Pontifícia Universidade Católica
do Rio de Janeiro

Carlos Andrés Gamboa Rodríguez

Conservative-solution methodologies for stochastic programming: A distributionally robust optimization approach

Tese de Doutorado

Thesis presented to the Programa de Pós–graduação em Engenharia de Produção of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia de Produção.

Advisor : Prof. Davi Michel Valladão
Co-Advisor: Prof. Alexandre Street de Aguiar

Rio de Janeiro
March 2021

Pontifícia Universidade Católica
do Rio de Janeiro

Carlos Andrés Gamboa Rodríguez

Conservative-solution methodologies for stochastic programming: A distributionally robust optimization approach

Thesis presented to the Programa de Pós–graduação em Engenharia de Produção of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia de Produção. Approved by the Examination Committee.

Prof. Davi Michel Valladão
Advisor
Departamento de Engenharia Industrial – PUC-Rio

Prof. Alexandre Street de Aguiar
Co-Advisor
Departamento de Engenharia Elétrica – PUC-Rio

Prof. Bernardo Kulnig Pagnoncelli
Universidad Adolfo Ibañez – UAI

Prof. Bruno Fânzeres dos Santos
Departamento de Engenharia Industrial – PUC-Rio

Dr. Bruno da Costa Flach
IBM Research Brazil – IBM

Prof. Tito Homem-de-Mello
Universidad Adolfo Ibañez – UAI

Rio de Janeiro, March 17th, 2021

Carlos Andrés Gamboa Rodríguez

Carlos Andrés Gamboa Rodríguez received in 2013 his B.Sc. degree in Mathematics from the National University of Colombia, Bogotá – Colombia (UN). In 2017, he received his M.Sc. degree in Industrial Engineering from Pontifical Catholic University, Rio de Janeiro – Brazil (PUC-Rio). During his years at academia, he has been actively participating on research projects in the Laboratory of Applied Mathematical Programming and Statistics (LAMPS) at the Electrical Engineering Department of PUC-Rio.

Ficha Catalográfica

Gamboa Rodríguez, Carlos Andrés

Conservative-solution methodologies for stochastic programming: A distributionally robust optimization approach /Carlos Andrés Gamboa Rodriguez; advisor: Davi Michel Valladão; co-advisor: Alexandre Street de Aguiar. – 2021.

v., 85 f: il. color. ; 30 cm

Tese (doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Industrial.

Inclui bibliografia

1. Engenharia Industrial – Teses. 2. Programação estocástica dois-estágios;. 3. Otimização robusta à distribuições;. 4. Métodos exatos de limites baseados na partição;. 5. Métodos de decomposição;. 6. Métrica de Wasserstein;. I. Valladão, Davi Michel. II. Street de Aguiar, Alexandre. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Industrial. IV. Título.

CDD: 658.5

Acknowledgments

Thank my advisor Prof. Davi Michel Valladão and co-advisor Alexandre Street for their support and encouragement for this work.

To my parents, family, and friends who supported me even with my absence in family life.

To LAMPS - Laboratory of Applied Mathematical Programming and Statistics, all colleagues, faculty, and staff of the Department of Industrial Engineering of PUC-Rio, for the fellowship, learning, and support.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Abstract

Gamboa Rodríguez, Carlos Andrés; Valladão, Davi Michel (Advisor); Street de Aguiar, Alexandre (Co-Advisor). **Conservative-solution methodologies for stochastic programming: A distributionally robust optimization approach**. Rio de Janeiro, 2021. Tese de Doutorado 85p. – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

Two-stage stochastic programming is a mathematical framework widely used in real-life applications such as power system operation planning, supply chains, logistics, inventory management, and financial planning. Since most of these problems cannot be solved analytically, decision-makers make use of numerical methods to obtain a near-optimal solution. Some applications rely on the implementation of non-converged and therefore sub-optimal solutions because of computational time or power limitations. In this context, the existing methods provide an optimistic solution whenever convergence is not attained. Optimistic solutions often generate high disappointment levels because they consistently underestimate the actual costs in the approximate objective function. To address this issue, we have developed two conservative-solution methodologies for two-stage stochastic linear programming problems with right-hand-side uncertainty and rectangular support: When the actual data-generating probability distribution is known, we propose a DRO problem based on partition-adapted conditional expectations whose complexity grows exponentially with the uncertainty dimensionality; When only historical observations of the uncertainty are available, we propose a DRO problem based on the Wasserstein metric to incorporate ambiguity over the actual data-generating probability distribution. For this latter approach, existing methods rely on dual vertex enumeration of the second-stage problem rendering the DRO problem intractable in practical applications. In this context, we propose algorithmic schemes to address the computational complexity of both approaches. Computational experiments are presented for the farmer problem, aircraft allocation problem, and the stochastic unit commitment problem.

Keywords

Two-stage stochastic programming; Distributtionally robust optimization; Exact bound partition methods; Decomposition methods; Wasserstein metric;

Resumo

Gamboa Rodríguez, Carlos Andrés; Valladão, Davi Michel; Street de Aguiar, Alexandre. **Metodologias para obtenção de soluções conservadoras para programação estocástica: Uma abordagem de otimização robusta à distribuições**. Rio de Janeiro, 2021. Tese de Doutorado 85p. – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

A programação estocástica dois estágios é uma abordagem matemática amplamente usada em aplicações da vida real, como planejamento da operação de sistemas de energia, cadeias de suprimentos, logística, gerenciamento de inventário e planejamento financeiro. Como a maior parte desses problemas não pode ser resolvida analiticamente, os tomadores de decisão utilizam métodos numéricos para obter uma solução quase ótima. Em algumas aplicações, soluções não convergidas e, portanto, sub-ótimas terminam sendo implementadas devido a limitações de tempo ou esforço computacional. Nesse contexto, os métodos existentes fornecem uma solução otimista sempre que a convergência não é atingida. As soluções otimistas geralmente geram altos níveis de arrependimento porque subestimam os custos reais na função objetivo aproximada. Para resolver esse problema, temos desenvolvido duas metodologias de solução conservadora para problemas de programação linear estocástica dois estágios com incerteza do lado direito e suporte retangular: Quando a verdadeira distribuição de probabilidade da incerteza é conhecida, propomos um problema DRO (*Distributionally Robust Optimization*) baseado em esperanças condicionais adaptadas à uma partição do suporte cuja complexidade cresce exponencialmente com a dimensionalidade da incerteza; Quando apenas observações históricas da incerteza estão disponíveis, propomos um problema de DRO baseado na métrica de Wasserstein a fim de incorporar ambiguidade sobre a real distribuição de probabilidade da incerteza. Para esta última abordagem, os métodos existentes dependem da enumeração dos vértices duais do problema de segundo estágio, tornando o problema DRO intratável em aplicações práticas. Nesse contexto, propomos esquemas algorítmicos para lidar com a complexidade computacional de ambas abordagens. Experimentos computacionais são apresentados para o problema do fazendeiro, o problema de alocação de aviões, e o problema do planejamento da operação do sistema elétrico (*unit ommitmnet problem*).

Palavras-chave

Programação estocástica dois-estágios; Otimização robusta à distribuições; Métodos exatos de limites baseados na partição; Métodos de decomposição; Métrica de Wasserstein;

Table of contents

List of figures

List of tables

1
Introduction

Two-stage stochastic programming is a mathematical framework to model decision making under uncertainty. In this context, first-stage decisions are made under uncertainty, while the second-stage (or recourse) decisions are intended to correct the impact of the first-stage decisions after observing the uncertainty realization. This framework is widely used in real-world applications covering the planning of power system operations and supply chain, logistics, inventory, and financial planning, to mention a few. Since there is no analytical solution to most two-stage stochastic optimization problems, efficient numerical methods are of paramount importance.

One of the first numerical methods explored in the stochastic programming literature utilized bounds for the expected recourse function. Initiated by [2], this approach considers minimization problems of which the recourse is a convex function of the uncertainty. It relies on a partition of the uncertainty support, the law of total probability, and Jensen's and Edmundson-Madanski's inequalities in order to obtain lower and upper bounds for the optimal value of the problem. Partition refinement methods ensure improvements to the approximation error with a monotone sequence of limits (see [3–7]). It is worth mentioning that this numerical approach provides an exact solution for the stochastic optimization problem. Indeed, it is based on numerical integration techniques known for a long time which is why the referenced works are fairly old.

With the advancement of computers, this approach was replaced by sampling-based techniques, (see, e.g., [8] for a review). In data-driven settings the so-called *Sample Average Approximation* (SAA) approach, a terminology introduced by [9], is the most popular method to solve two-stage stochastic problems because of its asymptotic convergence and tractability. For a cost minimization problem, the SAA method is based on the approximation of the expected cost by the sample average cost. The obtained solution is an statistical estimator whose quality can be assessed by an interval confidence level. The randomness of the obtained solution may be a practical difficulty in the SAA method. On the other hand, the approximated solution of the SAA is optimistic since this method considers just a subset of scenarios in the decision-making process. Moreover, as mentioned in [1,10], in many settings the SAA solution tends to display a poor out-of-sample performance and may be highly unstable for a finite moderate sample.

Indeed, in a centralized system setting, a sampling method is not appropriate because of the randomness of the obtained solution. Moreover, to obtain a converged solution without randomness – exact solution – or low variability – SAA solution for a large sample – entails computational overburden. Additionally, some practical applications have a computing-time budget, whereby non-converged sub-optimal solutions are usually implemented. It is important to note that all existing methods (exact bounds or sampling-based) generates optimistic non-converged sub-optimal solutions that usually generates high disappointment levels: out-of-sample costs significantly higher than in-sample estimates. In this context, we believe that an audited decision with computing-time budget must have a solution methodology that provides conservative non-converged sub-optimal solutions at every step the algorithm to avoid practical consequences of cost disappointment.

To address this issue, our main objective is to study solution methodologies to obtain exact conservative solutions for two-stage stochastic linear programming problems with right-hand-side uncertainty and rectangular support. In this context, we focus on Distributionally Robust Optimization (DRO), which is a mathematical framework that minimizes the worst-case of the expected cost over an ambiguity set comprising plausible data-generating probability distributions. We developed two DRO approaches assuming we know: (i) the actual data-generating (continuous) probability distribution; (ii) historical observations of the uncertainty.

The proposed conservative solution methodologies are motivated by the idea that distributionally robust solutions avoid out-of-sample disappointment. In both cases, we face a computational challenge because of the exponential complexity of the problem. To address this issue, we propose algorithmic schemes based on exact decomposition methods by exploring specific characteristics for computational tractability.

We summarize below our main contributions to conservative solution methodologies for two-stage stochastic problems with right-hand-side uncertainty and rectangular support:

- Raise awareness on the negative impact of optimistic solution methods to solve two-stage stochastic optimization problems.

- A new conservative solution framework for two-stage stochastic linear optimization problems based on a deterministic partition refinement algorithm and exact bounds, assuming a continuous probability distribution.
  - Reformulation of the upper-bound, conservative, distributionally robust problem into a linear programming model.
  - Development of two acceleration procedures, i.e., 1) an exact decomposition approach based on the column and constraint generation algorithm to solve medium-scaled problems and 2) an approximative simplex-based partitioning scheme to find robust solutions for large-scale instances.
- A new decomposition scheme to solve the DRO problem with a Wasserstein ambiguity set and the rectangular uncertainty support.
  - We develop a novel master-oracle decomposition framework based on a new and exact MILP-based oracle subproblem.
  - We develop three decomposition methods, namely, Column-Constraint Generation, Single-cut Benders, and Multi-cut Benders.
  - We illustrate the computational performance of the proposed methods for a unit commitment problem with 5, 14, and 54 thermal generators over a 24-hour uncertainty dimension.

The remainder of this thesis is organized as follows: Chapter 2 presents the novel conservative partition-refinement method for the case when the actual data-generating distribution is known. In Chapter 3 is presented the DRO approach with Wasserstein ambiguity set for the case when only historical data is available. Chapter 4 presents numerical experiments based on the unit commitment problem for assessing the proposed solution methodology presented in Chapter 3. Finally, in Chapter 5, relevant conclusions are provided for both conservative solution methodologies.

2
Moment-based distributionally robust optimization approach

One of the first numerical methods explored in the stochastic programming literature utilized bounds for the expected recourse function. It relies on a partition of the uncertainty support, the law of total probability, and Jensen's and Edmundson-Madanski inequalities in order to obtain lower and upper bounds. Starting from the law of total probability, previous works have provided an optimistic solution by applying Jensen's inequality to each partition cell to obtain the lower approximation problem in a computationally tractable fashion. The optimal value of this proxy provides a lower bound to the original problem because it replaces the expected recourse function by a weighted average recourse evaluation. A deterministic optimality gap is obtained by fixing the current solution and computing an upper-bound that applies Edmundson-Madanski inequality to each cell of the partition and averages its results with the associated probability mass. Given a refining partition sequence, this method converges to the true optimal in an optimistic manner, i.e., at each iteration, the solution underestimates the actual recourse costs. The available partition refinement methods in literature only provide optimistic solutions when the method is not converged. The available partition refinement methods in literature only provide optimistic solutions when the method is not converged.

By assuming that the actual data-generating distribution is known, we developed a conservative solution methodology for the class of two-stage stochastic linear optimization problems with right-hand-side uncertainty and rectangular support. We formulated a distributionally robust optimization model based on a generalization of Edmundson-Madanski inequality (see [3, 11, 12]), and solved it to obtain a conservative solution and a tighter upper bound to the original problem. The distributionally robust model minimizes the worst expected cost over every extreme probability measure with known partition-adapted conditional expectations. A deterministic optimality gap was obtained by solving the lower-bound problem that applied Jensen's inequality to each partition cell and computing the distance between both limits.

Nevertheless, in the presence of high dimensional uncertainty vectors, the proposed method is challenged because of the exponential growth of the number of linear constraints and variables. To handle this, we propose different solution schemes depending on the uncertainty's dimensionality: (i) for problems with low-dimensional uncertainty, we developed a deterministic equivalent linear programming model, (ii) for medium-sized uncertainty dimensionality, we propose a column and constraint generation algorithm [13], and (iii) to handle high dimensional uncertainty, we propose a simplex-based heuristic method whose complexity grows linearly with the uncertainty dimension. For the latter, we prove convergence when the recourse function is monotone over the uncertainty.

2.1 Theoretical background for the moment-based approach

In this section, we introduce the nomenclature and provide a short review of the existing partition methods from which our contributions of the first-moment-based approach are derived. To clarify the notation, we begin by introducing some definitions of probability theory. We assume a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and the random vector $\boldsymbol{\xi} : \Omega \longrightarrow \mathbb{R}^{d_\xi}$, which is called the uncertain variable. Given the measurable space $(\mathbb{R}^{d_\xi}, \mathcal{B})$, where $\mathcal{B}$ denotes the Borel σ-algebra of $\mathbb{R}^{d_\xi}$, we denote by P the distribution of $\boldsymbol{\xi}$ which is a probability for the measurable space $(\mathbb{R}^{d_\xi}, \mathcal{B})$. This probability is defined by $P(A) = \mathbb{P}(\{\omega \in \Omega \,:\, \boldsymbol{\xi}(\omega) \in A\}) = \mathbb{P}(\boldsymbol{\xi} \in A)$, for all $A \in \mathcal{B}$.

2.2 Notation and definitions

We study two-stage stochastic linear programming problems of the form:

$$z^* := \min_{\mathbf{x} \in X} \{ f(\mathbf{x}) := \mathbf{c}^\top \mathbf{x} + \mathbb{E}[Q(\mathbf{x}, \boldsymbol{\xi})] \} \tag{2-1}$$

where

$$Q(\mathbf{x}, \boldsymbol{\xi}) = \min_{\mathbf{y} \geq 0} \mathbf{q}^\top \mathbf{y} \tag{2-2a}$$

$$\text{s.t.} \quad \mathbf{W}\mathbf{y} = \mathbf{H}(\mathbf{x})\boldsymbol{\xi} + \mathbf{r}(\mathbf{x}). \tag{2-2b}$$

is known as the recourse function or second-stage problem [1]. This model corresponds to a linear optimization problem that minimizes the cost where

[1]While this formulation (which we borrow from [14]) is not in the form usually associated with two-stage models in which the second-stage constraints are written as $\mathbf{W}\mathbf{y} = \mathbf{h}(\xi) - \mathbf{T}(\xi)\mathbf{x})$, it easy to see that we can rewrite these equations as in (2-2b) by representing the random elements of $\mathbf{h}$ and $\mathbf{T}$ as $\boldsymbol{\xi}$ and defining $\mathbf{H}(\mathbf{x})$ and $\mathbf{r}(\mathbf{x})$ appropriately.

$\mathbf{x}$ denotes the first-stage decisions, $\mathbf{y}$ denotes the second-stage decisions and the expectation $\mathbb{E}[Q(\mathbf{x}, \boldsymbol{\xi})]$ represents the expected cost of the recourse. We assume that the random vector $\boldsymbol{\xi}$ has known probability distribution P with support in $\Xi \subseteq \mathbb{R}^{d_\xi}$ and the expectation $\mathbb{E}[\cdot]$ is taken with respect to the distribution P[2]. We denote the set of feasible first-stage decisions by $X \subseteq \mathbb{R}^{d_x}$. Here, $\mathbf{c} \in \mathbb{R}^{d_x}$, $\mathbf{q} \in \mathbb{R}^{d_y}$, $\mathbf{W} \in \mathbb{R}^{m_y \times d_y}$.
In the problem (2-2), $\mathbf{H}(\mathbf{x}) \in \mathbb{R}^{m_y \times d_\xi}$ and $\mathbf{r}(\mathbf{x}) \in \mathbb{R}^{m_y}$ represent a decision-dependent matrix and vector, respectively. We assume that $\mathbf{H} : X \to \mathbb{R}^{m_y \times d_\xi}$ and $\mathbf{r} : X \to \mathbb{R}^{m_y}$ are affine functions of $\mathbf{x}$. Moreover, we assume that problem (2-1) has complete recourse, i.e., problem (2-2) is feasible for every $\mathbf{x} \in X$, and every realization of the unknown data $\boldsymbol{\xi}$; the uncertainty support Ξ is compact and the expectation $\mathbb{E}[Q(\mathbf{x}, \boldsymbol{\xi})]$ exists.

The existing partition-based method [2,4,5,7,11,15] makes use of a partition of the support of the uncertainty to generate a monotonic sequence of limits. These limits are given by the optimal objective values of the upper- and lower-bound problems obtained by the classical inequalities of Edmundson-Mandanski and Jensen, respectively.

According to previous reported partition methods, we start from the partition of the uncertainty support Ξ. The set $\mathscr{P}_n = \{\Xi^k : k = 1, \ldots, n\}$ of cells Ξ^k, is a partition of the support $\Xi \subset \mathbb{R}^{d_\xi}$ if:

1. $P\left(\bigcap_{k\in K} \Xi^k\right) = 0, \quad \forall K \subseteq \{1, \ldots, n\},$
2. $\bigcup_{k=1}^{n} \Xi^k = \Xi.$

Since most real-life applications are based on bounded probability distributions for the uncertainty, in this work, we assume rectangular support for the data-generating probability distribution. Moreover, we assume rectangular partitions with cells of the form $\Xi^k = \bigtimes_{i=1}^{d_\xi} [a_i^k, b_i^k]$, because the partition refinement procedure for this partition regards the method computational tractable.

Finally, even thought the disappointment concept has been widely studied by the risk-averse theory and there exist several definitions under this concept, in this work, we assume the following definition:

Definition 1. *The disappointment is the difference between the out-of-sample cost and the in-sample estimate.*

[2]Sometimes we write $\mathbb{E}^{\widetilde{P}}[\cdot]$ to emphasise that the expectation $\mathbb{E}[\cdot]$ is taken with respect to the probability distribution $\widetilde{P}$.

2.3
Existing lower bound

According to [2], if we consider a partition of the uncertainty support Ξ, by the law of total probability, we have that the expectation in (2-1) can be expressed by $\mathbb{E}[Q(\mathbf{x},\boldsymbol{\xi})] = \sum_{k=1}^{n} p^k \mathbb{E}[Q(\mathbf{x},\boldsymbol{\xi})|\boldsymbol{\xi} \in \Xi^k]$, where $p^k = P(\Xi^k)$ is the probability mass of the cell k, $k = 1,\ldots,n$. Then (2-1) is equivalent to the following linear programming problem:

$$\min_{\mathbf{x}\in X} \left\{ f(\mathbf{x}) := \mathbf{c}^\top \mathbf{x} + \sum_{k=1}^{n} p^k \mathbb{E}[Q(\mathbf{x},\boldsymbol{\xi})|\boldsymbol{\xi} \in \Xi^k] \right\} \tag{2-3}$$

For the first-moment-based approach, we consider as nominal value of $\boldsymbol{\xi}$ the conditional mean $\overline{\boldsymbol{\xi}}^k = \mathbb{E}[\boldsymbol{\xi}|\boldsymbol{\xi} \in \Xi^k]$, for each cell $k = 1,\ldots,n$. Given that $Q(\mathbf{x},\boldsymbol{\xi})$ is convex in $\boldsymbol{\xi}$ for all $\mathbf{x} \in X$, by Jensen's inequality we have that

$$Q(\mathbf{x},\overline{\boldsymbol{\xi}}^k) \leq \mathbb{E}[Q(\mathbf{x},\boldsymbol{\xi})|\boldsymbol{\xi} \in \Xi^k], \quad k = 1,\ldots,n,$$

which gives the following lower-bound for the optimal objective value of (2-3):

$$\begin{aligned} z_n^L := & \min_{\mathbf{x}\in X} \left\{ f_n^L(\mathbf{x}) := \mathbf{c}^\top \mathbf{x} + \sum_{k=1}^{n} p^k Q(\mathbf{x},\overline{\boldsymbol{\xi}}^k) \right\} \\ & \leq \min_{\mathbf{x}\in X} \left\{ \mathbf{c}^\top \mathbf{x} + \sum_{k=1}^{n} p^k \mathbb{E}[Q(\mathbf{x},\boldsymbol{\xi})|\boldsymbol{\xi} \in \Xi^k] \right\}. \end{aligned} \tag{2-4}$$

Note that the lower-bound problem on the left side of Eq. (2-4) underestimates the expected cost, because it only considers the finite set of conditional mean scenarios $\{\overline{\boldsymbol{\xi}}^k : k = 1,\ldots,n\}$ to approximate $\mathbb{E}[Q(\mathbf{x},\boldsymbol{\xi})]$.

We can explicitly write the lower-bound problem as the following deterministic-equivalent linear program:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y}^k} \quad & \mathbf{c}^\top \mathbf{x} + \sum_{k=1}^{n} p^k \mathbf{q}^\top \mathbf{y}^k \\ \text{s.t.} \quad & \mathbf{W}\mathbf{y}^k = \mathbf{H}(\mathbf{x})\overline{\boldsymbol{\xi}}^k + \mathbf{r}(\mathbf{x}), \quad k = 1,\ldots,n, \\ & \mathbf{y}^k \geq 0, \quad k = 1,\ldots,n, \\ & \mathbf{x} \in X. \end{aligned} \tag{2-5}$$

Note that the number of blocks of linear constraints for (2-5) is the same as the partition size, meaning that the lower-bound problem does not demand a high computational effort. When n is not sufficiently large, the optimal solution $\mathbf{x}_n^L$ of (2-5) represents an optimistic decision because the lower approximation from bellow to the conditional expectation $\mathbb{E}[Q(\mathbf{x},\boldsymbol{\xi})|\boldsymbol{\xi} \in \Xi^k]$ by $Q(\mathbf{x},\overline{\boldsymbol{\xi}}^k)$, $k = 1,\ldots,n$.

It is worth mentioning that generally the computing of $p^k = P(\Xi^k)$ and $\overline{\boldsymbol{\xi}}^k = \mathbb{E}[\boldsymbol{\xi}|\boldsymbol{\xi} \in \Xi^k$ is not easy. By assuming independence or linear correlation, this computing derives in a uni-dimensional numerical integration problem which is easily solved by computational tools. However, in general, this computing can be done by using the importance sample technique in a tractable fashion. Under this technique, we can compute the probability mass p^k and the conditional expectation $\overline{\boldsymbol{\xi}}^k$ by generating samples from an independent multivariate uniform probability distribution.

2.4 Existing upper bound

To derive an upper bound for the expectation $\mathbb{E}[Q(\mathbf{x}, \boldsymbol{\xi})]$, Edmundson-Madanski [16] initially proposed an inequality based on a convex combination of the value of $Q(\mathbf{x}, \cdot)$ at the extreme points of the convex hull of the support Ξ. Therefore, if $\{\mathbf{e}_j \ : \ j = 1, \ldots, 2^{d_\xi}\}$ is the set of extreme points of $\Xi = \bigtimes_{i=1}^{d_\xi} [a_i, b_i]$, since every point $\boldsymbol{\xi} \in \Xi$ can be expressed by a convex combination of the vertices of Ξ, it is always possible to find $p_j(\boldsymbol{\xi}) \geq 0$, $j = 1, \ldots, 2^{d_\xi}$, such that

$$\sum_{j=1}^{2^{d_\xi}} p_j(\boldsymbol{\xi}) \cdot \mathbf{e}_j = \boldsymbol{\xi} \quad \forall \boldsymbol{\xi} \in \Xi \tag{2-6}$$

and

$$\sum_{j=1}^{2^{d_\xi}} p_j(\boldsymbol{\xi}) = 1 \quad \forall \boldsymbol{\xi} \in \Xi. \tag{2-7}$$

The weights $p_j(\boldsymbol{\xi})$, $j = 1, \ldots, 2^{d_\xi}$, can be interpreted as conditional probabilities, i.e., $p_j(\boldsymbol{\xi}) = P(\mathbf{e} = \mathbf{e}_j \,|\, \boldsymbol{\xi})$, considering a random vector $\mathbf{e}$ with support in the set of extreme points $\{\mathbf{e}_j \ : \ j = 1, \ldots, 2^{d_\xi}\}$. Given that (2-6) and (2-7) are true and $Q(\mathbf{x}, \cdot)$ is a convex function for any given $\mathbf{x}$, the following inequality holds for any set of conditional probabilities $\{(p_1(\boldsymbol{\xi}), \ldots, p_{2^{d_\xi}}(\boldsymbol{\xi}))\}_{\boldsymbol{\xi} \in \Xi}$:

$$\mathbb{E}[Q(\mathbf{x}, \boldsymbol{\xi})] = \int_\Xi Q\left(\mathbf{x}, \sum_{j=1}^{2^{d_\xi}} p_j(\boldsymbol{\xi})\mathbf{e}_j\right) P(d\boldsymbol{\xi}) \leq \sum_{j=1}^{2^{d_\xi}} Q(\mathbf{x}, \mathbf{e}_j) \int_\Xi p_j(\boldsymbol{\xi}) P(d\boldsymbol{\xi}).$$

Therefore,

$$\begin{aligned}
\mathbb{E}[Q(\mathbf{x},\boldsymbol{\xi})] \leq \quad & \max_{p_j(\xi)} \sum_{j=1}^{2^{d_\xi}} Q(\mathbf{x},\mathbf{e}_j) \int_{\Xi} p_j(\boldsymbol{\xi}) dP(d\boldsymbol{\xi}) \\
& \text{s.t.} \sum_{j=1}^{2^{d_\xi}} \mathbf{e}_j p_j(\boldsymbol{\xi}) = \boldsymbol{\xi} \quad \forall \boldsymbol{\xi} \in \Xi \\
& \sum_{j=1}^{2^{d_\xi}} p_j(\boldsymbol{\xi}) = 1 \quad \forall \boldsymbol{\xi} \in \Xi.
\end{aligned} \tag{2-8}$$

Note that (2-8) is a semi-infinite optimization problem that allows us to derive an upper bound through a tractable finite relaxation. If we replace the two set of constraints in (2-8) with their expected value, i.e.,

$$\sum_{j=1}^{2^{d_\xi}} \mathbf{e}_j \int_{\Xi} p_j(\boldsymbol{\xi}) dP(d\boldsymbol{\xi}) = \overline{\boldsymbol{\xi}}$$

$$\sum_{j=1}^{2^{d_\xi}} \int_{\Xi} p_j(\boldsymbol{\xi}) dP(d\boldsymbol{\xi}) = 1,$$

and denote $\delta_j = \int_{\Xi} p_j(\boldsymbol{\xi}) P(d\boldsymbol{\xi})$, for $j = 1, \ldots, 2^{d_\xi}$, we obtain an upper bound based on the following tractable finite linear optimization problem:

$$\begin{aligned}
\max_{\delta \in \mathcal{D}(\overline{\xi})} \mathbb{E}^{\delta}\left[Q(\mathbf{x},\mathbf{e})\right] := \max_{\delta \geq 0} \quad & \sum_{j=1}^{2^{d_\xi}} Q(\mathbf{x},\mathbf{e}_j)\, \delta_j \\
\text{s.t.} \quad & \sum_{j=1}^{2^{d_\xi}} \mathbf{e}_j\, \delta_j = \overline{\boldsymbol{\xi}}, \\
& \sum_{j=1}^{2^{d_\xi}} \delta_j = 1,
\end{aligned} \tag{2-9}$$

where $\mathcal{D}(\overline{\boldsymbol{\xi}}) = \left\{ \delta \in \mathbb{R}_{+}^{2^{d_\xi}} \ : \ \sum_{j=1}^{2^{d_\xi}} \mathbf{e}_j\, \delta_j = \overline{\boldsymbol{\xi}}, \quad \sum_{j=1}^{2^{d_\xi}} \delta_j = 1 \right\}$. Thus, as per the above developments, the following inequality holds:

$$\mathbb{E}[Q(\mathbf{x},\boldsymbol{\xi})] \leq \max_{\delta \in \mathcal{D}(\overline{\xi})} \mathbb{E}^{\delta}\left[Q(\mathbf{x},\mathbf{e})\right]. \tag{2-10}$$

Existing partition-based methods used the derived the upper bound (2-10) for a first-stage decision found by the lower bound problem (2-5). So, supposing that $\mathbf{x}_n^L$ is the solution of (2-5), in [4,7,12] the following upper bound is proposed:

$$z^* \leq f_n^U(\mathbf{x}_n^L) := \mathbf{c}^\top \mathbf{x}_n^L + \sum_{k=1}^{n} p^k \max_{\delta \in \mathcal{D}(\overline{\xi}^k)} \left(\sum_{j=1}^{2^{d_\xi}} \delta_j^k Q(\mathbf{x}_n^L, \mathbf{e}_j^k) \right), \tag{2-11}$$

where,
$$Q(\mathbf{x}_n^L, \mathbf{e}_j^k) := \min_{\mathbf{y}\geq 0} \quad \mathbf{q}^\top \mathbf{y}$$
$$\text{s.t.} \quad \mathbf{W}\mathbf{y} = \mathbf{H}(\mathbf{x}_n^L)\mathbf{e}_j^k + \mathbf{r}(\mathbf{x}_n^L), \tag{2-12}$$

for all $j = 1, \ldots, 2^{d_\xi}$ and $k = 1, \ldots, n$.

The existing partition-based method is summarized in the following pseudo-algorithm:

Algorithm 1 Existing partition-based methods

Require: $\mathscr{P}^1 = \{\Xi\}$, $\epsilon > 0$ stopping criteria.

Ensure: $\mathbf{x}_n^L$ the optimistic solution of the lower bound problem.

1: Solve (2-5) to determine the optimal solution $\mathbf{x}_n^L$.
2: Evaluate $\mathbf{x}_n^L$ in the recourse problem (2-12) for $j = 1, \ldots, 2^{d_\xi}$, and $k = 1, \ldots, n$, to get an upper bound for the optimal objective value of (2-3), according to Eq. (2-11).
3: **if** $\frac{f_n^U(\mathbf{x}_n^L) - z_n^L}{f_n^U(\mathbf{x}_n^L)} \leq \epsilon$ **then**
4: stop and return $\mathbf{x}_n^L$ as the optimal solution.
5: **else**
6: refine the partition $\mathscr{P}_n$ to $\mathscr{P}_{n+1}$ and return to 1
7: **end if**

The existing partition-based methods are based on the optimistic lower bound problem, as it underestimates the recourse cost by relying only on the partition-adapted conditional means within the recourse function assessment. Nevertheless, optimistic solutions can generate high disappointment levels when they are evaluated for adverse scenarios of the realization of the uncertainty as shown in Section 2.9.1 of the computational experiments. However, as it is pointed by [17], conservative solutions obtained by solving distributionally robust optimization problems avoid out-of-sample disappointments that quantify the risk that the actual expected cost of the candidate decision exceeds its predicted cost.

In this next section, we present a distributionally robust optimization approach based on the first moment which is a new conservative solution framework for two-stage stochastic linear optimization problems.

2.5
Proposed upper-bound problem reformulation

Motivated by the concept explored in [1], i.e., conservative solutions obtained by solving distributionally robust optimization problems – under certain conditions– avoid out-of-sample disappointments, our thrust is to obtain pre-convergence conservative solutions in a computationally efficient manner. For that, we develop in this section a deterministic equivalent reformulation of the upper-bound (distributionally robust) problem that is computationally tractable for low-dimensional uncertainty. We propose a column-and-constraint (C&CG) generation algorithm for medium-sized uncertainty dimensionality and, to handle high dimensional uncertainty, we also propose a simplex-based heuristic method whose complexity grows linearly with the uncertainty dimension. In the presence of monotone recourse functions with regard to an uncertain parameter, we prove convergence of the proposed simplex-based heuristic method. The assumption of monotonicity of the recourse function is reasonable in light of the fact that most real-world applications belong to this class of problem.

2.5.1
Deterministic equivalent model for upper-bound problem

Following the upper approximation of the expected recourse function (2-10), we define the upper-bound problem

$$z_n^U := \min_{\mathbf{x}\in X}\left\{ f_n^U(\mathbf{x}) := \mathbf{c}^\top \mathbf{x} + \sum_{k=1}^{n} p^k \max_{\delta\in\mathcal{D}(\overline{\boldsymbol{\xi}}^k)} \mathbb{E}^\delta\left[Q(\mathbf{x},\mathbf{e}^k)\right]\right\}, \tag{2-13}$$

where $\mathbf{e}^k$ is the random vector whose support is the set of vertex of the hypercube Ξ^k and $\mathcal{D}(\overline{\boldsymbol{\xi}}^k)$ is the ambiguity set generated by the conditional mean $\overline{\boldsymbol{\xi}}^k$ of the cell Ξ^k.

The solutions obtained from the upper-bound problem can be seen as robust or conservative solutions. Note that problem (2-13) is a distributionally robust optimization problem, where the conditional-probability distribution within each cell is selected to represent the worst-case distribution preserving the conditional-average information of the cell [18]. Since the upper-bound problem (2-13) overestimates the actual cost (2-3), the upper-bound solution has a mathematical certificate against disappointment. This is specially useful when non-converged solutions are actually implemented owing to time or computational-power limitations.

However, solving the upper-bound problem to obtain the conservative solution $\mathbf{x}_n^U$ requires significant computational effort. The number of variables of the inner problem grows exponentially with the uncertainty dimension. In this work, we propose a new framework to obtain a conservative solution by solving the upper-bound problem. To the best of our knowledge, no existing partition-based method optimizes the upper bound problem.

For a given cell Ξ^k of the rectangular partition, the upper-bound for the conditional-expected recourse cost (2-9) can be recast according to its dual formulation as follows:

$$\begin{aligned}\max_{\delta\in\mathcal{D}(\overline{\boldsymbol{\xi}}^k)} \mathbb{E}^\delta\left[Q(\mathbf{x},\mathbf{e}^k)\right] = \min_{\pi,\boldsymbol{\eta}^k} \quad & \pi + [\boldsymbol{\eta}^k]^\top \overline{\boldsymbol{\xi}}^k \\ \text{s.t.} \quad & \pi + [\boldsymbol{\eta}^k]^\top \mathbf{e}_j^k \geq Q(\mathbf{x},\mathbf{e}_j^k), \quad j = 1,\ldots,2^{d_\xi}.\end{aligned} \tag{2-14}$$

where $\pi^k \in \mathbb{R}$ and $\boldsymbol{\eta}^k \in \mathbb{R}^{d_\xi}$ be the dual variables associated with the first and second linear constraints of (2-9), respectively.

Replacing the worst-case conditional-expected recourse cost of (2-13) with its dual formulation (2-14), the upper-bound problem (2-13) can be recast as the following linear optimization problem:

$$
\begin{aligned}
\min_{\pi^k,\boldsymbol{\eta}^k,\mathbf{x}} \quad & \mathbf{c}^\top \mathbf{x} + \sum_{k=1}^{n} p^k(\pi^k + [\boldsymbol{\eta}^k]^\top \overline{\boldsymbol{\xi}}^k) \\
\text{s.t.} \quad & \pi^k + [\boldsymbol{\eta}^k]^\top \mathbf{e}_j^k \geq Q(\mathbf{x}, \mathbf{e}_j^k), \quad j = 1, \ldots, 2^{d_\xi}, \, k = 1, \ldots, n \\
& \mathbf{x} \in X.
\end{aligned}
\tag{2-15}
$$

Since the recourse function is a minimization problem, we are able to obtain the deterministic equivalent

$$
\begin{aligned}
\min_{\pi^k,\boldsymbol{\eta}^k,\mathbf{x},\mathbf{y}} \quad & \mathbf{c}^\top \mathbf{x} + \sum_{k=1}^{n} p^k(\pi^k + [\boldsymbol{\eta}^k]^\top \overline{\boldsymbol{\xi}}^k) \\
\text{s.t.} \quad & \left.\begin{aligned}
& \pi^k + [\boldsymbol{\eta}^k]^\top \overline{\boldsymbol{\xi}}^k \geq \mathbf{q}^\top \mathbf{y}_{\overline{\boldsymbol{\xi}}^k} \\
& \mathbf{W}\mathbf{y}_{\overline{\boldsymbol{\xi}}^k} = \mathbf{H}(\mathbf{x})\overline{\boldsymbol{\xi}}^k + \mathbf{r}(\mathbf{x}) \\
& \left.\begin{aligned}
\pi^k + [\boldsymbol{\eta}^k]^\top \mathbf{e}_j^k \geq \mathbf{q}^\top \mathbf{y}_j^k \\
\mathbf{W}\mathbf{y}_j^k = \mathbf{H}(\mathbf{x})\mathbf{e}_j^k + \mathbf{r}(\mathbf{x})
\end{aligned}\right\} \forall j \in J^k
\end{aligned}\right\} \quad \forall k = 1, \ldots, n \\
& \mathbf{x} \in X.
\end{aligned}
\tag{2-16}
$$

which is an equivalent formulation of the original problem (2-15) whenever J^k comprises all vertices of cell Ξ^k. This equivalent formulation consists in replacing the recourse function by the second-stage objective function, including all second stage decisions as variables and adding all second stage feasibility constraints. Indeed, if the left-hand-side (LHS) of the first block of constraints is greater than or equal to the second stage cost of a feasible second-stage solution, then the LHS is greater than or equal to the minimum second stage cost given by the recourse function.

Note for instance that (2-16) is a linear programming problem whenever X is a polyhedral set and can be efficiently solved whenever for problems with low dimension uncertainty vector. However, problem (2-16) is in general an intractable problem for medium- and large-scale instances as it relies on an exponential set of constraints. To handle this challenge, we present two acceleration procedures: 1) an exact decomposition approach based on the column and constraint generation algorithm [13] to solve medium-dimensional problems and 2) an heuristic procedure for high-dimensional uncertainty based on a simplex-based heuristic method using a circumscribed simplex for each partition cell. For the latter, we prove convergence whenever the recourse function over the uncertainty dimension is monotone.

2.6 Proposed conservative partition refining (CPR) method for two-stage stochastic programming

In this section, we present a new conservative solution framework for two-stage stochastic linear optimization problems that solves the upper-bound problem. The aim is to obtain a conservative solution that avoids disappointment, i.e., the objective function cost estimate is the upper limit for the actual expected cost. We propose a tractable reformulation, namely the deterministic equivalent model, for the upper-bound problem and prove convergence of the proposed methodology. We developed a simple and efficient partition refinement algorithm based on the structure of the optimality gap of each iteration.

We start from a sequential procedure to split the uncertainty support to obtain a refined partition. For iteration n and partition $\mathscr{P}_n$, we solve the upper-bound problem (2-15) and obtain the conservative solution $\mathbf{x}_n^U$. Then, we solve the lower-bound problem for the same partition $\mathscr{P}_n$ and compute the optimality GAP – the difference of the optimal values of the upper- and lower-bound problems. The proposed CPR method is outlined in the following pseudo-algorithm:

Algorithm 2 The CPR method

Require: $\mathscr{P}^1 = \{\Xi\}$, $\epsilon > 0$ stopping criteria.
Ensure: $\mathbf{x}_n^U$ the conservative solution of the upper bound problem.
1: Solve (2-15) to determine the optimal solution $\mathbf{x}_n^U$ and the optimal objective value z_n^U of the upper bound problem for the partition $\mathscr{P}_n$.
2: Solve (2-5) to determine the optimal objective value z_n^L of the lower bound problem for the partition $\mathscr{P}_n$.
3: **if** $\frac{z_n^U - z_n^L}{z_n^U} \leq \epsilon$ **then**
4:     stop and return $\mathbf{x}_n^U$ as the optimal solution.
5: **else**
6:     refine the partition $\mathscr{P}_n$ to $\mathscr{P}_{n+1}$ and return to 1
7: **end if**

The convergence of the CPR method is ensured by the **Theorem 1**, whose proof is presented in the Appendix A.1. It states that the sequence of optimal solutions and objective values of the lower- and upper-bound problems converges to the optimal solution $\mathbf{x}^*$ and the optimal objective value z^* of the two-stage stochastic optimization problem (2-1), respectively. Note that the proof of the convergence does not depend on the refinement-partition procedure, however, the choosing of this procedure could enhance the velocity of convergence.

Theorem 1. *:*

1.1 *The optimal objective value sequence* $\{z_n^U\}_{n=1}^{\infty}$ *corresponds to conservative solutions given by the upper bound problem, for a family of partitions* $\{\mathscr{P}_n\}_{n=1}^{\infty}$ *such that* $\mathscr{P}_{n+1}$ *refines* $\mathscr{P}_n$*, is non-increasing.*

1.2 *The optimal objective value sequence* $\{z_n^L\}_{n=1}^{\infty}$ *corresponds to optimistic solutions given by the lower bound problem, for a family of partitions* $\{\mathscr{P}_n\}_{n=1}^{\infty}$ *such that* $\mathscr{P}_{n+1}$ *refines* $\mathscr{P}_n$*, is non-decreasing.*

1.3 *We have that the sequences* $\{z_n^U\}_{n=1}^{\infty}$ *and* $\{z_n^L\}_{n=1}^{\infty}$ *are convergent, i.e.,* $z_n^L \longrightarrow z^* \longleftarrow z_n^U$*, as* $n \longrightarrow \infty$*. Also the sequence* $\{x_n^U\}_{n=1}^{\infty}$ *and* $\{x_n^L\}_{n=1}^{\infty}$ *converge to* $\mathbf{x}^* \in \arg\min_{\mathbf{x}\in X} f(\mathbf{x})$.

It is worth mentioning that the upper bound sequence obtained by computing the objective function of the upper-bound problem at the optimistic solution defines a non-increasing sequence. This fact ensures the convergence of the existing partition-based method. However, the optimistic solution is sub-optimal for the upper-bound problem, so the existing partition-based method derives an upper bound less tight.

Next, we propose a simple and efficient partition refinement algorithm that defines the sequential upper- and lower-bounds and, consequently, generates the sequence of conservative solutions $\{\mathbf{x}_n^U\}_{n=1}^{\infty}$.

2.7 Accelerating methods for the moment-based approach

In this section, we present two accelerating methods for the distributionally robust optimization approach based on the first moment. These methods are proposed regarding the dimensionality. For problems with the medium-size uncertainty dimensionality a column and constraint generation algorithm is proposed. For problems with a high-dimensional uncertainty we propose a simplex-based heuristic method based on the extension of the uncertainty support.

2.7.1 Column and constraint generation algorithm for the moment-based approach

For problems with medium-sized uncertainty dimension, we adapted the column and constraint generation framework to iteratively identify a subset of constraints that ensures feasibility to the original problem. The goal of this algorithm is to obtain an equivalent formulation of (2-15) with significantly fewer constraints and variables.

According to [13], the column and constraint generation procedure is implemented in a master-oracle scheme. In our context, the master problem is a relaxation of an equivalent formulation of the original problem (2-15). Given the solution of the master (relaxed) problem, the oracle finds the vertex that generates the worst infeasibility and adds the associated block of linear constraints and decision variables to the master problem. This iterative procedure stops whenever the oracle asserts that the master solution is feasible for the original problem.

For the initial iteration $\ell = 0$ of the column-and-constraint algorithm, we start with an empty set $J^k = \emptyset$ of constraints, i.e., the initial master problem is a relaxation of (2-16). For initialization purposes, we also add a block of constraints associated with the conditional expectation $\overline{\boldsymbol{\xi}}^k$ adapted to cell Ξ^k. This additional constraint does not affect the original feasible set since it can be represented as an weighted average of all constraints associated to each vertex.

For a given iteration m, with an updated set J^k, we solve (2-16) and obtain a candidate solution $(\pi^*_m, \boldsymbol{\eta}^*_m, \mathbf{x}^*_m)$. Then, we solve the oracle problem

$$\textbf{ORACLE}: \quad \vartheta^* = \max_j \{Q(\mathbf{x}^*_m, \mathbf{e}^k_j) - \pi^*_m - [\boldsymbol{\eta}^*_m]^\top \mathbf{e}^k_j\}, \tag{2-17}$$

to find the highest constraint violation given the current solution. To efficiently solve the oracle problem, we replace the recourse function by its dual representation and combine them in a single maximization problem

$$\begin{aligned} \max_{j,\boldsymbol{\theta}} \quad & \boldsymbol{\theta}^\top (\mathbf{H}(\mathbf{x}^*_m)\mathbf{e}^k_j + \mathbf{r}(\mathbf{x}^*_m)) - \pi^* - [\boldsymbol{\eta}^*_m]^\top \mathbf{e}^k_j \\ \text{s.t.} \quad & \mathbf{W}^\top \boldsymbol{\theta} \leq \mathbf{q}. \end{aligned} \tag{2-18}$$

Following [19], we avoid solving (2-18) by enumeration of vertices $\{\mathbf{e}^k_j : j = 1, \ldots, 2^{d_\xi}\}$ of the cell Ξ^k by introducing binary variables to represent each vertex and transform (2-18) into a MILP equivalent formulation presented in details in Appendix A.2. The algorithm stops whenever the oracle optimal value is non-negative, i.e., the current solution $(\pi^*_{m+1}, \boldsymbol{\eta}^*_{m+1}, \mathbf{x}^*_{m+1})$ is feasible for the original problem. If the oracle optimal value is positive, then update the set J^k including the oracle solution i^* and repeat the process solving again the master problem.

We summarize the proposed column and constraint generation algorithm in the following pseudo-algorithm:.

Algorithm 3 Column and Constraint Generation framework

Require: $\overline{\xi}^k$, $m = 0$, set $J^k := \emptyset$ of the constraints generated by the vertices of the cell Ξ^k.

Ensure: $\mathbf{x}_n^U$ the conservative solution of the upper bound problem.

1: Initialization. Solve the problem (2-16) to derive an optimal solution. $(\pi_0^*, \boldsymbol{\eta}_0^*, \mathbf{x}_0^*)$ with $J^k := \emptyset$.
2: Maximize infeasibility. Find j^* as the optimal solution of (2-17).
3: **if** $\vartheta^* \leq 0$ **then**
4: return $\mathbf{x}_m^* = \mathbf{x}_n^U$ as the optimal solution.
5: **else**
6: Update $m = m + 1$. Do $J^k \cup \{j^*\}$ and go to the next step
7: **end if**
8: Update $(\pi_{m+1}^*, \boldsymbol{\eta}_{m+1}^*, \mathbf{x}_{m+1}^*)$ as the optimal solution of the master problem (2-16) and go to 2.

2.7.2
Simplex-based heuristic method to handle high-dimensional uncertainty

There are some instances of high dimensional uncertainty that make the column and constraint generation computationally intractable. Thus, it is necessary to appeal to an alternative solution method to obtain an upper bound and corresponding conservative solution.

To handle high dimensional uncertainty, we propose a heuristic solution method that extends the original box uncertainty support to a circumscribed simplex polyhedral where the number of vertexes depends linearly on the uncertainty dimension. Under the extended support, we reformulate (2-14) to obtain an upper-bound problem whose complexity grows linearly with the uncertainty dimension.

The proposed extension is the minimum volume simplex that contains the original cell (see Fig. 2.1) and one selected vertex $\hat{\xi}^k$ coinciding with the cell's vertexes. This simplex has the property that the length of the edges that contain the original vertex $\hat{\xi}$ is equal to the sum of the length of the projection of the original support along to each dimension, respectively.

For example in Fig. 2.1, it is represented the extension of the support $[a_1, b_1] \times [a_2, b_2]$ by the simplex represented in green. We considered two possibilities $\{(a_1, a_2), (b_1, b_2)\}$ for the original vertex $\hat{\xi}^k$, indicated by the black point, that corresponds to the extreme events of the realization of the

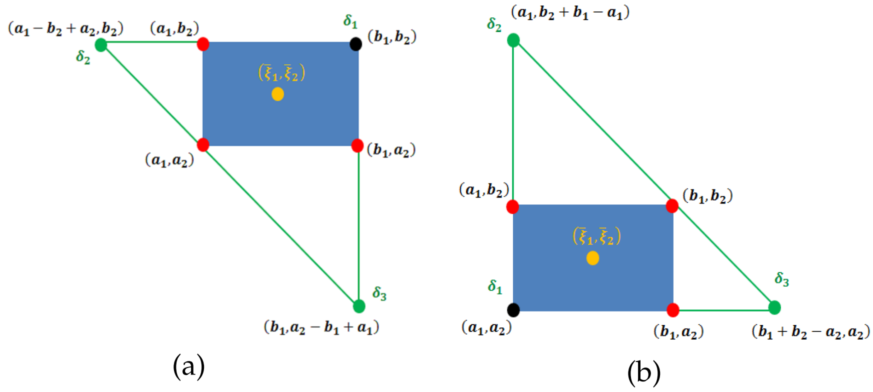

Figure 2.1: The blue region in the figure corresponds to the original support $[a_1, b_1] \times [a_2, b_2]$, and the simplex, represented in green, is its extension. The vertex in black denotes the selected $\hat{\xi}$ of the realization of the uncertainty for the recourse and the point highlighted at the middle indicates the mean. (a) $\hat{\xi} = (b_1, b_2)$; (b) $\hat{\xi} = (a_1, a_2)$

uncertainty. In both cases, the length of the edges of the simplex that contain the vertex $\hat{\xi} \in \{(a_1, a_2), (b_1, b_2)\}$ is equal to $(b_1 - a_1) + (b_2 - a_2)$, the sum of dimensions. We use this property as a simple rule to create the simplex.

The proposed simplex-based heuristic method solves the sequential partition refinement problem by considering the extension of each cell of the partition. Based on the extension of cell Ξ^k, problem (2-14) can be rewritten as follows:

$$
\begin{aligned}
\max_{\delta \in \mathcal{R}(\bar{\xi}^k)} \mathbb{E}^{\delta}\left[Q(\mathbf{x}, \mathbf{v}^k)\right] := \min_{\pi, \boldsymbol{\eta}} \quad & \pi + \boldsymbol{\eta}^\top \bar{\xi}^k \\
\text{s.t.} \quad & \pi + \boldsymbol{\eta}^\top \mathbf{v}_j^k \geq Q(\mathbf{x}, \mathbf{v}_j^k), \quad j = 1, \ldots, (d_\xi + 1),
\end{aligned}
\tag{2-19}
$$

where $\mathcal{R}(\bar{\xi}^k) = \left\{ \delta_j^k \in \mathbb{R}^{d_\xi + 1} : \sum_{j=1}^{d_\xi+1} \delta_j^k \mathbf{v}_j^k = \bar{\xi}^k, \sum_{j=1}^{d_\xi+1} \delta_j^k = 1 \right\}$, $\mathbf{v}^k$ is a random vector with support in the set $\{\mathbf{v}_j^k : j = 1, \ldots, d_\xi + 1\}$ of vertexes of the simplex that contains the cell Ξ^k and $\mathbf{v}_1^k := \hat{\xi}^k$ is the original vertex of the hypercube Ξ^k. Note that the vertex $\mathbf{v}_j^k$ for $j = 2, \ldots, (d_\xi + 1)$ different from $\hat{\xi}^k$ in just one component according to the construction of the simplex that contains the hypercube Ξ^k. That said, we propose the simplex-based heuristic problem

$$
\begin{aligned}
\tilde{z}_n^U := \min_{\pi^k, \boldsymbol{\eta}^k, \mathbf{x}} \quad & \mathbf{c}^\top \mathbf{x} + \sum_{k=1}^{n} p^k (\pi^k + [\boldsymbol{\eta}^k]^\top \bar{\xi}^k) \\
\text{s.t.} \quad & \pi^k + [\boldsymbol{\eta}^k]^\top \mathbf{v}_j^k \geq Q(\mathbf{x}, \mathbf{v}_j^k), \quad j = 1, \ldots, (d_\xi + 1), k = 1, \ldots, n \\
& \mathbf{x} \in X.
\end{aligned}
\tag{2-20}
$$

Note that the number of blocks of linear constraints of the problem

(2-20) is $n \cdot (d_{\xi} + 1)$, i.e., it depends linearly on the uncertainty dimension. With this alternative upper-bound problem we can solve the sequential partition refinement problem described in Algorithm 2 solving (2-20) instead (2-15).

In particular, we proof convergence of the partition refinement problem using the simplex-based heuristic method if we will assume that the recourse function $Q(\mathbf{x}, \cdot)$ is monotone over the uncertainty vector $\boldsymbol{\xi} \in \Xi$, and the selected original vertex $\hat{\boldsymbol{\xi}}^k$ of the hypercube Ξ^k is the worst-case of the recourse, i.e., $\hat{\boldsymbol{\xi}}^k \in \arg\max\limits_{\boldsymbol{\xi} \in \Xi^k} Q(\mathbf{x}, \boldsymbol{\xi})$ for any $\mathbf{x} \in X$.

Proposition 1. *Let $Q(\mathbf{x}, \xi_i)$ be a monotonic function of ξ_i, for all $i = 1, \ldots, d_{\xi}$. Assuming that, for any $\mathbf{x}$, $\hat{\boldsymbol{\xi}}^k \in \arg\max_{\boldsymbol{\xi} \in \Xi^k} Q(\mathbf{x}, \boldsymbol{\xi})$ generates the worst-case recourse cost given the cell Ξ^k. Then, the sequences $\{\tilde{z}_n^U\}_{n=1}^{\infty}$ and $\{\tilde{\mathbf{x}}_n^U\}_{n=1}^{\infty}$ converge to the optimal objective value and optimal solution of* (2-1)*, respectively, where $\tilde{\mathbf{x}}_n^U$ is the optimal solution of* (2-20).

The proof of **Proposition1** is presented in the Appendix A.3.

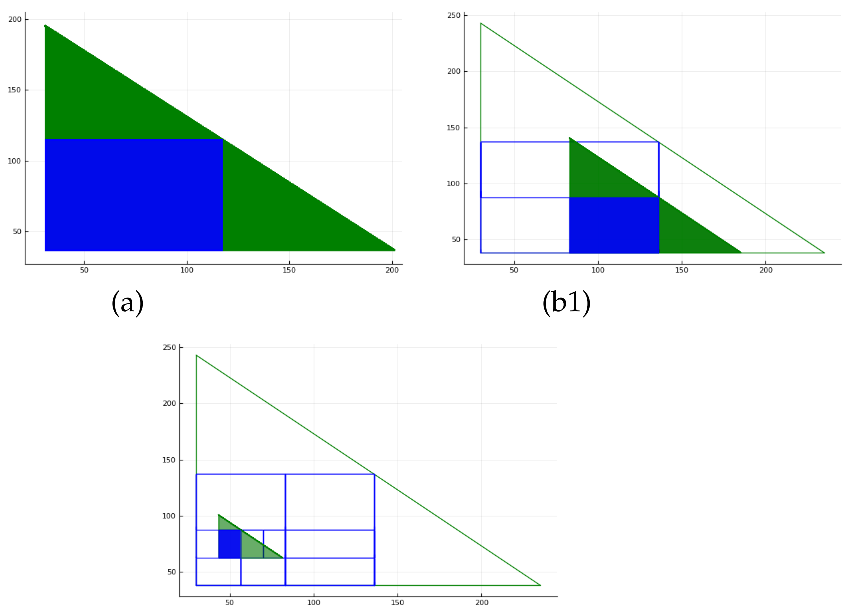
(a) (b1) (b2)

Figure 2.2: (a) Extension of original support $[a_1, b_1] \times [a_2, b_2]$ for newsvendor problem for two items; (b1) and (b2) The extension of the chosen cell (highlighted in blue) in the refinement procedure is the same as that of original support for different partition sizes.

Note that a reformulation similar to (2-16) can be obtained by just considering J^k as the number of vertexes of the circumscribed simplex. For a polyhedral set X, this reformulation is a linear programming problem whose complexity grows linearly with the uncertainty vector dimension.

2.8 Solution algorithm with worst-case partition refinement (SAWPR)

Any partition-based method is supported on the refinement procedure of the partition to improve the upper and lower approximations. Next, we detail the three basic steps of the proposed partition refining procedure as (i) selection of the cell Ξ^{k*} to be split, (ii) selection of the uncertainty dimension $i^* \in \{1, \ldots, d_\xi\}$ to refine the partition, and (iii) selection of the cutting point.

In the first step of the partition refinement procedure, we aim to select the cell with the highest contribution to the current optimality gap, which is composed of the difference of upper and lower approximations of the expected recourse function. Given that the partition influences the gap through the expected recourse function, we select the cell with the maximum contribution to the difference of upper and lower approximations of the expected recourse function. Let us define the selected cell as

$$k^* \in \arg\max_k \left\{ p^k \left(\max_{\xi \in \mathcal{D}(\bar{\xi}^k)} \mathbb{E}^\delta \left[Q(\mathbf{x}_n^U, \mathbf{e}^k) \right] - Q(\mathbf{x}_n^L, \bar{\xi}^k) \right) \right\}. \tag{2-21}$$

The second step of the partition refinement procedure is the selection of the uncertainty direction. Most existing partition-based methods consider the direction with the highest metric of non-linearity of $Q(\mathbf{x}, \cdot)$ as the optimal direction to split the optimal cell to refine the partition. This is motivated by the fact the lower and upper approximations coincide for an affine function. Initially, [18] proposed to use dual (subgradient) information at the endpoints of the cell, while [4] compares the difference between the upper and lower approximations. However, the use of dual information increases the computational burden of the sequential partition refinement problem.

In this work, we selected the uncertainty direction by solving an optimization problem that resembles the robust optimization model with an uncertainty budget proposed by [20]. For the selected cell k^*, we formulated an adversary problem in (2-22) that aimed to find the uncertainty realization with the highest cost for the optimistic solution given by the lower-bound problem. The intent was to select the dimension where the conditional mean was less representative of the entire conditional distribution adapted to that cell. We imposed a unitary budget constraint, i.e., only one component of the random vector is allowed to change around its nominal value (conditional mean). Hence, we selected the uncertainty dimension i^* such that $\hat{\xi}_{i^*} \neq \bar{\xi}^k_{i^*}$, where

$$\begin{aligned} \hat{\boldsymbol{\xi}} \in \underset{\boldsymbol{\xi}, \mathbf{z}}{\arg\max} \quad & Q(\mathbf{x}^L_n, \boldsymbol{\xi}) \\ \text{s.t.} \quad & \bar{\xi}^k_i - z_i(\bar{\xi}^k_i - a^k_i) \leq \xi_i \quad \forall i, \\ & \xi_i \leq \bar{\xi}^k_i + z_i(b^k_i - \bar{\xi}^k_i) \quad \forall i, \\ & \sum_{i=1}^{d_\xi} z_i \leq 1, \\ & 0 \leq z_i \leq 1 \quad \forall i. \end{aligned} \tag{2-22}$$

Since (2-22) is a maximization problem with a convex objective function and the feasible set is a polyhedron, there is a vertex that is optimal. By construction, the number of vertices of the feasible polyhedral is $2\,d_\xi$, i.e., grows linearly on the uncertainty dimension. Thus, to efficiently solve (2-22) it suffices to enumerate the vertices and cast the one with the highest recourse cost.

Note also that the vertices are intuitive and easily identifiable since they are defined as the conditional mean $\bar{\boldsymbol{\xi}}^k$ projected on the faces of the hypercube Ξ^k. To illustrate this concept, let us consider $\Xi = \bigtimes_{i=1}^{3}[a^k_i, b^k_i]$ (see **Fig.** 2.3) and a given optimistic solution $\mathbf{x}^L_n$. A unitary uncertainty budget leads to a "diamond" shaped polyhedral inscribed in the hypercube Ξ^k. In other words, the vertices of the feasible set differ from the center (conditional mean $\bar{\boldsymbol{\xi}}^k$) in just one component i, which can assume any extreme value $\{a^k_i, b^k_i\}$. Therefore, the number of vertexes is $2\,d_\xi$.

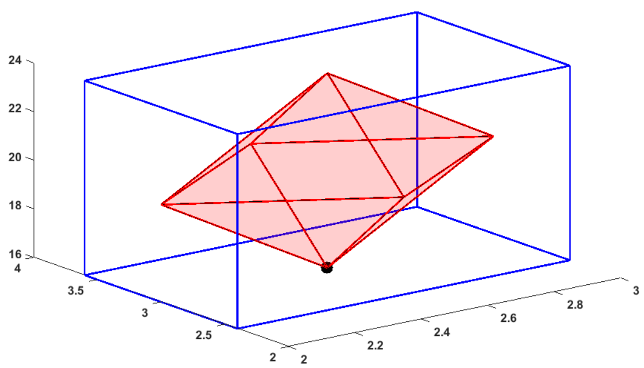

Figure 2.3: Ambiguity set representing the feasible region of the problem (2-22) for the case $d_\xi = 3$

The third and last step of the partition refinement procedure is the selection of the cutting point. For simplicity and computational efficiency, we assumed that the cutting point is the component of the conditional mean $\bar{\xi}_{i^*}$ along to the uncertainty dimension i^* the same as [18, 21].

Finally, we summarize the proposed partition refinement procedure, namely, solution algorithm with worst-case partition refinement (SAWPR) as follows:

Algorithm 4 SAWPR

Require: $\mathbf{x}^*$ optimal solution of (2-13)

Ensure: Ξ^{k^*} and Ξ^{n+1}

1: Determine $k*$ the optimal cell to be split as the cell with the largest error contribution given by (2-21).

2: Find i^* as the component $\hat{\xi}_{i^*}$ of the optimal solution $\hat{\xi}$ of (2-22) such that $\hat{\xi}_{i^*} \neq \bar{\xi}^k_{i^*}$.

3: $\Xi^{k^*} \Longleftarrow [a^{k^*}_{i^*}, \bar{\xi}^{k^*}_{i^*}] \times \bigtimes_{\{1,2,\ldots,d_\xi\} \setminus \{i^*\}} [a^{k^*}_t, b^{k^*}_t]$

4: $\Xi^{n+1} \Longleftarrow [\bar{\xi}^{k^*}_{i^*}, b^{k^*}_{i^*}] \times \bigtimes_{\{1,2,\ldots,d_\xi\} \setminus \{i^*\}} [a^{k^*}_t, b^{k^*}_t]$

In Fig. 2.4 we present the refinement procedure for the farmer problem instance for three types of crops as an illustrative example of the SAWPR algorithm. Note that the cells of the partition are clustered around the region of the uncertainty support corresponding to scenarios of less land productivity which represents in particular, the worst-case for this instance.

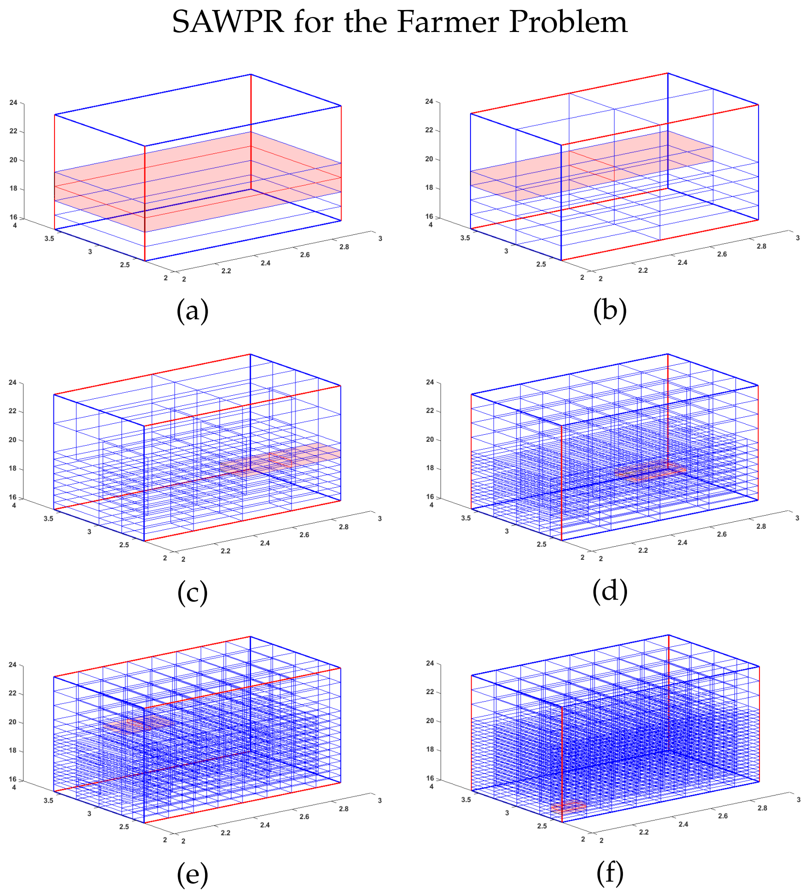

Figure 2.4: Partition of the support $\Xi = \times_{i=1}^{3}[a_i^k, b_i^k]$ with: (a) 4 cells, (b) 20 cells, (c) 100 cells, (d) 300 cells, (e) 500 cells, and (f) 900 cells. The shaded region shows the chosen cell and the red edges show the direction chosen to refine the partition.

2.9
Empirical study for the moment-based approach

We present an empirical study of the CPR method using of the two acceleration approaches (column and constraint generation method and the propose simplex-based heuristic method of the extension of the uncertainty support), applied to an aircraft allocation problem [22] and several instances of the farmer problem [18] for different quantities of cultivated crops, which are cost-minimizing problems belonging to the problem class for which the method was designed.

For the first application, different types of aircraft must be allocated to a certain route for the transportation of passengers. The number of allocated aircraft is the first-stage decision, and the recourse of the problem is defined by the number of bumped passengers when demand for seats outstrips capacity. The right-hand side uncertainty corresponds to the unknown demand of passengers modeled by a uniform probability distribution. For this problem, the dimension of the uncertainty is 5, the number of first-stage variables was 21, and the number of second stage-variables is 10 for each uncertainty realization.

Regarding the second application, the farmer problem is an example of a production model under uncertainty where the first-stage decisions correspond to the land allocation destined to rise different types of crops, and the recourse consists in trading the cultivated products in the local market to satisfy a given supply. In this problem, we assume a uniform probability distribution to model the uncertainty in the land productivity for growing each crop. The number of first-stage variables is equal to the uncertainty dimension, since they correspond to the land allocation to cultivate each type of crop while the number of second-stage variables for each uncertainty realization is twice the quantity of cultivated products because they correspond to the sold and buy quantities of each crop in the local market.

To generate different instances of the farmer problem, we varied the uncertainty dimension considering eight and 20 types of crops to create a computational experiment for each of the following situations:

– It is impossible to solve in reasonable time the deterministic equivalent linear model for the enumerative case considering the 2^{d_ξ} vertexes of the hypercubes of the rectangular partition, but it is possible to handle the medium-sized uncertainty dimensionality with the proposed column and constraint generation algorithm.
– It is impossible to obtain a conservative solution solving the reformulated upper-bound problem using the column and constraint generation algorithm, and it is necessary to appeal to a heuristic solution method.

For both applications, the refinement algorithm and the algorithm for the sequential partition-based method is implemented in JuMP [23], a modeling language for mathematical optimization embedded in the Julia programming language and Gurobi was used as the linear optimization solver to run the computational experiments on a Intel Core i7, 4.0-GHz processor with 32 GB of RAM.

2.9.1 Computational results for the moment-based approach

In general, the optimal solutions of the upper-bound problem of both the CPR method solved by the column and constraint generation algorithm and the simplex-based heuristic method of the extension of the uncertainty support represents a conservative decision policy. Indeed, these solutions are obtained from the approximation of the conditional expectation $\mathbb{E}[Q(\mathbf{x},\boldsymbol{\xi})|\boldsymbol{\xi} \in \Xi^k]$ by the worst distributionally expectation

$$\max_{\delta \in \mathcal{D}(\overline{\boldsymbol{\xi}}^k)} \mathbb{E}^{\delta}\left[Q(\mathbf{x},\mathbf{e}^k)\right]$$

and

$$\max_{\delta \in \mathcal{R}(\overline{\boldsymbol{\xi}}^k)} \mathbb{E}^{\delta}\left[Q(\mathbf{x},\mathbf{v}^k)\right],$$

considering the marginal distribution of the random vectors $\mathbf{e}^k$ and $\mathbf{v}^k$ with support in the set $\{\mathbf{e}_j^k : j = 1,\ldots,2^{d_\xi}\}$ and $\{\mathbf{v}_j^k : j = 1,\ldots,d_\xi + 1\}$, respectively, for all $k = 1,\ldots,n$.

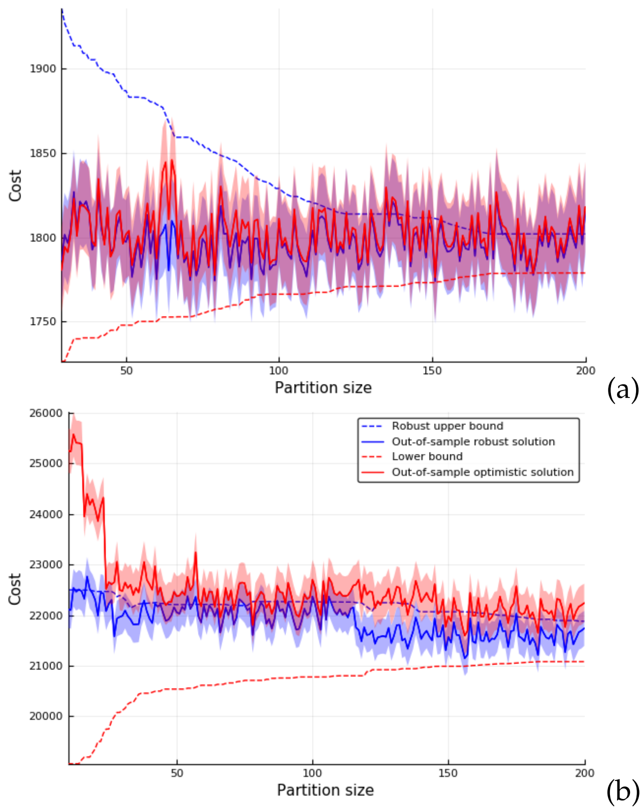

Figure 2.5: Disappointment in out-of-sample analysis for aircraft allocation problem: (a) low recourse cost; (b) high recourse cost.

As mentioned above, the existing partition-based method do not solve the upper-bound problem, instead, it determines the optimistic solution $\mathbf{x}_n^L$ given by the lower-bound problem. Nevertheless, depending on the recourse cost, sometimes it is necessary to obtain a conservative solution because the optimistic solution generates a significant disappointment. To study the disappointment of a given conservative and optimistic solution in an out-of-sample analysis, we performed a Monte Carlo simulation with a number N of scenarios to estimate the actual cost by the sample-average cost [24].

For the aircraft allocation problem, we consider two instances: (a) low recourse cost; (b) high recourse cost. For the low recourse cost, we consider a negligible cost for unattended demand. For the high recourse cost, we consider a significantly high deficit cost associated with unattended demand. Given these two instances, we depict in Fig. 2.5 the disappointment by confronting the out-of-sample cost against the cost estimate given by objective value of the upper (conservative) and lower (optimistic) bound problems. For didactic purposes, the results associated with the optimistic solution are presented in red while the results for the conservative solution are presented in blue. For each partition size, the cost estimate given by optimal values of the upper and lower objective values are represented by the dashed lines. The solid line corresponds to the out-of-sample cost evaluation and the shaded area corresponds to the associated 95% confidence interval.

For the a low recourse cost instance, Fig. 2.5 (a), we observe that both (conservative and optimistic) solutions have similar out-of-sample performances. However, for the high recourse cost instance, Fig. 2.5 (b), the conservative solution out-performs the optimistic one for any partition size. Note also that, for both instances, the optimistic solution leads to significant disappointment–the out-of-sample cost evaluation is significantly higher than the cost estimate given by the optimal value of the lower bound problem. On the other hand, the conservative solution has a mathematical guarantee against disappointment, which is corroborated by this empirical study–there is no statistically significant disappointment. This means that optimistic solution methods might not be suitable for applications with high recourse cost due to poor out-of-sample performance and potentially high disappointment levels.

In Fig. 2.6, we present the numerical results for the farmer problem with eight crops considering an high recourse cost instance. The out-of-sample evaluation of the conservative solution shows a very similar behavior when compared to the optimal objective value of the upper-bound problem. Conversely, the out-of-sample cost evaluation for the optimistic solution is significantly higher than the optimal objective value, i.e., it shows a high disappointment level. Moreover, the conservative solution significantly out-performs the optimistic one in an out-of-sample analysis. It is important to note that the above mentioned effects are amplified whenever the algorithm

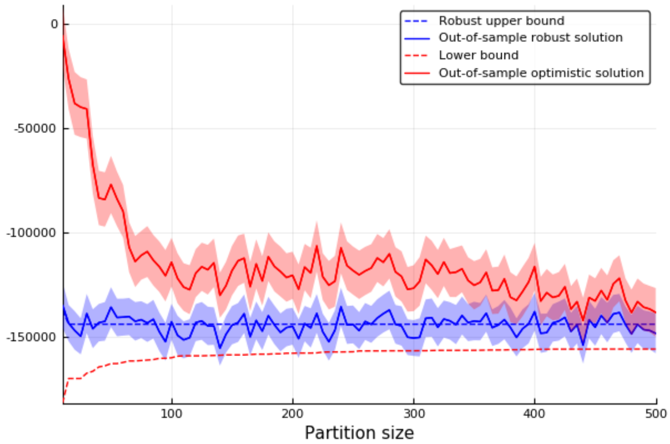

Figure 2.6: Disappointment in out-of-sample analysis of the conservative solution and optimistic solution of the farmer problem with eight crops.

is far from converging.

This result alerts us to use of the optimistic solution for two-stage stochastic linear programming problems when the recourse cost is significantly high. If the method stops before convergence–due to time or computational-power limitations–the optimistic solution is not reliable given its poor out-of-sample performance and high disappointment level. On the other hand, the conservative decision is robust even for non-converged solutions.

Finally, Fig. 2.7 presents the numerical results for the farmer problem with 20 crops obtained by the simplex-based heuristic method of the extension of the uncertainty support, considering an high recourse cost. As in the upper-bound problem (2-15), solved by the column and constraint generation algorithm, the solution given by the simplex-based heuristic method (2-20) also avoids disappointments and significantly out-perform the optimistic solution. As before, the non-converged optimistic solution presented a significant disappointment and poor out-of-sample performance.

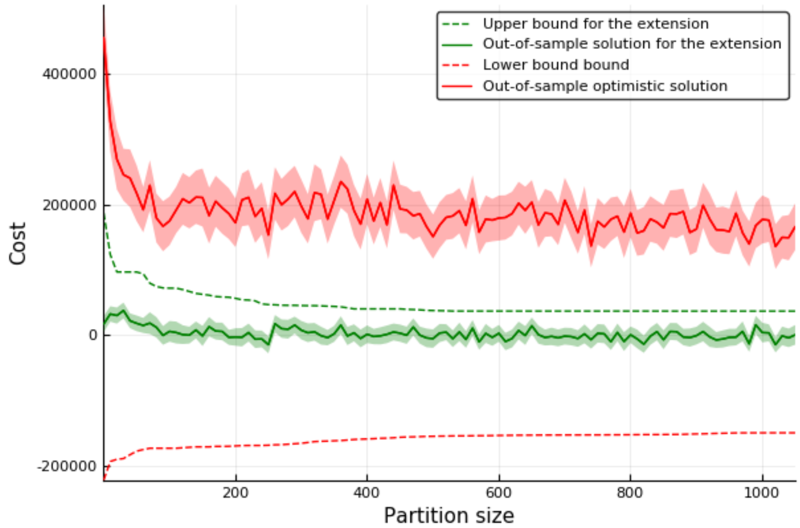

Figure 2.7: Upper bound and lower bound obtained by the simplex-based heuristic method of the extension of the uncertainty support for the farmer problem with 20 crops.

3 Distributionally robust optimization approach based on the Wasserstein metric

Distributionally Robust Optimization (DRO) is a mathematical framework to incorporate ambiguity in the characterization of the true data-generating distribution. In general, this approach considers an ambiguity set $\mathcal{P}$ to evaluate the worst-case for the expectation $\mathbb{E}^{\widetilde{P}}[Q(\mathbf{x}, \boldsymbol{\xi})]$ over all probability distributions[1] $\widetilde{P} \in \mathcal{P}$. Then, for a DRO minimization problem, the optimal decision $\mathbf{x} \in X$ is the one that minimizes the highest expected cost $\mathbb{E}^{\widetilde{P}}[Q(\mathbf{x}, \boldsymbol{\xi})]$ over all $\widetilde{P} \in \mathcal{P}$.

Research on DRO models—formulations, algorithms and applications—has grown enormously in the past few years; a recent survey can be found in [25]. One particular setting that has received considerable attention in the literature is when the ambiguity set $\mathcal{P}$ is defined as the set of distributions that are not "too far" from some reference distribution. Of course, such a notion requires defining an appropriate way to measure the distance between distributions. While there are multiple ways to measure such distance, the *Wasserstein* distance has been popular due do its theoretical properties and practical performance. The thrust of the two-stage *data-driven* DRO with Wasserstein metric (DD-DRO-W) framework is to center the ambiguity set on the empirical distribution corresponding to the data and to set a radius around it to enclose the true data-generating distribution with a high confidence level [14]. The data-driven Wasserstein distance has been widely used in DRO partly because the associate ambiguity set collects both discrete and continuous probability distributions, even though the Wasserstein ball is centered around the empirical discrete distribution.

Despite their practical appeal, DD-DRO-W problems are hard to solve, but specific characteristics can be explored for computational tractability. For instance, by considering a compact support set for the uncertainty, [26] presents a finite-dimensional non-convex reformulation of the worst-case expectation problem, which has high computational burden and no global optimality guarantee. For a continuous data-generating distribution

[1] We use the notation $\mathbb{E}^{\widetilde{P}}[\cdot]$ to emphasize that the expectation is taken over the probability distribution $\widetilde{P}$.

supported on a polyhedron, [27] presents a semi-infinite linear reformulation of a two-stage distributionally robust unit commitment problem for the integration of renewable energy over the data-driven Wasserstein ball. An exact decomposition scheme is proposed by [28], but considering an ambiguity set with only discrete probability distributions. In [29], the authors present a general conic programming reformulation for a two-stage distributionally robust optimization problem with Wasserstein-based ambiguity set. Particularly, for linear programming problems with right-hand-sided uncertainty, the authors present a tractable reformulation for the case of unbounded uncertainty supports, an unsuited assumption for a large range of applications. By assuming a compact support set for the uncertainty, [14] present tractable reformulations for a number of cases, except for linear programming problems with right-hand-sided uncertainty. For this case, the resulting reformulation requires pre-computing all dual vertices of the recourse problem, but such number grows exponentially with the size of the problem.

In this context, we study decomposition methods applied to two-stage DD-DRO-W with right-hand-sided uncertainty and *rectangular* support. As an alternative to the proposed dual vertex enumeration in [14], in Section 3.4, we propose a novel finite reformulation that explores the rectangular uncertainty support. We develop an exact decomposition (oracle-master) scheme based on the Column-and-Constraint Generation (C&CG) method. Moreover, by considering variations of the proposed scheme, we derive two other alternative decomposition methods, namely, Multi-cut Benders and Single-cut Benders. Whereas Benders' methods consider a local linear approximation of the recourse function for a given scenario, the C&CG method tends to converge faster since co-optimizes first-stage and the recourse for the uncertainty realizations selected by the oracle problem.

It is important to mention that just prior to the submission of this paper we came across a paper by [30]—which was developed independently of our work and very recently made available online—which solves a similar but different class of problems. Unlike our framework that considers all problems with uncertainty (with rectangular support) on the right-hand side, the authors do not allow randomness in the technology matrix, i.e., the matrix multiplying the first stage variable. On the other hand, [30] consider problems in which the uncertainty has either bounded or unbounded support. They provide a master-oracle scheme that resembles our Benders Multi-cut method with a similar master but with a different MILP oracle reformulation, where the uncertain parameter does not appear as a coefficient of the first-stage variable. Also, [30] do not propose an algorithm similar to C&CG which turns out to be the most efficient in our numerical experiments.

We compare the proposed C&CG method with two variants of the Benders algorithm applied to the newly proposed finite reformulation. Furthermore, we also compare the proposed methods with the existing formulation provided in [14] using the vertex enumeration algorithm available in [31,32] and the decomposition method presented in [28]. To do that, we present results for the unit commitment problem with 5, 14, and 54 thermal generators over a 24-hour uncertainty dimension.

3.1 Theoretical background for the Wasserstein-metric-based approach

In this section, we provide a brief review of elementary notions of convex analysis and present the definition of the Wasserstein metric.

3.1.1 Dual norm

Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$. The associated dual norm denoted by $\|\cdot\|_*$, is defined as

$$\|\mathbf{x}\|_* := \sup_{\mathbf{y}}\{\mathbf{y}^\top \mathbf{x} \,:\, \|\mathbf{y}\| \leq 1\}.$$

By the Hölder's inequality it holds that the dual of the ℓ_p-norm is the ℓ_q-norm, where q satisfies $\frac{1}{p} + \frac{1}{q} = 1$. Particularly, the dual of the ℓ_1-norm (i.e., $\|\mathbf{x}\| = \sum_{i=1}^{d} |x_i|$) is the ℓ_∞-norm (i.e., $\|\mathbf{x}\| = \max_{i \leq d} |x_i|$). Another basic property that will be useful later is that identity $\|\mathbf{x}\| = \|\mathbf{x}\|_{**}$ holds for all $\mathbf{x} \in \mathbb{R}^d$.

3.1.2 Wasserstein metric

The Wasserstein metric is a probability distance that measures the separation between probability distributions. This metric is defined on the space $\mathcal{M}(\Xi)$ of probability distributions P supported on Ξ with $\mathbb{E}^P[\|\xi\|] < \infty$.

Definition 2. *(Wasserstein metric [33]) The Wasserstein metric* $d_w : \mathcal{M}(\Xi) \times \mathcal{M}(\Xi) \longrightarrow \mathbb{R}_+$ *is defined via*

$$d_w(P,P') := \inf_{\Pi} \left\{ \int_{\Xi^2} \|\xi - \xi'\| \Pi(d\xi, d\xi') : \begin{array}{l} \Pi \textit{ is a join distribution of } \xi \textit{ and } \xi' \\ \textit{with marginals } P \textit{ and } P', \textit{ respectively} \end{array} \right\} \tag{3-1}$$

for all distributions P, $P' \in \mathcal{M}(\Xi)$, *where* $\|\cdot\|$ *represents an arbitrary norm on* $\mathbb{R}^{d_\xi}$.

According to [1], the decision variable Π in (3-1) can be viewed as a transportation plan for moving a probability mass described by distribution P to another one described by P' (see Fig. 3.1).

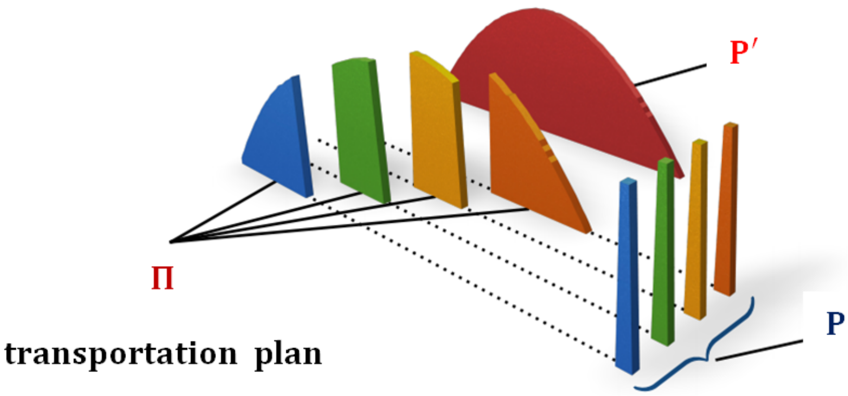

Figure 3.1: Wasserstein distance as a transportation problem, adapted from [1].

In the case of two discrete probability distributions $P = \{p_i\}_{i=1}^m$ and $P' = \{p'_j\}_{j=1}^n$, the resulting problem (3-1) is a linear program given by

$$\begin{aligned} \min_{\pi_{ij}} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n} \|\xi_i - \xi'_j\| \pi_{ij} \\ \text{s.t.} \quad & \sum_{i=1}^{m} \pi_{ij} = p'_j, \quad \forall j = 1,\dots,n, \\ & \sum_{j=1}^{n} \pi_{ij} = p_i, \quad \forall i = 1,\dots,m, \end{aligned}$$

which is an easy problem since linear programming is a dominated technology. In the case of a continuous probability distribution P and a discrete probability distribution $P' = \{p'_j\}_{j=1}^n$, the resulting problem (3-1) is the following infinite optimization program:

$$
\begin{aligned}
\inf_{\Pi \in \mathcal{M}(\Xi^2)} \quad & \sum_{j=1}^{n} \int_{\Xi} \|\boldsymbol{\xi} - \boldsymbol{\xi}'_j\| \Pi(d\boldsymbol{\xi}, \boldsymbol{\xi}_j) \\
\text{s.t.} \quad & \int_{\Xi} \Pi(d\boldsymbol{\xi}, \boldsymbol{\xi}'_j) = p_j, \quad \forall j = 1, \ldots, n, \\
& \sum_{j=1}^{n} \Pi(d\boldsymbol{\xi}, \boldsymbol{\xi}_j) = P(d\boldsymbol{\xi}), \quad , \forall \boldsymbol{\xi} \in \Xi,
\end{aligned}
$$

which is a hard complex problem. In general, computing the Wasserstein distance when both probability distributions are not finite is $\#P$-hard (see [34]). As we will see in the next section, the DRO problem over the Wasserstein ball has a convex reduction that reduces to a linear program by considering the ℓ_1-norm or ℓ_∞-norm.

3.2 Distributionally robust optimization over Wasserstein ambiguity sets

Distributionally Robust Optimization (DRO) is a mathematical framework to incorporate ambiguity in the characterization of the true data-generating distribution. In general, this approach considers an ambiguity set $\mathcal{P}$ to evaluate the worst-case for the expectation $\mathbb{E}^{\widetilde{P}}[Q(\mathbf{x}, \boldsymbol{\xi})]$ over all probability distributions[2] $\widetilde{P} \in \mathcal{P}$. Then, for a DRO minimization problem, the optimal decision $\mathbf{x} \in X$ is the one that minimizes the highest expected cost $\mathbb{E}^{\widetilde{P}}[Q(\mathbf{x}, \boldsymbol{\xi})]$ over all $\widetilde{P} \in \mathcal{P}$.

Research on DRO models—formulations, algorithms and applications—has grown enormously in the past few years; a recent survey can be found in [25]. One particular setting that has received considerable attention in the literature is when the ambiguity set $\mathcal{P}$ is defined as the set of distributions that are not "too far" from some reference distribution. Of course, such a notion requires defining an appropriate way to measure the distance between distributions. While there are multiple ways to measure such distance, the *Wasserstein* distance has been popular due do its theoretical properties and practical performance. The thrust of the two-stage *data-driven* DRO with Wasserstein metric (DD-DRO-W) framework is to center the ambiguity set on the empirical distribution corresponding to the data and to set a radius around it to enclose the true data-generating distribution with a high confidence level [14]. The data-driven Wasserstein distance has been widely used in DRO partly because the associate ambiguity set collects both discrete and continuous probability distributions, even though the Wasserstein ball is centered around the empirical discrete distribution.

[2]We use the notation $\mathbb{E}^{\widetilde{P}}[\cdot]$ to emphasize that the expectation is taken over the probability distribution $\widetilde{P}$.

3.3
Problem statement

In this section, we introduce the data-driven Wasserstein-based ambiguity set (Wasserstein ball). Next, we present the data-driven two-stage distributionally robust optimization problem with right-hand-side uncertainty and a Wasserstein-based ambiguity set.

3.3.1
Wasserstein-based ambiguity set

We assume that we can access a finite data-set of training samples $\{\widehat{\boldsymbol{\xi}}_n\}_{n\leq N}$ that are generated by the true (but unknown) probability distribution. We consider the reference distribution

$$\widehat{P}_N = \frac{1}{N}\sum_{n=1}^{N} \delta_{\widehat{\boldsymbol{\xi}}_n} \tag{3-2}$$

where $\frac{1}{N}$ is the probability mass of the point $\widehat{\boldsymbol{\xi}}_n$ and $\delta_{\widehat{\boldsymbol{\xi}}_n}$ denotes the Dirac's delta function that concentrates unit mass at $\widehat{\boldsymbol{\xi}}_n$.

For a given set $\Xi \subseteq \mathbb{R}^{d_\xi}$, let $\mathcal{M}(\Xi)$ be the space of probability distributions $\widetilde{P}$ supported on Ξ such that $\mathbb{E}^{\widetilde{P}}[||\boldsymbol{\xi}||] < \infty$. For a given $\delta > 0$, the data-driven Wasserstein ball is defined by

$$\mathbb{B}_\delta(\widehat{P}_N) = \{\widetilde{P} \in \mathcal{M}(\Xi) \,:\, d_w(\widetilde{P}, \widehat{P}_N) \leq \delta\} \tag{3-3}$$

where $d_w(\widetilde{P}, \widehat{P}_N)$ denotes the Wasserstein distance between the probability distribution $\widetilde{P} \in \mathcal{M}(\Xi)$ and the reference distribution $\widehat{P}_N$

Even though computing the Wasserstein distance is generally $\#P$-hard, this metric is often used in Distributionally Robust Optimization (DRO) because of its nice properties—in particular, the ambiguity set defined by the Wasserstein distance (Wasserstein ball) contains all probability distributions (continuous and discrete) whose Wasserstein distance to the reference distribution $\widehat{P}_N$ is less or equal to δ.

By using the Wasserstein ball (3-3), the two-stage distributionally robust optimization problem with right-hand-side uncertainty can be defined as follows:

$$\min_{\mathbf{x}\in X} \mathbf{c}^\top \mathbf{x} + \sup_{\widetilde{P}\in\mathbb{B}_\delta(\widehat{P}_N)} \mathbb{E}^{\widetilde{P}}[Q(\mathbf{x},\boldsymbol{\xi})], \tag{3-4}$$

We call henceforth the inner problem $\sup_{\widetilde{P}\in\mathbb{B}_\delta(\widehat{P}_N)} \mathbb{E}^{\widetilde{P}}[Q(\mathbf{x},\boldsymbol{\xi})]$ as the worst-case expectation problem.

As shown in [14] and [35], if (3-4) is such that $\sup_{\boldsymbol{\xi}\in\Xi} |Q(\mathbf{x},\boldsymbol{\xi})| < \infty$ for all $\mathbf{x} \in X$, i.e., has relatively complete recourse with bounded recourse value, then it is equivalent to the semi-infinite optimization problem

$$\min_{\mathbf{x},\lambda,s_n} \mathbf{c}^\top \mathbf{x} + \lambda\delta + \frac{1}{N}\sum_{n=1}^{N} s_n \tag{3-5a}$$

$$\text{s.t.} \quad Q(\mathbf{x},\boldsymbol{\xi}) - \lambda\|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}_n\| \leq s_n, \quad \forall n \leq N, \forall \boldsymbol{\xi} \in \Xi \tag{3-5b}$$

$$\lambda \geq 0, \tag{3-5c}$$

$$\mathbf{x} \in \mathbf{X}. \tag{3-5d}$$

In the next section, by considering the primal ℓ_1-norm in (3-5b), we propose a (finite) linear programming reformulation for the semi-infinite problem (3-5) exploring the particular case of rectangular uncertainty support. We argue that this development is also valid for the ℓ_∞-norm.

3.4 A new finite reformulation for a rectangular uncertainty support

In this section, we derive a finite reformulation for the semi-infinite problem (3-5) by considering the ℓ_1-norm (or equivalently the ℓ_∞-norm). We start by stating formally our assumptions:

Assumption 1. *We assume that: (i) the second stage-problem belongs to the class of parametric linear programs with right-hand-side uncertainty as in* (2-2)*; (ii) the set* $\{\mathbf{y} \in \mathbb{R}_+^{d_y} : \mathbf{W}\mathbf{y} = \mathbf{H}(\mathbf{x})\boldsymbol{\xi} + \mathbf{r}(\mathbf{x})\}$ *is nonempty and the optimal value of* (2-2) *is bounded for all* $\mathbf{x} \in X$ *and* $\boldsymbol{\xi} \in \Xi$*, and (iii) the uncertainty support* $\Xi \subseteq \mathbb{R}^{d_\xi}$ *is a hypercube, i.e.,* $\Xi = \bigtimes_{i=1}^{d_\xi}[a_i, b_i]$.

Within this framework, for any given $\mathbf{x} \in X$, the constraint (3-5b) is equivalent to

$$\sup_{\boldsymbol{\xi}\in\Xi} \left(Q(\mathbf{x},\boldsymbol{\xi}) - \lambda\|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}_n\|_1 \right) \leq s_n, \quad \forall n \leq N. \tag{3-6}$$

By representing the norm with linear inequalities and considering the dual formulation of $Q(\mathbf{x},\boldsymbol{\xi})$, the left-hand side of constraint (3-6) is equivalent to the following nonlinear optimization problem (with bilinear term $\boldsymbol{\theta}^\top\boldsymbol{\xi}$):

$$
\begin{aligned}
\sup_{\theta,\alpha,\xi} \quad & \theta^\top \mathbf{r}(\mathbf{x}) + [\boldsymbol{H}^\top(\mathbf{x})\theta]^\top \xi - \lambda \mathbf{1}^\top \alpha \\
\text{s.t.} \quad & \alpha \geq \xi - \widehat{\xi}_n, \\
& \alpha \geq \widehat{\xi}_n - \xi, \\
& \mathbf{W}^\top \theta \leq \mathbf{q}, \\
& \xi \in \Xi.
\end{aligned}
\tag{3-7}
$$

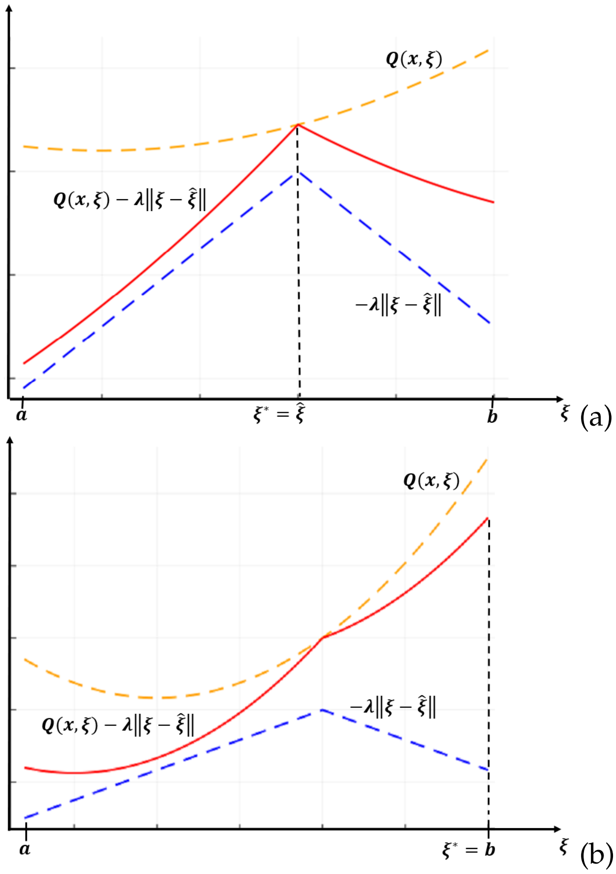

Figure 3.2: Illustrative example of the optimal solution ξ^* of (3-7) equals to $\widehat{\xi}_n$ (a) and extreme point of the uncertainty support (b).

As illustrated by Fig. 3.2 and stated in Proposition 2, under Assumption 1, for a given scenario $\widehat{\xi}_n$, $n \leq N$, the i-th component of the optimal solution ξ^* of the equivalent inner problem (3-7) is an extreme point of the interval $[a_i, b_i]$ or the i-th component of the nominal value $(\widehat{\xi}_n)_i$.

Proposition 2. *Suppose Assumption 1 holds. Then, there exists an optimal solution* ξ^* *to* (3-7) *such that* $\xi_i^* \in \{(\widehat{\xi}_n)_i, a_i, b_i\}$, $i = 1, \ldots, d_\xi$.

Proof. *Consider the rightmost problem in* (3-7), *and write it as*

$$
\sup_{\theta : \mathbf{W}^\top \theta \leq \mathbf{q}} \quad \theta^\top \mathbf{r}(\mathbf{x}) + J(\theta)
$$

where

$$J(\boldsymbol{\theta}) := \sup_{\boldsymbol{\alpha}\geq 0} \; -\lambda \mathbf{1}^\top \boldsymbol{\alpha} + V(\boldsymbol{\theta},\boldsymbol{\alpha}), \tag{3-8}$$

and $V(\boldsymbol{\theta},\boldsymbol{\alpha})$ *is given by*

$$\begin{aligned} \max_{\boldsymbol{\xi}\in\Xi} \quad & [\boldsymbol{H}^\top(\mathbf{x})\boldsymbol{\theta}]^\top \boldsymbol{\xi} \\ s.t. \quad & \boldsymbol{\xi} \leq \widehat{\boldsymbol{\xi}}_n + \boldsymbol{\alpha}, \\ & \boldsymbol{\xi} \geq \widehat{\boldsymbol{\xi}}_n - \boldsymbol{\alpha}, \end{aligned} \tag{3-9}$$

i.e.,

$$V(\boldsymbol{\theta},\boldsymbol{\alpha}) = \max_{\boldsymbol{\xi}\in\Xi} \mathbf{w}(\boldsymbol{\theta})^\top \boldsymbol{\xi} - \mathbb{I}_C(\boldsymbol{\xi},\boldsymbol{\alpha}) \tag{3-10}$$

where $\mathbf{w}(\boldsymbol{\theta}) := \boldsymbol{H}^\top(\mathbf{x})\boldsymbol{\theta}$*, and* $\mathbb{I}_C(\boldsymbol{\xi},\boldsymbol{\alpha})$ *is the indicator function of the set*

$$C := \{(\boldsymbol{\xi},\boldsymbol{\alpha}) \,:\, |\xi_i - (\widehat{\boldsymbol{\xi}}_n)_i| \leq \alpha_i, \quad i = 1,\ldots,d_\xi\}.$$

Since the set C is convex, it follows that $-\mathbb{I}_C(\boldsymbol{\xi},\boldsymbol{\alpha})$ *is concave and therefore* $V(\boldsymbol{\theta},\boldsymbol{\alpha})$ *is concave in* $\boldsymbol{\alpha}$*. Moreover, it is easy to see that the problem in* (3-10) *can be decomposed per coordinate* ξ_i*, i.e.,* $V(\boldsymbol{\theta},\boldsymbol{\alpha}) = \sum_{i=1}^{d_\xi} V_i(\boldsymbol{\theta},\alpha_i)$*, where*

$$V_i(\boldsymbol{\theta},\alpha_i) = \max\{w_i(\boldsymbol{\theta})\xi_i \,:\, |\xi_i - (\widehat{\boldsymbol{\xi}}_n)_i| \leq \alpha_i,\ \xi_i \in [a_i, b_i]\}. \tag{3-11}$$

The solution to (3-11) *is trivial to determine: if* $w_i(\boldsymbol{\theta}) > 0$*, then* $\xi_i^*(\alpha_i) = \min\{b_i,\ (\widehat{\boldsymbol{\xi}}_n)_i + \alpha_i\}$*; otherwise,* $\xi_i^*(\alpha_i) = \max\{a_i,\ (\widehat{\boldsymbol{\xi}}_n)_i - \alpha_i\}$*. It follows that the function* $J(\boldsymbol{\theta})$ *defined in* (3-8) *can be decomposed as*

$$J(\boldsymbol{\theta}) := \sum_{i=1}^{d_\xi} \sup_{\alpha_i\geq 0} \begin{cases} -\lambda\alpha_i + w_i(\boldsymbol{\theta})\min\{b_i,\ (\widehat{\boldsymbol{\xi}}_n)_i + \alpha_i\} & \text{if } w_i(\boldsymbol{\theta}) > 0 \\ -\lambda\alpha_i + w_i(\boldsymbol{\theta})\max\{a_i,\ (\widehat{\boldsymbol{\xi}}_n)_i - \alpha_i\} & \text{otherwise.} \end{cases} \tag{3-12}$$

We see that, when $w_i(\boldsymbol{\theta}) > 0$*, the expression inside the sup in* (3-12) *is maximized at* $\alpha_i = 0$ *if* $\lambda > w_i(\boldsymbol{\theta})$*, and it is maximized at all values of* $\alpha_i \geq b_i - (\widehat{\boldsymbol{\xi}}_n)_i$ *if* $\lambda \leq w_i(\boldsymbol{\theta})$*. Similarly, when* $w_i(\boldsymbol{\theta}) \leq 0$*, the expression inside the sup in* (3-12) *is maximized at* $\alpha_i = 0$ *if* $\lambda > |w_i(\boldsymbol{\theta})|$*, and it is maximized at all values of* $\alpha_i \geq (\widehat{\boldsymbol{\xi}}_n)_i - a_i$ *if* $\lambda \leq |w_i(\boldsymbol{\theta})|$*. The value of* $J(\boldsymbol{\theta})$ *is then given by*

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{d_\xi} \begin{cases} -w_i(\boldsymbol{\theta})(\widehat{\boldsymbol{\xi}}_n)_i & \text{if } \lambda > |w_i(\boldsymbol{\theta})| \\ -\lambda(b_i - (\widehat{\boldsymbol{\xi}}_n)_i) + w_i(\boldsymbol{\theta})b_i & \text{if } \lambda \leq |w_i(\boldsymbol{\theta})| \text{ and } w_i(\boldsymbol{\theta}) > 0 \\ -\lambda((\widehat{\boldsymbol{\xi}}_n)_i - a_i) + w_i(\boldsymbol{\theta})a_i & \text{if } \lambda \leq |w_i(\boldsymbol{\theta})| \text{ and } w_i(\boldsymbol{\theta}) \leq 0 \end{cases}$$

Moreover, by substituting the optimal values of α_i *found above into* (3-11)*, we conclude that an optimal solution to the maximization problem in* (3-11) *is* *given by*

$$\begin{cases} \zeta_i^* = (\widehat{\boldsymbol{\xi}}_n)_i & \text{if } \lambda > |w_i(\boldsymbol{\theta})| \\ \zeta_i^* = b_i & \text{if } \lambda \leq |w_i(\boldsymbol{\theta})| \text{ and } w_i(\boldsymbol{\theta}) > 0 \\ \zeta_i^* = a_i & \text{if } \lambda \leq |w_i(\boldsymbol{\theta})|) \text{ and } w_i(\boldsymbol{\theta}) \leq 0. \end{cases}$$

That is, $\zeta_i^* \in \{(\widehat{\boldsymbol{\xi}}_n)_i, a_i, b_i\}$ regardless of the value of $\boldsymbol{\theta}$. We conclude that there always exists an optimal solution $\boldsymbol{\xi}^*$ to (3-7) *such that $\zeta_i^* \in \{(\widehat{\boldsymbol{\xi}}_n)_i, a_i, b_i\}$, $i = 1, \ldots, d_\xi$.* $\square$

Let us introduce the notation $\widehat{\Xi}_n = \bigtimes_{i=1}^{d_\xi}\{a_i, (\widehat{\boldsymbol{\xi}}_n)_i, b_i\} = \{\xi_\ell^*\}_{\ell \in \mathcal{L}_n}$ and $\mathcal{L}_n = \{1, \ldots, |\widehat{\Xi}_n|\}$, for all $n = 1 \ldots, N$. According to **Proposition 2**, the set $\widehat{\Xi}_n$ comprises the (eligible) candidates for optimal solution of the sup problem on the left side of (3-6) for any feasible pair $(\mathbf{x}, \lambda)$, and for each scenario $\widehat{\boldsymbol{\xi}}_n$, $n = 1, \ldots, N$. With this notation at hand, by replacing the infinite set Ξ with the finite set $\widehat{\Xi}_n$, for all $n = 1, \ldots, N$, in the constraint (3-5b), the problem (3-5) reduces to the following linear program:

$$\min_{\mathbf{x}, \lambda, s_n} \mathbf{c}^\top \mathbf{x} + \lambda\delta + \frac{1}{N}\sum_{n=1}^{N} s_n \tag{3-13a}$$

$$\text{s.t.} \quad Q(\mathbf{x}, \boldsymbol{\xi}_\ell^*) - \lambda\|\boldsymbol{\xi}_\ell^* - \widehat{\boldsymbol{\xi}}_n\|_1 \leq s_n, \quad \forall \ell \in \mathcal{L}_n, \forall n \leq N, \tag{3-13b}$$

$$\lambda \geq 0, \tag{3-13c}$$

$$\mathbf{x} \in \mathbf{X}, \tag{3-13d}$$

which scales with the number of candidate solutions in $\mathcal{L}_n$, for all $n = 1, \ldots, N$. In the next section, we propose decomposition schemes to handle the extensive linear program (3-13).

Note that **Proposition 2** also establishes an equivalence of the worst-case expectation problem in the left-hand side of constraint (3-6) with the distribution separation problem (see [28]) based on the Wasserstein set $\widehat{\mathbb{B}}_\delta(\widehat{\mathbb{P}}_N)$ defined below, which comprises all finite distributions supported on the finite set $\widehat{\Xi} := \bigcup_{n \leq N} \widehat{\Xi}_n$:

$$
\begin{aligned}
\widehat{\mathbb{B}}_\delta(\widehat{\mathbb{P}}_N) := \Big\{ \{v_\ell\}_{\ell \le N \cdot 3^{d_\xi}} \, : \, & \sum_{\ell=1}^{N \cdot 3^{d_\xi}} \sum_{n=1}^{N} \|\boldsymbol{\xi}^*_\ell - \widehat{\boldsymbol{\xi}}_n\|_1 \pi_{\ell n} \le \delta, \\
& \sum_{n=1}^{N} \pi_{\ell n} = v_\ell, \, \forall \ell \le N \cdot 3^{d_\xi}, \\
& \sum_{\ell=1}^{N \cdot 3^{d_\xi}} \pi_{\ell n} = 1/N, \, \forall n \le N, \\
& \sum_{\ell=1}^{N \cdot 3^{d_\xi}} v_\ell = 1, \\
& v_\ell \ge 0, \, \forall \ell \le N \cdot 3^{d_\xi}, \\
& \pi_{\ell n} \ge 0, \, \forall \ell \le N \cdot 3^{d_\xi}, \forall n \le N \Big\}.
\end{aligned}
$$

The distribution separation problem can be formulated as:

$$
\max \left\{ \sum_{\ell=1}^{N \cdot 3^{d_\xi}} v_\ell Q(\mathbf{x}, \boldsymbol{\xi}^*_\ell) \, : \, \{v_\ell\}_{\ell \le N \cdot 3^{d_\xi}} \in \widehat{\mathbb{B}}_\delta(\widehat{\mathbb{P}}_N) \right\}. \tag{3-14}
$$

Therefore, we have the following corollary.

Corollary 1. *For each $\mathbf{x} \in X$, the optimal value of the worst-case expectation problem of the DD-DRO-W problem equals the optimal value of the distribution separation problem which comprises the finite distributions supported on the set $\widehat{\Xi}$ within Wasserstein ambiguity set.*

Proof. *We have the following sequence of equivalences:*

$$
\sup_{\widetilde{P} \in \mathbb{B}_\delta(\widehat{P}_N)} \mathbb{E}^{\widetilde{P}}[Q(\mathbf{x}, \boldsymbol{\xi})] =
$$

$$
\begin{cases}
\min\limits_{\lambda, s_n} & \lambda\delta + \sum\limits_{n=1}^{N} \dfrac{1}{N} s_n \\
s.t. & \sup\limits_{\boldsymbol{\xi} \in \Xi} \Big(Q(\mathbf{x}, \boldsymbol{\xi}) - \lambda \|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}_n\|_1 \Big) \le s_n, \quad \forall n \le N, \\
& \lambda \ge 0.
\end{cases} \tag{3-15}
$$

$$
= \begin{cases}
\min\limits_{\lambda, s_n} & \lambda\delta + \sum\limits_{n=1}^{N} \dfrac{1}{N} s_n \\
s.t. & \sup\limits_{\boldsymbol{\xi} \in \widehat{\Xi}_n} \Big(Q(\mathbf{x}, \boldsymbol{\xi}) - \lambda \|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}_n\|_1 \Big) \le s_n, \quad \forall n \le N, \\
& \lambda \ge 0.
\end{cases} \tag{3-16}
$$

$$= \max \left\{ \sum_{\ell=1}^{N \cdot 3^{d_\xi}} v_\ell Q(\mathbf{x}, \boldsymbol{\xi}_\ell^*) \; : \; \{v_\ell\}_{\ell \leq N \cdot 3^{d_\xi}} \in \widehat{\mathbb{B}}_\delta(\widehat{\mathbb{P}}_N) \right\}. \tag{3-17}$$

Equality (3-15) *is derived by [17, 35]. Equality* (3-16) *holds by* ***Proposition 2****. Finally, equality* (3-17) *is obtained by applying again the result from [17, 35] to the convex reduction of the worst-expectation problem over the Wasserstein ambiguity set for the discrete distributions supported on the set* $\widehat{\Xi}$*.*

□

Note that $|\widehat{\Xi}| = N \cdot 3^{d_\xi}$ by definition of set $\widehat{\Xi}_n$, so the decomposition methodology proposed [28] may be computationally intractable even for moderately high dimensions. Therefore, for low uncertainty dimensionality d_ξ, we can address the problem (3-13) by using the decomposition method presented in [28] which uses the distribution separation problem (3-17). However, as the problem (3-17) has a large number of variables for high uncertainty dimensionality, it may not be possible to solve it due to memory or time constraints. That precise issue was encountered by [36], who circumvented the problem by using machine learning techniques to select the random variables with most impact on the model, and then applying the algorithm of [28] only to those variables.

3.5 Decomposition methods

In this section, we present the proposed numerical schemes to solve the tractable reformulation (3-13) of the two-stage distributionally robust optimization problem with right-hand-side uncertainty under a data-driven Wasserstein based ambiguity set. We show that this problem is suitable to be solved by three exact decomposition methods: the Benders multi-cut and single-cut methods and the column and constraint generation method (C&CG).

In general, a decomposition method can be implemented in a master-oracle scheme; see, e.g., [13] for an application of that technique to a robust two-stage model. In our context, the master problem is a relaxation of the equivalent linear program (3-13) of the two-stage DD-DRO-W problem. Given the solution of the master (relaxed) problem, the oracle identifies the worst infeasibility to add the corresponding constraint (or block of constraints and variables) to the master problem. This iterative procedure stops whenever the oracle asserts that the master solution is feasible for the original problem (3-13).

3.5.1 Column and constraint generation method

We start by proposing a solution methodology to address the problem (3-13) by using the column and constraint generation method (C&CG). We develop a iterative procedure based on lower and upper bounding approximations of the linear program (3-13) which converges to its optimal value and optimal solution.

By considering the primal formulation (2-2) of the recourse function $Q(\mathbf{x}, \boldsymbol{\xi})$, the problem

$$\min_{\mathbf{x},\mathbf{y}_\ell,\lambda,s_n} \mathbf{c}^\top \mathbf{x} + \lambda\delta + \frac{1}{N}\sum_{n=1}^{n} s_n \tag{3-18a}$$

$$\text{s.t.} \quad \mathbf{q}^\top \mathbf{y}_\ell - \lambda \|\boldsymbol{\xi}_\ell^* - \widehat{\boldsymbol{\xi}}_n\|_1 \leq s_n, \quad \forall \ell \in \mathcal{L}_n^K, \forall n \leq N, \tag{3-18b}$$

$$\mathbf{W}\mathbf{y}_\ell = \mathbf{H}(\mathbf{x})\boldsymbol{\xi}_\ell^* + \mathbf{r}(\mathbf{x}), \quad \forall \ell \in \mathcal{L}_n^K, \forall n \leq N, \tag{3-18c}$$

$$\mathbf{y}_\ell \geq 0, \quad \forall \ell \in \mathcal{L}_n^K \forall n \leq N, \tag{3-18d}$$

$$\lambda \geq 0, \tag{3-18e}$$

$$\mathbf{x} \in X, \tag{3-18f}$$

is equivalent to (3-13) if the set $\mathcal{L}_n^K$ equals $\mathcal{L}_n$. Instead, if $\mathcal{L}_n^K \subset \mathcal{L}_n$ is a subset defined at iteration K of the iterative procedure, problem (3-18) represents a relaxation of (3-13). Therefore, its optimal value is a valid lower bound, LB, for the optimal value of problem (3-13). Henceforth, we call the relaxed problem (3-18) as the master problem.

For the current optimal solution $(\mathbf{x}^K, \lambda^K, \mathbf{s}^K)$ of the master problem (3-18), where $\mathbf{s}^K = (s_n^K)_{n\leq N}$, we need an oracle problem to find the worst-case uncertainty realization $\boldsymbol{\xi}_\ell^* \in \widehat{\Xi}_n$ maximizing the infeasibility of constraint (3-18b). Based on this result, we update the subset $\mathcal{L}_n^K$, for all $n = 1, \ldots, N$, and add the block of linear constraints (3-18b)-(3-18c) and variables (3-18d) - new columns - to the master problem. With this in mind, let us consider the left-hand-side of the inequality (3-6). We have the following equivalences:

$$\sup_{\boldsymbol{\xi}\in\Xi} \left(Q(\mathbf{x}^K,\boldsymbol{\xi}) - \lambda^K \|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}_n\|_1 \right) \tag{3-19a}$$

$$= \left\{ \begin{array}{lll} \sup\limits_{\boldsymbol{\xi}\in\Xi} & \min\limits_{\mathbf{y}\geq 0} & \mathbf{q}^\top\mathbf{y} - \lambda^K \|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}_n\|_1 \\ & \text{s.t.} & \mathbf{W}\mathbf{y} = \mathbf{H}(\mathbf{x}^K)\boldsymbol{\xi} + \mathbf{r}(\mathbf{x}^K) \end{array} \right\} \tag{3-19b}$$

$$= \left\{ \begin{array}{ll} \max\limits_{\boldsymbol{\theta},\boldsymbol{\xi}} & \boldsymbol{\theta}^\top \mathbf{r}(\mathbf{x}^K) + [\boldsymbol{H}^\top(\mathbf{x}^K)\boldsymbol{\theta}]^\top\boldsymbol{\xi} - \lambda^K\|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}_n\|_1 \\ \text{s.t.} & \mathbf{W}^\top\boldsymbol{\theta} \leq \mathbf{q} \\ & \boldsymbol{\xi} \in \Xi. \end{array} \right\} \tag{3-19c}$$

We consider the bilinear problem (3-19c) as the oracle problem which can be reduced to a MILP problem by linearizing the products of binary and continuous variables. To that end, let us consider the inner problem in the left-hand side of inequality (3-6) and introduce the notation $\boldsymbol{\Delta}^+ = \mathbf{b} - \widehat{\boldsymbol{\xi}}_n$, $\boldsymbol{\Delta}^- = \widehat{\boldsymbol{\xi}}_n - \mathbf{a}$[3]. With this notation at hand, according to **Proposition 2**, the optimal solution $\boldsymbol{\xi}^*$ of problem (3-7) - which is equivalent to problem (3-19c) - can be expressed by:

$$\boldsymbol{\xi}^* = \widehat{\boldsymbol{\xi}}_n + \text{diag}(\boldsymbol{\Delta}^+)\mathbf{z}^+ - \text{diag}(\boldsymbol{\Delta}^-)\mathbf{z}^-,$$

where $\mathbf{z}^+, \mathbf{z}^- \in \{0,1\}^{d_\xi}$ are binary vector variables and $\text{diag}(\boldsymbol{\Delta}^i) = \text{diag}(\Delta_1^i, \ldots, \Delta_{d_\xi}^i)$, denotes the diagonal matrix of the vector $\boldsymbol{\Delta}^i$, for $i = +, -$, i.e.,

$$\begin{aligned} \boldsymbol{\xi}^* &= \begin{bmatrix} \widehat{\xi}_1 \\ \vdots \\ \widehat{\xi}_{d_\xi} \end{bmatrix} + \begin{bmatrix} \Delta_1^+ & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Delta_{d_\xi}^+ \end{bmatrix} \begin{bmatrix} z_1^+ \\ \vdots \\ z_{d_\xi}^+ \end{bmatrix} - \begin{bmatrix} \Delta_1^- & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Delta_{d_\xi}^- \end{bmatrix} \begin{bmatrix} z_1^- \\ \vdots \\ z_{d_\xi}^- \end{bmatrix}. \\ &= \begin{bmatrix} \widehat{\xi}_1 + \Delta_1^+ z_1^+ - \Delta_1^- z_1^- \\ \vdots \\ \widehat{\xi}_{d_\xi} + \Delta_{d_\xi}^+ z_{d_\xi}^+ - \Delta_{d_\xi}^- z_{d_\xi}^- \end{bmatrix} \end{aligned}$$

Thus, the decision variable $\boldsymbol{\xi}$ of the optimization problem (3-7) can be replaced by binary decision variables $\mathbf{z}^i$, $i = +, -$, and the optimal value of problem (3-7) equals:

[3]The i-th component of the vectors $\mathbf{a} = (a_i)_{i\leq d_\xi}$ and $\mathbf{b} = (b_i)_{i\leq d_\xi}$ coincides with the lower and upper limit of the interval $[a_i, b_i]$, respectively, along on the i-th dimension of the uncertainty support.

$$
\begin{aligned}
\max_{\boldsymbol{\theta},\mathbf{z}^+,\mathbf{z}^-} \quad & \boldsymbol{\theta}^\top \mathbf{r}(\mathbf{x}^K) + [\boldsymbol{H}^\top(\mathbf{x}^K)\boldsymbol{\theta}]^\top[\widehat{\boldsymbol{\xi}}_n + \mathrm{diag}(\boldsymbol{\Delta}^+)\mathbf{z}^+ - \mathrm{diag}(\boldsymbol{\Delta}^-)\mathbf{z}^-] - \lambda^K \mathbf{1}^\top \boldsymbol{\alpha} \\
\text{s.t.} \quad & \mathbf{W}^\top \boldsymbol{\theta} \le \mathbf{q} \\
& \widehat{\boldsymbol{\xi}}_n + \mathrm{diag}(\boldsymbol{\Delta}^+)\mathbf{z}^+ - \mathrm{diag}(\boldsymbol{\Delta}^-)\mathbf{z}^- \le \boldsymbol{\alpha} \\
& \widehat{\boldsymbol{\xi}}_n + \mathrm{diag}(\boldsymbol{\Delta}^-)\mathbf{z}^- - \mathrm{diag}(\boldsymbol{\Delta}^+)\mathbf{z}^+ \le \boldsymbol{\alpha}.
\end{aligned} \tag{3-20}
$$

However, problem (3-20) has bilinear terms of products of binary and continuous variables which can be linearized by disjunctive constraint following [37]. For the detailed exact linearized mixed integer linear programming (MILP) formulation, see Appendix A.

Let us denote by $\rho_n(\mathbf{x}^K, \lambda^K)$ the optimal value of the oracle problem (3-20) (or equivalently problem (3-19c)). For a fixed $(\mathbf{x}^K, \lambda^K)$ we claim that

$$
s_n^K \le \rho_n(\mathbf{x}^K, \lambda^K), \quad \forall n \le N. \tag{3-21}
$$

Indeed, we have that constraint (3-18b) is active for the $\ell \in \mathcal{L}_n^K$ that maximizes the left-hand-side of constraint (3-18b) over the set $\mathcal{L}_n^K$, for all $n = 1, \ldots, N$, i.e.,

$$
\max_{\ell \in \mathcal{L}_n^K} \left(\mathbf{q}^\top \mathbf{y}_\ell^* - \lambda^K \|\boldsymbol{\xi}_\ell^* - \widehat{\boldsymbol{\xi}}_n\|_1 \right) = s_n^K, \, \forall n \le N, \tag{3-22}
$$

where $\mathbf{y}_\ell^*$ denotes the optimal solution of the primal second-stage variable $\mathbf{y}_\ell$, for all $\ell \in \mathcal{L}_n^K$, $n = 1, \ldots, N$, within problem (3-18). Therefore, we have the following valid inequality:

$$
\begin{aligned}
\rho_n(\mathbf{x}^K, \lambda^K) &= \sup_{\boldsymbol{\xi} \in \Xi} \left(Q(\mathbf{x}^K, \boldsymbol{\xi}) - \lambda^K \|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}_n\|_1 \right) && \text{(3-23a)} \\
&\ge \max_{\ell \in \mathcal{L}_n^K} \left(\mathbf{q}^\top \mathbf{y}_\ell^* - \lambda^K \|\boldsymbol{\xi}_\ell^* - \widehat{\boldsymbol{\xi}}_n\|_1 \right) = s_n^K, && \text{(3-23b)}
\end{aligned}
$$

for all $n = 1, \ldots, N$. Equality (3-23a) holds because the equivalence between problems (3-19c) and (3-19a), whereas (3-23b) follows from the fact that the optimization problem on the right-hand side of (3-23a) is less constrained than that of (3-23b). Hence, a valid upper bound UB for the problem (3-13) can be obtained as

$$
UB = \mathbf{c}^\top \mathbf{x}^K + \lambda^K \delta + \frac{1}{N} \sum_{n=1}^{N} \rho_n(\mathbf{x}^K, \lambda^K). \tag{3-24}
$$

The algorithm converges whenever the $UB - LB \le \varepsilon$, i.e, the current solution $(\mathbf{x}^K, \lambda^K, \mathbf{s}_n^K)$ lies within a user-defined tolerance level ε.

We summarize the column and constraint generation algorithm in the following pseudo-code:

Algorithm 5 Column and constraint generation method

Initialization: Set $K = 0$, $UB \longleftarrow +\infty$ and $LB \longleftarrow -\infty$, and $\mathcal{L}_n^K \longleftarrow \{n\}$ for $n = 1, ..., N$
while $UB - LB > \varepsilon$ **do**
  Solve the Master problem (3-18);
  Store the Master solution: $(LB, \mathbf{x}^K, \lambda^K, \mathbf{s}^K)$;
  for n=1 **to** N **do**
    Solve the MILP version of the Oracle problem (3-20) for $(\mathbf{x}^K, \lambda^K)$;
    Store the Oracle solution: $\rho_n(\mathbf{x}^K, \lambda^K), (\boldsymbol{\theta}^*, \boldsymbol{\xi}^*)$;
    if $\rho_n(\mathbf{x}^K, \lambda^K) > s_n^K$ **then**
      $\ell \longleftarrow N(K+1) + n$;
      Update $\mathcal{L}_n^{K+1} \longleftarrow \mathcal{L}_n^K \cup \{\ell\}$ and make $\boldsymbol{\xi}_\ell^* \longleftarrow \boldsymbol{\xi}^*$
    end if
  end for
  Set $UB \longleftarrow \min\{UB,\ \mathbf{c}^\top \mathbf{x}^K + \lambda^K \delta + \frac{1}{N}\sum_{n=1}^N \rho_n(\mathbf{x}^K, \lambda^K)\}$
  Update $K \longleftarrow K+1$, add the block of linear constraints (3-18b)–(3-18c) and variables (3-18d)
end while
return $\mathbf{x}^K$, UB, LB

For initialization purpose, we can solve the second-stage problem

$$\min_{\mathbf{y} \geq 0} \mathbf{q}^\top \mathbf{y} \tag{3-25a}$$

$$\text{s.t. } \mathbf{W}\mathbf{y} = \mathbf{H}(\overline{\mathbf{x}})\widehat{\boldsymbol{\xi}}_n + \mathbf{r}(\overline{\mathbf{x}}) \quad : \boldsymbol{\theta}_n \tag{3-25b}$$

for the optimal solution $\overline{\mathbf{x}}$ of the deterministic equivalent problem for the average scenario, $\overline{\boldsymbol{\xi}} = \frac{1}{N}\sum_{n \leq N} \widehat{\boldsymbol{\xi}}_n$, and cast the dual variable $\boldsymbol{\theta}_n$ of the constraint (3-25b)—which is a dual vertex—for all $n = 1, \ldots, N$. We then initialize the algorithm (5) with $\boldsymbol{\xi}_n^* \longleftarrow \widehat{\boldsymbol{\xi}}_n$ for all $n = 1, \ldots, N$.

3.5.2 Multi-cut Benders

We can also address the problem (3-13) by using a multi-cut Benders algorithm. By strong duality, we can assess $Q(\mathbf{x}, \boldsymbol{\xi}_\ell^*)$ by:

$$Q(\mathbf{x}, \boldsymbol{\xi}_\ell^*) = \max_{d \in \mathcal{D}} \left\{ \boldsymbol{\theta}_d^\top \mathbf{r}(\mathbf{x}) + [\boldsymbol{H}^\top(\mathbf{x})\boldsymbol{\theta}_d]^\top \boldsymbol{\xi}_\ell^* \right\}, \tag{3-26}$$

for all $\ell \in \mathcal{L}_n$ and $\mathbf{x} \in X$, where $\{\boldsymbol{\theta}_d\}_{d \in \mathcal{D}}$ is the set of vertices of the dual polyhedron $\{\boldsymbol{\theta} : \mathbf{W}^\top \boldsymbol{\theta} \leq \mathbf{q}\}$, hereafter referred to as dual vertices. Thus, by replacing $Q(\mathbf{x}, \boldsymbol{\xi}_\ell^*)$ with the enumeration of the affine functions

$\left\{\boldsymbol{\theta}_d^\top \mathbf{r}(\mathbf{x}) + [\boldsymbol{H}^\top(\mathbf{x})\boldsymbol{\theta}_d]^\top \boldsymbol{\xi}_\ell^*\right\}_{d\in\mathcal{D}}$ in (3-13b), the problem

$$\min_{\mathbf{x},\lambda,s_n} \mathbf{c}^\top \mathbf{x} + \lambda\delta + \frac{1}{N}\sum_{n=1}^{N} s_n \tag{3-27a}$$

$$\text{s.t.} \quad \begin{aligned} &\boldsymbol{\theta}_d^\top \mathbf{r}(\mathbf{x}) + [\boldsymbol{H}^\top(\mathbf{x})\boldsymbol{\theta}_d]^\top \boldsymbol{\xi}_\ell^* - \lambda\|\boldsymbol{\xi}_\ell^* - \widehat{\boldsymbol{\xi}}_n\|_1 \le s_n, \\ &\qquad\qquad\qquad\qquad \forall (d,\ell) \in \mathcal{D}\times\mathcal{L}_n,\ \forall n \le N, \end{aligned} \tag{3-27b}$$

$$\lambda \ge 0, \tag{3-27c}$$

$$\mathbf{x} \in X, \tag{3-27d}$$

meets problem (3-13).

We can derive an alternative master problem from the linear programming relaxation of the equivalent problem (3-27). Observe that by solving the oracle problem (3-19c), we obtain an uncertainty realization and a dual vertex that maximizes the infeasibility of constraint (3-27b). Let

$$(\boldsymbol{\theta}_n^k, \boldsymbol{\xi}_n^k) \in \arg\max_{\boldsymbol{\theta},\boldsymbol{\xi}} \left\{ \boldsymbol{\theta}^\top \mathbf{r}(\mathbf{x}^k) + [\boldsymbol{H}^\top(\mathbf{x}^k)\boldsymbol{\theta}]^\top \boldsymbol{\xi} - \lambda^k \|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}_n\|_1 \;\middle|\; \begin{array}{c} \mathbf{W}^\top \boldsymbol{\theta} \le \mathbf{q}, \\ \boldsymbol{\xi} \in \Xi \end{array} \right\} \tag{3-28}$$

denote the optimal solution of the oracle problem, where $(\mathbf{x}^k, \lambda^k)$ is the optimal solution, at iteration $k \le K$, of the master problem for the finite extensive equivalent form (3-27). Therefore, a linear programming relaxation of (3-27) can be derived by substituting constraint (3-27b) with optimality cuts:

$$[\boldsymbol{\theta}_n^k]^\top \mathbf{r}(\mathbf{x}) + [\boldsymbol{H}^\top(\mathbf{x})\boldsymbol{\theta}_n^k]^\top \boldsymbol{\xi}_n^k - \lambda\|\boldsymbol{\xi}_n^k - \widehat{\boldsymbol{\xi}}_n\|_1 \le s_n,\ \forall k \le K,\ n \le N, \tag{3-29}$$

Note that for K sufficiently large, the relaxed problem equals (3-27) if

$$\{(\boldsymbol{\theta}_n^k, \boldsymbol{\xi}_n^k) \,|\, k \le K\} = \{(\boldsymbol{\theta}_d, \boldsymbol{\xi}_\ell^*) \,|\, d \in \mathcal{D}, \ell \in \mathcal{L}_n\},$$

for all $n = 1, \ldots, N$.

3.5.3 Single-cut Benders

For the purpose of constructing the single-cut Benders algorithm, we develop an equivalent formulation that incorporate additional valid constraints to construct the average cut. The equivalent formulation for (3-27) is obtained by replacing (3-27b) with

$$\boldsymbol{\theta}_d^\top \mathbf{r}(\mathbf{x}) + [\boldsymbol{H}^\top(\mathbf{x})\boldsymbol{\theta}_d]^\top \boldsymbol{\xi}_\ell^* - \lambda\|\boldsymbol{\xi}_\ell^* - \widehat{\boldsymbol{\xi}}_n\|_1 \leq s_n, \forall (d,\ell) \in \mathcal{D} \times \mathcal{L}, \forall n \leq N, \tag{3-30}$$

where $\mathcal{L} = \cup_{n \leq N} \mathcal{L}_n$. To see that such equivalence holds, notice that

$$\max_{\ell \in \mathcal{L}} Q(\mathbf{x}, \boldsymbol{\xi}_\ell^*) - \lambda\|\boldsymbol{\xi}_\ell^* - \widehat{\boldsymbol{\xi}}_n\|_1 \leq \sup_{\xi \in \Xi} Q(\mathbf{x}, \boldsymbol{\xi}) - \lambda\|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}_n\|_1 \tag{3-31a}$$

$$= \max_{\ell \in \mathcal{L}_n} Q(\mathbf{x}, \boldsymbol{\xi}_\ell^*) - \lambda\|\boldsymbol{\xi}_\ell^* - \widehat{\boldsymbol{\xi}}_n\|_1 \tag{3-31b}$$

$$\leq \max_{\ell \in \mathcal{L}} Q(\mathbf{x}, \boldsymbol{\xi}_\ell^*) - \lambda\|\boldsymbol{\xi}_\ell^* - \widehat{\boldsymbol{\xi}}_n\|_1. \tag{3-31c}$$

The first inequality (3-31a) is valid since $\{\xi_\ell\}_{\ell \in \mathcal{L}} \subseteq \Xi$, while the second equality (3-31b) is guaranteed by Proposition 2. Finally, the last inequality (3-31c) holds since $\{\xi_\ell\}_{\ell \in \mathcal{L}_n} \subseteq \{\xi_\ell\}_{\ell \in \mathcal{L}}$. It follows that

$$\max_{\ell \in \mathcal{L}} Q(\mathbf{x}, \boldsymbol{\xi}_\ell^*) - \lambda\|\boldsymbol{\xi}_\ell^* - \widehat{\boldsymbol{\xi}}_n\|_1 = \max_{\ell \in \mathcal{L}_n} Q(\mathbf{x}, \boldsymbol{\xi}_\ell^*) - \lambda\|\boldsymbol{\xi}_\ell^* - \widehat{\boldsymbol{\xi}}_n\|_1$$

and consequently

$$\begin{aligned} &\max_{(d,\ell) \in \mathcal{D} \times \mathcal{L}} \boldsymbol{\theta}_d^\top \mathbf{r}(\mathbf{x}) + [\boldsymbol{H}^\top(\mathbf{x})\boldsymbol{\theta}_d]^\top \boldsymbol{\xi}_\ell^* - \lambda\|\boldsymbol{\xi}_\ell^* - \widehat{\boldsymbol{\xi}}_n\|_1 \\ = &\max_{(d,\ell) \in \mathcal{D} \times \mathcal{L}_n} \boldsymbol{\theta}_d^\top \mathbf{r}(\mathbf{x}) + [\boldsymbol{H}^\top(\mathbf{x})\boldsymbol{\theta}_d]^\top \boldsymbol{\xi}_\ell^* - \lambda\|\boldsymbol{\xi}_\ell^* - \widehat{\boldsymbol{\xi}}_n\|_1. \end{aligned}$$

Now, we can rewrite the problem (3-27) as a finite (extensive) linear program with average cuts by replacing (3-27b) with (3-30) and taking the average of the latter over n:

$$\min_{\mathbf{x}, \lambda, \beta} \mathbf{c}^\top \mathbf{x} + \lambda\delta + \beta \tag{3-32a}$$

$$\text{s.t.} \quad \frac{1}{N} \sum_{n=1}^{N} \Big([\boldsymbol{\theta}_d]^\top \mathbf{r}(\mathbf{x}) + [\boldsymbol{H}^\top(\mathbf{x})\boldsymbol{\theta}_d]^\top \boldsymbol{\xi}_\ell - \lambda\|\boldsymbol{\xi}_\ell - \widehat{\boldsymbol{\xi}}_n\|_1 \Big) \leq \beta, \forall (d,\ell) \in \mathcal{D} \times \mathcal{L}, \tag{3-32b}$$

$$\lambda \geq 0, \tag{3-32c}$$

$$\mathbf{x} \in X. \tag{3-32d}$$

Accordingly, we derive an alternative master-oracle scheme that uses the same oracle (3-28), but with the modified master

$$\min_{\mathbf{x},\lambda,\beta} \mathbf{c}^\top \mathbf{x} + \lambda\delta + \beta \tag{3-33a}$$

$$\text{s.t.} \quad \frac{1}{N}\sum_{n=1}^{N}\left([\boldsymbol{\theta}_n^k]^\top \mathbf{r}(\mathbf{x}) + [\boldsymbol{H}^\top(\mathbf{x})\boldsymbol{\theta}_n^k]^\top \boldsymbol{\xi}_n^k - \lambda\|\boldsymbol{\xi}_n^k - \widehat{\boldsymbol{\xi}}_n\|_1\right) \leq \beta, \forall k \leq K, \tag{3-33b}$$

$$\lambda \geq 0, \tag{3-33c}$$

$$\mathbf{x} \in X, \tag{3-33d}$$

where $(\boldsymbol{\theta}_n^k, \boldsymbol{\xi}_n^k)$ in this case denote the optimal solution of the oracle (3-28) for the optimal solution $(\mathbf{x}^k, \lambda^k)$ of the problem (3-33) at iteration $k-1$. Note that the resulting master problem (3-33) is less constrained than the linear programming relaxation (3-27). Therefore, the multi-cut algorithm provides a tighter lower bound than the single-cut master problem (3-33). In turn, the single-cut version of the algorithm relies on a reduced number of constraints (cuts). While this implies a lower computational effort to the master problem, it is very likely that the multi-cut will converge in fewer iterations. Thus, there is a tradeoff between these two versions of the algorithm that should be empirically studied, as we do in the next section.

4
Conservative solution for the stochastic unit commitment problem

The unit commitment problem is a major part of power system operation planning. The stochastic extension of this problem has become the standard procedure to address the uncertainties associated with the incorporation of renewable power generation. As usual, the stochastic unit commitment (SUC) problem has been assessed by the SAA method, however, solving this problem for large, real-world networks with a large number of scenarios leads to prohibitive computational times, especially when considering that the SUC problem needs to be solved in a few hours in order to decide the day-ahead operation.

As mentioned in [1, 10], a non-converged SAA solution can generate high disappointment levels because of its instability for a finite moderate sample. Given that the unit commitment problem relies on the implementation of a non-converged SAA solution, we developed a DRO model over the Wasserstein ambiguity set to obtain instead a conservative solution, which has a mathematical certificate against disappointment. The numerical results show the superiority of the DRO approach over the SAA solution.

We start from the formulation of the stochastic unit commitment problem and, then we present the numerical results.

4.1
Data-driven unit commitment formulation

As discussed in [35], when the probability distribution of the electricity load can not be accurately estimated, the obtained unit commitment decision can be biased. In the literature, DRO models have been already been applied to the unit commitment problem (see [38] and references therein). However, to the best of our knowledge, existing works rely mainly on approximations, such as affine policies, for the second stage variables. Our paper aims to address precisely this issue.

We use a classic formulation of a data-driven two-stage stochastic unit commitment problem [39]. The first stage comprises the commitment decisions while the second stage accounts for the dispatch decisions and power flow. We develop the problem by minimizing the expected total generation cost, and the uncertainty on the right-hand-side of the problem in the net electricity load parameter (load subtracted by uncertain renewable injections).

We summarize the notations by sets, parameters, first-stage variables and second-stage variables listed as follows:

Set	**Description**
$\mathcal{I}_b$	Set of electricity generators that are located in bus b
$\mathcal{T}$	Set of time of periods
$\mathcal{L}$	Set of lines
$\mathcal{B}$	Set of nodes
	$\mathcal{LI}_b = \{l \in \mathcal{L} : l = (k,b),\, k \in \mathcal{B}\}$
	$\mathcal{LO}_b = \{l \in \mathcal{L} : l = (b,k),\, k \in \mathcal{B}\}$

Table 4.1: Description of the sets.

Parameter	**Description**
N	Number of electricity load scenarios
S_l	Susceptance of the line $l \in \mathcal{L}$
C_i^u	Fixed cost of unit $i \in \mathcal{I}$
C_i^{SU}	Start-up cost of unit $i \in \mathcal{I}$
C_i^{SD}	Shut-down cost of unit $i \in \mathcal{I}$
R_i^U	Ramp-up limit of unit $i \in \mathcal{I}$
R_i^D	Ramp-down limit of unit $i \in \mathcal{I}$
$\overline{P}_i$	Maximum power generation of unit $i \in \mathcal{I}$
$\underline{P}_i$	Minimum power generation of unit $i \in \mathcal{I}$
$\widehat{\xi}_{b,t,n}$	The electricity load in bus $b \in \mathcal{B}$ in time $t \in \mathcal{T}$ corresponding to scenario $n \leq N$

Table 4.2: Description of the parameters.

Variable	**Description**
$u_{i,t}$	Binary commit variable: 1 if the thermal generator i is on in time t; 0 otherwise
$v_{i,t}$	Start-up variable for unit $i \in \mathcal{I}$ in time $t \in \mathcal{T}$
$w_{i,t}$	Shut-down variable for unit $i \in \mathcal{I}$ in time $t \in \mathcal{T}$

Table 4.3: Description of the first-stage variables.

Variable	Description
$p_{i,t,n}$	Power generation of unit $i \in \mathcal{I}$ in time $t \in \mathcal{T}$ in scenario $n \leq N$
$f_{l,t,n}$	Power flow of line $l \in \mathcal{L}$ in time $t \in \mathcal{T}$ in scenario $n \leq N$
$\theta_{b,t,n}$	Phase angle of node $b \in \mathcal{B}$ in time $t \in \mathcal{T}$ in scenario $n \leq N$

Table 4.4: Description of the second-stage variables.

Based on this notation, the mathematical formulation of the data-driven two-stage stochastic unit commitment problem is as follows:

$$\min_{u_{i,t},v_{i,t},w_{i,t},p_{i,t}} \quad \sum_{i\in\mathcal{I}}\sum_{t\in\mathcal{T}} \Big[\, C_i^u u_{i,t} + C_i^{SU} v_{i,t} + C_i^{SD} w_{i,t} + \frac{1}{N}\sum_{n=1}^{N} \left[C_i^P p_{i,t,n}\right] \Big] \tag{4-1a}$$

$$s.t. \quad \sum_{i\in\mathcal{I}_b} p_{i,t,n} + \sum_{l\in\mathcal{LI}_b} f_{l,t,n} - \sum_{l\in\mathcal{LO}_b} f_{l,t,n} = \widehat{\xi}_{b,t,n}, \quad \forall b \in \mathcal{B},\ \forall t \in \mathcal{T},\ \forall n \leq N \tag{4-1b}$$

$$f_{l,t,n} = S_l(\theta_{b,t,n} - \theta_{d,t,n}), \qquad l = (b,d) \in \mathcal{L},\ \forall t \in \mathcal{T},\ \forall n \leq N \tag{4-1c}$$

$$v_{i,t} - w_{i,t} = u_{i,t} - u_{i,t-1}, \qquad \forall i \in \mathcal{I},\ \forall t \in \mathcal{T}, \tag{4-1d}$$

$$v_{i,t} \leq u_{i,t}, \qquad \forall i \in \mathcal{I},\ \forall t \in \mathcal{T}, \tag{4-1e}$$

$$w_{i,t} \leq 1 - u_{i,t}, \qquad \forall i \in \mathcal{I},\ \forall t \in \mathcal{T}, \tag{4-1f}$$

$$p_{i,t} - pi,t-1 \leq R_i^U u_{i,t-1} + \bar{P}_i v_{i,t}, \qquad \forall i \in \mathcal{I},\ \forall t \in \mathcal{T}, \tag{4-1g}$$

$$p_{i,t-1} - pi,t \leq R_i^D u_{i,t} + \underline{P}_i w_{i,t}, \qquad \forall i \in \mathcal{I},\ \forall t \in \mathcal{T}, \tag{4-1h}$$

$$\underline{P}_i u_{i,t} \leq p_{i,t} \leq \bar{P}_i u_{i,t}, \qquad \forall i \in \mathcal{I},\ \forall t \in \mathcal{T}, \tag{4-1i}$$

$$0 \leq v_{i,t} \leq 1, \qquad \forall i \in \mathcal{I},\ \forall t \in \mathcal{T}, \tag{4-1j}$$

$$0 \leq w_{i,t} \leq 1, \qquad \forall i \in \mathcal{I},\ \forall t \in \mathcal{T}, \tag{4-1k}$$

$$u_{i,t} \in \{0,1\}, \qquad \forall i \in \mathcal{I},\ \forall t \in \mathcal{T}. \tag{4-1l}$$

In the above formulation, constraint (4-1b) ensures balancing the amount of power that flows into and out of each bus. Constraint (4-1c) represents the linearized Kirchhoff's law for a a DC power flow approximation according to which the power flow on a line l is proportional to the phase angle difference between the two end buses of the line. Constraints (4-1d)-(4-1e) are the start-up and shut-down operational constraints for each thermal unit. Constraints (4-1f) and (4-1g) are the the ramping up and ramping down constraints, respectively. Finally, constraint (4-1h) is the minimum and maximum power generation of the unit $i \in \mathcal{I}$ in time $t \in \mathcal{T}$.

We consider the distributionally robust optimization version of the problem (4-1). Regarding to the empirical distribution of historical data, we construct the Wasserstein-based ambiguity set for a given confidence level $\delta > 0$. We then consider the formulation (3-4) of the problem (4-1), where the first-stage variable $\mathbf{x}$ is $[u_{i,t}, v_{i,t}, w_{i,t}]_{i\in\mathcal{I},t\in\mathcal{T}}$ and $Q(\mathbf{x}, \xi)$ represents the economic dispatch problem.

For illustrative purposes, we start from a single-bus version of the problem (4-1) by substituting constraint (4-1b) by

$$\sum_{b\in\mathcal{B}} \sum_{i\in\mathcal{I}_b} p_{i,t,n} = \sum_{b\in\mathcal{B}} \widehat{\xi}_{b,t,n}, \tag{4-2}$$

and vanishing the constraint (4-1c).

4.2 Numerical experiments for the Wasserstein-based approach

In this section, first, we analyze the computational efficiency of each decomposition method described in Chapter 3, by assessing the DD-DRO-W unit commitment problem single-bus. Then, given the superiority of the C&CG decomposition method, we benchmark this method with the SAA. For this, we consider the stochastic unit commitment problem with power flow constraints, for a system consisting of 4 buses, 4 transmission lines, 14 thermal generators, and 2 wind farms.

4.2.1 Computational efficiency analysis

For performance comparison purposes, we solve a single-bus system with $I = 5, 14, 54$ thermal generators over a 24-hour operational time. We develop scenarios for the electricity load over 24-hour span time by setting a deterministic profile distribution of the load and discounting wind power generation. The source of data for the wind power generation is the Global Energy Competition (GEFCom) [40,41]. For reproducibility purposes, the system data is presented in Appendix B.

With the scenarios at hand, we construct the Wasserstein ball around the empirical distribution. Although the number of training samples N can be estimated for a given confidence level $\delta > 0$, for illustration purposes in our computational experiments we consider a fixed number of training samples equal to 100 and a fixed parameter δ equal to 3.

The computational experiments were implemented using JuMP [23], a modeling language for mathematical optimization embedded in the Julia programming language. The solver Gurobi 7.5.2 was used as the MIP solver to run the computational experiments on an Intel Core i7, a 4.0-GHz processor with 32 GB of RAM.

For illustrative purposes, we report in Table 4.5 the gap—the difference between the upper and lower bounding approximation—at iteration K and the computing time of each algorithm, for the system with 5 thermal generators. We used $\varepsilon = 0.0$.

The C&CG method is clearly superior. The results show that the oracle sub problem asserts feasibility faster for the current solution of the master problem $(\mathbf{x}^K, \lambda^K)$ of the C&CG method. This is the main advantage of this algorithm, whenever the computational burden of the master problem can be dealt with.

Iteration	**C&CG**	**Benders multi-cut**	**Benders single-cut**
K	$UB - LB$	$UB - LB$	$UB - LB$
1	1273245	1404389	1404388
2	0	147243	282048
3		135559	135559
4		8471	48166
5		8064	19806
6		10656	19806
7		1670	19490
8		1274	16295
9		5548	16295
10		756	16295
11		1278	16295
12		2946	16295
13		760	15564
14		799	15564
15		863	15041
16		799	14310
17		0	13777
40			9142
80			3857
160			2218
165			0
Time (CPUs)	3394	26622	148127

Table 4.5: Data-driven Unit Commitment problem with 5 generators.

Table 4.6 reports the computing time of convergence of each algorithm for the systems with 14 and 54 thermal generators, respectively. Symbol (-) means that there is no computing time to report. For the system with 14 thermal generators, all algorithms converged in moderate computing time, whereas for the system with 54 thermal generators, only the C&CG

algorithm converged, which confirms the superiority of this last method even for these large systems.

Method	**Syst-14** **Time** (CPUs)	**Syst-54** **Time** (CPUs)
C&CG	15365	24225
Bender's multi-cut	27833	-
Bender's single-cut	38379	-

Table 4.6: Computing time.

We also benchmark the proposed solution methodology with the existing solution approaches applied to the class of parametric linear programs with right-hand-side uncertainty. First, we consider the tractable reformulation of the TS-DRO-W problem derived from the convex reduction develop in [14] which scales with the number of dual vertices. We use the algorithm for enumerating vertices embedding in the computational tool Polyhedral.jl of the programming language Julia. However, the enumerating algorithm does not converge in reasonable time for the considered instances.

Given the equivalence of the TS-DRO-W problem with the TS-DRO problem under the Wasserstein ambiguity set which comprises the finite distributions supported on the set $\widehat{\Xi}$, we utilize the algorithmic decomposition method presented in [28] for two-stage distributionally robust mixed binary programs to the address unit commitment instances. Nevertheless, the resulting distribution separation problem could not be solved, because the enumeration of the set $\widehat{\Xi}$ has $N \cdot 3^{d_\xi}$ complexity which is out-of-memory for the considered uncertainty dimensionality.

4.3 SAA Benchmark

Once we have an efficient algorithmic scheme to solve the DD-DRO-W problem – the C&CG algorithm presented in the Chapter 3 –, we can benchmark this approach with the SAA method. For the empirical estimator $\widehat{P}_N$, we obtain the conservative solution by assessing the DD-DRO over the Wasserstein ball centered in $\widehat{P}_N$. Following [1], we consider the smallest Wasserstein ball that contains the actual data-generating distribution with confidence $1-\beta$ for some $\beta \in (0,1)$. The SAA solution is obtained by assessing the deterministic equivalent of this problem. We compare both approaches by an out-of-sample analysis.

Let $\mathbf{x}^N_{wass}$ and $\mathbf{x}^N_{SAA}$, the corresponding optimal solution of the DD-DRO-W and the SAA problem, respectively, for $\widehat{P}_N$. Note that this solution inherit the randomness from the empirical estimator $\widehat{P}_N$. We define the reduction metric

$$r^N = \left(1 - \frac{\mathbf{c}^\top \mathbf{x}^N_{wass} + \mathbb{E}[Q(\mathbf{x}^N_{wass}, \xi)]}{\mathbf{c}^\top \mathbf{x}^N_{SAA} + \mathbb{E}[Q(\mathbf{x}^N_{SAA}, \xi)]}\right) \times 100\% \tag{4-3}$$

by the percentage difference between the out-of-sample cost of the SAA and the conservative DD-DRO-W solution. We reference the probability

$$P^\infty\left(r^N \leq \tau\right) \tag{4-4}$$

as the out-of-sample disappointment of the SAA solution over the conservative solution $\mathbf{x}^N_{wass}$ for the empirical estimator $\widehat{P}_N$. That is, the cumulative distribution function (c.d.f) of the estimator r^N under the sample path distribution P^∞.

We generate realizations of the estimator r^N from the optimal solutions $\mathbf{x}^N_{wass}$ and $\mathbf{x}^N_{SAA}$ of the data-driven unit commitment problem (4-1). These solutions correspond to different realizations of the empirical estimator $\widehat{P}_N$ defined by batches independent and identically distributed (i.i.d), each of which composed of $N = 100$ samples. We construct the Wasserstein ball with confidence 95%. Fig. 4.1 displays the c.d.f (4-4) of the estimator r^N.

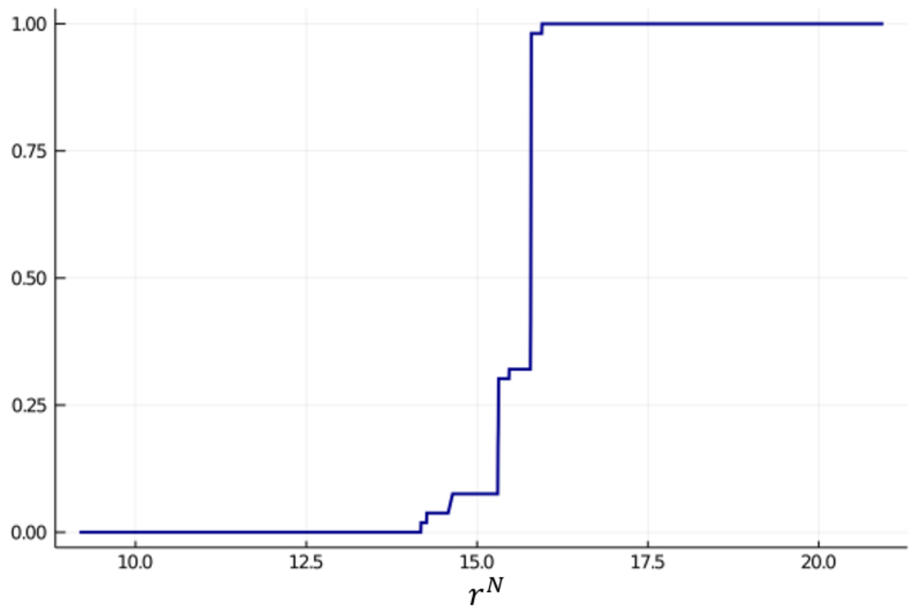

Figure 4.1: Cumulative distribution function of the estimator r^N for the data-driven unit commitment problem.

As shown in Fig. 4.1, the out-of-sample disappointment is high for values between 15% and 17.5% — $P^\infty(15\% \leq r^N \leq 17.5\%) \approx 0.75$. Besides showing the instability of the SAA solution for changes in the data, this result reveals a concern for the power system operator since a poor out-of-sample performance in the order of 15% translates into substantial cost overruns.

Even though the existing methods render the DD-DRO-W problem intractable, we have developed exact decomposition schemes – C&CG mainly – to assess this problem. As we have seen, the obtained conservative solution alleviates the concern of poor out-of-sample performance, in settings where the asymptotic convergence of the SAA method is no guarantee because of computational time limitations.

5
Conclusion

We have developed two conservative solution methodologies for the class of two-stage stochastic linear optimization problems with right-hand-side uncertainty and rectangular support: When the true data-generating probability distribution is known, we propose an exact solution method based on a partition-refinement algorithm of the support and a DRO problem that minimizes the worst-expected cost over all extreme probability distributions with known partition-adapted conditional expectations; When only historical observations of the uncertainty are available, we proposed a DRO problem based on the Wasserstein metric to incorporate ambiguity in the data-generating probability distribution and developed a novel solution method to solve this problem.

5.1
Conclusion of the moment-based approach

Considering that the complexity of the moment-based approach grows exponentially over an uncertainty dimension, for computational tractability, we reformulated the upper-bound problem and proposed algorithmic schemes: (i) for problems with low-dimensional uncertainty, we developed a deterministic equivalent linear programming model, (ii) for medium-sized uncertainty dimensionality, we proposed a column and constraint generation algorithm, and (iii) to handle high dimensional uncertainty, we proposed a simplex-based heuristic method whose complexity grew linearly with the uncertainty dimension.

Out-of-sample computational experiments show that our moment-based method avoid disappointment in comparison to the non-converged sub-optimal optimistic solution given by the lower-bound problem based on Jensen's inequality when the cost of recourse was high. This raises awareness regarding the use of the optimistic solution provided by the partition-based method to solve two-stage stochastic optimization problems when the recourse cost is significantly high. Many practical applications that exhibit this type of recourse cost and depend on the implementation of a non-converged sub-optimal solution because of computational time limitations will benefit from the developed framework to obtain a conservative solution.

5.2 Conclusion of the Wasserstein-based approach

We have presented also a new algorithmic approach to solve distributionally robust optimization problems with right-hand-side uncertainty over Wasserstein balls. Our model assumes distributions supported on an rectangular set within a Wasserstein ball, which is built around the empirical distribution of observed data of the uncertainty. The proposed approach and formulation allow us solving the extensive form of a convex reformulation of the problem very efficiently.

This issue is of paramount importance because of the existing algorithmic schemes are computationally intractable for instances with high uncertain dimensionality or an exponential number of dual vertices. Instead, we have proposed a finite extensive equivalent form of the problem which is solved by using an exact decomposition algorithm that converges in a finite number of iterations.

We tested the proposed algorithms with a two-stage unit commitment problem for the day-ahead scheduling of power generation by assuming an uncertain energy load. We analyzed the computational performance of each algorithm by varying the size of the considered systems. In particular, there are no tractable formulations in the literature for these instances, which demonstrates that our proposed approach is a substantial advance to address the class of DRO problems with right-hand side uncertainty.

The results show that when the second-stage problem has a complex polyhedral structure, the C&CG has the best computational performance among the three methods that were tested. The superiority of C&CG stems from the fact that it optimizes the actual value of the second-stage function for each uncertainty realization, unlike the approaches based on Benders methods that optimize only a piecewise linear approximation of this function.

References

[1] P. Mohajerin Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, no. 1, pp. 115–166, Sep 2018. [Online]. Available: https://doi.org/10.1007/s10107-017-1172-1

[2] C. C. Huang, W. T. Ziemba, and A. Ben-Tal, "Bounds on the expectation of a convex function of a random variable: With applications to stochastic programming," *Operations Research*, vol. 25, no. 2, pp. 315–325, 1977.

[3] J. R. Birge and R. J. West, "Designing approximation schemes for stochastic optimization problems, in particular for stochastic programs with recourse," *Stochastic Programming 84 Part I*, pp. 54–102, 1986.

[4] K. Frauendorfer, "Solving slp recourse problems with arbitrary multivariate distributions: The dependent case," *Mathematics of Operations Research*, vol. 13, no. 3, pp. 377–394, 1988. [Online]. Available: http://www.jstor.org/stable/3690018

[5] N. Edirisinghe and W. T. Ziemba, "Implementing bounds-based approximations in convex-concave two-stage stochastic programming," *Mathematical Programming*, vol. 75, no. 2, pp. 295–325, 1996.

[6] N. C. P. Edirisinghe and W. T. Ziemba, "Bounds for two-stage stochastic programs with fixed recourse," *Mathematics of Operations Research*, vol. 19, no. 2, pp. 292–313, 1994.

[7] N. C. P. Edirisinghe and G.-M. You, "Second-order scenario approximation and refinement in optimization under uncertainty," *Annals of Operations Research*, vol. 64, no. 1, pp. 143–178, Dec 1996. [Online]. Available: https://doi.org/10.1007/BF02187644

[8] T. Homem-de-Mello and G. Bayraksan, "Monte Carlo sampling-based methods for stochastic optimization," *Surveys in Operations Research and Management Science*, vol. 19, pp. 56–85, 2014.

[9] A. Kleywegt, A. Shapiro, and T. Homem-de-Mello, "The sample average approximation method for stochastic discrete optimization," *SIAM Journal on Optimization*, vol. 12, no. 2, pp. 479–502, 2001.

[10] D. Bertsimas, V. Gupta, and N. Kallus, "Robust sample average approximation," *Mathematical Programming*, vol. 171, no. 1-2, pp. 217–282, 2018.

[11] H. Gassmann and W. T. Ziemba, *A tight upper bound for the expectation of a convex function of a multivariate random variable*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1986, pp. 39–53. [Online]. Available: https://doi.org/10.1007/BFb0121113

[12] K. Frauendorfer and P. Kall, "A solution method for slp recourse problems with arbitrary multivariate distributions - the independent case," *Problems of Control and Information Theory*, vol. 17, 01 1988.

[13] B. Zeng and L. Zhao, "Solving two-stage robust optimization problems using a column-and-constraint generation method," *Operations Research Letters*, vol. 41, no. 5, pp. 457 – 461, 2013.

[14] P. Mohajerin Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, no. 1, pp. 115–166, Sep 2018. [Online]. Available: https://doi.org/10.1007/s10107-017-1172-1

[15] P. Kall and D. Stoyan, "Solving stochastic programming problems with recourse including error bounds," *Mathematische Operationsforschung und Statistik. Series Optimization*, vol. 13, no. 3, pp. 431–447, 1982. [Online]. Available: https://doi.org/10.1080/02331938208842805

[16] A. Madansky, "Bounds on the expectation of a convex function of a multivariate random variable," *Ann. Math. Statist.*, vol. 30, no. 3, pp. 743–746, 09 1959. [Online]. Available: https://doi.org/10.1214/aoms/1177706203

[17] W. Wiesemann, D. Kuhn, and M. Sim, "Distributionally robust convex optimization," *Operations Research*, vol. 62, no. 6, pp. 1358–1376, 2014. [Online]. Available: https://doi.org/10.1287/opre.2014.1314

[18] J. Birge and Louveaux, *Introduction to Stochastic Programming*, 2nd ed. New York Dordrecht Heidelberg London: Springer, 2011.

[19] A. Street, F. Oliveira, and J. M. Arroyo, "Contingency-constrained unit commitment with $n-k$ security criterion: A robust optimization approach," *IEEE Transactions on Power Systems*, vol. 26, no. 3, pp. 1581–1590, Aug 2011.

[20] D. Bertsimas and M. Sim, "The price of robustness," *Operations Research*, vol. 52, no. 1, pp. 35–53, Jannuary-February 2004. [Online]. Available: http://dx.doi.org/10.1287/opre.1030.0065

[21] P. Kall, A. Ruszczyński, and K. Frauendorfer., *Approximation techniques in stochastic programming*. Berlin: Springer-Verlag, 1988, pp. 33–64. [Online]. Available: https://doi.org/10.1007/BFb0121113

[22] G. B. Dantzig, *Linear Programming and Extensions*. Princeton University Press, 1991. [Online]. Available: http://www.jstor.org/stable/j.ctt1cx3tvg

[23] I. Dunning, J. Huchette, and M. Lubin, "Jump: A modeling language for mathematical optimization," *SIAM Review*, vol. 59, no. 2, pp. 295–320, 2017. [Online]. Available: https://doi.org/10.1137/15M1020575

[24] D. Bertsimas, V. Gupta, and N. Kallus, "Robust sample average approximation," *Mathematical Programming*, vol. 171, no. 1, pp. 217–282, Sep 2018. [Online]. Available: https://doi.org/10.1007/s10107-017-1174-z

[25] H. Rahimian and S. Mehrotra, "Distributionally robust optimization: A review," 2019, arXiv preprint arXiv:1908.05659.

[26] D. Wozabal, "A framework for optimization under ambiguity," *Annals of Operations Research*, vol. 193, no. 1, pp. 21 – 47, 2012. [Online]. Available: https://doi.org/10.1007/s10479-010-0812-0

[27] C. Zhao, "Data-driven risk-averse stochastic program and renewable energy integration," Ph.D. dissertation, University of Florida, 2014.

[28] M. Bansal, K.-L. Huang, and S. Mehrotra, "Decomposition algorithms for two-stage distributionally robust mixed binary programs," *SIAM Journal on Optimization*, vol. 28, no. 3, pp. 2360–2383, 2018. [Online]. Available: https://doi.org/10.1137/17M1115046

[29] G. A. Hanasusanto and D. Kuhn, "Conic programming reformulations of two-stage distributionally robust linear programs over wasserstein

balls," *Operations Research*, vol. 66, no. 3, pp. 849–869, 2018. [Online]. Available: https://doi.org/10.1287/opre.2017.1698

[30] D. Duque, S. Mehrotra, and D. P. Morton, "Distributionally robust two-stage stochastic programming," *Optimization online*, 2020.

[31] T. S. Motzkin, H. Raiffa, G. L. Thompson, and R. M. Thrall, "The double description method," *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 51–73, 1953.

[32] K. Fukuda and A. Prodon, "Double description method revisited," in *Franco-Japanese and Franco-Chinese Conference on Combinatorics and Computer Science*. Springer, 1995, pp. 91–111.

[33] L. Kantorovich and G. Rubinshtein, "On a space of totally additive functions, vestn," *Vestn Lening. Univ*, vol. 13, pp. 52–59, 1958.

[34] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh, *Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning*. INFORMS TutORials in Operations Research, 2019, ch. 6, pp. 130–166. [Online]. Available: https://pubsonline.informs.org/doi/abs/10.1287/educ.2019.0198

[35] C. Zhao and Y. Guan, "Data-driven risk-averse stochastic optimization with wasserstein metric," *Operations Research Letters*, vol. 46, no. 2, pp. 262 – 267, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167637718300506

[36] E. Guevara, F. Babonneau, T. Homem-de-Mello, and S. Moret, "A machine learning and distributionally robust optimization framework for strategic energy planning under uncertainty," *Applied Energy*, vol. 271, p. 115005, 2020.

[37] J. Fortuny-Amat and B. McCarl, "A representation and economic interpretation of a two-level programming problem," *Journal of the operational Research Society*, vol. 32, no. 9, pp. 783–792, 1981.

[38] R. Zhu, H. Wei, and X. Bai, "Wasserstein metric based distributionally robust approximate framework for unit commitment," *IEEE Transactions on Power Systems*, vol. 34, no. 4, pp. 2991–3001, 2019.

[39] A. Papavasiliou and S. S. Oren, "Multiarea stochastic unit commitment for high wind penetration in a transmission constrained network," *Operations Research*, vol. 61, no. 3, pp. 578–592, 2013.

[40] T. Hong, P. Pinson, and S. Fan, "Global energy forecasting competition 2012," *International Journal of Forecasting*, vol. 30, no. 2, pp. 357 – 363, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169207013000745

[41] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, "Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond," *International Journal of Forecasting*, vol. 32, no. 3, pp. 896 – 913, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169207016000133

A
Appendix

A.1
Proof of the Theorem

If $\{f_n\}_{n=1}^{\infty}$ is a sequence of functions pointwise converging to f, i.e., if $f(\mathbf{x}) = \lim_{n\longrightarrow\infty} f(\mathbf{x})$ for all $\mathbf{x} \in X$, then f_n epi-converges to f, if the sequence $\{f_n\}_{n=1}^{\infty}$ is monotone increasing or monotone decreasing [3] and f is continuous.

Epi-convergence is a kind of convergence very useful for approximate minimization problems in the following sense:
Suppose a sequence of function $\{f_n\}_{n=1}^{\infty}$ epi-converges to f. Let $z^* := \min_{\mathbf{x}\in X} f(\mathbf{x})$, $z_n^* := \min_{\mathbf{x}\in X} f_n(\mathbf{x})$, and $\mathbf{x}^* \in \arg\min_{\mathbf{x}\in X} f(\mathbf{x})$, $\mathbf{x}_n^* \in \arg\min_{\mathbf{x}\in X} f_n(\mathbf{x})$ the correspond minimizers, respectively. Then, under some assumptions, we can ensure that

$$\lim_{n\longrightarrow\infty} z_n^* = z^*, \quad \textit{and} \quad \lim_{n\longrightarrow\infty} \mathbf{x}_n^* = \mathbf{x}^*.$$

Before prove the **Theorem 1**, we will present the following lema that will be necessary to do the proof.

Lema 1. *let $G = \{(\boldsymbol{\xi}, \eta) \in \mathbb{R}^{d_\xi+1} : \boldsymbol{\xi} \in \Xi, \eta = Q(\mathbf{x}, \boldsymbol{\xi})\}$ be the graph of the function $Q(\mathbf{x}, \cdot)$ and let $co(G)$ be its convex hull. Then*

$$sup\{\eta \in \mathbb{R} : (\bar{\boldsymbol{\xi}}, \eta) \in co(G)\} = \max_{\delta\in\mathcal{D}(\bar{\boldsymbol{\xi}})} \mathbb{E}^{\delta}[Q(\mathbf{x}, \mathbf{e})] \tag{A-1}$$

***Proof of Lema* 1.** *It is clear that*

$$\max_{\delta\in\mathcal{D}(\bar{\boldsymbol{\xi}})} \mathbb{E}^{\delta}\left[Q(\mathbf{x}, \mathbf{e})\right] \le sup\{\eta \in \mathbb{R} : (\bar{\boldsymbol{\xi}}, \eta) \in co(G)\}.$$

On the other hand, let $(\bar{\boldsymbol{\xi}}, \eta) \in co(G)$. Then there exist S, $\boldsymbol{\xi}^s \in \Xi$, and probabilities $\mathbb{P}(\boldsymbol{\xi} = \boldsymbol{\xi}^s) \ge 0$, $s = 1, \ldots, S$ (parameters of the convex combination), such that $\sum_{s=1}^{S} \mathbb{P}(\boldsymbol{\xi} = \boldsymbol{\xi}^s)\boldsymbol{\xi}^s = \bar{\boldsymbol{\xi}}$, $\sum_{s=1}^{S} \mathbb{P}(\boldsymbol{\xi} = \boldsymbol{\xi}^s) Q(\mathbf{x}, \boldsymbol{\xi}^s) = \eta$. Now for every s, there exist conditional probabilities $\mathbb{P}(\mathbf{e} = \mathbf{e}_j | \boldsymbol{\xi} = \boldsymbol{\xi}^s) \ge 0$, $j = 1, \ldots, 2^{d_\xi}$ such that $\sum_{j=1}^{2^{d_\xi}} \mathbb{P}(\mathbf{e} = \mathbf{e}_j | \boldsymbol{\xi} = \boldsymbol{\xi}^s)\mathbf{e}_j = \boldsymbol{\xi}^s$, i.e., $\hat{\boldsymbol{\xi}} = \sum_s \sum_j \mathbb{P}(\boldsymbol{\xi} = \boldsymbol{\xi}^s)\mathbb{P}(\mathbf{e} = \mathbf{e}_j | \boldsymbol{\xi} = \boldsymbol{\xi}^s)\mathbf{e}_j$. Since $Q(\mathbf{x}, \cdot)$ is convex

$$\begin{aligned}\eta &= \sum_{s=1}^{S} \mathbb{P}(\boldsymbol{\xi} = \boldsymbol{\xi}^s) Q(\mathbf{x}, \boldsymbol{\xi}^s) \\ &= \sum_{s} \mathbb{P}(\boldsymbol{\xi} = \boldsymbol{\xi}^s) Q\left(\sum_{j} \mathbb{P}(\mathbf{e} = \mathbf{e}_j | \boldsymbol{\xi} = \boldsymbol{\xi}^s) \mathbf{e}_j \right) \\ &\leq \sum_{s} \sum_{j} \mathbb{P}(\boldsymbol{\xi} = \boldsymbol{\xi}^s) \mathbb{P}(\mathbf{e} = \mathbf{e}_j | \boldsymbol{\xi} = \boldsymbol{\xi}^s) Q(\mathbf{x}, \mathbf{e}_j).\end{aligned}$$

Setting $\delta_j = \sum_s \mathbb{P}(\boldsymbol{\xi} = \boldsymbol{\xi}^s)\mathbb{P}(\mathbf{e} = \mathbf{e}_j | \boldsymbol{\xi} = \boldsymbol{\xi}^s)$ *yields* $\bar{\boldsymbol{\xi}} = \sum_j \delta_j \mathbf{e}_j$, $\sum_j \delta_j = 1$, $\delta_j \geq 0$, *and*

$$\eta \leq \sum_j \delta_j Q(\mathbf{x}, \mathbf{e}_j) \leq \max_{\delta \in \mathcal{D}(\bar{\boldsymbol{\xi}})} \mathbb{E}^{\delta} \left[Q(\mathbf{x}, \mathbf{e}) \right].$$

Therefore

$$sup\{\eta \in \mathbb{R} \,:\, (\bar{\boldsymbol{\xi}}, \eta) \in co(G)\} \leq \max_{\delta \in \mathcal{D}(\bar{\boldsymbol{\xi}})} \mathbb{E}^{\delta} \left[Q(\mathbf{x}, \mathbf{e}) \right].$$

Making use of the properties of epi-convergent functions and the lema above, we will prove the **Theorem 1**.

***Proof of Theorem* 1.**

1.1 *Let* $\Xi' \subseteq \Xi$, *let* G' *be the graph of the function* $Q(\mathbf{x}, \boldsymbol{\xi})$ *with* $\boldsymbol{\xi} \in \Xi'$ *and let* $\bar{\boldsymbol{\xi}}' = \mathbb{E}[\boldsymbol{\xi} | \boldsymbol{\xi} \in \Xi']$ *be the conditional mean, then*

$$sup\{\eta \,:\, (\bar{\boldsymbol{\xi}}, \eta) \in co(G))\} \geq sup\{\eta \,:\, (\bar{\boldsymbol{\xi}}', \eta) \in co(G'))\}$$

which implies

$$\max_{\delta \in \mathcal{D}(\bar{\boldsymbol{\xi}})} \mathbb{E}^{\delta}[Q(\mathbf{x}, \mathbf{e})] \geq \max_{\delta \in \mathcal{D}(\bar{\boldsymbol{\xi}}')} \mathbb{E}^{\delta}[Q(\mathbf{x}, \mathbf{e}')],$$

where $\mathbf{e}'$ *is the random variable with support on the vertex of the cell* Ξ'.

Suppose that $\mathscr{P}_{n+1}$ *refines* $\mathscr{P}_n$, *then there exist* $\Xi^k \in \mathscr{P}_n$ *such that* $\Xi^k = \Xi^{k'} \cup \Xi^{k''}$ *with* $\Xi^{k'}$, $\Xi^{k''} \in \mathscr{P}_{n+1}$.

Let $\mathbf{e}^k$, $\mathbf{e}^{k'}$, *and* $\mathbf{e}^{k''}$ *be the random variables with support in the set of vertex of the cells* Ξ^k, $\Xi^{k'}$ *and* $\Xi^{k''}$, *respectively, and let* $p^{k'} = \mathbb{P}(\boldsymbol{\xi} \in \Xi^{k'})$ *and* $p^{k''} = \mathbb{P}(\boldsymbol{\xi} \in \Xi^{k''})$. *By the above we have that*

$$p^{k'} \left(\max_{\delta \in \mathcal{D}(\bar{\boldsymbol{\xi}}^k)} \mathbb{E}^{\delta} \left[Q(\mathbf{x}, \mathbf{e}^k) \right] \right) \geq p^{k'} \left(\max_{\delta \in \mathcal{D}(\bar{\boldsymbol{\xi}}^{k'})} \mathbb{E}^{\delta} \left[Q(\mathbf{x}, \mathbf{e}^{k'}) \right] \right),$$

and

$$p^{k''}\left(\max_{\delta\in\mathcal{D}(\bar{\boldsymbol{\xi}}^{k})}\mathbb{E}^{\delta}\left[Q(\mathbf{x},\mathbf{e}^{k})\right]\right)\geq p^{k''}\left(\max_{\delta\in\mathcal{D}(\bar{\boldsymbol{\xi}}^{k''})}\mathbb{E}^{\delta}\left[Q(\mathbf{x},\mathbf{e}^{k''})\right]\right),$$

then,

$$\begin{aligned}\underbrace{(p^{k'}+p^{k''})}_{p^{k}}\left(\max_{\delta\in\mathcal{D}(\bar{\boldsymbol{\xi}}^{k})}\mathbb{E}^{\delta}\left[Q(\mathbf{x},\mathbf{e}^{k})\right]\right)\geq & p^{k'}\left(\max_{\delta\in\mathcal{D}(\bar{\boldsymbol{\xi}}^{k'})}\mathbb{E}^{\delta}\left[Q(\mathbf{x},\mathbf{e}^{k'})\right]\right)\\ &+p^{k''}\left(\max_{\delta\in\mathcal{D}(\bar{\boldsymbol{\xi}}^{k''})}\mathbb{E}^{\delta}\left[Q(\mathbf{x},\mathbf{e}^{k''})\right]\right).\end{aligned}$$

Since the other cells in $\mathscr{P}_n$ and $\mathscr{P}_{n+1}$ are the same, it follows that

$$\sum_{k=1}^{n}p^{k}\left(\max_{\delta\in\mathcal{D}(\bar{\boldsymbol{\xi}}^{k})}\mathbb{E}^{\delta}\left[Q(\mathbf{x},\mathbf{e}^{k})\right]\right)\geq\sum_{k=1}^{n+1}p^{k}\left(\max_{\delta\in\mathcal{D}(\bar{\boldsymbol{\xi}}^{k})}\mathbb{E}^{\delta}\left[Q(\mathbf{x},\mathbf{e}^{k})\right]\right),$$

which implies $f_n^U(\mathbf{x})\geq f_{n+1}^U(\mathbf{x})$ for all $\mathbf{x}\in X$, therefore

$$z_n^U=\min_{\mathbf{x}\in X}f_n^U(\mathbf{x})\geq\min_{\mathbf{x}\in X}f_{n+1}^U(\mathbf{x})=z_{n+1}^U.$$

1.2 *It is true that*

$$\bar{\boldsymbol{\xi}}^{k}=\frac{p^{k'}\bar{\boldsymbol{\xi}}^{k'}+p^{k''}\bar{\boldsymbol{\xi}}^{k''}}{p^{k}}$$

By convexity it holds that

$$Q(\mathbf{x},\bar{\boldsymbol{\xi}}^{k})\leq\frac{p^{k'}}{p^{k}}Q(\mathbf{x},\bar{\boldsymbol{\xi}}^{k'})+\frac{p^{k''}}{p^{k}}Q(\mathbf{x},\bar{\boldsymbol{\xi}}^{k''}).$$

Since the others cells in $\mathscr{P}_n$ and $\mathscr{P}_{n+1}$ are the same, it follows that

$$\sum_{k=1}^{n}p^{k}Q(\mathbf{x},\bar{\boldsymbol{\xi}}^{k})\leq\sum_{k=1}^{n+1}p^{k}Q(\mathbf{x},\bar{\boldsymbol{\xi}}^{k})$$

which implies $f_n^L(\mathbf{x})\leq f_{n+1}^L(\mathbf{x})$ for all $\mathbf{x}\in X$, therefore

$$z_n^L=\min_{\mathbf{x}\in X}f_n^L(\mathbf{x})\leq\min_{\mathbf{x}\in X}f_{n+1}^L(\mathbf{x})=z_{n+1}^L.$$

1.3 *We have that*

$$\max_{\delta\geq 0}\Big\{\sum_j \delta_j Q(\mathbf{x},\mathbf{e}_j^k)\Big|\sum_j \delta_j = 1, \sum_j \delta_j \mathbf{e}_j = \bar{\boldsymbol{\xi}}^k\Big\} \leq \max_{\delta\geq 0}\Big\{\sum_j \delta_j Q(\mathbf{x},\mathbf{e}_j^k)\Big|\sum_j \delta_j = 1\Big\}.$$

Let $M^k := \max_{\boldsymbol{\xi}\in\Xi^k} Q(\mathbf{x},\boldsymbol{\xi}) = \max_{\delta\geq 0}\Big\{\sum_j \delta_j Q(\mathbf{x},\mathbf{e}_j^k)\Big|\sum_j \delta_j = 1\Big\}$. *By convexity* $\max_{\boldsymbol{\xi}\in\Xi^k} Q(\mathbf{x},\boldsymbol{\xi}) = Q(\mathbf{x},\mathbf{e}_j^k)$ *for some j.*

By the existence of $\mathbb{E}[Q(\mathbf{x},\boldsymbol{\xi})]$, *it holds that*

$$\sum_{k=1}^{n} p^k M^k \longrightarrow \mathbb{E}[Q(\mathbf{x},\boldsymbol{\xi})] \quad \text{as } n \longrightarrow \infty,$$

therefore

$$\sum_{k=1}^{n} p^k \max_{\delta\in\mathcal{D}(\bar{\boldsymbol{\xi}}^k)} \mathbb{E}^{\delta}\left[Q(\mathbf{x},\mathbf{e}^k)\right] \longrightarrow \mathbb{E}[Q(\mathbf{x},\boldsymbol{\xi})] \quad \text{as } n \longrightarrow \infty,$$

which implies $f_n^U(\mathbf{x}) \longrightarrow f(\mathbf{x})$ *as* $n \longrightarrow \infty$ *for all* $\mathbf{x} \in X$. *Since the sequence* $\{f_n^U\}_{n=1}^{\infty}$ *is non-increasing and the objective function* f *is continuous, we have that* f_n^U *epi-converges to* f. *So, if* $\mathbf{x}^* = \lim_{n\longrightarrow\infty} \mathbf{x}_n^U$ *then* $\mathbf{x}^* \in \arg\min_{\mathbf{x}\in X} f(\mathbf{x})$. *Therefore*

$$\lim_{n\longrightarrow\infty} z_n^U = \lim_{n\longrightarrow\infty} f_n^U(\mathbf{x}_n^U) = f(\mathbf{x}^*) = z^*.$$

On the another hand, we have that

$$m^k := \min_{\xi\in\Xi^k} Q(\mathbf{x},\boldsymbol{\xi}) \leq Q(\mathbf{x},\bar{\boldsymbol{\xi}}^k).$$

By the existence of $\mathbb{E}[Q(\mathbf{x},\boldsymbol{\xi})]$ *it holds that*

$$\sum_{k=1}^{n} p^k m^k \longrightarrow \mathbb{E}[Q(\mathbf{x},\boldsymbol{\xi})] \quad \text{as } n \longrightarrow \infty$$

so

$$\sum_{k=1}^{n} p^k Q(\mathbf{x},\bar{\boldsymbol{\xi}}^k) \longrightarrow \mathbb{E}[Q(\mathbf{x},\boldsymbol{\xi})] \quad \text{for all } \mathbf{x} \in X,$$

which implies $f_n^L(\mathbf{x}) \longrightarrow f(\mathbf{x})$ *for all* $\mathbf{x} \in X$. *Since the sequence* $\{f_n^L\}_{n=1}^{\infty}$ *is no-decreasing and the objective function* f *is continuous, we have that*

f_n^L epi-converges to f. So, if $\mathbf{x}^* = \underset{n\longrightarrow\infty}{lim} \mathbf{x}_n^L$ then $\mathbf{x}^* \in \arg\min_{\mathbf{x}\in X} f(\mathbf{x})$. Therefore

$$\underset{n\longrightarrow\infty}{lim} f_n^L(\mathbf{x}_n^L) = f(\mathbf{x}^*) = z^*.$$

A.2 Oracle MILP problem for the moment-based approach

Let $(\mathbf{x}_m^*, \boldsymbol{\eta}_m^* \pi_m^*)$ be the optimal solution of the master problem (2-16) at iteration m. By assuming $\Xi = \times_{i=1}^{d_\xi}[a_i, b_i]$ and defining $\mathbf{a} = [a_1, \ldots, a_{d_\xi}]^\top$ and $\mathbf{b} = [b_1, \ldots, b_{d_\xi}]^\top$, any extreme point $\mathbf{e}_j^k$, $j = 1, \ldots, 2^{d_\xi}$, of cell $k = 1, \ldots, n$ can be expressed by $\text{diag}(\mathbf{a})\mathbf{z} + \text{diag}(\mathbf{b})(\mathbf{1} - \mathbf{z})$, where $\mathbf{z} \in \{0,1\}^{d_\xi}$ is a vector binary variable. Then, problem (2-17) is equivalent to:

$$\begin{cases} \max\limits_{\boldsymbol{\theta},\mathbf{z}} & \boldsymbol{\theta}^\top \mathbf{r}(\mathbf{x}_m^*) + [\mathbf{H}^\top(\mathbf{x}_m^*)\boldsymbol{\theta} - \boldsymbol{\eta}_m^*]^\top[\text{diag}(\mathbf{a})\mathbf{z} + \text{diag}(\mathbf{b})(\mathbf{1} - \mathbf{z})] - \pi^* \\ \text{s.t.} & \mathbf{W}^\top\boldsymbol{\theta} \le \mathbf{q} \end{cases}$$
$$= \begin{cases} \max\limits_{\boldsymbol{\theta},\mathbf{z}} & \left\{ \begin{array}{l} \sum\limits_{i=1}^{m_y} \theta_i r_i(\mathbf{x}_m^*) + \sum\limits_{j=1}^{d_\xi}\sum\limits_{i=1}^{m_y} H_{j,i}(\mathbf{x}_m^*)\underbrace{\theta_i z_j}_{w_i^j} a_j - \sum\limits_{j=1}^{d_\xi} a_j \eta_j^* z_j \\ \sum\limits_{j=1}^{d_\xi}\sum\limits_{i=1}^{m_y} H_{j,i}(\mathbf{x}_m^*)(\theta_i b_j - \underbrace{\theta_i z_j}_{w_i^j} b_j) - \sum\limits_{j=1}^{d_\xi} b_j \eta_j^*(1 - z_j) - \pi^* \end{array} \right\} \\ \text{s.t.} & \sum\limits_{i=1}^{m_y} W_{ji}\theta_i \le q_j, \quad j = 1, \ldots, d_y. \end{cases} \tag{A-2}$$

By introducing the auxiliary variable $\mathbf{w} = [w_1, \ldots, d_\xi]^\top$ where $w_i^j = \theta_i z_j$, $i = 1, \ldots, m_y$, $j = 1, \ldots, d_\xi$, problem (A-2) is equivalent to the following MILP problem:

$$
\begin{aligned}
\max_{\boldsymbol{\theta},\mathbf{z}} \quad & \left\{ \begin{array}{l} \sum_{i=1}^{m_y} \theta_i r_i(\mathbf{x}_m^*) + \sum_{j=1}^{d_\xi} \sum_{i=1}^{m_y} H_{j,i}(\mathbf{x}_m^*) w_i^j a_j - \sum_{j=1}^{d_\xi} a_j \eta_j^* z_j \\ \sum_{j=1}^{d_\xi} \sum_{i=1}^{m_y} H_{j,i}(\mathbf{x}_m^*)(\theta_i b_j - w_i^j b_j) - \sum_{j=1}^{d_\xi} b_j \eta_j^* (1 - z_j) - \pi^* \end{array} \right\} \\
\text{s.t.} \quad & \sum_{i=1}^{m_y} W_{ji}\theta_i \leq q_j, \quad j = 1, \ldots, d_y, \\
& |w_i^j - \theta_i| \leq (1 - z_j)M \quad j = 1, \ldots, d_\xi, \quad i = 1, \ldots, m_y, \\
& |w_i^j| \leq M z_j, \quad j = 1, \ldots, d_\xi, \quad i = 1, \ldots, m_y, \\
& z_j \in \{0, 1\}, \quad j = 1, \ldots, d_\xi,
\end{aligned} \tag{A-3}
$$

Where $M \in \mathbb{R}$ is sufficiently large and the products $w_i^j = \theta_i z_j$, $i = 1, \ldots, m_y$, $j = 1, \ldots, d_\xi$ are linearized.

A.3 Proof of the Proposition

***Proof of Proposition* 1.** *By duality we have that*

$$
\max_{\delta \in \mathcal{R}(\bar{\boldsymbol{\xi}}^k)} \mathbb{E}^\delta \left[Q(\mathbf{x}, \mathbf{v}^k) \right] = \max_{\delta^k \geq 0} \left\{ \sum_{j=1}^{d_\xi + 1} \delta_j^k Q(\mathbf{x}, \mathbf{v}_j^k) \; : \; \sum_{j=1}^{d_\xi + 1} \delta_j^k \mathbf{v}_j^k = \bar{\boldsymbol{\xi}}^k, \sum_{j=1}^{d_\xi + 1} \delta_j^k = 1 \right\}.
$$

Since the recourse function $Q(\mathbf{x}, \xi_i)$ is monotonic for all $i = 1, \ldots, d_\xi$ we have that

$$
Q(\mathbf{x}, \mathbf{v}_j^k) \leq Q(\mathbf{x}, \hat{\boldsymbol{\xi}}^k), \quad \forall j = 1, \ldots, (d_\xi + 1),
$$

since $\mathbf{v}_j^k$ only differs from $\hat{\boldsymbol{\xi}}^k$ in just one component. So, $\sum_{j=1}^{d_\xi+1} \delta_j^k Q(\mathbf{x}, \mathbf{v}_j^k) \leq Q(\mathbf{x}, \hat{\boldsymbol{\xi}}^k)$ for all $\mathbf{x} \in X$. By other hand, by the existence of $\mathbb{E}[Q(\mathbf{x}, \boldsymbol{\xi})]$, it holds that

$$
\sum_{k=1}^{n} p^k Q(\mathbf{x}, \hat{\boldsymbol{\xi}}^k) \longrightarrow \mathbb{E}[Q(\mathbf{x}, \boldsymbol{\xi})] \quad \textit{as } n \longrightarrow \infty,
$$

therefore

$$
\sum_{k=1}^{n} p^k \sum_{j=1}^{d_\xi + 1} \delta_j^k Q(\mathbf{x}, \mathbf{v}_j^k) \longrightarrow \mathbb{E}[Q(\mathbf{x}, \boldsymbol{\xi})] \quad \textit{as } n \longrightarrow \infty, \quad \forall \mathbf{x} \in X.
$$

Since $\tilde{f}_n(\mathbf{x}) := \mathbf{c}^\top \mathbf{x} + \sum_{k=1}^{n} p^k \sum_{j=1}^{d_\xi+1} \delta_j^k Q(\mathbf{x}, \mathbf{v}_j^k)$ is a sequence of decreasing *and continuous functions pointwise converging to* $f(\mathbf{x}) = \mathbf{c}^\top \mathbf{x} + \mathbb{E}[Q(\mathbf{x}, \boldsymbol{\xi})]$*, then* $\tilde{f}_n(\mathbf{x})$ *epi-converge to* $f(\mathbf{x})$*. Therefore, by epi-convergence,* $\{\tilde{z}_n^U\}_{n=1}^{\infty}$ *and* $\{\tilde{\mathbf{x}}_n^U\}_{n=1}^{\infty}$ *converge to the optimal value and optimal solution of* (2-1)*, respectively.*

A.4
Oracle MILP problem for the Wasserstein-based approach

Since the optimal solution $\boldsymbol{\xi}^*$ of problem (3-7) is a vertex of the hypercube Ξ or the median $\widehat{\boldsymbol{\xi}}_n$, the decision variable $\boldsymbol{\alpha}$ can be expressed by $\boldsymbol{\alpha} = \text{diag}(\boldsymbol{\Delta}^+)\mathbf{z}^+ + \text{diag}(\boldsymbol{\Delta}^-)\mathbf{z}^-$. For a given $(\mathbf{x}_m^*, \lambda_m^*)$, by introducing auxiliary variables $\overline{\mathbf{w}}^j = z_j^+ \boldsymbol{\theta}$ and $\underline{\mathbf{w}}^j = z_j^- \boldsymbol{\theta}$, the objective function of problem (3-7) is equivalent to

$$
\begin{aligned}
&\boldsymbol{\theta}^\top \mathbf{r}(\mathbf{x}_m^*) + [\mathbf{H}^\top(\mathbf{x}_m^*)\boldsymbol{\theta}]^\top [\overline{\boldsymbol{\xi}} + \text{diag}(\boldsymbol{\Delta}^+)\mathbf{z}^+ - \text{diag}(\boldsymbol{\Delta}^-)\mathbf{z}^-] \\
&- \lambda_m^* \mathbf{1}^\top \boldsymbol{\alpha} \\
= &\sum_{i=1}^{m_y} \theta_i r_i(\mathbf{x}_m^*) + \sum_{j=1}^{d_\xi} \sum_{i=1}^{m_y} H_{j,i}(\mathbf{x}_m^*) \theta_i \overline{\xi}_j + \sum_{j=1}^{d_\xi} \sum_{i=1}^{m_y} H_{j,i}(\mathbf{x}_m^*) \underbrace{\theta_i z_j^+}_{\overline{w}_i^j} \Delta_j^+ \\
&- \sum_{j=1}^{d_\xi} \sum_{i=1}^{m_y} H_{j,i}(\mathbf{x}_m^*) \underbrace{\theta_i z_j^-}_{\underline{w}_i^j} \Delta_j^- - \lambda_m^* \sum_{j=1}^{d_\xi} \alpha_j \\
= &\sum_{i=1}^{m_y} \theta_i r_i(\mathbf{x}_m^*) + \sum_{j=1}^{d_\xi} \sum_{i=1}^{m_y} H_{j,i}(\mathbf{x}_m^*) \theta_i \overline{\xi}_j + \sum_{j=1}^{d_\xi} \sum_{i=1}^{m_y} H_{j,i}(\mathbf{x}_m^*) \overline{w}_i^j \Delta_j^+ \\
&- \sum_{j=1}^{d_\xi} \sum_{i=1}^{m_y} H_{j,i}(\mathbf{x}_m^*) \underline{w}_i^j \Delta_j^- - \lambda_m^* \sum_{j=1}^{d_\xi} \alpha_j.
\end{aligned}
$$

Then, the oracle $f^k(\mathbf{x}_m^*, \lambda_m^*)$ has a MILP equivalent given by:

$$
\begin{array}{ll}
\max\limits_{\boldsymbol{\theta},\mathbf{z}^+,\overline{\mathbf{w}}^j,\mathbf{z}^-,\underline{\mathbf{w}}^j,\boldsymbol{\alpha}} & \left\{ \begin{array}{l} \sum\limits_{i=1}^{m_y} \theta_i r_i(\mathbf{x}_m^*) + \sum\limits_{j=1}^{d_\xi}\sum\limits_{i=1}^{m_y} H_{j,i}(\mathbf{x}_m^*)\theta_i\overline{\xi}_j + \sum\limits_{j=1}^{d_\xi}\sum\limits_{i=1}^{m_y} H_{j,i}(\mathbf{x}_m^*)\overline{w}_i^j\Delta_j^+ \\ - \sum\limits_{j=1}^{d_\xi}\sum\limits_{i=1}^{m_y} H_{j,i}(\mathbf{x}_m^*)\underline{w}_i^j\Delta_j^- - \lambda_m^* \sum\limits_{j=1}^{d_\xi}\alpha_j \end{array} \right\} \\
\text{s.t.} & \sum\limits_{i=1}^{m_y} W_{ji}\theta_i \le q_j, \quad j = 1,\dots,d_y, \\
& \overline{\xi}_j - \overline{\xi}_j^k + \Delta_j^+ z_j^+ - \Delta_j^- z_j^- \le \alpha_j, \quad j = 1,\dots,d_\xi, \\
& \overline{\xi}_j^k - \overline{\xi}_j + \Delta_j^- z_j^- - \Delta_j^+ z_j^+ \le \alpha_j, \quad j = 1,\dots,d_\xi, \\
& |\overline{w}_i^j - \theta_i| \le (1 - z_j^+)M \quad j = 1,\dots,d_\xi, \quad i = 1,\dots,m_y, \\
& |\underline{w}_i^j - \theta_i| \le (1 - z_j^-)M \quad j = 1,\dots,d_\xi, \quad i = 1,\dots,m_y, \\
& |\overline{w}_i^j| \le M z_j^+, \quad j = 1,\dots,d_\xi, \quad i = 1,\dots,m_y, \\
& |\underline{w}_i^j| \le M z_j^-, \quad j = 1,\dots,d_\xi, \quad i = 1,\dots,m_y, \\
& z_j^+ + z_j^- \le 1 \quad j = 1,\dots,d_\xi, \\
& z_j^+, z_j^- \in \{0,1\}, \quad j = 1,\dots,d_\xi,
\end{array}
\tag{A-4}
$$

where $M \in \mathbb{R}$ is sufficiently large and the products $\theta_i z_j^+$ and $\theta_i z_j^-$, for $j = 1,\dots d_\xi$, $i = 1,\dots,m_y$, are linearized.

A.5
The Farmer problem

One farmer specializes in raising N types of crops. He has $L\ (km^2)$ of land and he must decide how much land will be allocated to devote each crop. The fixed cost to rise the i-th type of crop is c_i per ton (T), for $i = 1,\dots,N$. By another hand, he must attend some restrictions related to his plantation; he must have at least $h_i\ (T)$ of the i-th type of crop, for $i = 1,\dots,N$. Those quantities can be obtained by own plantation or buying them in a local market. The purchase price for the i-th product is s_i per ton (T), for $i = 1,\dots,N$. Additionally, every excess of the i-th type of crop can be sold at the selling price of r_i per ton (T), for $i = 1,\dots,N$.

Let

ξ_i = productivity land for rising the i-th type of crop,

x_i = acres of land devoted to rise the i-th type of crop,

w_i = tons of the i-th type of crop sold,

y_i = tons of the i-th type of crop purchased

Since the farmer wants to minimize the cost, the two-stage stochastic linear optimization problem is:

$$\begin{aligned} \min_{x\geq 0} \quad & \sum_{i=1}^{N} c_i \cdot x_i + \mathbb{E}[Q(x,\boldsymbol{\xi})] \\ \text{s.t.} \quad & \sum_{i=1}^{N} x_i \leq L, \end{aligned} \tag{A-5}$$

where

$$\begin{aligned} Q(x,\boldsymbol{\xi}) = \min_{y,w} \quad & \sum_{i=1}^{N} s_i \cdot y_i - \sum_{i=1}^{N} r_i \cdot w_i \\ \text{s.t.} \quad & \xi_i \cdot x_i + y_i - w_i \geq h_i, \quad \forall i = 1,\ldots,N \\ & \mathbf{y} \geq 0, \mathbf{w} \geq 0. \end{aligned} \tag{A-6}$$

To simplify the problem, we assume that the productivity land for raising each crop is represented by independent random variables with uniform distribution.

For the two instances of the farmer's problem (eight and 20 types of crops) the data to run the computational experiments is reported bellow:

	Plantation cost	Purchase price	Selling price
i = 1	92	667	28
i = 2	80	905	35
i = 3	92	1024	41
i = 4	88	660	42
i = 5	91	974	25
i = 6	80	1041	35
i = 7	92	978	40
i = 8	93	997	45
Total land (L) : 3500			

Table A.1: Parameters of the Farmer's problem for 8 crops.

Uncertainty	a_i	b_i
i = 1	2,296	8,575
i = 2	1,981	6,820
i = 3	1,079	7,730
i = 4	2,470	6,232
i = 5	1,305	6,390
i = 6	0,900	6,445
i = 7	2,043	5,430
i = 8	0,409	4,214

Table A.2: Uncertainty support of the random vector for the Farmer´s problem with 8 crops.

	Plantation cost	Purchase price	Selling price
i = 1	81	853	137
i = 2	86	570	64
i = 3	85	817	135
i = 4	83	983	58
i = 5	81	1001	70
i = 6	93	547	131
i = 7	91	728	106
i = 8	85	966	135
i = 9	89	875	158
i = 10	94	913	87
i = 11	82	1010	118
i = 12	95	844	140
i = 13	92	698	133
i = 14	94	834	109
i = 15	92	724	107
i = 16	93	702	150
i = 17	84	980	95
i = 18	94	771	66
i = 19	85	752	77
i = 20	81	907	151
Total land (L) : 5200			

Table A.3: Parameters of the Farmer′s problem for 20 crops.

Uncertainty	a_i	b_i
i = 1	0,963	4,05
i = 2	0,844	2,72
i = 3	2,070	4,32
i = 4	2,640	4,60
i = 5	1,730	1,48
i = 6	2,400	4,99
i = 7	2,410	5,33
i = 8	1,980	4,03
i = 9	1,180	4,79
i = 10	2,060	3,27
i = 11	2,900	4,34
i = 12	1,100	4,39
i = 13	2,690	3,43
i = 14	1,260	6,12
i = 15	2,750	3,91
i = 16	1,150	5,36
i = 17	1,920	4,41
i = 18	1,760	5,68
i = 19	0,547	4,71
i = 20	0,508	3,91

Table A.4: Uncertainty support of the random vector for the Farmer´s problem with 20 crops.

A.6 The aircraft allocation problem

The aircraft allocation problem was one of the first stochastic linear programs ever formulated by Dantzig [22]. In this problem aircraft of different types are allocated on routes in order to minimize the operating costs. Besides the operating cost, there are costs associated with bumping passengers due to insufficient capacity to meet demand.

Let

I = set of available aircrafts,

R = set of routes,

$R(i)$ = subset of routes serviced by aircraft of type i,

b_i = number of aircraft available of type i,

c_{ir} = cost of operating an aircraft of type i along route r,

t_{ir} = passenger capacity of aircraft i on route r,

h_r = passenger demand on route r,

q_r = revenue lost per bumped passenger on route r,

x_{ir} = number of aircraft of type i assigned to route r,

y_r = number of bumped passengers on route r,

z_r = number of empty seats on route r.

we can set up the aircraft allocation problem with the following model:

$$\begin{aligned} \min_{x} \quad & \sum_{i \in I} \sum_{r \in R(i)} c_{ir} \cdot x_{ir} + \mathbb{E}[Q(x, \xi)] \\ \text{s.t.} \quad & \sum_{r \in R(i)} x_{ir} \leq b_i, \quad \forall i \in I, \end{aligned} \tag{A-7}$$

where

$$\begin{aligned} Q(x, \xi) = \min_{y,z} \quad & \sum_{r \in R} q_r \cdot y_r \\ \text{s.t.} \quad & \sum_{i \in I, r \in R(i)} t_{ir} \cdot x_{ir} + y_r - z_r = h_r, \quad \forall r \in R, \\ & x_{ir} \geq 0, \quad \forall i \in I, r \in R(i), \quad y_r \geq 0 \quad \forall r \in R. \end{aligned} \tag{A-8}$$

The input data to execute this computational experiment was taken from the following site:

Low recourse cost:

$$\mathbf{q} = (11, 13, 0, 13, 8, 7, 0, 7, 0, 12)^\top$$

High recourse cost:

$$\mathbf{q} = (301, 349, 239, 700, 254, 493, 348, 70, 474, 361)^\top$$