Contextual decision-making under uncertainty

Juan M. Morales, Salvador Pineda, Miguel Á. Muñoz and Adrián Esteban-Pérez

juan.morales@uma.es, spineda@uma.es

University of Malaga
OASYS group
oasys.uma.es

European Research Council

Established by the European Commission

Online Seminar Series Machine Learning NeEDS Mathematical Optimization
March 22, 2021

About Málaga

- Over 300 sunny days per year (known as Costa del Sol)
- University of Málaga was stablished in 1972 and currently has 40000 students and 2500 faculty members
- Málaga is becoming the Silicon Valley of the south of Spain
- Andalusia Technology Park includes over 600 companies (Oracle, Ericsson, IBM, TDK, Huawei, Microsoft, Cisco), 20.000 employees and a turnover of 2.000 M in 2018

Google To Open A Cybersecurity 'Centre Of Excellence' In Malaga

By **Chris King** - 11 February 2021 @ 21:05

About OASYS

- **O**ptimization and **A**nalytics for **S**ustainable energ**Y** **S**ystems
- Established in 2018
- 2 professors, 2 Postdoc, 3 PhD students, 3 research assistants
- Research topics:
  - Mathematical models for decision-making under uncertainty
  - Use of large amounts of data for Smart Energy Grids
  - Forecasting and optimization for Sustainable Energy Systems
  - Algorithms for the efficient solution of large-scale optimization problems
  - Game theory for the analysis of energy markets
- More info: oasys.uma.es

Introduction: Conditional Stochastic Problem

Solve

$$J^* := \inf_{x \in X} \mathbb{E}_{\mathbb{Q}} \left[f(x, y) \,\middle/\, z = z_0 \right] = \mathbb{E}_{\mathbb{Q}_{\widetilde{\Xi}}} \left[f(x, y) \right]$$

with optimal solution x^*.

- The decision variable $x \in X \subseteq \mathbb{R}^n$.
- A random vector y with support set $\Xi_y \subseteq \mathbb{R}^{d_y}$, whereby we model the uncertainty affecting the value of the decision. Example: demand...
- Covariates (or features) modeled by a random vector z with support set $\Xi_z \subseteq \mathbb{R}^{d_z}$. Example: weather forecast, hashtag twitter...
- Side information: $\widetilde{\Xi} = \{(z, y) : z = z_0\}$

Introduction: Conditional Stochastic Problem

- At 10 am we have to decide how much ice cream to make (decision x)
- At 10 am we do not know the demand in the afternoon (uncertain parameter y)
- We can use some available information such as the temperature at 10 am (contextual data z)
- Obviously there is a relationship between the morning temperature (z) and the ice cream demand in the afternoon (y).
- We would like to use such a relation to make better decisions about ice cream quantity

Introduction: Conditional Stochastic Problem

Fundamental challenge

- The true joint distribution $\mathbb{Q}$ and the true conditional distribution $\mathbb{Q}_{\cong}$ are unknown!
- All that we have is a data sample of size T, i.e., $(\mathsf{z}_t, \mathsf{y}_t)_{t=1}^{T}$.

Goal

We want to exploit the info on the features, z_0 (the context), in our favor to prescribe better decisions.

Some possible solution approaches

(Traditional) Forecasting

Compute $\widehat{y}(z_0) \approx \mathbb{E}[y|z = z_0]$ in the hope that

$$(arg)\inf_{x \in X} \mathbb{E}_{\mathbb{Q}}\left[f(x,y)\,\middle/\, z = z_0\right] \approx (arg)\inf_{x \in X} f(x, \widehat{y}(z_0))$$

Uncertainty, and thus its impact, are ignored.

Decision rule

Find $\widehat{x}^*(\cdot) : \Xi_z \longrightarrow X$ such that

$$arg\inf_{x \in X} \mathbb{E}_{\mathbb{Q}}\left[f(x,y)\,\middle/\, z = z_0\right] \approx \widehat{x}^*(z_0)$$

The image of $x(\cdot)$ should be a subset of X to guarantee decision feasibility.

Some possible solution approaches

Smart predict

Find $\widehat{y}(\cdot) : \Xi_z \longrightarrow \Xi_y$ (very possibly different from $\mathbb{E}[y|z]$) such that

$$(arg)\inf_{x\in X}\mathbb{E}_{\mathbb{Q}}\left[f(x,y)\,\middle/\,z=z_0\right]\approx(arg)\inf_{x\in X}f(x,\widehat{y}(z_0))$$

Approximate conditional distribution $\mathbb{Q}_{\widetilde{\Xi}}$

Take $\widehat{\mathbb{Q}}_{\widetilde{\Xi}}$ as a proxy of $\mathbb{Q}_{\widetilde{\Xi}}$ in the hope that

$$(arg)\inf_{x\in X}\mathbb{E}_{\mathbb{Q}}\left[f(x,y)\,\middle/\,z=z_0\right]=(arg)\inf_{x\in X}\mathbb{E}_{\mathbb{Q}_{\widetilde{\Xi}}}\left[f(x,y)\right]\approx(arg)\inf_{x\in X}\mathbb{E}_{\widehat{\mathbb{Q}}_{\widetilde{\Xi}}}\left[f(x,y)\right]$$

In any case, we must *infer* from the data sample $(z_t, y_t)_{t=1}^{T}$

Some possible solution approaches

Parametric approach (bilevel optimization)

A bilevel framework for decision-making under uncertainty with contextual information

M. A. Muñoz, S. Pineda, J. M. Morales
OASYS Group, University of Malaga, Malaga, Spain

Abstract

In this paper we propose a novel approach for data-driven decision-making under uncertainty in the presence of contextual information. Given a finite collection of observations of the uncertain parameters and potential explanatory variables (i.e., the contextual information), our approach fits a parametric model to those data that is specifically tailored to maximizing the decision value, while accounting for possible feasibility constraints. From a mathematical point of view, our framework translates into a bilevel program, for which we provide both a fast regularization procedure and a big-M-based reformulation to aim for a local and a global optimal solution, respectively. We showcase the benefits of moving from the traditional scheme for model estimation (based on statistical quality metrics) to decision-guided prediction using the problem of a strategic producer competing *la Cournot* in a market for an homogeneous product. In particular, we include a realistic case study, based on data from the Iberian electricity market, whereby we compare our approach with alternative ones available in the technical literature and analyze the conditions (in terms of the firm's cost structure and production capacity) under which our approach proves to be more advantageous to the producer.

Non-parametric approach (distributionally robust optimization)

Distributionally robust stochastic programs with side information based on trimmings

Adrián Esteban-Pérez · Juan M. Morales

Received: date / Accepted: date

Abstract We consider stochastic programs conditional on some covariate information, where the only knowledge of the possible relationship between the uncertain parameters and the covariates is reduced to a finite data sample of their joint distribution. By exploiting the close link between the notion of *trimmings* of a probability measure and the *partial* mass transportation problem, we construct a data-driven Distributionally Robust Optimization (DRO) framework to hedge the decision against the intrinsic error in the process of inferring conditional information from limited joint data. We show that our approach is computationally as tractable as the standard (without side information) Wasserstein-metric-based DRO. Furthermore, our DRO framework can be conveniently used to address data-driven decision-making problems under contaminated samples and naturally produces distributionally robust versions of some local nonparametric predictive methods, such as Nadaraya-Watson kernel regression and K-nearest neighbors, which are often used in the context of conditional stochastic optimization. Leveraging results from empirical point processes and optimal transport, we show that our approach enjoys performance guarantees. Finally, the theoretical results are illustrated using a single-item newsvendor problem and a portfolio allocation problem with side information.

Keywords Distributionally Robust Optimization · Trimmings · Side information · Partial Mass Transportation Problem · Newsvendor problem · Portfolio optimization

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 755705). This work was also supported in part by the Spanish Ministry of Economy, Industry and Competitiveness and the European Regional Development Fund (ERDF) through project ENE2017-83775-P.

Adrián Esteban-Pérez
Department of Applied Mathematics, University of Málaga, Málaga, 29071, Spain
E-mail: adrianesteban@uma.es

Juan M. Morales (corresponding author)
Department of Applied Mathematics, University of Málaga, Málaga, 29071, Spain

A *parametric* family of functions

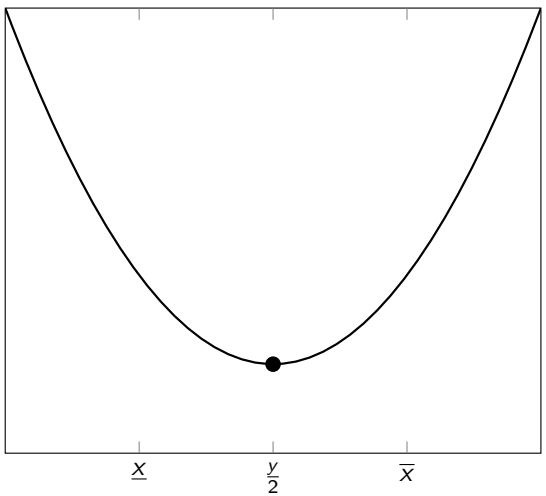
- Forecasting approach (FO)
  - learns the relation between y and z ignoring f and X
- Decision rule approach (DR)
  - learns the relation between x^* and z
- Bilevel approach (BL)
  - learns the relation between y and z taking into account f and X

Application: strategic producer

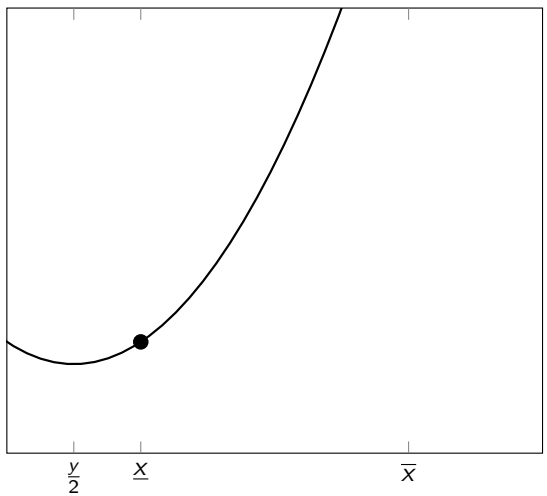
The optimal quantity of a strategic producer (x^*) can be determined as:

$$x^* = \arg \min_{\underline{x} \leq x \leq \overline{x}} x^2 - yx$$

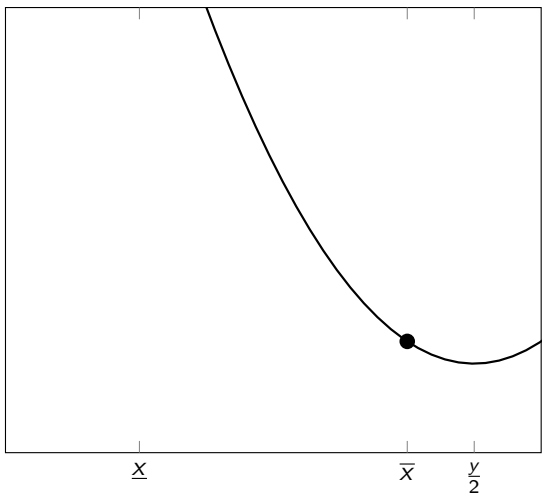
where y is an uncertain parameters that depends on market conditions



$x^* = \frac{y}{2}$



$x^* = \underline{x}$

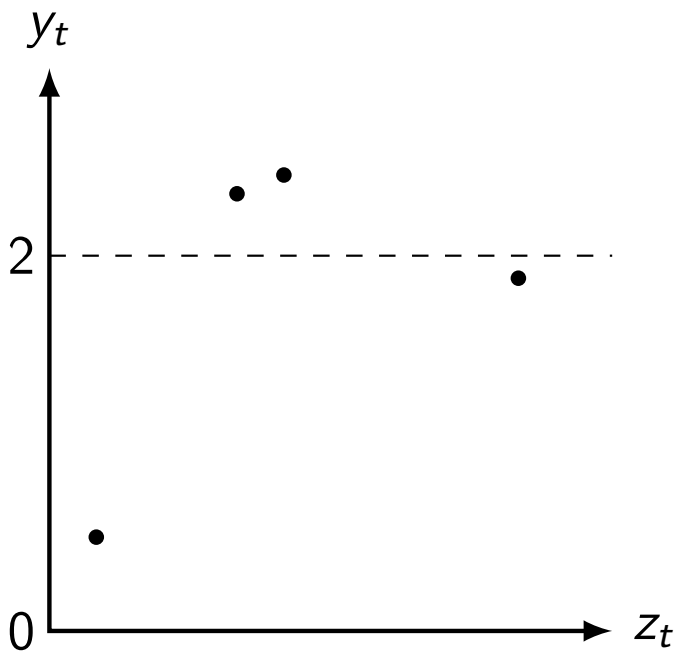


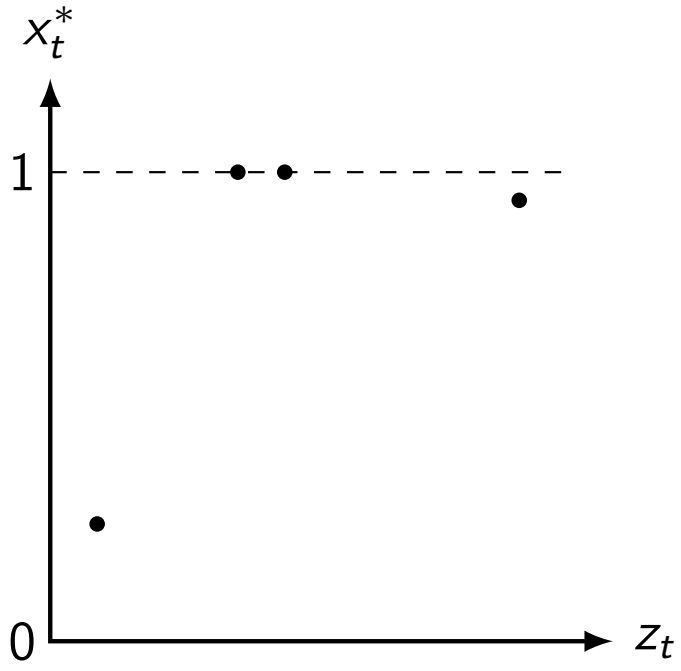
$x^* = \overline{x}$

Illustrative example

t	z_t	y_t	$y_t/2$	x_t^*
1	1	0.50	0.25	0.25
2	4	2.33	1.16	1.00
3	5	2.43	1.21	1.00
4	10	1.88	0.94	0.94

$$0 \leq x \leq 1$$





Income = 21.16 (100%)

Forecasting approach (FO)

- We assume a linear function: $\hat{y}_t = a + bz_t$

Forecasting approach (FO)

- We assume a linear function: $\hat{y}_t = a + bz_t$
- We compute a, b to minimize errors as follows:

$$a^*, b^* = \arg\min_{a,b} \sum_{t \in \mathcal{T}} (y_t - (a + bz_t))^2$$

Forecasting approach (FO)

- We assume a linear function: $\hat{y}_t = a + bz_t$
- We compute a, b to minimize errors as follows:

$$a^*, b^* = \arg\min_{a,b} \sum_{t \in \mathcal{T}} (y_t - (a + bz_t))^2$$

- We estimate $\hat{y}$ for a new time period $\tilde{t}$ as

$$\hat{y}_{\tilde{t}} = a^* + b^* z_{\tilde{t}}$$

Forecasting approach (FO)

- We assume a linear function: $\hat{y}_t = a + bz_t$
- We compute a, b to minimize errors as follows:

$$a^*, b^* = \arg\min_{a,b} \sum_{t \in \mathcal{T}} (y_t - (a + bz_t))^2$$

- We estimate $\hat{y}$ for a new time period $\tilde{t}$ as

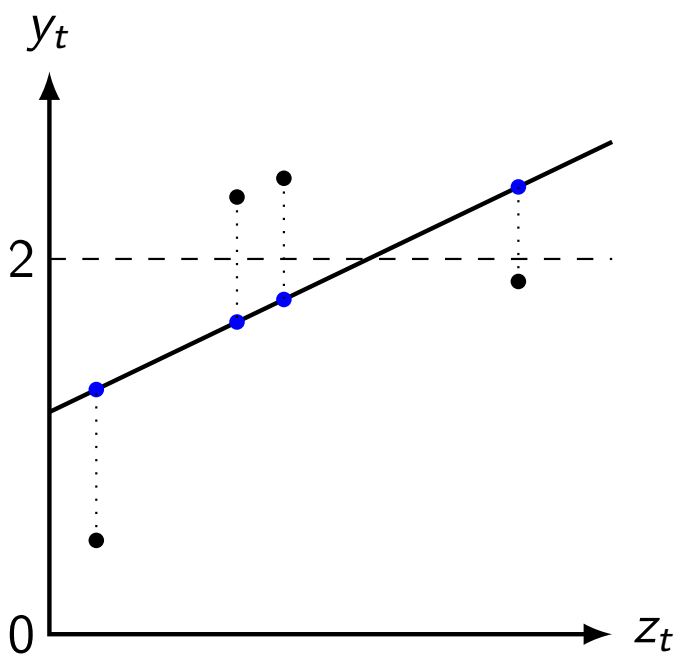
$$\hat{y}_{\tilde{t}} = a^* + b^* z_{\tilde{t}}$$

- We determine the produced quantity for a new time period $\tilde{t}$ as

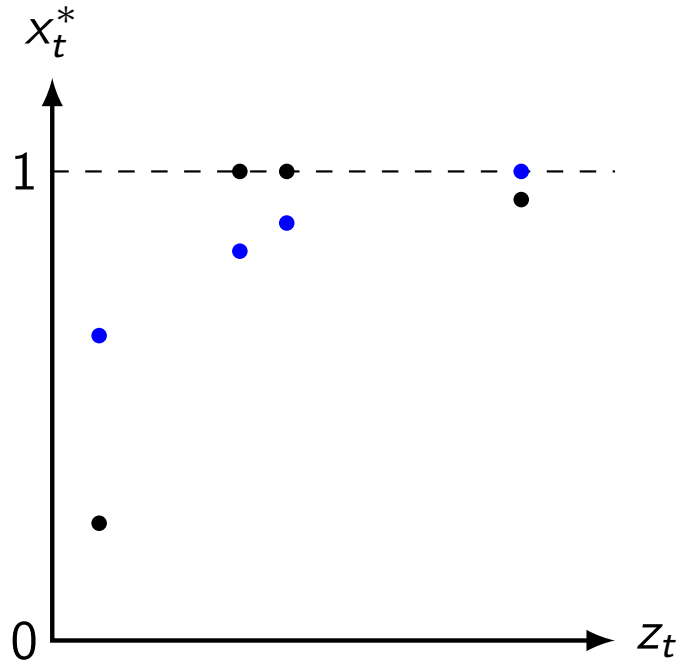
$$x^*_{\tilde{t}} = \arg\min_{\underline{x} \leq x_{\tilde{t}} \leq \overline{x}} x^2_{\tilde{t}} - \hat{y}_{\tilde{t}} x_{\tilde{t}}$$

Illustrative example

Forecasting approach (FO): $\hat{y}_t = 1.184 + 0.12 z_t$



RMSE = 0.665



Income = 20.14 (95.2%)

Decision-rule approach (DR)

- We assume a linear function so that $x_t^* = a + bz_t$

Decision-rule approach (DR)

- We assume a linear function so that $x_t^* = a + bz_t$
- We compute a, b to minimize the objective function

$$a^*, b^* = \arg\min_{a,b} \sum_{t \in \mathcal{T}} x_t^2 - y_t x_t$$
$$\text{s.t. } \underline{x} \leq x_t \leq \overline{x}, \quad \forall t \in \mathcal{T}$$
$$x_t = a + bz_t, \quad \forall t \in \mathcal{T}$$

Decision-rule approach (DR)

- We assume a linear function so that $x_t^* = a + bz_t$
- We compute a, b to minimize the objective function

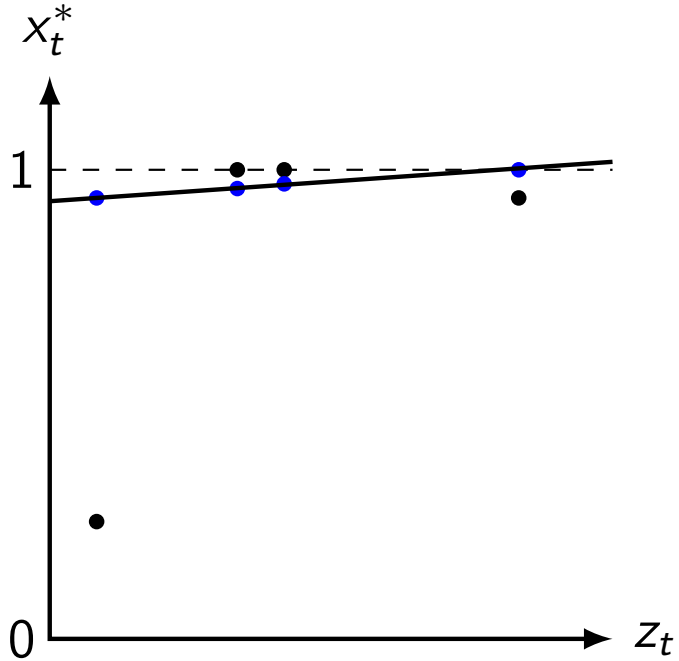
$$a^*, b^* = \arg\min_{a,b} \sum_{t \in \mathcal{T}} x_t^2 - y_t x_t$$
$$\text{s.t. } \underline{x} \leq x_t \leq \overline{x}, \quad \forall t \in \mathcal{T}$$
$$x_t = a + bz_t, \quad \forall t \in \mathcal{T}$$

- We determine the produced quantity for a new time period $\tilde{t}$ as

$$x_{\tilde{t}}^* = a^* + b^* z_{\tilde{t}}$$

Illustrative example

Decision-rule approach (DR): $\hat{x}_t^* = 0.933 + 0.007 z_t$



Income = 20.02 (94.6%)

Infeasibilities may happen!!

A bit of bilevel programming

John and Peter have a deal to go to the movies. First John decides the movie theater, and then Peter decides which movie they will watch. John prefers action over terror movies, and terror over romantic movies. Peter prefers romantic over terror, and terror over action movies. Which theater do you think John would choose?

Theater A
Spiderman
Notting Hill

Theater B
The Exorcist
The Matrix

A bit of bilevel programming

Bilevel optimization is a special kind of optimization where one problem is embedded (nested) within another and can be formulated as follows

$$
\begin{aligned}
\min_{\alpha} \quad & F_0(\alpha, \beta) \\
\text{s.t.} \quad & F_i(\alpha, \beta) \leq 0, \qquad i = 1, \ldots, I \\
& H_j(\alpha, \beta) = 0, \qquad j = 1, \ldots, J \\
& \min_{\beta} \quad f_0(\alpha, \beta) \\
& \text{s.t.} \quad f_k(\alpha, \beta) \leq 0, \qquad k = 1, \ldots, K \\
& \qquad\quad h_l(\alpha, \beta) = 0, \qquad l = 1, \ldots, L
\end{aligned}
$$

Bilevel approach (BL)

- We assume a linear function so that $\hat{y}_t = a + bz_t$

Bilevel approach (BL)

- We assume a linear function so that $\hat{y}_t = a + bz_t$
- We compute a, b by solving the following bilevel problem:

$$a^*, b^* = \arg\min_{a,b} \sum_{t \in \mathcal{T}} \hat{x}_t^2 - y_t \hat{x}_t$$

$$\text{s.t. } \hat{x}_t = \arg\min_{\underline{x} \leq x_t \leq \overline{x}} x_t^2 - (a + bz_t)x_t, \quad \forall t \in \mathcal{T}$$

Bilevel approach (BL)

- We assume a linear function so that $\hat{y}_t = a + bz_t$
- We compute a, b by solving the following bilevel problem:

$$a^*, b^* = \arg\min_{a,b} \sum_{t \in \mathcal{T}} \hat{x}_t^2 - y_t \hat{x}_t$$
$$\text{s.t. } \hat{x}_t = \arg\min_{\underline{x} \le x_t \le \overline{x}} x_t^2 - (a + bz_t)x_t, \quad \forall t \in \mathcal{T}$$

- We estimate $\hat{y}$ for a new time period $\tilde{t}$ as

$$\hat{y}_{\tilde{t}} = a^* + b^* z_{\tilde{t}}$$

Bilevel approach (BL)

- We assume a linear function so that $\hat{y}_t = a + bz_t$
- We compute a, b by solving the following bilevel problem:

$$a^*, b^* = \arg\min_{a,b} \sum_{t \in \mathcal{T}} \hat{x}_t^2 - y_t \hat{x}_t$$
$$\text{s.t. } \hat{x}_t = \arg\min_{\underline{x} \le x_t \le \overline{x}} x_t^2 - (a + bz_t)x_t, \quad \forall t \in \mathcal{T}$$

- We estimate $\hat{y}$ for a new time period $\tilde{t}$ as

$$\hat{y}_{\tilde{t}} = a^* + b^* z_{\tilde{t}}$$

- We determine the produced quantity for a new time period $\tilde{t}$ as

$$x_{\tilde{t}}^* = \arg\min_{\underline{x} \le x_{\tilde{t}} \le \overline{x}} x_{\tilde{t}}^2 - \hat{y}_{\tilde{t}} x_{\tilde{t}}$$

Bilevel approach (BL)

We replace the lower-level by its KKT to obtain a single-level problem:

$$
\begin{aligned}
a^*, b^* = \arg\min_{a,b} \quad & \sum_{t \in \mathcal{T}} \hat{x}_t^2 - y_t \hat{x}_t \\
\text{s.t.} \quad & 2\hat{x}_t - a - b z_t - \lambda_{1t} + \lambda_{2t} = 0, \quad \forall t \in \mathcal{T} \\
& \underline{x} \leq x_{\tilde{t}} \leq \overline{x}, \quad \forall t \in \mathcal{T} \\
& \lambda_{1t}, \lambda_{2t} \geq 0, \quad \forall t \in \mathcal{T} \\
& \lambda_{1t}(\hat{q}_t - \underline{q}) = 0, \quad \forall t \in \mathcal{T} \\
& \lambda_{2t}(\overline{q} - \hat{q}_t) = 0, \quad \forall t \in \mathcal{T}
\end{aligned}
$$

where $\lambda_{1t}, \lambda_{2t}$ are dual variables of the lower-level problem

Bilevel approach (BL)

We can use Fortuny-Amat to reformulate complementarity conditions:

$$
\begin{aligned}
a^*, b^* = \arg\min_{a,b} \ & \sum_{t\in\mathcal{T}} \hat{x}_t^2 - y_t \hat{x}_t \\
\text{s.t. } & 2\hat{x}_t - a - bz_t - \lambda_{1t} + \lambda_{2t} = 0, \ \ \forall t \in \mathcal{T} \\
& 0 \le \hat{x}_t - \underline{x} \le (1 - u_{1t}) M^P, \ \ \forall t \in \mathcal{T} \\
& 0 \le \overline{x} - \hat{x}_t \le (1 - u_{2t}) M^P, \ \ \forall t \in \mathcal{T} \\
& 0 \le \lambda_{1t} \le u_{1t} M^D, \ \ \forall t \in \mathcal{T} \\
& 0 \le \lambda_{2t} \le u_{2t} M^D, \ \ \forall t \in \mathcal{T} \\
& u_{1t}, u_{2t} \in \{0, 1\}, \ \ \forall t \in \mathcal{T}
\end{aligned}
$$

where u_{1t}, u_{2t} are binary variables and M^P, M^D large enough constants

Bilevel approach (BL)

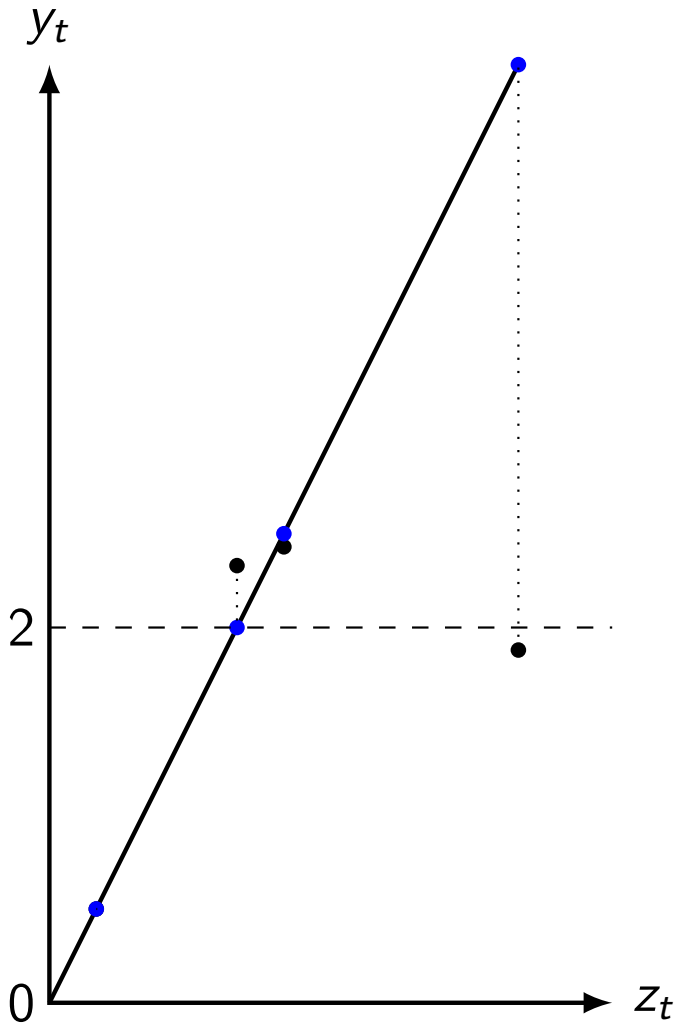
Alternatively, we can use the following regularization approach:

$$
\begin{aligned}
a^*, b^* = \arg\min_{a,b} \quad & \sum_{t \in \mathcal{T}} \hat{x}_t^2 - y_t \hat{x}_t \\
\text{s.t.} \quad & 2\hat{x}_t - a - b z_t - \lambda_{1t} + \lambda_{2t} = 0, \quad \forall t \in \mathcal{T} \\
& \underline{x} \leq \hat{x}_t \leq \overline{x}, \quad \forall t \in \mathcal{T} \\
& \lambda_{1t}, \lambda_{2t} \geq 0, \quad \forall t \in \mathcal{T} \\
& \lambda_{1t}(\hat{x}_t - \underline{x}) + \lambda_{2t}(\overline{x} - \hat{x}_t) \leq \epsilon
\end{aligned}
$$

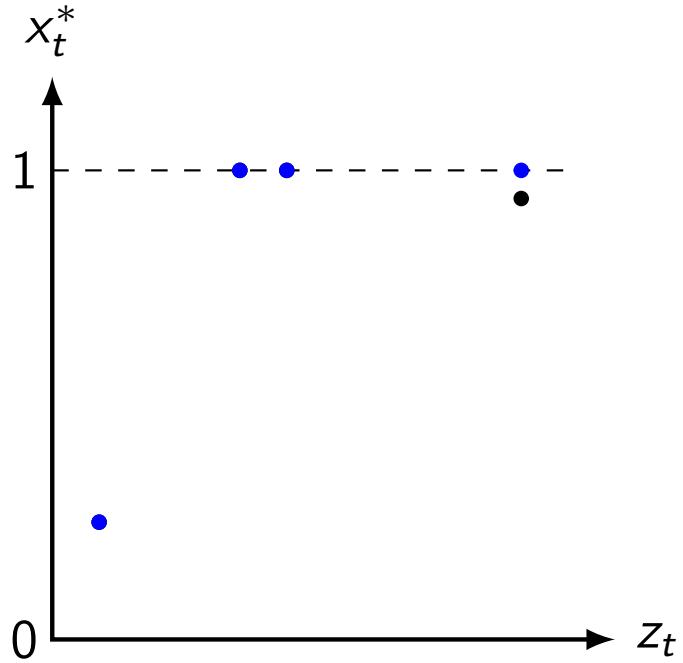
where paremeter ϵ is iteratively reduced to 0

Illustrative example

Bilevel approach (BL): $\hat{y}_t = 0.000 + 0.5 z_t$



RMSE = 0.745



Income = 21.13 (99.9%)

Illustrative example

	RMSE	Income
Forecasting (FO)	0.665	95.2%
Decision-rule (DR)	-	94.6%
Bilevel (BL)	0.745	99.9%

- FO minimizes forecast error, but yields suboptimal decisions
- DR simplifies decision-making, but also yields suboptimal decisions
- BL finds the relation between y and z that derives best decisions

Case study

- Real data from Iberian electricity market is used to approximate the inverse demand function
- Wind and solar power forecasts is used as contextual information
- Three different generation technologies: base (nuclear), medium (carbon) and peak (gas)
- 43 sets of 200 hours (160 hours as training and 40 hours as test) are used to compute average results

Case study

	Base	Medium	Peak
Relative income FO	96.0%	77.3%	41.6%
Relative income DR	94.6%	62.6%	18.9%
Relative income BL	96.3%	80.0%	58.7%
Infeasible cases DR	4.9%	1.7%	0.1 %

- All methods provide similar incomes for the base unit since it is at full capacity most of the time
- The uncertainty of the inverse demand function significantly affects the operation of medium and peak units
- The proposed BL approach obtains the highest incomes for the three generating technologies
- DR approach lead to a significant number of infeasible cases

Conclusions

- Forecasting approach (FO)
  - learns the relation between y and z ignoring f and X
  - 👍 wide variety of learning techniques can be applied
  - 👎 obtained decisions may be suboptimal
- Decision rule approach (DR)
  - learns the relation between x^* and z
  - 👍 decisions are quickly obtained wihtout solving an optimization problem
  - 👎 obtained decisions may be infeasible
- Bilevel approach (BL)
  - learns the relation between y and z taking into account f and X
  - 👍 best possible decisions using available contextual information
  - 👎 bilevel problem can be only solved under certain assumptions

Some possible solution approaches

Parametric approach (bilevel optimization)

A bilevel framework for decision-making under uncertainty with contextual information

M. A. Muñoz, S. Pineda, J. M. Morales
OASYS Group, University of Malaga, Malaga, Spain

Abstract

In this paper we propose a novel approach for data-driven decision-making under uncertainty in the presence of contextual information. Given a finite collection of observations of the uncertain parameters and potential explanatory variables (i.e., the contextual information), our approach fits a parametric model to those data that is specifically tailored to maximizing the decision value, while accounting for possible feasibility constraints. From a mathematical point of view, our framework translates into a bilevel program, for which we provide both a fast regularization procedure and a big-M-based reformulation to aim for a local and a global optimal solution, respectively. We showcase the benefits of moving from the traditional scheme for model estimation (based on statistical quality metrics) to decision-guided prediction using the problem of a strategic producer competing *la Cournot* in a market for an homogeneous product. In particular, we include a realistic case study, based on data from the Iberian electricity market, whereby we compare our approach with alternative ones available in the technical literature and analyze the conditions (in terms of the firm's cost structure and production capacity) under which our approach proves to be more advantageous to the producer.

Non-parametric approach (distributionally robust optimization)

Distributionally robust stochastic programs with side information based on trimmings

Adrián Esteban-Pérez · Juan M. Morales

Received: date / Accepted: date

Abstract We consider stochastic programs conditional on some covariate information, where the only knowledge of the possible relationship between the uncertain parameters and the covariates is reduced to a finite data sample of their joint distribution. By exploiting the close link between the notion of *trimmings* of a probability measure and the *partial* mass transportation problem, we construct a data-driven Distributionally Robust Optimization (DRO) framework to hedge the decision against the intrinsic error in the process of inferring conditional information from limited joint data. We show that our approach is computationally as tractable as the standard (without side information) Wasserstein-metric-based DRO. Furthermore, our DRO framework can be conveniently used to address data-driven decision-making problems under contaminated samples and naturally produces distributionally robust versions of some local nonparametric predictive methods, such as Nadaraya-Watson kernel regression and K-nearest neighbors, which are often used in the context of conditional stochastic optimization. Leveraging results from empirical point processes and optimal transport, we show that our approach enjoys performance guarantees. Finally, the theoretical results are illustrated using a single-item newsvendor problem and a portfolio allocation problem with side information.

Keywords Distributionally Robust Optimization · Trimmings · Side information · Partial Mass Transportation Problem · Newsvendor problem · Portfolio optimization

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 755705). This work was also supported in part by the Spanish Ministry of Economy, Industry and Competitiveness and the European Regional Development Fund (ERDF) through project ENE2017-83775-P.

Adrián Esteban-Pérez
Department of Applied Mathematics, University of Málaga, Málaga, 29071, Spain
E-mail: adrianesteban@uma.es

Juan M. Morales (corresponding author)
Department of Applied Mathematics, University of Málaga, Málaga, 29071, Spain

Some possible solution approaches

Smart predict

Find $\widehat{y}(\cdot) : \Xi_z \longrightarrow \Xi_y$ (very possibly different from $\mathbb{E}[y|z]$) such that

$$(arg)\inf_{x\in X} \mathbb{E}_{\mathbb{Q}}\left[f(x,y)\,\middle/\, z = z_0\right] \approx (arg)\inf_{x\in X} f(x,\widehat{y}(z_0))$$

Approximate conditional distribution $\mathbb{Q}_{\widetilde{\Xi}}$

Take $\widehat{\mathbb{Q}}_{\widetilde{\Xi}}$ as a proxy of $\mathbb{Q}_{\widetilde{\Xi}}$ in the hope that

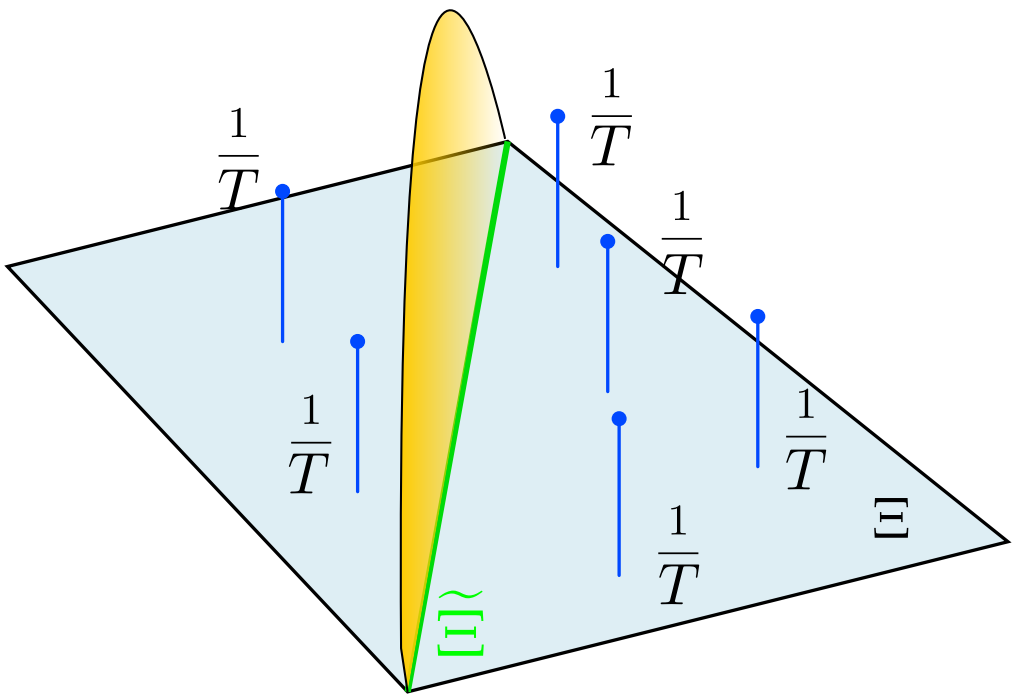
$$(arg)\inf_{x\in X} \mathbb{E}_{\mathbb{Q}}\left[f(x,y)\,\middle/\, z = z_0\right] = (arg)\inf_{x\in X} \mathbb{E}_{\mathbb{Q}_{\widetilde{\Xi}}}\left[f(x,y)\right] \approx (arg)\inf_{x\in X} \mathbb{E}_{\widehat{\mathbb{Q}}_{\widetilde{\Xi}}}\left[f(x,y)\right]$$

In any case, we must *infer* from the data sample $(z_t, y_t)_{t=1}^{T}$

Approximate conditional distribution $\mathbb{Q}_{\widetilde{\Xi}}$

Training data: A sample of size T, joint limited data: $(z_t, y_t)_{t=1}^T$

$$\mathbb{Q}_{\widetilde{\Xi}} \approx \sum_{t=1}^{T} w_T^t(z_0)\delta_{\widehat{y}^t} = \widehat{\mathbb{Q}}_{\widetilde{\Xi}} \quad \text{“Empirical” conditional distribution}$$

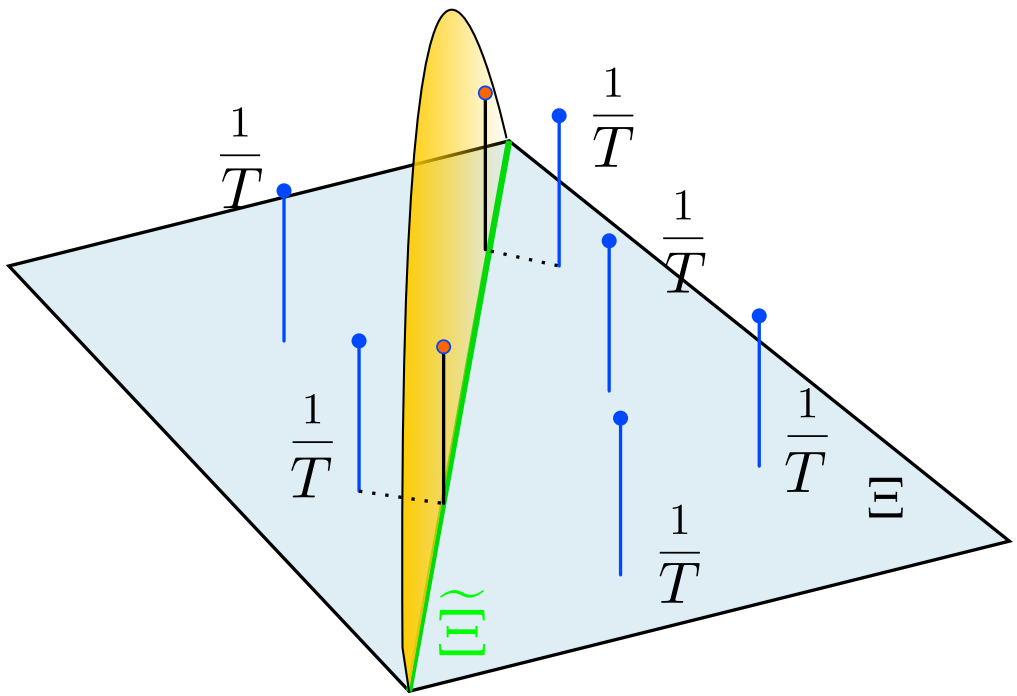


Bertsimas and Kallus 2019 (non-parametric)

Solve:

$$\inf_{x \in X} \sum_{t=1}^{T} w_T^t(z_0) f(x, \widehat{y}^t) \quad \left(\approx \inf_{x \in X} \mathbb{E}_{\mathbb{Q}_{\widehat{\Xi}}} [f(x, y)] \right),$$

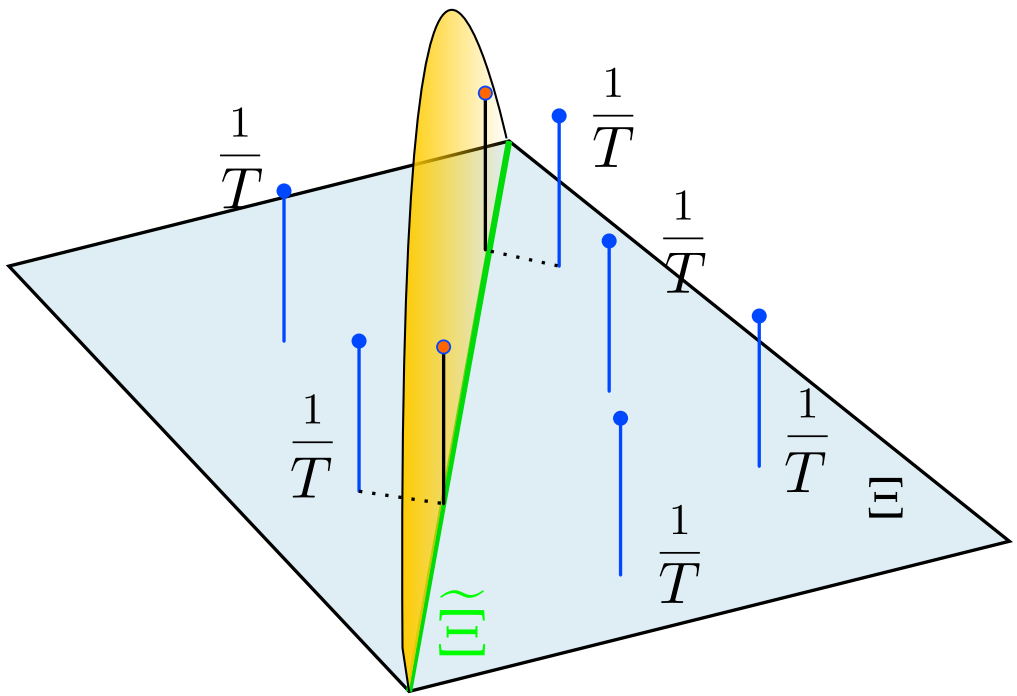
where $w_T^t(z_0)$ is given by a local predictive method ($K-$NN, Nadaraya-Watson...)



Bertsimas and Kallus 2019 (non-parametric)

Example: 2-NN

$$\inf_{x \in X} \frac{1}{2} f(x, \widehat{y}_{1^{st}}) + \frac{1}{2} f(x, \widehat{y}_{2^{nd}})$$

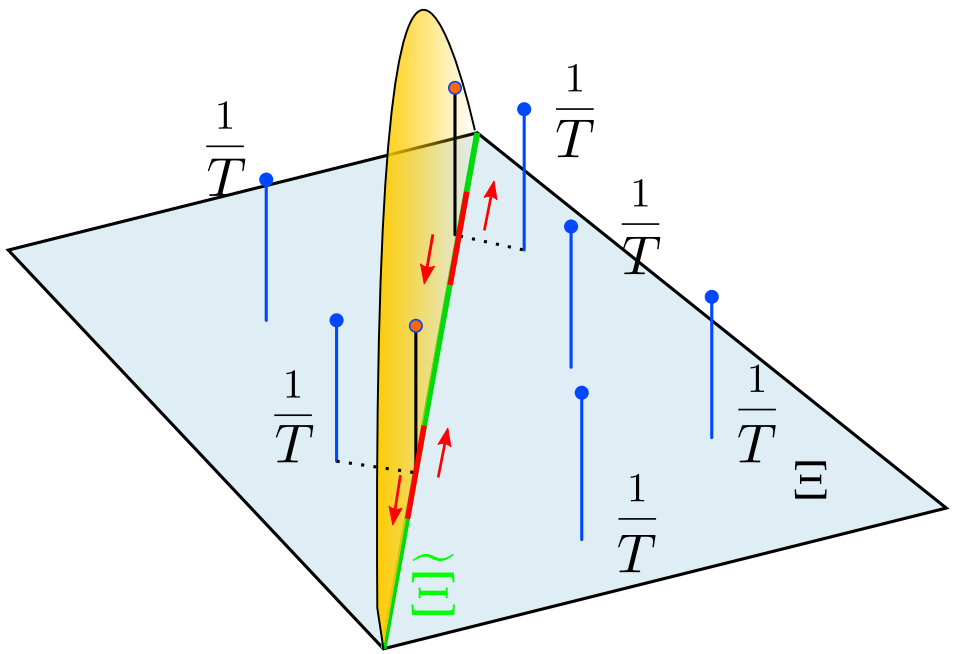


Bertsimas, McCord, and Sturt 2019

Solve:

$$\inf_{x \in X} \sum_{t=1}^{T} w_T^t(z_0) \sup_{y \in \mathcal{U}_T^t} [f(x, y)] \tag{1}$$

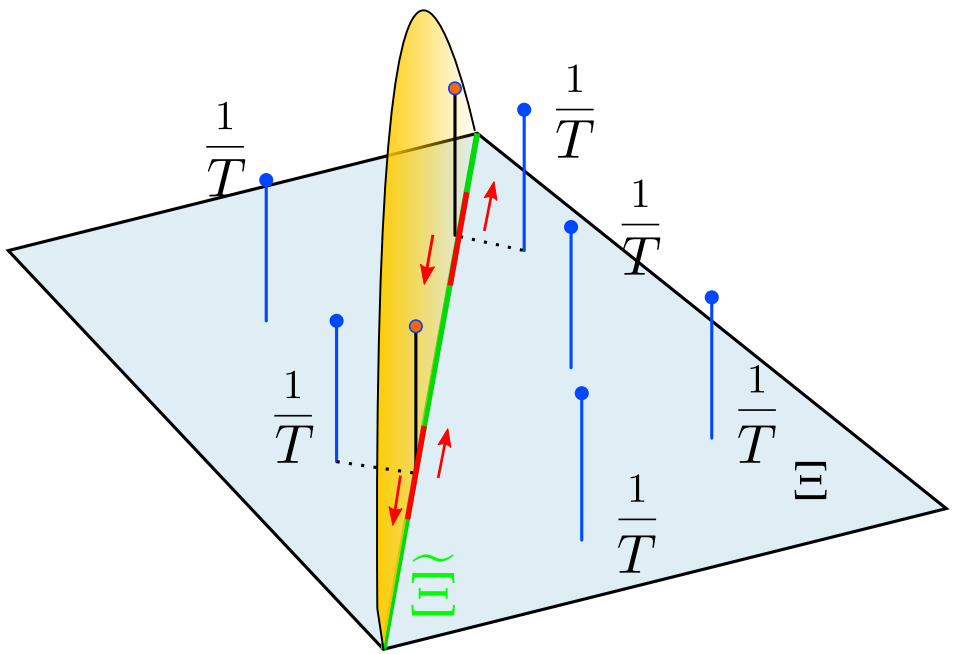
where $\mathcal{U}_T^t := \{y \in \Xi_y \ / \ \|y - \hat{y}^t\|_p \leqslant \varepsilon_T\}$.



Bertsimas, McCord, and Sturt 2019

A DRO approach: Wasserstein ball centered at the "empirical" conditional distribution given by a local predictive method.

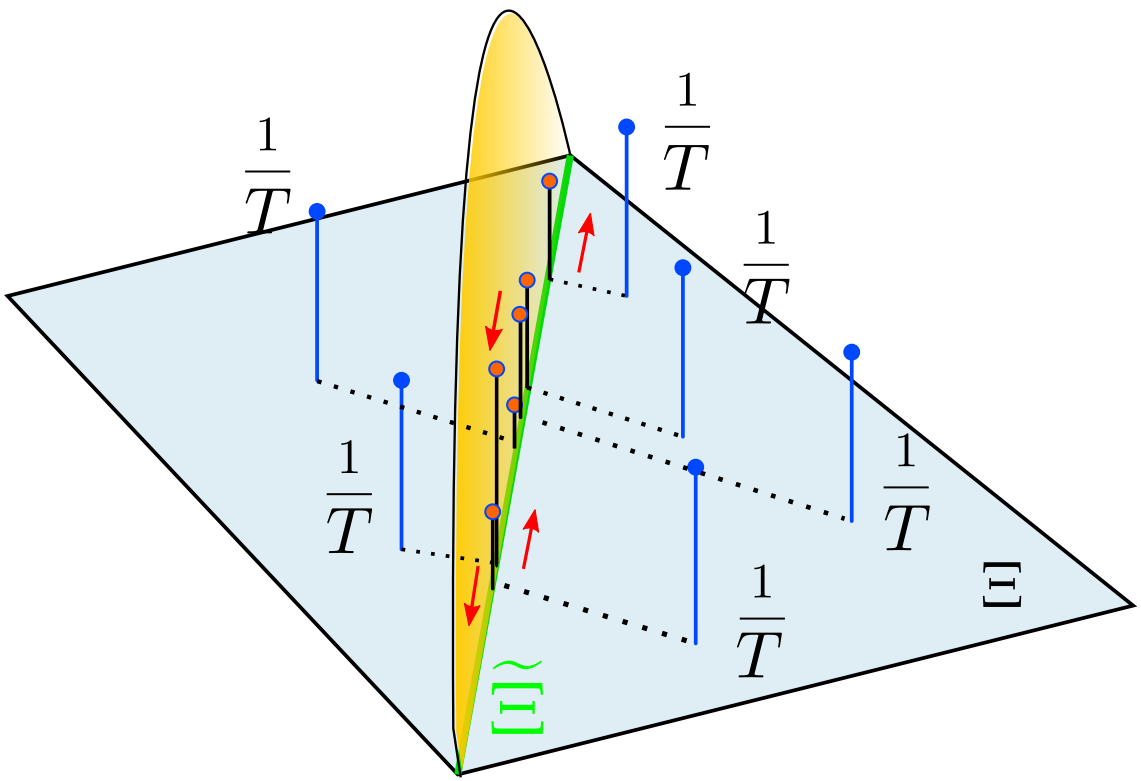
$$\mathcal{U}_T := \left\{ Q_{\widetilde{\Xi}}(\widetilde{\Xi}) = 1 : \mathcal{W}_p(\widehat{\mathbb{Q}}_{\Xi}, Q_{\widetilde{\Xi}}) \leq \widetilde{\rho}^{1/p} \right\}$$



Our approach: Esteban-Pérez and Morales 2020

$$\mathcal{U}_T := \left\{ Q_{\widetilde{\Xi}}(\widetilde{\Xi}) = 1 : \mathcal{W}_p(\mathcal{R}_{1-\alpha}(\widehat{\mathbb{Q}}_T), Q_{\widetilde{\Xi}}) \leq \widetilde{\rho}^{1/p} \right\}$$

where $\mathcal{R}_{1-\alpha}(\widehat{\mathbb{Q}}_T)$, $\alpha \in [0,1]$, is the *set of all* $(1-\alpha)$*-trimmings* of $\widehat{\mathbb{Q}}_T$ (all probability distributions $\sum_{t=1}^{T} b_t \delta_{\boldsymbol{\xi}_t}$ with $0 \leq b_t \leq \frac{1}{T\alpha}$, $\forall t = 1, \ldots, T$, and $\sum_{t=1}^{T} b_t = 1$).



Our DRO approach

$$\left(\mathsf{P}_{(\alpha_T,\widetilde{\rho}_T)}\right) \quad \inf_{x\in X} \sup_{Q_{\widetilde{\Xi}}} \ \mathbb{E}_{Q_{\widetilde{\Xi}}}\left[f(x,\boldsymbol{\xi})\right] \tag{2a}$$

$$\text{s.t. } \mathcal{W}_p^p(\mathcal{R}_{1-\alpha_T}(\widehat{\mathbb{Q}}_T), Q_{\widetilde{\Xi}}) \leq \widetilde{\rho}_T \tag{2b}$$

$$Q_{\widetilde{\Xi}}(\widetilde{\Xi}) = 1 \tag{2c}$$

- $\mathcal{W}_p^p(\mathcal{R}_{1-\alpha_T}(\widehat{\mathbb{Q}}_T), Q_{\widetilde{\Xi}}) \leq \widetilde{\rho}_T$: Partial mass transportation constraint.
- α_T : Amount of mass of $\widehat{\mathbb{Q}}_T$ (in per unit) transported. Natural choice: $\alpha_T := K_T/T$ (Distributionally Robust K_T-NN version).
- $\widetilde{\rho}_T$: Transportation budget (degree of robustness).

Key features of the proposed framework

- Fully data-driven single-phase method (no prior estimate phase by a local predictive method).
- The relationship of the random vector and the features is encoded in the framework and its impact on the objective function is considered.
- Performance guarantees are available under mild assumptions and a suitable choice of the parameters α_T and $\widetilde{\rho}_T$ (Esteban-Pérez and Morales 2020).
- As tractable as the standard DRO model based on the Optimal Transport (without side information).

Single-item Newsvendor problem with side information

- h: unit holding cost; b: unit backorder cost; x: order quantity (decision variable)
- y: r.v. (demand); $\{z = z_0\}$: side information (week day, weather,...)

$$\min \mathbb{E}\left[h(x-y)^+ + b(y-x)^+ \Big/ z = z_0\right]$$

**Equivalent to estimate a quantile $b/(b+h)$.

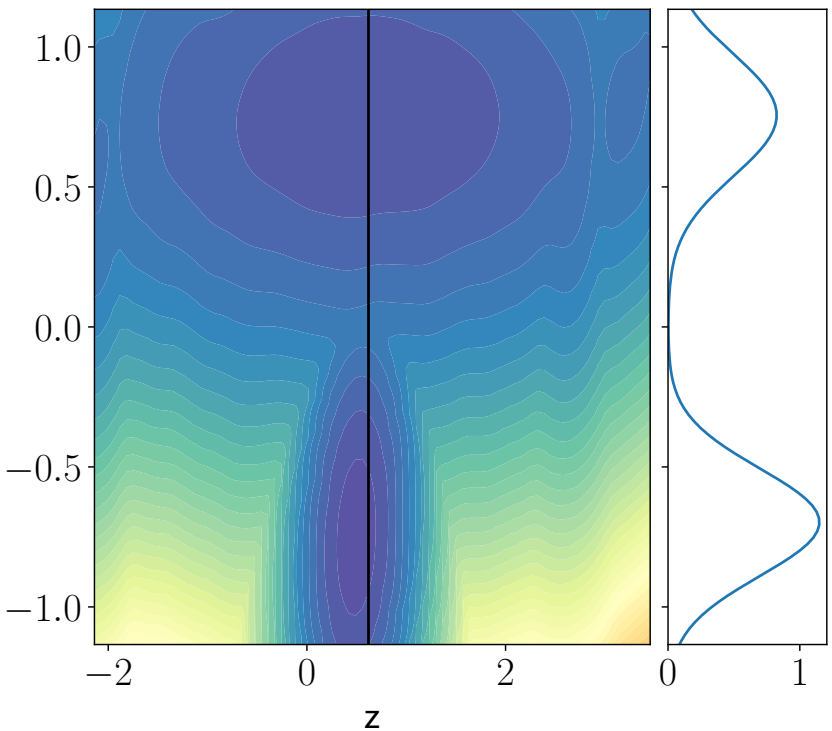
We compare with:

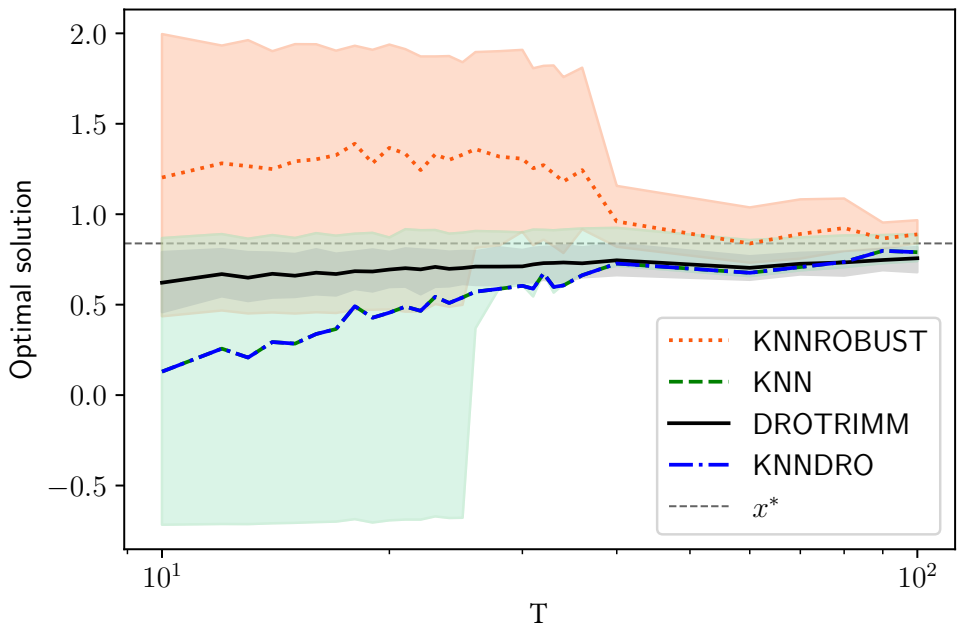
- KNN: KNN, i.e. Conditional SAA (Bertsimas and Kallus 2019).
- KNNROBUST: 2-step procedure; KNN+uncertainty set around the K nearest neighbors (Bertsimas, McCord, and Sturt 2019).
- KNNDRO: 2-step procedure; KNN+DRO Wasserstein ball centered at the K nearest neighbors (Bertsimas, McCord, and Sturt 2019).

Our method: DROTRIMM

Numerical experiments I

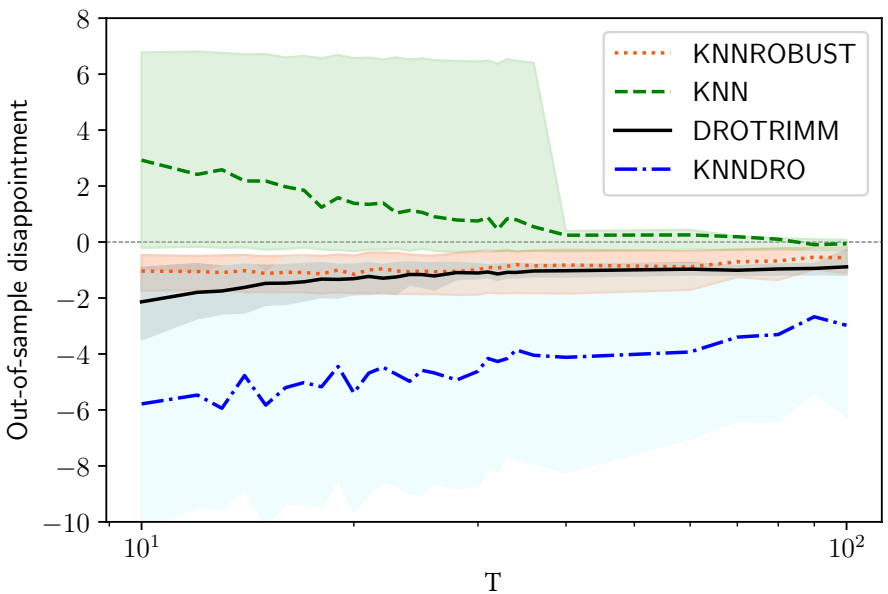
Single-item Newsvendor problem with side information

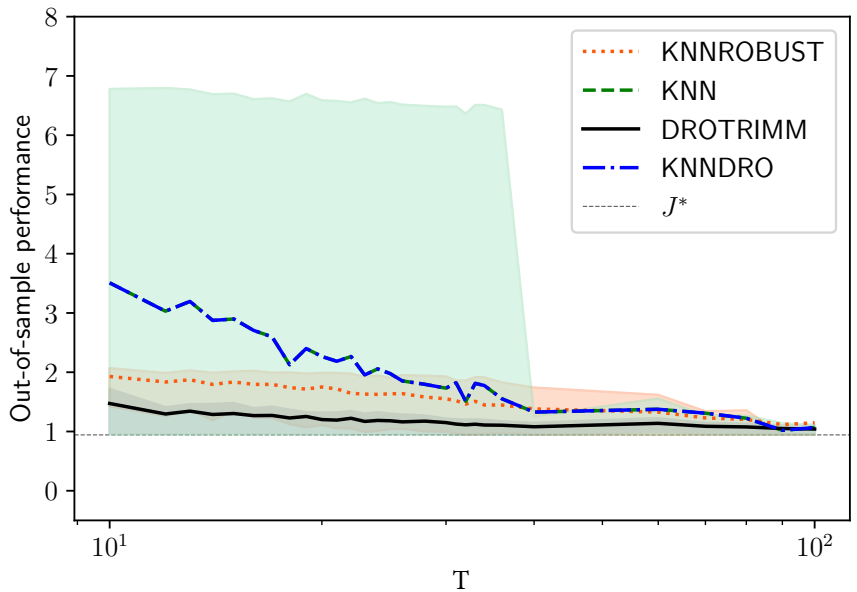




Numerical experiments I

Single-item Newsvendor problem with side information





Conclusions

- We have extended the DRO approach based on optimal transport to the Conditional Stochastic Optimization framework.
- Tractability is preserved.
- Performance guarantees are available.
- Significant improvements over alternative approaches.

References I

Bertsimas, Dimitris and Nathan Kallus (2019). "From predictive to prescriptive analytics". In: *Management Science*, mnsc.2018.3253.

Bertsimas, Dimitris, Christopher McCord, and Bradley Sturt (2019). "Dynamic optimization with side information". In: arXiv: 1907.07307.

Esteban-Pérez, Adrián and Juan M. Morales (2020). "Distributionally robust stochastic programs with side information based on trimmings". In: arXiv: 2009.10592.

Thanks for the attention!! Questions??

Juan Miguel Morales (juan.morales@uma.es)

Salvador Pineda (spineda@uma.es)

OASYS
Optimization and Analytics for Sustainable energY Systems

More info: oasys.uma.es

erc
European Research Council
Established by the European Commission

Supported by the project FlexAnalytics - Advanced Analytics to Empower the Small Flexible Consumers of Electricity.

Trimmed distributions

Trimmed distributions and Trimming Sets

Given $0 \leqslant \alpha \leqslant 1$ and probability measures $P, Q \in \mathbb{R}^d$, we say that Q is an $(1-\alpha)$-*trimming* of P if $Q \ll P$, and $\frac{dQ}{dP} \leqslant \frac{1}{\alpha}$.

- The *set of all* $(1-\alpha)$-*trimmings*, $\mathcal{R}_{1-\alpha}(P)$, has nice properties (convex, compact under the weak topology...).
- Particular case: $\mathcal{R}_{1-\alpha}(\widehat{\mathbb{Q}}_T)$ is the set of all probability distributions in the form $\sum_{t=1}^{T} b_t \delta_{\boldsymbol{\xi}_t}$ such that $0 \leq b_t \leq \frac{1}{T\alpha}$, $\forall t = 1, \ldots, T$, and $\sum_{t=1}^{T} b_t = 1$.
- We exploit the close link between Trimming sets and the Partial Optimal Transport to build our DRO framework.

Constructing the ambiguity set

Exploiting local information around $z = z_0$

The local information (in the joint sample data) must be selected accounting for the impact of its uncertainty on the objective function.

Using Partial Optimal Transport (POT)

Take some amount of mass (namely α) from the empirical joint distribution and transport it to another distribution supported in $\widetilde{\Xi}$ in the *cheapest* way:

$$\min_{R \in \mathcal{R}_{\mathbf{1}-\alpha}(\widehat{\mathbb{Q}}_T)} \mathcal{W}_p(R, Q_{\widetilde{\Xi}})$$

Constructing the ambiguity set

Ambiguity set

All distributions with support $\widetilde{\Xi}$ that result from a partial optimal transport of mass α from $\widehat{\mathbb{Q}}_T$ to $\widetilde{\Xi}$ within a budget $\widetilde{\rho}^{1/p}$:

$$\mathcal{W}_p(\mathcal{R}_{1-\alpha}(\widehat{\mathbb{Q}}_T), Q_{\widetilde{\Xi}}) \leq \widetilde{\rho}^{1/p}$$