) regression, and Random Forests (RF), . . .
Consistency and Asymptotic Optimality

**Theorem**

*Under the above assumptions, we have*

1. $\hat{v}_n^{ER}(x) \xrightarrow{p} v^*(x)$,

2. $\mathbb{D}\left(\hat{S}_n^{ER}(x), S^*(x)\right) \xrightarrow{p} 0$, and

3. $\sup_{z \in \hat{S}_n^{ER}(x)} g(z; x) \xrightarrow{p} v^*(x)$ for a.e. $x \in \mathcal{X}$,

where $\xrightarrow{p}$ denotes convergence in probability, and for sets $A, B \subseteq \mathbb{R}^{d_z}$, let $\mathbb{D}(A, B) := \sup_{v \in A} \text{dist}(v, B)$ denote the deviation of $A$ from $B$, where $\text{dist}(v, B) := \inf_{w \in B} \|v - w\|$.
RATES OF CONVERGENCE

Similar results can be obtained on rate of convergence by strengthening the assumptions.
Rates of Convergence

Similar results can be obtained on rate of convergence by strengthening the assumptions

Assumption

There is a constant $0 < \alpha \leq 1$ (that is independent of the number of samples $n$, but could depend on the dimension $d_x$ of the covariates $X$) such that the regression procedure satisfies the following asymptotic convergence rate criteria:

1. **Pointwise error rate**: $\|f^*(x) - \hat{f}_n(x)\|^2 = O_p(n^{-\alpha})$ for a.e. $x \in \mathcal{X}$,

2. **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 $\alpha = 1$

kNN, RF satisfy assumption with $\alpha = O(1)$

Kannan, Luedtke & Bayraksan

**Data-Driven SAA with Covariates**

CRM OUU Workshop, Montréal
Rates of Convergence

Similar results can be obtained on rate of convergence by strengthening the assumptions.

Assumption

There is a constant $0 < \alpha \leq 1$ (that is independent of the number of samples $n$, but could depend on the dimension $d_x$ of the covariates $X$) such that the regression procedure satisfies the following asymptotic convergence rate criteria:

1. **Pointwise error rate**: $\|f^*(x) - \hat{f}_n(x)\|^2 = O_p(n^{-\alpha})$ for a.e. $x \in \mathcal{X}$,

2. **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 $\alpha = 1$

kNN, RF satisfy assumption with $\alpha = \frac{O(1)}{d_x}$
Theorem

Under the above assumptions, plus continuity assumptions and a functional Central Limit Theorem assumption, we have

$$\hat{v}_n^{ER}(x) = v^*(x) + \tilde{o}_p\left(n^{-\frac{\alpha}{2}}\right) \quad \text{for a.e. } x \in \mathcal{X},$$

where $\tilde{o}$ notation hides polylogarithmic factors in $n$.  

Rates of Convergence
Finite Sample Guarantees

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

  Recall that
  - $\hat{S}_{n}^{ER}(x)$ — set of optimal solutions to ER-SAA at covariate value $x$
  - $S^{\kappa}(x)$ — set of $\kappa$-optimal solutions to the “True” problem at $x$

- **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 approximation are $\kappa$-optimal to the true problem with probability $\geq 1 - \delta$**
Finite Sample Guarantees

Two effects:

1. Sample size required for “Full-Information” SAA to be close to “True” Problem

\[
n \geq \frac{O(1)\sigma_c^2(x)}{\kappa^2} \left[ d_z \log \left( \frac{O(1)D}{\kappa} \right) + \log \left( \frac{O(1)}{\delta} \right) \right] \left\uparrow \right._{N_C}
\]

from classical SAA analysis.

2. Sample size required for ER-SAA to be close to “Full-Information” SAA

[Note: For brevity, some variance terms are omitted throughout this section]
Finite Sample Guarantees

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

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

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

$$n \geq N_C + \frac{O(1)}{\kappa^2} d_y \left[ s \log \left( \frac{O(1)}{\delta} \right) + s \log(d_x) \right]$$
If $f^*$ is **Lipschitz continuous** and we use kNN regression with $k = \lceil O(1)n^\gamma \rceil$ for some constant $\gamma \in (0, 1)$, then holds if

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

$$+ \left( \frac{O(1)d_y}{\kappa^2} \right)^{d_x} \left[ \frac{d_x}{2} \log \left( \frac{O(1)d_x d_y}{\kappa^2} \right) + \log \left( \frac{O(1)}{\delta} \right) \right]$$
Finite Sample Guarantees: What does it all mean?

- **Prediction** of the regression function $f^*$ introduces additional terms that depend on the dimensions $d_y$ and $d_x$ of the random vector $Y$ and the covariates $X$.

- **Assuming that** the regression function $f^*$ satisfies the necessary structural properties, using **OLS regression or the Lasso** for the regression step can yield sample size estimates that depend modestly on the accuracy $\kappa$ and the dimensions $d_x$ and $d_y$.

- The sample size estimates for **kNN regression** are valid under mild assumptions on the regression function $f^*$, but more heavily dependent on $d_x$ and $d_y$.
Computational Results on ER-SAA
Two-Stage Resource Allocation Model (Luedtke, 2014)

- Meet demands of $|\mathcal{J}| = 30$ customers using 20 resources
- Uncertain demands $Y$ generated according to

$$Y_j = \varphi_j^* + \sum_{l \in \mathcal{L}^*} \zeta_{jl}^* \cdot (X_l)^\theta + \varepsilon_j, \quad \forall j \in \mathcal{J},$$

where

- $\theta \in \{0.5, 1, 2\}$ is a fixed parameter that determines the model class,
- $\varepsilon_j \sim \mathcal{N}(0, \sigma_j^2)$ is an additive error, and
- Covariate dimension $d_x \in \{3, 10, 100\}$
- $|\mathcal{L}^*| = 3$, i.e., the demands truly depend only on 3 covariates
Computational Setup

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

  \[ Y_j = \varphi_j + \sum_{l \in L} \zeta_{jl} \cdot X_l + \eta_j, \quad \forall j \in \mathcal{J}, \]

  where $\eta_j$ are zero-mean errors, using OLS or Lasso regression

- Estimate optimality gap of solutions $\hat{z}_{ER}^n(x)$

  \[ \hat{z}_{ER}^n(x) \in \arg \min_{z \in Z} \frac{1}{n} \sum_{i=1}^{n} c(z, \hat{f}_n(x) + \hat{\varepsilon}_n) \]
Effect of Varying Covariate Dimension:

with correct ($\theta = 1$) or incorrect ($\theta \neq 1$) regression fit
Results with correct model class ($\theta = 1$)

Red (E): ER-SAA + OLS
Black (k): Reweighted SAA with kNN (Bertsimas and Kallus, 2019)

Boxes: 25 and 75 percentiles of Upper Confidence Bounds (UCB)
Whiskers: 2 and 98 percentiles
Results with misspecified model class ($\theta \neq 1$)

$p = 0.5$

$p = 2$
Effect of Prediction Model:

with correct ($\theta = 1$) or incorrect ($\theta \neq 1$) regression fit
Effect of Prediction Model ($\theta = 1$)

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

![Graph for $d_x = 10$](image1)

![Graph for $d_x = 100$](image2)
Effect of Prediction Model with misspecified model ($\theta \neq 1$)

$p = 0.5$

$p = 2$
Outline

1. Introduction and Motivation

2. Empirical Residuals SAA
   - Theoretical Analysis
   - Computational Results

3. Distributionally Robust ER-SAA
   - Theoretical Analysis
   - Computational Results

4. Conclusions and Future Work
Empirical Residuals-Based DRO

DRO can regularize small-sample ER-SAA, yielding solutions with better out-of-sample performance
Empirical Residuals-Based DRO

DRO can regularize small-sample ER-SAA, yielding solutions with better out-of-sample performance.

Given “estimated” empirical distribution

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

Consider Empirical Residuals-Based DRO

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

where \( \hat{P}_n(x) \) is an ambiguity set containing \( \hat{P}_n^{ER}(x) \).
How to Form the Ambiguity Set $\hat{P}_n(x)$

Let $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ denote the support of $Y$

1. **Wasserstein-$p$, $p \in [1, +\infty]$** (Esfahani and Kuhn, 2018)

\[
\hat{P}_n(x) = \{ Q \in \mathcal{P}(\mathcal{Y}) : d_{W,p}(Q, \hat{P}_n^{ER}(x)) \leq \zeta_n(x) \}.
\]
How to Form the Ambiguity Set $\hat{P}_n(x)$

Let $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ denote the support of $Y$

1. **Wasserstein-$p$, $p \in [1, +\infty]$**  
   
   \[ \hat{P}_n(x) = \left\{ Q \in \mathcal{P}(\mathcal{Y}) : d_{W,p}(Q, \hat{P}_n^{ER}(x)) \leq \zeta_n(x) \right\}. \]

2. **Sample-robust**  
   
   \[ \hat{P}_n(x) = \left\{ Q = \frac{1}{n} \sum_{i=1}^{n} \delta_{\bar{y}^i} : \|\bar{y}^i - (\hat{f}_n(x) + \hat{\epsilon}_n^i)\| \leq \mu_n(x), \bar{y}^i \in \mathcal{Y}, \forall i \right\}. \]
How to Form the Ambiguity Set $\hat{P}_n(x)$

Let $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ denote the support of $Y$

1. **Wasserstein-$p$, $p \in [1, +\infty]$** (Esfahani and Kuhn, 2018)

   $$\hat{P}_n(x) = \left\{ Q \in \mathcal{P}(\mathcal{Y}) : d_{W,p}(Q, \hat{P}_n^{ER}(x)) \leq \zeta_n(x) \right\}.$$

2. **Sample-robust** (Bertsimas, Shtern, and Sturt, 2018)

   $$\hat{P}_n(x) = \left\{ Q = \frac{1}{n} \sum_{i=1}^{n} \delta_{\tilde{y}^i} : \|\tilde{y}^i - (\hat{f}_n(x) + \hat{\varepsilon}_n^i)\| \leq \mu_n(x), \tilde{y}^i \in \mathcal{Y}, \forall i \right\}.$$

3. **$\phi$-Divergence (or other) with the same support as $\hat{P}_n^{ER}(x)$**

   Kullback-Leibler divergence, Variation distance, Hellinger distance, ...
Some Theoretical Results on DRO-ER-SAA
Let us focus on Wasserstein-$p$, $p \in [1, +\infty)$.

Given $Q_1, Q_2 \in \mathcal{P}(S)$, let $\Pi(Q_1, Q_2)$ denote the set of joint distributions with marginals $Q_1$ and $Q_2$. The $p$-Wasserstein distance $d_{W,p}(Q_1, Q_2)$ between $Q_1$ and $Q_2$ with respect to the $\ell_2$-norm $\| \cdot \|$ is given by

$$d_{W,p}(Q_1, Q_2) := \left( \inf_{\pi \in \Pi(Q_1, Q_2)} \int_{S^2} \| y_1 - y_2 \|^p d\pi(y_1, y_2) \right)^{1/p}$$

\[1\] Our results can be extended to Wasserstein distances defined using $\ell_q$-norms with $q \neq 2$. 
Preliminaries – Notation

Optimal value

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

Optimal solution

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

Note: Some details will be skipped for brevity. For precise statements, please see arXiv:2012.01088
Toward 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

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

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

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

and

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

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 $\kappa_{2,n}^2(\alpha, x) = O(n^{-1} \log(\alpha^{-1}))^{O(1)/d_x}$
Toward Convergence Theory for Wasserstein ER-DRO

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

\[
\zeta_n(\alpha, x) := 2\kappa_{p,n}(\alpha, x) + \bar{\kappa}_{p,n}(\alpha)
\]

as the radius of the Wasserstein ambiguity set, where

\[
\bar{\kappa}_{p,n}(\alpha) = \text{traditional Wasserstein DRO radius that is used if we know } f^*(\text{Kuhn et al., 2019})
\]

**Note:** Radius guarantees

\[
P\left\{ d_W(\hat{P}_{n}^{ER}(x), P_{Y|X=x}) > \zeta_n(\alpha, x) \right\} \leq \alpha
\]
Wasserstein ER-DRO Results: Finite Sample Certificate Guarantee

The optimal value \( \hat{v}_n^{DRO}(x) \) provides the following certificate on the out-of-sample cost of \( \hat{z}_n^{DRO}(x) \):

“Informal” Theorem [Finite Sample Certificate Guarantee]

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

\[
P \left\{ \mathbb{E}_\varepsilon \left[ c(\hat{z}_n^{DRO}(x), f^*(x) + \varepsilon) \right] \leq \hat{v}_n^{DRO}(x) \right\} \geq 1 - \alpha
\]
Wasserstein ER-DRO Results: Rate of Convergence

“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, under appropriate assumptions, the sequence \( \{\hat{z}_n^{DRO}(x)\} \) of solutions satisfies

\[
\hat{v}_n^{DRO}(x) = v^*(x) + O_p(\zeta_n(\alpha_n, x))
\]

\[
E[\epsilon \cdot c(\hat{z}_n^{DRO}(x), f^*(x) + \epsilon)] = v^*(x) + O_p(\zeta_n(\alpha_n, x))
\]

where \( v^*(x) \) is the optimal value of the true conditional stochastic program (SP).
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 $\zeta_n$
  - **Approach 2**: Use the data $\mathcal{D}_n$ (including covariates) to choose $\zeta_n$ independently of the new covariate realization $X = x$
  - **Approach 3**: Use both the data $\mathcal{D}_n$ and the new covariate realization $X = x$ to choose the radius $\zeta_n(x)$

- Approach 3 is more data intensive than Approaches 1 & 2
Computational Results on DRO-ER-SAA
Numerical Study: Mean-Risk Portfolio Optimization

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

where \( Z := \{ z \in \mathbb{R}^d_+ : \sum_i z_i = 1 \} \).

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

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

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

where $\bar{\epsilon}_j \sim \mathcal{N}(0, 0.02j)$, $\omega \sim \mathcal{N}(0, 0.02)$, $\theta \in \{0.5, 1, 2\}$, $\text{dim}(X) \in \{10, 100\}$.
Numerical Study: Mean-Risk Portfolio Optimization

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

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

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

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

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

where $\eta_j$ are zero-mean errors

- Estimate optimality gap of solutions $\hat{z}_{nER}^*(x)$ and $\hat{z}_{nDRO}^*(x)$
Results with OLS and Correct Model Class ($\theta = 1$)

$\mathbf{I}^*$: Ideal Wasserstein radius (only for benchmarking)

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

$\mathbf{E}$: 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: $\{1.5, 2, 3, 5\} \times (\text{dim}(X) + 1)$
Results with OLS and Misspecified Model Class ($\theta \neq 1$)

$d_x = 10$

$
\begin{align*}
\theta &= 0.5 \\
\theta &= 2
\end{align*}$

$d_x = 100$
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: $\{0.5, 0.8, 1.2, 1.5\} \times (\dim(X) + 1)$
Conclusions and Future Work

Outline

1. Introduction and Motivation

2. Empirical Residuals SAA
   ▶ Theoretical Analysis
   ▶ Computational Results

3. Distributionally Robust ER-SAA
   ▶ Theoretical Analysis
   ▶ Computational Results

4. Conclusions and Future Work
Extension: Handling Heteroscedastic Errors
(arXiv:2101.03139)

- **Assumption thus far**: true model is $Y = f^*(X) + \varepsilon$ with errors $\varepsilon$ 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 $X, \varepsilon$ independent
  - Estimate $f^*$ and $Q^*$ $\implies$ estimate samples of $\varepsilon$
  - Theoretical results for ER-SAA and ER-DRO readily generalize

Thanks to Erick Delage
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
- DRO variant outperforms with limited data, benefit diminishes with increased 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
- DRO variant outperforms with limited data, benefit diminishes with increased data

Future research directions

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

(bayraksan.1@osu.edu)

Preprints available at

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

DRO-ER-SAA:
https://arxiv.org/pdf/2012.01088.pdf
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.

T. Bazier-Matte and E. Delage. Generalization bounds for regularized portfolio selection with market side information. *INFOR: Information Systems and Operational Research*, pages 1–28, 2020.

D. Bertsimas and N. Kallus. From predictive to prescriptive analytics. *Management Science*, pages 1–20, 2019.

D. Bertsimas, S. Shtern, and B. Sturt. A data-driven approach for multi-stage linear optimization. *Optimization Online. URL: http://www.optimization-online.org/DB_FILE/2018/11/6907.pdf*, 2018.
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”. *Management Science*, pages 1–18, 2021. Published in Articles in Advance.

P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2): 115–166, 2018.

T. Homem-de Mello and G. Bayraksan. Monte carlo sampling-based methods for stochastic optimization. *Surveys in Operations Research and Management Science*, 19(1):56–85, 2014.
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.

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](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.