La collana della Scuola di Alta Formazione Dottorale accoglie le migliori tesi di dottorato dell'Università degli Studi di Bergamo, insignite della dignità di stampa e sottoposte a procedura di *blind peer review*.

This work deals with decision making problems plagued by the presence of uncertainty. The first analyzed problem aims at separating two sets of points with non-disjoint convex closures. To this end, robust and distributionally robust Support Vector Machine (SVM) models are formulated and their efficiency is evaluated on real-world databases.
Being distributionally robust models notoriously hard to solve, the second chapter of this work proposes approximation techniques providing bounds on objective function optimal values. This is done through scenario grouping and via ϕ-divergences and Wasserstein distance.
The last problem this work investigates aims at detecting the optimal assortment a retailer shall offer to maximize profits when strong preferences among products are observed. A deterministic approximation is recovered to solve the original intractable stochastic formulation.

Daniel Faccini obtained his PhD in Applied Economics & Management (34th cycle) from the University of Bergamo and the University of Pavia. His research interests are focused on decision making problems under uncertainty, with applications to Machine Learning and Revenue Management. He conducted part of his research at Georgia Institute of Technology (Atlanta, U.S.A) - School of Industrial and Systems Engineering.

ISBN: 978-88-97413-71-4
DOI: 10.13122/978-88-97413-71-4

Daniel Faccini MODELS AND APPROXIMATIONS FOR OPTIMIZATION PROBLEMS

50

Collana della Scuola di Alta Formazione Dottorale
- 50 -

Daniel Faccini

MODELS AND APPROXIMATIONS for Optimization Problems under Uncertainty

UNIVERSITÀ DEGLI STUDI DI BERGAMO

2023

Collana della Scuola di Alta Formazione Dottorale

Diretta da Paolo Cesaretti

Ogni volume è sottoposto a *blind peer review*.

ISSN: 2611-9927

Sito web: https://aisberg.unibg.it/handle/10446/130100

Daniel Faccini

MODELS AND APPROXIMATIONS FOR OPTIMIZATION PROBLEMS
under Uncertainty with Applications to Support Vector Machine and Revenue Management

Università degli Studi di Bergamo

2023

Models and Approximations for Optimization Problems under Uncertainty with Applications to Support Vector Machine and Revenue Management / Daniel Faccini. – Bergamo : Università degli Studi di Bergamo, 2023.
(Collana della Scuola di Alta Formazione Dottorale; 50)

ISBN: 978-88-97413-71-4
DOI: 10.13122/978-88-97413-71-4

Progetto grafico: Servizi Editoriali – Università degli Studi di Bergamo
© 2018 Università degli Studi di Bergamo
via Salvecchio, 19
24129 Bergamo
Cod. Fiscale 80004350163
P. IVA 01612800167

https://aisberg.unibg.it/handle/10446/130115

Acknowledgments

"*Happiness only real when shared*"

Christopher McCandless

I would like to express my deepest gratitude towards my Ph.D. supervisor, Prof. Francesca Maggioni, whose patience and guidance made all of this possible. She has been a wise and trustworthy adviser, a real mentor to me, and I could not have completed this work without her unwavering support.

There are so many other people I will be forever indebted to. Foremost are my sisters, who thought me to think critically, dream fearlessly and be excited for new adventures in life. I thank them for their endless love and encouragement. They have always been my role models and this work is dedicated to them. Besides, I will never be able to express how grateful I am to my nieces, Veronica and Alice, who fill my everyday life with happiness and goofiness. No matter how badly I fail, I known they will always treat me like a winner. Also, I thank my parents. They have sacrificed a lot to bring me up to this stage of my life and any of my achievement is their achievement. Without their inspiration, I would not be the person I am today.

I am also immensely obliged to my friends. Lisa, for being all ears to my problems, never judging or complaining; Greta, for making me feel there is nothing I am not worthy of, even in the darkest hours; and Sara, for always cheering me up when I am feeling low, making hard times easier or good times funnier. Another special friend I had the incredible joy and pleasure of working with over this journey is Irina: thank you for the morning coffee chats and for patiently dealing with my Ph.D. frustration.

Finally, I want to extend my appreciation to all those who could not be mentioned here, yet always believed in me. We made it, all together.

DANIEL

Table of Contents

Introduction

Every day people are asked to make decisions with respect to a future that has yet to be known. Neglecting this uncertainty, however, may lead to solutions that are far from optimal and exhibiting remarkable sensitivity to perturbations.

A natural way of addressing the uncertain nature of a decision problem is through *Stochastic Programming* (SP) [27,78,146], which assumes the uncertainty to have a known probabilistic description. The main difficulty associated with this approach is the need to provide the probability distribution function of the underlying stochastic parameter, a requirement that creates a heavy burden on the user because in many real world situations such information is unavailable or hard to obtain. A more recent approach addressing the uncertain nature of a problem without making specific assumptions on probability distributions is *Robust Optimization* (RO) [9], in which instead of seeking to immunize the solution in some probabilistic sense, the decision-maker builds a solution that is feasible for any realization of the uncertainty in a given set. The major drawback of RO approaches is their excess of conservatism.

The mutual need of protecting the decision-maker from the ambiguity of the underlying probability distribution while avoiding overly conservative solutions is fulfilled by *Distributionally Robust Optimization* (DRO) [43,67,139,178], in which optimal decisions are sought for the worst-case probability distribution within a family that is well-described by certain properties. This approach has seen numerous applications for a wide variety of management inspired optimization problems, and in Chapter 1 of this work we provide practical evidence on how successful DRO is in determining well-performing optimal solutions in the context of *Machine Learning* (ML). Specifically, in this chapter, our attention is centered on the study of binary classification problems, whose goal is to categorize data points into one of two *buckets*: true or false, healthy or unhealthy, defaulting or not defaulting, to name but a few. Already gathered observations' features are exploited to detect the classifier, which should have a good generalization ability and therefore minimize the misclassification error of new unseen data. Nonetheless, as highlighted before, assuming such observations to be not corrupted by noise is often impractical: uncertainty, indeed, may easily manifest due to limited precision of collecting instruments, human measuring mistakes or sampling errors. As a consequence, the problem of designing classifiers not facing deterioration when there are some perturbations in the data set is an interesting problem that lately has gained considerable attention. To deal with the binary classification problem under feature uncertainty of the input data, we propose RO and DRO versions of one of the deter-

ministic *Support Vector Machine* (SVM) formulations presented in [96]. We will first consider input observations to be bounded within hyperrectangles and hyperellipsoids. Secondly, we will formulate a moment-based distributionally robust counterpart assuming each observation to be unknown but we will mitigate the degree of conservatism enforcing limits on the observations' deviations along directions detected by means of principal component analysis. With experiments crossing many different applications fields, ranging from business to physical science, we will show that the solutions of robust and distributionally robust models provide more accurate solutions compared to their deterministic counterpart. Finally, managerial insights which can be valuable for practitioners will be provided.

Despite its importance and broad applicability, addressing uncertainty through DRO often leads to computationally intractable models, especially for multistage problems that involve sequences of decisions over time and suffer from the curse of dimensionality. In these cases, providing bounds for their optimal objective function values can be very useful in practice, as may help in evaluating whether the DRO approach is worth the additional computational effort or if simplified approaches should be preferred. Therefore, Chapter 2 of this work is devoted to the definition of novel bounding schemes for multistage mixed-integer DRO programs, defined on a finite scenario tree and formed via ϕ-divergences (*i.e.*, variation distance, Cressie-Read power divergence family, J-divergence, and χ-divergence of order $a > 1$) or Wasserstein distance. The approach we suggest divides the whole sample space into independent subgroups, which being of smaller size can be solved more efficiently. We then provide conditions on ways to combine the optimal values of the subgroups to obtain lower bounds on the optimal value of the original problem. Our approach does not require any special problem structure, such as convexity and linearity, so suggested bounds can be applied to a wide range of DRO problems including two-stage and multistage, with or without integer variables, nested or non-nested formulations. The effectiveness of the proposed bounds is investigated on a multistage mixed-integer production planning problem, providing a discussion of the insights gained.

The third and last chapter of this work is devoted to the study of the assortment planning problem [110], a critical task faced by many industries operating onto diversified markets and which consists in determining the optimal subset of products a retailer should offer to potential buyers over a selling horizon. Uncertainty highly affects this class of management problems too: indeed, one of the major challenges in assortment planning is to understand which is the "right" demand model to use for describing the behavioral process that leads customers to choose, and hence to determine products purchasing probabilities. Classical choice models (*e.g.*, Multinomial Logit, Nested Logit, etc.) have failed, so far, to properly describe strong substitution behaviors in the form of 100% buydown effects, exhibited by customers with net preferences over one product with respect to another, thus annulling

the latter purchasing probability whenever the former belongs to the same assortment. Therefore, we first propose a novel choice model able to capture 100% buydown effects and which we call *Dominance Multinomial Logit* (DMNL); then, we use it to formulate the assortment planning problem via *Dynamic Programming* (DP). The resulting model, however, suffers from the curse of dimensionality and its computational burden increases exponentially in the number of products available. For this reason, we recover a deterministic approximation called DMNL *Sales Based Linear Program* (DMNL-SBLP) that captures 100% buydown effects and avoids at the same time the exponential number of variables. Preliminary numerical experiments using synthetic data compare our novel approach with the classical discrete choice model.

Chapter 1. Robust and Distributionally Robust Optimization Models for Linear Support Vector Machine

In collaboration with Francesca Maggioni [1] and Florian A. Potra [2]. Released in a different version on *Computers & Operations Research*.

1.1 Introduction

Binary pattern separation is one of the main *Machine Learning* (ML) tasks [86]. Its aim is to classify observations into one of two classes and it is a critical problem in many practical application fields, such as robotics [35], environmental engineering [42, 118, 119], nutrition [151], neural and medical image analysis [190] and computer security [23]. From the ML standpoint, a great variety of algorithms have been devised to address the classification problem: *Decision Trees* (DT) [136], *Logistic Regression* (LR) classifiers [45], k-*Nearest Neighbors* (NN) classifiers [49], and *Support Vector Machines* (SVM), which though simple and intuitive have proved to be one of the most effective estimation techniques [183]. A recent comparison of ML methods for binary classification is found in [5].

SVM is a supervised ML algorithm tracing back to the seminal contribution of [169], which has received significant attention in the optimization literature and has strong orientation towards real-world applications [101]. Given a set of training observations, each labeled as belonging to one of two classes, SVM goal is to detect a hyperplane induced from the available examples that is able to predict the category of new unlabeled observations. The most basic version of the SVM is the *Hard Margin*-SVM (HM-SVM) that assumes that there exists a hyperplane geometrically separating data points into the two classes, such that no observation is misclassified and margins are maximized. When the data is linearly inseparable, the *Soft Margin* SVM (SM-SVM) introduces slack variables into the constraints and aims at finding a separating hyperplane that not only achieves the maximum margin between the two classes but also minimizes the training error of misclassification [10, 39]. Many variations to the classical SVM approach have been proposed over time to enhance classifiers predictive power, see for instance [22, 81, 91, 96, 111, 140, 170]. In this chapter, we specifically focus

[1]Department of Management, Information & Production Engineering, University of Bergamo
Dalmine, BG, IT. Francesca.Maggioni@unibg.it

[2]Department of Mathematics and Statistics, University of Maryland
Baltimore County, MA, USA. Potra@math.umbc.edu

our attention on the SVM variant presented in [96], whose computational experience proved to detect separators with higher levels of accuracy compared to the standard ones.

An underlying assumption of classical SVM approaches is that the input observations are not corrupted with noise and, therefore, all problem data are known exactly at the moment of classifying [34]. This assumption, however, is not always practical. Indeed, real-world observations are often plagued by uncertainty (*e.g.*, due to limited precision of collecting instruments, measurement mistakes in data gathering, sampling errors, etc.) and disregarding it might lead to solutions that are far from optimal, as well as to major fluctuations of performances [68]. Therefore, the problem of designing classifiers not facing deterioration when there are some perturbations in the data set is an interesting problem that has gained considerable attention from the scientific community. One of the main paradigms to deal with problems affected with uncertain data is given by *Robust Optimization* (RO) (see [9] and [12]). Another way to handle uncertainty is given by *Distributionally Robust Optimization* (DRO) pioneered in [139] and [187], which can be regarded as a natural generalization of *Stochastic Programming* (SP) and RO. In DRO optimal decisions are sought for the worst-case probability distribution within a family of possible distributions defined by certain properties. The two most widely used types of ambiguity sets in the DRO literature are moment-based and statistical distance-based sets. While moment-based ambiguity sets contain all probability distributions that satisfy certain general moment conditions, the statistical distance-based approach considers distributions that are close in the sense of a chosen distance to a nominal distribution (*e.g.*, the empirical one). Popular choices to measure the dissimilarity between two probability distributions are Wasserstein distance or ϕ-divergences. A growing literature in these directions both from theoretical and applied points of view can be found in [1, 6, 43, 65, 67, 145, 147, 178, 185, 192].

In this chapter, we deal with the binary classification problem under feature uncertainty of the input data, introducing robust and distributionally robust versions of one of the deterministic formulations presented in [96] (Formulation II), aiming at obtaining a classifier that has good generalization properties and reduces the error of misclassification of new unseen data. The main contributions of the chapter are four-fold and can be summarized as follows:

- To develop box and ellipsoidal robust counterparts of the deterministic model associated with the Formulation II proposed in [96]. We assume each input observation to be bounded within hyperrectangles and hyperellipsoids.

- To formulate a new moment-based distributionally robust counterpart associated with the Formulation II proposed in [96]. We still assume each observation to be unknown but we mitigate

the degree of conservatism enforcing limits on the deviations along directions detected by means of *Principal Component Analysis* (PCA) [73].

- To provide extensive numerical experiments based on real-world databases [46] with the aim of understanding the advantage of explicitly considering the uncertainty versus deterministic approaches.

- To provide managerial insights on how to choose between robust and distributionally robust approaches to model uncertainty, depending on the data set dimension.

The chapter is organized as follows. Section 1.2 provides a literature review, while Section 1.3 presents basic facts and notation. In Section 1.4 we introduce new robust and distributionally robust optimization models for SVM, along with tractable reformulations. Section 1.5 presents experiments attempting to evaluate the accuracy of the proposed formulations versus deterministic approaches. Finally, conclusions and future works are provided in Section 1.6.

1.2 Literature Review

The extensive connections among RO, DRO and SVM have been explored by a number of authors. In [53] a minimax model for data bounded by hyper-rectangles is presented. The model looks for a linear hyperplane that minimizes the worst-case loss over input data in given intervals, and a tractable reformulation in the form of *Linear Programming* (LP) is provided. In [19–21] *Second Order Cone Programming* (SOCP) formulations are derived to design linear classifiers when the uncertainty of input observations is described by multivariate normal distributions. Geometrically, these solutions correspond to a minimax strategy with hyper-ellipsoids around the training instances, rather than hyper-rectangles. Similar approaches are provided in [162, 163], where the additive perturbations of the uncertain data are assumed to be bounded by the general w-norm. A related model is [22] that, assuming the data to be subject to additive noises bounded by the general w-norm, constructs classifiers by focusing on the more trust-worthy data that are less uncertain. A more general case for bounded uncertainty sets is studied in [181], where the linkage between regularization and robustness is also showed. The authors proved that, even though traditional SVM methods do not explicitly consider individual data uncertainties, the objective function regularization term aimed at maximizing the classifier margins represents a kind of intrinsic robustness. Other important insights about stability of SVM against uncertainty with bounded sets are due to [161], while the work developed in [80, 116] demonstrate how robust classification can be used to handle situations with imbalanced training data.

For other models with polyhedral uncertainty sets see [54, 61]. Detailed reviews of the existing literature on RO in ML are found in [33].

RO and DRO are also used for solving *Chance-Constrained* (CC)-SVM, to ensure bounded probabilities of misclassification for the uncertain data. In [87] authors consider the case of binary classification, where only the mean and covariance matrix of the classes are assumed to be known. The minimax probabilistic decision hyperplane is then determined by optimizing the worst-case probabilities over all possible class-conditional distributions. Besides, the model presented in [150] treats all input observations as random variables for which only finite mean and covariance matrices are known, and then looks for the hyperplane able to correctly classify the observations, with high probability, even for the worst distributions. Both of these CC-SVM are relaxed using Chebyshev inequality (see [109]) to yield a SOCP whose solution is guaranteed to satisfy the original problem. In a similar fashion, the Bernstein bounding scheme (see [125]) is used in [7, 18]. Under the same assumptions of known moments, equivalent results have been obtained in [175], where authors propose a different proof for obtaining the equivalent SOCP formulation and also provide reformulations in the form of *Semidefinite Programming* (SDP) models. Analogously, *Pearson divergence* distributionally robust CC-SVM is discussed in [149]. Another related work is [174], which investigates the stochastic sub-gradient descent method to solve distributionally robust CC-SVM on large-scale data sets.

In [71] risk averse theory is linked to SVM, showing that the minimization of a convex risk functional in place of the traditional hinge-loss objective function (*i.e.*, minimization of the empirical risk) straightforwardly treats a class of DRO problems. This corresponds to build an ambiguity set for the population distribution based on samples, and then searching for the classifier that minimizes the sum of the regularization term and the hinge-loss function for the worst-case distribution within the set. Authors also prove that under a specific class of risk functionals the distributionally robustified models can be reformulated as tractable convex optimization problems. Risk averse SVM is further investigated in [171] where the authors, instead of using a single measure of risk as SVM objective function, propose group differentiation by employing a different risk functional for every single class. Other related studies are [70, 157, 164, 182]. In a similar fashion, DRO for classification problems with Wasserstein ambiguity set has been investigated in [85, 89]. Instead of solving an optimization problem minimizing the hinge-losses of misclassified samples, the proposed formulation minimizes the worst-case expected prediction error with respect to distributions belonging to a Kantorovich ball, which is centered on the empirical distribution based on samples. Related works are [93, 100]. Learning and classification algorithms have also been proposed under the ϕ-divergence measures, see

for instance [47, 48], and with ambiguity sets measured via maximum mean discrepancy, see [154]. All these approaches for linear SVM models are summarized in Table 1.

While all these approaches have dealt mainly with input data features uncertainty, there have also been attempts to model uncertainty in observation labels, see [24, 33, 112, 156, 179] and [13], where robust methods are employed to construct a new family of classifiers protecting against uncertainty in both features and labels for the three most widely used classification algorithms (*i.e.*, SVM, LR, and DT). RO is also employed in [66] to address the problem of corruption in missing data (see [64]), sensitivity to outliers in input samples (see [60, 79, 88, 92, 183]) and to adversarial training (see [98, 141, 180, 191]), where it is assumed that data become corrupted during the classification phase.

The approaches presented so far to hedge against uncertainty have also been successfully applied to many SVM variants. Robust counterparts have indeed been developed for the *Twin Support Vector Machine* (T-SVM), firstly proposed by [81]. See, for instance, [31, 99, 108, 126] and references therein. An alternative formulation, known as ν-*Support Vector Machine* (ν-SVM), was designed in [140], and models to hedge against uncertainty are proposed in [158, 176]. Another popular variant of SVM is the so called *One-Class Support Vector Machine* (OC-SVM) pioneered in [111], with robust reformulations that can be found in [97, 165–167]. There also has been a recent surge of interest in the ML community for developing distributionally robust SVM models aiming at fairness, which represents the need of a classifier performance to be invariant under certain sensitive perturbations of the inputs. Fairness in ML goes beyond the scope of this article, so we refer to [72, 158, 160, 177] and references therein. For a comprehensive survey of RO developments in the field of SVM we refer the reader to [152, 153].

The approach we propose in this chapter substantially differs from the literature in several perspectives. Foremost, the deterministic variant we aim at robustifying is the one proposed in [96], which with the inclusion of a line search step showed to outperform the classical formulation in prediction accuracy. Besides, two streams of distributionally robust approaches have emerged from the review of SVM literature. The first poses the SVM problem as a CC program and then looks for bounding schemes that find solutions guaranteed to satisfy the probabilistic constraint in the worst-case distribution. The second stream, instead, aims at minimizing in the objective function the worst-case expected prediction error with respect to distributions belonging to a prespecified ambiguity set. Our proposal does not fall into any of these branches, since we are not dealing with CC programs or with uncertainty into the objective function, rather we consider input data to be random variables with unknown distributions, and then we optimize over the worst one affecting the coefficients of the constraints left-hand sides. Furthermore, we provide exact reformulations rather than approximations.

	Uncertainty			*Methodology*									
	Features Uncertainty	*Labels Uncertainty*	*Missing Data Uncertainty*	*Box RO*	*Ellipsoidal RO*	*Bounded by norm RO*	*Polyhedral RO*	*Chance Constraints*	*Moments DRO*	*Wasserstein DRO*	*ϕ-divergences DRO*	*Mean Discrepancy DRO*	*Risk Averse*
Lanckriet et al. (2002), [87]	✓							✓	✓				
El Ghaoui et al. (2003), [53]	✓	✓		✓									
Fung et al. (2003), [61]	✓						✓						
Bhattacharyya et al. (2004, 2005), [19–21]	✓		✓		✓								
Bi and Zhang (2005), [22]	✓					✓							
Shivaswamy et al. (2006), [150]	✓		✓					✓	✓				
Trafalis and Gilbert (2006, 2007), [162, 163]	✓					✓							
Takeda and Kanamori (2009), [157]	✓												✓
Bhadra et al. (2009), [18]	✓							✓	✓				
Xu et al. (2009), [181]	✓					✓							
Trafalis and Alwazzi (2010), [161]	✓					✓							
Ben-Tal et al. (2011), [7]	✓							✓	✓				
Pant et al. (2011), [116]	✓					✓							
Tsyurmasto et al. (2013), [164]	✓												✓
Fan et al. (2014), [54]	✓						✓						
Gotoh et al. (2014), [70]	✓												✓
Katsumata and Takeda (2015), [80]	✓					✓							
Lee and Mehrotra (2015), [89]	✓									✓			
Gotoh and Uryasev (2017), [71]	✓												✓
Wang et al. (2017, 2018), [174, 175]	✓							✓	✓				
Duchi et al. (2019, 2021), [47, 48]	✓										✓		
Bertsimas et al. (2019), [13]	✓	✓				✓							
Kuhn et al. (2019), [85]	✓									✓			
Staib and Jegelka (2019), [154]	✓											✓	
Vitt et al. (2019), [171]	✓												✓
Li at al. (2020), [93]	✓									✓			
Shen at al. (2020), [149]	✓							✓			✓		

Table 1: Linear SVM Literature Review.

1.3 Basic Facts and Notation

In the following, all vectors will be column vectors. We use ";" for adjoining elements in a column and "," for adjoining elements in a row. Vector components are identified as being subscripted, while superscripts specify to which observation we are referring to. Vector e of arbitrary dimension has all entries equal to one, while $\mathcal{I}$ and $\mathbf{0}$ denote, respectively, the identity matrix and the square null matrix of dimension n. We denote by $\mathbb{R}^n$ the n-dimensional real space, by $\mathbb{R}^n_+$ the set of non-negative vectors of dimension n, by $\mathbb{N}$ the set of natural numbers and by $\mathrm{diag}(a) \in \mathbb{R}^{n\times n}$ the matrix whose n diagonal entries are the elements of vector a and off-diagonal components are all equal to zero. For any vector $a \in \mathbb{R}^n$, $|a| \in \mathbb{R}^n_+$ represents the vector of absolute values of the components of a, *i.e.*, $|a| := [|a_1|; |a_2|; \ldots; |a_n|]$. For any vector $a \in \mathbb{R}^n$ and $1 \leq w < \infty$, its w-norm is defined as $\|a\|_w$ with:

$$\|a\|_w := \left(\sum_{p=1}^{n} |a_p|^w \right)^{\frac{1}{w}}$$

and

$$\|a\|_\infty := \max_{p=1,\ldots,n} |a_p|.$$

Finally, the indicator function $\mathbb{1}(\alpha \in \mathbb{R}) = 1$ if $\alpha > 0$, and 0 otherwise.

1.3.1 The Classification Problem

Let X and Y be two sets of points such that $X := \left\{x^{(1)}, x^{(2)}, \ldots, x^{(I)}\right\} \subseteq \mathbb{R}^n$ and $Y := \left\{y^{(1)}, y^{(2)}, \ldots, y^{(J)}\right\} \subseteq \mathbb{R}^n$.

The *Hard Margin* SVM (HM-SVM) separating hyperplane is defined by a pair $(a \in \mathbb{R}^n, \gamma \in \mathbb{R})$ such that all vectors in X lie on one side of the hyperplane, all the vectors in Y lie on the opposite side and the distance between the separating hyperplane and the nearest data point of each class is maximized [169]. The HM-SVM optimization problem is defined as follows:

$$\begin{aligned} \min_{a,\gamma} \quad & \|a\|_w \\ \text{s.t.} \quad & a^\top x^{(i)} \leq \gamma - 1 \qquad i = 1, \ldots, I \\ & a^\top y^{(j)} \geq \gamma + 1 \qquad j = 1, \ldots, J, \end{aligned} \tag{1.1}$$

whose solution maximizes the distance between the hyperplanes $(a, \gamma - 1)$ and $(a, \gamma + 1)$ computed using the dual norm $\|\cdot\|_v$ with $\frac{1}{v} + \frac{1}{w} = 1$. The dual norm of the 1-norm is the infinity norm, and vice versa.

Soft Margin SVM (SM-SVM) relaxes the condition of perfect separability, introducing slack variables in the constraints and penalizing in the objective function data points belonging to the wrong side of

the hyperplane. Specifically, let $z_X := [z_{x^{(1)}}; \ldots; z_{x^{(I)}}] \in \mathbb{R}^I_+$ and $z_Y := [z_{y^{(1)}}; \ldots; z_{y^{(J)}}] \in \mathbb{R}^J_+$ be the non-negative vectors of errors of group X and Y. Observation $x^{(i)} \in \mathbb{R}^n$ will be correctly classified if $0 \leq z_{x^{(i)}} \leq 1$, or misclassified if $z_{x^{(i)}} > 1$. Similarly, for every observation $y^{(j)} \in \mathbb{R}^n$. The SM-SVM optimization problem is then defined as follows [39]:

$$\begin{aligned} \min_{a,\gamma,z_X,z_Y} \quad & \|a\|_w + \nu \left(e^\top z_X + e^\top z_Y\right) & \\ \text{s.t.} \quad & a^\top x^{(i)} \leq \gamma - 1 + z_{x^{(i)}} & i = 1, \ldots, I \\ & a^\top y^{(j)} \geq \gamma + 1 - z_{y^{(j)}} & j = 1, \ldots, J \\ & z_X \geq 0, \ z_Y \geq 0, & \end{aligned} \tag{1.2}$$

where the user-defined penalty parameter $\nu \geq 0$ is introduced to allow a trade-off between the margin maximization and tolerating misclassification.

In order to achieve superior pattern separation, rather than minimizing the classification error with respect to a single hyperplane, in [96] it is proposed to separate the sets X and Y by firstly finding two parallel hyperplanes H_1 and H_2 that satisfy the following properties:

(P1) all points of X lie on one side of H_1;

(P2) all points of Y lie on the opposite side of H_2;

(P3) the intersection of convex hulls of X and Y is contained in the region between H_1 and H_2.

Through line search, hyperplane H_3 is then constructed parallel to (and lying between) H_1 and H_2, such that most of the points of X lie on the same side of H_3 and most of the points of Y lie on the opposite side of H_3. A point that fails to do so is called a *misclassified point*. Therefore, H_3 should be determined so that the number of misclassified points is minimized. In [96] five different deterministic formulations are proposed for obtaining hyperplane H_3, and since Formulation II proves to outperform the others, we restrict our attention to it. This formulation employs as starting point the hyperplane separating algorithm detected by model (1.2), in which the hyperplane margins are measured by means of the ∞-norm, and hence requires the minimization of $\|a\|_1$ into the objective function:

$$\begin{aligned} \min_{a,\gamma,z_X,z_Y} \quad & \|a\|_1 + \nu \left(e^\top z_X + e^\top z_Y\right) & \\ \text{s.t.} \quad & a^\top x^{(i)} \leq \gamma - 1 + z_{x^{(i)}} & i = 1, \ldots, I \\ & a^\top y^{(j)} \geq \gamma + 1 - z_{y^{(j)}} & j = 1, \ldots, J \\ & z_X \geq 0, \ z_Y \geq 0. & \end{aligned} \tag{1.3}$$

Once the starting hyperplane (a, γ) of (1.3) is obtained, it is shifted in order to determine hyperplanes H_1 and H_2 that satisfy properties **(P1)-(P3)**. Specifically, $H_1 := (a, \gamma - 1 + \omega_1)$ and $H_2 := (a, \gamma + 1 - \omega_2)$, where:

$$\omega_1 := \max\left\{z_{x^{(i)}} \;\middle|\; i = 1, \dots, I\right\}, \qquad \omega_2 := \max\left\{z_{y^{(j)}} \;\middle|\; j = 1, \dots, J\right\}. \tag{1.4}$$

The following minimization problem is finally solved using the line search method (see [114]), with the aim of obtaining the scalar $b \in \mathbb{R}$ that defines the hyperplane $H_3 := (a, b)$, parallel to and lying between H_1 and H_2 and minimizing the overall number of misclassified points:

$$\begin{aligned} \min_{b} \quad & \sum_{i=1}^{I} \mathbb{1}\left(a^\top x^{(i)} - b\right) + \sum_{j=1}^{J} \mathbb{1}\left(b - a^\top y^{(j)}\right) \\ \text{s.t.} \quad & \gamma + 1 - \omega_2 \leq b \leq \gamma - 1 + \omega_1. \end{aligned} \tag{1.5}$$

Specifically, as the objective of (1.5) is not continuous, we divide the interval $[b_{\min}, b_{\max}] := [\gamma + 1 - \omega_2, \gamma - 1 + \omega_1]$ into $k_{\max}$ sub-intervals of equal length and denote $s_k := \sum_{i=1}^{I} \mathbb{1}\left(a^\top x^{(i)} - b_k\right) + \sum_{j=1}^{J} \mathbb{1}\left(b_k - a^\top y^{(j)}\right)$, with $b_k = b_{\min} + k \cdot \frac{b_{\max} - b_{\min}}{k_{\max}}$, $k = 0, \dots, k_{max}$. The final solution of (1.5) is then given by b_{k^*} with $k^* \in \arg\min\{s_0, \dots, s_k, \dots, s_{k_{\max}}\}$.

1.4 Robust and Distributionally Robust Support Vector Machine Models

The basic assumption of the deterministic model (1.3)-(1.5) presented in [96] is that all input observations of both groups X and Y are always provided exactly, ignoring any type of uncertainty associated with lack of data or with data that cannot be fully trusted. However, when the given values differ significantly from the true ones, the predictive power of the deterministic classifier might be unsatisfactory. Therefore in this section, rather than dealing with a countable set of well-defined data points, we handle data features as uncertain and formulate robust counterparts to model problem (1.3)-(1.5) with uncertainty sets in the form of hyperrectangles (Section 1.4.1) and hyperellipsoids (Section 1.4.1). Moreover, we propose a distributionally robust counterpart to the deterministic formulation (1.3)-(1.5) that enforces limits on the observations first-order deviations along directions detected by means of PCA (Section 1.4.2).

1.4.1 Robust Support Vector Machine

In this section, we assume the uncertainty of every input observation $x^{(i)} \in X \subseteq \mathbb{R}^n,\ i = 1, \dots, I$ to be represented by the uncertainty set $\mathcal{U}\left(x^{(i)}\right)$. Equivalently for every observation $y^{(j)} \in Y \subseteq \mathbb{R}^n,\ j = 1, \dots, J$. Then, the robust counterpart of model (1.3) that optimizes over worst-case realizations on

all possible observations in $\mathcal{U}\left(x^{(i)}\right), \mathcal{U}\left(y^{(j)}\right), i = 1, \ldots, I, j = 1, \ldots, J$ corresponds to the following optimization model:

$$\begin{aligned} \min_{a,\gamma,z_X,z_Y} \quad & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) \\ \text{s.t.} \quad & \max_{x \in \mathcal{U}(x^{(i)})} \left[a^\top x\right] \leq \gamma - 1 + z_{x^{(i)}} \qquad i = 1, \ldots, I \\ & \min_{y \in \mathcal{U}(y^{(j)})} \left[a^\top y\right] \geq \gamma + 1 - z_{y^{(j)}} \qquad j = 1, \ldots, J \\ & z_X \geq 0, \ z_Y \geq 0. \end{aligned} \tag{1.6}$$

The size of the uncertainty sets $\mathcal{U}\left(x^{(i)}\right), \mathcal{U}\left(y^{(j)}\right), i = 1, \ldots, I, j = 1, \ldots, J$ reflects the degree of data uncertainty. If:

$$\mathcal{U}\left(x^{(i)}\right) := \left\{x^{(i)}\right\}, \ i = 1, \ldots, I \quad \text{and} \quad \mathcal{U}\left(y^{(j)}\right) := \left\{y^{(j)}\right\}, \ j = 1, \ldots, J,$$

then the robust formulation (1.6) reduces to the deterministic model (1.3).

Robust Support Vector Machine with Interval Data Uncertainty

First, we consider uncertainty sets having the form of hyperrectangles. Let $\zeta_{x^{(i)}}, \zeta_{y^{(j)}} \in \mathbb{R}^n_+$ define the perturbation vectors of input observations $x^{(i)}$ and $y^{(j)}$, respectively; further, let $\rho_X, \rho_Y \in \mathbb{R}_+$ be global measures of uncertainty for group X and Y, respectively. Then, the hyperrectangular uncertainty sets $\mathcal{U}_\mathcal{B}\left(x^{(i)}\right)$ and $\mathcal{U}_\mathcal{B}\left(y^{(j)}\right)$ centered around $x^{(i)}$ and $y^{(j)}$ are defined, respectively, as:

$$\mathcal{U}_\mathcal{B}\left(x^{(i)}\right) := \left\{ x \in \mathbb{R}^n \ \middle| \ x^{(i)} - \rho_X \zeta_{x^{(i)}} \leq x \leq x^{(i)} + \rho_X \zeta_{x^{(i)}} \right\} \quad i = 1, \ldots, I, \tag{1.7}$$

$$\mathcal{U}_\mathcal{B}\left(y^{(j)}\right) := \left\{ y \in \mathbb{R}^n \ \middle| \ y^{(j)} - \rho_Y \zeta_{y^{(j)}} \leq y \leq y^{(j)} + \rho_Y \zeta_{y^{(j)}} \right\} \quad j = 1, \ldots, J. \tag{1.8}$$

Depending on how reliable the decision maker considers the available data, parameters ρ_X and ρ_Y allow to tailor the degree of conservatism. When uncertainty sets are described by means of (1.7)-(1.8), model (1.6) can be reformulated by the following linear program (see [53] and derivation in Appendix A):

$$\begin{aligned} \min_{a,\gamma,z_X,z_Y} \quad & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) \\ \text{s.t.} \quad & a^\top x^{(i)} + \rho_X \zeta_{x^{(i)}}^\top |a| \leq \gamma - 1 + z_{x^{(i)}} \qquad i = 1, \ldots, I \\ & a^\top y^{(j)} - \rho_Y \zeta_{y^{(j)}}^\top |a| \geq \gamma + 1 - z_{y^{(j)}} \qquad j = 1, \ldots, J \\ & z_X \geq 0, \ z_Y \geq 0, \end{aligned} \tag{1.9}$$

where the number of continuous variables is $n + 1 + I + J$ and the number of constraints is $2(I + J)$ of which $I + J$ are non-negative. As in the deterministic case, once the solution (a, γ, z_X, z_Y) of (1.9)

is obtained, the final hyperplane H_3 is recovered through line search:

$$\begin{aligned} \min_{b} \quad & \sum_{i=1}^{I} \mathbb{1}\left(a^\top x^{(i)} + \rho_X \zeta_{x^{(i)}}^\top |a| - b\right) + \sum_{j=1}^{J} \mathbb{1}\left(b - a^\top y^{(j)} + \rho_Y \zeta_{y^{(j)}}^\top |a|\right) \\ \text{s.t.} \quad & \gamma + 1 - \omega_2 \leq b \leq \gamma - 1 + \omega_1, \end{aligned} \tag{1.10}$$

with ω_1, ω_2 as in (1.4) and where robustness fails for those points whose hyperrectangle intersects the hyperplane H_3. Consequently, all those points either lying on the wrong side of (a, b) or whose hyperrectangles intersect H_3 will be considered misclassified. To summarize, the geometrical interpretation of the proposed approach is sketched in Figure 1. For the sake of clarity, we restrict our attention to the bidimensional case ($n = 2$). Points of group X are represented by filled black dots, while points of group Y by empty white circles.

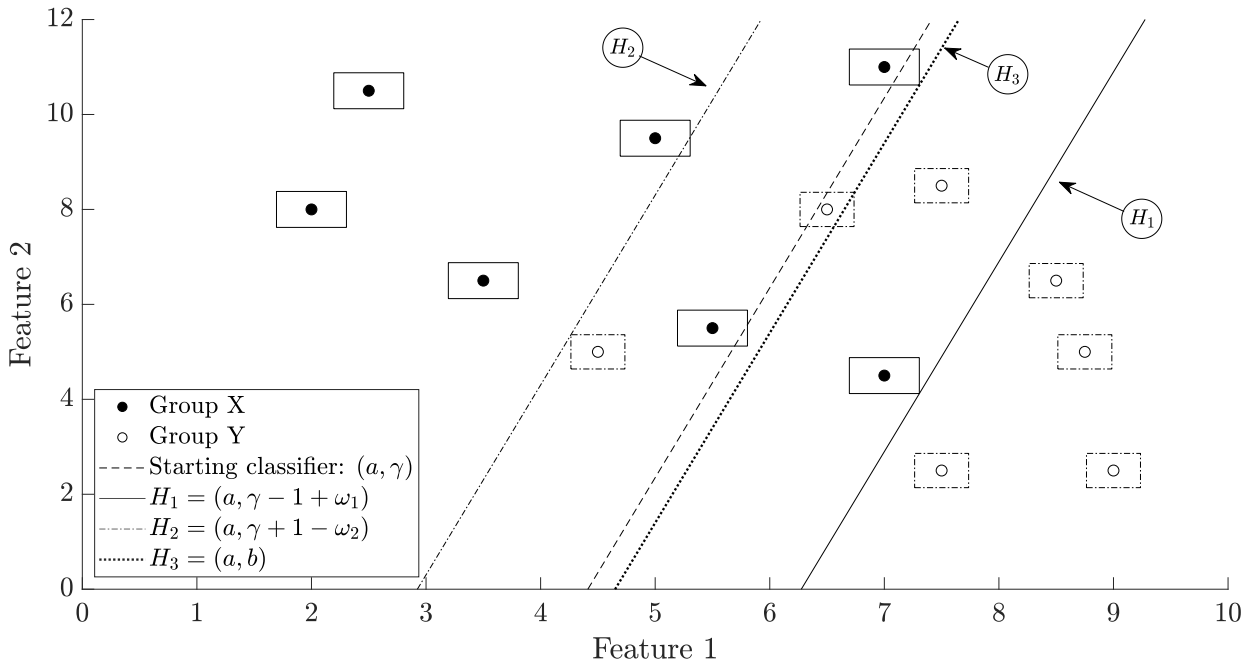


Figure 1: Input observations of groups X and Y bounded by boxes and separating hyperplanes H_1, H_2 and H_3.

After building boxes around every observation, we detect the starting hyperplane (a, γ) by means of model (1.9). We then shift it to the right and to the left, by amounts ω_1 and ω_2 respectively, to detect H_1 and H_2 such that all boxes of group X lie on one side of H_1 and all boxes of group Y lie on the opposite side of H_2. Through model (1.10), the final classifier H_3 is found such that the overall number of misclassified boxes is minimized.

Robust Support Vector Machine with Ellipsoidal Data Uncertainty

It is well known that intervals perturbations assumption can lead to overly conservative solutions. Therefore, to alleviate this drawback, in this section we propose an alternative robust formulation that considers uncertainty sets having the form of hyperellipsoids. This latter choice turns into less conservative models with respect to the hyperrectangles case since disregards those situations under which all features jointly assume extreme interval values. Moreover, this choice does not hinder the tractability of the associated reformulation, leading to a SOCP.

Let $\Sigma_{x^{(i)}}, \Sigma_{y^{(j)}} \in \mathbb{R}^{n\times n}$ be positive definite covariance matrices associated to points $x^{(i)}$ and $y^{(j)}$, respectively; further, let $\rho_X, \rho_Y \in \mathbb{R}_+$ denote the radii of the ellipsoids of groups X and Y, respectively. Then, the ellipsoidal uncertainty sets $\mathcal{U}_\mathcal{E}\left(x^{(i)}\right)$ and $\mathcal{U}_\mathcal{E}\left(y^{(j)}\right)$ centered around $x^{(i)}$ and $y^{(j)}$ are defined, respectively, as:

$$\mathcal{U}_\mathcal{E}\left(x^{(i)}\right) := \left\{ x \in \mathbb{R}^n \;\middle|\; \left(x - x^{(i)}\right)^\top \Sigma_{x^{(i)}}^{-1} \left(x - x^{(i)}\right) \leq \rho_X^2 \right\} \quad i = 1, \ldots, I, \tag{1.11}$$

$$\mathcal{U}_\mathcal{E}\left(y^{(j)}\right) := \left\{ y \in \mathbb{R}^n \;\middle|\; \left(y - y^{(j)}\right)^\top \Sigma_{y^{(j)}}^{-1} \left(y - y^{(j)}\right) \leq \rho_Y^2 \right\} \quad j = 1, \ldots, J. \tag{1.12}$$

According to [19] (see derivation in Appendix A), when uncertainty sets are described by means of (1.11)-(1.12), model (1.6) can be reformulated by the following SOCP:

$$\begin{aligned} \min_{a,\gamma,z_X,z_Y} \quad & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) \\ \text{s.t.} \quad & a^\top x^{(i)} + \rho_X \|\Sigma_{x^{(i)}}^{\frac{1}{2}} a\|_2 \leq \gamma - 1 + z_{x^{(i)}} \qquad i = 1, \ldots, I \\ & a^\top y^{(j)} - \rho_Y \|\Sigma_{y^{(j)}}^{\frac{1}{2}} a\|_2 \geq \gamma + 1 - z_{y^{(j)}} \qquad j = 1, \ldots, J \\ & z_X \geq 0,\ z_Y \geq 0, \end{aligned} \tag{1.13}$$

where:

$$\|\Sigma_{x^{(i)}}^{\frac{1}{2}} a\|_2 := \sqrt{a^\top \Sigma_{x^{(i)}} a} \quad \text{and} \quad \|\Sigma_{y^{(j)}}^{\frac{1}{2}} a\|_2 := \sqrt{a^\top \Sigma_{y^{(j)}} a}.$$

The number of continuous variables is $n + 1 + I + J$, while the number of constraints is $2(I + J)$ of which $I + J$ are non-negative. From the solution of problem (1.13), we get hyperplanes H_1 and H_2 which satisfy properties **(P1)-(P3)** with ω_1, ω_2 as in (1.4). To find H_3 we finally solve the following minimization problem using line search:

$$\begin{aligned} \min_b \quad & \sum_{i=1}^{I} \mathbb{1}\left(a^\top x^{(i)} + \rho_X \|\Sigma_{x^{(i)}}^{\frac{1}{2}} a\|_2 - b\right) + \sum_{j=1}^{J} \mathbb{1}\left(b - a^\top y^{(j)} + \rho_Y \|\Sigma_{y^{(j)}}^{\frac{1}{2}} a\|_2\right) \\ \text{s.t.} \quad & \gamma + 1 - \omega_2 \leq b \leq \gamma - 1 + \omega_1, \end{aligned} \tag{1.14}$$

where robustness fails for those points whose hyperellipsoid intersects the decision hyperplane H_3. To summarize, the geometrical interpretation of the proposed approach is sketched in Figure 2. After building ellipsoids around every observation, we detected the starting hyperplane (a, γ) by means of model (1.13). We then shift it to the right and to the left, by amounts ω_1 and ω_2 respectively, to detect H_1 and H_2 such that all ellipsoids of group X lie on one side of H_1 and all ellipsoids of group Y lie on the opposite side of H_2. Through line search of model (1.14), the final classifier H_3 is found such that the overall number of misclassified ellipsoids is minimized.

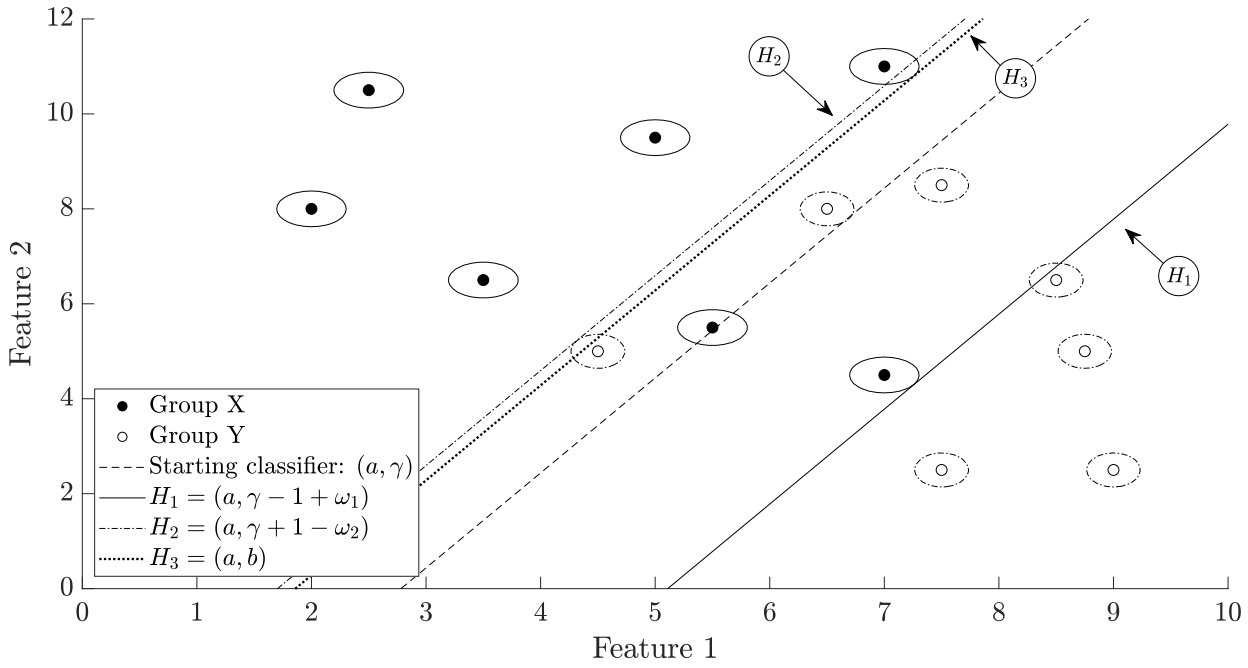


Figure 2: Input observations of groups X and Y bounded by ellipsoids and separating hyperplanes H_1, H_2 and H_3.

1.4.2 Distributionally Robust Support Vector Machine

Solutions obtained considering uncertainty sets having the form of hyperellipsoids can still be too conservative. One way to overcome this limitation would consist in resorting to other types of uncertainty sets such as polyhedral, conic, convex constraints (see [69]) or combinations of them (*e.g.*, box + ellipsoidal, box + polyhedral, box + ellipsoidal + polyhedral, see [94]). However, these specific approaches would require precise knowledge of the instances under analysis and would be highly problem-dependent. Moreover, conic uncertainty sets would require the use of conic duality while convex constraints sets the use of Fenchel duality.

Therefore, with the aim of providing progressively less conservative models that do not lose generalization ability and still protect against uncertainty, in this section we employ the most recent techniques of moment-based DRO.

In this section we treat all input observations $x^{(i)},\ y^{(j)},\ i = 1,\dots,I,\ \ j = 1,\dots,J$ as random variables, for which the exact probability distributions $\mathbb{P}^{\text{true}}_{x^{(i)}},\ i = 1,\dots,I$ and $\mathbb{P}^{\text{true}}_{y^{(j)}},\ j = 1,\dots,J$ are unknown. To hedge against uncertainty, for each input observation $x^{(i)}$ we optimize against the worst-case expectation under all possible distributions $\mathbb{P}$ belonging to the ambiguity set $\mathcal{D}\Big(x^{(i)}\Big)$. Equivalently for $y^{(j)}$ and $\mathcal{D}\Big(y^{(j)}\Big)$. Accordingly, the distributionally robust counterpart of model (1.3) can be formulated as follows:

$$\begin{aligned}
\min_{a,\gamma,z_X,z_Y} \quad & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) \\
\text{s.t.} \quad & \sup_{\mathbb{P}\in\mathcal{D}(x^{(i)})} \mathbb{E}_{\mathbb{P}}\Big[a^\top x\Big] \le \gamma - 1 + z_{x^{(i)}} \qquad i = 1,\dots,I \\
& \inf_{\mathbb{P}\in\mathcal{D}(y^{(j)})} \mathbb{E}_{\mathbb{P}}\Big[a^\top y\Big] \ge \gamma + 1 - z_{y^{(j)}} \qquad j = 1,\dots,J \\
& z_X \ge 0,\ z_Y \ge 0.
\end{aligned} \tag{1.15}$$

The choice of the specific ambiguity set $\mathcal{D}$ when modeling a problem is context dependent. This decision depends on the data being represented by the set, as well as the needs of the modeler. Hereby, a formulation that protects against uncertainty not losing generalization ability is sought and we assume that estimates are easily available from a prior statistical analysis of the uncertain data. In the following, namely, we will focus on principal directions and variance information since shared by many different distributions, while disregard higher order moments which are often unavailable (see [115]).

We consider the general moment-based ambiguity set proposed in [178] where the support and a list of partial moments describing the uncertainty are available:

$$\mathcal{D}\left(x^{(i)}\right) := \left\{ \mathbb{P}\in\mathcal{P}^n_+ \;\middle|\; \begin{array}{c} \mathbb{P}\Big(x\in\mathcal{U}_{\mathcal{B}}\Big(x^{(i)}\Big)\Big) = 1 \\ \mathbb{E}_{\mathbb{P}}\Big[g_p(x)\Big] \le \Big(\varrho_X\Big)_p \quad p = 1,\dots,n \end{array} \right\} \quad i = 1,\dots,I, \tag{1.16}$$

with $\mathcal{P}^n_+$ representing the set of probabilities distributions on $\mathbb{R}^n$. Specifically, the first constraint in set (1.16) requires every realization to be constrained within its support set $\mathcal{U}_{\mathcal{B}}\Big(x^{(i)}\Big)$ defined in (1.7). The second group of constraints in (1.16) characterizes the moments information via n functions $g_p(\cdot)$, and enforces the generalized moment $\mathbb{E}_{\mathbb{P}}\Big[g_p(x)\Big]$ not to exceed a given threshold $\Big(\varrho_X\Big)_p \in \mathbb{R}_+,\ p = 1,\dots,n$. While several generalized moment functions to describe moment information were suggested in the literature, in this chapter we employ the piecewise linear formulation

proposed by [3], which can be interpreted as the first-order deviations of the uncertain parameter with respect to the nominal value $x^{(i)}$ along certain projections $f_X^{(p)} \in \mathbb{R}^n$. Namely:

$$g_p\Big(x\Big) := \Big| {f_X^{(p)}}^\top \Big(x - x^{(i)}\Big)\Big| \quad p = 1, \ldots, n. \tag{1.17}$$

To determine projections $F_X := \Big[f_X^{(1)}, \ldots, f_X^{(n)}\Big] \in \mathbb{R}^{n\times n}$ and thresholds $\varrho_X := \Big[(\varrho_X)_1; \ldots; (\varrho_X)_n\Big] \in \mathbb{R}^n_+$ we adopt a strategy based on PCA (see [143]). The same approach holds for observations of group Y.

1. Given an unbiased estimate of the covariance matrix Σ_X:

$$\Sigma_X := \frac{\left(\sum\limits_{i=1}^{I} {x^{(i)}}^\top x^{(i)}\right) - \left(\sum\limits_{i=1}^{I} x^{(i)}\right)^\top \left(\sum\limits_{i=1}^{I} x^{(i)}\right)}{I-1}, \tag{1.18}$$

   we perform PCA onto Σ_X. Performing PCA enables capturing meaningful information about the available data. Specifically, it enables detecting the directions that manifest the most variations. We obtain:

$$\Sigma_X = F_X \cdot \Lambda_X \cdot F_X^\top, \tag{1.19}$$

   where $F_X \in \mathbb{R}^{n\times n}$ stands for the orthogonal transformation matrix and $\Lambda_X := \mathrm{diag}(\lambda_X) \in \mathbb{R}^{n\times n}_+$ is a diagonal matrix including variance information λ_X after transformation (*i.e.*, along the principal directions F_X).

2. To determine the maximum deviations allowed along the n principal directions given by thresholds $\varrho_X := \Big[(\varrho_X)_1; \ldots; (\varrho_X)_n\Big] \in \mathbb{R}^n_+$ we set:

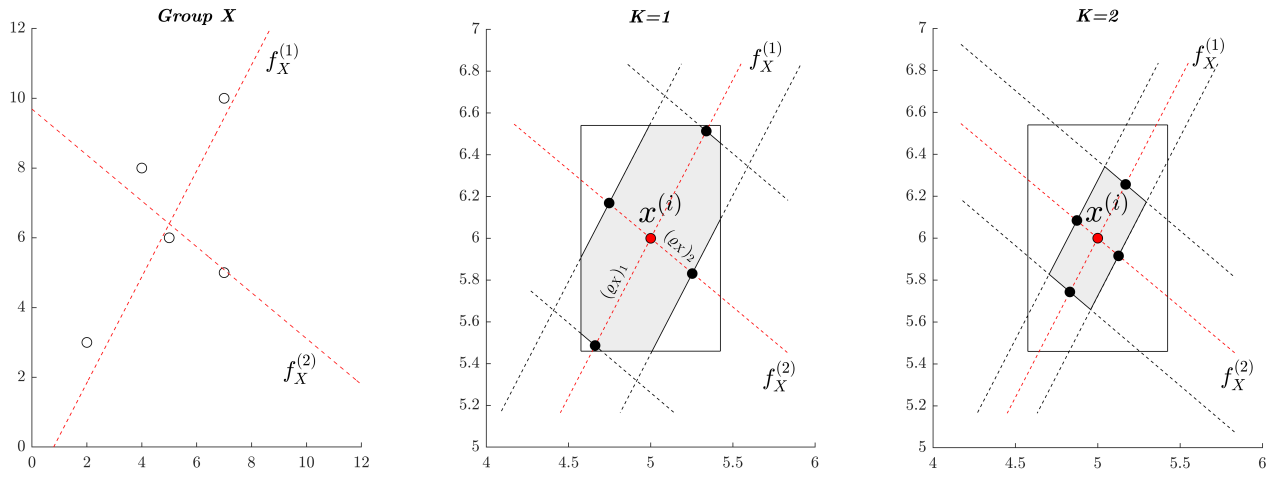
$$(\varrho_X)_p := \frac{\rho_X \sqrt{(\lambda_X)_p}}{K} \qquad p = 1, \ldots, n, \tag{1.20}$$

   where λ_X has been obtained from PCA and $K \in \mathbb{N} \setminus \{0\}$ is a scale parameter.

An attractive feature of this moment-based approach, is that one can control the model degree of conservatism simply by adjusting values of the limits $(\varrho_X)_p$, $p = 1, \ldots, n$. So, depending on specific applications and problem instances, one can opt for a more conservative strategy and tune lower values for the scale parameter K, or opt for more aggressive approaches setting higher values of K and allowing less dispersion.

Figure 3 provides a graphical representation of the procedure for a single observation $x^{(i)}$. Given a starting group X, PCA is performed to detect principal directions $F_X = \Big[f_X^{(1)}, f_X^{(2)}\Big] \in \mathbb{R}^{2\times 2}$. Then, for every observation $x^{(i)}$ the box support $\mathcal{U}_{\mathcal{B}}\Big(x^{(i)}\Big)$ is defined and limits $\varrho_X = \Big[(\varrho_X)_1; (\varrho_X)_2\Big] \in \mathbb{R}^2_+$

on variations along principal direction $f_X^{(1)}$ and direction $f_X^{(2)}$ are enforced. As shown, tuning higher values for the scale parameter K turns into a less conservative strategy compared to lower values of K, as allowing less dispersion.



(a) **(b)** **(c)**

Figure 3: Given group X (a), principal directions $f_X^{(1)}$ and $f_X^{(2)}$ are detected. For every point $x^{(i)}$, limits $(\varrho_X)_1$, $(\varrho_X)_2$ on variations along them are enforced together with the box support; K may be fixed to 1 (b) or 2 (c).

Notice that, although there are no theoretical guarantees to ensure that any moment-based ambiguity set contains the true distribution with high probability, on this purpose [43] proposed confidence regions for the mean and covariance matrix of the uncertainty using historical samples. Recently, other methods adopting data-driven robust optimization have also been suggested (see for instance [14], [113], [142]) employing hypothesis tests to determine the size of the ambiguity sets in order to ensure them to be statistically interpretable. Confidence regions can also be constructed from historical observations using resampling techniques, such as jackknifing or bootstrapping [37]. Unfortunately, these strategies cannot be trivially applied to the ambiguity set used in this work but represent an interesting future research direction to obtain a probabilistic guarantee for the true distribution to be contained in $\mathcal{D}$.

Tractable Reformulation of the Distributionally Robust Model

Model (1.15) is intractable due to the infinite number of probability distributions contained in every ambiguity set (1.16); therefore, in this section, we reformulate this problem as a tractable deterministic optimization model. Introducing the auxiliary random vector $\varphi_X := [(\varphi_X)_1; \ldots; (\varphi_X)_n] \in \mathbb{R}_+^n$ the ambiguity set given in (1.16) can be equivalently re-formulated as the projection of an extended

ambiguity set $\bar{\mathcal{D}}\left(x^{(i)}\right)$:

$$\bar{\mathcal{D}}\left(x^{(i)}\right) := \left\{ \mathbb{Q} \in \mathcal{P}_+^n \;\middle|\; \begin{array}{c} \mathbb{Q}\Big(x, \varphi_X \in \bar{\mathcal{U}}_{\mathcal{B}}\Big(x^{(i)}\Big)\Big) = 1 \\ \mathbb{E}_{\mathbb{Q}}\Big[(\varphi_X)_p\Big] \leq \Big(\varrho_X\Big)_p \quad p = 1, \ldots, n \end{array} \right\} \quad i = 1, \ldots, I, \tag{1.21}$$

with lifted support set defined as:

$$\bar{\mathcal{U}}_{\mathcal{B}}\Big(x^{(i)}\Big) := \left\{ \Big(x, \varphi_X\Big) \in \mathbb{R}^n \times \mathbb{R}_+^n \;\middle|\; \begin{array}{c} x \in \mathcal{U}_{\mathcal{B}}\Big(x^{(i)}\Big) \\ g_p(x) \leq (\varphi_X)_p \quad p = 1, \ldots, n \end{array} \right\} \quad i = 1, \ldots, I. \tag{1.22}$$

Using (1.7) and (1.17) the lifted support set (1.22) can be equally expressed as:

$$\bar{\mathcal{U}}_{\mathcal{B}}\Big(x^{(i)}\Big) = \left\{ \Big(x, \varphi_X\Big) \;\middle|\; \begin{array}{c} x \leq x^{(i)} + \rho_X \zeta_{x^{(i)}} \\ x \geq x^{(i)} - \rho_X \zeta_{x^{(i)}} \\ (\varphi_X)_p \geq 0 \quad p = 1, \ldots, n \\ {f_X^{(p)}}^\top x - {f_X^{(p)}}^\top x^{(i)} \leq (\varphi_X)_p \quad p = 1, \ldots, n \\ {f_X^{(p)}}^\top x^{(i)} - {f_X^{(p)}}^\top x \leq (\varphi_X)_p \quad p = 1, \ldots, n \end{array} \right\} \quad i = 1, \ldots, I, \tag{1.23}$$

or equivalently in matrix form:

$$\bar{\mathcal{U}}_{\mathcal{B}}\Big(x^{(i)}\Big) = \left\{ \Big(x, \varphi_X\Big) \;\middle|\; C_X x + D\varphi_X \leq h_{x^{(i)}} \right\} \quad i = 1, \ldots, I, \tag{1.24}$$

where:

$$C_X := \begin{bmatrix} \mathcal{I} \\ -\mathcal{I} \\ \mathbf{0} \\ F_X^\top \\ -F_X^\top \end{bmatrix} \in \mathbb{R}^{5n \times n}, \qquad D := \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ -\mathcal{I} \\ -\mathcal{I} \\ -\mathcal{I} \end{bmatrix} \in \mathbb{R}^{5n \times n},$$

$$h_{x^{(i)}} := \begin{bmatrix} x^{(i)} + \rho_X \zeta_{x^{(i)}} \\ -x^{(i)} + \rho_X \zeta_{x^{(i)}} \\ \mathbf{0} \\ F_X^\top x^{(i)} \\ -F_X^\top x^{(i)} \end{bmatrix} \in \mathbb{R}^{5n}.$$

An analogous reformulation can be performed for every observation of group Y and a distributionally robust formulation equivalent to (1.15) is then given as follows:

$$\begin{aligned} \min_{a,\gamma,z_X,z_Y} \quad & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) \\ \text{s.t.} \quad & \sup_{\mathbb{Q}\in\bar{\mathcal{D}}(x^{(i)})} \mathbb{E}_{\mathbb{Q}}\Big[a^\top x\Big] \le \gamma - 1 + z_{x^{(i)}} \qquad i = 1,\dots,I \\ & \inf_{\mathbb{Q}\in\bar{\mathcal{D}}(y^{(j)})} \mathbb{E}_{\mathbb{Q}}\Big[a^\top y\Big] \ge \gamma + 1 - z_{y^{(j)}} \qquad j = 1,\dots,J \\ & z_X \ge 0,\ z_Y \ge 0. \end{aligned} \tag{1.25}$$

It is worth noticing that the second group of constraints of formulation (1.25) can also be expressed as:

$$\inf_{\mathbb{Q}\in\bar{\mathcal{D}}(y^{(j)})} \mathbb{E}_{\mathbb{Q}}\Big[a^\top y\Big] \ge \gamma + 1 - z_{y^{(j)}} \quad \Leftrightarrow \quad \sup_{\mathbb{Q}\in\bar{\mathcal{D}}(y^{(j)})} \mathbb{E}_{\mathbb{Q}}\Big[-a^\top y\Big] \le -\gamma - 1 + z_{y^{(j)}} \quad j = 1,\dots,J.$$

For every $i = 1,\dots,I$, the left-hand side of the distributionally robust constraint of model (1.25) coincides with the optimal value of the following moment problem:

$$\begin{aligned} \sup_{\mathbb{Q}\in\bar{\mathcal{D}}(x^{(i)})} \mathbb{E}_{\mathbb{Q}}\Big[a^\top x\Big] = \sup_{\mathbb{Q}} \quad & \int_{\bar{\mathcal{U}}_{\mathcal{B}}(x^{(i)})} q(x,\varphi_X)\Big(a^\top x\Big)\,\mathrm{d}x\,\mathrm{d}\varphi_X \\ \text{s.t.} \quad & \int_{\bar{\mathcal{U}}_{\mathcal{B}}(x^{(i)})} q(x,\varphi_X)\,\mathrm{d}x\,\mathrm{d}\varphi_X = 1 \\ & \int_{\bar{\mathcal{U}}_{\mathcal{B}}(x^{(i)})} q(x,\varphi_X)\varphi_X\,\mathrm{d}x\,\mathrm{d}\varphi_X \le \varrho_X, \end{aligned} \tag{1.26}$$

where the decision variable is $q(x,\varphi_X)$. Introducing the multipliers $\eta_{x^{(i)}} \in \mathbb{R}$ and $\beta_{x^{(i)}} \in \mathbb{R}^n_+$, the Lagrangian reformulation of (1.26) is:

$$\sup_{\mathbb{Q}} \quad \int_{\bar{\mathcal{U}}_{\mathcal{B}}(x^{(i)})} q(x,\varphi_X)\left(a^\top x - \eta_{x^{(i)}} - \beta_{x^{(i)}}^\top \varphi_X\right)\mathrm{d}x\,\mathrm{d}\varphi_X + \eta_{x^{(i)}} + \beta_{x^{(i)}}^\top \varrho_X. \tag{1.27}$$

If there exists (x,φ_X) such that $a^\top x - \beta_{x^{(i)}}^\top \varphi_X \ge \eta_{x^{(i)}}$, then (1.27) is unbounded above because $q(x,\varphi_X) \ge 0,\ \forall (x,\varphi_X) \in \bar{\mathcal{U}}_{\mathcal{B}}\Big(x^{(i)}\Big)$. On the contrary, when $a^\top x - \beta_{x^{(i)}}^\top \varphi_X \le \eta_{x^{(i)}}$, then $\forall (x,\varphi_X) \in \bar{\mathcal{U}}_{\mathcal{B}}\Big(x^{(i)}\Big)$ the function admits a solution given by $\eta_{x^{(i)}} + \beta_{x^{(i)}}^\top \varrho_X$. The dual of (1.26) then becomes:

$$\begin{aligned} \min_{\eta_{x^{(i)}},\,\beta_{x^{(i)}}\ge 0} \quad & \eta_{x^{(i)}} + \beta_{x^{(i)}}^\top \varrho_X \\ \text{s.t.} \quad & a^\top x - \beta_{x^{(i)}}^\top \varphi_X \le \eta_{x^{(i)}} \quad \forall (x,\varphi_X) \in \bar{\mathcal{U}}_{\mathcal{B}}\Big(x^{(i)}\Big). \end{aligned} \tag{1.28}$$

The robust set of constraints of model (1.28) can be equivalently reformulated as:

$$a^\top x - \beta_{x^{(i)}}^\top \varphi_X \le \eta_{x^{(i)}} \quad \forall (x,\varphi_X) \in \bar{\mathcal{U}}_{\mathcal{B}}\Big(x^{(i)}\Big) \quad \Leftrightarrow \quad \max_{(x,\varphi_X)\,\in\,\bar{\mathcal{U}}_{\mathcal{B}}(x^{(i)})} \Big[a^\top x - \beta_{x^{(i)}}^\top \varphi_X\Big] \le \eta_{x^{(i)}}$$

where the dual of the left-hand side maximization problem is equal to:

$$\begin{array}{ll} \min\limits_{\pi_{x^{(i)}} \geq 0} & \pi_{x^{(i)}}^\top h_{x^{(i)}} \\ \text{s.t.} & C_X^\top \pi_{x^{(i)}} \geq a \\ & D^\top \pi_{x^{(i)}} \geq -\beta_{x^{(i)}}, \end{array} \tag{1.29}$$

with $\pi_{x^{(i)}} \in \mathbb{R}_+^{5n}$. Combining (1.28) with (1.29), and repeating for all $i = 1, \ldots, I$ and $j = 1, \ldots, J$, a tractable distributionally robust formulation of problem (1.3) is:

$$\begin{array}{lll} \min\limits_{a,\gamma,z_X,z_Y,\eta_X,\eta_Y,\beta_X,\beta_Y,\pi_X,\pi_Y} & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) & \\ \text{s.t.} & \eta_{x^{(i)}} + \beta_{x^{(i)}}^\top \varrho_X \leq \gamma - 1 + z_{x^{(i)}} & i = 1, \ldots, I \\ & \pi_{x^{(i)}}^\top h_{x^{(i)}} \leq \eta_{x^{(i)}} & i = 1, \ldots, I \\ & C_X^\top \pi_{x^{(i)}} \geq a & i = 1, \ldots, I \\ & D^\top \pi_{x^{(i)}} \geq -\beta_{x^{(i)}} & i = 1, \ldots, I \\ & \eta_{y^{(j)}} + \beta_{y^{(j)}}^\top \varrho_Y \leq -\gamma - 1 + z_{y^{(j)}} & j = 1, \ldots, J \\ & \pi_{y^{(j)}}^\top h_{y^{(j)}} \leq \eta_{y^{(j)}} & j = 1, \ldots, J \\ & C_Y^\top \pi_{y^{(j)}} \geq -a & j = 1, \ldots, J \\ & D^\top \pi_{y^{(j)}} \geq -\beta_{y^{(j)}} & j = 1, \ldots, J \\ & z_X \geq 0,\ z_Y \geq 0 & \\ & \pi_X \geq 0,\ \pi_Y \geq 0,\ \beta_X \geq 0,\ \beta_Y \geq 0, & \end{array} \tag{1.30}$$

where $\eta_X := \left[\eta_{x^{(1)}}; \ldots; \eta_{x^{(I)}}\right] \in \mathbb{R}^I$, $\eta_Y := \left[\eta_{y^{(1)}}; \ldots; \eta_{y^{(J)}}\right] \in \mathbb{R}^J$, $\beta_X := \left[\beta_{x^{(1)}}; \ldots; \beta_{x^{(I)}}\right] \in \mathbb{R}_+^{nI}$, $\beta_Y := \left[\beta_{y^{(1)}}; \ldots; \beta_{y^{(J)}}\right] \in \mathbb{R}_+^{nJ}$, $\pi_X := \left[\pi_{x^{(1)}}; \ldots; \pi_{x^{(I)}}\right] \in \mathbb{R}_+^{5nI}$ and $\pi_Y := \left[\pi_{y^{(1)}}; \ldots; \pi_{y^{(J)}}\right] \in \mathbb{R}_+^{5nJ}$. The number of variables of linear formulation (1.30) is $n+1+I(2+6n)+J(2+6n)$, while the number of constraints is $n + I(5+6n) + J(5+6n)$ of which $I(1+6n) + J(1+6n)$ are non-negativity constraints. From the solution of optimization problem (1.30), the hyperplanes H_1 and H_2, satisfying properties **(P1)-(P3)** are obtained. We find H_3 solving the minimization problem via line search:

$$\begin{array}{ll} \min\limits_{b} & \sum\limits_{i=1}^{I} \mathbb{1}\left(\eta_{x^{(i)}} + \beta_{x^{(i)}}^\top \varrho_X - b\right) + \sum\limits_{j=1}^{J} \mathbb{1}\left(b + \eta_{y^{(j)}} + \beta_{y^{(j)}}^\top \varrho_Y\right) \\ \text{s.t.} & \gamma + 1 - \omega_2 \leq b \leq \gamma - 1 + \omega_1. \end{array} \tag{1.31}$$

To summarize, Figure 4 provides the geometrical interpretation of the proposed approach. First, principal directions are detected for group X. Each nominal observation $x^{(i)}$ is therefore bounded by a box support and limits on $x^{(i)}$ deviations along principal directions are enforced. The same holds for every observation $y^{(j)}$ of group Y and its principal directions. Then, the starting hyperplane (a, γ) is detected by means of model (1.30), and it is shifted to the right and to the left by amounts ω_1 and ω_2, respectively, to identify H_1 and H_2. Through line search given by (1.31), the final classifier H_3 is found such that the overall number of misclassified realizations is minimized.

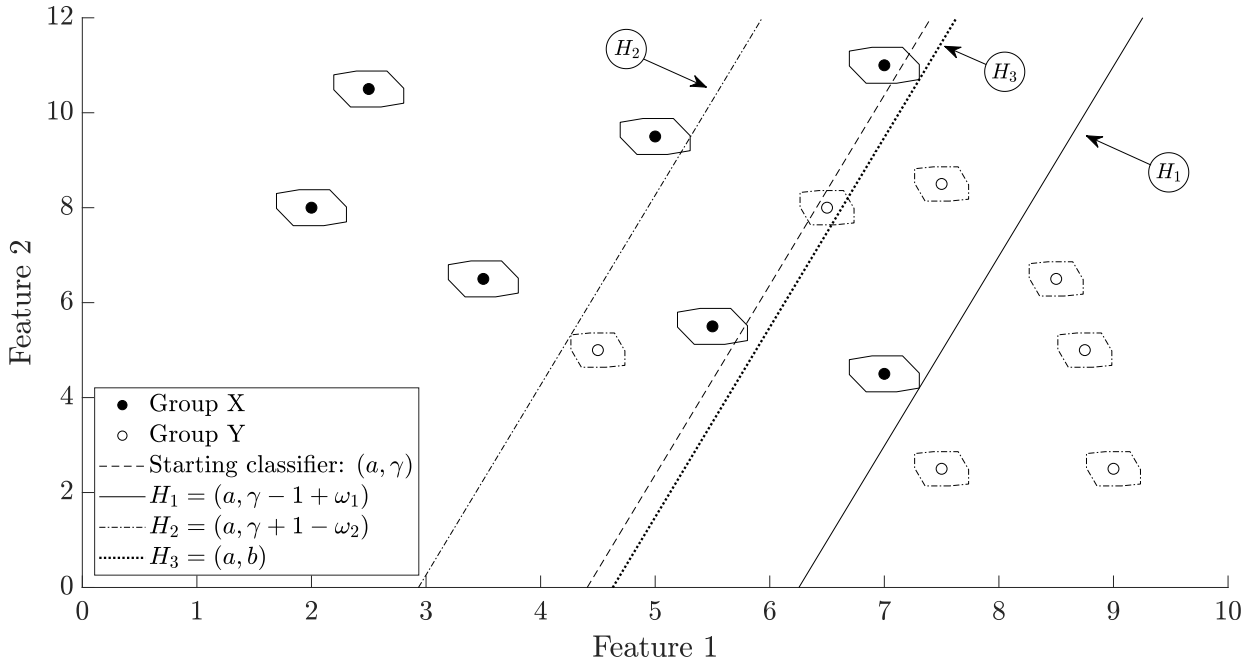


Figure 4: Input observations of groups X and Y and separating hyperplanes H_1, H_2 and H_3.

Notice that in the approaches described above we limited our attention to the the problem of linearly separating two sets of points; nonetheless, those formulations can be also applied to the multiclass separation problem (with number of classes $\kappa > 2$) by iteratively solving a sequence of two classes separation problems. Examples of these heuristic methods are the *one-versus-all* and *one-versus-one* schemes (see [2]). While the former approach detects $\kappa - 1$ classifiers, each of which solves the problem of separating points in a particular class from all the points not in that class, the latter alternative computes $\kappa(\kappa - 1)/2$ classifiers, one for every possible pair of classes.

1.5 Numerical Results

In this section, we evaluate the performance of robust and distributionally robust optimization models compared to their deterministic counterparts. The proposed SVM formulations are tested on ten

real-world databases, all of which are publicly available and can be downloaded from [46]. The data sets used are listed in Table 2, where the number of features $n \in [4, 279]$, while the number of observations considered $I + J \in [68; 4{,}435]$. For multiclass data sets (*i.e.*, Arrhythmia, Dermatology, Heart Disease, and Landsat Satellite), we adopted the *one-versus-all* scheme and detected the classifier separating the first class from the remaining ones. This was done to ensure a fair comparison with the results reported in [13], where the same approach was implemented. Clearly, models presented in Section 4 could be also used to identify the remaining $\kappa - 2$ hyperplanes under the *one-versus-all* scheme as well as the $\kappa(\kappa - 1)/2$ classifiers of the *one-versus-one* technique.

The computations have been performed on a 64-bit machine with 8 GB of RAM, a 1.8 GHz Intel i7 processor, and numerical results are obtained under MATLAB environment using MOSEK solver (version 8.1.0.72).

Data set	*Application Field*	*Observations*	*Features*	*Class Balancing*
Arrhythmia	Life Sciences	68	279	$70.59\% - 29.41\%$
Breast Cancer	Life Sciences	683	9	$65.89\% - 34.11\%$
Breast Cancer Diagnostic	Life Sciences	569	30	$62.74\% - 37.26\%$
Dermatology	Life Sciences	358	34	$68.99\% - 31.01\%$
Heart Disease	Life Sciences	297	13	$53.87\% - 46.13\%$
Parkinson	Life Sciences	195	22	$75.38\% - 24.62\%$
Climate Model Crashes	Physical Sciences	540	18	$91.48\% - 8.52\%$
Landsat Satellite	Physical Sciences	4,435	36	$95.47\% - 4.53\%$
Ozone Level Detection One	Physical Sciences	1,848	72	$96.92\% - 3.08\%$
Blood Transfusion	Business	748	4	$76.20\% - 23.80\%$

Table 2: Summary of data sets from UCI Machine Learning Repository.

For every data set, we first split the overall number of observations $(I + J)$ at our disposal into two disjoint subsets: the former (called *training set*) contains 75% of the observations (of which I_{tr} belong to the first class and J_{tr} to the second), the latter (called *testing set*) contains what is left ($I_{\text{ts}} + J_{\text{ts}}$ observations). The observations of the training set are randomly chosen with the only requirement of maintaining the original class balancing, a partition strategy known in the literature as proportional (or stratified) random sampling, *i.e.*:

$$\frac{I_{\text{tr}}}{I_{\text{tr}} + J_{\text{tr}}} = \frac{I}{I + J} \quad \text{and} \quad \frac{J_{\text{tr}}}{I_{\text{tr}} + J_{\text{tr}}} = \frac{J}{I + J}.$$

We refer the reader to [36] for a deeper discussion on proportional random sampling steady performances. For the sake of illustration, we show how to construct training sets on the data set "Breast Cancer Diagnostic". This database lists in total $I + J = 569$ observations, of which $I = 212$ represent malignant instances and $J = 357$ are observations of benign tumors. The class balancing is therefore

$62.74\% - 37.26\%$. In the generation of the training set we randomly select $I_{\text{tr}} + J_{\text{tr}} = 427$ observations (75% of 569), with $I_{\text{tr}} = 268$ belonging to the malignant group and $J_{\text{tr}} = 159$ to the benign one. By doing so the class balancing is not altered. It is worth noticing that in our computational experiments we did not implement any feature reduction algorithm (such as *feature selection* or *feature extraction*), which means that if an observation belongs to the training set, the entirety of its features will be considered during the training phase. Nonetheless, including such dimensionality reduction approaches could constitute a promising future research direction. Once the partition procedure is complete, different final separating hyperplanes are obtained solving, sequentially, the deterministic (1.2)-(1.5), box robust (1.9)-(1.10), ellipsoidal robust (1.13)-(1.14) and distributionally robust (1.30)-(1.31) formulations over the training set. Specifically, we first set the user-defined penalty parameter ν equally distributed in log space from 10^{-3} to 10^{0} with 5 discretization points, similarly to what done in [96], and $k_{\max} = 10^4$. Then, we solve the deterministic formulation under every candidate value ν_i with $i \in \{1, \ldots, 5\}$, record the hyperplane $H_3^{\nu_i}$ and compute the associated misclassification error ε^{ν_i}. The final deterministic hyperplane H_3 is chosen to be the one minimizing the misclassification error ε^{ν_i}, *i.e.*, $H_3 = H_3^{\nu^*}$ with $\nu^* \in \arg\min\{\varepsilon^{\nu_1}, \ldots, \varepsilon^{\nu_5}\}$. The same procedure is repeated for the box formulation, where we additionally set perturbation vectors $\zeta_{x^{(i)}}$ and $\zeta_{y^{(j)}}$ equal to the standard deviation vectors σ_X and σ_Y of the training groups X and Y, *i.e.*, $\zeta_{x^{(i)}} = \sigma_X,\ i = 1, \ldots, I_{\text{tr}}$ and $\zeta_{y^{(j)}} = \sigma_Y,\ j = 1, \ldots, J_{\text{tr}}$. Similarly, for the ellipsoidal robust formulation where covariance matrices are given by $\Sigma^{\frac{1}{2}}_{x^{(i)}} = \text{diag}(\sigma_X),\ i = 1, \ldots, I_{\text{tr}}$ and $\Sigma^{\frac{1}{2}}_{y^{(j)}} = \text{diag}(\sigma_Y),\ j = 1, \ldots, J_{\text{tr}}$. For the distributionally robust model, we first perform PCA on the training sets and fix the parameter $K \in \mathbb{N} \setminus \{0\}$ to tune the maximum deviations allowed along principal directions for each observation. For all our test problems, the results we get with values of K larger than 2 are worsening in terms of accuracy levels. Thus, we use $K \in \{1, 2\}$. It is worth recalling that setting $K = 1$ allows more dispersion compared to $K = 2$. For all the robust and distributionally robust formulations, procedures are repeated considering increasing levels of $\rho_X, \rho_Y \in \{0.1, 0.2, 0.3\}$. After detecting the final separating hyperplanes using the training data and under the different formulations, we measure their prediction accuracy by reporting the out-of-sample misclassification error on observations belonging to the testing set (*i.e.*, computing testing errors). In order to get stable results, the experiments are performed over 100 different compositions of the hold-out 75%-25% and results are averaged. Furthermore, the procedure is repeated under different hold-outs: 50%-50% and 25%-75%.

These perturbation assumptions for robust and distributionally robust models imply that all the sources of information about features might follow the same form of uncertainty. This is a simplifying assumption driven by the unavailability of explicit details on input data gathering, especially in the

medical field where data often comes from heterogeneous sources (*e.g.*, medical imaging, pathology reports, physician notes, genetic assays, lab results, etc.). Naturally, precise knowledge of special structure of input instances would be desired, as would allow taking into account non-homogeneous sources and would therefore lead to wiser choices of perturbations parameters.

For each formulation and every considered data set, we report in Table 3 mean out-of-sample testing errors and standard deviations[3] for the first hold-out 75%-25%. The solutions under our robust and distributionally robust approaches have intuitive practical appeal, and offer important operational insights. Foremost, by adjusting the radius parameters ρ_X, ρ_Y all robust and distributionally robust formulations are always able to improve prediction accuracies compared to their deterministic counterpart. Therefore, numerical experiments demonstrate that accounting for uncertainty proves to always be beneficial in terms of SVM predictive power.

Furthermore, it can be noted that once the optimal degree of conservatism is identified, then departing from it translates into a progressive worsening of performances. For instance, for the data set "Breast Cancer" the highest-accuracy model is the distributionally robust with $K = 1$ (out-of-sample testing error rate equal to 3.12%). It follows that opting for progressively more conservative models (ellipsoidal and box robust, in the order) gradually increases out-of-sample testing errors (3.31% and 3.36%, respectively); same conclusion can be drawn solving a less conservative model (distributionally robust with $K = 2$, with an out-of-sample testing error rate setting around 3.25%). For ease of visualization, Figure 5a reports the lowest out-of-sample testing error rate achieved by every formulation under the "Breast Cancer" data set (data from Table 3). In a similar fashion, for the data set "Heart Disease" the highest-accuracy model happens to be the ellipsoidal robust (out-of-sample testing error rate equal to 16.20%) and solving less conservative models (distributionally robust with $K = 1, 2$) gradually increases out-of-sample testing errors (16.28% and 16.50%, respectively); same conclusion can be drawn solving a more conservative model (box robust, with an out-of-sample testing error rate setting around 16.38%), see Figure 5b. The same trends are confirmed under every data set, whose plots are reported in Figure 6.

Furthermore, we compare the performance of our models with the accuracy scores reported in [13], that we consider literature benchmark results for robust classification with feature uncertainty. Such comparison highlights that our classifiers perform favorably relative to the standard SVM feature-robust formulation for the majority of the considered problems: 8 out of 10 data sets, as shown in Table 6.

[3]For each method and every data set, the best result is underlined. Overall, for every single data set, we indicate in bold the lowest out-of-sample testing error rate achieved.

	Deterministic	$\rho_X = \rho_Y$	Box RO	Ellipsoidal RO	DRO $K = 1$	DRO $K = 2$
Arrhythmia	$25.65\% \pm 0.107$	0.1	$23.65\% \pm 0.104$	$24.82\% \pm 0.102$	$23.65\% \pm 0.097$	$23.41\% \pm 0.090$
		0.2	$23.53\% \pm 0.092$	$23.06\% \pm 0.102$	$23.29\% \pm 0.095$	$23.65\% \pm 0.093$
		0.3	$23.06\% \pm 0.088$	$\mathbf{23.00\% \pm 0.089}$	$23.53\% \pm 0.090$	$23.65\% \pm 0.090$
Average CPU seconds	0.560		0.955	1.338	125.292	104.712
Breast Cancer	$3.49\% \pm 0.012$	0.1	$3.58\% \pm 0.013$	$3.53\% \pm 0.012$	$3.34\% \pm 0.011$	$3.51\% \pm 0.012$
		0.2	$3.47\% \pm 0.012$	$3.43\% \pm 0.014$	$3.24\% \pm 0.012$	$3.33\% \pm 0.011$
		0.3	$3.36\% \pm 0.013$	$3.31\% \pm 0.012$	$\mathbf{3.12\% \pm 0.012}$	$3.25\% \pm 0.012$
Average CPU seconds	0.244		0.324	1.118	7.627	7.449
Breast Cancer Diagnostic	$4.89\% \pm 0.015$	0.1	$3.90\% \pm 0.016$	$4.45\% \pm 0.015$	$4.66\% \pm 0.016$	$4.70\% \pm 0.015$
		0.2	$3.97\% \pm 0.015$	$\mathbf{3.89\% \pm 0.015}$	$4.06\% \pm 0.016$	$4.12\% \pm 0.017$
		0.3	$4.04\% \pm 0.015$	$4.09\% \pm 0.014$	$4.10\% \pm 0.015$	$4.23\% \pm 0.015$
Average CPU seconds	0.261		0.330	0.622	20.383	17.094
Dermatology	$0.56\% \pm 0.008$	0.1	$0.34\% \pm 0.007$	$0.34\% \pm 0.008$	$0.21\% \pm 0.007$	$0.31\% \pm 0.008$
		0.2	$0.24\% \pm 0.007$	$0.19\% \pm 0.006$	$0.21\% \pm 0.007$	$0.30\% \pm 0.008$
		0.3	$0.20\% \pm 0.006$	$\mathbf{0.13\% \pm 0.005}$	$0.29\% \pm 0.008$	$0.35\% \pm 0.009$
Average CPU seconds	0.357		0.608	1.072	9.958	9.331
Heart Disease	$16.68\% \pm 0.039$	0.1	$16.38\% \pm 0.037$	$16.38\% \pm 0.036$	$16.28\% \pm 0.039$	$16.50\% \pm 0.041$
		0.2	$17.81\% \pm 0.045$	$\mathbf{16.20\% \pm 0.035}$	$16.61\% \pm 0.039$	$16.88\% \pm 0.037$
		0.3	$21.57\% \pm 0.043$	$16.49\% \pm 0.037$	$18.16\% \pm 0.040$	$17.32\% \pm 0.040$
Average CPU seconds	0.228		0.269	1.002	3.319	3.238
Parkinson	$14.13\% \pm 0.043$	0.1	$13.38\% \pm 0.032$	$\mathbf{13.00\% \pm 0.037}$	$14.31\% \pm 0.039$	$14.29\% \pm 0.039$
		0.2	$14.42\% \pm 0.031$	$13.21\% \pm 0.033$	$13.75\% \pm 0.038$	$14.00\% \pm 0.038$
		0.3	$15.50\% \pm 0.037$	$13.79\% \pm 0.033$	$13.60\% \pm 0.035$	$13.94\% \pm 0.036$
Average CPU seconds	0.212		0.314	0.611	2.851	2.811
Climate Model Crashes	$4.99\% \pm 0.016$	0.1	$4.80\% \pm 0.013$	$4.67\% \pm 0.013$	$\mathbf{4.34\% \pm 0.017}$	$4.41\% \pm 0.013$
		0.2	$6.01\% \pm 0.011$	$4.48\% \pm 0.013$	$4.38\% \pm 0.016$	$4.76\% \pm 0.019$
		0.3	$8.50\% \pm 0.004$	$4.61\% \pm 0.014$	$5.18\% \pm 0.015$	$5.62\% \pm 0.021$
Average CPU seconds	0.252		0.317	0.540	8.234	8.002
Landsat Satellite	$0.43\% \pm 0.001$	0.1	$0.44\% \pm 0.002$	$0.42\% \pm 0.001$	$0.46\% \pm 0.002$	$\mathbf{0.36\% \pm 0.001}$
		0.2	$0.42\% \pm 0.002$	$0.39\% \pm 0.001$	$0.37\% \pm 0.001$	$0.40\% \pm 0.001$
		0.3	$0.43\% \pm 0.002$	$0.41\% \pm 0.002$	$0.37\% \pm 0.001$	$0.49\% \pm 0.001$
Average CPU seconds	0.906		1.041	1.250	1,142.028	1,128.582
Ozone Level Detection One	$6.19\% \pm 0.013$	0.1	$5.32\% \pm 0.012$	$4.97\% \pm 0.009$	$4.80\% \pm 0.012$	$3.15\% \pm 0.001$
		0.2	$4.84\% \pm 0.008$	$3.86\% \pm 0.007$	$4.11\% \pm 0.007$	$3.06\% \pm 0.001$
		0.3	$4.57\% \pm 0.008$	$3.81\% \pm 0.004$	$3.72\% \pm 0.006$	$\mathbf{3.06\% \pm 0.001}$
Average CPU seconds	0.628		0.819	0.993	683.121	677.719
Blood Transfusion	$23.49\% \pm 0.026$	0.1	$23.21\% \pm 0.010$	$23.28\% \pm 0.013$	$22.87\% \pm 0.013$	$23.02\% \pm 0.015$
		0.2	$23.43\% \pm 0.007$	$\mathbf{22.55\% \pm 0.010}$	$22.78\% \pm 0.014$	$22.80\% \pm 0.014$
		0.3	$23.53\% \pm 0.008$	$23.36\% \pm 0.005$	$23.46\% \pm 0.021$	$23.09\% \pm 0.016$
Average CPU seconds	0.255		0.305	0.927	7.158	7.040

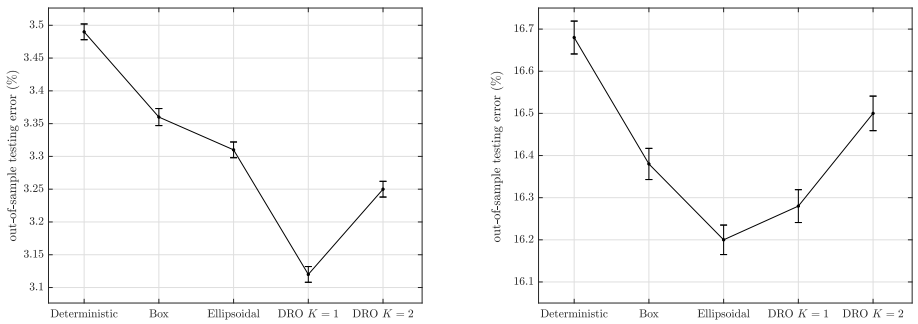
Table 3: Average out-of-sample testing errors and standard deviations over 100 runs of the deterministic, robust and distributionally robust models, for the different considered data sets. Hold-out 75%-25%.

	Deterministic	$\rho_X = \rho_Y$	Box RO	Ellipsoidal RO	DRO $K=1$	DRO $K=2$
Arrhythmia	$26.76\% \pm 0.080$	0.1	$25.44\% \pm 0.075$	$25.76\% \pm 0.075$	$25.44\% \pm 0.075$	$25.50\% \pm 0.076$
		0.2	$25.12\% \pm 0.072$	$25.00\% \pm 0.075$	$25.62\% \pm 0.076$	$25.59\% \pm 0.075$
		0.3	$\mathbf{24.00\% \pm 0.069}$	$24.32\% \pm 0.069$	$25.65\% \pm 0.074$	$25.59\% \pm 0.073$
Average CPU seconds	0.433		0.681	0.792	36.200	32.092
Breast Cancer	$3.70\% \pm 0.010$	0.1	$3.56\% \pm 0.007$	$3.54\% \pm 0.008$	$3.42\% \pm 0.008$	$3.64\% \pm 0.008$
		0.2	$3.43\% \pm 0.007$	$3.40\% \pm 0.008$	$3.40\% \pm 0.008$	$3.55\% \pm 0.008$
		0.3	$3.32\% \pm 0.007$	$3.29\% \pm 0.008$	$\mathbf{3.28\% \pm 0.007}$	$3.38\% \pm 0.008$
Average CPU seconds	0.240		0.242	0.846	5.036	4.984
Breast Cancer Diagnostic	$5.60\% \pm 0.011$	0.1	$5.08\% \pm 0.012$	$4.64\% \pm 0.010$	$5.18\% \pm 0.013$	$5.25\% \pm 0.012$
		0.2	$5.19\% \pm 0.014$	$4.45\% \pm 0.011$	$5.32\% \pm 0.012$	$5.42\% \pm 0.012$
		0.3	$5.29\% \pm 0.017$	$\mathbf{4.41\% \pm 0.010}$	$5.44\% \pm 0.012$	$5.39\% \pm 0.012$
Average CPU seconds	0.260		0.302	0.593	9.382	9.229
Dermatology	$0.90\% \pm 0.008$	0.1	$0.57\% \pm 0.008$	$0.81\% \pm 0.012$	$0.42\% \pm 0.006$	$0.49\% \pm 0.008$
		0.2	$0.41\% \pm 0.006$	$0.46\% \pm 0.007$	$0.46\% \pm 0.007$	$0.47\% \pm 0.008$
		0.3	$0.32\% \pm 0.005$	$\mathbf{0.21\% \pm 0.004}$	$0.47\% \pm 0.008$	$0.50\% \pm 0.008$
Average CPU seconds	0.246		0.255	0.617	8.198	7.283
Heart Disease	$18.74\% \pm 0.027$	0.1	$18.11\% \pm 0.022$	$18.38\% \pm 0.027$	$18.23\% \pm 0.029$	$18.49\% \pm 0.028$
		0.2	$19.37\% \pm 0.033$	$17.82\% \pm 0.025$	$18.56\% \pm 0.031$	$18.41\% \pm 0.029$
		0.3	$24.57\% \pm 0.053$	$\mathbf{17.82\% \pm 0.024}$	$19.47\% \pm 0.032$	$18.68\% \pm 0.029$
Average CPU seconds	0.227		0.256	0.521	2.478	2.347
Parkinson	$15.62\% \pm 0.036$	0.1	$14.28\% \pm 0.031$	$14.55\% \pm 0.027$	$15.55\% \pm 0.039$	$15.36\% \pm 0.037$
		0.2	$15.57\% \pm 0.030$	$14.47\% \pm 0.025$	$15.05\% \pm 0.035$	$15.29\% \pm 0.036$
		0.3	$17.26\% \pm 0.041$	$\mathbf{14.23\% \pm 0.023}$	$15.22\% \pm 0.037$	$15.19\% \pm 0.034$
Average CPU seconds	0.206		0.244	0.504	1.845	1.788
Climate Model Crashes	$5.61\% \pm 0.015$	0.1	$5.02\% \pm 0.010$	$5.21\% \pm 0.012$	$5.34\% \pm 0.014$	$5.42\% \pm 0.015$
		0.2	$6.19\% \pm 0.011$	$4.90\% \pm 0.011$	$5.33\% \pm 0.014$	$5.52\% \pm 0.014$
		0.3	$8.46\% \pm 0.003$	$\mathbf{4.87\% \pm 0.010}$	$5.73\% \pm 0.014$	$6.37\% \pm 0.011$
Average CPU seconds	0.239		0.254	0.517	5.518	5.433
Landsat Satellite	$0.51\% \pm 0.002$	0.1	$0.47\% \pm 0.001$	$0.44\% \pm 0.001$	$0.48\% \pm 0.001$	$\mathbf{0.41\% \pm 0.001}$
		0.2	$0.47\% \pm 0.001$	$0.43\% \pm 0.001$	$0.47\% \pm 0.002$	$0.43\% \pm 0.001$
		0.3	$0.48\% \pm 0.002$	$0.44\% \pm 0.001$	$0.42\% \pm 0.001$	$0.48\% \pm 0.002$
Average CPU seconds	0.654		0.684	0.846	522.252	484.864
Ozone Level Detection One	$6.27\% \pm 0.018$	0.1	$5.93\% \pm 0.015$	$5.81\% \pm 0.014$	$3.71\% \pm 0.009$	$5.81\% \pm 0.015$
		0.2	$5.79\% \pm 0.014$	$5.11\% \pm 0.011$	$3.07\% \pm 0.001$	$4.47\% \pm 0.005$
		0.3	$5.71\% \pm 0.014$	$4.42\% \pm 0.008$	$\mathbf{3.06\% \pm 0.001}$	$4.36\% \pm 0.005$
Average CPU seconds	0.489		0.502	0.667	310.425	291.384
Blood Transfusion	$23.93\% \pm 0.016$	0.1	$23.15\% \pm 0.005$	$22.94\% \pm 0.007$	$23.48\% \pm 0.011$	$23.53\% \pm 0.009$
		0.2	$23.45\% \pm 0.004$	$\mathbf{22.88\% \pm 0.006}$	$23.55\% \pm 0.013$	$23.59\% \pm 0.008$
		0.3	$23.60\% \pm 0.003$	$23.44\% \pm 0.005$	$23.88\% \pm 0.021$	$23.91\% \pm 0.016$
Average CPU seconds	0.216		0.230	0.492	4.952	4.935

Table 4: Average out-of-sample testing errors and standard deviations over 100 runs of the deterministic, robust and distributionally robust models, for the different considered data sets. Hold-out 50%-50%.

	Deterministic	$\rho_X = \rho_Y$	Box RO	Ellipsoidal RO	DRO $K = 1$	DRO $K = 2$
Arrhythmia	$33.18\% \pm 0.068$	0.1	$31.12\% \pm 0.074$	$31.98\% \pm 0.070$	$32.16\% \pm 0.083$	$32.06\% \pm 0.082$
		0.2	$29.82\% \pm 0.069$	$31.37\% \pm 0.073$	$32.12\% \pm 0.082$	$32.14\% \pm 0.082$
		0.3	$\mathbf{29.04\% \pm 0.064}$	$30.29\% \pm 0.075$	$32.06\% \pm 0.082$	$32.24\% \pm 0.082$
Average CPU seconds	0.423		0.535	0.780	7.646	6.182
Breast Cancer	$4.81\% \pm 0.013$	0.1	$4.35\% \pm 0.011$	$4.25\% \pm 0.009$	$4.47\% \pm 0.013$	$4.35\% \pm 0.010$
		0.2	$3.96\% \pm 0.008$	$4.03\% \pm 0.009$	$4.31\% \pm 0.011$	$4.22\% \pm 0.010$
		0.3	$\mathbf{3.74\% \pm 0.007}$	$3.81\% \pm 0.008$	$3.91\% \pm 0.009$	$3.94\% \pm 0.009$
Average CPU seconds	0.226		0.240	0.608	2.621	2.552
Breast Cancer Diagnostic	$6.35\% \pm 0.013$	0.1	$5.16\% \pm 0.009$	$5.17\% \pm 0.010$	$6.02\% \pm 0.011$	$6.02\% \pm 0.011$
		0.2	$5.19\% \pm 0.012$	$4.96\% \pm 0.009$	$6.00\% \pm 0.012$	$6.04\% \pm 0.012$
		0.3	$6.03\% \pm 0.015$	$\mathbf{4.94\% \pm 0.009}$	$6.04\% \pm 0.011$	$6.06\% \pm 0.012$
Average CPU seconds	0.250		0.258	0.575	3.586	3.541
Dermatology	$2.03\% \pm 0.014$	0.1	$1.12\% \pm 0.010$	$1.02\% \pm 0.011$	$0.66\% \pm 0.007$	$0.74\% \pm 0.007$
		0.2	$0.76\% \pm 0.008$	$0.65\% \pm 0.008$	$0.69\% \pm 0.008$	$0.74\% \pm 0.008$
		0.3	$0.54\% \pm 0.006$	$\mathbf{0.46\% \pm 0.006}$	$0.72\% \pm 0.008$	$0.76\% \pm 0.007$
Average CPU seconds	0.215		0.239	0.590	2.148	2.130
Heart Disease	$20.90\% \pm 0.027$	0.1	$20.50\% \pm 0.030$	$20.48\% \pm 0.026$	$20.05\% \pm 0.028$	$20.22\% \pm 0.027$
		0.2	$21.14\% \pm 0.036$	$19.72\% \pm 0.025$	$20.42\% \pm 0.029$	$20.60\% \pm 0.030$
		0.3	$23.90\% \pm 0.045$	$\mathbf{19.67\% \pm 0.025}$	$20.81\% \pm 0.035$	$20.75\% \pm 0.032$
Average CPU seconds	0.222		0.229	0.492	1.414	1.269
Parkinson	$17.92\% \pm 0.044$	0.1	$16.67\% \pm 0.036$	$16.55\% \pm 0.039$	$17.87\% \pm 0.041$	$18.12\% \pm 0.046$
		0.2	$17.87\% \pm 0.039$	$16.39\% \pm 0.039$	$18.29\% \pm 0.043$	$17.92\% \pm 0.042$
		0.3	$19.82\% \pm 0.043$	$\mathbf{16.26\% \pm 0.035}$	$18.47\% \pm 0.046$	$17.87\% \pm 0.041$
Average CPU seconds	0.206		0.227	0.461	1.099	1.037
Climate Model Crashes	$7.61\% \pm 0.022$	0.1	$\mathbf{6.40\% \pm 0.016}$	$7.26\% \pm 0.020$	$7.50\% \pm 0.024$	$7.88\% \pm 0.023$
		0.2	$6.61\% \pm 0.012$	$6.90\% \pm 0.018$	$7.84\% \pm 0.028$	$9.04\% \pm 0.032$
		0.3	$8.06\% \pm 0.009$	$6.55\% \pm 0.015$	$8.51\% \pm 0.032$	$9.04\% \pm 0.036$
Average CPU seconds	0.226		0.239	0.491	2.352	2.304
Landsat Satellite	$0.59\% \pm 0.002$	0.1	$0.51\% \pm 0.001$	$0.50\% \pm 0.002$	$0.51\% \pm 0.002$	$0.51\% \pm 0.002$
		0.2	$0.49\% \pm 0.001$	$0.49\% \pm 0.002$	$0.50\% \pm 0.002$	$0.50\% \pm 0.002$
		0.3	$0.53\% \pm 0.002$	$0.48\% \pm 0.002$	$\mathbf{0.47\% \pm 0.002}$	$0.49\% \pm 0.002$
Average CPU seconds	0.414		0.423	0.701	122.338	118.907
Ozone Level Detection One	$6.43\% \pm 0.012$	0.1	$6.26\% \pm 0.017$	$6.20\% \pm 0.018$	$4.25\% \pm 0.010$	$5.12\% \pm 0.017$
		0.2	$5.08\% \pm 0.014$	$5.33\% \pm 0.017$	$3.32\% \pm 0.006$	$4.74\% \pm 0.017$
		0.3	$5.06\% \pm 0.011$	$4.70\% \pm 0.010$	$\mathbf{3.08\% \pm 0.001}$	$4.73\% \pm 0.019$
Average CPU seconds	0.393		0.404	0.612	97.110	96.562
Blood Transfusion	$24.09\% \pm 0.014$	0.1	$23.43\% \pm 0.006$	$23.17\% \pm 0.005$	$23.42\% \pm 0.028$	$23.61\% \pm 0.011$
		0.2	$23.57\% \pm 0.004$	$\mathbf{23.03\% \pm 0.008}$	$23.25\% \pm 0.036$	$24.13\% \pm 0.021$
		0.3	$23.66\% \pm 0.002$	$23.45\% \pm 0.005$	$23.43\% \pm 0.045$	$25.42\% \pm 0.026$
Average CPU seconds	0.214		0.223	0.490	2.622	2.587

Table 5: Average out-of-sample testing errors and standard deviations over 100 runs of the deterministic, robust and distributionally robust models, for the different considered data sets. Hold-out 25%-75%.



(a) Breast Cancer **(b) Heart Disease**

Figure 5: Lowest out-of-sample testing error rates over changes of ρ_X, ρ_Y per formulation under the data sets: (a) Breast Cancer; (b) Heart Disease. Vertical error bars represents standard errors. Data of Table 3.

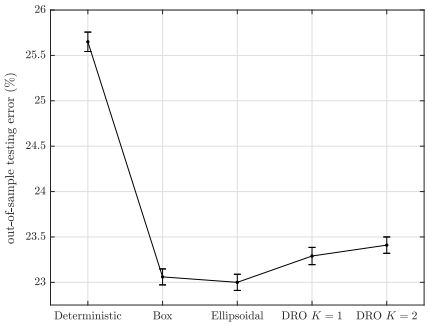
Overall, experimental results show that the robustification of the deterministic formulation (1.2)-(1.5) proposed in [96] leads to more powerful decision boundaries compared to classical approaches.

Data set	*Table 3*	*Ref. [13]*
Arrhythmia	**23.00%**	29.23%
Breast Cancer	**3.12%**	4.26%
Breast Cancer Diagnostic	**3.89%**	4.04%
Dermatology	**0.13%**	1.13%
Heart Disease	**16.20%**	16.61%
Parkinson	**13.00%**	16.41%
Climate Model Crashes	4.34%	**4.07%**
Landsat Satellite	**0.36%**	1.87%
Ozone Level Detection One	3.06%	**2.98%**
Blood Transfusion	**22.55%**	23.62%

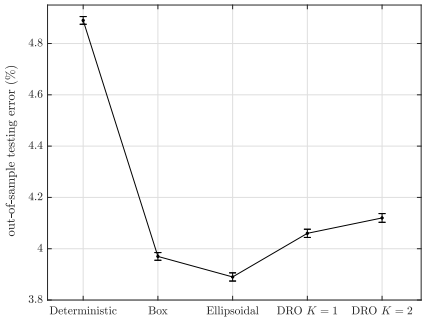
Table 6: Out-of-sample testing error rates comparison. Data of Table 3 against accuracy scores from [13]. For each data set, we indicate in bold the lowest out-of-sample testing error rate achieved.

In Tables 4 and 5 we present the results under the 50%-50% and 25%-75% hold-outs. We observe that, with respect to the 75%-25% hold-out, robust and distributionally robust methods significantly outperform the deterministic formulation in terms of prediction accuracy with improvements that increase as the training sample size decreases. This confirms that robust and distributionally robust methods produce high-quality classifiers when the uncertainty increases during the training phase, and therefore their ability to recover the truth from the data increases. To this end, Table 7 shows the robust and distributionally robust improvements in out-of-sample testing errors over their deterministic counterpart. For every data set, we report the best performing model under each hold-out with its average out-of-sample testing error, which we refer to as τ^*. We also compute the improvement ratios δ as follows:

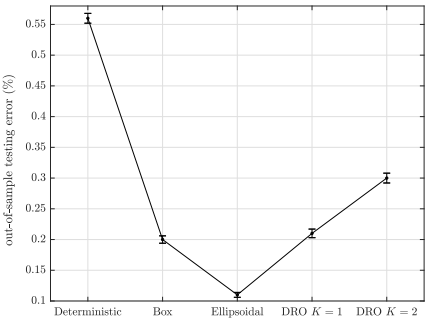
$$\delta := \frac{\tau^{\text{det}} - \tau^*}{\tau^{\text{det}}}$$



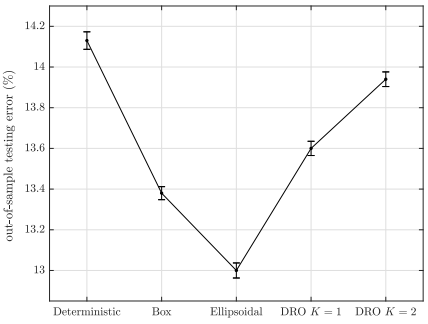
(a) Arrhythmia



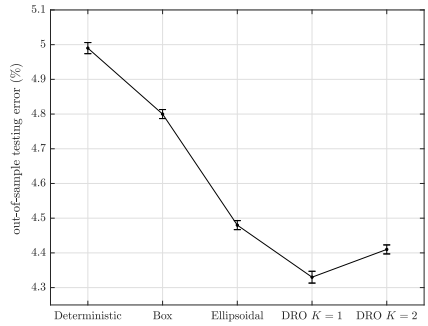
(b) Breast Cancer Diagnostic



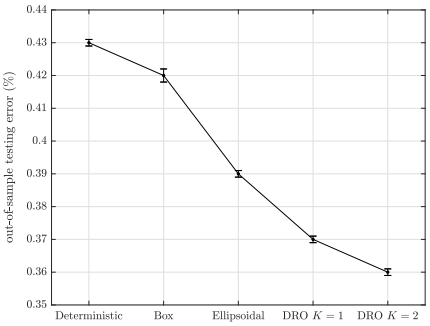
(c) Dermatology



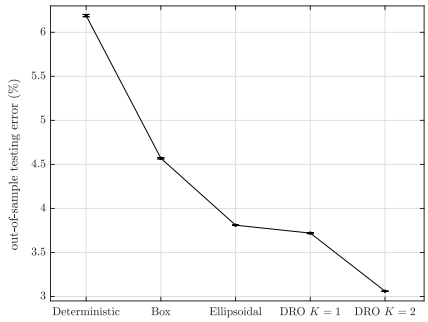
(d) Parkinson



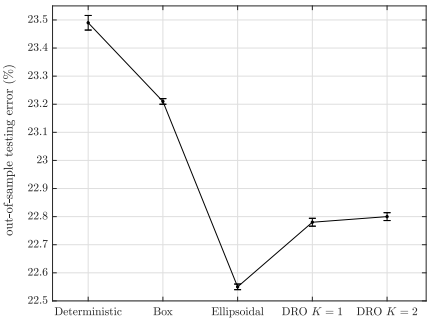
(e) Climate Model Crashes



(f) Landsat Satellite



(g) Ozone Level Detection One



(h) Blood Transfusion

Figure 6: Lowest out-of-sample testing error rates over changes of ρ_X, ρ_Y per formulation under the data sets: (a) Arrhythmia; (b) Breast Cancer Diagnostic; (c) Dermatology; (d) Parkinson; (e) Climate Model Crashes; (f) Landsat Satellite; (g) Ozone Level Detection One; (h) Blood Transfusion. Vertical error bars represents standard errors. Graphics refer to data of Table 3.

	75%-25%		50%-50%		25%-75%	
	BEST MODEL	p-value	BEST MODEL	p-value	BEST MODEL	p-value
Arrhythmia	Ellipsoidal	3.26E-02	Box	5.69E-04	Box	3.60E-09
Breast Cancer	DRO $K=1$	1.38E-02	DRO $K=1$	1.60E-05	Box	3.49E-11
Breast Cancer Diagnostic	Ellipsoidal	2.54E-11	Ellipsoidal	1.39E-17	Ellipsoidal	1.54E-15
Dermatology	Ellipsoidal	9.50E-06	Ellipsoidal	4.66E-14	Ellipsoidal	1.20E-17
Heart Disease	Ellipsoidal	8.73E-02	Ellipsoidal	2.84E-04	Ellipsoidal	2.26E-05
Parkinson	Ellipsoidal	2.10E-04	Ellipsoidal	5.15E-05	Ellipsoidal	1.66E-04
Climate Model Crashes	DRO $K=1$	8.70E-03	Ellipsoidal	1.50E-06	Box	7.51E-07
Landsat Satellite	DRO $K=2$	6.07E-05	DRO $K=2$	6.20E-04	DRO $K=1$	1.38E-05
Ozone Level Detection One	DRO $K=2$	3.30E-43	DRO $K=1$	3.98E-32	DRO $K=1$	5.52E-47
Blood Transfusion	Ellipsoidal	8.02E-05	Ellipsoidal	1.70E-09	Ellipsoidal	1.18E-11

Table 8: p-values of the best performing robust model on hold-outs 75%-25%, 50%-50%, 25%-75%.

with τ^{det} being the average out-of-sample testing error of the deterministic model of each data set.

	75%-25%			50%-50%			25%-75%		
	BEST MODEL	τ^*	δ	BEST MODEL	τ^*	δ	BEST MODEL	τ^*	δ
Arrhythmia	Ellipsoidal	23.00%	10.32%	Box	24.00%	10.33%	Box	29.04%	12.47%
Breast Cancer	DRO $K=1$	3.12%	10.54%	DRO $K=1$	3.28%	11.30%	Box	3.74%	22.25%
Breast Cancer Diagnostic	Ellipsoidal	3.89%	20.45%	Ellipsoidal	4.41%	21.25%	Ellipsoidal	4.94%	22.20%
Dermatology	Ellipsoidal	0.13%	76.79%	Ellipsoidal	0.21%	76.67%	Ellipsoidal	0.46%	77.34%
Heart Disease	Ellipsoidal	16.20%	2.88%	Ellipsoidal	17.82%	4.91%	Ellipsoidal	19.67%	5.43%
Parkinson	Ellipsoidal	13.00%	7.96%	Ellipsoidal	14.23%	8.91%	Ellipsoidal	16.26%	9.29%
Climate Model Crashes	DRO $K=1$	4.34%	13.03%	Ellipsoidal	4.87%	13.19%	Box	6.40%	15.90%
Landsat Satellite	DRO $K=2$	0.36%	17.17%	DRO $K=2$	0.41%	18.58%	DRO $K=1$	0.47%	20.88%
Ozone Level Detection One	DRO $K=2$	3.06%	50.52%	DRO $K=1$	3.06%	51.27%	DRO $K=1$	3.08%	52.09%
Blood Transfusion	Ellipsoidal	22.55%	4.00%	Ellipsoidal	22.88%	4.39%	Ellipsoidal	23.03%	4.40%

Table 7: Robust improvements with respect to the deterministic model on hold-outs 75%-25%, 50%-50%, 25%-75%.

To highlight the statistical significance of our results, under each data set, we also display the p-values for the best performing method against the result of its deterministic counterpart, see Table 8. Reported p-values are calculated performing a paired-sample t-test under the assumption of the null hypothesis that the difference in accuracy of the deterministic and robust or distributionally robust classifier is zero. All results are found to be significant with respect to the typical 5% threshold, except for the "Heart Disease" with 75%-25% hold-out that starts rejecting the null hypothesis at a significance level equal to 8.73%. We recall that the smaller the p-value, the more significant is the difference in accuracy.

In Figure 7, for the considered hold-outs, we report the number of data sets for which every formulation gave the lowest out-of-sample testing error rate. Histograms clearly underline that for greater training sets (75% of that overall data) less conservative models tend to perform better with respect to the most conservative model (*i.e.*, box). Conversely, as the cardinality of the training set progressively diminishes (down to 25% of that overall data, under the most extreme circumstance) best predictions are

obtained using more conservative models. We recall, indeed, that distributionally robust formulations represent more aggressive approaches, since they extract relevant information on the given data and exploit it to define per group perturbation directions.

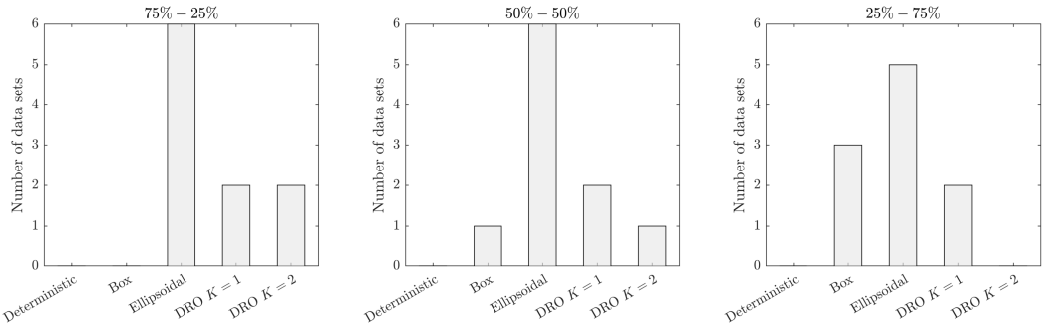


Figure 7: Number of data sets for which every formulation gave the lowest out-of-sample testing error rate. Data of Tables 3, 4, and 5.

To provide advice to final users on when it is valuable to use robust rather than distributionally robust models in practical applications, Figure 8 plots the best performing method against the dimension of the training data set (25%, 50%, 75%). Additionally, the circle sizes are proportional to the values of robust improvements δ from Table 7.

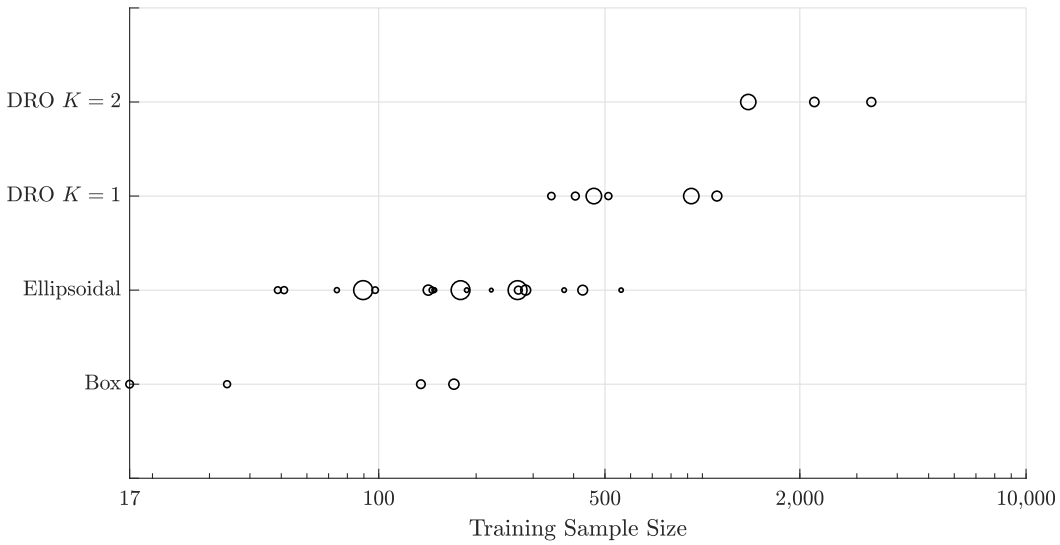


Figure 8: Best performing models versus dimension of the training samples. Data are from Tables 3, 4, and 5. Horizontal axis is in log-scale.

We observe that distributionally robust models outperform robust formulations for the majority of the training sets in the region of high dimensionality (*i.e.*, training sample size with more than approximately 500 observations). On the contrary, robust models beat distributionally robust methods for the majority of training sets in the region of low dimensionality (*i.e.*, training sample size with less than approximately 500 observations). It is also insightful to compare distributionally robust formulations

with distinct degrees of conservatism. Indeed, we observe that more aggressive distributionally robust models (obtained setting $K = 2$) outperform more conservative formulations for the training sets with more than approximately $1{,}000$ observations. This confirms our previous conclusion related to the value of the information at disposal during the training phase, which makes opting for more aggressive models when data might be considered more trustworthy.

Tables 3, 4, and 5 also present the average CPU time (in seconds) required to find a solution for each method over 500 runs. Numerical results show that solutions for the deterministic and robust formulations were obtained within a few seconds. Contrariwise, higher computational times are observed for the distributionally robust formulations, especially for data sets with larger numbers of observations (*e.g.*, Landsat Satellite and Ozone Level Detection One) or greater number of features (*e.g.*, Arrhythmia). In these cases, deterministic as well as box and ellipsoidal RO methods are several orders of magnitude faster. Therefore, we can conclude that a satisfactory trade-off between accuracy and performing speed is provided by ellipsoidal formulations.

The crucial takeaway message of this work is that hedging against uncertainty in the input observations via robust and distributionally robust approaches offer substantial benefits compared to deterministic formulations and can improve the classification accuracy up to 77.34% (see Table 7). Furthermore, accuracy results recorded by robust and distributionally robust classifiers are more stable, showing less variability when compared to the separators obtained under the deterministic approach. The proposed formulations, overall, allow finding a trade-off between increasing the average performance accuracy and protecting against uncertainty, enabling the decision maker to chose the strategy that is appropriate for each decision making setting.

1.6 Conclusions

In this work we have presented new optimization models for SVM under uncertainty. Since the consideration of uncertainty is critical to enhance classifiers predictive power, we have formulated robust models with uncertainty regions in the form of both box and ellipsoids, and distributionally robust models that enforce limits on first-order deviations of each input observations along principal directions. We have conducted extensive computational tests on real-world databases with several fields of application. The proposed robust and distributionally robust models have proved to have stronger prediction ability compared to their corresponding deterministic one. Numerical experiments have also shown that as the information at disposal during the training step increases, better prediction accuracy is achieved with more aggressive models (such as the distributionally robust) that account for

a higher degree of information. Contrariwise, assuming to have information when such is unreliable has led to poor results. Indeed, as the training sample size gets smaller and the available amount of data is scarce, the utility of implementing distributionally robust approaches has decreased and more conservative models (*i.e.*, box and ellipsoidal robust formulations) have performed better. Overall, taking uncertainty into account during the training phase –to reasonable extents– has always enhanced the classifier's predictive power. Further research activity could be focused on different interesting directions such as: 1) distributionally robust formulations with ellipsoidal supports for SVM; 2) consider the use of different kernel functions for non-linear classifiers under uncertainty; 3) consider the uncertainty in the labels; 4) extend SVM formulations to DRO with different ambiguity sets, such as the ones induced by ϕ-divergences and Wasserstein distance. On this last regard, the use of distance-based approaches such the Wasserstein-1 metric [115] and an appropriate choice of robustness level could guarantee the inclusion of the true distribution within the ambiguity set with a prescribed level of confidence.

Chapter 2. Bounds for Multistage Mixed-Integer Distributionally Robust Optimization

In collaboration with Güzin Bayraksan[1], Francesca Maggioni[2] and Ming Yang [3]. Released in a different version on *SIAM Journal on Optimization*.

2.1 Introduction

Multistage stochastic programming has been widely used to solve important problems arising in various fields including finance [26], transportation [11], energy [120, 173] and the environment [189], among others. Despite their wide applicability, this class of problems suffers from two main issues. First, traditional models assume that the underlying stochastic process that governs the uncertain parameters is known. This is rarely true in real life. Second, multistage stochastic programs—particularly those involving mixed-integer variables and nonlinear terms—are notoriously difficult to solve. To alleviate the first issue, *Distributionally Robust Optimization* (DRO) can be used, where the assumed-known distribution is replaced by an ambiguity set of distributions [144]. Unfortunately, the resulting multistage problem is still extremely challenging to solve and can become even more challenging depending on the type of DRO used. Due to the exponential growth of the problem in the number stages, approximation techniques that provide bounds on the optimal value for multistage DRO problems can be very useful in practice. In this situation, easy-to-compute bounds and approximations are desirable. This chapter investigates easy-to-compute lower bounds (for minimization problems) through scenario grouping and convolution for a class of multistage DRO formed using ϕ-divergences and Wasserstein distance. Most of the literature on DRO focuses on static, two-stage, or chance-constrained settings [128], and there is relatively little work on multistage DRO. Many of these works investigate different ways of forming ambiguity sets in the multistage setting, which can be more complicated relative to the static/two-stage setting [123, 144]. Moment-based [17, 148, 184], nested Wasserstein [121], modified χ^2 distance [122], general ϕ-divergences [117], L_∞-norm [74], Wasserstein [50] and ∞-Wasserstein distance [16] have been examined to form multistage DRO. Most papers assume linear models with

[1]Department of Integrated Systems Engineering, The Ohio State University
Columbus, OH, USA. Bayraksan.1@osu.edu

[2]Department of Management, Information & Production Engineering, University of Bergamo
Dalmine, BG, IT. Francesca.Maggioni@unibg.it

[3]Department of Integrated Systems Engineering, The Ohio State University
Columbus, OH, USA. Yang.3149@osu.edu

continuous decision variables, except for [16, 184], which consider mixed-integer decision variables. Our bounds do not assume any problem structure such as linearity, convexity, and continuity. Majority of the existing works also focus on solution methods through nested Benders' decomposition or its sampling-based variant, stochastic dual dynamic programming [50, 74, 117, 122, 184]. Linear decision rules have also been used to approximately solve these problems [16, 17].

The approach presented in this work divides the sample space into subgroups. The subgroups, being of smaller size, can be solved more efficiently. They can be solved either by the traditional expected-value objective approach or a DRO approach. Then, the optimal values of subgroups can be combined, *e.g.*, using DRO, to form lower bounds on the optimal value. It turns out that not all combinations of subgroups optimal values yield lower bounds. We provide conditions on ways to combine optimal values of the subgroups to obtain lower bounds for ambiguity sets formed via many commonly used ϕ-divergences and Wasserstein distance. The Wasserstein setting is more complicated since requires an appropriate distance between subgroups. We define such a distance between subgroups to ensure lower bounds, and discuss how to apply these bounds in the multistage setting.

2.1.1 Related Work

Bounding techniques have a rich history in the stochastic programming literature, and these have been successfully applied to traditional multistage stochastic programs with expected-value objectives. For instance, [51] considers two-stage bounds-based distributional approximations for multistage stochastic linear programs (*i.e.*, moment-based approximations derived as solutions to certain generalized moment problems), relaxing the nonanticipativity constraints. Nonanticipativity is regained progressively via a disaggregation procedure. In [52], the authors propose tight upper and lower bounds to stochastic convex programs with random right-hand sides. Using a constraint aggregation procedure, a group of stages from the end of the multistage stochastic program are aggregated to form a single stage, and error bounds are developed. In [84], the author elaborates an approximation scheme that integrates stage-aggregation and discretization through coarsening of sigma-algebras to ensure computational tractability, while providing deterministic error bounds.

Bounds for multistage stochastic linear programs via scenario tree decomposition were proposed for the first time in [102], by solving pair subproblems, measuring the quality of the deterministic solution, and introducing rolling horizon measures. In [103], the authors extend the bounding approach of [25, 102, 137] for stochastic multistage mixed-integer linear programs, solving a sequence of group subproblems made by a subset of reference scenarios plus a subset of scenarios from the finite support. They show the monotonicity of the chains of lower bounds in terms of the cardinality of reference

scenarios and of the remaining scenarios in each subgroup. A scalable bounding framework for general multistage stochastic programs, extending the work of [137], has been investigated in [138]. This framework scales well with problem size and obtains high-quality solutions within a reasonable time frame.

An alternative approach to bound the original multistage stochastic program is to construct two approximating trees, a lower tree and an upper tree, the solutions of which lead to upper and lower bounds for the optimal value of the original *continuous* problem. Results in this direction were first obtained by [59], followed by [58, 83]. In [83], barycentric discretizations are adopted in a more general setting for convex multistage stochastic programs with a generalized non-convex dependence on the random variables. In [105], the authors generalize the bounding ideas of [58, 59, 83] to not necessarily Markovian scenario processes and derive valid lower and upper bounds for the convex case. They construct new discrete probability measures directly from the simulated data of the whole scenario process based on the concepts of first order and convex order stochastic dominance.

Bounds for risk-averse multistage mixed-integer stochastic programs via scenario tree decomposition were first proposed by [104] and [106]. In particular, [104] considers multistage convex problems with concave risk functional applied to the total cost over the planning horizon. New refinement chains of lower bounds are constructed, where each bound can be computed by solving sets of group subproblems less complex than the original one, and recalculating the probabilities of each scenario in the group accordingly. A monotonically nondecreasing behavior in the cardinality of scenarios of each subproblem is proved. In [106], the authors consider a dynamic risk functional in the objective function, formed by a convex combination of mean and *Conditional Value-at-Risk* (mean-CVaR). Lower bounds by using convolution of mean-CVaRs with different parameters are obtained through various scenario partition strategies, and a solution algorithm for mean-CVaR multistage mixed-integer stochastic problems is provided; see also [107] for algorithmic use of these bounds.

2.1.2 Summary of Contributions

This work, to the best of our knowledge for the first time, introduces new *lower-bound* (LB) criteria for multistage DRO through scenario tree decomposition. Upper bounds are also examined. Our work is similar in spirit to [104, 106], but we consider a large class of DRO formed on finite scenario trees, where the ambiguity sets are constructed using a ϕ-divergence or Wasserstein distance on a finite support [50]. We provide conditions on how the optimal values of subgroup problems can be combined to yield lower bounds by directly using the ambiguity sets. We first present our results in the two-stage setting and then discuss how to apply these LB criteria in the multistage setting. Finally,

we investigate the effectiveness of the proposed bounds on a multistage mixed-integer production planning problem. The proposed approach has the important advantage to split a given problem into independent scenario groups. This allows to tackle problems for which simple linear relaxations leave large optimality gaps, problems lacking special structure, and large-scale multistage problems that are not solvable by commercial solvers.

The chapter is organized as follows. In Section 2.2, basic facts on multistage DRO, the construction of ambiguity sets via ϕ-divergences and Wasserstein distance as well as their relation to risk-averse optimization are recalled. Section 2.3 contains the main results of the chapter, namely the LB criteria for ϕ-divergences and Wasserstein distance, their extension to multistage problems and upper bounds. Section 2.4 reports numerical results on a multistage mixed-integer production problem and provides a discussion of insights gained. Section 2.5 concludes the chapter and outlines future research directions.

2.2 Basic Facts and Notation

2.2.1 Multistage DRO

We consider a finite-horizon sequential decision making problem under uncertainty. Decisions are made at discrete stages $t \in \mathcal{T} := \{0, 1, \ldots, T\}$, where T denotes the planning horizon. The decision process begins with initial decision $\mathbf{x}_0 \in \mathbb{R}_+^{n_0} \times \mathbb{Z}_+^{n_0'}$ at stage $t = 0$, called the *first-stage* decision, and is followed by sequential decisions $\mathbf{x}_t \in \mathbb{R}_+^{n_t} \times \mathbb{Z}_+^{n_t'}$ at stages $t \in \mathcal{T} \setminus \{0\}$. The history of the decision process, at a given point in time, is denoted by $\mathbf{x}^t := (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_t)$, $t \in \mathcal{T}$. The uncertainty is described by a random process $\boldsymbol{\xi} := \{\boldsymbol{\xi}_0, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_T\} \in \mathbb{R}^{d_0} \times \ldots \times \mathbb{R}^{d_T}$, where $\boldsymbol{\xi}_t$, $t \in \mathcal{T}$ is defined on a probability space $(\Omega_t, \mathscr{F}_t, Q_t)$ with support $\Omega_t \in \mathbb{R}^{d_t}$, σ-algebra $\mathscr{F}_t$ (with $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \ldots \subseteq \mathscr{F}_T = \mathscr{F}$) and nominal distribution Q_t. We assume $\boldsymbol{\xi}_0$ is a degenerate random vector (*i.e.*, constant) and $\boldsymbol{\xi}$ is a random parameter evolving as a discrete-time stochastic process with finite support. The history of the random process up to stage t is denoted by $\boldsymbol{\xi}^t := (\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_t)$, $t \in \mathcal{T}$. The following represents the nested formulation of a multistage DRO (see [144]):

$$
\begin{aligned}
\min_{\mathbf{x}_0 \in \mathcal{X}_0(\boldsymbol{\xi}_0)} \; c_0(\mathbf{x}_0, \boldsymbol{\xi}_0) \; + \; & \max_{P_1 | \boldsymbol{\xi}^0 \in \mathcal{P}_{1|\boldsymbol{\xi}^0}} \; \mathbb{E}_{P_1|\boldsymbol{\xi}^0} \Bigg[\min_{\mathbf{x}_1 \in \mathcal{X}_1(\mathbf{x}_0, \boldsymbol{\xi}_1)} c_1(\mathbf{x}_1, \boldsymbol{\xi}_1) \\
& + \; \max_{P_2 | \boldsymbol{\xi}^1 \in \mathcal{P}_{2|\boldsymbol{\xi}^1}} \; \mathbb{E}_{P_2|\boldsymbol{\xi}^1} \Bigg[\min_{\mathbf{x}_2 \in \mathcal{X}_2(\mathbf{x}_1, \boldsymbol{\xi}_2)} c_2(\mathbf{x}_2, \boldsymbol{\xi}_2) \\
& + \; \ldots \\
& + \; \max_{P_T | \boldsymbol{\xi}^{T-1} \in \mathcal{P}_{T|\boldsymbol{\xi}^{T-1}}} \; \mathbb{E}_{P_T|\boldsymbol{\xi}^{T-1}} \Big[\min_{\mathbf{x}_T \in \mathcal{X}_T(\mathbf{x}_{T-1}, \boldsymbol{\xi}_T)} c_T(\mathbf{x}_T, \boldsymbol{\xi}_T) \Big] \Bigg] \Bigg]
\end{aligned}
\tag{2.1}
$$

where the mixed-integer first-stage feasibility set is given by $\mathcal{X}_0 \subseteq \mathbb{R}^{n_0}_+ \times \mathbb{Z}^{n'_0}_+$ and, for $t \in \mathcal{T} \setminus \{0\}$, $\mathcal{X}_t : \mathbb{R}^{n_{t-1}}_+ \times \mathbb{Z}^{n'_{t-1}}_+ \times \mathbb{R}^{d_t} \to \mathbb{R}^{n_t}_+ \times \mathbb{Z}^{n'_t}_+$ are $\mathscr{F}_t$-measurable mixed-integer point-to-set mappings. The possibly nonlinear cost functions are given by $c_0 : \mathbb{R}^{n_0}_+ \times \mathbb{Z}^{n'_0}_+ \to \mathbb{R}$ in the first stage and by $c_t : \mathbb{R}^{n_t}_+ \times \mathbb{Z}^{n'_t}_+ \times \mathbb{R}^{d_t} \to \mathbb{R}$ in stages $t \in \mathcal{T} \setminus \{0\}$, which are $\mathscr{F}_t$-measurable. We assume all relevant optimization problems in the chapter have finite optimal solutions. Set $\mathcal{P}_{t|\boldsymbol{\xi}^{t-1}}$ denotes the conditional ambiguity set at period $t \in \mathcal{T} \setminus \{0\}$, conditioned on the history $\boldsymbol{\xi}^{t-1}$, and it is defined as:

$$\mathcal{P}_{t|\boldsymbol{\xi}^{t-1}} := \left\{ P_{t|\boldsymbol{\xi}^{t-1}} \in \mathcal{M}(\Omega_{t|\boldsymbol{\xi}^{t-1}}) : \ \Delta(P_{t|\boldsymbol{\xi}^{t-1}}, Q_{t|\boldsymbol{\xi}^{t-1}}) \leq \rho_t \right\}, \tag{2.2}$$

where $\rho_t \geq 0$ is a given *radius*, also called the *level of robustness*. Above, $\mathcal{M}(\Omega_{t|\boldsymbol{\xi}^{t-1}})$ represents a class of probability distributions defined on the support $\Omega_{t|\boldsymbol{\xi}^{t-1}}$, $Q_{t|\boldsymbol{\xi}^{t-1}}$ denotes the nominal conditional probability measure at stage t, $t \in \mathcal{T} \setminus \{0\}$, conditioned on the history of the process $\boldsymbol{\xi}^{t-1}$, and $\Delta(\cdot, \cdot)$ denotes a measure of similarity or distance between $P_{t|\boldsymbol{\xi}^{t-1}}$ and $Q_{t|\boldsymbol{\xi}^{t-1}}$. We are interested in building ambiguity sets using existing data via ϕ-divergences and Wasserstein distance, which we recall in the next sections. Before we do so, let us define notation that is used throughout the chapter.

2.2.2 Scenario Tree and Nominal Probability Notation

Because we assume $\boldsymbol{\xi}$ has finite support, the information structure can be described in the form of a *scenario tree* $\mathfrak{T}$ with $T+1$ levels (stages). Let Ω_t be the set of ordered nodes of the tree $\mathfrak{T}$ at stage $t \in \mathcal{T}$ and let $\Omega := \Omega_1 \times \ldots \times \Omega_T$. By assumption, we have a discrete number $|\Omega_t|$ of nodes at each stage $t \in \mathcal{T}$. Each stage-t ($t > 0$) node n is connected to a unique node at stage $t-1$, called *ancestor* and denoted $a(n)$. Similarly, each stage-t ($t < T$) node n is connected to nodes at stage $t+1$ called *successors* or *children*, where $\mathscr{B}(n)$ denotes the set of children nodes of n. With $q_{a(n),n}$ we denote the conditional nominal probability of the random process at node n given its history up to the ancestor node $a(n)$. A *scenario* $\omega_i, \ i = 1, \ldots, |\Omega_T|$ is a path through nodes from the root node at $t = 0$ to a leaf node at $t = T$. We indicate with q_{ω_i} the probability of a scenario ω_i passing through nodes $n_0, n_1, \ldots, n_T$ (where n_t, $t = 0, \ldots, T$ represent generic nodes at stage t), defined as $q_{\omega_i} := q_{n_0,n_1} \cdot q_{n_1,n_2} \cdot \ldots \cdot q_{n_{T-1},n_T}$. We also indicate with q^n_t the nominal probability of node n at stage t. So, if node n at stage t is reachable through node n_0 at stage 0, node n_1 at stage $1, \ldots,$ node n_{t-1} at stage $t-1$, then $q^n_t := q_{n_0,n_1} \cdot q_{n_1,n_2} \cdot \ldots \cdot q_{n_{t-1},n_t}$. Moreover, $\sum_{n \in \Omega_t} q^n_t = 1, \ t \in \mathcal{T}$ and $\sum_{m \in \mathscr{B}(n)} q_{n,m} = 1, \ n \in \Omega_t, \ \ t = 0, \ldots, T-1$.

Assumption. For simplicity, from here until Section 2.3.5, we consider two-stage DRO and only point to changes for multistage case. That is, we set $T = 1$ in (2.1), drop $\boldsymbol{\xi}_0$ as it is a constant, and let $\boldsymbol{\xi}_1 \equiv \boldsymbol{\xi}$ be defined on a probability space $(\Omega, \mathscr{F}, Q)$ with finite support $\Omega := \{\omega_1, \omega_2, \ldots, \omega_{|\Omega|}\}$, filtration $\mathscr{F}$

and probability Q. The probability of scenario $\omega_i \in \Omega$ can be specified as $q_{\omega_i} \geq 0$ with $\sum_{i=1}^{|\Omega|} q_{\omega_i} = 1$. Similarly, we simply use P with probability of scenario ω_i as $p_{\omega_i} \geq 0$ to define ambiguity set (2.2). So, (2.2) becomes:

$$\mathcal{P} = \left\{ P : \ \Delta(P, Q) \leq \rho, \ \sum_{i=1}^{|\Omega|} p_{\omega_i} = 1, \ p_{\omega_i} \geq 0, \ \omega_i \in \Omega \right\}. \tag{2.3}$$

From now on, we also use shorthand notation $[m]$ denote the set $\{1, 2, \ldots, m\}$. So, $i \in [|\Omega|]$ is equivalent to $i \in \{1, \ldots, |\Omega|\}$.

2.2.3 ϕ-Divergences

For this class of ambiguity sets, Δ in (2.2) is given by

$$\Delta_\phi(P, Q) := \sum_{i=1}^{|\Omega|} q_{\omega_i} \phi\left(\frac{p_{\omega_i}}{q_{\omega_i}}\right),$$

where the convex ϕ-divergence function $\phi(u) \geq 0$ takes value 0 when both $p_{\omega_i} > 0$ and $q_{\omega_i} > 0$ have the same value; *i.e.*, $\phi(1) = 0$. When $q_{\omega_i} = 0$, it holds that $0 \cdot \phi(p_{\omega_i}/0) = p_{\omega_i} \lim_{u \to \infty} (\phi(u)/u)$ and $0 \cdot \phi(0/0) = 0$. Accordingly, ambiguity set $\mathcal{P}_{t|\boldsymbol{\xi}^{t-1}}$ in (2.2) can be built using some of the well-known ϕ-divergences described in Tables 9 and 10. These include *Variation Distance* (VD) and J-divergence, along with two families of ϕ-divergences, namely, the *Cressie-Read* (CR) power divergence family and the χ-divergence family of order $a > 1$. CR power divergence family includes some of the most widely used ϕ-divergences as a special case—*e.g.*, the modified χ^2 distance and the *Kullback-Leibler* (KL) divergence—when its parameter θ takes specific values or when the limit of θ tends to 0 or 1. These special cases are listed in Table 10.

Divergence	$\phi(u)$	$\phi(u), u \geq 0$	$\Delta_\phi(P, Q)$
Variation Distance	ϕ_v	$\lvert u - 1 \rvert$	$\sum \lvert p_{\omega_i} - q_{\omega_i} \rvert$
Cressie-Read Power Divergence	ϕ^θ_{CR}	$\frac{1-\theta+\theta u - u^\theta}{\theta(1-\theta)}, \theta \neq 0, 1$	$\frac{1-\sum p^\theta_{\omega_i} q^{1-\theta}_{\omega_i}}{\theta(1-\theta)}, \theta \neq 0, 1$
J-Divergence	ϕ_J	$(u-1)\log u$	$\sum (p_{\omega_i} - q_{\omega_i}) \log\left(\frac{p_{\omega_i}}{q_{\omega_i}}\right)$
χ-*Divergence of order* $a > 1$	ϕ^a_χ	$\lvert u - 1 \rvert^a$	$\sum q_{\omega_i} \left\lvert 1 - \frac{p_{\omega_i}}{q_{\omega_i}} \right\rvert^a$

Table 9: Common ϕ-divergences.

Equivalence of the well-known divergences in Table 10 and the CR power divergence family in Table 9 is achieved when the radius ρ_t in the ambiguity set (2.2) formed via a divergence in Table 10 is set to an adjusted value $c \cdot \rho^\theta_{t,CR}$, where $\rho^\theta_{t,CR}$ is the radius of the CR divergence in Table 9. Values of coefficient c corresponding to certain θ are listed in the last column of Table 10. For example, when

the radius of the modified χ^2 distance, denoted $\rho_{t,m\chi^2}$, equals $2 \cdot \rho_{t,CR}^{\theta=2}$, where $\rho_{t,CR}^{\theta=2}$ represents the radius formed via CR power divergence with $\theta = 2$, the two ambiguity sets are equivalent.

θ	*Corresponding Divergence*	$\phi(u)$	$\phi(u), u \geq 0$	$\Delta_\phi(P, Q)$	$\phi_{CR}^{\theta}(u)$	c
2	*Modified χ^2 Distance*	$\phi_{m\chi^2}$	$(u-1)^2$	$\sum \frac{(p_{\omega_i}-q_{\omega_i})^2}{q_{\omega_i}}$	$\frac{1}{2}(u^2-2u+1) = \frac{1}{2}(u-1)^2$	2
$\frac{1}{2}$	*Hellinger Distance*	ϕ_H	$(\sqrt{u}-1)^2$	$\sum(\sqrt{p_{\omega_i}} - \sqrt{q_{\omega_i}})^2$	$4(\frac{1}{2}+\frac{1}{2}t-\sqrt{u}) = 2(1-\sqrt{u})^2$	$\frac{1}{2}$
-1	*χ^2 Distance*	ϕ_{χ^2}	$\frac{1}{u}(u-1)^2$	$\sum \frac{(p_{\omega_i}-q_{\omega_i})^2}{p_{\omega_i}}$	$\frac{1}{2}(-2+u+\frac{1}{u}) = \frac{1}{2}(\sqrt{u}-\frac{1}{\sqrt{u}})^2$	2
$\to 1$	*Kullback-Leibler Divergence*	ϕ_{KL}	$u \log u - u + 1$	$\sum p_{\omega_i} \log\left(\frac{p_{\omega_i}}{q_{\omega_i}}\right)$	$u(\log u - 1) + 1$	1
$\to 0$	*Burg Entropy*	ϕ_B	$-\log u + u - 1$	$\sum q_{\omega_i} \log\left(\frac{q_{\omega_i}}{p_{\omega_i}}\right)$	$-\log u + u - 1$	1

Table 10: Some special cases of CR power divergence family. Kullback-Leibler divergence and Burg entropy are obtained by taking the limit of θ to 1 and 0, respectively.

2.2.4 Wasserstein Distance

Let $\boldsymbol{\eta} \in (\Omega, \mathscr{F}, Q)$ be a random variable taking values $(\eta_{\omega_1}, \ldots, \eta_{\omega_{|\Omega|}})$. We quantify distributions P close to nominal distribution Q taking values on support Ω via Wasserstein distance (see [50]), where Δ in (2.2) is defined by

$$\begin{aligned}
\Delta_W(P, Q) := \min_{\{z_{\omega_i,\omega_j}\}_{i,j=1}^{|\Omega|}} \quad & \sum_{j=1}^{|\Omega|}\sum_{i=1}^{|\Omega|} d_{\omega_i,\omega_j} z_{\omega_i,\omega_j} \\
\text{s.t.} \quad & \sum_{i=1}^{|\Omega|} z_{\omega_i,\omega_j} = q_{\omega_j} \qquad j = 1, \ldots, |\Omega| \\
& \sum_{j=1}^{|\Omega|} z_{\omega_i,\omega_j} = p_{\omega_i} \qquad i = 1, \ldots, |\Omega| \\
& z_{\omega_i,\omega_j} \geq 0 \qquad i, j = 1, \ldots, |\Omega|
\end{aligned}$$

with $d_{\omega_i,\omega_j} := ||\eta_{\omega_i} - \eta_{\omega_j}||_\varsigma$ a distance between the two scenarios ω_i and ω_j using ς-norm (*e.g.*, $\varsigma \in \{1, 2, \infty\}$). Ambiguity set $\mathcal{P}_{t|\boldsymbol{\xi}^{t-1}}$ in (2.2) can be built accordingly.

2.2.5 Relation to Risk-Averse Optimization

Because the ambiguity sets considered in this chapter are compact convex subsets of (conditional) probability measures and optimal values are assumed to be real-valued, DRO is equivalent to *Risk-Averse Stochastic Optimization* (RASO) with the objective function expressed by a coherent risk measure; see *e.g.,* [6, 130, 144]. Let us recall coherent risk measures. Let $\mathcal{Z} := \mathfrak{L}_\infty(\Omega, \mathscr{F}, Q)$ be the space of bounded and $\mathscr{F}$-measurable random variables with respect to sample space Ω and probability distribution Q, and let $\boldsymbol{\eta} \in \mathcal{Z}$ be a random variable with values $(\eta_{\omega_1}, \ldots, \eta_{\omega_{|\Omega|}})$. First defined by [4], a function $\mathcal{R}(\boldsymbol{\eta}) : \mathcal{Z} \to \mathbb{R}$ is called a *coherent measure of risk* if it satisfies the following properties:

1. *Convexity*: $\mathcal{R}(\lambda\boldsymbol{\eta}^1 + (1-\lambda)\boldsymbol{\eta}^2) \leq \lambda\mathcal{R}(\boldsymbol{\eta}^1) + (1-\lambda)\mathcal{R}(\boldsymbol{\eta}^2)$ for all $\boldsymbol{\eta}^1, \boldsymbol{\eta}^2 \in \mathcal{Z}$, $\lambda \in [0,1]$;
2. *Monotonocity*: $\boldsymbol{\eta}^1 \geq \boldsymbol{\eta}^2$ implies $\mathcal{R}(\boldsymbol{\eta}^1) \geq \mathcal{R}(\boldsymbol{\eta}^2)$ for all $\boldsymbol{\eta}^1, \boldsymbol{\eta}^2 \in \mathcal{Z}$;
3. *Translation Equivariance*: $\mathcal{R}(\boldsymbol{\eta} + \lambda) = \mathcal{R}(\boldsymbol{\eta}) + \lambda$ for all $\boldsymbol{\eta} \in \mathcal{Z}$, $\lambda \in \mathbb{R}$;
4. *Positive Homogeneity*: $\mathcal{R}(\lambda \cdot \boldsymbol{\eta}) = \lambda \cdot \mathcal{R}(\boldsymbol{\eta})$ for all $\boldsymbol{\eta} \in \mathcal{Z}$, $\lambda > 0$.

Coherent measures of risk can be interpreted as worst-case expectations from a compact convex set of probability measures through their dual representation:

$$\mathcal{R}(\boldsymbol{\eta}) := \max_{P \in \mathcal{P}} \mathbb{E}_P\Big[\boldsymbol{\eta}\Big].$$

Therefore, it follows that a RASO can be re-written as a DRO:

$$\min_{\mathbf{x}} \mathcal{R}\Big(c(\mathbf{x}, \boldsymbol{\xi})\Big) := \min_{\mathbf{x}} \max_{P \in \mathcal{P}} \mathbb{E}_P\Big[c(\mathbf{x}, \boldsymbol{\xi})\Big].$$

The above conclusion is straightforwardly extended to the multistage setting by recursively using conditional ambiguity sets, which we recall in Section 2.3.5; see *e.g.,* [134, 135, 144] for nested coherent composite risk measures in multistage setting.

2.3 Lower Bounds for DRO

The aim of this section is to provide lower bounds for DRO formed by ϕ-divergences and the Wasserstein distance. For this purpose, instead of dealing with the whole sample space Ω, whose large cardinality may lead to computational concerns, the lower bound is achieved by dividing the sample space Ω into subgroups that can be considered separately and then combining the solutions of the subgroups using DRO with possibly another radius. To perform such a division, we consider the approaches presented in [104] and summarized below.

2.3.1 Dissecting the Scenario Tree

We construct a collection of m_l subsets, each of cardinality l, of the space Ω:

$$\left(\Omega_1^{(l)}, \Omega_2^{(l)}, \ldots, \Omega_{m_l}^{(l)}\right),$$

with the property that their union covers the whole space $\Omega = \cup_{g=1}^{m_l} \Omega_g^{(l)}$. For each $\Omega_g^{(l)}$, $g \in [m_l]$, there corresponds a probability measure $Q_g^{(l)}$. Therefore:

$$\left(Q_1^{(l)}, Q_2^{(l)}, \ldots, Q_{m_l}^{(l)}\right),$$

represents a dissection of the probability measure $Q = \sum_{g=1}^{m_l} \pi_g^{(l)} Q_g^{(l)}$ with $\sum_{g=1}^{m_l} \pi_g^{(l)} = 1$ and $\pi_g^{(l)} \geq 0$ for all $g \in [m_l]$. For instance, when $l = 1$, then $\Omega_g^{(1)} = \{\omega_g\}$, $Q_g^{(1)} = \{\delta_{\omega_g}\}$, $g \in [|\Omega|]$, where $\{\delta_{\omega_g}\}$ represents the Dirac measure at scenario ω_g. Hence, $Q_g^{(1)} = \{\delta_{\omega_g}\}$ has probability 1 for scenario ω_g and probability zero for all other scenarios. Each measure $Q_g^{(l)}$ in the collection is given by:

$$Q_g^{(l)} := \sum_{\omega_i \in \Omega_g^{(l)}} (q_{\omega_i})_g^{(l)} \cdot \delta_{\omega_i},$$

where $(q_{\omega_i})_g^{(l)}$ denotes the nominal probability of scenario ω_i within subgroup g with $\sum_{\omega_i \in \Omega_g^{(l)}} (q_{\omega_i})_g^{(l)} = 1$. Below, we provide details of probability measures $Q_g^{(l)}$—and hence details of $(q_{\omega_i})_g^{(l)}$—based on different constructions. Collections of subgroups can be constructed principally in two ways: by keeping one or several scenarios fixed in all subsets, or by choosing them disjoint.

Fixed Scenarios

We first consider the case where one or more scenarios appear in all subsets. Without loss of generality, we assume that the first $f < l$ scenarios of Ω ($\Omega_f = \{\omega_1, \ldots, \omega_f\}$) are fixed. Consequently the number of subgroups with cardinality l is $m_l = \frac{|\Omega| - f}{l - f} \in \mathbb{N}$. Then, the probability measures $Q_g^{(l)}$ can be calculated as follows:

$$Q_g^{(l)} := \sum_{i=1}^{f} q_{\omega_i} \cdot \delta_{\omega_i} + \sum_{\omega_i \in \Omega_g^{(l)} \setminus \Omega_f} \frac{q_{\omega_i}}{\pi_g^{(l)}} \cdot \delta_{\omega_i},$$

with weights

$$\pi_g^{(l)} := \frac{\sum_{\omega_i \in \Omega_g^{(l)} \setminus \Omega_f} q_{\omega_i}}{1 - \sum_{i=1}^{f} q_{\omega_i}},$$

for all $g \in [m_l]$.

Disjoint Partitions

Alternatively, one may also consider disjoint partitions: $\Omega = \cup_{g=1}^{m_l} \Omega_g^{(l)}$ with $\Omega_{g_1}^{(l)} \cap \Omega_{g_2}^{(l)} = \emptyset$ for $g_1 \neq g_2$. Consequently, the number of subgroups with cardinality l is $m_l = \frac{|\Omega|}{l} \in \mathbb{N}$. In this case, probability measures $Q_g^{(l)}$ are given by:

$$Q_g^{(l)} := \sum_{\omega_i \in \Omega_g^{(l)}} \frac{q_{\omega_i}}{\pi_g^{(l)}} \cdot \delta_{\omega_i},$$

with weights

$$\pi_g^{(l)} := \sum_{\omega_i \in \Omega_g^{(l)}} q_{\omega_i},$$

for all $g \in [m_l]$.

Let $\Omega_{t,g}^{(l)}$ denote set of nodes of subgroup g at stage $t \in \mathcal{T}$. In the multistage setting, under both dissection strategies, probabilities of non-leaf nodes are adjusted as follows:

$$\left(q_{t-1}^{n}\right)_{g}^{(l)} := \sum_{m \in \mathscr{B}(n):\ m \in \Omega_{t,g}^{(l)}} (q_t^m)_g^{(l)} \qquad n \in \Omega_{t-1,g}^{(l)},\ t \in \mathcal{T} \setminus \{0\}.$$

Example

Figure 9 displays a sample space $\Omega = \{\omega_i\}_{i=1}^{15}$ with 15 scenarios. We divide it into 7 subsets $\Omega_g^{(3)}, g \in [7]$ each of them of cardinality $l = 3$ with scenario ω_1 fixed. That is, $\Omega_f = \{\omega_1\}$, $\Omega_1^{(3)} = \{\omega_1, \omega_2, \omega_3\}$, $\Omega_2^{(3)} = \{\omega_1, \omega_4, \omega_5\}$, and so forth. Assuming equal probability for each scenario, *i.e.*, $q_{\omega_i} = 1/15$, $i \in [15]$, the probability of scenario ω_1 within each scenario group g is $(q_{\omega_1})_g^{(3)} = \frac{1}{15}$ and the probability of other two scenarios in the same group is $(q_{\omega_i})_g^{(3)} = \frac{7}{15}$. The weight of each group is $\pi_g^{(3)} = \frac{1}{7},\ g \in [7]$.

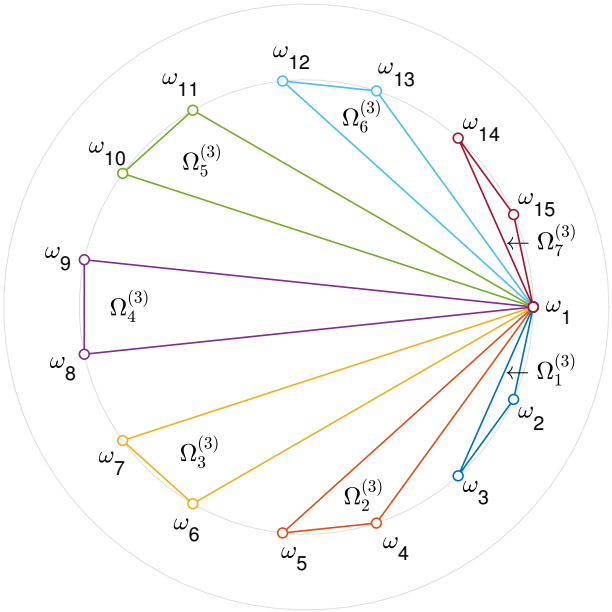


Figure 9: A graphical representation of the sample space Ω, with $|\Omega| = 15$ scenarios, divided into $m_3 = 7$ subsets of cardinality $l = 3$, with one fixed scenario $\Omega_f = \{\omega_1\}$.

2.3.2 Convolution of Risk Measures Induced by DRO

Given a collection of subsets of the scenario tree, our next step is to solve the resulting subgroup problems and combine them judiciously to form lower bounds on the optimal value of DRO. Toward this end, we use convolution of risk measures induced by the considered ambiguity sets [106]. We describe this process next.

We denote as $\mathscr{G}$ the σ-algebra generated by the collection of subsets $\Omega = \cup_{g=1}^{m_l} \Omega_g^{(l)}$, where each subset $\Omega_g^{(l)}$ corresponds to an elementary event of $\mathscr{G}$. We solve subgroup g using DRO with radius $\bar{\rho}_g$ and denote these radii collectively by $\bar{\boldsymbol{\rho}} = \{\bar{\rho}_g\}_{g=1}^{m_l}$. We then combine the optimal values of subgroups using DRO with radius $\bar{\bar{\rho}}$. We denote the ambiguity set used to combine the subgroups as $\widetilde{\mathcal{P}}_{\bar{\bar{\rho}}}^{\mathscr{G}}$ and the induced risk measure on this collection of subsets as $\widetilde{\mathcal{R}}_{\bar{\bar{\rho}}}^{\mathscr{G}}$. Here, we refer to $\bar{\boldsymbol{\rho}}$ and $\bar{\bar{\rho}}$ equivalently as *risk parameters*. We define:

$$\widetilde{\mathcal{R}}_{\bar{\bar{\rho}},\bar{\boldsymbol{\rho}}}(\boldsymbol{\eta}) := \left(\widetilde{\mathcal{R}}_{\bar{\bar{\rho}}}^{\mathscr{G}} \circ \widetilde{\mathcal{R}}_{\bar{\boldsymbol{\rho}}}^{\mathscr{F}|\mathscr{G}}\right)(\boldsymbol{\eta})$$

as convolution of the one-step conditional risk measure $\widetilde{\mathcal{R}}_{\bar{\boldsymbol{\rho}}}^{\mathscr{F}|\mathscr{G}} : \mathcal{Z} \to \mathfrak{L}_\infty(\Omega, \mathscr{G}, Q)$ and the risk measure on the collection of subsets $\widetilde{\mathcal{R}}_{\bar{\bar{\rho}}}^{\mathscr{G}} : \mathfrak{L}_\infty(\Omega, \mathscr{G}, Q) \to \mathbb{R}$ (see [106]). Note that $\widetilde{\mathcal{R}}_{\bar{\boldsymbol{\rho}}}^{\mathscr{F}|\mathscr{G}}$ can be represented in terms of $\mathcal{R}_{\bar{\rho}_g}^{(l)} : \mathfrak{L}_\infty(\Omega, \sigma(\Omega_g^{(l)}), Q) \to \mathbb{R}$ for each subgroup $g \in [m_l]$ with risk parameter $\bar{\rho}_g$, $\sigma(\Omega_g^{(l)})$ the σ-algebra on $\Omega_g^{(l)}$, and $\mathcal{P}_{\bar{\rho}_g}^{(l)}$ the corresponding ambiguity set. We denote the ambiguity set associated with the one-step conditional risk measure $\widetilde{\mathcal{R}}_{\bar{\boldsymbol{\rho}}}^{\mathscr{F}|\mathscr{G}}$ as $\widetilde{\mathcal{P}}_{\bar{\boldsymbol{\rho}}}^{\mathscr{F}|\mathscr{G}} := \cup_{g=1}^{m_l} \mathcal{P}_{\bar{\rho}_g}^{(l)}$. Similarly, we denote the ambiguity set corresponding to the convolution as $\widetilde{\mathcal{P}}_{\bar{\bar{\rho}},\bar{\boldsymbol{\rho}}}$.

The ambiguity sets mentioned above can be formulated as follows. First, given the subset $\Omega_g^{(l)}$ for subgroup g, the ambiguity set associated with DRO using radius $\bar{\rho}_g \geq 0$ inducing risk measure $\mathcal{R}_{\bar{\rho}_g}^{(l)}$ on this subgroup is:

$$\mathcal{P}_{\bar{\rho}_g}^{(l)} := \left\{ \bar{P}_g^{(l)} : \Delta(\bar{P}_g^{(l)}, Q_g^{(l)}) \leq \bar{\rho}_g, \sum_{\omega_i \in \Omega_g^{(l)}} (\bar{p}_{\omega_i})_g^{(l)} = 1, \ \ (\bar{p}_{\omega_i})_g^{(l)} \geq 0, \forall \omega_i \in \Omega_g^{(l)} \right\},$$

where $(\bar{p}_{\omega_i})_g^{(l)}$ represents the probability $\bar{P}_g^{(l)}$ assumes for scenario $\omega_i \in \Omega_g^{(l)}$. Hence, the ambiguity set corresponding to the one-step conditional risk measure $\widetilde{\mathcal{R}}_{\bar{\boldsymbol{\rho}}}^{\mathscr{F}|\mathscr{G}}$ is:

$$\widetilde{\mathcal{P}}_{\bar{\boldsymbol{\rho}}}^{\mathscr{F}|\mathscr{G}} := \left\{ \bar{P} : \Delta(\bar{P}_g^{(l)}, Q_g^{(l)}) \leq \bar{\rho}_g, \forall g \in [m_l], \sum_{\omega_i \in \Omega_g^{(l)}} (\bar{p}_{\omega_i})_g^{(l)} = 1, \forall g \in [m_l], (\bar{p}_{\omega_i})_g^{(l)} \geq 0, \forall \omega_i \in \Omega_g^{(l)}, \forall g \in [m_l] \right\}.$$

Above, $\bar{P} := \{\bar{P}_g^{(l)}\}_{g=1}^{m_l}$. Next, the ambiguity set associated with DRO using radius $\bar{\bar{\rho}} \geq 0$ inducing risk measure $\widetilde{\mathcal{R}}_{\bar{\bar{\rho}}}^{\mathscr{G}}$ on the collection of subsets is:

$$\widetilde{\mathcal{P}}_{\bar{\bar{\rho}}}^{\mathscr{G}} := \left\{ \bar{\bar{P}} : \Delta(\bar{\bar{P}}, \bar{\bar{Q}}) \leq \bar{\bar{\rho}}, \sum_{g \in [m_l]} \bar{\bar{p}}_g^{(l)} = 1, \ \bar{\bar{p}}_g^{(l)} \geq 0, \forall g \in [m_l] \right\},$$

where $\bar{\bar{Q}}$ is the nominal distribution composed of the weights $\pi_g^{(l)}$ detailed in Section 2.3.1 and $\bar{\bar{p}}_g^{(l)}$ represents the probability $\bar{\bar{P}}$ assumes for the subgroup g at cardinality l. Finally, the ambiguity set corresponding to the convolution $\widetilde{\mathcal{R}}_{\bar{\bar{\rho}},\bar{\boldsymbol{\rho}}}$ becomes:

$$\begin{aligned} \widetilde{\mathcal{P}}_{\bar{\bar{\rho}},\bar{\boldsymbol{\rho}}} := \Big\{ P' : \ & p'_{\omega_i,g} = \bar{\bar{p}}_g^{(l)} \cdot (\bar{p}_{\omega_i})_g^{(l)}, \ \forall \omega_i \in \Omega_f, g \in [m_l], \ p'_{\omega_i} = \sum_{g \in [m_l]} p'_{\omega_i,g}, \ \forall \omega_i \in \Omega_f, \\ & \text{and } \ p'_{\omega_i} = \bar{\bar{p}}_g^{(l)} \cdot (\bar{p}_{\omega_i})_g^{(l)}, \ \forall \omega_i \in \Omega_g^{(l)} \setminus \Omega_f, g \in [m_l], \ \bar{\bar{P}} \in \widetilde{\mathcal{P}}_{\bar{\bar{\rho}}}^{\mathscr{G}}, \ \ \bar{P} \in \widetilde{\mathcal{P}}_{\bar{\boldsymbol{\rho}}}^{\mathscr{F}|\mathscr{G}} \Big\}. \end{aligned} \tag{2.4}$$

Recall Ω_f denotes the set of fixed scenarios (Section 2.3.1), and if $\Omega_f = \emptyset$, disjoint partitions are used. For any fixed scenario $\omega_i \in \Omega_f$, its probability p'_{ω_i} is found by summing up its group probabilities $p'_{\omega_i,g}$ for all subgroups $g \in [m_l]$. Notice that in (2.4) the condition $\sum_{\omega_i \in \Omega} p'_{\omega_i} = 1$ always holds because the respective ambiguity sets $\widetilde{\mathcal{P}}^{\mathscr{G}}_{\bar{\bar{\rho}}}$ and $\widetilde{\mathcal{P}}^{\mathscr{F}|\mathscr{G}}_{\bar{\boldsymbol{\rho}}}$ require $\sum_{g\in[m_l]} \bar{\bar{p}}^{(l)}_g = 1$ and $\sum_{\omega_i \in \Omega^{(l)}_g} (\bar{p}_{\omega_i})^{(l)}_g = 1$:

$$\begin{aligned}
\sum_{\omega_i\in\Omega} p'_{\omega_i} &= \sum_{\omega_i\in\Omega_f} p'_{\omega_i} + \sum_{\omega_i\in(\Omega_f)^C} p'_{\omega_i} = \sum_{\omega_i\in\Omega_f}\sum_{g\in[m_l]} p'_{\omega_i,g} + \sum_{\omega_i\in(\Omega_f)^C} p'_{\omega_i} \\
&= \sum_{\omega_i\in\Omega_f}\sum_{g\in[m_l]} \bar{\bar{p}}^{(l)}_g \cdot (\bar{p}_{\omega_i})^{(l)}_g + \sum_{g\in[m_l]}\sum_{\omega_i\in\Omega^{(l)}_g\setminus\Omega_f} \bar{\bar{p}}^{(l)}_g \cdot (\bar{p}_{\omega_i})^{(l)}_g \\
&= \sum_{g\in[m_l]}\sum_{\omega_i\in\Omega^{(l)}_g} \bar{\bar{p}}^{(l)}_g \cdot (\bar{p}_{\omega_i})^{(l)}_g = \sum_{g\in[m_l]} \left(\bar{\bar{p}}^{(l)}_g \cdot \sum_{\omega_i\in\Omega^{(l)}_g} (\bar{p}_{\omega_i})^{(l)}_g\right) = \sum_{g\in[m_l]} \left(\bar{\bar{p}}^{(l)}_g \cdot 1\right) = 1,
\end{aligned}$$

where the complement set with respect to Ω is denoted by $(\cdot)^C$.

Example, Continued

Figure 10 shows the computation of the convolution $\widetilde{\mathcal{R}}_{\bar{\bar{\rho}},\bar{\boldsymbol{\rho}}}(\cdot) = \left(\widetilde{\mathcal{R}}^{\mathscr{G}}_{\bar{\bar{\rho}}} \circ \widetilde{\mathcal{R}}^{\mathscr{F}|\mathscr{G}}_{\bar{\boldsymbol{\rho}}}\right)(\cdot)$ induced by the ambiguity set $\widetilde{\mathcal{P}}_{\bar{\bar{\rho}},\bar{\boldsymbol{\rho}}}$. Here, $\left(\bar{\bar{p}}^{(3)}_1, \ldots, \bar{\bar{p}}^{(3)}_7\right) \in \mathbb{R}^7$, and for each subgroup $g \in [7]$ we have $\left((\bar{p}_{\omega_1})^{(3)}_g, (\bar{p}_{\omega_{2g}})^{(3)}_g, (\bar{p}_{\omega_{2g+1}})^{(3)}_g\right) \in \mathbb{R}^3$. So, the probability measure $P' = \sum_{\omega_i\in\Omega} p'_{\omega_i}\delta_{\omega_i}$ in (2.4) is given by the probabilities $\left(p'_{\omega_1} = \sum_{g=1}^{7} \bar{\bar{p}}^{(3)}_g \cdot (\bar{p}_{\omega_1})^{(3)}_g,\ p'_{\omega_2} = \bar{\bar{p}}^{(3)}_1 \cdot (\bar{p}_{\omega_2})^{(3)}_1, \ldots,\ p'_{\omega_5} = \bar{\bar{p}}^{(3)}_2 \cdot (\bar{p}_{\omega_5})^{(3)}_2,\ \ldots,\ p'_{\omega_{15}} = \bar{\bar{p}}^{(3)}_7 \cdot (\bar{p}_{\omega_{15}})^{(3)}_7\right) \in \mathbb{R}^{15}$.

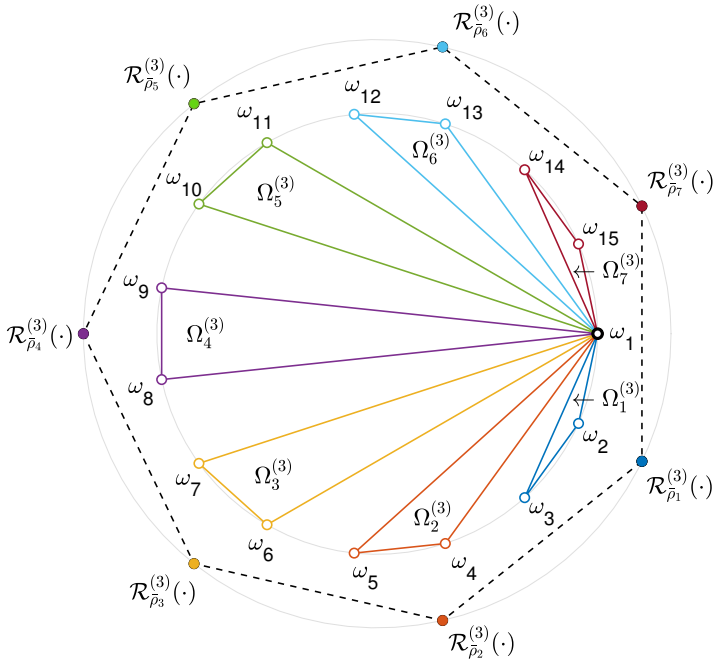


Figure 10: Computation of the risk measure $\tilde{\mathcal{R}}^{\mathscr{G}}_{\bar{\bar{\rho}}}(\cdot)$ (dashed line) combining the optimal values of subgroups obtained using risk measures $\mathcal{R}^{(l)}_{\bar{\rho}_g}(\cdot)$, $g \in [7]$ induced by DRO with $\mathcal{P}^{(l)}_{\bar{\rho}_g}$.

In the rest of this section, we denote the nominal probabilities of the *fixed* scenarios after dissection as $q_{\omega_i,g} = \pi_g^{(l)} \cdot (q_{\omega_i})_g^{(l)}, \forall \omega_i \in \Omega_f,\ g \in [m_l]$, where $q_{\omega_i} = \sum_{g\in[m_l]} q_{\omega_i,g}$ for any $\omega_i \in \Omega_f$. We use $\bar{\rho}_{max}$ to denote the maximal value of $\bar{\rho}_g$ among subgroups $g \in [m_l]$ (*i.e.*, $\bar{\rho}_{max} = \max_{g\in[m_l]} \bar{\rho}_g$). Subscript ϕ_{CR}^θ is used to represent all relevant ambiguity sets and risk measures induced by CR power divergence family with parameter θ. For instance, $\mathcal{R}_{\phi_{CR}^\theta(\rho)}(\boldsymbol{\eta})$ denotes the risk measure induced by the CR power divergence with ambiguity set $\mathcal{P}_{\phi_{CR}^\theta(\rho)}$ using radius ρ, and $\widetilde{\mathcal{R}}_{\phi_{CR}^\theta(\bar{\bar{\rho}},\bar{\boldsymbol{\rho}})}(\boldsymbol{\eta}) = \left(\widetilde{\mathcal{R}}^{\mathscr{G}}_{\phi_{CR}^\theta(\bar{\bar{\rho}})} \circ \widetilde{\mathcal{R}}^{\mathscr{F}|\mathscr{G}}_{\phi_{CR}^\theta(\bar{\boldsymbol{\rho}})}\right)(\boldsymbol{\eta})$ denotes the risk measure after convolution using radii $\bar{\bar{\rho}}$ and $\bar{\boldsymbol{\rho}}$, and so forth. Similarly, we use subscripts ϕ_v, ϕ_J, ϕ_χ^a, and W to denote VD, J-divergence, χ-divergence of order $a > 1$ and Wasserstein distance, respectively. These notations are used in the subsequent results and their proofs.

2.3.3 Lower-Bound Criteria for ϕ-Divergences

We now present LB criteria for DRO formed via some commonly used ϕ-divergences listed in Table 9 through scenario grouping. We begin with CR power divergence family and present the proof in detail. The LB criteria for the special cases of the CR power divergence family in Table 10 can be acquired from the below result.

Proposition 1. (LB criteria for CR power divergences). Consider the convolution of DRO formed by CR power divergence with parameter $\theta \neq 0, 1$. For $0 < \theta < 1$, radii $\rho,\ \bar{\rho}_g,\ \bar{\bar{\rho}} \in \left[0,\ \frac{1}{\theta} + \frac{1}{(1-\theta)}\right],\ g \in [m_l]$; for $\theta < 0$ or $\theta > 1$, $\rho,\ \bar{\rho}_g,\ \bar{\bar{\rho}} \geq 0$ and we further suppose that the support Ω is dissected by disjoint partitions (*i.e.*, $\Omega_f = \emptyset$). If

$$\begin{cases} \bar{\bar{\rho}} + \bar{\rho}_{max} \leq \rho & \text{when} \quad \theta \in (0,1), \\ \bar{\bar{\rho}} + \bar{\rho}_{max} - \theta(1-\theta) \cdot \bar{\bar{\rho}} \cdot \bar{\rho}_{max} \leq \rho & \text{when} \quad \theta < 0 \text{ or } \theta > 1 \text{ and } \Omega_f = \emptyset, \\ \bar{\bar{\rho}} + \bar{\rho}_{max} \leq \rho & \text{when} \quad \theta \to 0 \text{ or } \theta \to 1, \end{cases}$$

then $\widetilde{\mathcal{R}}_{\phi_{CR}^\theta(\bar{\bar{\rho}},\bar{\boldsymbol{\rho}})}(\boldsymbol{\eta}) \leq \mathcal{R}_{\phi_{CR}^\theta(\rho)}(\boldsymbol{\eta})$ for all $\boldsymbol{\eta} \in \mathcal{Z}$.

Proof. Let $P' \in \widetilde{\mathcal{P}}_{\phi_{CR}^\theta(\bar{\bar{\rho}},\bar{\boldsymbol{\rho}})}$. Then there exists $\bar{\bar{P}} \in \widetilde{\mathcal{P}}^{\mathscr{G}}_{\phi_{CR}^\theta(\bar{\bar{\rho}})}$ and $\bar{P} \in \widetilde{\mathcal{P}}^{\mathscr{F}|\mathscr{G}}_{\phi_{CR}^\theta(\bar{\boldsymbol{\rho}})}$ such that $\sum_{g\in[m_l]} \bar{\bar{p}}_g^{(l)} = 1$, $\sum_{\omega_i\in\Omega_g^{(l)}} (\bar{p}_{\omega_i})_g^{(l)} = 1$ and, by the definition of $\Delta_{\phi_{CR}^\theta}$ from Table 9, we have

$$\frac{1 - \sum\limits_{g\in[m_l]} \left(\bar{\bar{p}}_g^{(l)}\right)^\theta \left(\pi_g^{(l)}\right)^{1-\theta}}{\theta(1-\theta)} \leq \bar{\bar{\rho}}, \quad \frac{1 - \sum\limits_{\omega_i\in\Omega_g^{(l)}} \left((\bar{p}_{\omega_i})_g^{(l)}\right)^\theta \left((q_{\omega_i})_g^{(l)}\right)^{1-\theta}}{\theta(1-\theta)} \leq \bar{\rho}_g. \tag{a}$$

We now show the steps to find the criteria for $\Delta_{\phi_{CR}^\theta}(P', Q) \leq \rho$.

1. When $\theta \in (0,1)$, writing out $\Delta_{\phi^{\theta}_{CR}}(P', Q)$, we obtain

$$\frac{1-\sum_{\omega_i\in\Omega}\left(p'_{\omega_i}\right)^{\theta}\left(q_{\omega_i}\right)^{1-\theta}}{\theta(1-\theta)}$$

$$=\frac{1-\sum_{\omega_i\in\Omega_f}\left(p'_{\omega_i}\right)^{\theta}\left(q_{\omega_i}\right)^{1-\theta}-\sum_{\omega_i\in(\Omega_f)^C}\left(p'_{\omega_i}\right)^{\theta}\left(q_{\omega_i}\right)^{1-\theta}}{\theta(1-\theta)}$$

$$=\frac{1-\sum_{\omega_i\in\Omega_f}\left(\sum_{g\in[m_l]}p'_{\omega_i,g}\right)^{\theta}\left(\sum_{g\in[m_l]}q_{\omega_i,g}\right)^{1-\theta}-\sum_{\omega_i\in(\Omega_f)^C}\left(p'_{\omega_i}\right)^{\theta}\left(q_{\omega_i}\right)^{1-\theta}}{\theta(1-\theta)}$$

$$\leq\frac{1-\sum_{\omega_i\in\Omega_f}\sum_{g\in[m_l]}\left(p'_{\omega_i,g}\right)^{\theta}\left(q_{\omega_i,g}\right)^{1-\theta}-\sum_{\omega_i\in(\Omega_f)^C}\left(p'_{\omega_i}\right)^{\theta}\left(q_{\omega_i}\right)^{1-\theta}}{\theta(1-\theta)} \tag{b}$$

$$=\frac{1-\sum_{\substack{g\in[m_l]\\ \omega_i\in\Omega_f}}\left(\bar{\bar{p}}^{(l)}_g(\bar{p}_{\omega_i})^{(l)}_g\right)^{\theta}\left(\pi^{(l)}_g(q_{\omega_i})^{(l)}_g\right)^{1-\theta}-\sum_{\substack{g\in[m_l]\\ \omega_i\in\Omega^{(l)}_g\setminus\Omega_f}}\left(\bar{\bar{p}}^{(l)}_g(\bar{p}_{\omega_i})^{(l)}_g\right)^{\theta}\left(\pi^{(l)}_g(q_{\omega_i})^{(l)}_g\right)^{1-\theta}}{\theta(1-\theta)}$$

$$=\frac{1-\sum_{g\in[m_l]}\sum_{\omega_i\in\Omega^{(l)}_g}\left(\bar{\bar{p}}^{(l)}_g(\bar{p}_{\omega_i})^{(l)}_g\right)^{\theta}\left(\pi^{(l)}_g(q_{\omega_i})^{(l)}_g\right)^{1-\theta}}{\theta(1-\theta)} \tag{c}$$

$$=\frac{1-\sum_{g\in[m_l]}\left(\bar{\bar{p}}^{(l)}_g\right)^{\theta}\left(\pi^{(l)}_g\right)^{1-\theta}}{\theta(1-\theta)}+\frac{\sum_{g\in[m_l]}\left(\bar{\bar{p}}^{(l)}_g\right)^{\theta}\left(\pi^{(l)}_g\right)^{1-\theta}-\sum_{g\in[m_l]}\sum_{\omega_i\in\Omega^{(l)}_g}\left(\bar{\bar{p}}^{(l)}_g(\bar{p}_{\omega_i})^{(l)}_g\right)^{\theta}\left(\pi^{(l)}_g(q_{\omega_i})^{(l)}_g\right)^{1-\theta}}{\theta(1-\theta)}$$

$$=\frac{1-\sum_{g\in[m_l]}\left(\bar{\bar{p}}^{(l)}_g\right)^{\theta}\left(\pi^{(l)}_g\right)^{1-\theta}}{\theta(1-\theta)}+\sum_{g\in[m_l]}\left[\left(\bar{\bar{p}}^{(l)}_g\right)^{\theta}\left(\pi^{(l)}_g\right)^{1-\theta}\frac{1-\sum_{\omega_i\in\Omega^{(l)}_g}\left((\bar{p}_{\omega_i})^{(l)}_g\right)^{\theta}\left((q_{\omega_i})^{(l)}_g\right)^{1-\theta}}{\theta(1-\theta)}\right]$$

$$\leq\bar{\bar{\rho}}+\sum_{g\in[m_l]}\left[\left(\bar{\bar{p}}^{(l)}_g\right)^{\theta}\left(\pi^{(l)}_g\right)^{1-\theta}\bar{\rho}_g\right]\leq\bar{\bar{\rho}}+\sum_{g\in[m_l]}\left[\left(\bar{\bar{p}}^{(l)}_g\right)^{\theta}\left(\pi^{(l)}_g\right)^{1-\theta}\right]\cdot\bar{\rho}_{max} \tag{d}$$

$$=\bar{\bar{\rho}}+\bar{\rho}_{max}-\theta(1-\theta)\frac{1-\sum_{g\in[m_l]}\left(\bar{\bar{p}}^{(l)}_g\right)^{\theta}\left(\pi^{(l)}_g\right)^{1-\theta}}{\theta(1-\theta)}\cdot\bar{\rho}_{max}, \tag{e}$$

where inequality (b) follows from Hölder's inequality applied on the fixed scenarios. The first inequality in (d) follows from (a) and the second from definition of $\bar{\rho}_{max}$. Let us denote the right-hand side of (e) as A. Since $-\theta(1-\theta)$ is negative, $A \leq \bar{\bar{\rho}} + \bar{\rho}_{max}$. Therefore, if $\bar{\bar{\rho}} + \bar{\rho}_{max} \leq \rho$, the result follows.

2. When $\theta < 0$ or $\theta > 1$, we can no longer apply Hölder's inequality on the fixed scenarios in (b). However, when $\Omega_f = \emptyset$ (*i.e.*, disjoint partitions are used), we can directly start from (c) and follow the steps to (e). Since $-\theta(1-\theta)$ is positive, $A \leq \bar{\bar{\rho}} + \bar{\rho}_{max} - \theta(1-\theta)\cdot\bar{\bar{\rho}}\cdot\bar{\rho}_{max}$ by (a). Therefore, if $\bar{\bar{\rho}} + \bar{\rho}_{max} - \theta(1-\theta)\cdot\bar{\bar{\rho}}\cdot\bar{\rho}_{max} \leq \rho$, the result follows.

3. When $\theta \to 1$, the CR power divergence is equivalent to Kullback-Leibler divergence. Detailed proof is provided in Appendix B.

When $\theta \to 0$, the proof is similar to the $\theta \to 1$ case and hence omitted. □

In Appendix B, we provide a detailed proof of Kullback-Leibler divergence (*i.e.*, $\theta \to 1$ limit case). Proof of Proposition 1 and Appendix B reveal that when $\theta \in (0, 1)$ or in the limit cases of Kullback-Leibler divergence ($\theta \to 1$) and Burg entropy ($\theta \to 0$), the scenario tree Ω can be dissected in any way, either using disjoint partitions or fixed scenarios. However, when $\theta < 0$ or $\theta > 1$, the above result is valid for disjoint partitions.

LB criteria for other ϕ-divergences in Table 9 can be obtained using a similar proof technique. Below, we provide the results and relegate the proofs to the Appendix B.

Proposition 2. (LB criterion for variation distance). Consider the convolution of DRO formed by variation distance, where radii $\rho, \bar{\rho}_g, \bar{\bar{\rho}} \in [0, 2],\ g \in [m_l]$.

$$\bar{\bar{\rho}} \cdot \bar{\rho}_{max} + \bar{\bar{\rho}} + \bar{\rho}_{max} \leq \rho,$$

then $\widetilde{\mathcal{R}}_{\phi_v(\bar{\bar{\rho}},\bar{\boldsymbol{\rho}})}(\boldsymbol{\eta}) \leq \mathcal{R}_{\phi_v(\rho)}(\boldsymbol{\eta})$ for all $\boldsymbol{\eta} \in \mathcal{Z}$.

Proposition 3. (LB criterion for J-divergence). Consider the convolution of DRO formed by J-divergence, where radii $\rho,\ \bar{\rho}_g,\ \bar{\bar{\rho}} \geq 0,\ g \in [m_l]$. If

$$\bar{\bar{\rho}} + \bar{\rho}_{max} \leq \rho,$$

then $\widetilde{\mathcal{R}}_{\phi_J(\bar{\bar{\rho}},\bar{\boldsymbol{\rho}})}(\boldsymbol{\eta}) \leq \mathcal{R}_{\phi_J(\rho)}(\boldsymbol{\eta})$ for all $\boldsymbol{\eta} \in \mathcal{Z}$.

Proposition 4. (LB criterion for χ-divergence of order $a > 1$). Consider the convolution of DRO formed by χ-divergence of order $a > 1$, where radii $\rho,\ \bar{\rho}_g,\ \bar{\bar{\rho}} \geq 0,\ g \in [m_l]$ and suppose the support Ω is dissected by disjoint partitions (*i.e.*, $\Omega_f = \emptyset$). If

$$\left[\left(\bar{\bar{\rho}}\right)^{\frac{1}{a}} + (\bar{\rho}_{max})^{\frac{1}{a}} + \left(\bar{\bar{\rho}} \cdot \bar{\rho}_{max}\right)^{\frac{1}{a}}\right]^a \leq \rho,$$

then $\widetilde{\mathcal{R}}_{\phi^a_\chi(\bar{\bar{\rho}},\bar{\boldsymbol{\rho}})}(\boldsymbol{\eta}) \leq \mathcal{R}_{\phi^a_\chi(\rho)}(\boldsymbol{\eta})$ for all $\boldsymbol{\eta} \in \mathcal{Z}$.

Note that Proposition 4 is a general result that applies to all values of $a > 1$. For certain values of a, a tighter inequality might be available (*e.g.*, when $a = 2$).

2.3.4 Lower-Bound Criterion for Wasserstein Distance

We now provide a LB criterion for Wasserstein distance introduced in Section 2.2.4 through scenario decomposition. The main idea is the same: to find criteria that guarantee the convoluted ambiguity set being a subset of the ambiguity set of the original problem. Recall that Wasserstein distance needs a distance d_{ω_i,ω_j} between any two scenarios ω_i and ω_j. To apply it to scenario groups, we need a distance between subgroups as well. We provide such a distance between subgroups and a criterion for radii $\bar{\bar{\rho}}$, $\bar{\boldsymbol{\rho}}$ to ensure lower bounds below.

Proposition 5. (LB criterion for Wasserstein distance). Consider the convolution of DRO formed by Wasserstein distance, where radii $\rho,\ \bar{\rho}_g,\ \bar{\bar{\rho}} \geq 0,\ g \in [m_l]$. Let the distance between scenario groups be defined as

$$d_{g_1,g_2} := \begin{cases} \max\limits_{\omega_i \in \Omega_{g_1}^{(l)},\ \omega_j \in \Omega_{g_2}^{(l)}} \{d_{\omega_i,\omega_j}\} & \text{when} \quad g_1 \neq g_2 \\ 0 & \text{when} \quad g_1 = g_2, \end{cases} \tag{2.5}$$

where $g_1,\ g_2 \in [m_l]$. If

$$\bar{\bar{\rho}} + \bar{\rho}_{max} \leq \rho,$$

then $\widetilde{\mathcal{R}}_{W(\bar{\bar{\rho}},\bar{\boldsymbol{\rho}})}(\boldsymbol{\eta}) \leq \mathcal{R}_{W(\rho)}(\boldsymbol{\eta})$ for all $\boldsymbol{\eta} \in \mathcal{Z}$.

Proof. First assume the support Ω is dissected using *disjoint* partitions (*i.e.*, $\Omega_f = \emptyset$). Given $P' \in \widetilde{\mathcal{P}}_{W(\bar{\bar{\rho}},\bar{\boldsymbol{\rho}})}$ formed after scenario grouping, the Wasserstein distance between P' and Q for the original problem can be written as

$$\Delta_W(P',Q) := \min_{z \geq 0} \Bigg\{ \sum_{\omega_i \in \Omega} \sum_{\omega_j \in \Omega} d_{\omega_i,\omega_j} z_{\omega_i,\omega_j} : \sum_{\omega_i \in \Omega} z_{\omega_i,\omega_j} = q_{\omega_j},\ \omega_j \in \Omega, \\ \sum_{\omega_j \in \Omega} z_{\omega_i,\omega_j} = p'_{\omega_i},\ \omega_i \in \Omega \Bigg\}, \tag{f}$$

and the Wasserstein distance for ambiguity set $\widetilde{\mathcal{P}}^{\mathscr{G}}_{W(\bar{\bar{\rho}})}$ can be written as

$$\Delta_W(\bar{\bar{P}},\bar{\bar{Q}}) := \min_{y \geq 0} \Bigg\{ \sum_{g_1 \in [m_l]} \sum_{g_2 \in [m_l]} d_{g_1,g_2}\ y_{g_1,g_2} : \sum_{g_1 \in [m_l]} y_{g_1,g_2} = \pi^{(l)}_{g_2},\ g_2 \in [m_l], \\ \sum_{g_2 \in [m_l]} y_{g_1,g_2} = \bar{\bar{p}}^{(l)}_{g_1},\ g_1 \in [m_l] \Bigg\}. \tag{g}$$

Similarly, for each subgroup $g \in [m_l]$, we have

$$\Delta_W(\bar{P}^{(l)}_g, Q^{(l)}_g) := \min_{x \geq 0} \Bigg\{ \sum_{\omega_i \in \Omega^{(l)}_g} \sum_{\omega_j \in \Omega^{(l)}_g} d_{\omega_i,\omega_j} (x_{\omega_i,\omega_j})_g : \sum_{\omega_i \in \Omega^{(l)}_g} (x_{\omega_i,\omega_j})_g = (q_{\omega_j})^{(l)}_g,\ \omega_j \in \Omega^{(l)}_g, \\ \sum_{\omega_j \in \Omega^{(l)}_g} (x_{\omega_i,\omega_j})_g = (\bar{p}_{\omega_i})^{(l)}_g,\ \omega_i \in \Omega^{(l)}_g \Bigg\}. \tag{h}$$

Let us now define one way to obtain z_{ω_i,ω_j} in (f) by using the transportation decisions y_{g_1,g_2} and $(x_{\omega_i,\omega_j})_g$ in (g) and (h), respectively:

$$z_{\omega_i,\omega_j} = y_{g,g}\ (x_{\omega_i,\omega_j})_g, \qquad \omega_i,\omega_j \in \Omega_g^{(l)},\ g \in [m_l] \tag{i}$$

$$\sum_{\omega_i \in \Omega_{g_1}^{(l)}} z_{\omega_i,\omega_j} = y_{g_1,g_2} \sum_{\omega_i \in \Omega_{g_2}^{(l)}} (x_{\omega_i,\omega_j})_{g_2}, \qquad \omega_j \in \Omega_{g_2}^{(l)},\ g_2 \in [m_l] \tag{j}$$

$$\sum_{\omega_j \in \Omega_{g_2}^{(l)}} z_{\omega_i,\omega_j} = y_{g_1,g_2} \sum_{\omega_j \in \Omega_{g_1}^{(l)}} (x_{\omega_i,\omega_j})_{g_1}, \qquad \omega_i \in \Omega_{g_1}^{(l)},\ g_1 \in [m_l]. \tag{k}$$

With the transformation above, we can show that the constraints in (f) of the Wasserstein distance $\Delta_W(P', Q)$ are all satisfied, even though z_{ω_i,ω_j} formed through (i)–(k) may not be optimal to $\Delta_W(P', Q)$. For instance, the first set of constraints in (f) for all $\omega_j \in \Omega$ (or equivalently all $\omega_j \in \Omega_g^{(l)},\ g \in [m_l]$) are satisfied by (j) and the first sets of constraints in (g) and (h):

$$\begin{aligned}
\sum_{\omega_i \in \Omega} z_{\omega_i,\omega_j} &= \sum_{g_1 \in [m_l]} \sum_{\omega_i \in \Omega_{g_1}^{(l)}} z_{\omega_i,\omega_j} \\
&= \sum_{g_1 \in [m_l]} \left(y_{g_1,g} \sum_{\omega_i \in \Omega_g^{(l)}} (x_{\omega_i,\omega_j})_g \right) \\
&= \left(\sum_{g_1 \in [m_l]} y_{g_1,g} \right) \left(\sum_{\omega_i \in \Omega_g^{(l)}} (x_{\omega_i,\omega_j})_g \right) \\
&= \pi_g^{(l)}\ (q_{\omega_j})_g^{(l)} \\
&= q_{\omega_j},
\end{aligned}$$

where g denotes the subgroup scenario ω_j belongs to. The second set of constraints in (f) can be shown similarly by using (k) and the second sets of constraints in (g) and (h). Hence, all feasible solutions to constraints in (g) and (h) are also feasible to the constraints in (f).

Let $\mathbb{Z}$, $\mathbb{Y}$, and $\mathbb{X}$ denote the feasible regions given by the constraints in (f), (g), and (h), respectively, each supplemented with their nonnegativity constraints $z \geq 0$, $y \geq 0$, and $x \geq 0$. We now show steps to find criteria for $\Delta_W(P', Q) \leq \rho$:

$$\begin{aligned}
&\min_{z \in \mathbb{Z}} \sum_{\omega_i \in \Omega} \sum_{\omega_j \in \Omega} d_{\omega_i,\omega_j} z_{\omega_i,\omega_j} \\
=&\min_{z \in \mathbb{Z}} \sum_{g \in [m_l]} \sum_{\omega_i \in \Omega_g^{(l)}} \left(\sum_{\omega_j \in \Omega_g^{(l)}} d_{\omega_i,\omega_j} z_{\omega_i,\omega_j} + \sum_{\omega_j \in \left(\Omega_g^{(l)}\right)^C} d_{\omega_i,\omega_j} z_{\omega_i,\omega_j} \right) \\
\leq&\min_{\substack{z \in \mathbb{Z} \\ y \in \mathbb{Y} \\ x \in \mathbb{X}}} \sum_{g \in [m_l]} \sum_{\omega_i \in \Omega_g^{(l)}} \left(\sum_{\omega_j \in \Omega_g^{(l)}} d_{\omega_i,\omega_j} y_{g,g} (x_{\omega_i,\omega_j})_g + \sum_{\omega_j \in \left(\Omega_g^{(l)}\right)^C} d_{\omega_i,\omega_j} z_{\omega_i,\omega_j} \right)
\end{aligned} \tag{l}$$

$$\leq \min_{\substack{z\in\mathbb{Z}\\ y\in\mathbb{Y}\\ x\in\mathbb{X}}} \sum_{g\in[m_l]} \left(\sum_{\omega_i\in\Omega_g^{(l)}} \sum_{\omega_j\in\Omega_g^{(l)}} d_{\omega_i,\omega_j} y_{g,g}(x_{\omega_i,\omega_j})_g + \sum_{g_2\in[m_l]} \sum_{\omega_i\in\Omega_g^{(l)}} \sum_{\omega_j\in\Omega_{g_2}^{(l)}} d_{g,g_2} z_{\omega_i,\omega_j} \right) \tag{m}$$

$$\leq \min_{\substack{y\in\mathbb{Y}\\ x\in\mathbb{X}}} \sum_{g\in[m_l]} \left(y_{g,g} \sum_{\omega_i\in\Omega_g^{(l)}} \sum_{\omega_j\in\Omega_g^{(l)}} d_{\omega_i,\omega_j}(x_{\omega_i,\omega_j})_g \right) + \sum_{g\in[m_l]} \sum_{g_2\in[m_l]} d_{g,g_2} y_{g,g_2} \tag{n}$$

$$\leq \min_{\substack{y\in\mathbb{Y}\\ x\in\mathbb{X}}} \sum_{g\in[m_l]} \left(\pi_g^{(l)} \sum_{\omega_i\in\Omega_g^{(l)}} \sum_{\omega_j\in\Omega_g^{(l)}} d_{\omega_i,\omega_j}(x_{\omega_i,\omega_j})_g \right) + \sum_{g\in[m_l]} \sum_{g_2\in[m_l]} d_{g,g_2} y_{g,g_2} \tag{o}$$

$$= \sum_{g\in[m_l]} \left(\pi_g^{(l)} \min_{x\in\mathbb{X}} \sum_{\omega_i\in\Omega_g^{(l)}} \sum_{\omega_j\in\Omega_g^{(l)}} d_{\omega_i,\omega_j}(x_{\omega_i,\omega_j})_g \right) + \min_{y\in\mathbb{Y}} \sum_{g\in[m_l]} \sum_{g_2\in[m_l]} d_{g,g_2} y_{g,g_2} \tag{p}$$

$$= \sum_{g\in[m_l]} \pi_g^{(l)} \Delta_W(\bar{P}_g^{(l)}, Q_g^{(l)}) + \Delta_W(\bar{\bar{P}}, \bar{\bar{Q}}) \tag{q}$$

$$\leq \sum_{g\in[m_l]} \pi_g^{(l)} \cdot \bar{\rho}_g + \bar{\bar{\rho}}$$

$$\leq \bar{\bar{\rho}} + \bar{\rho}_{max}, \tag{r}$$

where (l) follows from (i). Note that this is an inequality because decisions z obtained through this transformation may not be optimal. Inequality (m) follows (2.5). By summing over $\omega_j \in \Omega_{g_2}^{(l)}$ on both sides of (j), we can show $y_{g,g_2} = \sum_{\omega_i\in\Omega_g^{(l)}} \sum_{\omega_j\in\Omega_{g_2}^{(l)}} z_{\omega_i,\omega_j}$, $\forall g, g_2 \in [m_l]$. Then (n) follows. Inequality (o) follows from $y_{g,g} \leq \pi_g^{(l)}$ (see (g)). Equality (p) is due to the fact that the resulting problem is separable. Finally, the equality in (q) follows from the definition of $\Delta_W(\bar{P}_g^{(l)}, Q_g^{(l)})$ and $\Delta_W(\bar{\bar{P}}, \bar{\bar{Q}})$, and the following inequalities follows by construction. Therefore, similar to the statement at the end of the proof of Proposition 1, when $\theta \in (0,1)$, if $\bar{\bar{\rho}} + \bar{\rho}_{max} \leq \rho$, then $\widetilde{\mathcal{R}}_{W(\bar{\bar{\rho}},\bar{\boldsymbol{\rho}})}(\boldsymbol{\eta}) \leq \mathcal{R}_{W(\rho)}(\boldsymbol{\eta})$ for all $\boldsymbol{\eta} \in \mathcal{Z}$.

We now consider the case when the support Ω is dissected using *fixed* scenarios ($\Omega_f \neq \emptyset$). Recall in (2.4), for any fixed scenario $\omega_i \in \Omega_f$, we have $p'_{\omega_i,g} = \bar{\bar{p}}_g^{(l)} \ (\bar{p}_{\omega_i})_g^{(l)}, \ g \in [m_l]$ and $p'_{\omega_i} = \sum_{g\in[m_l]} p'_{\omega_i,g}$. Hence splitting the two constraints in (f), we have

$$\sum_{\omega_i\in\Omega} z_{\omega_i,\omega_j} = q_{\omega_j} = \sum_{g\in[m_l]} q_{\omega_j,g} = \sum_{g\in[m_l]} \pi_g^{(l)} \ (q_{\omega_j})_g^{(l)}, \qquad \omega_j \in \Omega_f, \tag{s}$$

$$\sum_{\omega_i\in\Omega} z_{\omega_i,\omega_j} = q_{\omega_j} = \pi_g^{(l)} \ (q_{\omega_j})_g^{(l)}, \qquad \omega_j \in (\Omega_f)^C, \tag{t}$$

$$\sum_{\omega_j\in\Omega} z_{\omega_i,\omega_j} = p'_{\omega_i} = \sum_{g\in[m_l]} p'_{\omega_i,g} = \sum_{g\in[m_l]} \bar{\bar{p}}_g^{(l)} \ (\bar{p}_{\omega_i})_g^{(l)}, \qquad \omega_i \in \Omega_f, \tag{u}$$

$$\sum_{\omega_j\in\Omega} z_{\omega_i,\omega_j} = p'_{\omega_i} = \bar{\bar{p}}_g^{(l)} \ (\bar{p}_{\omega_i})_g^{(l)}, \qquad \omega_i \in (\Omega_f)^C. \tag{v}$$

Define a finite expanded support $\widetilde{\Omega} := \{\omega_{1_{(1)}}, \omega_{1_{(2)}}, \ldots, \omega_{1_{(m_l)}}, \omega_{2_{(1)}}, \omega_{2_{(2)}}, \ldots, \omega_{2_{(m_l)}}, \ldots, \omega_{f_{(1)}}, \omega_{f_{(2)}}, \ldots, \omega_{f_{(m_l)}}, \omega_{f+1}, \omega_{f+2}, \ldots, \omega_{|\Omega|}\}$, where the fixed scenarios $\omega_i \in \Omega_f = \{\omega_1, \ldots,$

$\omega_f\}$ in different subgroups are considered to have different "atoms" and the rest of the scenarios $\omega_i \in (\Omega_f)^C$ are left as before. We again use the same subgroups $\Omega_g^{(l)}$ but on the expanded support $\widetilde{\Omega}$. Then, we have the following Wasserstein distance on $\widetilde{\Omega}$:

$$\widetilde{\Delta}_W(P', Q) := \min_{\tilde{z}\geq 0} \Bigg\{ \sum_{\omega_i\in\widetilde{\Omega}} \sum_{\omega_j\in\widetilde{\Omega}} d_{\omega_i,\omega_j} \tilde{z}_{\omega_i,\omega_j} : \quad \sum_{\omega_i\in\widetilde{\Omega}} \tilde{z}_{\omega_i,\omega_j} = \pi_g^{(l)}\ (q_{\omega_j})_g^{(l)}, \forall \omega_j \in \widetilde{\Omega}, \\ \sum_{\omega_j\in\widetilde{\Omega}} \tilde{z}_{\omega_i,\omega_j} = \bar{\bar{p}}_g^{(l)}\ (\bar{p}_{\omega_i})_g^{(l)}, \forall \omega_i \in \widetilde{\Omega} \Bigg\}. \tag{w}$$

For any $\tilde{z} \geq 0$ feasible to Wasserstein distance $\widetilde{\Delta}_W(P', Q)$—including the optimal $\tilde{z}$—on the expanded support $\widetilde{\Omega}$, we can generate a feasible solution $z \geq 0$ to (s)–(v). First, for any non-fixed scenarios $\omega_i, \omega_j \in (\Omega_f)^C$, we set $z_{\omega_i,\omega_j} = \tilde{z}_{\omega_i,\omega_j}$ and observe the constraints in (w) are the same as (t) and (v). Next, for any fixed scenario $\omega_i \in \Omega_f$ and non-fixed scenario $\omega_j \in (\Omega_f)^C$, we set (i) $z_{\omega_i,\omega_i} = \sum_{g_1\in[m_l]} \sum_{g_2\in[m_l]} \tilde{z}_{\omega_{i_{(g_1)}},\omega_{i_{(g_2)}}}$, (ii) $z_{\omega_i,\omega_j} = \sum_{g\in[m_l]} \tilde{z}_{\omega_{i_{(g)}},\omega_j}$, and (iii) $z_{\omega_j,\omega_i} = \sum_{g\in[m_l]} \tilde{z}_{\omega_j,\omega_{i_{(g)}}}$. Then (s) and (u) are also satisfied. Furthermore, with this z, the objective functions of two Wasserstein distances coincide: $\sum_{\omega_i\in\Omega} \sum_{\omega_j\in\Omega} d_{\omega_i,\omega_j} z_{\omega_i,\omega_j} = \sum_{\omega_i\in\widetilde{\Omega}} \sum_{\omega_j\in\widetilde{\Omega}} d_{\omega_i,\omega_j} \tilde{z}_{\omega_i,\omega_j}$ because any distance involving $\omega_{i_{(g)}} \in \widetilde{\Omega}$ is equivalent to distance involving $\omega_i \in \Omega_f$, *e.g.*, $d_{\omega_{i_{(g_1)}},\omega_{i_{(g_2)}}} = 0$ for all $\omega_i \in \Omega_f$, $g_1, g_2 \in [m_l]$. As a result, $\Delta_W(P', Q) \leq \widetilde{\Delta}_W(P', Q)$. Observe $\widetilde{\Delta}_W(P', Q)$ is obtained as "disjoint" partitions on $\widetilde{\Omega}$. Then, following similar steps to the disjoint partition, we show $\widetilde{\Delta}_W(P', Q) \leq \bar{\bar{\rho}} + \bar{\rho}_{max}$. Therefore, if $\bar{\bar{\rho}} + \bar{\rho}_{max} \leq \rho$ the result follows. □

Remark. Although the above propositions use $\bar{\rho}_{max}$, these results can also be obtained using the individual $\bar{\rho}_g$ values for each scenario group $g \in [m_l]$. For instance, the condition for Wasserstein would be $\bar{\bar{\rho}} + \sum_{g\in[m_l]} \pi_g^{(l)} \cdot \bar{\rho}_g \leq \rho$ and the result for J-divergence would be $\bar{\bar{\rho}} + \sum_{g\in[m_l]} \left(\bar{\bar{p}}_g^{(l)} \cdot \bar{\rho}_g\right) \leq \rho$. In our numerical results, we use the same value for each group, *i.e.*, $\bar{\rho}_g = \bar{\rho}_{max}$ for all $g \in [m_l]$.

2.3.5 Lower Bounds for Multistage Optimization Problems

We now extend the LBs obtained in Sections 2.3.3 and 2.3.4 to multistage DRO. Due to the correspondence between DRO and RASO, here we focus on a RASO formulation and present our results and proofs using the properties of conditional coherent risk measures. We first recall the results in [134]. For a multistage decision horizon with stages $t \in \mathcal{T}$ let $\mathcal{Z}_t := \mathfrak{L}_\infty(\Omega_t, \mathscr{F}_t, Q_t)$ with $\mathscr{F}_0 = \{\Omega_T, \emptyset\}$. The mapping $\mathcal{R}_{\rho_{t+1}}^{\mathscr{F}_{t+1}|\mathscr{F}_t} : \mathcal{Z}_{t+1} \to \mathcal{Z}_t$ is called one-step conditional risk measure if it satisfies properties presented in Section 2.2.5 for corresponding spaces $\mathcal{Z}_t$ and $\mathcal{Z}_{t+1}$ for all $t \in \{0, \ldots, T-1\}$. The risk involved in a sequence of random variables $\boldsymbol{\eta}_t \in \mathcal{Z}_t$, $t \in \mathcal{T}$ adapted to the filtration $\mathscr{F}_t$, $t \in \mathcal{T}$ can be evaluated by a time-consistent dynamic risk measure $\mathfrak{R}_{\boldsymbol{\rho}}$ induced by a measure of similarity between

distributions Δ using radii $\boldsymbol{\rho}$, defined as follows:

$$\mathfrak{R}_{\boldsymbol{\rho}}\left(\boldsymbol{\eta}_0,\ldots,\boldsymbol{\eta}_T\right) := \boldsymbol{\eta}_0 + \mathcal{R}_{\rho_1}^{\mathscr{F}_1|\mathscr{F}_0}\left(\boldsymbol{\eta}_1 + \mathcal{R}_{\rho_2}^{\mathscr{F}_2|\mathscr{F}_1}\left(\boldsymbol{\eta}_2 + \ldots + \mathcal{R}_{\rho_T}^{\mathscr{F}_T|\mathscr{F}_{T-1}}\left(\boldsymbol{\eta}_T\right)\right)\right), \tag{2.6}$$

where $\boldsymbol{\rho} := (\rho_1,\ldots,\rho_T)$. It is not necessary to use the same measure of distribution distance Δ at each stage of the problem. Also, by changing the radii ρ_t we can choose how close we remain to the nominal distributions at different stages. Setting $\boldsymbol{\eta}_t := c_t(\mathbf{x}_t,\boldsymbol{\xi}_t)$ at stages $t \in \mathcal{T}$ and using (2.6), the multistage RASO problem can be formulated as

$$\min_{\mathbf{x}_0\in\mathcal{X}_0(\boldsymbol{\xi}_0)} \quad c_0(\mathbf{x}_0,\boldsymbol{\xi}_0) + \mathcal{R}_{\rho_1}(\mathcal{Q}(\mathbf{x}_0,\boldsymbol{\xi})) \tag{2.7}$$

where

$$\mathcal{Q}(\mathbf{x}_0,\boldsymbol{\xi}) := \min_{\mathbf{x}_t\in\mathcal{X}_t(\mathbf{x}_{t-1},\boldsymbol{\xi}_t),\ t\in\mathcal{T}\setminus\{0\}} \quad \mathfrak{R}_{\rho_2,\ldots,\rho_T}\Big(c_1(\mathbf{x}_1,\boldsymbol{\xi}_1),\ldots,c_T(\mathbf{x}_T,\boldsymbol{\xi}_T)\Big). \tag{2.8}$$

Let $\mathbf{x}_0^*$ and z^* be an optimal first-stage solution and the optimal value of (2.7)–(2.8), respectively. Our first approach, which we refer to as *first-level LB*, is formed as follows. Consider the collection of subsets $\Omega = \cup_{g=1}^{m_l}\Omega_g^{(l)}$ and its induced σ-algebra $\mathscr{G}$. We solve problem (2.7)–(2.8) with sample space $\Omega_g^{(l)}$ where $\mathcal{R}_{\rho_1}$ is replaced by $\mathcal{R}_{\bar{\rho}_g}^{(l)}$ and let $z_g^{*(l)}$ be its optimal value. Also let $\boldsymbol{\zeta}_{LB} := \{z_g^{*(l)}\}_{g=1}^{m_l}$ be a $\mathscr{G}$-measurable random variable with probabilities $\pi_g^{(l)}$, $g \in [m_l]$. A LB on z^* can be obtained by applying the LB risk measure $\widetilde{\mathcal{R}}_{\bar{\bar{\rho}}}^{\mathscr{G}}$ introduced in the previous section to $\boldsymbol{\zeta}_{LB}$, hence computing $\widetilde{\mathcal{R}}_{\bar{\bar{\rho}}}^{\mathscr{G}}(\boldsymbol{\zeta}_{LB})$. There is no need to make any changes to the radii from stage 2 to stage T; hence the name 'first-level' LB. Observe that the scenario tree can be dissected at finer partitions than just the first stage (*e.g.*, a single scenario at every stage $t \in \mathcal{T}$), but the convolution is performed only at the first stage.

Proposition 6. (First-level LB criteria for multistage problems). Given problem (2.7)–(2.8), assume that the risk measure at each stage is induced by a ϕ-divergence. Consider risk measure $\widetilde{\mathcal{R}}_{\bar{\bar{\rho}}}^{\mathscr{G}} : \mathfrak{L}_\infty(\Omega,\mathscr{G},Q) \to \mathbb{R}$ and the one-step conditional risk measure $\widetilde{\mathcal{R}}_{\bar{\boldsymbol{\rho}}}^{\mathscr{F}|\mathscr{G}} : \mathfrak{L}_\infty(\Omega,\mathscr{F},Q) \to \mathfrak{L}_\infty(\Omega,\mathscr{G},Q)$ (*i.e.*, $\mathcal{R}_{\bar{\rho}_g}^{(l)} : \mathfrak{L}_\infty(\Omega,\sigma(\Omega_g^{(l)}),Q) \to \mathbb{R}$, $\Omega_g^{(l)}$, $g \in [m_l]$), where these risk measures are induced by the same type of measure of similarity Δ. If $\bar{\bar{\rho}}$ and $\bar{\rho}_{max}$ satisfy the criteria from one of the Propositions 1–4 with $\rho = \rho_1$, then $z^* \geq \widetilde{\mathcal{R}}_{\bar{\bar{\rho}}}^{\mathscr{G}}(\boldsymbol{\zeta}_{LB})$.

Proof. If $\mathbf{x}_0^*$ is an optimal first-stage solution of (2.7)–(2.8), then it is a feasible first-stage solution for each subgroup problem, $g \in [m_l]$. Thus, we have

$$c_0(\mathbf{x}_0^*,\boldsymbol{\xi}_0) + \mathcal{R}_{\bar{\rho}_g}^{(l)}(\mathcal{Q}(\mathbf{x}_0^*,\boldsymbol{\xi})) \geq z_g^{*(l)},\ g \in [m_l],$$

or equivalently

$$c_0(\mathbf{x}_0^*,\boldsymbol{\xi}_0) + \widetilde{\mathcal{R}}_{\bar{\boldsymbol{\rho}}}^{\mathscr{F}|\mathscr{G}}(\mathcal{Q}(\mathbf{x}_0^*,\boldsymbol{\xi})) \geq \boldsymbol{\zeta}_{LB}.$$

Both sides of this inequality are $\mathscr{G}$-measurable, and since $\widetilde{\mathcal{R}}^{\mathscr{G}}_{\tilde{\rho}}$ is a coherent risk measure that satisfies the monotonicity property (see Section 2.2.5), we obtain

$$\widetilde{\mathcal{R}}^{\mathscr{G}}_{\tilde{\rho}}\left(c_0(\mathbf{x}_0^*, \boldsymbol{\xi}_0) + \widetilde{\mathcal{R}}^{\mathscr{F}|\mathscr{G}}_{\boldsymbol{\rho}}\left(\mathcal{Q}(\mathbf{x}_0^*, \boldsymbol{\xi})\right)\right) \geq \widetilde{\mathcal{R}}^{\mathscr{G}}_{\tilde{\rho}}(\boldsymbol{\zeta}_{LB}).$$

We can now apply translation equivariance property (see Section 2.2.5) to the left-hand side of above inequality to get

$$\widetilde{\mathcal{R}}^{\mathscr{G}}_{\tilde{\rho}}\left(\widetilde{\mathcal{R}}^{\mathscr{F}|\mathscr{G}}_{\boldsymbol{\rho}}\left(c_0(\mathbf{x}_0^*, \boldsymbol{\xi}_0) + \mathcal{Q}(\mathbf{x}_0^*, \boldsymbol{\xi})\right)\right) \geq \widetilde{\mathcal{R}}^{\mathscr{G}}_{\tilde{\rho}}(\boldsymbol{\zeta}_{LB}).$$

Since the criteria from Propositions 1–4 are satisfied, we obtain

$$\mathcal{R}_{\rho_1}\left(c_0(\mathbf{x}_0^*, \boldsymbol{\xi}_0) + \mathcal{Q}(\mathbf{x}_0^*, \boldsymbol{\xi})\right) \geq \widetilde{\mathcal{R}}^{\mathscr{G}}_{\tilde{\rho}}\left(\widetilde{\mathcal{R}}^{\mathscr{F}|\mathscr{G}}_{\boldsymbol{\rho}}\left(c_0(\mathbf{x}_0^*, \boldsymbol{\xi}_0) + \mathcal{Q}(\mathbf{x}_0^*, \boldsymbol{\xi})\right)\right) \geq \widetilde{\mathcal{R}}^{\mathscr{G}}_{\tilde{\rho}}\left(\boldsymbol{\zeta}_{LB}\right).$$

Using once more the translation equivariance property, we reach

$$z^* = c_0(\mathbf{x}_0^*, \boldsymbol{\xi}_0) + \mathcal{R}_{\rho_1}\left(\mathcal{Q}(\mathbf{x}_0^*, \boldsymbol{\xi})\right) \geq \widetilde{\mathcal{R}}^{\mathscr{G}}_{\tilde{\rho}}\left(\boldsymbol{\zeta}_{LB}\right),$$

which concludes the proof. See also [106]. □

First-level bounding scheme cannot be applied to the Wasserstein distance case because we do not have a distance between subgroups in a multistage setting. Thus, we introduce the following definition which will allow us to reduce the computation of the distance between subgroups in a multistage setting only to a single stage, in a recursive way.

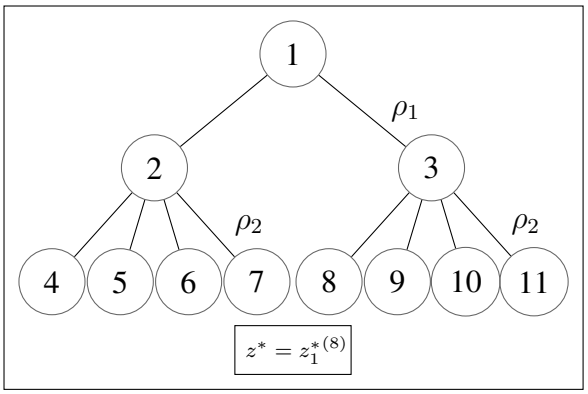
Definition 1. Let $\Omega^{(l)}_{t,g}$ denote set of nodes of subgroup $\Omega^{(l)}_g$ at stage $t \in \mathcal{T}$. We say the scenario tree $\mathfrak{T}$ is dissected up to stage $\tau \in \mathcal{T} \setminus \{0\}$ if:

1. for every subgroup $\Omega^{(l)}_g,\ g \in [m_l]$ all nodes at stage τ have the same ancestor, $i.e., a(n') = a(n''),\ n', n'' \in \Omega^{(l)}_{\tau,g},\ g \in [m_l]$;

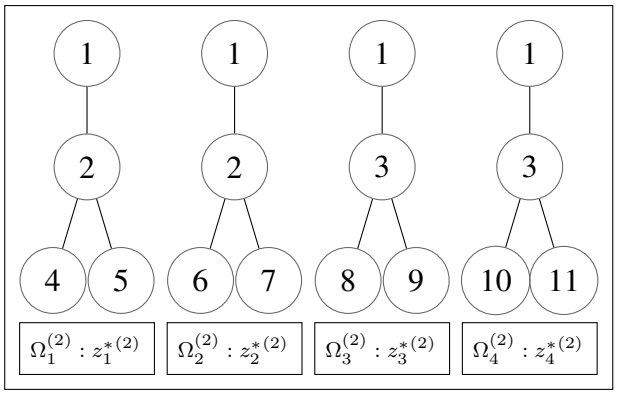
2. all the successors of stage-τ nodes belong to the same subgroup, $i.e., \mathscr{B}(n) \in \Omega^{(l)}_{\tau+1,g},\ n \in \Omega^{(l)}_{\tau,g},\ g \in [m_l], \tau \neq T$;

3. subgroups are disjoint, $i.e., \Omega^{(l)}_{\tau,g_1} \cap \Omega^{(l)}_{\tau,g_2} = \emptyset,\ g_1, g_2 \in [m_l],\ g_1 \neq g_2$.

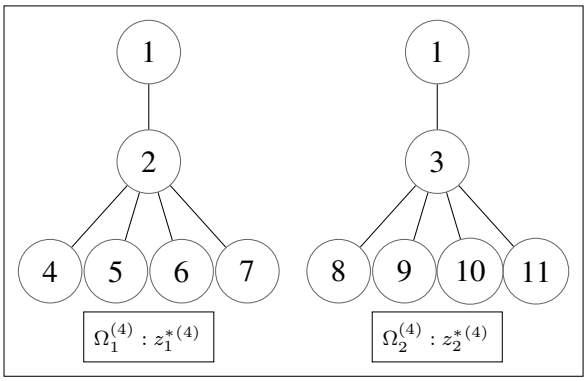
Notice that according to Definition 1 we split the ambiguity sets at stage τ while the ambiguity sets at subsequent stages are not modified. Furthermore, all the subgroups involved in the splitting of a given ambiguity set at stage τ share the same path up to stage $\tau - 1$. For example, in Figure 11b the scenario tree depicted in Figure 11a is dissected up to stage $\tau = 2$. Consequently, subgroups $\Omega^{(2)}_1$ and $\Omega^{(2)}_2$ share the same nodes 1 and 2 up to stage $\tau - 1 = 1$. Similarly, for subgroups $\Omega^{(2)}_3$ and $\Omega^{(2)}_4$. Analogously, Figure 11c shows a dissection up to stage $\tau = 1$.



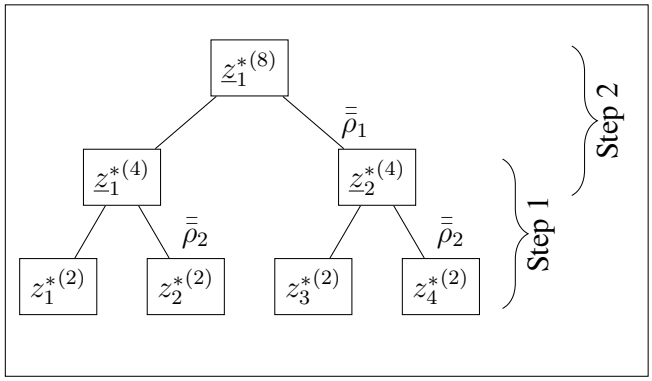
(a) Scenario tree $\mathfrak{T}$ with $\mathcal{T} := \{0, 1, 2\}$.



(b) Tree $\mathfrak{T}$ dissected up to stage $\tau = 2$.



(c) Tree $\mathfrak{T}$ dissected up to stage $\tau = 1$.



(d) Multi-level LB criteria.

Figure 11: Visual representation of the multi-level bounding scheme.

Notice that Proposition 6 can also be applied to the Wasserstein distance case if the scenario tree is dissected at $\tau = 1$ according to Definition 1. In this situation, we are able to compute distances among subgroups because the ambiguity sets of the subsequent stages are unchanged. For the other situations, we propose the following *multi-level* LB scheme, which works for both ϕ-divergences and the Wasserstein distance.

Proposition 7. (Multi-level LB criteria for multistage problems). Let the scenario tree $\mathfrak{T}$ be dissected up to stage τ in subgroups $\Omega_g^{(l)}$, $g \in [m_l]$, according to Definition 1. Let $z_g^{*(l)}$, $g \in [m_l]$ be the optimal values of problem (2.7)–(2.8) with sample space $\Omega_g^{(l)}$, where ρ_t is replaced by $\bar{\rho}_{t,g}$, $t = 1, \ldots, \tau$ and $\bar{\rho}_{t,max} := \max_{g \in [m_l]} \bar{\rho}_{t,g}$. Also let $\boldsymbol{\zeta}_{LB} := \{z_g^{*(l)}\}_{g=1}^{m_l}$ be a $\mathscr{G}$-measurable random variable with probabilities $\pi_g^{(l)}$, $g \in [m_l]$. If $\bar{\rho}_{t,g} = \rho_t$, $g \in [m_l]$ for all $t = \tau + 1, \ldots, T$ and $\bar{\bar{\rho}}_t$, $\bar{\rho}_{t,max}$ satisfy the criteria from one of the Propositions 1–5 with respect to ρ_t for all $t = 1, \ldots, \tau$, then

$$\widetilde{\mathcal{R}}_{\bar{\bar{\rho}}_1}^{\mathscr{F}_1|\mathscr{F}_0}\left(\widetilde{\mathcal{R}}_{\bar{\bar{\rho}}_2}^{\mathscr{F}_2|\mathscr{F}_1}\left(\cdots\left(\widetilde{\mathcal{R}}_{\bar{\bar{\rho}}_\tau}^{\mathscr{G}|\mathscr{F}_{\tau-1}}\left(\boldsymbol{\zeta}_{LB}\right)\right)\right)\right) \leq z^*$$

where

$$\widetilde{\mathcal{R}}_{\bar{\bar{\rho}}_\tau}^{\mathscr{G}|\mathscr{F}_{\tau-1}}\left(\boldsymbol{\zeta}_{LB}\right) := \left\{\widetilde{\mathcal{R}}_{\bar{\bar{\rho}}_\tau}^{\mathscr{F}^s}\left(\boldsymbol{\zeta}_{LB}^s\right)\right\}_{s=1}^{|\Omega_{\tau-1}|}$$

with $\Omega_s^{(k)}$, $s \in [|\Omega_{\tau-1}|]$ dissection of the scenario tree $\mathfrak{T}$ up to stage $\tau-1$, $\mathscr{S}^s$ the σ-algebra induced by the collection of subsets $\Omega_s^{(k)} = \bigcup_{\Omega_g^{(l)} \subseteq \Omega_s^{(k)}} \Omega_g^{(l)}$ and $\boldsymbol{\zeta}_{LB}^s := \{z_g^{*(l)}\}_{\Omega_g^{(l)} \subseteq \Omega_s^{(k)}}$.

Proof. Let $\Omega_g^{(l)}$, $g \in [m_l]$ be a dissection of the scenario tree $\mathfrak{T}$ up to stage τ and let $\Omega_s^{(k)}$, $s \in [|\Omega_{\tau-1}|]$ be a dissection of the scenario tree $\mathfrak{T}$ up to stage $\tau-1$. Let $\mathbf{x}_{i,s}^*$, $i = 0, \ldots, \tau-1$ optimal solutions of a group $\Omega_s^{(k)}$ be given. Then, they are a feasible solution for each subgroup problem $\Omega_g^{(l)} \subseteq \Omega_s^{(k)}$. Thus, for all $\Omega_g^{(l)} \subseteq \Omega_s^{(k)}$ we have

$$c_0(\mathbf{x}_{0,s}^*, \boldsymbol{\xi}_0) + \mathcal{R}_{\bar{\rho}_{1,g}}^{(l)}\left(c_1(\mathbf{x}_{1,s}^*, \boldsymbol{\xi}_1) + \mathcal{R}_{\bar{\rho}_{2,g}}^{(l)}\left(\ldots + \mathcal{R}_{\bar{\rho}_{\tau,g}}^{(l)}\left(\mathcal{Q}(\mathbf{x}_{\tau-1,s}^*, \boldsymbol{\xi})\right)\right)\right) \geq z_g^{*(l)}$$

with

$$\mathcal{Q}(\mathbf{x}_{\tau-1,s}^*, \boldsymbol{\xi}) := \min_{\substack{\mathbf{x}_\tau \in \mathcal{X}_\tau(\mathbf{x}_{\tau-1,s}^*, \boldsymbol{\xi}_t) \\ \mathbf{x}_t \in \mathcal{X}_t(\mathbf{x}_{t-1}, \boldsymbol{\xi}_t),\ t \in \{\tau+1, \ldots, T\}}} \mathfrak{R}_{\rho_{\tau+1}, \ldots, \rho_T}\left(c_\tau(\mathbf{x}_\tau, \boldsymbol{\xi}_\tau), \ldots, c_T(\mathbf{x}_T, \boldsymbol{\xi}_T)\right)$$

or equivalently defining $\boldsymbol{\zeta}_{LB}^s := \{z_g^{*(l)}\}_{\Omega_g^{(l)} \subseteq \Omega_s^{(k)}}$, $\bar{\boldsymbol{\rho}}_t^s = \{\bar{\rho}_{t,g}\}_{\Omega_g^{(l)} \subseteq \Omega_s^{(k)}}$ and $\mathscr{S}^s$ the σ-algebra induced by the collection of subsets $\Omega_g^{(l)}$ such that $\Omega_s^{(k)} = \bigcup_{\Omega_g^{(l)} \subseteq \Omega_s^{(k)}} \Omega_g^{(l)}$:

$$c_0(\mathbf{x}_{0,s}^*, \boldsymbol{\xi}_0) + \widetilde{\mathcal{R}}_{\bar{\boldsymbol{\rho}}_1^s}^{(k)|\mathscr{S}^s}\left(c_1(\mathbf{x}_{1,s}^*, \boldsymbol{\xi}_1) + \widetilde{\mathcal{R}}_{\bar{\boldsymbol{\rho}}_2^s}^{(k)|\mathscr{S}^s}\left(\ldots + \widetilde{\mathcal{R}}_{\bar{\boldsymbol{\rho}}_\tau^s}^{(k)|\mathscr{S}^s}\left(\mathcal{Q}(\mathbf{x}_{\tau-1,s}^*, \boldsymbol{\xi})\right)\right)\right) \geq \boldsymbol{\zeta}_{LB}^s$$

Both sides of the inequality are $\mathscr{S}^s$-measurable. Then, since $\widetilde{\mathcal{R}}_{\bar{\bar{\rho}}_\tau}^{\mathscr{S}^s}$ is a coherent risk measure that satisfies the monotonicity property, we obtain

$$\widetilde{\mathcal{R}}_{\bar{\bar{\rho}}_\tau}^{\mathscr{S}^s}\left(c_0(\mathbf{x}_{0,s}^*, \boldsymbol{\xi}_0) + \widetilde{\mathcal{R}}_{\bar{\boldsymbol{\rho}}_1^s}^{(k)|\mathscr{S}^s}\left(c_1(\mathbf{x}_{1,s}^*, \boldsymbol{\xi}_1) + \widetilde{\mathcal{R}}_{\bar{\boldsymbol{\rho}}_2^s}^{(k)|\mathscr{S}^s}\left(\ldots + \widetilde{\mathcal{R}}_{\bar{\boldsymbol{\rho}}_\tau^s}^{(k)|\mathscr{S}^s}\left(\mathcal{Q}(\mathbf{x}_{\tau-1,s}^*, \boldsymbol{\xi})\right)\right)\right)\right) \geq \widetilde{\mathcal{R}}_{\bar{\bar{\rho}}_\tau}^{\mathscr{S}^s}(\boldsymbol{\zeta}_{LB}^s).$$

Because of translation equivariance property we have

$$\sum_{i=0}^{\tau-1} c_i(\mathbf{x}_{i,s}^*, \boldsymbol{\xi}_i) + \widetilde{\mathcal{R}}_{\bar{\bar{\rho}}_\tau}^{\mathscr{S}^s}\left(\widetilde{\mathcal{R}}_{\bar{\boldsymbol{\rho}}_\tau^s}^{(k)|\mathscr{S}^s}\left(\mathcal{Q}(\mathbf{x}_{\tau-1,s}^*, \boldsymbol{\xi})\right)\right) \geq \widetilde{\mathcal{R}}_{\bar{\bar{\rho}}_\tau}^{\mathscr{S}^s}(\boldsymbol{\zeta}_{LB}^s) := \underline{z}_s^{*(k)}$$

and since by hypothesis $\bar{\bar{\rho}}_\tau$, $\bar{\rho}_{\tau,max}$ satisfy the criteria from one of the Propositions 1–5 with respect to ρ_τ, we get

$$z_s^{*(k)} = \sum_{i=0}^{\tau-1} c_i(\mathbf{x}_{i,s}^*, \boldsymbol{\xi}_i) + \mathcal{R}_{\rho_\tau}^{(k)}\left(\mathcal{Q}(\mathbf{x}_{\tau-1,s}^*, \boldsymbol{\xi})\right) \geq \underline{z}_s^{*(k)}.$$

Repeating for all $s \in [|\Omega_{\tau-1}|]$ leads to

$$\left[z_1^{*(k)}, \ldots, z_{|\Omega_{\tau-1}|}^{*(k)}\right]^\top \geq \left[\underline{z}_1^{*(k)}, \ldots, \underline{z}_{|\Omega_{\tau-1}|}^{*(k)}\right]^\top = \widetilde{\mathcal{R}}_{\bar{\bar{\rho}}_\tau}^{\mathscr{G}|\mathscr{F}_{\tau-1}}(\boldsymbol{\zeta}_{LB}).$$

Let $\Omega_d^{(j)}$, $d \in [|\Omega_{\tau-2}|]$ be a dissection of the scenario tree $\mathfrak{T}$ up to stage $\tau-2$. Let $\mathbf{x}_{i,d}^*$, $i = 0, \ldots, \tau-2$ optimal solutions of group $\Omega_d^{(j)}$ be given. Then, they are a feasible solution for each subproblem $\Omega_s^{(k)} \subseteq \Omega_d^{(j)}$. Following the steps above and defining $\underline{\boldsymbol{\zeta}}_{LB}^d := \{\underline{z}_s^{*(k)}\}_{\Omega_s^{(k)} \subseteq \Omega_d^{(j)}}$ and $\mathscr{D}^d$ the σ-algebra induced by the collection of subsets $\Omega_d^{(j)} = \bigcup_{\Omega_s^{(k)} \subseteq \Omega_d^{(j)}} \Omega_s^{(k)}$ we have:

$$z_d^{*(j)} = \sum_{i=0}^{\tau-2} c_i(\mathbf{x}_{i,d}^*, \boldsymbol{\xi}_i) + \mathcal{R}_{\rho_{\tau-1}}^{(j)}\left(\mathcal{Q}(\mathbf{x}_{\tau-2,d}^*, \boldsymbol{\xi})\right) \geq \widetilde{\mathcal{R}}_{\bar{\bar{\rho}}_{\tau-1}}^{\mathscr{D}^d}(\underline{\boldsymbol{\zeta}}_{LB}^d) := \underline{z}_d^{*(j)}.$$

Repeating for all $d \in [|\Omega_{\tau-2}|]$ leads to

$$\left[z_1^{*(j)}, \ldots, z_{|\Omega_{\tau-2}|}^{*(j)}\right]^\top \geq \left[\underline{z}_1^{*(j)}, \ldots, \underline{z}_{|\Omega_{\tau-2}|}^{*(j)}\right]^\top = \widetilde{\mathcal{R}}_{\bar{\rho}_{\tau-1}}^{\mathscr{F}_{\tau-1}|\mathscr{F}_{\tau-2}}\left(\widetilde{\mathcal{R}}_{\bar{\rho}_\tau}^{\mathscr{G}|\mathscr{F}_{\tau-1}}(\zeta_{LB})\right).$$

Repeating the same procedure going backwards for other $\tau - 2$ times, the thesis follows. □

See Figure 11d for a graphical representation of the multi-level bounding approach where the tree of Figure 11a has been dissected at stage $\tau = 2$ and, subsequently, at stage $\tau - 1 = 1$.

Remark. When the multi-level LB scheme is applied to the Wasserstein case, the distance at time τ between groups g_1, g_2 is computed as follows:

$$d_{g_1,g_2} = \begin{cases} \max\limits_{i \in \Omega_{\tau,g_1}^{(l)},\, k \in \Omega_{\tau,g_2}^{(l)}} \{d_{i,k}\} & \text{when} \quad g_1 \neq g_2 \\ 0 & \text{when} \quad g_1 = g_2 \end{cases}$$

where g_1 and g_2 are chosen such that $\forall n_1 \in \Omega_{\tau,g_1}^{(l)}, n_2 \in \Omega_{\tau,g_2}^{(l)} : a(n_1) = a(n_2)$ and $d_{i,k}$ is the distance between nodes i and k.

Remark. The bounding methodology just described is also applicable for the case where the risk measure is applied to the whole scenario cost as a time-inconsistent objective function, given as:

$$\mathfrak{R}_\rho^{whole}(\boldsymbol{\eta}_0, \ldots, \boldsymbol{\eta}_T) := \boldsymbol{\eta}_0 + \mathcal{R}_\rho(\boldsymbol{\eta}_1 + \boldsymbol{\eta}_2 + \ldots + \boldsymbol{\eta}_T).$$

In this case we would apply the first-level bounding scheme.

2.3.6 Upper Bounds for Multistage Optimization Problems

Finding an upper bound of an optimization problem is of critical importance when an optimal solution is not available.

In general, upper bounds are obtained by constraining some decision variables to be equal to pre-determined fixed values. In this work, upper bounds are obtained by using optimal solutions of single scenario subproblems. Using the procedure described in Section 2.3.1, we solve each single scenario group $\Omega_g^{(1)}$, $g \in [|\Omega_T|]$ obtaining $(\hat{\mathbf{x}}_{0,g}, \ldots, \hat{\mathbf{x}}_{T,g})$ as optimal solution. Let UB_g^t, $t \in \mathcal{T}$ be the optimal value of the original problem (2.1) where the variables up to stage t are set to $\hat{\mathbf{x}}_{i,g}$ for $i = 0, \ldots, t$. From an algorithmic perspective, this approach requires us to solve problems of smaller dimension than the original one. The best available upper bound is obtained by taking the minimum value of UB_g^t over all $g \in [|\Omega_T|]$, *i.e.*, $UB^t := \min_{g \in [|\Omega_T|]} UB_g^t$. See [103] for the formal definition and the proof.

2.4 Case Study: a Multistage Production Problem

2.4.1 Formulation

To show the effectiveness of the proposed approach, we consider a mixed-integer variant of the inventory management problem introduced in [105]. The problem can be summarized as follows. Consider a single product inventory system, comprised of a warehouse and a factory equipped with production machinery. At each time step $t = 0, \dots, T-1$, production can be performed by starting up machinery. Random demands coming from customers have to be satisfied from the existing inventory. If the random demand exceeds the stock, it will be satisfied by rapid orders from a different source that come at a higher price. The goal is to minimize the total costs of the factory for the entire planning period. In addition to the scenario tree and nominal probability notation defined in Section 2.2.1 (recall q_t^n denotes nominal probability of node n at stage t), we use the following notation for this problem.

Deterministic parameters

c_t : unit production cost at the factory at time $t \in \mathcal{T} \setminus \{T\}$;

k_t : machinery start-up up cost at time $t \in \mathcal{T} \setminus \{T\}$;

b_t : unit procurement cost from another retailer at time $t \in \mathcal{T} \setminus \{0\}$;

s_t : unit selling price at time $t \in \mathcal{T} \setminus \{0\}$;

h_t : unit inventory holding cost from time t to $t+1$, $t \in \mathcal{T} \setminus \{T\}$;

o : unit final value of the inventory;

e_t : maximal production capacity of factory at time $t \in \mathcal{T} \setminus \{T\}$;

v_1 : amount of the product in the warehouse at root node 1.

Stochastic parameters

ξ_n : stochastic demand for the product at node $n \in \Omega_t$, $t \in \mathcal{T}$.

Decision variables

$x_n \in \mathbb{R}_+$: amount to be produced by the factory at node $n \in \Omega_t$, $t \in \mathcal{T} \setminus \{T\}$;

$y_n \in \{0,1\}$: machinery start-up indicator at node $n \in \Omega_t$, $t \in \mathcal{T} \setminus \{T\}$;

$v_n \in \mathbb{R}$: amount of the product in the warehouse at node $n \in \Omega_t$, $t \in \mathcal{T} \setminus \{0\}$;

$v_n^+ / v_n^- \in \mathbb{R}_+$: positive/negative part of v_n;

$F_n \in \mathbb{R}$: auxiliary cost variable at node $n \in \Omega_t$, $t \in \mathcal{T} \setminus \{0\}$.

For every node $n \in \Omega_t$, $t \in \mathcal{T} \setminus \{T\}$, if v_n is positive (*i.e.*, $v_n = v_n^+$) an inventory holding cost $h_t v_n^+$ is paid to carry the stock to the next period. Otherwise, for every node $n \in \Omega_t$, $t \in \mathcal{T} \setminus \{0\}$, if v_n is negative (*i.e.*, $v_n = v_n^-$) a procurement cost $b_t v_n^-$ to buy extra stock from another retailer is incurred. Finally, for every leaf node $n \in \Omega_T$, the final stock is valued at ov_n^+. The multistage mixed-integer risk-neutral stochastic model is

$$\min \quad c_0 x_1 + k_0 y_1 + h_0 v_1 + \sum_{t=1}^{T} \sum_{n \in \Omega_t} q_t^n F_n \tag{2.9}$$

$$\text{s.t.} \quad F_n = c_t x_n + k_t y_n + h_t v_n^+ + b_t v_n^- - s_t \xi_n, \qquad n \in \Omega_t,\ t \in \mathcal{T} \setminus \{0, T\} \tag{2.10}$$

$$F_n = b_T v_n^- - s_T \xi_n - o v_n^+, \qquad n \in \Omega_T \tag{2.11}$$

$$0 \le x_n \le e_t y_n, \qquad n \in \Omega_t,\ t \in \mathcal{T} \setminus \{T\} \tag{2.12}$$

$$v_n = v_{a(n)}^+ + x_{a(n)} - \xi_n, \qquad n \in \Omega_t,\ t \in \mathcal{T} \setminus \{0\} \tag{2.13}$$

$$v_n = v_n^+ - v_n^-, \qquad n \in \Omega_t,\ t \in \mathcal{T} \setminus \{0\} \tag{2.14}$$

$$v_n^+ \ge 0,\ v_n^- \ge 0, \qquad n \in \Omega_t,\ t \in \mathcal{T} \setminus \{0\} \tag{2.15}$$

$$y_n \in \{0, 1\}, \qquad n \in \Omega_t,\ t \in \mathcal{T} \setminus \{T\}. \tag{2.16}$$

The objective function (2.9) and constraints (2.10) and (2.11) denote the expected total cost obtained from production, procurement cost from external retailers, and inventory holding, as well as the profits from selling and the final value of the inventory. Constraint (2.12) imposes lower and upper levels on the factory production. Constraints (2.13), (2.14), and (2.15) define the dynamics of the inventory. Finally, constraint (2.16) defines the binary decision variables related to starting up the machinery.

We assume that the distribution of the scenario process is described by a six-stage ($T = 5$) scenario tree with 5 branches from the root, 4 from each of the second stage nodes, and 3 from each of the third, fourth, fifth stages nodes, resulting in $|\Omega_T| = 5 \times 4 \times 3 \times 3 \times 3 = 540$ scenarios and 806 nodes; see [104] for details on scenario tree generation.

The value of the process at the root node ($n = 1$) is $\xi_1 = 65$. At each period $t = 0, 1, 2, 3, 4$, the maximal production capacity of the factory is $e_t = 567$ units and the setup cost is $k_t = 75$. The initial inventory is $v_1 = 10$, the final value of the inventory is $o = 2$ per unit, and the values of production price c_t, selling price s_t, inventory holding cost h_t and procurement cost b_t at time period t are presented in Table 11.

t	0	1	2	3	4	5
c_t	3.5	3.6	2.3	2.8	3	-
s_t	-	10.7	10.5	10.9	10.6	10
h_t	2	1.9	2.1	2.2	2.1	
b_t	-	4	3.1	4.9	7	7.5

Table 11: Production price c_t, selling price s_t, holding cost h_t from time t to time $t+1$, and procurement cost b_t for extra stock from another retailer at time t.

2.4.2 Computation of Bounds

This section presents computational results on a DRO version of the production problem described above, considering different ambiguity sets using ϕ-divergences (VD and the modified χ^2 distance) and the Wasserstein distance. All the considered multistage DRO are implemented in a nested fashion, and we use the same value of the radii $\rho_t = 0.50$ over all stages $t \in \mathcal{T} \setminus \{0\}$. The problems derived from our case study were solved under AMPL environment using the CPLEX solver 12.8.0.0. Computations have been performed on a 64-bit machine with 8 GB of RAM and a 1.8 GHz Intel i7 processor.

Variation distance case

Table 12 lists the LBs obtained by applying Proposition 6(first-level LB) using VD. We choose subsets $\Omega_g^{(l)}$ to be disjoint with $l = 1, 3, 9, 27, 54, 108$. The instance $l = 540$ refers to the original problem, which we report as a benchmark. The first group of bounds $(l = 108, m_l = 5)$ has been obtained by solving $m_l = 5$ subproblems, each composed of $l = 108$ consecutive scenarios. The group of bounds mainly follow the structure of the scenario tree. Only at $l = 54$, we split the 5 scenarios at $t = 1$ individually and group the 4 scenarios at $t = 2$ two by two consecutively. When $l = 1$ and $m_l = 540$, each subgroup forms a deterministic problem with only one scenario. According to Proposition 2, we set $\bar{\rho}_g = \bar{\rho}_{max}$, $g \in [m_l]$ and choose the combinations $(\bar{\bar{\rho}}, \bar{\rho}_{max})$ with $\bar{\bar{\rho}} \in \{0.00, 0.25, 0.50\}$ and $\bar{\rho}_{max} = \frac{\rho_1 - \bar{\bar{\rho}}}{1+\bar{\bar{\rho}}}$. The overall problem, *i.e.*, the full tree with 540 scenarios, is solved within 14.125 seconds and with optimal value $z^* = -1523.49$. To measure the quality of the obtained LBs (LB), an optimality gap information is computed as $\%GAP = \frac{z^* - LB}{z^*} \cdot 100$. In the following, for each dissection with given cardinality l, the best lower bound is highlighted in bold.

From numerical results of Table 12 obtained considering disjoint subgroups, we observe that the tightest bounds are achieved for greater values of $\bar{\bar{\rho}}$ and l, at the cost of increasing CPU running times. Indeed, overall, the best calculated LB is given by -1528.74 (obtained setting $\bar{\bar{\rho}} = 0.50$, $\bar{\rho}_{max} = 0.00$ when $l = 108$ and $m_l = 5$), while the worst LB is given by -1836.90 (obtained setting $\bar{\bar{\rho}} = 0.00$,

$\bar{\rho}_{max} = 0.50$ when $l = 1$ and $m_l = 540$, which is the partition into atoms). Results show monotonic increases in CPU time per subproblem with both the dimension of each subproblem (cardinality l) and the values of $\bar{\bar{\rho}}$. These results also show that very high-quality LBs can be obtained saving considerable time with respect to the original DRO problem. For example, when $l = 54$ a $\%GAP = -0.50\%$ can be achieved with about 4 times faster overall computation time.

l	m_l	$\bar{\bar{\rho}}$	$\bar{\rho}_{max}$	*LB*	*CPU time overall*	*CPU time per subpr.*	$\%GAP$
540	1	0.00	0.50	−1523.49	14.125	14.250	-
108	5	0.00	0.50	−1544.66	5.469	1.094	−1.39%
		0.25	0.20	−1536.16	5.797	1.159	−0.83%
		0.50	**0.00**	**−1528.74**	**5.938**	**1.188**	**−0.34%**
54	10	0.00	0.50	−1591.31	3.281	0.328	−4.45%
		0.25	0.20	−1557.46	3.478	0.348	−2.23%
		0.50	**0.00**	**−1531.07**	**3.563**	**0.356**	**−0.50%**
27	20	0.00	0.50	−1623.40	2.063	0.103	−6.56%
		0.25	0.20	−1575.28	2.344	0.117	−3.40%
		0.50	**0.00**	**−1543.54**	**2.656**	**0.133**	**−1.32%**
9	60	0.00	0.50	−1703.05	3.266	0.054	−11.79%
		0.25	0.20	−1632.46	3.625	0.060	−7.15%
		0.50	**0.00**	**−1581.98**	**4.109**	**0.068**	**−3.84%**
3	180	0.00	0.50	−1780.91	8.484	0.047	−16.90%
		0.25	0.20	−1690.56	9.563	0.053	−10.97%
		0.50	**0.00**	**−1622.65**	**13.359**	**0.074**	**−6.51%**
1	540	0.00	0.50	−1836.90	20.336	0.038	−20.57%
		0.25	0.20	−1735.04	22.266	0.041	−13.89%
		0.50	**0.00**	**−1656.70**	**27.141**	**0.050**	**−8.74%**

Table 12: Collections of LBs with disjoint subsets $\Omega_g^{(l)}$ obtained by applying Proposition 6 (first-level LB) to the multistage inventory problem with VD.

Table 13 provides detailed results obtained by keeping fixed the worst scenario (ω_1) in all subsets $\Omega_g^{(l)}$. VD focuses on a convex combination of CVaR and the worst case [77, 127]. So when the worst-case scenario is fixed at each subgroup, we may get better LBs. The cardinality l of each subproblem has been chosen to have the ratio $m_l = \frac{540-1}{l-1} \in \mathbb{N}$ and, specifically, we consider the values $m_l \in \{2, 8, 12, 50, 78\}$. The combinations $(\bar{\bar{\rho}}, \bar{\rho}_{max})$ are chosen as for the disjoint case. Results

l	m_l	$\bar{\overline{\rho}}$	$\bar{\rho}_{max}$	*LB*	*CPU time overall*	*CPU time per subpr.*	$\%GAP$
540	1	0.00	0.50	−1523.49	14.125	14.250	-
78	7	**0.00**	**0.50**	**−1544.10**	**4.844**	**0.692**	**−1.35%**
		0.25	0.20	−1544.43	5.469	0.781	−1.37%
		0.50	0.00	−1548.96	5.734	0.819	−1.67%
50	11	0.00	0.50	−1564.61	4.453	0.405	−2.70%
		0.25	0.20	−1562.56	4.859	0.442	−2.57%
		0.50	**0.00**	**−1571.56**	**5.031**	**0.457**	**−3.16%**
12	49	0.00	0.50	−1624.06	4.500	0.092	−6.60%
		0.25	0.20	−1608.13	4.844	0.099	−5.56%
		0.50	**0.00**	**−1583.24**	**5.563**	**0.114**	**−3.92%**
8	77	0.00	0.50	−1642.35	6.641	0.086	−7.80%
		0.25	0.20	−1624.56	7.563	0.098	−6.63%
		0.50	**0.00**	**−1594.00**	**7.672**	**0.100**	**−4.63%**
2	539	0.00	0.50	−1724.44	30.609	0.057	−13.19%
		0.25	0.20	−1695.82	38.453	0.071	−11.31%
		0.50	**0.00**	**−1650.21**	**39.500**	**0.073**	**−8.32%**

Table 13: Collections of LBs obtained by keeping the worst scenario (ω_1) fixed in all subsets $\Omega_g^{(l)}$ and applying Proposition 6 (first-level LB) to the multistage inventory problem with VD.

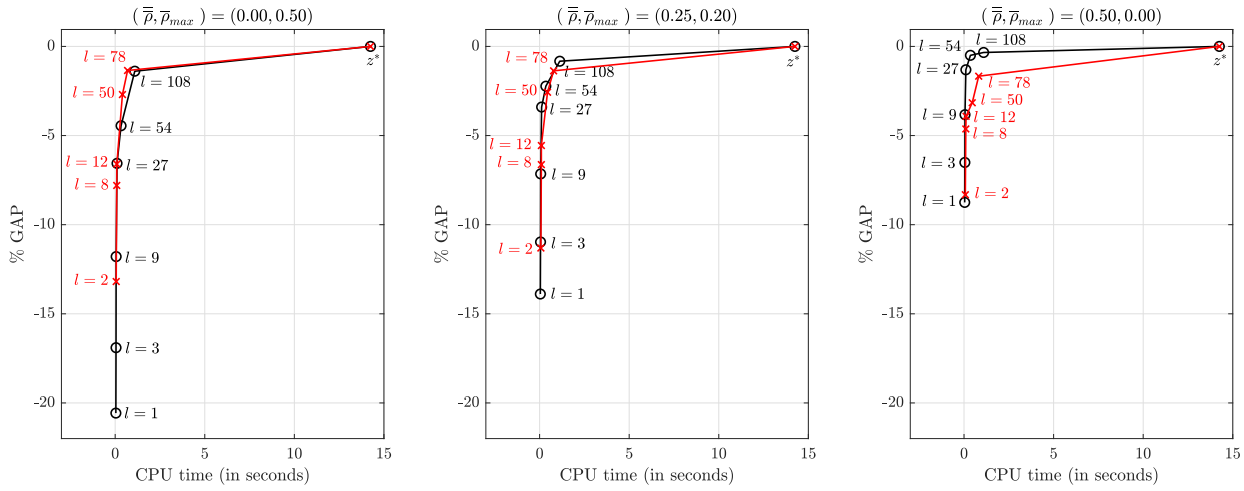


Figure 12: Percentage gaps from the optimal value z^* versus CPU time (per subproblem, in seconds) for VD for different combinations of $(\bar{\overline{\rho}}, \bar{\rho}_{max})$ under disjoint partitions (black lines) with cardinality $l = 1, 3, 9, 27, 54, 108$ (results refer to Table 12) and under subgroups with scenario ω_1 fixed (red lines) and cardinality $l = 2, 8, 12, 50, 78$ (results refer to Table 13).

show that fixing the worst scenario improves the quality of the LB for those partitions with small cardinalities l and greater values of $\bar{\rho}_{max}$ (see for instance $l = 1, 3, 9$ in Table 12 and $l = 2, 8$ in Table 13). It is also interesting to notice that when the worst scenario is fixed, although the tightest bounds are still obtained for greater values of l, tighter bounds are obtained setting progressively smaller value of $\bar{\bar{\rho}}$ and larger $\bar{\rho}_{max}$ (see, for instance, $l = 50$ and $l = 78$ in Table 13).

Figure 12 shows the LBs of Tables 12 and 13 versus increasing complexity measured in CPU seconds per subproblem for different combinations of $\left(\bar{\bar{\rho}}, \bar{\rho}_{max}\right)$. From the results we observe that the LBs improve monotonically in the number of scenarios l in each subproblem for the studied problem.

Modified χ^2 case

In Table 14, we construct collections of LBs applying Proposition 6 (first-level LB) to the multistage inventory problem with modified χ^2 distance. Subsets $\Omega_g^{(l)}$ are chosen to be disjoint, following the structure of the scenario tree with $l = 1, 3, 9, 27, 54, 108$ as before. According to Proposition 1 we set $\bar{\rho}_g = \bar{\rho}_{max}$, $g \in [m_l]$ and choose the combinations $(\bar{\bar{\rho}}, \bar{\rho}_{max})$ with $\bar{\bar{\rho}} \in \{0.00, 0.25, 0.50\}$ and $\bar{\rho}_{max} = \frac{\rho_1 - \bar{\bar{\rho}}}{1+\bar{\bar{\rho}}}$. The overall problem, *i.e.*, the full tree with 540 scenarios, was unsolvable within a time limit of 86400 CPU seconds (or 24 hours) and values of percentage deviations ($\%GAP$) with respect to the optimal value could not be explicitly computed. Therefore, to measure the quality of the obtained LB, a new optimality gap is computed as follows: $\%GAP^* = \frac{LB^* - LB}{LB^*} \cdot 100$, with LB^* representing the best observed lower bound. Being a problem too large to be solved exactly within the prespecified time limit, the bounding methodology proposed in this chapter is particularly helpful. It is worth noting that when $l = 108$ the solver also could not solve the subproblems within the time limit, and therefore $l = 108$ results are not reported for ease of presentation.

From the numerical results given in Table 14 and plotted in Figure 13, we observe that regardless of the cardinality l of each subproblem, the best strategy to get tighter LBs is to set the cardinality l and $\bar{\bar{\rho}}$ as large as possible. Indeed, overall, the best calculated LB is given by -1497.95 (obtained setting $\bar{\bar{\rho}} = 0.50$, $\bar{\rho}_{max} = 0.00$ when $l = 54$ and $m_l = 10$), although it requires considerable effort in terms of CPU time (25395.048 seconds overall). A drastic reduction in computation time can be obtained by using smaller subgroups of cardinality $l = 27$ without sacrificing the quality of the LB too much. Again by setting $\bar{\bar{\rho}} = 0.50$, $\bar{\rho}_{max} = 0.00$, a LB within 1.82% of the best LB is obtained in approximately 23.5 times faster overall computation time. On the other hand, the worst lower bound is given by -1836.90 (obtained setting $\bar{\bar{\rho}} = 0.00$, $\bar{\rho}_{max} = 0.50$ when $l = 1$ and $m_l = 540$ in just 211.031 CPU seconds. Results also show monotonic increases in CPU time per subproblem with both the dimension l of each subproblem and the values of $\bar{\bar{\rho}}$.

l	m_l	$\bar{\overline{\rho}}$	$\bar{\rho}_{max}$	*LB*	*CPU time overall*	*CPU time per subpr.*	$\%GAP^*$
540	1	0.00	0.50	-	-	-	-
54	10	0.00	0.50	-1559.56	25231.639	2523.164	-4.11%
		0.25	0.20	-1514.99	25392.418	2539.242	-1.14%
		0.50	**0.00**	$\mathbf{-1497.95}$	**25395.048**	**2539.505**	**-**
27	20	0.00	0.50	-1597.77	962.151	48.108	-6.66%
		0.25	0.20	-1543.95	1019.859	50.993	-3.07%
		0.50	**0.00**	$\mathbf{-1525.26}$	**1078.266**	**53.913**	$\mathbf{-1.82\%}$
9	60	0.00	0.50	-1686.68	118.181	1.970	-12.60%
		0.25	0.20	-1607.16	163.063	2.718	-7.29%
		0.50	**0.00**	$\mathbf{-1577.86}$	**165.172**	**2.753**	$\mathbf{-5.33\%}$
3	180	0.00	0.50	-1774.09	115.151	0.640	-18.43%
		0.25	0.20	-1678.34	196.344	1.091	-12.04%
		0.50	**0.00**	$\mathbf{-1642.55}$	**203.219**	**1.129**	$\mathbf{-9.65\%}$
1	540	0.00	0.50	-1836.90	211.031	0.391	-22.63%
		0.25	0.20	-1732.42	253.547	0.470	-15.65%
		0.50	**0.00**	$\mathbf{-1693.06}$	**299.313**	**0.554**	$\mathbf{-13.03\%}$

Table 14: Collections of LBs with disjoint subsets $\Omega_g^{(l)}$ obtained by applying Proposition 6 (first-level LB) to the multistage inventory problem with modified χ^2 (time limit = 86400 CPU sec.s).

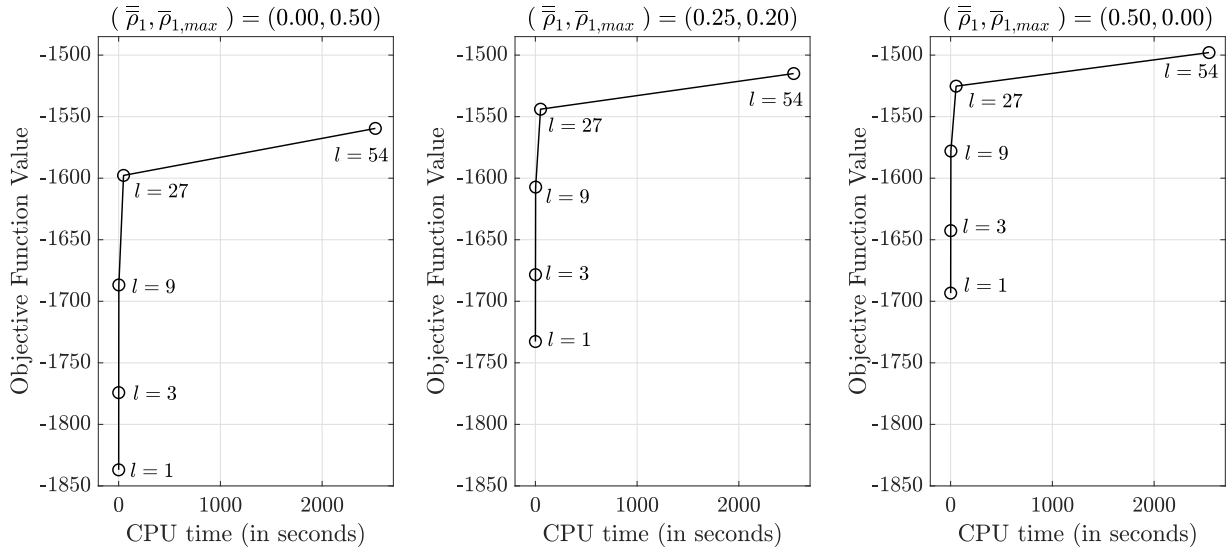


Figure 13: LBs versus CPU time (per subproblem, in seconds) for modified χ^2 for different combinations of $(\bar{\overline{\rho}}, \bar{\rho}_{max})$ and under disjoint partitions with cardinality $l = 1, 3, 9, 27, 54$ (results refer to Table 14).

We now apply the bounding scheme proposed in Proposition 7 (multi-level LB) to the multistage inventory problem with modified χ^2 distance. The results are reported in Table 15. Following this partition method, regardless of the cardinality l of each subproblem and given τ, the last stage where the scenario tree is partitioned, the best strategy to get tighter LBs is to set $\bar{\bar{\rho}}_t = \rho_t$, $t = 1, \ldots, \tau - 1$ (which are therefore the only results we report, for ease of exposition). We allow, instead, changes in $\bar{\bar{\rho}}_\tau$ and $\bar{\rho}_{\tau,max}$ taking values $0.00, 0.25$ and 0.50. Overall, the best lower bound is given by -1504.41, obtained setting $\bar{\bar{\rho}}_2 = 0.25$ and $\bar{\rho}_{2,max} = 0.20$ when $l = 54$, $m_l = 10$ and $\tau = 2$. Comparing these results with bounds of Table 14, we conclude that this multi-level bounding technique becomes particularly useful by allowing to get tighter LBs as partitions progressively contain smaller-dimension subproblems. For instance, at lower values of l better LBs are obtained with approximately similar overall computation times.

l	m_l	τ	$\{\bar{\bar{\rho}}_t\}_{t=1}^{\tau-1}$	$\{\bar{\rho}_{t,max}\}_{t=1}^{\tau-1}$	$\bar{\bar{\rho}}_\tau$	$\bar{\rho}_{\tau,max}$	LB	*CPU time overall*	*CPU time per subpr.*	$\%GAP^*$
540	1	-	-	-	-	-	-	-	-	-
54	10	2	0.50	0.00	0.00	0.50	-1536.64	27231.188	2723.119	-2.58%
					0.25	**0.20**	**-1504.41**	**28557.600**	**2855.760**	**-0.43%**
					0.50	0.00	-1505.45	30232.425	3023.243	-0.50%
27	20	2	0.50	0.00	0.00	0.50	-1571.74	1002.295	50.115	-4.93%
					0.25	0.20	-1523.02	1051.172	52.559	-1.67%
					0.50	**0.00**	**-1506.41**	**1112.942**	**55.647**	**-0.56%**
9	60	3	0.50	0.00	0.00	0.50	-1593.95	155.574	2.593	-6.41%
					0.25	0.20	-1538.46	157.103	2.618	-2.70%
					0.50	**0.00**	**-1518.76**	**161.375**	**2.690**	**-1.39%**
3	180	4	0.50	0.00	0.00	0.50	-1602.07	158.351	0.880	-6.95%
					0.25	0.20	-1555.76	158.932	0.883	-3.86%
					0.50	**0.00**	**-1538.21**	**160.105**	**0.889**	**-2.69%**
1	540	5	0.50	0.00	0.00	0.50	-1597.22	254.552	0.471	-6.63%
					0.25	0.20	-1560.59	255.380	0.473	-4.18%
					0.50	**0.00**	**-1547.23**	**256.578**	**0.475**	**-3.29%**

Table 15: Collections of LBs with disjoint subsets $\Omega_g^{(l)}$ obtained by applying Proposition 7 (multi-level LB) to the multistage inventory problem with modified χ^2 (time limit = 86400 CPU sec.s).

Given that the problem was too large to be solved exactly, to evaluate the quality of obtained lower bounds, we resort to the upper bounding methodology described in Section 2.3.6. The only upper bounds we were able to compute were $UB^3 = -1453.29$ (within 72131.000 CPU seconds) and $UB^4 = -1411.91$ (within 125.844 CPU seconds). Although UB^3 performs better than UB^4, it

requires a considerable larger computational effort. All the other upper bounds ($UB^t,\ t = 0, 1, 2$) went out of memory because the number of fixed variables was not enough to reduce the dimension of the scenario tree to a computationally tractable size. The difference between the best upper and lower bounds obtained, respectively -1453.29 and -1497.95, is 44.66 (*i.e.*, of 2.98% of the LB) and gives information about the range where we should expect to find the total cost of the full DRO problem. Results confirm the goodness of the proposed lower bounds.

Wasserstein distance case

In Table 16, we construct collections of LBs by applying Proposition 7 (multi-level LB) for the Wasserstein distance. Given τ, the last stage where the tree is dissected, according to Proposition 7 we set $\bar{\bar{\rho}}_t = \rho_t$, $t = 1, \ldots, \tau - 1$ and choose the combinations $(\bar{\bar{\rho}}_\tau, \bar{\rho}_{\tau,max})$ with $\bar{\bar{\rho}}_\tau \in \{0.00, 0.25, 0.50\}$ and $\bar{\rho}_{\tau,max} = \rho_\tau - \bar{\bar{\rho}}_\tau$. For $t = \tau + 1, \ldots, T$ we set $\bar{\rho}_{t,g} = \rho_t$, $g \in [m_l]$. When $\tau = 1$ we work directly with $(\bar{\bar{\rho}}_1, \bar{\rho}_{1,max})$, see for instance Table 16 with $l = 108$ and $m_l = 5$. The overall problem, *i.e.*, the full tree with 540 scenarios, is solved within 85.810 seconds and with optimal value $z^* = -1706.63$. From the results in Table 16, we observe that, overall, the best calculated LB is given by -1711.93 (obtained setting $\bar{\bar{\rho}}_1 = 0.50$, $\bar{\rho}_{1,max} = 0.00$ when $l = 108$ and $m_l = 5$). On the contrary, the worst LB is given by -1808.79 (obtained setting $\bar{\bar{\rho}}_5 = 0.00$, $\bar{\rho}_{5,max} = 0.50$ when $l = 1$ and $m_l = 540$). Results still show monotonic increases in CPU time with both the dimension l of each subproblem and the values of $\bar{\bar{\rho}}_\tau$ for this problem.

Figure 14 shows the percentage deviation of the LBs reported in Table 16 versus increasing complexity measured in CPU seconds per subproblem and different combinations of $(\bar{\bar{\rho}}_\tau, \bar{\rho}_{\tau,max})$. We observe that LBs improve monotonically in the number of scenarios l in each subproblem.

2.4.3 Discussions

Some insights gained from the numerical experiments are as follows.

- Generally, the greater the number of scenarios per subproblem, the sharper the obtained LBs.
- For the first-level bounding scheme, when subtrees have a single node at time $t = 1$, it is best not to waste any of the robustness budget ρ_1 on $\bar{\rho}_{max}$. This is because, for the subproblem at time $t = 1$ there is not really an ambiguity set of distributions. Thus, in this case, the best LBs are obtained by setting $\bar{\rho}_{max} = 0$ and using the largest possible value of $\bar{\bar{\rho}}$. This can be seen, *e.g.*, in Tables 12, 14 where the 5 scenarios of the original tree at stage $t = 1$ are always split in a single-scenario manner. More generally, when subtrees have multiple nodes at time

l	m_l	τ	$\{\bar{\bar{\rho}}_t\}_{t=1}^{\tau-1}$	$\{\bar{\rho}_{t,max}\}_{t=1}^{\tau-1}$	$\bar{\bar{\rho}}_\tau$	$\bar{\rho}_{\tau,max}$	LB	*CPU time overall*	*CPU time per subpr.*	$\%GAP^*$
540	1	-	-	-	-	-	-1706.63	85.810	85.810	-
108	5	1	-	-	0.00	0.50	-1724.33	9.000	1.800	-1.04%
					0.25	0.25	-1718.13	9.797	1.959	-0.67%
					0.50	**0.00**	$\mathbf{-1711.93}$	**9.984**	**1.997**	$\mathbf{-0.31\%}$
54	10	2	0.50	0.00	**0.00**	**0.50**	$\mathbf{-1725.08}$	**5.569**	**0.557**	$\mathbf{-1.08\%}$
					0.25	0.25	-1726.07	5.840	0.584	-1.14%
					0.50	0.00	-1727.02	6.183	0.618	-1.20%
27	20	2	0.50	0.00	0.00	0.50	-1737.88	4.455	0.223	-1.83%
					0.25	0.25	-1735.05	4.672	0.234	-1.67%
					0.50	**0.00**	$\mathbf{-1732.22}$	**4.946**	**0.247**	$\mathbf{-1.50\%}$
9	60	3	0.50	0.00	0.00	0.50	-1763.31	9.075	0.151	-3.32%
					0.25	0.25	-1760.22	9.164	0.153	-3.14%
					0.50	**0.00**	$\mathbf{-1757.12}$	**9.414**	**0.157**	$\mathbf{-2.96\%}$
3	180	4	0.50	0.00	0.00	0.50	-1792.85	12.216	0.068	-5.05%
					0.25	0.25	-1790.31	12.261	0.068	-4.90%
					0.50	**0.00**	$\mathbf{-1787.78}$	**12.351**	**0.069**	$\mathbf{-4.75\%}$
1	540	5	0.50	0.00	0.00	0.50	-1808.79	28.082	0.052	-5.99%
					0.25	0.25	-1807.04	28.173	0.052	-5.88%
					0.50	**0.00**	$\mathbf{-1805.29}$	**28.305**	**0.052**	$\mathbf{-5.78\%}$

Table 16: Collections of LBs with disjoint subsets $\Omega_g^{(l)}$ ($l = 1, 3, 9, 27, 54, 108$) obtained applying Proposition 7 (multi-level LB) to the multistage inventory problem with Wasserstein distance.

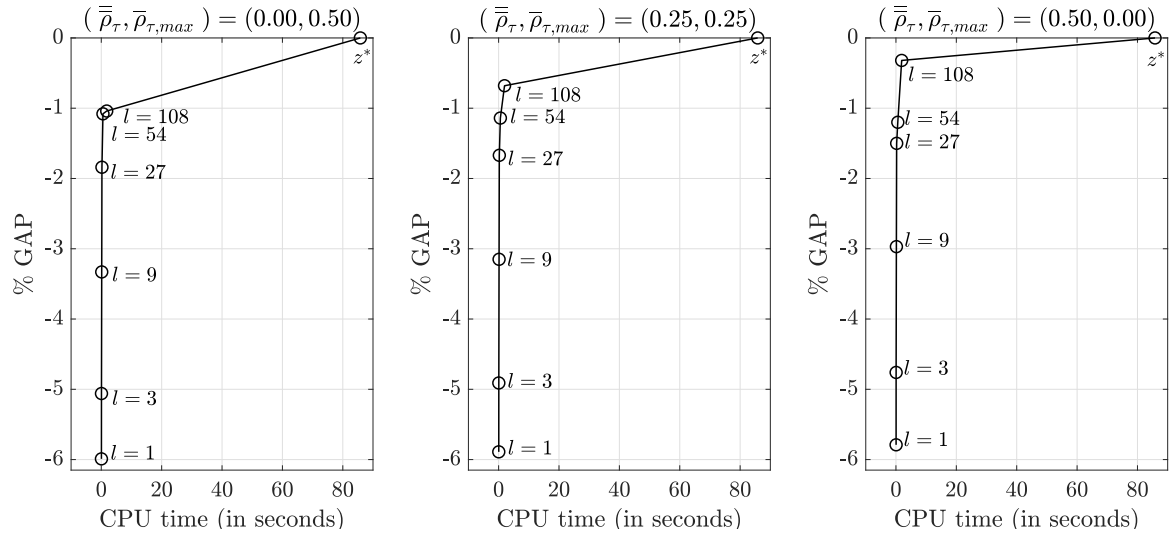


Figure 14: Percentage gaps from the optimal value z^* versus CPU time (per subproblem, in seconds) for Wasserstein distance for different combinations of $(\bar{\bar{\rho}}_\tau, \bar{\rho}_{\tau,max})$ and under partitions with cardinality $l = 1, 3, 9, 27, 54, 108$ (results refer to Table 16).

$t = 1$, numerical results show that more importance should be assigned to $\bar{\bar{\rho}}$ at the expense of $\bar{\rho}_{max}$ as the cardinality l of subgroups decreases (see, *e.g.*, Table 13 when $l = 2, 8, 12$), while progressively more importance should be assigned to $\bar{\rho}_{max}$ at the expense of $\bar{\bar{\rho}}$ as the cardinality l of subgroups increases (see, *e.g.*, Table 13 when $l = 50, 78$).

- Similarly, for the multi-level bounding scheme, when subtrees have a single node at time τ, it is best not to waste any of the robustness budget ρ_τ on $\bar{\rho}_{\tau,max}$ and use the largest possible value of $\bar{\bar{\rho}}_\tau$. This can be seen, *e.g.*, in Tables 15, 16 when $l = 1, 3, 9, 27, 108$. Contrariwise, when there are multiple nodes at time τ more importance should be assigned to $\bar{\rho}_{\tau,max}$ at the expense of $\bar{\bar{\rho}}_\tau$ (see, *e.g.*, Tables 15, 16 when $l = 54$). For dissections with smaller cardinality l, the multi-level bounding scheme described in Proposition 7 appears to be more effective than the first-level bounding scheme described in Proposition 6, which is instead more useful when the number of scenarios in each subgroup is larger.

- We observe some gains in fixing a worst-case scenario using VD at small cardinalities l and higher values of $\bar{\rho}_{max}$, with a slight increase in computation time due to having slightly larger subproblems to solve. This strategy can be useful, *e.g.*, when using ϕ-divergences that can pop scenarios [6]. Divergences that can pop scenarios (like the CR power divergence with $\theta < 1$ and $\theta \to 1$) can make the worst-case scenarios to have positive probabilities even if they have a nominal probability of zero. Thus, in large-scale versions of such problems, due to computational bottleneck when small cardinalities l are needed, it may be possible to obtain better LBs by fixing worst-case scenarios.

- Finally, even though above we empirically observe monotonicity of LBs in the subgroups' cardinality l for fixed values of $(\bar{\bar{\rho}}, \bar{\rho}_{max})$, we found cases (not shown here for brevity) where monotonicity in l is not satisfied. This is in contrast to the risk-neutral stochastic optimization setting, where the monotonicity of the LBs in l is guaranteed [103].

2.5 Conclusions

In this work new LB criteria for multistage mixed-integer DRO problems—formed by creating ambiguity sets associated with various commonly used ϕ-divergences and the Wasserstein distance on a finite scenario tree—are derived. Conditions on the way the scenario tree is dissected and the convolutions are formed to ensure a LB on the optimal value are established. The scenario tree can be dissected either by disjoint partitions or by fixing certain scenarios in each subgroup, except for CR

power divergences with $\theta < 0$ and $\theta > 1$ and χ-divergence of order $a > 1$, for which the results are established under disjoint partitions. A comparison with classical upper bounds shows the effectiveness of the proposed LB criteria. Our results do not require any structural properties, and thus they are applicable to a broad class of problems. Numerical results on a multistage production problem show that high-quality LBs can be obtained with a small computation time using the proposed bounding methodology.

Future work could include combining the proposed bounds with sampling-based bounds (see, *e.g.*, [124]). Devising new hybrid sampling-based algorithms that could utilize the proposed bounds to be used within *Stochastic Dual Dynamic Programming* (SDDP) type algorithms (see, *e.g.*, [122]) merit further research. It would also be interesting to investigate if the concept of ineffective and effective scenarios, defined in [127], can be used to further increase the efficiency of the proposed bounding methods. Ineffective scenarios can be removed from the problem without altering the optimal value. Therefore, if such scenarios can be identified, these can significantly speed up the proposed bounds. Finally, it should be highlighted that the proposed approach has the important advantage of dividing a given problem into subproblems, the solution of which are independent from one another and might be easily parallelized. Such parallel implementations might significantly decrease running times and therefore merit further computational research.

Chapter 3. Assortment Optimization & Revenue Management with Dominated Alternatives

In collaboration with Anton J. Kleywegt[1].

3.1 Introduction

We consider a seller who has a set of products with fixed attributes, and who can offer any subset of these products in the market. The seller's *Assortment Optimization* (AO) problem is the problem of choosing a single subset (assortment) of these products to offer in the market with the objective of maximizing his/her expected profit. The considered AO problem applies in a setting in which each product has a given revenue and the supply (or inventory) of each product is not constraining. On the other hand, the seller's *Revenue Management* (RM) problem is the problem of choosing, over a given time horizon, what subset of these products to offer in the market at each point in time with the objective of maximizing the expected profit over the time horizon. Similar to the assortment optimization problem, the revenue management problem considered applies in a setting in which each product has a given profit per unit. Unlike the assortment optimization problem, in the revenue management problem it may be optimal to change the offered assortment over time. A possible reason is that the supplies of products are limited (*i.e.*, the inventory is constraining). See [110] and [168] for a comprehensive overview.

Modern AO and RM problems include a model of demand for each product as a function of the assortment that is offered. This model of demand (also known as *discrete choice model*) given any set of products that is offered, specifies the probability that a customer will choose each product in the set (or the fraction of customers who will choose each product in the set), as well as the probability that a customer chooses no product in the set. Various discrete choice models have been studied. Important characteristics of such discrete choice models are (a) how well their expressive abilities match the choice behavior being modeled (too little expressive ability is too restrictive to obtain a good model, whereas too much expressive ability often results in intractability and overfitting), (b) how easy it is to interpret the model, (c) how much data are needed to calibrate a useful model, (d) how tractable the model calibration problem is, and (e) how tractable the resulting assortment optimization and revenue management problems are. Of course, there are trade-offs between these model characteristics. One

[1]H. Milton Stewart School of Industrial & Systems Engineering, Georgia Institute of Technology
Atlanta, GA, USA. Anton@isye.gatech.edu

of the most widely used discrete choice models is the *Multinomial Logit* (MNL) model. The MNL model has many of the desirable properties (see [28]): it is easy to interpret the MNL model; the MNL model needs less data for calibration than more general models such as the nested logit model, the Markov chain choice model, and the mixed logit model; the model calibration problem is an unconstrained convex optimization problem that is much easier to solve than the calibration problems for the other models mentioned; and the resulting assortment optimization and revenue management problems are relatively easy to solve.

However, the MNL model also has shortcomings. Its major drawback arises when the structure that it imposes does not accurately fit the choice behavior being modeled. For example, one property of its structure is called *Independence from Irrelevant Alternatives* (IIA). This is the property that the relative choice probabilities of two alternatives do not depend on the presence or absence in the choice set of other alternatives (see [129]).

In this work we are concerned with another structural property that is shared by many discrete choice models, including the MNL model. This is the property that if the assortment contains products i and j, where i dominates j, then according to the model the probability that a customer chooses j is positive, whereas the desired model would set the probability that a customer chooses j to be zero when a dominating product i is in the assortment (100% buydown effect of i over j). For example, consider a setting with 3 alternatives, i, j, and k. Suppose that these alternatives have 2 relevant attributes, say price and quality, and suppose that i has slightly lower price and slightly better quality than j, whereas j has both higher price and better quality than k, such that (1) if the assortment contains only j and k, then each is chosen with probability $1/2$, (2) if the assortment contains only i and k, then i is chosen with probability $2/3$ and k is chosen with probability $1/3$, and (3) if the assortment contains only i and j, then i is chosen with probability 1. Suppose an MNL model is chosen to satisfy conditions (1) and (2). Then it follows from the IIA property that the model will predict that if the assortment contained only i and j, then i would be chosen with probability $2/3$ and j would be chosen with probability $1/3$, which is very different from the desired choice probabilities of i being chosen with probability 1 and j being chosen with probability 0. More general choice models such as the nested logit model and the Markov chain choice model exhibit the same structural shortcoming.

Electronic commerce offers many examples in which a seller's assortment contains pairs of products with one dominating the other, and real-life examples of 100% buydowns are given below.

- *Bimbo Store* sells infant care products. As shown in Figure 15, the same wipes (*Huggies Pure Baby Wipes* — 56 pieces per bag) were offered both in a box with 10-bags (15a) and in a box with 12-bags (15b). Both boxes were sold for 11.99 euros each at the same time.

- *Fitvia* sells nutritional products online. As shown in Figure 16, the same snack ("*Chocolate Protein Snack*" — 50 grams per bag) was offered both in a single pack (16a) or as part of a three-pack combo (16b). The single pack was sold for 5.90 euros each, and at the same time the three-pack combo was sold for 4.90 euros each.

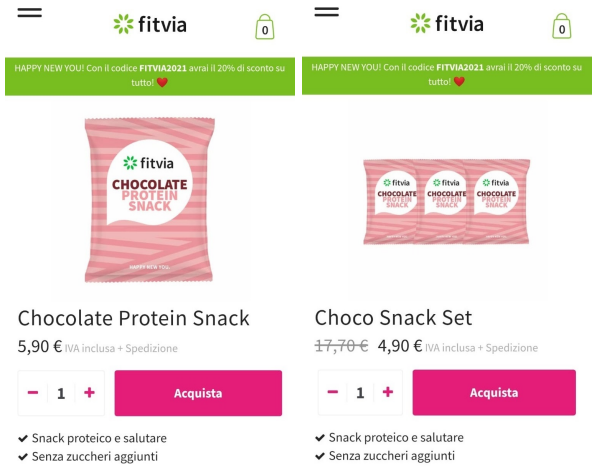
- *Lyft* sells urban transportation services, including ride-hailing services, mostly through its mobile phone app. As shown in Figure 17, at the same point in time, Lyft offered several ride-hailing alternatives for the same origin-destination pair (from 836 Juniper St., Atlanta to Westside Provisions, Atlanta). Both the *Shared Saver* service and the *Shared* service may require the rider to share the car with riders who made separate trip requests. The *Shared Saver* service may require the rider to wait longer before a car is assigned to the rider (to give Lyft extra time to find a better match between rider and car), and may require the rider to walk a small distance to a better pickup spot. With the "*Shared*" service, the pickup spot is the rider's current location. The "*Lyft*" service can be regarded as the full-service alternative; the rider does not share the car with riders who made separate trip requests, and a car is assigned to the rider after a very short time. In the example, the *Shared Saver* had a price of 6.00 dollars and the estimated arrival time was between 9:28PM and 9:30PM, the "*Shared*" service had a price of 6.19 dollars and the estimated arrival time was 9:28PM, and the "*Lyft*" service had a price of 6.00 dollars and the estimated arrival time was 9:24PM. In this example, the "*Lyft*" service dominated both of the other alternatives.

- *Marionnaud* is a retailer of beauty products that sells both in stores as well as online. Figure 18 shows a setting in which Marionnaud offered the same fragrance (*Yves Saint Laurent: Libre* — eau de parfum, 50 milliliter) both as a single bottle (18a) and as part of a gift set that also included a lipstick and an eyeliner (18b). The single bottle was sold for 102.00 euros and at the same time the gift set was sold for 73.85 euros.

With the purpose of overcoming all of these limitations, sophisticated modifications to the original formulation are proposed within the AO and RM literature (see [32, 40, 63]) and our work falls into this novel stream of research. In this chapter we propose a more flexible variant of the classical AO and RM models under MNL, able to capture dominance relations in the form of 100% buydown effect. Starting from a *Dynamic Programming* (DP) formulation and appraising that its computational burden increases exponentially in the problem dimension, we will work with a compact and tractable deterministic approximation. We strongly believe that our research would provide a preliminary exploration on a wide class of extensions to the standard MNL model, helping in depicting consumers

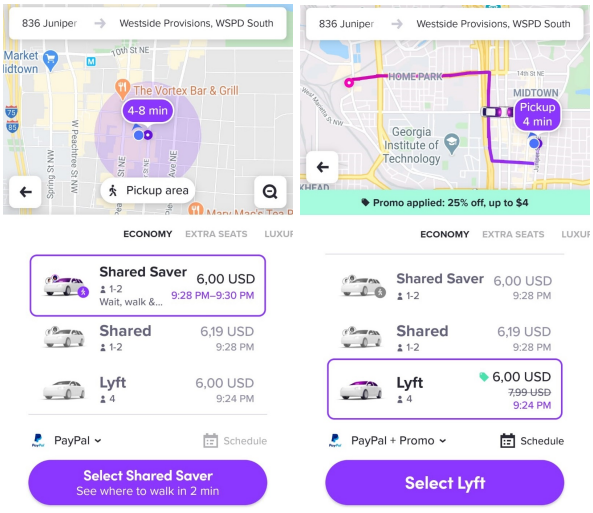
(a) Wipes 10 packs. (b) Wipes 12 packs.

Figure 15: *Huggies* example.



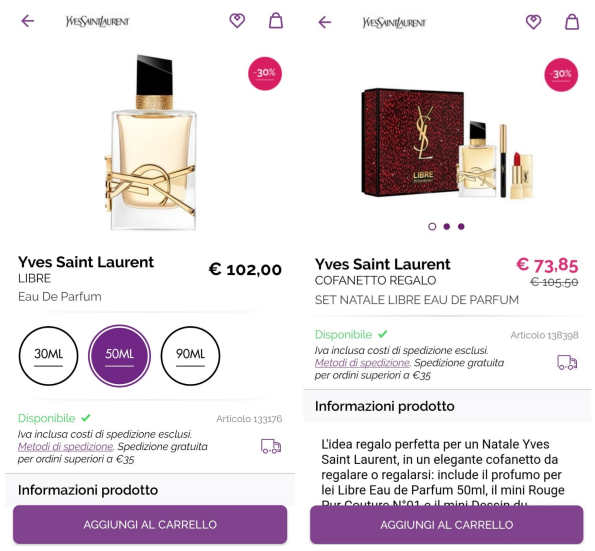
(a) Single pack. (b) 3-pack combo.

Figure 16: *Fitvia* example.



(a) *Lyft Shared Saver*. (b) *Lyft* regular ride.

Figure 17: *Lyft* example.



(a) Fragrance bottle. (b) Fragrance gift set.

Figure 18: *YSL* example.

choice behaviors that –as shown– apply to many different realities (the beauty, fitness, and service industry to name but a few). The result of this study, therefore, will be valuable both for the academic literature and the industry practitioners, who will be able to develop better tools aimed at taking more intelligent decisions.

The chapter is organized as follows. We start in Section 3.2 with a brief review of related work. Section 3.3 introduces the problems set-up, defines the dominance structures (Subsection 3.3.1) as well as the AO and RM problems (Subsections 3.3.2 and 3.3.3, respectively). Section 3.4 provides

a deterministic AO and RM tractable reformulation, along with a conversion algorithm proving its equivalence with the original model. In Section 3.5 the obtained solutions are used to construct a booking limit policy. Section 3.6 presents preliminary numerical experiments using synthetic data to compare our novel approach with the classical discrete choice model. Finally, we conclude in Section 3.7, highlighting some directions for future research.

3.2 Literature Review

The extensive connections between *Operations Research* (OR), AO and RM have been explored by a number of authors. See surveys by [82] and [75] for comprehensive overviews of this literature.

Traditional demand models assume that each customer arrives into a system with the intention of purchasing a specific product; if this product is available, then the customer makes the purchase, otherwise he/she leaves the system. This modeling assumption is known as the *independent demand model* (see [110]). In many applications, however, the potential buyer observes a set of available products (*i.e.*, the assortment) before making his/her selection. The independent demand model does not account for this customer choice behavior and, therefore, its use during seller's decision making process may lead to progressive deterioration of revenue performance (see [38]), especially when products in the offer sets are close substitutes (see [76]). For this reason, moving from independent to *choice-based demand models* has been a trend both in the academic literature and in industry practice, and several studies have validated empirically the significant leverage that is obtained when accounting for choice behaviors in highly competitive markets: see, *e.g.*, [90, 172, 186]; etc.

Loosely speaking, a choice model can be thought of as a conditional probability distribution that for any offer set yields the probability that an arriving customer purchases a given product in that set (see [55]). Therefore, the specification of a choice model –either parametric or non-parametric– is a critical task to make accurate revenues and sale predictions. While non-parametric models may be interpreted as data-driven approaches, by parametric choice models we mean that the family of underlying distributions is described by a fixed number of parameters, being independent of the training data set volume.

Several parametric choice-based demand models have been active object of investigation, including the MNL (see [63, 95, 159]); the MNL robust variants (see [131, 133]); the mixed MNL (see [30, 132]); the spiked-MNL model (see [29]); the nested logit (see [41, 56]); the Markov chain choice model (see [57]); and rank-based choice models (see [15, 55]). Among these models, the MNL model is widely used in the literature as a benchmark, because of the desirable properties it possesses, and which were mentioned in the Introduction.

Choice-based demand models are used in AO and RM with the aim of helping the decision maker (*i.e.*, seller) in choosing which assortments of products to offer to customers during a given selling horizon. The problem is formulated through DP, whose computational burden however increases exponentially due to the curse of dimensionality. A possible solution to this limitation was proposed by [188], who suggested to approximate the DP value functions with affine functions, and proposed a column generation algorithm to solve the resulting approximated DP problem. Another well-known way to deal with the intractability of the DP formulation was devised by [62], who formulated a *Choice-Based Deterministic Linear Program* (CDLP) as an approximation of the original DP, whose solutions were proved by [95] to be asymptotically optimal for the original DP problem. Nonetheless, although the CDLP problem size is smaller compared to the starting DP, its number of decision variables could still be exponential in the input number of products. This motivated the suggestion by [95] of solving the CDLP using column generation. In a different fashion, [159] proposed a new approach called *Segment-based Deterministic Concave Program* (SDCP), which is a concise relaxation of the CDLP. Most recently, [63] presented a linear program with a polynomial number of variables proved to be equivalent to the CDLP under the standard MNL model, called the *Sales-Based Linear Program* (SBLP). The authors also proved that an optimal solution of the SBLP can be converted in polynomial time to an optimal solution of the CDLP under MNL, and vice versa. Analogously, [32, 40, 44] proposed different SBLP variants, all demonstrating to be equivalent to the CDLP under the spiked-MNL model. Our research falls into this novel promising application field and aims at devising, first, a choice-based demand model incorporating strong dominance effects (which we call DMNL), which violating the IIA property cannot be represented by the standard MNL choice model. Secondly, we aim at proposing a new SBLP variant, proving to be equivalent to the CDLP under the DMNL model.

3.3 Model Formulations

Let $\mathcal{J} := \{1, \dots, n\}$ denote the seller's set of candidate products. The seller can offer any subset $A \subseteq \mathcal{J}$ in the market. The subset that the seller offers is called its assortment. Each potential customer chooses one product from the assortment or chooses to buy nothing from the assortment. If assortment A is offered, then each customer chooses product $j \in A$ with probability $P^{\text{true}}_{j:A}$, or chooses nothing from the assortment (the no-purchase alternative or outside alternative) with probability $P^{\text{true}}_{0:A}$. Each product $j \in \mathcal{J}$ contributes a profit per unit of r_j. Thus, if assortment A is offered, then the seller's expected profit per customer is equal to $\sum_{j \in A} r_j P^{\text{true}}_{j:A}$. For any two products $j, j' \in \mathcal{J}$, if product j' dominates product j, we write $j' \succ j$.

3.3.1 Dominance Structure

We make the following assumptions regarding the dominance relation:

1. The dominance relation is *irreflexive*, that is, no product j dominates itself, $j \not\succ j$.

2. The dominance relation is *antisymmetric*, that is, there are no distinct products j, j', $j \neq j'$, such that j dominates j' and j' dominates j.

3. The dominance relation is *transitive*, that is, if product j dominates product j', and product j' dominates product j'', then product j also dominates product j''.

Note that properties 1 and 2 imply that the dominance relation is *asymmetric*, that is, there are no products j, j' (distinct or not), such that $j \succ j'$ and $j' \succ j$.

The dominance relation can be represented with a directed graph as follows: for each product $j \in \mathcal{J}$, the graph has a node, also called j. For every pair of products $j, j' \in \mathcal{J}$ such that $j \succ j'$, there is an initial arc (j, j') in the graph (some of these arcs are redundant, and will be removed). Note that properties 1, 2, and 3 imply that this (initial) graph contains no cycles. Therefore, without loss of generality, we assume that the products (and nodes) are indexed and that the nodes are sorted in *topological order*. Specifically, the nodes are indexed such that for every arc (i, j) in the graph, it holds that $i > j$. Figure 19 shows an example of such an initial graph for a setting with $\mathcal{J} = \{1, \ldots, 6\}$.

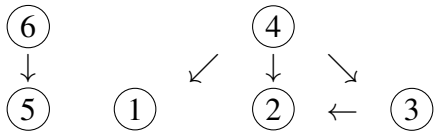


Figure 19: Example of initial dominance graph with $n = 6$ products.

It is convenient to remove unnecessary arcs. Specifically, for any two products j and j', if the initial graph contains a path from j to j' with more than one arc, then we remove the direct arc from j to j'. The resulting parsimonious graph is denoted with $\mathfrak{G}$. Figure 20 shows the parsimonious graph for the same example as in Figure 19.

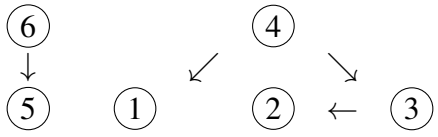


Figure 20: Example of parsimonious dominance graph $\mathfrak{G}$.

A graph $\mathfrak{G}$ can have multiple components. For example, the graphs in Figures 19 and 20 have a component containing nodes 1 to 4 and a component containing nodes 5 and 6.

For each product $j \in \mathcal{J}$, let $\mathcal{D}^-(j)$ denote the set of products that dominate product j, so,

$$\mathcal{D}^-(j) \;:=\; \{j' \in \mathcal{J} \;:\; j' \succ j\}$$

and let $\mathcal{D}^+(j)$ denote the set of products dominated by product j, that is,

$$\mathcal{D}^+(j) \;:=\; \{j' \in \mathcal{J} \;:\; j' \prec j\}.$$

Thus, for each $j \in \mathcal{J}$ and $j' \in \mathcal{D}^-(j)$, there is at least one path from j' to j in $\mathfrak{G}$, and for each $j \in \mathcal{J}$ and $j' \in \mathcal{D}^+(j)$, there is at least one path from j to j' in $\mathfrak{G}$. The notation $\mathcal{P}$ for a path in $\mathfrak{G}$ will also be used for the nodes (products) in the path, including the start node and the end node of the path. A path may be a singleton, that is, $\mathcal{P} = \{j\}$ is a valid path. For example, if $\mathcal{P}$ is a path from j to j' in $\mathfrak{G}$, where j may be equal to j', and $j'' \in \mathcal{P}$, then $j'' = j$ or $j'' \in \mathcal{D}^+(j)$. Let $\mathfrak{P}$ denote the set of maximal paths in $\mathfrak{G}$, that is,

$$\mathfrak{P} \;:=\; \left(\bigcup_{\left\{ \substack{j \in \mathcal{J}: \\ \mathcal{D}^-(j) = \varnothing} \right\}} \; \bigcup_{\left\{ \substack{j' \in \mathcal{D}^+(j): \\ \mathcal{D}^+(j') = \varnothing} \right\}} \Big\{ \text{paths } \mathcal{P} \text{ in } \mathfrak{G} \text{ from } j \text{ to } j' \Big\} \right) \bigcup_{\left\{ \substack{j \in \mathcal{J}: \\ \mathcal{D}^-(j) = \varnothing, \\ \mathcal{D}^+(j) = \varnothing} \right\}} \Big\{ \{j\} \Big\}.$$

For the graph $\mathfrak{G}$ in Figure 20, $\mathfrak{P} = \Big\{\{4,1\}, \{4,3,2\}, \{6,5\}\Big\}$. For the graph $\mathfrak{G}$ in Figure 21, $\mathfrak{P} = \Big\{\{8,1\}, \{8,6,5,3,2\}, \{8,6,4,3,2\}, \{7,5,3,2\}, \{7,4,3,2\}\Big\}$.

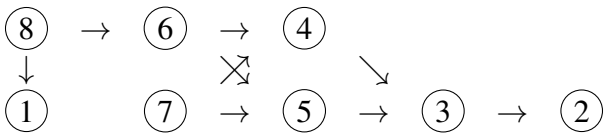


Figure 21: Example of a non-arborescence dominance graph $\mathfrak{G}$.

An important property of the dominance graph $\mathfrak{G}$ is the number $|\mathfrak{P}|$ of maximal paths. Next we give an example that shows that $|\mathfrak{P}|$ may be exponential in the number $|\mathcal{J}|$ of nodes, and in the Appendix C we establish that $|\mathfrak{P}| \leq \exp(|\mathcal{J}|/e)$. Thereafter we point out that for a class of dominance graphs that is important in applications, it holds that $|\mathfrak{P}| \leq |\mathcal{J}|$.

Example 1. Suppose that $\mathcal{J} = \{1, 2, \ldots, 2n\}$. In general, each maximal path goes from a $j \in \mathcal{J}$ such that $\mathcal{D}^-(j) = \varnothing$ to a $j' \in \mathcal{J}$ such that $\mathcal{D}^+(j') = \varnothing$. Let $\{j \in \mathcal{J} \;:\; \mathcal{D}^-(j) = \varnothing\} = \{2n-1, 2n\}$, and $\{j \in \mathcal{J} \;:\; \mathcal{D}^+(j) = \varnothing\} = \{1, 2\}$. For each $i \in \{2, \ldots, n\}$, there are arcs $(2i{-}1, 2i{-}3)$, $(2i{-}1, 2i{-}2)$, $(2i, 2i-3)$, and $(2i, 2i-2)$. Figure 22 shows the resulting dominance graph $\mathfrak{G}$ for $n = 6$. Note that for each sequence $(b_n, \ldots, b_1) \in \{0,1\}^n$, there is a maximal path $(2n - b_n, \ldots, 2i - b_i, \ldots, 2 - b_1)$ in $\mathfrak{G}$. Thus, $|\mathfrak{P}| = 2^n$.

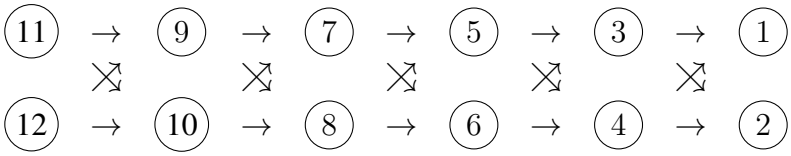


Figure 22: Dominance graph $\mathfrak{G}$ of Example 1 for $n = 6$.

Many of the dominance graphs in applications have the property that each component is either an *arborescence* or an *anti-arborescence*.

Definition 2. An arborescence is a directed and acyclic graph with a node r, called the root, such that for any other node i, the graph has exactly one directed path from r to i. An anti-arborescence is a directed and acyclic graph with a node r, called the root, such that for any other node i, the graph has exactly one directed path from i to r.

If each component is either an *arborescence* or an *anti-arborescence*, then the number of maximal paths is at most $|\mathcal{J}|$.

3.3.2 Assortment Optimization Problem

In the assortment optimization problem, the set $\mathcal{J}$ of candidate products, the unit profits r_j for all products $j \in \mathcal{J}$, and the expected number of customers are input. Hence the assortment optimization problem is

$$\max \left\{ \sum_{j \in A} r_j P_{j:A}^{\text{true}} \ : \ A \subseteq \mathcal{J} \right\}. \tag{3.1}$$

Next we consider two demand-based choice models for estimating $P_{j:A}^{\text{true}}$.

1. In the *Multinomial Logit* (MNL) model, each product $j \in \mathcal{J}$ has a preference weight $v_j > 0$, and the no-purchase alternative has preference weight $v_0 > 0$. For any assortment $A \subseteq \mathcal{J}$ and any product $j \in \mathcal{J}$, the MNL model specifies the probability $\hat{P}_{j:A}$ that a customer chooses j if A is offered, as follows:

$$\hat{P}_{j:A} \ := \ \begin{cases} \dfrac{v_j}{v_0 + \sum_{j' \in A} v_{j'}} & \text{if} \quad j \in A \\ \\ 0 & \text{if} \quad j \notin A \end{cases} \tag{3.2}$$

and

$$\hat{P}_{0:A} \ := \ \frac{v_0}{v_0 + \sum_{j' \in A} v_{j'}}. \tag{3.3}$$

The corresponding assortment optimization problem is

$$\max \left\{ \sum_{j \in A} r_j \hat{P}_{j:A} \ : \ A \subseteq \mathcal{J} \right\} \tag{3.4}$$

with $A^{\mathrm{MNL}} \in \operatorname{argmax} \left\{ \sum_{j \in A} r_j \hat{P}_{j:A} \ : \ A \subseteq \mathcal{J} \right\}$ optimal assortment under the MNL model.

2. The MNL model has many desirable properties, but a serious shortcoming of the MNL model is that $\hat{P}_{j:A} > 0$, even when $j, j' \in A$ and $j' \succ j$. Hence, we propose a discrete choice model with the property that if $j, j' \in A$ and $j' \succ j$, then the probability of choosing j from assortment A is zero. We call this choice model the *Dominance Multinomial Logit* (DMNL) model. This choice model retains the desirable properties of the MNL model: it is as easy to interpret the DMNL model as the MNL model; the DMNL model has the same number of parameters as the MNL model; the model calibration problems for the DMNL model and the MNL model are the same. In addition, the DMNL model overcomes one of the most serious structural shortcomings of the MNL model.

   For any assortment $A \subseteq \mathcal{J}$ and any product $j \in \mathcal{J}$, the DMNL model specifies the probability $P_{j:A}$ that a customer chooses j if A is offered, as follows[2]:

$$P_{j:A} \ := \ \begin{cases} \dfrac{v_j \mathbb{1}\{j'' \not\succ j \ \forall \ j'' \in A\}}{v_0 + \sum_{j' \in A} v_{j'} \mathbb{1}\{j'' \not\succ j' \ \forall \ j'' \in A\}} & \text{if} \quad j \in A \\ \\ 0 & \text{if} \quad j \notin A \end{cases} \tag{3.5}$$

   and

$$P_{0:A} \ := \ \frac{v_0}{v_0 + \sum_{j' \in A} v_{j'} \mathbb{1}\{j'' \not\succ j' \ \forall \ j'' \in A\}}. \tag{3.6}$$

   The corresponding assortment optimization problem is

$$\max \left\{ \sum_{j \in A} r_j P_{j:A} \ : \ A \subseteq \mathcal{J} \right\} \tag{3.7}$$

   with $A^{\mathrm{DMNL}} \in \operatorname{argmax} \left\{ \sum_{j \in A} r_j P_{j:A} \ : \ A \subseteq \mathcal{J} \right\}$ optimal assortment under the DMNL model. An alternative formulation of problem (3.7) that sometimes is easier to use is the following. Let

$$\mathcal{A} \ := \ \left\{ A \subseteq \mathcal{J} \ : \ j' \not\succ j \ \forall \ j, j' \in A \right\}$$

   denote the collection of assortments consisting only of products that do not dominate each other. Then the assortment optimization problem (3.7) is equivalent to

$$\max_{A \in \mathcal{A}} \sum_{j \in A} r_j \hat{P}_{j:A}. \tag{3.8}$$

[2]The indicator function $\mathbb{1}\{\text{event}\}$ takes value 1 if the event happens (true statement) and 0 otherwise.

Note that the choice model in (3.8) is the MNL choice model. Problems (3.7) and (3.8) are equivalent, because every $A \in \mathcal{A}$ has the same objective value in (3.7) and (3.8), and for every $A \subseteq \mathcal{J}$, there is an $A' = \{j \in A \ : \ j' \not\succ j \ \forall \ j' \in A\} \in \mathcal{A}$, obtained by removing all the dominated products from A, such that the objective value of A in (3.7) is equal to the objective value of A' in (3.8).

We will show in the Numerical Experiments (Section 3.6) that using an MNL model instead of a DMNL in the presence of dominance relations can lead to arbitrarily bad relative revenue performances.

3.3.3 Revenue Management Problem

We consider the revenue management problem with the same data as the assortment optimization problem. Additionally, there is a set $\mathcal{R}$ of resources used to produce the products. The seller has an initial supply of b_r units of resource r. Each unit of product j consumes $a_{rj} \in \{0, 1, \ldots, b_r\}$ units of resource r. For each $j \in \mathcal{J}$, let $a^j := (a_{rj}, r \in \mathcal{R})$. The resources can be used to sell products over a selling horizon with discrete periods indexed by $t = 0, 1, \ldots, T-1$. The seller chooses an assortment $A_t \subseteq \mathcal{J}$ at the beginning of each period t. In each period t, either one customer arrives with probability $0 \leq \lambda \leq 1$, or no customer arrives. The customer chooses from the assortment A_t according to a choice model $P_{j:A}$, as described in (3.5)-(3.6).

The revenue management problem can be formulated as a *Markov Decision Process* (MDP), as follows: At the beginning of any period t, let $y_r(t) \geq 0$ denote the amount of resource r that the seller still has available, and let $y(t) := (y_r(t), r \in \mathcal{R})$ denote the state at the beginning of period t. For any given state $y = (y_r, r \in \mathcal{R})$, let $\mathcal{J}(y) := \{j \in \mathcal{J} \ : \ a_{rj} \leq y_r \ \forall \ r \in \mathcal{R}\}$ denote the set of products that can be produced with the available resources. At the beginning of period t, the seller chooses an assortment $A_t \subseteq \mathcal{J}(y(t))$ to maximize the seller's expected revenue over the selling horizon. Let $b := (b_r, r \in \mathcal{R})$ denote the seller's initial supply of resources, that is, $y(0) = b$. Let $Y := \{y \in \mathbb{N}^{|\mathcal{R}|} \ : \ y_r \leq b_r \, \forall \, r \in \mathcal{R}\}$ denote the state space of the MDP, and let $V : Y \times \{0, 1, \ldots, T\} \mapsto \mathbb{R}$ denote the optimal value function[3], with $V(y, T) = 0$ for all $y \in Y$. Then V satisfies the following optimality equation for all $y \in Y$ and $t \in \{0, \ldots, T-1\}$:

$$V(y,t) \ = \ \max_{A \subseteq \mathcal{J}(y)} \left\{ \sum_{j \in A} \lambda P_{j:A} \Big[r_j - \Big(V(y, t+1) - V(y - a^j, t+1) \Big) \Big] \right\} + V(y, t+1). \quad (3.9)$$

Problem (3.9) is intractable for instances with a large number of resources, due to curse of dimensionality. Therefore, we recover a fluid approximation often used in the RM literature, called the *Choice-based Deterministic Linear Program* (CDLP) in which customer arrivals and choices are

[3] We denote by $\mathbb{N}^n$ the n-dimensional natural space and by $\mathbb{R}$ the set of real numbers.

replaced by their means, and resource supplies are real-valued rather than integer valued. For any assortment $A \subseteq \mathcal{J}$, let $\alpha(A)$ denote the fraction of time T that assortment A is offered. Then the CDLP is given by:

$$\begin{aligned} \max \quad & \lambda T \sum_{A \subseteq \mathcal{J}} \alpha(A) \sum_{j \in A} r_j P_{j:A} \\ \text{s.t.} \quad & \sum_{A \subseteq \mathcal{J}} \alpha(A) \leq 1 \\ & \lambda T \sum_{A \subseteq \mathcal{J}} \alpha(A) \sum_{j \in A} a_{rj} P_{j:A} \leq b_r \qquad \forall\, r \in \mathcal{R} \\ & \alpha(A) \geq 0 \qquad\qquad\qquad\qquad \forall\, A \subseteq \mathcal{J}. \end{aligned} \tag{CDLP}$$

The objective function is the expected total revenue over the time horizon. The first constraint specifies that the sum of the fractions of time that different assortments are offered is less than 1. The capacity constraints for every resource follow, and we finally enforce non-negativity on all decision variables. Note that problem (3.7) happens to be a special case of problem (CDLP). We recall indeed that an AO problem consists in looking for a single assortment A, with products not constrained in capacity. Hence, removing the resource constraint from formulation (CDLP) leads to the same solution as of problem (3.7). For this reason, from now on, we focus our attention on the CDLP problem only.

3.4 The DMNL Sales-Based Linear Program

The number of decision variables in problem (CDLP) is exponential in the number of products $|\mathcal{J}|$. This motivates the development of an equivalent linear program –much smaller in size and easier to solve– which we call the DMNL *Sales-Based Linear Program* (SBLP). For each $j \in \mathcal{J}$, let $x_j \geq 0$ denote the total amount of product j that is sold over the time horizon $[0, T]$, and let $x := (x_j, j \in \mathcal{J})$. Then the SBLP is as follows:

$$\begin{aligned} \max_{x} \quad & \sum_{j \in \mathcal{J}} r_j x_j \\ \text{s.t.} \quad & x_0 + \sum_{j \in \mathcal{J}} x_j \leq \lambda T \\ & \sum_{j \in \mathcal{J}} a_{rj} x_j \leq b_r \qquad \forall\, r \in \mathcal{R} \\ & \sum_{j \in \mathcal{P}} \frac{x_j}{v_j} \leq \frac{x_0}{v_0} \qquad \forall\, \mathcal{P} \in \mathfrak{P} \\ & x_0 \geq 0, \quad x_j \geq 0 \quad \forall\, j \in \mathcal{J}. \end{aligned} \tag{SBLP}$$

The objective function aims at maximizing total expected revenues from the sale of all products. The first constraint is the so called balance constraint and represents the fact the number of no-purchase

customers plus the overall number of sales does not exceed the number of arrivals (*i.e.*, λT denotes the expected number of arriving customers). The capacity constraints for every resource follow. The third set of constraints, named dominating alternatives constraints, requires that no products dominating each other are jointly offered. We finally enforce non-negativity on all decision variables.

The overall number of variables in problem (SBLP) is $|\mathcal{J}| + 1$. The objective function and the constraints are linear functions of the decision variables. The number of dominating alternative constraints is at most $\exp(|\mathcal{J}|/e)$. However, if each graph component is either an *arborescence* or an *anti-arborescence*, then the number of such constraints is at most $|\mathcal{J}|$.

3.4.1 Conversion of SBLP Feasible Solutions into CDLP Feasible Solutions

We now provide an Algorithm that converts a feasible solution x of problem (SBLP) into a feasible solution $\alpha(A)$, $A \subseteq \mathcal{J}$ of problem (CDLP). A practical example illustrates the algorithm in the Appendix C.

Algorithm 1: Converting a SBLP solution into a CDLP solution.

Data: SBLP solution x

Set $k \leftarrow 0$, $\alpha(A) = 0$ for all $A \subseteq \mathcal{J}$
while $\{j \in \mathcal{J} : x_j > 0\} \neq \varnothing$ **do**
  Set $k \leftarrow k + 1$
  Set $A_k \leftarrow \{j \in \mathcal{J} : x_j > 0\}$
  Set $D_k \leftarrow \{j \in A_k : \mathcal{D}^-(j) \cap A_k = \varnothing\}$
  Set $Y_k \leftarrow \min\left\{\frac{x_j}{v_j} : j \in D_k\right\}$
  Set $j_k \in \operatorname{argmin}\left\{\frac{x_j}{v_j} : j \in D_k\right\}$
  Set $\alpha(A_k) \leftarrow \frac{v_0 + \sum_{j \in D_k} v_j}{\lambda T} Y_k$
  for all $j \in D_k$ **do**
    $x_j \leftarrow x_j - \lambda T \alpha(A_k) \frac{v_j}{v_0 + \sum_{j' \in D_k} v_{j'}}$
  end
end

Result: Output $\alpha(A)$ for all $A \subseteq \mathcal{J}$

Before we establish that given an optimal solution of the SBLP the output of Algorithm 1 is an optimal solution of the CDLP, we note some basic properties of Algorithm 1 in Lemma 1.

Lemma 1. The output of Algorithm 1 satisfies the following properties:

1. For each iteration k, the following holds:

   a) The sales quantity of product $j \in A_k \setminus D_k$ while assortment A_k is offered for $T\alpha(A_k)$ units of time is 0.

b) The sales quantity of product $j \in D_k$ while assortment A_k is offered for $T\alpha(A_k)$ units of time is $\lambda T\alpha(A_k)\frac{v_j}{v_0+\sum_{j'\in D_k} v_{j'}}$, which is equal to the amount subtracted from x_j in Step 10.

2. The total sales quantity of each product $j \in \mathcal{J}$ resulting from the CDLP solution α produced by Algorithm 1 is equal to

$$\begin{aligned} x_j &= \lambda T \sum_{\{k\,:\,j\in D_k\}} \alpha(A_k)\frac{v_j}{v_0+\sum_{j'\in D_k} v_{j'}} \\ &= \lambda T \sum_{\{k\,:\,j\in A_k\}} \alpha(A_k) P_{j:A_k} \\ &= \lambda T \sum_{A\subseteq\mathcal{J}} \alpha(A) P_{j:A}. \end{aligned} \tag{3.10}$$

3. Consider any path $\mathcal{P}^* \in \operatorname{argmax}\left\{\sum_{j\in\mathcal{P}} \frac{x_j}{v_j} \;:\; \mathcal{P}\in\mathfrak{P}\right\}$. Then $|\mathcal{P}^*\cap D_k| = 1$ for each k, that is, for each offered assortment A_k, one product on path $\mathcal{P}^*$ is being sold while assortment A_k is offered.

Proof. Properties of Lemma 1 are proven as follows.

1. The sales quantity of $j \in A_k$ while A_k is offered for $T\alpha(A_k)$ units of time is

$$\begin{aligned} \lambda T\alpha(A_k)P_{j:A_k} &= \lambda T\alpha(A_k)\frac{v_j \mathbb{1}\{j''\not\succ j\ \forall\, j''\in A_k\}}{v_0+\sum_{j'\in A_k} v_{j'}\mathbb{1}\{j''\not\succ j'\ \forall\, j''\in A_k\}} \\ &= \begin{cases} \lambda T\alpha(A_k)\frac{v_j}{v_0+\sum_{j'\in D_k} v_{j'}} & \text{if} \quad j\in D_k \\ 0 & \text{if} \quad j\in A_k\setminus D_k. \end{cases} \end{aligned}$$

2. We show by induction that at the end of each iteration k it holds that $x_j \geq 0$ for all $j \in \mathcal{J}$, and that in Step 10 of each iteration k, at least one component x_j is reduced to zero. Note that at the beginning of iteration 1 it holds that $x_j \geq 0$ for all $j \in \mathcal{J}$. Assume that at the beginning of iteration k it holds that $x_j \geq 0$ for all $j \in \mathcal{J}$. Note that for any $j \in \mathcal{J}\setminus D_k$, x_j does not change during iteration k, and thus it holds that $x_j \geq 0$ at the end of iteration k. Next note that for each $j \in D_k$ the calculation in Step 10 results in

$$\begin{aligned} x_j - \lambda T\alpha(A_k)\frac{v_j}{v_0+\sum_{j'\in D_k} v_{j'}} &= x_j - \lambda T\frac{v_0+\sum_{j\in D_k} v_j}{\lambda T} Y_k \frac{v_j}{v_0+\sum_{j'\in D_k} v_{j'}} \\ &= x_j - Y_k v_j \;\geq\; x_j - \frac{x_j}{v_j} \;=\; 0. \end{aligned}$$

where the first equality follows from Step 8, and the inequality follows from Step 6. Therefore it holds that $x_j \geq 0$ at the end of iteration k for all $j \in \mathcal{J}$. Specifically, note that for each $j_k \in \operatorname{argmin}\left\{\frac{x_j}{v_j} : j \in D_k\right\}$ the calculation in Step 10 reduces to

$$x_{j_k} - \lambda T \alpha(A_k) \frac{v_{j_k}}{v_0 + \sum_{j' \in D_k} v_{j'}} = x_{j_k} - Y_k v_{j_k} = x_{j_k} - \frac{x_{j_k}}{v_{j_k}} v_{j_k} = 0$$

where the first equality follows from Step 8, and the third equality follows from Steps 6 and 7. Since Algorithm 1 stops when all x_j have been reduced to 0, it follows that the total sales quantity of each product $j \in \mathcal{J}$ resulting from the CDLP solution α produced by Algorithm 1 is given by (3.10).

3. We show by induction on k that at the beginning of each iteration k it holds that $\sum_{j \in \mathcal{P}^*} \frac{x_j}{v_j} \geq \sum_{j \in \mathcal{P}} \frac{x_j}{v_j}$ for all paths $\mathcal{P}$, and that $\mathcal{P}^* \cap D_k \neq \varnothing$ for each k. It follows from the definition of $\mathcal{P}^*$ that $\sum_{j \in \mathcal{P}^*} \frac{x_j}{v_j} \geq \sum_{j \in \mathcal{P}} \frac{x_j}{v_j}$ for all paths $\mathcal{P}$ at the beginning of iteration 1. Assume that $\sum_{j \in \mathcal{P}^*} \frac{x_j}{v_j} \geq \sum_{j \in \mathcal{P}} \frac{x_j}{v_j}$ for all paths $\mathcal{P}$ at the beginning of iteration k. Note that there is an iteration k only if $\{j \in \mathcal{P}^* : x_j > 0\} \neq \varnothing$. Let $j^* := \max\{j \in \mathcal{P}^* : x_j > 0\}$. We show by contradiction that $j^* \in D_k$. Suppose that $j^* \notin D_k$, that is, $\mathcal{D}^-(j^*) \cap A_k \neq \varnothing$. Thus, there is a $j' \in \mathcal{D}^-(j^*) \cap A_k$, that is, $j' \succ j^*$, $x_{j'} > 0$, and $j' \notin \mathcal{P}^*$. Let $\mathcal{P}'$ be the path from j' to j^* and from there $\mathcal{P}'$ coincides with $\mathcal{P}^*$. Then $\sum_{j \in \mathcal{P}^*} \frac{x_j}{v_j} < \sum_{j \in \mathcal{P}^*} \frac{x_j}{v_j} + \frac{x_{j'}}{v_{j'}} \leq \sum_{j \in \mathcal{P}'} \frac{x_j}{v_j}$, which contradicts the induction hypothesis. Thus $\mathcal{P}^* \cap D_k = \{j^*\}$.

   Next we show that $\sum_{j \in \mathcal{P}^*} \frac{x_j}{v_j} \geq \sum_{j \in \mathcal{P}} \frac{x_j}{v_j}$ for all paths $\mathcal{P}$ at the end of iteration k. Note that for all $j \in D_k$, Algorithm 1 reduces $\frac{x_j}{v_j}$ by the same amount $\lambda T \alpha(A_k) \frac{1}{v_0 + \sum_{j' \in D_k} v_{j'}} = Y_k = \frac{x_{j_k}}{v_{j_k}}$ in Step 10 of iteration k. Also, note that for each path $\mathcal{P}$ such that $\mathcal{P} \cap D_k \neq \varnothing$, there is exactly one product in $\mathcal{P} \cap D_k$, and thus Algorithm 1 reduces both $\sum_{j \in \mathcal{P}^*} \frac{x_j}{v_j}$ and $\sum_{j \in \mathcal{P}} \frac{x_j}{v_j}$ by the same amount $\frac{x_{j_k}}{v_{j_k}}$ in Step 10 of iteration k. Thus, at the end of iteration k it holds that $\sum_{j \in \mathcal{P}^*} \frac{x_j}{v_j} \geq \sum_{j \in \mathcal{P}} \frac{x_j}{v_j}$ for all paths $\mathcal{P}$ such that $\mathcal{P} \cap D_k \neq \varnothing$. Next consider any path $\mathcal{P}$ such that $\mathcal{P} \cap D_k = \varnothing$. If $\{j \in \mathcal{P} : x_j > 0\} = \varnothing$, then $\sum_{j \in \mathcal{P}^*} \frac{x_j}{v_j} \geq \sum_{j \in \mathcal{P}} \frac{x_j}{v_j} = 0$ at the end of iteration k. Otherwise, let $j'' := \max\{j \in \mathcal{P} : x_j > 0\}$. Since $j'' \notin D_k$, there is a $j' \in \mathcal{D}^-(j'') \cap A_k$, that is, $j' \succ j''$, $x_{j'} > 0$, and $j' \notin \mathcal{P}$. Let $\mathcal{P}'$ be the path from j' to j'' and from there $\mathcal{P}'$ coincides with $\mathcal{P}$. Note that if $j' \notin D_k$, then the argument can be repeated, and therefore we can assume without loss of generality that $j' \in D_k$. At the beginning of iteration k it holds that $\sum_{j \in \mathcal{P}'} \frac{x_j}{v_j} \geq \sum_{j \in \mathcal{P}} \frac{x_j}{v_j} + \frac{x_{j'}}{v_{j'}}$, and Algorithm 1 reduces $\sum_{j \in \mathcal{P}'} \frac{x_j}{v_j}$ by $\frac{x_{j_k}}{v_{j_k}} \leq \frac{x_{j'}}{v_{j'}}$ in Step 10 of iteration k, and $\sum_{j \in \mathcal{P}} \frac{x_j}{v_j}$ remains unchanged in iteration k. Thus at the end of iteration k it holds that $\sum_{j \in \mathcal{P}'} \frac{x_j}{v_j} \geq \sum_{j \in \mathcal{P}} \frac{x_j}{v_j}$. Since $\mathcal{P}' \cap D_k \neq \varnothing$, at the end of iteration k it holds that $\sum_{j \in \mathcal{P}^*} \frac{x_j}{v_j} \geq \sum_{j \in \mathcal{P}'} \frac{x_j}{v_j}$, and thus $\sum_{j \in \mathcal{P}^*} \frac{x_j}{v_j} \geq \sum_{j \in \mathcal{P}} \frac{x_j}{v_j}$.

□

Theorem 2. Given any feasible solution for the SBLP, Algorithm 1 computes a feasible solution for the CDLP, such that the two solutions represent the same sales amount for each product and have the same objective value.

Proof. First we show that given a feasible solution x for the SBLP, Algorithm 1 computes a feasible solution α for the CDLP. It follows from the SBLP capacity constraint and from (3.10) that

$$\sum_{j\in\mathcal{J}} a_{rj}x_j \;\leq\; b_r \qquad \Rightarrow \quad \lambda T \sum_{A\subseteq\mathcal{J}} \alpha(A) \sum_{j\in A} a_{rj} P_{j:A} \;\leq\; b_r$$

for all $r \in \mathcal{R}$, and thus the solution α from Algorithm 1 satisfies the capacity constraint of the CDLP. Recall the SBLP demand constraint

$$x_0 + \sum_{j\in\mathcal{J}} x_j \;\leq\; \lambda T. \tag{3.11}$$

It follows from the SBLP scale constraint

$$\sum_{j\in\mathcal{P}} \frac{x_j}{v_j} \;\leq\; \frac{x_0}{v_0} \qquad \forall\, \mathcal{P} \in \mathfrak{P} \tag{3.12}$$

and from (3.10) that the left side of (3.11) satisfies

$$\begin{aligned}
x_0 + \sum_{j\in\mathcal{J}} x_j \;&\geq\; v_0 \sum_{j\in\mathcal{P}^*} \frac{x_j}{v_j} + \sum_{j\in\mathcal{J}} x_j \\
&=\; \sum_{j\in\mathcal{P}^*} \frac{v_0}{v_j} \lambda T \sum_{\{k\,:\,j\in D_k\}} \alpha(A_k) \frac{v_j}{v_0 + \sum_{j'\in D_k} v_{j'}} + \sum_{j\in\mathcal{J}} \lambda T \sum_{A\subseteq\mathcal{J}} \alpha(A) P_{j:A} \\
&=\; \lambda T \left[\sum_{j\in\mathcal{P}^*} \sum_{\{k\,:\,j\in D_k\}} \alpha(A_k) \frac{v_0}{v_0 + \sum_{j'\in D_k} v_{j'}} + \sum_{A\subseteq\mathcal{J}} \alpha(A) \sum_{j\in A} P_{j:A} \right] \\
&=\; \lambda T \left[\sum_{k} \sum_{j\in\mathcal{P}^*\cap D_k} \alpha(A_k) P_{0:A_k} + \sum_{A\subseteq\mathcal{J}} \alpha(A) \sum_{j\in A} P_{j:A} \right] \\
&=\; \lambda T \left[\sum_{k} \alpha(A_k) P_{0:A_k} + \sum_{A\subseteq\mathcal{J}} \alpha(A) \sum_{j\in A} P_{j:A} \right] \\
&=\; \lambda T \left[\sum_{A\subseteq\mathcal{J}} \alpha(A) P_{0:A} + \sum_{A\subseteq\mathcal{J}} \alpha(A) \sum_{j\in A} P_{j:A} \right] \\
&=\; \lambda T \sum_{A\subseteq\mathcal{J}} \alpha(A) \left[P_{0:A} + \sum_{j\in A} P_{j:A} \right] \;=\; \lambda T \sum_{A\subseteq\mathcal{J}} \alpha(A)
\end{aligned} \tag{3.13}$$

where the fourth equality follows from the result in Lemma 1 that $|\mathcal{P}^* \cap D_k| = 1$ for all k. It follows from (3.11) and (3.13) that

$$\lambda T \sum_{A\subseteq\mathcal{J}} \alpha(A) \;\leq\; x_0 + \sum_{j\in\mathcal{J}} x_j \;\leq\; \lambda T \qquad \Rightarrow \quad \sum_{A\subseteq\mathcal{J}} \alpha(A) \;\leq\; 1$$

and thus the solution α from Algorithm 1 satisfies the time constraint of the CDLP. It follows from (3.10) that

$$\sum_{j\in\mathcal{J}} r_j x_j \;=\; \sum_{j\in\mathcal{J}} r_j \lambda T \sum_{A\subseteq\mathcal{J}} \alpha(A) P_{j:A} \;=\; \lambda T \sum_{A\subseteq\mathcal{J}} \alpha(A) \sum_{j\in A} r_j P_{j:A}$$

and thus the objective values of the SBLP solution x and the CDLP solution α output by Algorithm 1 are the same. □

3.4.2 Conversion of CDLP Feasible Solutions into SBLP Feasible Solutions

Given any feasible solution α for the CDLP, let

$$x_j \;=\; \lambda T \sum_{\{A\subseteq\mathcal{J}\,:\,j\in A\}} \alpha(A) P_{j:A} \tag{3.14}$$

for all $j \in \mathcal{J}$, and let

$$x_0 \;=\; \lambda T \left(\sum_{A\subseteq\mathcal{J}} \alpha(A) P_{0:A} + \left[1 - \sum_{A\subseteq\mathcal{J}} \alpha(A) \right] \right). \tag{3.15}$$

Theorem 3. Given any feasible solution α for the CDLP, x computed in (3.14)–(3.15) is a feasible solution for the SBLP, and the two solutions represent the same amount of sales for each product and have the same objective value.

Proof.

$$x_0 + \sum_{j\in\mathcal{J}} x_j \;=\; \lambda T \left(\sum_{A\subseteq\mathcal{J}} \alpha(A) P_{0:A} + \left[1 - \sum_{A\subseteq\mathcal{J}} \alpha(A) \right] \right) + \sum_{j\in\mathcal{J}} \lambda T \sum_{\{A\subseteq\mathcal{J}\,:\,j\in A\}} \alpha(A) P_{j:A}$$

$$= \lambda T \left(\left[1 - \sum_{A\subseteq\mathcal{J}} \alpha(A) \right] + \sum_{A\subseteq\mathcal{J}} \alpha(A) \left(P_{0:A} + \sum_{j\in A} P_{j:A} \right) \right) \;=\; \lambda T.$$

and thus x satisfies the SBLP demand constraint. Next, note that $\forall\, r \in \mathcal{R}$ it holds that

$$\sum_{j\in\mathcal{J}} a_{rj} x_j \;=\; \sum_{j\in\mathcal{J}} a_{rj} \lambda T \sum_{\{A\subseteq\mathcal{J}\,:\,j\in A\}} \alpha(A) P_{j:A} \;=\; \lambda T \sum_{A\subseteq\mathcal{J}} \alpha(A) \sum_{j\in A} a_{rj} P_{j:A} \;\le\; b_r$$

where the inequality follows from the CDLP capacity constraint, and thus x satisfies the SBLP capacity constraint. Also, for any $A \subseteq \mathcal{J}$ and $\mathcal{P} \in \mathfrak{P}$, if $\mathcal{P} \cap A \neq \varnothing$, let $j(\mathcal{P}, A) := \max\{j \in \mathcal{P} \cap A\}$. Then,

for any $\mathcal{P} \in \mathfrak{P}$ it holds that

$$\begin{aligned}
\sum_{j\in\mathcal{P}} \frac{x_j}{v_j} &= \sum_{j\in\mathcal{P}} \frac{1}{v_j} \lambda T \sum_{\{A\subseteq\mathcal{J} \,:\, j\in A\}} \alpha(A) P_{j:A} \\
&= \lambda T \sum_{j\in\mathcal{P}} \frac{1}{v_j} \sum_{\{A\subseteq\mathcal{J} \,:\, j\in A\}} \alpha(A) \frac{v_j \mathbb{1}\{j'' \not\succ j \;\forall\, j'' \in A\}}{v_0 + \sum_{j'\in A} v_{j'} \mathbb{1}\{j'' \not\succ j' \;\forall\, j'' \in A\}} \\
&= \lambda T \sum_{\{A\subseteq\mathcal{J} \,:\, \mathcal{P}\cap A\neq\varnothing\}} \frac{1}{v_{j(\mathcal{P},A)}} \alpha(A) \frac{v_{j(\mathcal{P},A)} \mathbb{1}\{j'' \not\succ j(\mathcal{P},A) \;\forall\, j'' \in A\}}{v_0 + \sum_{j'\in A} v_{j'} \mathbb{1}\{j'' \not\succ j' \;\forall\, j'' \in A\}} \\
&= \lambda T \sum_{\{A\subseteq\mathcal{J} \,:\, \mathcal{P}\cap A\neq\varnothing\}} \frac{1}{v_0} \alpha(A) \frac{v_0 \mathbb{1}\{j'' \not\succ j(\mathcal{P},A) \;\forall\, j'' \in A\}}{v_0 + \sum_{j'\in A} v_{j'} \mathbb{1}\{j'' \not\succ j' \;\forall\, j'' \in A\}} \\
&\leq \lambda T \sum_{\{A\subseteq\mathcal{J} \,:\, \mathcal{P}\cap A\neq\varnothing\}} \frac{1}{v_0} \alpha(A) P_{0:A} \;\leq\; \frac{1}{v_0} \lambda T \sum_{A\subseteq\mathcal{J}} \alpha(A) P_{0:A} \;\leq\; \frac{x_0}{v_0}
\end{aligned}$$

where the last inequality follows from the CDLP time constraint $\sum_{A\subseteq\mathcal{J}} \alpha(A) \leq 1$. Thus x satisfies the SBLP scale constraint. Last, note that

$$\sum_{j\in\mathcal{J}} r_j x_j = \sum_{j\in\mathcal{J}} r_j \lambda T \sum_{\{A\subseteq\mathcal{J} \,:\, j\in A\}} \alpha(A) P_{j:A} = \lambda T \sum_{A\subseteq\mathcal{J}} \alpha(A) \sum_{j\in A} r_j P_{j:A}$$

and thus the objective values of the SBLP solution x and the CDLP solution α are the same. □

3.5 Revenue Management Policies

In this section we construct policies based on a SBLP solution x^* and/or a CDLP solution α^*.

3.5.1 Time-Based Policy

A time-based policy can be obtained from a CDLP solution α as follows. Consider any time $t \in [0,T]$, and as before let $y_r(t) \geq 0$ denote the amount of resource r that the seller has available at time t, let $y(t) := (y_r(t), r \in \mathcal{R})$, and let $\mathcal{J}(y(t)) := \{j \in \mathcal{J} : a_{rj} \leq y_r(t) \;\forall\, r \in \mathcal{R}\}$ denote the set of products that can be produced with the available resources. Let $k(t)$ be such that

$$T \sum_{k=1}^{k(t)-1} \alpha^*(A_k) \;\leq\; t \;<\; T \sum_{k=1}^{k(t)} \alpha^*(A_k).$$

Then the time-based policy offers assortment $A^{\text{TB}}(t) := \mathcal{J}(y(t)) \cap A_{k(t)}$ at time t. Note that Algorithm 1 produces a nested sequence of assortments $A_1 \supset A_2 \supset \cdots \supset A_K$, and thus $A^{\text{TB}}(t_1) \supset A^{\text{TB}}(t_2)$ for all $t_1 < t_2$, that is, the time-based policy offers a nested sequence of assortments for every sample path.

3.5.2 Booking Limit Policy

Consider any time $t \in [0,T]$, and let $x_j(t) \geq 0$ denote the amount of product j that the seller has sold during time $(0,t)$, and let $x(t) := (x_j(t), j \in \mathcal{J})$. The booking limit policy offers assortment $A^{\text{BL}}(t) := \left\{j \in \mathcal{J} \,:\, x_j(t) < x_j^*\right\}$ at time t. Since $x_j(t)$ is nondecreasing in t, it follows that for each product j there is a random time $\tau_j \in [0,T]$ such that product j is offered during time interval $(0,\tau_j]$, and thus $A^{\text{BL}}(t_1) \supset A^{\text{BL}}(t_2)$ for all $t_1 < t_2$, that is, the booking limit policy offers a nested sequence of assortments for every sample path. Also note that since $x_j(t) \leq x_j^*$ for all t, it follows that $\sum_{j\in\mathcal{J}} a_{rj} x_j(t) \leq b_r$ for all t, that is, the booking limit policy does not exceed the capacity constraint.

3.6 Preliminary Numerical Results

In this section, we first demonstrate the hazard of using an MNL model when the true underlying model is actually a DMNL. Then, we conduct additional experiments with synthetic data. The computations have been performed on a 64-bit machine with 8 GB of RAM, a 1.8 GHz Intel i7 processor, and numerical results are obtained under MATLAB environment using MOSEK solver (version 8.1.0.72).

3.6.1 Relative Revenue Performance of the MNL and DMNL Models

We first consider a simple two-product example and show that a non-negligible gap in the relative revenue performance can be observed solving an AO problem using an MNL model instead of DMNL model when the universe of products $\mathcal{J}$ contains pairs of dominating/dominated alternatives. Hence, we show that the relative revenue performance

$$\rho := \frac{\sum\limits_{j\in A^{\text{DMNL}}} r_j P^{\text{true}}_{j:A^{\text{DMNL}}}}{\sum\limits_{j\in A^{\text{MNL}}} r_j P^{\text{true}}_{j:A^{\text{MNL}}}}$$

can be arbitrarily large. We consider the following instance, very similar to the worst-case instance considered in [32]. Consider any $\varepsilon \in (0,1)$. A seller can offer two products L and H, that is, $\mathcal{J} = \{L,H\}$. The products L and H are sold at prices r_L and r_H respectively, such that $0 < r_L/r_H < \varepsilon$. Suppose that $L \succ H$, that is, if both L and H are offered, then customers buy only L. The preference parameters are $v_0 = 1$ and $0 < v_L = v_H < r_L/(r_H - r_L)$.

- Using the DMNL model, we have $\mathcal{A} = \left\{\varnothing, \{L\}, \{H\}\right\}$. Since $v_L = v_H$ and $r_L < r_H$ it holds that

$$r_L P_{L:\{L\}} \;=\; r_L \frac{v_L}{v_0 + v_L} \;<\; r_H P_{H:\{H\}} \;=\; r_H \frac{v_H}{v_0 + v_H}$$

  and thus assortment $A^{\text{DMNL}} = \{H\}$ is always optimal.

Setting 1. First, consider a setting in which the MNL model is used with the correct parameter values. For what concerns the MNL model, the feasible assortments are $\varnothing$, $\{L\}$, $\{H\}$, and $\{L, H\}$. Since $v_H = v_L$ and $r_L < r_H$ it holds that

$$r_L \hat{P}_{L:\{L\}} \;=\; r_L \, \frac{v_L}{v_0 + v_L} \;<\; r_H \hat{P}_{H:\{H\}} \;=\; r_H \, \frac{v_H}{v_0 + v_H}$$

and thus, based on the MNL model, assortment $\{H\}$ is preferred over $\{L\}$. Moreover, since $v_H < r_L/(r_H - r_L)$ it holds that

$$r_H \hat{P}_{H:\{H\}} \;=\; r_H \, \frac{v_H}{v_0 + v_H} \;<\; r_L \hat{P}_{L:\{L,H\}} + r_H \hat{P}_{H:\{L,H\}} \;=\; \frac{r_L v_L + r_H v_H}{v_0 + v_L + v_H}$$

and thus, based on the MNL model, assortment $A^{\mathrm{MNL}} = \{L, H\}$ is preferred over $\{H\}$.

Let us assume that $P^{\mathrm{true}}_{H:\{H\}} = P^{\mathrm{true}}_{L:\{L\}} = P^{\mathrm{true}}_{L:\{L,H\}}$ and that $P^{\mathrm{true}}_{H:\{L,H\}} = 0$ because of the dominance relation between H and L. Then it holds that:

$$\rho = \frac{\sum\limits_{j \in A^{\mathrm{DMNL}}} r_j P^{\mathrm{true}}_{j:A^{\mathrm{DMNL}}}}{\sum\limits_{j \in A^{\mathrm{MNL}}} r_j P^{\mathrm{true}}_{j:A^{\mathrm{MNL}}}} \;=\; \frac{r_H P^{\mathrm{true}}_{H:\{H\}}}{r_L P^{\mathrm{true}}_{L:\{L,H\}} + r_H P^{\mathrm{true}}_{H:\{L,H\}}} \;=\; \frac{r_H}{r_L} \;>\; \frac{1}{\varepsilon} \quad \Longleftrightarrow \quad \rho > \frac{1}{\varepsilon}.$$

Therefore, in this instance $\rho \to \infty$ for $\varepsilon \to 0$.

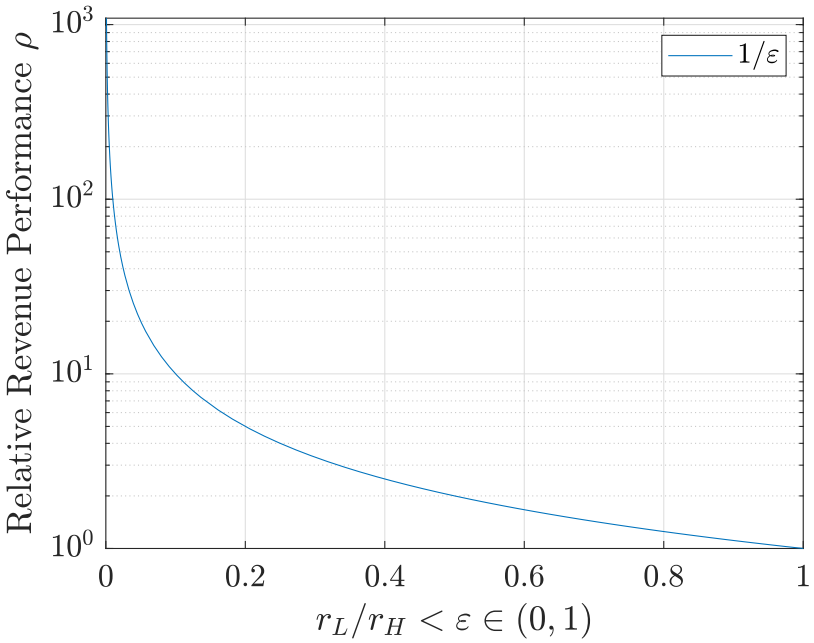


Figure 23: Relative revenue performance ρ versus ε. Vertical axis in log-scale.

Setting 2. Next, consider a setting in which the MNL model is used with parameter values calibrated with data and maximum likelihood estimation. Customers arrive one at a time, indexed with $k = 1, 2, \ldots$, and each customer arrival is observed, whether the customer chooses to buy or not (that is, no-purchase customers are observed). Customer choices are independent according to the DMNL

model. After each customer arrival, the data are updated with the customer's choice, parameter estimates are updated, and an updated assortment is chosen. Specifically, to each customer k, the seller offers one assortment $A^{(k)} \subset \{L, H\}$. The empty assortment is clearly inferior, and as shown in [159][Proposition 6], an optimal assortment under the MNL model is nested-by-revenue, and therefore it suffices to consider $\{H\}$ or $\{L, H\}$ for each $A^{(k)}$. Let $N_H^{(k)} = 1$ if customer k chooses product H, and let $N_0^{(k)} = 1$ if customer k chooses not to purchase. After each customer arrival, the seller calibrates the MNL model (3.2) using maximum likelihood estimation (MLE) with the historical data. Let $\hat{v}_H^{(k)}$, $\hat{v}_L^{(k)}$, and $v_0 = 1$ denote the estimated parameters after customer k. Regardless of the assortments offered, the MLE estimated parameter of product H is given by $\hat{v}_H^{(k)} = \sum_{k'=1}^{k} N_H^{(k')} / \sum_{k'=1}^{k} N_0^{(k')}$ (the denominator will be zero for the first few customers, but never for infinitely many customers, so set the denominator to some positive constant if $\sum_{k'=1}^{k} N_0^{(k')} = 0$). According to the estimated MNL model, it is optimal to next offer assortment $A^{(k+1)} = \{H\}$ if $\tilde{v}_H^{(k)} \geq r_L/(r_H - r_L)$; otherwise, it is optimal to offer assortment $A^{(k+1)} = \{L, H\}$. Thus, if for any k it holds that $\hat{v}_H^{(k)} < r_L/(r_H - r_L)$, then $A^{(k+1)} = \{L, H\}$, hence $N_H^{(k+1)} = 0$, and thus $\hat{v}_H^{(k+1)} = \sum_{k'=1}^{k+1} N_H^{(k')} / \sum_{k'=1}^{k+1} N_0^{(k')} \leq \sum_{k'=1}^{k} N_H^{(k')} / \sum_{k'=1}^{k} N_0^{(k')} = \hat{v}_H^{(k)} < r_L/(r_H - r_L)$, and hence $A^{(k')} = \{L, H\}$ for all $k' > k$. As shown in Setting 1, if $A^{(k)} = \{L, H\}$ then the relative revenue performance is greater than $1/\varepsilon$.

Next, we show that if the seller uses the MNL model, then $A^{(k)} = \{L, H\}$ for all but finitely many customers, with probability one. We consider two cases.

Case 1: The seller offers assortment $A^{(1)} = \{L, H\}$ to the first customer. Due to dominance, it follows that $N_H^{(1)} = 0$. Then the estimated parameter $\hat{v}_H^{(1)} = N_H^{(1)}/N_0^{(1)} = 0 < r_L/(r_H - r_L)$. It follows that the seller will offer assortment $A^{(k)} = \{L, H\}$ and $\hat{v}_H^{(k)} = 0$ for all k.

Case 2: The seller offers assortment $A^{(1)} = \{H\}$ to the first customer. As pointed out above, either there is a K such that $A^{(k)} = \{L, H\}$ for all $k > K$, or $\hat{v}_H^{(k)} \geq r_L/(r_H - r_L)$ and $A^{(k)} = \{H\}$ for all k. Next we show that, w.p.1, there is a K such that $A^{(k)} = \{L, H\}$ for all $k > K$. Suppose that $A^{(k)} = \{H\}$ for all k. Then, by the Strong Law of Large Numbers, w.p.1, $\sum_{k'=1}^{k} N_H^{(k')}/k \to v_H/(v_0 + v_H)$ and $\sum_{k'=1}^{k} N_0^{(k')}/k \to v_0/(v_0 + v_H)$ as $k \to \infty$. Thus, if $A^{(k)} = \{H\}$ for all k, then, w.p.1,

$$\hat{v}_H^{(k)} = \frac{\sum_{k'=1}^{k} N_H^{(k')}}{\sum_{k'=1}^{k} N_0^{(k')}} \to \frac{v_H}{v_0} < \frac{r_L}{r_H - r_L}.$$

Therefore, the event that $\hat{v}_H^{(k)} \geq r_L/(r_H - r_L)$ and $A^{(k)} = \{H\}$ for all k has probability 0.

3.6.2 Experiments with Synthetic Data

In this section, we use synthetic data to compare the performance of DMNL and MNL choice models applied to RM problems with dominated alternatives. We assume that whenever an assortment $A \subseteq \mathcal{J}$ is offered, then the true probability of purchasing a product $j \in A$ (that is, $P^{\text{true}}_{j:A}$) is uniformly distributed among the non-dominated alternatives belonging to A (including the null alternative). This hypothesis reflects the 100% buydown effect, according to which if two products are jointly offered and one dominates the other, then the latter purchasing probability equals zero.

In what follows we conduct several experiments keeping unaltered the universe of products $\mathcal{J} = \{1, \dots, 30\}$ but progressively enforcing new dominance relations, each representing a 100% buydown effect and depicted via edges in graphs of Figure 24. Recall that the dominance relation is transitive, so if $j \succ j' \succ j''$ then $j \succ j''$ and we only highlight parsimonious graphs accordingly. At the most extreme case (see "Instance a", Figure 24a) no dominance relations exist among products (*i.e.*, no buydown effects are assumed), while at the other extreme (see "Instance f", Figure 24f) there exists a product dominating all the others (*i.e.*, product $j = 30$). Details on the number of buydown effects assumed for every instance are provided in Table 17.

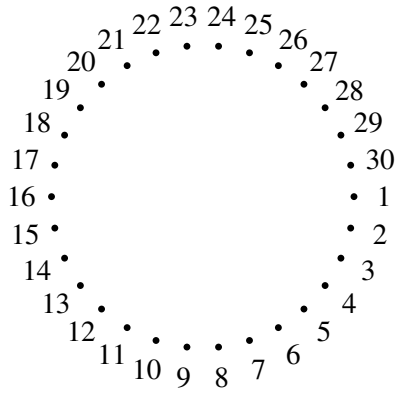
Instance	a	b	c	d	e	f
Number of buydown effects	0	20	50	74	98	170

Table 17: Number of dominance relations (buydown effects) among products of universe $\mathcal{J}$ for instances a to f of Figure 24.

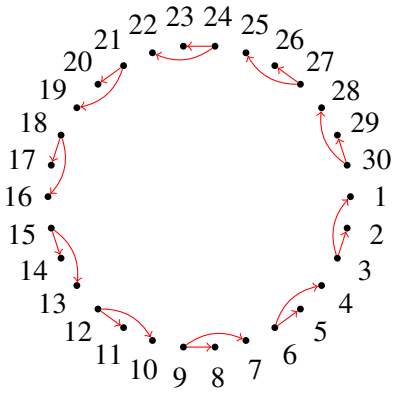
Problem data are as follows:

- There is a single resource $|\mathcal{R}| = 1$ with $b_1 = 1{,}000$ and $a_{1j} = 1$ for all $j \in \mathcal{J}$;
- Products unitary profits are $r = [14, 15, 19, 11, 14, 10, 14, 17, 11, 17, 13, 16, 11, 14, 10, 12, 12, 15, 12, 13, 18, 14, 10, 14, 15, 19, 13, 19, 16, 10]$;
- Dominance relations change for every instance and are graphically visualized via Figure 24.

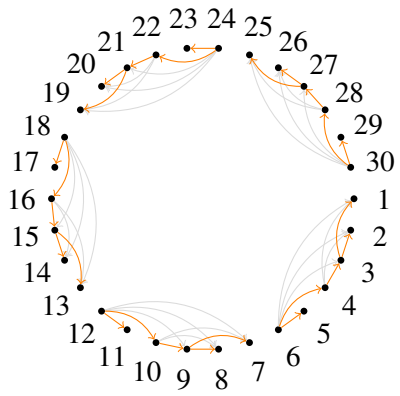
For every instance, we solve the DMNL-SBLP presented in Section 3.4 and then detect the CDLP optimal assortments A^{DMNL} by means of Algorithm 1. Similarly, we solve the standard SBLP presented in [63] and detect the corresponding CDLP optimal assortments A^{MNL}. For both models we set preference parameters $v_j = v_0 = 1$ for all $j \in \mathcal{J}$. Finally, given CDLP optimal assortments of both DMNL and MNL models, and knowing the true underlying purchasing probabilities $P^{\text{true}}_{j:A}$, we compute and compare out-of-sample total revenues, respectively, z^{DMNL} and z^{MNL}. Results are summarized in Table 18.



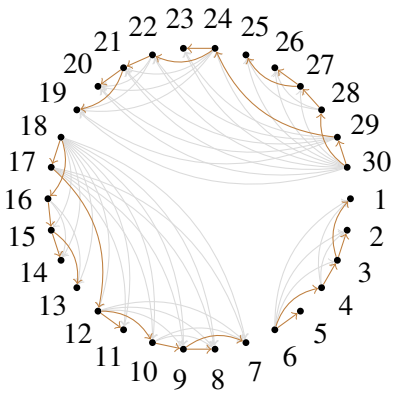
(a) Instance a: no dominance relations.



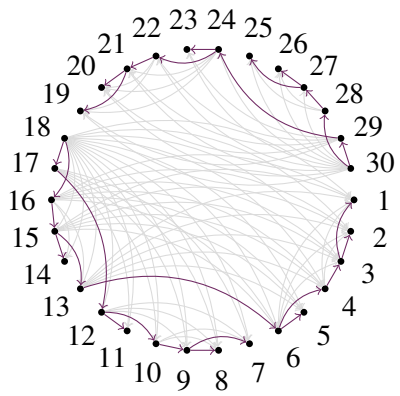
(b) Instance b.



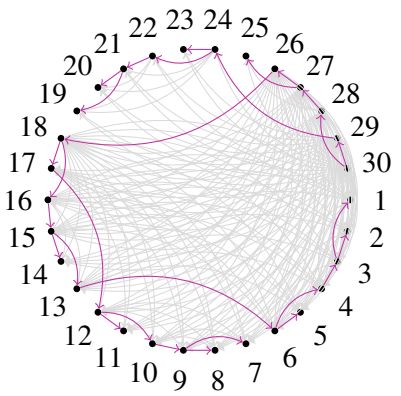
(c) Instance c.



(d) Instance d.



(e) Instance e.



(f) Instance f: one dominant product.

Figure 24: Dominance instances for the universe of products $\mathcal{J} = \{1, \ldots, 30\}$.

Instance	$\alpha\left(A^{\text{DMNL}}\right)$	z^{DMNL}	$\alpha\left(A^{\text{MNL}}\right)$	z^{MNL}	Δ
a	$\alpha\left(\{3,8,10,12,21,26,28,29\}\right)=1$	\$ 16,920.00	$\alpha\left(\{3,8,10,12,21,26,28,29\}\right)=1$	\$ 16,920.00	-
b	$\alpha\left(\{3,8,10,21,26,28,29\}\right)=1$	\$ 16,875.00	$\alpha\left(\{3,8,10,12,21,26,28,29\}\right)=1$	\$ 16,740.00	\$ 135.00
c	$\alpha\left(\{3,10,18,21,25,26,29\}\right)=1$	\$ 16,065.00	$\alpha\left(\{3,8,10,12,21,26,28,29\}\right)=1$	\$ 15,840.00	\$ 225.00
d	$\alpha\left(\{3,10,21,25,26\}\right)=1$	\$ 15,840.00	$\alpha\left(\{3,8,10,12,21,26,28,29\}\right)=1$	\$ 15,120.00	\$ 720.00
e	$\alpha\left(\{3,10,21,25,26\}\right)=1$	\$ 15,840.00	$\alpha\left(\{3,8,10,12,21,26,28,29\}\right)=1$	\$ 15,120.00	\$ 720.00
f	$\alpha\left(\{3,5,7,8,14,21,25\}\right)=1$	\$ 14,985.00	$\alpha\left(\{3,8,10,12,21,26,28,29\}\right)=1$	\$ 12,600.00	\$ 2,385.00

Table 18: Optimal assortments and out-of-sample revenues for DMNL and MNL models, with improvements Δ.

From Table 18 we observe that out-of-sample revenues scored by the DMNL model (z^{DMNL}) are always greater than (or equal to) the MNL out-of-sample results (z^{MNL}). The DMNL model successfully identifies optimal assortments that change across different instances to reflect and encode 100% buydown effects, demonstrating greater capacity to capture existing dominance relations among products. On the other hand, the MNL model –being unable to properly adapt to dominance relations– identifies optimal assortments that are unchanged across instances. Specifically, we notice that the relative revenue improvement

$$\Delta := z^{\text{DMNL}} - z^{\text{MNL}}$$

increases in the number of considered buydown effects, see Figure 25.

Indeed, under "Instance a", which is the situation with no dominance relations, the two models perform the same. However, the relative improvement Δ progressively grows as new dominance relations are gradually enforced. Finally, Δ registers the highest value (\$ 2,385.00) under "Instance f", the situation with more dominance relations among products belonging to the universe (highest number of buydown effects, see Table 17).

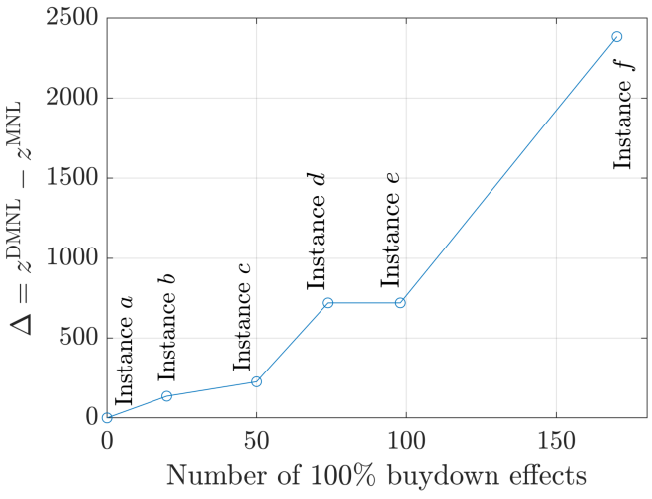


Figure 25: Relative improvement Δ versus number of buydown effects.

Overall, in our synthetic setting that deals with customers who reveal 100% buydown effects in their purchasing behavior, the DMNL model has proved to significantly outperform the MNL model in terms of out-of-sample accuracy. Therefore, it demonstrated to have greater generalization ability in suggesting assortments closer to the underlying truth.

3.7 Conclusions

In this work we propose a DMNL demand-based choice model, variant of the traditional MNL model. The proposal is motivated by the inability of traditional choice models to capture the recurrent phenomenon known as 100% buydown effect, which is the complete dominance of a product over another one and that annuls the purchasing probability of the latter. In this setting, under DMNL, we propose a deterministic approximation of the intractable dynamic revenue management problem, known as DMNL-SBLP. In the numerical experiments section we compare performance of RM problems under MNL and DMNL models, using synthetic data. An interesting future research direction would be evaluating the models using real-life data.

Acknowledgements

This research was carried out from August 2019 to February 2020, supported by the scholarship of the Bergamo University Ph.D. program in “Applied Economics & Management” and supervised by Prof. Francesca Maggioni.

Appendices

Appendix A

Appendix to Chapter 1

Box Robust Formulation

We now derive the box robust counterpart of formulation (1.3). Consider the problem:

$$\begin{array}{lll} \min\limits_{a,\gamma,z_X,z_Y} & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) & \\ \text{s.t.} & a^\top x \leq \gamma - 1 + z_{x^{(i)}} & \forall x \in \mathcal{U}_{\mathcal{B}}\Big(x^{(i)}\Big),\ i = 1, \ldots, I \\ & a^\top y \geq \gamma + 1 - z_{y^{(j)}} & \forall y \in \mathcal{U}_{\mathcal{B}}\Big(y^{(j)}\Big),\ j = 1, \ldots, J \\ & z_X \geq 0,\ z_Y \geq 0, & \end{array} \tag{A.1}$$

with:

$$\mathcal{U}_{\mathcal{B}}\Big(x^{(i)}\Big) := \Big\{ x \in \mathbb{R}^n \ \Big|\ x^{(i)} - \rho_X \zeta_{x^{(i)}} \leq x \leq x^{(i)} + \rho_X \zeta_{x^{(i)}} \Big\}, \tag{A.2}$$

$$\mathcal{U}_{\mathcal{B}}\Big(y^{(j)}\Big) := \Big\{ y \in \mathbb{R}^n \ \Big|\ y^{(j)} - \rho_Y \zeta_{y^{(j)}} \leq y \leq y^{(j)} + \rho_Y \zeta_{y^{(j)}} \Big\}, \tag{A.3}$$

where $\zeta_{x^{(i)}} \in \mathbb{R}^n_+$ and $\zeta_{y^{(j)}} \in \mathbb{R}^n_+$ define the perturbation vectors of observations $x^{(i)}$ and $y^{(j)}$, respectively, while $\rho_X \in \mathbb{R}_+$ and $\rho_Y \in \mathbb{R}_+$ are global measures of uncertainty. Formulation (A.1) can be equivalently re-stated as follows:

$$\begin{array}{lll} \min\limits_{a,\gamma,z_X,z_Y} & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) & \\ \text{s.t.} & \max\limits_{x \in \mathcal{U}_{\mathcal{B}}(x^{(i)})} \Big[a^\top x\Big] \leq \gamma - 1 + z_{x^{(i)}} & i = 1, \ldots, I \\ & \min\limits_{y \in \mathcal{U}_{\mathcal{B}}(y^{(j)})} \Big[a^\top y\Big] \geq \gamma + 1 - z_{y^{(j)}} & j = 1, \ldots, J \\ & z_X \geq 0,\ z_Y \geq 0. & \end{array} \tag{A.4}$$

The left-hand side of the first constraint in (A.4) can be re-written as follows:

$$\begin{array}{ll} \max\limits_{x} & a^\top x \\ \text{s.t.} & x \leq x^{(i)} + \rho_X \zeta_{x^{(i)}} \\ & x \geq x^{(i)} - \rho_X \zeta_{x^{(i)}} \end{array}$$

with dual given by:

$$\begin{array}{ll} \min\limits_{a^+,\, a^-} & \Big(x^{(i)} + \rho_X \zeta_{x^{(i)}}\Big)^\top a^+ - \Big(x^{(i)} - \rho_X \zeta_{x^{(i)}}\Big)^\top a^- \\ \text{s.t.} & a^+ - a^- = a \\ & a^+ \geq 0,\ a^- \geq 0, \end{array}$$

with a^+, a^- non-negative dual variables. The dual can equivalently be expressed as:

$$\begin{aligned} \min_{a^+,\, a^-} \quad & (a^+ - a^-)^\top x^{(i)} + \rho_X \zeta_{x^{(i)}}^\top (a^+ + a^-) \\ \text{s.t.} \quad & a^+ - a^- = a \\ & a^+ \geq 0,\ a^- \geq 0, \end{aligned}$$

which corresponds to $\min_a \ a^\top x^{(i)} + \rho_X \zeta_{x^{(i)}}^\top |a|$. Therefore the robust linear problem (A.4) now becomes:

$$\begin{aligned} \min_{a,\gamma,z_X,z_Y} \quad & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) \\ \text{s.t.} \quad & a^\top x^{(i)} + \rho_X \zeta_{x^{(i)}}^\top |a| \leq \gamma - 1 + z_{x^{(i)}} \quad i = 1, \dots, I \\ & \min_{y \in \mathcal{U}_{\mathcal{B}}(y^{(j)})} \left[a^\top y \right] \geq \gamma + 1 - z_{y^{(j)}} \qquad j = 1, \dots, J \\ & z_X \geq 0,\ z_Y \geq 0. \end{aligned} \tag{A.5}$$

Exploiting the following equivalence:

$$\min_{y \in \mathcal{U}_{\mathcal{B}}(y^{(j)})} \left[a^\top y \right] \geq \gamma + 1 - z_{y^{(j)}} \quad \Leftrightarrow \quad \max_{y \in \mathcal{U}_{\mathcal{B}}(y^{(j)})} \left[-a^\top y \right] \leq -\gamma - 1 + z_{y^{(j)}} \quad j = 1, \dots, J, \tag{A.6}$$

the same procedure can be followed for the second group of constraints of (A.4), leading to the following final robust formulation that corresponds to (1.9):

$$\begin{aligned} \min_{a,\gamma,z_X,z_Y} \quad & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) \\ \text{s.t.} \quad & a^\top x^{(i)} + \rho_X \zeta_{x^{(i)}}^\top |a| \leq \gamma - 1 + z_{x^{(i)}} \quad i = 1, \dots, I \\ & a^\top y^{(j)} - \rho_Y \zeta_{y^{(j)}}^\top |a| \geq \gamma + 1 - z_{y^{(j)}} \quad j = 1, \dots, J \\ & z_X \geq 0,\ z_Y \geq 0. \end{aligned} \tag{A.7}$$

Ellipsoidal Robust Formulation

We now derive the ellipsoidal robust counterpart of formulation (1.3). Let the problem:

$$\begin{aligned} \min_{a,\gamma,z_X,z_Y} \quad & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) \\ \text{s.t.} \quad & a^\top x \leq \gamma - 1 + z_{x^{(i)}} \qquad \forall x \in \mathcal{U}_{\mathcal{E}}\left(x^{(i)}\right),\ i = 1, \dots, I \\ & a^\top y \geq \gamma + 1 - z_{y^{(j)}} \qquad \forall y \in \mathcal{U}_{\mathcal{E}}\left(y^{(j)}\right),\ j = 1, \dots, J \\ & z_X \geq 0,\ z_Y \geq 0, \end{aligned} \tag{A.8}$$

be given, with:

$$\mathcal{U}_{\mathcal{E}}\left(x^{(i)}\right) := \left\{ x \in \mathbb{R}^n \;\middle|\; \begin{array}{c} x = x^{(i)} + \Sigma_{x^{(i)}}^{\frac{1}{2}} u \\ \\ \|u\|_2 \leq \rho_X \end{array} \right\}, \tag{A.9}$$

$$\mathcal{U}_{\mathcal{E}}\left(y^{(j)}\right) := \left\{ y \in \mathbb{R}^n \;\middle|\; \begin{array}{c} y = y^{(j)} + \Sigma_{y^{(j)}}^{\frac{1}{2}} u \\ \\ \|u\|_2 \leq \rho_Y \end{array} \right\}, \tag{A.10}$$

where $\Sigma_{x^{(i)}} \in \mathbb{R}^{n \times n}$ and $\Sigma_{y^{(j)}} \in \mathbb{R}^{n \times n}$ are positive definite covariance matrices for, respectively, $x^{(i)}$ and $y^{(j)}$, and with the scalars ρ_X, $\rho_Y \in \mathbb{R}_+$ denoting the radii of the ellipsoids centered in $x^{(i)}$ and $y^{(j)}$. Equivalently, uncertainty sets (A.9) and (A.10) may be expressed as follows:

$$\mathcal{U}_{\mathcal{E}}\left(x^{(i)}\right) := \left\{ x \in \mathbb{R}^n \;\middle|\; \left(x - x^{(i)}\right)^\top \Sigma_{x^{(i)}}^{-1} \left(x - x^{(i)}\right) \leq \rho_X^2 \right\}, \tag{A.11}$$

$$\mathcal{U}_{\mathcal{E}}\left(y^{(j)}\right) := \left\{ y \in \mathbb{R}^n \;\middle|\; \left(y - y^{(j)}\right)^\top \Sigma_{y^{(j)}}^{-1} \left(y - y^{(j)}\right) \leq \rho_Y^2 \right\}. \tag{A.12}$$

Once again, we can formulate our problem as:

$$\begin{array}{ll} \min\limits_{a,\gamma,z_X,z_Y} & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) \\ \text{s.t.} & \max\limits_{x \in \mathcal{U}_{\mathcal{E}}(x^{(i)})} \left[a^\top x\right] \leq \gamma - 1 + z_{x^{(i)}} \quad i = 1, \ldots, I \\ & \min\limits_{y \in \mathcal{U}_{\mathcal{E}}(y^{(j)})} \left[a^\top y\right] \geq \gamma + 1 - z_{y^{(j)}} \quad j = 1, \ldots, J \\ & z_X \geq 0, \; z_Y \geq 0. \end{array} \tag{A.13}$$

The left-hand side of the first constraint in (A.13) can be re-written as follows:

$$\begin{array}{ll} \max\limits_{x} & a^\top x \\ \text{s.t.} & x \in \mathcal{U}_{\mathcal{E}}\left(x^{(i)}\right) \end{array}$$

which is equivalent to:

$$\begin{array}{lll} a^\top x^{(i)} + & \max\limits_{u} & \left[a^\top \Sigma_{x^{(i)}}^{\frac{1}{2}} u\right] \\ & \text{s.t.} & \|u\|_2 \leq \rho_X. \end{array} \tag{A.14}$$

Applying the Cauchy-Schwarz inequality (see [155]) we get: $|a^\top \Sigma_{x^{(i)}}^{\frac{1}{2}} u| \leq \|a^\top \Sigma_{x^{(i)}}^{\frac{1}{2}}\|_2 \cdot \|u\|_2$. Therefore, since $\|u\|_2 \leq \rho_X$, problem (A.14) becomes:

$$a^\top x^{(i)} + \|a^\top \Sigma_{x^{(i)}}^{\frac{1}{2}}\|_2 \cdot \|u\|_2 \quad \Leftrightarrow \quad a^\top x^{(i)} + \rho_X \|\Sigma_{x^{(i)}}^{\frac{1}{2}} a\|_2. \tag{A.15}$$

Exploiting the equivalence:

$$\min_{y \in \mathcal{U}_{\mathcal{E}}(y^{(j)})} \left[a^\top y\right] \geq \gamma + 1 - z_{y^{(j)}} \quad \Leftrightarrow \quad \max_{y \in \mathcal{U}_{\mathcal{E}}(y^{(j)})} \left[-a^\top y\right] \leq -\gamma - 1 + z_{y^{(j)}} \quad j = 1, \ldots, J, \tag{A.16}$$

the same procedure can be followed for the second group of constraints of (A.13), leading to the final ellipsoidal robust formulation given by:

$$\begin{aligned} \min_{a,\gamma,z_X,z_Y} \quad & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) \\ \text{s.t.} \quad & a^\top x^{(i)} + \rho_X \|\Sigma_{x^{(i)}}^{\frac{1}{2}} a\|_2 \leq \gamma - 1 + z_{x^{(i)}} \quad i = 1, \ldots, I \\ & a^\top y^{(j)} - \rho_Y \|\Sigma_{y^{(j)}}^{\frac{1}{2}} a\|_2 \geq \gamma + 1 - z_{y^{(j)}} \quad j = 1, \ldots, J \\ & z_X \geq 0, \ z_Y \geq 0. \end{aligned} \tag{A.17}$$

Appendix B

Appendix to Chapter 2

Proof of LB Criterion for Kullback-Leibler Divergence

Recall from Section 2.2.3 that Kullback-Leibler (KL) divergence is a special case of the CR power divergence family when the parameter of this family $\theta \to 1$. We now prove the LB criteria $\bar{\bar{\rho}}+\bar{\rho}_{max} \leq \rho$ for KL divergence in Proposition 1 by directly working with the ambiguity set of KL divergence. We use subscript ϕ_{KL} $\mathcal{P}$ and Δ specific to KL.

Proof. Let $P' \in \widetilde{\mathcal{P}}_{\phi_{KL}(\bar{\bar{\rho}},\bar{\boldsymbol{\rho}})}$. Then there exists $\bar{\bar{P}} \in \widetilde{\mathcal{P}}^{\mathscr{G}}_{\phi_{KL}(\bar{\bar{\rho}})}$ and $\bar{P} \in \widetilde{\mathcal{P}}^{\mathscr{F}|\mathscr{G}}_{\phi_{KL}(\bar{\boldsymbol{\rho}})}$ such that $\sum_{g\in[m_l]} \bar{\bar{p}}^{(l)}_g = 1$, $\sum_{\omega_i\in\Omega^{(l)}_g} (\bar{p}_{\omega_i})^{(l)}_g = 1$ and

$$\sum_{g\in[m_l]} \left[\bar{\bar{p}}^{(l)}_g \log\left(\frac{\bar{\bar{p}}^{(l)}_g}{\pi^{(l)}_g}\right)\right] \leq \bar{\bar{\rho}}, \quad \sum_{\omega_i\in\Omega^{(l)}_g} \left[(\bar{p}_{\omega_i})^{(l)}_g \log\left(\frac{(\bar{p}_{\omega_i})^{(l)}_g}{(q_{\omega_i})^{(l)}_g}\right)\right] \leq \bar{\rho}_g.$$

We now show the steps to find the criteria for $\Delta_{\phi_{KL}}(P', Q) \leq \rho$. We have:

$$\begin{aligned}
&\sum_{\omega_i\in\Omega} \left[p'_{\omega_i} \log\left(\frac{p'_{\omega_i}}{q_{\omega_i}}\right)\right] \\
=&\sum_{\omega_i\in\Omega_f} \left[p'_{\omega_i} \log\left(\frac{p'_{\omega_i}}{q_{\omega_i}}\right)\right] + \sum_{\omega_i\in(\Omega_f)^C} \left[p'_{\omega_i} \log\left(\frac{p'_{\omega_i}}{q_{\omega_i}}\right)\right] \\
=&\sum_{\omega_i\in\Omega_f} \left[\sum_{g\in[m_l]} p'_{\omega_i,g} \log\left(\frac{\sum_{g\in[m_l]} p'_{\omega_i,g}}{\sum_{g\in[m_l]} q_{\omega_i,g}}\right)\right] + \sum_{\omega_i\in(\Omega_f)^C} \left[p'_{\omega_i} \log\left(\frac{p'_{\omega_i}}{q_{\omega_i}}\right)\right] \\
\leq&\sum_{\omega_i\in\Omega_f} \sum_{g\in[m_l]} \left[p'_{\omega_i,g} \log\left(\frac{p'_{\omega_i,g}}{q_{\omega_i,g}}\right)\right] + \sum_{\omega_i\in(\Omega_f)^C} \left[p'_{\omega_i} \log\left(\frac{p'_{\omega_i}}{q_{\omega_i}}\right)\right] \\
=&\sum_{\omega_i\in\Omega_f} \sum_{g\in[m_l]} \left[\bar{\bar{p}}^{(l)}_g (\bar{p}_{\omega_i})^{(l)}_g \log\left(\frac{\bar{\bar{p}}^{(l)}_g (\bar{p}_{\omega_i})^{(l)}_g}{\pi^{(l)}_g (q_{\omega_i})^{(l)}_g}\right)\right] + \sum_{g\in[m_l]} \sum_{\omega_i\in(\Omega_f)^C} \left[\bar{\bar{p}}^{(l)}_g (\bar{p}_{\omega_i})^{(l)}_g \log\left(\frac{\bar{\bar{p}}^{(l)}_g (\bar{p}_{\omega_i})^{(l)}_g}{\pi^{(l)}_g (q_{\omega_i})^{(l)}_g}\right)\right] \\
=&\sum_{g\in[m_l]} \sum_{\omega_i\in\Omega^{(l)}_g} \left[\bar{\bar{p}}^{(l)}_g\ (\bar{p}_{\omega_i})^{(l)}_g\ \left(\log\left(\frac{\bar{\bar{p}}^{(l)}_g}{\pi^{(l)}_g}\right) + \log\left(\frac{(\bar{p}_{\omega_i})^{(l)}_g}{(q_{\omega_i})^{(l)}_g}\right)\right)\right] \\
=&\sum_{g\in[m_l]} \left[\bar{\bar{p}}^{(l)}_g \log\left(\frac{\bar{\bar{p}}^{(l)}_g}{\pi^{(l)}_g}\right) \sum_{\omega_i\in\Omega^{(l)}_g} (\bar{p}_{\omega_i})^{(l)}_g\right] + \sum_{g\in[m_l]} \left[\bar{\bar{p}}^{(l)}_g \sum_{\omega_i\in\Omega^{(l)}_g} \left((\bar{p}_{\omega_i})^{(l)}_g \log\left(\frac{(\bar{p}_{\omega_i})^{(l)}_g}{(q_{\omega_i})^{(l)}_g}\right)\right)\right] \\
\leq&\sum_{g\in[m_l]} \left[\bar{\bar{p}}^{(l)}_g \log\left(\frac{\bar{\bar{p}}^{(l)}_g}{\pi^{(l)}_g}\right) 1\right] + \sum_{g\in[m_l]} \left[\bar{\bar{p}}^{(l)}_g \bar{\rho}_g\right] \\
\leq&\bar{\bar{\rho}} + \sum_{g\in[m_l]} \left[\bar{\bar{p}}^{(l)}_g\right] \cdot \bar{\rho}_{max} \\
=&\bar{\bar{\rho}} + \bar{\rho}_{max}
\end{aligned}$$

where the first inequality follows from the log sum inequality applied to fixed scenarios $\omega_i \in \Omega_f$ and the last set of inequalities follow from the facts that $\Delta_{\phi_{KL}}(\bar{P}_g^{(l)}, Q_g^{(l)}) \leq \bar{\rho}_g$ for all subgroups $g \in [m_l]$, the definition of $\bar{\rho}_{max}$, $\Delta_{\phi_{KL}}(\bar{\bar{P}}, \bar{\bar{Q}}) \leq \bar{\bar{\rho}}$, and $\sum_{\omega_i \in \Omega_g^{(l)}} (\bar{p}_{\omega_i})_g^{(l)} = \sum_{g \in [m_l]} \bar{\bar{p}}_g^{(l)} = 1$. Therefore, if $\bar{\bar{\rho}} + \bar{\rho}_{max} \leq \rho$, the result follows. □

Proof of LB Criterion for Variation Distance

Proof. Let $P' \in \widetilde{\mathcal{P}}_{\phi_v(\bar{\bar{\rho}}, \bar{\boldsymbol{\rho}})}$. Then there exists $\bar{\bar{P}} \in \widetilde{\mathcal{P}}^{\mathscr{G}}_{\phi_v(\bar{\bar{\rho}})}$ and $\bar{P} \in \widetilde{\mathcal{P}}^{\mathscr{F}|\mathscr{G}}_{\phi_v(\bar{\boldsymbol{\rho}})}$ such that $\sum_{g \in [m_l]} \bar{\bar{p}}_g^{(l)} = 1$, $\sum_{\omega_i \in \Omega_g^{(l)}} (\bar{p}_{\omega_i})_g^{(l)} = 1$ and

$$\sum_{g \in [m_l]} |\bar{\bar{p}}_g^{(l)} - \pi_g^{(l)}| \leq \bar{\bar{\rho}}, \quad \sum_{\omega_i \in \Omega_g^{(l)}} |(\bar{p}_{\omega_i})_g^{(l)} - (q_{\omega_i})_g^{(l)}| \leq \bar{\rho}_g$$

We now show the steps to find the criteria for $\Delta_{\phi_v}(P', Q) \leq \rho$. We have:

$$\begin{aligned}
&\sum_{\omega_i \in \Omega} \left| p'_{\omega_i} - q_{\omega_i} \right| \\
&= \sum_{\omega_i \in \Omega_f} \left| \sum_{g \in [m_l]} p'_{\omega_i, g} - \sum_{g \in [m_l]} q_{\omega_i, g} \right| + \sum_{\omega_i \in (\Omega_f)^C} \left| p'_{\omega_i} - q_{\omega_i} \right| \\
&\leq \sum_{\omega_i \in \Omega_f} \sum_{g \in [m_l]} \left| p'_{\omega_i, g} - q_{\omega_i, g} \right| + \sum_{\omega_i \in (\Omega_f)^C} \left| p'_{\omega_i} - q_{\omega_i} \right| \\
&= \sum_{g \in [m_l]} \sum_{\omega_i \in \Omega_g^{(l)}} \left| \bar{\bar{p}}_g^{(l)} (\bar{p}_{\omega_i})_g^{(l)} - \pi_g^{(l)} (q_{\omega_i})_g^{(l)} \right| \\
&\leq \sum_{g \in [m_l]} \sum_{\omega_i \in \Omega_g^{(l)}} \left(\left| \bar{\bar{p}}_g^{(l)} - \pi_g^{(l)} \right| \left| (\bar{p}_{\omega_i})_g^{(l)} - (q_{\omega_i})_g^{(l)} \right| + \left| \left(\bar{\bar{p}}_g^{(l)} - \pi_g^{(l)} \right) (q_{\omega_i})_g^{(l)} \right| + \left| \pi_g^{(l)} \left((\bar{p}_{\omega_i})_g^{(l)} - (q_{\omega_i})_g^{(l)} \right) \right| \right) \\
&= \sum_{g \in [m_l]} \left(\left| \bar{\bar{p}}_g^{(l)} - \pi_g^{(l)} \right| \sum_{\omega_i \in \Omega_g^{(l)}} \left| (\bar{p}_{\omega_i})_g^{(l)} - (q_{\omega_i})_g^{(l)} \right| \right) + \sum_{g \in [m_l]} \left(\left| \bar{\bar{p}}_g^{(l)} - \pi_g^{(l)} \right| \sum_{\omega_i \in \Omega_g^{(l)}} (q_{\omega_i})_g^{(l)} \right) \\
&\quad + \sum_{g \in [m_l]} \left(\pi_g^{(l)} \sum_{\omega_i \in \Omega_g^{(l)}} \left| (\bar{p}_{\omega_i})_g^{(l)} - (q_{\omega_i})_g^{(l)} \right| \right) \\
&\leq \sum_{g \in [m_l]} \left(\left| \bar{\bar{p}}_g^{(l)} - \pi_g^{(l)} \right| \bar{\rho}_g \right) + \bar{\bar{\rho}} + \sum_{g \in [m_l]} \left(\pi_g^{(l)} \bar{\rho}_g \right) \\
&\leq \bar{\bar{\rho}} \cdot \bar{\rho}_{max} + \bar{\bar{\rho}} + \bar{\rho}_{max},
\end{aligned}$$

where the equality on the second line above follows from the way $p'_{\omega_i, g}$, $q_{\omega_i, g}$ for fixed and p'_{ω_i}, q_{ω_i} for non-fixed scenarios are defined in set (2.4) and in Sections 2.3.1 and 2.3.2. The second inequality follows from, for any numbers a, b, c, d, that we have $|ac - bd| = |(a-b)(c-d) + (a-b)d + b(c-d)| \leq |(a-b)||(c-d)| + |(a-b)d| + |b(c-d)|$. The last two sets of inequalities follow from the facts that $\Delta_{\phi_v}(\bar{P}_g^{(l)}, Q_g^{(l)}) \leq \bar{\rho}_g$ for all subgroups $g \in [m_l]$, the definition of $\bar{\rho}_{max}$, $\Delta_{\phi_v}(\bar{\bar{P}}, \bar{\bar{Q}}) \leq \bar{\bar{\rho}}$, and $\sum_{\omega_i \in \Omega_g^{(l)}} (q_{\omega_i})_g^{(l)} = \sum_{g \in [m_l]} \pi_g^{(l)} = 1$. Therefore, if $\bar{\bar{\rho}} \cdot \bar{\rho}_{max} + \bar{\bar{\rho}} + \bar{\rho}_{max} \leq \rho$ the result follows. □

Proof of LB Criterion for J-Divergence

Proof. J-divergence is the sum of KL divergence and Burg entropy [8]. Burg entropy is similar to the KL divergence with q_{ω_i} and p_{ω_i} exchanged (see Table 9). Therefore, the proof of Burg entropy follows along the same lines as KL divergence. Splitting the J-divergence as sum of KL divergence and Burg entropy, following along the lines of the proof above and setting

$$a_{\omega_i} = \log\left(\frac{(\bar{p}_{\omega_i})_g^{(l)}}{(q_{\omega_i})_g^{(l)}}\right)$$

we obtain

$$\begin{aligned}
\Delta_{\phi_J}(P', Q) &\leq \bar{\bar{\rho}} + \sum_{g\in[m_l]}\left[\bar{\bar{p}}_g^{(l)} \sum_{\omega_i\in\Omega_g^{(l)}} \left((\bar{p}_{\omega_i})_g^{(l)} a_{\omega_i}\right)\right] - \sum_{g\in[m_l]}\left[\pi_g^{(l)} \sum_{\omega_i\in\Omega_g^{(l)}} \left((q_{\omega_i})_g^{(l)} a_{\omega_i}\right)\right] \\
&= \bar{\bar{\rho}} + \sum_{g\in[m_l]}\left[\bar{\bar{p}}_g^{(l)} \sum_{\omega_i\in\Omega_g^{(l)}} \left[(\bar{p}_{\omega_i})_g^{(l)} - (q_{\omega_i})_g^{(l)}\right] a_{\omega_i}\right] + \sum_{g\in[m_l]}\left[\left(\bar{\bar{p}}_g^{(l)} - \pi_g^{(l)}\right) \sum_{\omega_i\in\Omega_g^{(l)}} \left((q_{\omega_i})_g^{(l)} a_{\omega_i}\right)\right] \\
&\leq \bar{\bar{\rho}} + \sum_{g\in[m_l]} \bar{\bar{p}}_g^{(l)} \bar{\rho}_g + \left[\max_g \sum_{\omega_i\in\Omega_g^{(l)}} \left((q_{\omega_i})_g^{(l)} a_{\omega_i}\right)\right] \sum_{g\in[m_l]} \left(\bar{\bar{p}}_g^{(l)} - \pi_g^{(l)}\right) \\
&\leq \bar{\bar{\rho}} + \bar{\rho}_{max} \cdot 1 + 0,
\end{aligned}$$

where the inequalities on the last line follow from the facts that $\Delta_{\phi_J}(\bar{P}_g^{(l)}, Q_g^{(l)}) \leq \bar{\rho}_g$ for all subgroups $g \in [m_l]$, the definition of $\bar{\rho}_{max}$, and $\sum_{g\in[m_l]} \bar{\bar{p}}_g^{(l)} = \sum_{g\in[m_l]} \pi_g^{(l)} = 1$. Therefore, if $\bar{\bar{\rho}} + \bar{\rho}_{max} \leq \rho$, the result follows. □

Proof of LB Criterion for χ-Divergence of $a > 1$

Proof. Set $x_g^{(l)} = 1 - \frac{\bar{\bar{p}}_g^{(l)}}{\pi_g^{(l)}}$ and $(y_{\omega_i})_g^{(l)} = 1 - \frac{(\bar{p}_{\omega_i})_g^{(l)}}{(q_{\omega_i})_g^{(l)}}$. Let $P' \in \widetilde{\mathcal{P}}_{\phi_\chi^a(\bar{\bar{\rho}},\bar{\boldsymbol{\rho}})}$. Then there exists $\bar{\bar{P}} \in \widetilde{\mathcal{P}}^{\mathscr{G}}_{\phi_\chi^a(\bar{\bar{\rho}})}$ and $\bar{P} \in \widetilde{\mathcal{P}}^{\mathscr{F}|\mathscr{G}}_{\phi_\chi^a(\bar{\boldsymbol{\rho}})}$ such that

$$\sum_{g\in[m_l]} \pi_g^{(l)} |x_g^{(l)}|^a \leq \bar{\bar{\rho}}, \quad \sum_{\omega_i\in\Omega_g^{(l)}} (q_{\omega_i})_g^{(l)} |(y_{\omega_i})_g^{(l)}|^a \leq \bar{\rho}_g$$

for all subgroups $g \in [m_l]$. Since the scenario tree Ω is dissected using disjoint partitions (*i.e.*, $\Omega_f = \emptyset$), we have $q_{\omega_i} = \pi_g^{(l)} (q_{\omega_i})_g^{(l)}$, for all $\omega_i \in \Omega_g^{(l)}$ and $g \in [m_l]$. Then:

$$\begin{aligned}
&\sum_{\omega_i\in\Omega} q_{\omega_i} \left|1 - \frac{p'_{\omega_i}}{q_{\omega_i}}\right|^a \\
&= \sum_{g\in[m_l]} \sum_{\omega_i\in\Omega_g^{(l)}} q_{\omega_i} \left|1 - \frac{\bar{\bar{p}}_g^{(l)} (\bar{p}_{\omega_i})_g^{(l)}}{\pi_g^{(l)} (q_{\omega_i})_g^{(l)}}\right|^a \\
&= \sum_{g\in[m_l]} \sum_{\omega_i\in\Omega_g^{(l)}} q_{\omega_i} \left|x_g^{(l)} + (y_{\omega_i})_g^{(l)} - x_g^{(l)} (y_{\omega_i})_g^{(l)}\right|^a
\end{aligned}$$

$$
\begin{aligned}
&\leq \left[\left(\sum_{g\in[m_l]}\sum_{\omega_i\in\Omega_g^{(l)}} q_{\omega_i}\left|x_g^{(l)}\right|^a\right)^{\frac{1}{a}}+\left(\sum_{g\in[m_l]}\sum_{\omega_i\in\Omega_g^{(l)}} q_{\omega_i}\left|(y_{\omega_i})_g^{(l)}\right|^a\right)^{\frac{1}{a}}+\left(\sum_{g\in[m_l]}\sum_{\omega_i\in\Omega_g^{(l)}} q_{\omega_i}\left|x_g^{(l)}(y_{\omega_i})_g^{(l)}\right|^a\right)^{\frac{1}{a}}\right]^a\\
&=\left[\left(\sum_{g\in[m_l]}\pi_g^{(l)}\left|x_g^{(l)}\right|^a\sum_{\omega_i\in\Omega_g^{(l)}}(q_{\omega_i})_g^{(l)}\right)^{\frac{1}{a}}+\left(\sum_{g\in[m_l]}\pi_g^{(l)}\sum_{\omega_i\in\Omega_g^{(l)}}(q_{\omega_i})_g^{(l)}\left|(y_{\omega_i})_g^{(l)}\right|^a\right)^{\frac{1}{a}}\right.\\
&\quad\left.+\left(\sum_{g\in[m_l]}\pi_g^{(l)}\left|x_g^{(l)}\right|^a\sum_{\omega_i\in\Omega_g^{(l)}}(q_{\omega_i})_g^{(l)}\left|(y_{\omega_i})_g^{(l)}\right|^a\right)^{\frac{1}{a}}\right]^a\\
&\leq\left[\left(\sum_{g\in[m_l]}\pi_g^{(l)}\left|x_g^{(l)}\right|^a\right)^{\frac{1}{a}}+\left(\sum_{g\in[m_l]}\pi_g^{(l)}\bar{\rho}_g\right)^{\frac{1}{a}}+\left(\sum_{g\in[m_l]}\pi_g^{(l)}\left|x_g^{(l)}\right|^a\bar{\rho}_g\right)^{\frac{1}{a}}\right]^a\\
&\leq\left[(\bar{\bar{\rho}})^{\frac{1}{a}}+(\bar{\rho}_{max})^{\frac{1}{a}}+(\bar{\bar{\rho}}\cdot\bar{\rho}_{max})^{\frac{1}{a}}\right]^a,
\end{aligned}
$$

where the first inequality follows from Minkowski inequality and the last two set of inequalities follow from the facts that $\Delta_{\phi_\chi^a}(\bar{\bar{P}},\bar{\bar{Q}})\leq\bar{\bar{\rho}}$ and $\Delta_{\phi_\chi^a}(\bar{P}_g^{(l)},Q_g^{(l)})\leq\bar{\rho}_g$ for all subgroups $g\in[m_l]$, the definition of $\bar{\rho}_{max}$, and $\sum_{\omega_i\in\Omega_g^{(l)}}(q_{\omega_i})_g^{(l)}=\sum_{g\in[m_l]}\pi_g^{(l)}=1$. Therefore, if $\left[\left(\bar{\bar{\rho}}\right)^{\frac{1}{a}}+(\bar{\rho}_{max})^{\frac{1}{a}}+\left(\bar{\bar{\rho}}\cdot\bar{\rho}_{max}\right)^{\frac{1}{a}}\right]^a\leq\rho$ the result follows. □

Appendix C

Appendix to Chapter 3

Upper Bound on Number of Maximal Paths

First we show that a dominance graph with a maximal number of maximal paths has the same structure as the dominance graph in Example 1. Consider any dominance graph $\mathfrak{G}$ with set $\mathfrak{P}$ of maximal paths. For each $j \in \mathcal{J}$, let $\ell(j)$ denote the least number of arcs along any path in $\mathfrak{G}$ from any $j' \in \mathcal{J}$ such that $\mathcal{D}^-(j') = \varnothing$ to j. Let $L := \max\{\ell(j) \,:\, j \in \mathcal{J}\} + 1$. For each $\ell \in \{0, \ldots, L-1\}$, let $\mathcal{J}(\ell) := \{j \in \mathcal{J} \,:\, \ell(j) = \ell\}$. Next, let $\mathfrak{G}'$ denote the dominance graph with node set $\mathcal{J}$, and with an arc (j', j) from each $j' \in \mathcal{J}(\ell-1)$ to each $j \in \mathcal{J}(\ell)$, for each $\ell \in \{1, \ldots, L-1\}$. Let $\mathfrak{P}'$ denote the set of maximal paths in $\mathfrak{G}'$. Note that $|\mathfrak{P}'| \geq |\mathfrak{P}|$, and that $|\mathfrak{P}'| = |\mathcal{J}(0)| \times \cdots \times |\mathcal{J}(L-1)|$.
Next, we determine an upper bound on $|\mathcal{J}(0)| \times \cdots \times |\mathcal{J}(L-1)|$ subject to $|\mathcal{J}(0)| + \cdots + |\mathcal{J}(L-1)| = |\mathcal{J}|$, by relaxing the integrality requirement on each $|\mathcal{J}(\ell)|$. For any $\ell \in \{1, 2, \ldots\}$ and any $y \geq 0$, let

$$f_\ell(y) \;:=\; \max\left\{x_1 \times \cdots \times x_\ell \,:\, x_1 + \cdots + x_\ell = y,\; x_1, \ldots, x_\ell \geq 0\right\}.$$

Note that $f_1(y) = y$ and $f_{\ell+1}(y) = \max\{x \times f_\ell(y-x) \,:\, x \in [0, y]\}$ for all $\ell \in \{1, 2, \ldots\}$ and all $y \geq 0$. Next, we show by induction on ℓ that $f_\ell(y) = (y/\ell)^\ell$. Note that the induction hypothesis holds for $\ell = 1$. Consider

$$f_{\ell+1}(y) \;=\; \max\{x \times f_\ell(y-x) \,:\, x \in [0,y]\} \;=\; \max\left\{x \times \left(\frac{y-x}{\ell}\right)^\ell \,:\, x \in [0,y]\right\}.$$

Note that $g(x) := x \times \left(\frac{y-x}{\ell}\right)^\ell$ is maximized at $x^* = y/(\ell+1) \in [0, y]$, and thus $f_{\ell+1}(y) = (y/(\ell+1))^{\ell+1}$, and hence the induction hypothesis has been established. In words, this result shows that, ignoring integrality requirements, a dominance graph with a maximal number of maximal paths has the same number of nodes in each level ℓ.
Next, we determine the number of levels that maximizes the number of maximal paths for a given total number of nodes. For any $y > 0$, consider

$$\max\left\{f_\ell(y) = \left(\frac{y}{\ell}\right)^\ell \,:\, \ell > 0\right\}$$

Note that $\left(\frac{y}{\ell}\right)^\ell$ is maximized at $\ell^* = y/e$, and that

$$\max\left\{f_\ell(y) = \left(\frac{y}{\ell}\right)^\ell \,:\, \ell > 0\right\} \;=\; \exp(y/e).$$

Thus, given product set $\mathcal{J}$, the number of maximal paths is less than $\exp(|\mathcal{J}|/e)$.

Conversion of an SBLP Solution into a CDLP Solution: A Practical Example

This example is used to explain Algorithm 1. Suppose universe $\mathcal{J}$ has three products, so $\mathcal{J} = \{1, 2, 3\}$, whose profits per unit are $r_1 = 283$, $r_2 = 276$, and $r_3 = 286$. There is a single available resource (so $|\mathcal{R}| = 1$) with $b_1 = 560$, and it holds that $a^1 = a^2 = a^3 = 1$. Preference weight parameters are $v_1 = 10, v_2 = 7, v_3 = 2$, and $v_0 = 2$. Finally, $T = 1{,}080$ and $\lambda = 0.59$. The dominance relations between products are described by the following graph.

Solving problem (SBLP) leads to the following solution: $x_1 = 433.33$, $x_2 = 56.00$, $x_3 = 70.66$, and $x_0 = 86.66$ with an objective function value of \$158,300. With this solution as input, we invoke Algorithm 1 to recover optimal solutions of problem (CDLP).

- We begin by setting $k \leftarrow 0$, $\alpha(A) = 0$ for all $A \subseteq \mathcal{J}$ (Step 1).
- Since there exists at least a product $j \in \mathcal{J}$ such that $x_j > 0$ (namely, $j = 1, 2, 3$), we enter the loop and set $k \leftarrow 1$ (Step 3).
- We build set $A_1 = \{1, 2, 3\}$ (Step 4).
- We build set $D_1 = \{1, 3\}$ (Step 5).
- We set $Y_1 = \min\left\{\frac{x_1}{v_1}, \frac{x_3}{v_3}\right\} = \min\left\{\frac{433.33}{10}, \frac{70.66}{2}\right\} = 35.33$ (Step 6).
- We set $j_1 = 3$ (Step 7).
- We set $\alpha(A_1) = \frac{v_0+v_1+v_3}{\lambda T} \cdot Y_1 = \frac{2+10+2}{646.66} \cdot 35.33 = 0.7649$ (Step 8).
- We set $x_1 = 433.33 - 646.66 \cdot 0.7649 \cdot \frac{10}{2+10+2} = 80.00$ (Step 10).
- We set $x_3 = 70.66 - 646.66 \cdot 0.7649 \cdot \frac{2}{2+10+2} = 0$ (Step 10).
- Since there exists at least a product $j \in \mathcal{J}$ such that $x_j > 0$ (namely, $j = 1, 2$), we stay in the loop and set $k \leftarrow 2$ (Step 3).
- We build set $A_2 = \{1, 2\}$ (Step 4).
- We build set $D_2 = \{1, 2\}$ (Step 5).
- We set $Y_2 = \min\left\{\frac{x_1}{v_1}, \frac{x_2}{v_2}\right\} = \min\left\{\frac{80.00}{10}, \frac{56.00}{7}\right\} = 8.00$ (Step 6).
- We set $j_2 = 2$ (Step 7).
- We set $\alpha(A_2) = \frac{v_0+v_1+v_2}{\lambda T} \cdot Y_2 = \frac{2+10+7}{646.66} \cdot 8 = 0.2351$ (Step 8).

- We set $x_1 = 80 - 646.66 \cdot 0.2351 \cdot \frac{10}{2+10+7} = 0$ (Step 10).
- We set $x_2 = 56 - 646.66 \cdot 0.2351 \cdot \frac{7}{2+10+7} = 0$ (Step 10).
- Since $\{j \in \mathcal{J} : x_j > 0\} = \varnothing$ Algorithm 1 stops.

Output: $\alpha(A_1) = 0.7649$, $\alpha(A_2) = 0.2351$, and $\alpha(A) = 0$ for all remaining $A \subseteq \mathcal{J}$.

List of Figures

List of Tables

References

[1] B. Analui and G. C. Pflug. On distributionally robust multiperiod stochastic optimization. *Computational Management Science*, 11(3):197–220, 2014.

[2] Y. Anzai. *Pattern recognition and machine learning*. Elsevier, 2012.

[3] A. Ardestani-Jaafari and E. Delage. Robust optimization of sums of piecewise linear functions with application to inventory problems. *Operations Research*, 64(2):474–494, 2016.

[4] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

[5] P. Baumann, D. S. Hochbaum, and Y. T. Yang. A comparative study of the leading machine learning techniques and two new optimization algorithms. *European Journal of Operational Research*, 272(3):1041–1057, 2019.

[6] G. Bayraksan and D. K. Love. Data-driven stochastic programming using phi-divergences. In *The Operations Research Revolution*, pages 1–19. INFORMS, 2015.

[7] A. Ben-Tal, S. Bhadra, C. Bhattacharyya, and J. S. Nath. Chance constrained uncertain classification via robust optimization. *Mathematical Programming*, 127(1):145–173, 2011.

[8] A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[9] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*, volume 28. Princeton University Press, 2009.

[10] K. P. Bennett and O. L. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1(1):23–34, 1992.

[11] L. Bertazzi and F. Maggioni. A stochastic multi-stage fixed charge transportation problem: Worst-case analysis of the rolling horizon approach. *European Journal of Operational Research*, 267(2):555–569, 2018.

[12] D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.

[13] D. Bertsimas, J. Dunn, C. Pawlowski, and Y. D. Zhuo. Robust classification. *INFORMS Journal on Optimization*, 1(1):2–34, 2019.

[14] D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292, 2018.

[15] D. Bertsimas and V. V. Mišic. Data-driven assortment optimization. *Management Science*, 1:1–35, 2015.

[16] D. Bertsimas, S. Shtern, and B. Sturt. A data-driven approach for multi-stage linear optimization. *Available from Optimization Online*, 2018.

[17] D. Bertsimas, M. Sim, and M. Zhang. Adaptive distributionally robust optimization. *Management Science*, 65(2):604–618, 2019.

[18] S. Bhadra, J. S. Nath, A. Ben-Tal, and C. Bhattacharyya. Interval data classification under partial information: A chance-constraint approach. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 208–219. Springer, 2009.

[19] C. Bhattacharyya. Robust classification of noisy data using second order cone programming approach. In *Proceedings of the International Conference on Intelligent Sensing and Information Processing*, pages 433–438. IEEE, 2004.

[20] C. Bhattacharyya, L. Grate, M. I. Jordan, L. E. Ghaoui, and I. S. Mian. Robust sparse hyperplane classifiers: Application to uncertain molecular profiling data. *Journal of Computational Biology*, 11(6):1073–1089, 2004.

[21] C. Bhattacharyya, K. Pannagadatta, and A. J. Smola. A second order cone programming formulation for classifying missing data. In *Neural Information Processing Systems*, pages 153–160, 2005.

[22] J. Bi and T. Zhang. Support vector classification with input data uncertainty. *Advances in Neural Information Processing Systems*, 17(1):161–168, 2005.

[23] B. Biggio, I. Corona, B. Nelson, B. I. Rubinstein, D. Maiorca, G. Fumera, G. Giacinto, and F. Roli. Security evaluation of support vector machines in adversarial environments. In *Support Vector Machines Applications*, pages 105 – 153. Springer, 2014.

[24] B. Biggio, B. Nelson, and P. Laskov. Support vector machines under adversarial label noise. In *Asian Conference on Machine Learning*, pages 97–112, 2011.

[25] J. R. Birge. The value of the stochastic solution in stochastic linear programs with fixed recourse. *Mathematical Programming*, 24(1):314–325, 1982.

[26] J. R. Birge. Optimization methods in dynamic portfolio management. In *Financial Engineering, Handbooks in Operations Research and Management Science*, pages 845–865. Elsevier, 2007.

[27] J. R. Birge and F. Louveaux. *Introduction to stochastic programming*. Springer Science & Business Media, 2011.

[28] J. Blanchet, G. Gallego, and V. Goyal. A Markov chain approximation to choice modeling. *Operations Research*, 64(4):886–905, 2016.

[29] E. A. Boyd and R. Kallesen. Practice papers: The science of revenue management when passengers purchase the lowest available fare. *Journal of Revenue and Pricing Management*, 3(2):171–177, 2004.

[30] J. J. M. Bront, I. Méndez-Díaz, and G. Vulcano. A column generation algorithm for choice-based network revenue management. *Operations Research*, 57(3):769–784, 2009.

[31] Q. Cao, X. Fu, and Y. Guo. Fuzzy chance constrained twin support vector machine for uncertain classification. In *International Conference on Management Science and Engineering Management*, pages 1508–1521. Springer, 2017.

[32] Y. Cao, A. J. Kleywegt, and H. Wang. Network revenue management under a spiked multinomial logit choice model. Technical report, Georgia Institute of Technology, 2020.

[33] C. Caramanis and S. Mannor. Learning in the limit with adversarial disturbances. In *COLT*, pages 467–478. Citeseer, 2008.

[34] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408:189–215, 2020.

[35] E. Ceseracciu, M. Reggiani, Z. Sawacha, M. Sartori, F. Spolaor, C. Cobelli, and E. Pagello. SVM classification of locomotion modes using surface electromyography for applications in rehabilitation robotics. In *19th International Symposium in Robot and Human Interactive Communication*, pages 165–170. IEEE, 2010.

[36] T. Y. Chen, T. Tse, and Y.-T. Yu. Proportional sampling strategy: A compendium and some insights. *Journal of Systems and Software*, 58(1):65–81, 2001.

[37] M. R. Chernick. *Bootstrap methods: A guide for practitioners and researchers*, volume 619. John Wiley & Sons, 2011.

[38] W. L. Cooper and L. Li. On the use of buy up as a model of customer choice in revenue management. *Production and Operations Management*, 21(5):833–850, 2012.

[39] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[40] J. Dai, W. Ding, A. Kleywegt, X. Wang, and Y. Zhang. Choice based revenue management for parallel flights. *SSRN:2404193*, 2014.

[41] J. M. Davis, G. Gallego, and H. Topaloglu. Assortment optimization under variants of the nested logit model. *Operations Research*, 62(2):250–273, 2014.

[42] S. De Cosmis, R. De Leone, E. Kropat, S. Meyer-Nieberg, and S. Pickl. Electric load forecasting using support vector machines for robust regression. In *Proceedings of the Emerging M&S Applications in Industry & Academia: Modeling and Humanities Symposium*, pages 1–8, 2013.

[43] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

[44] W. Ding. *Estimation and optimization problems in revenue management with customer choice behavior*. PhD thesis, Georgia Institute of Technology, 2017.

[45] S. Dreiseitl and L. Ohno-Machado. Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, 35(5-6):352–359, 2002.

[46] D. Dua and C. Graff. UCI Machine Learning Repository.

[47] J. Duchi and H. Namkoong. Variance-based regularization with convex objectives. *The Journal of Machine Learning Research*, 20(1):2450–2504, 2019.

[48] J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

[49] S. A. Dudani. The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, 4:325–327, 1976.

[50] D. Duque and D. P. Morton. Distributionally robust stochastic dual dynamic programming. *SIAM Journal on Optimization*, 30(4):2841–2865, 2020.

[51] N. Edirisinghe. Bound-based approximations in multistage stochastic programming: Nonanticipativity aggregation. *Annals of Operations Research*, 85:103–127, 1999.

[52] N. Edirisinghe and W. Ziemba. Tight bounds for stochastic convex programs. *Operations Research*, 40(4):660–677, 1992.

[53] L. El Ghaoui, G. R. G. Lanckriet, and G. Natsoulis. Robust classification with interval data. Technical report, University of California, 2003.

[54] N. Fan, E. Sadeghi, and P. M. Pardalos. Robust support vector machines with polyhedral uncertainty of the input data. In *International Conference on Learning and Intelligent Optimization*, pages 291–305. Springer, 2014.

[55] V. F. Farias, S. Jagabathula, and D. Shah. A nonparametric approach to modeling choice with limited data. *Management Science*, 59(2):305–322, 2013.

[56] J. B. Feldman and H. Topaloglu. Capacity constraints across nests in assortment optimization under the nested logit model. *Operations Research*, 63(4):812–822, 2015.

[57] J. B. Feldman and H. Topaloglu. Revenue management under the markov chain choice model. *Operations Research*, 65(5):1322–1342, 2017.

[58] K. Frauendorfer, D. Kuhn, and M. Schürle. *Barycentric bounds in stochastic programming: Theory and application*, pages 67–96. Springer, New York, 2011.

[59] K. Frauendorfer and M. Schürle. *Multistage stochastic programming: Barycentric approximation*, pages 2527–2531. Springer, Boston, 2009.

[60] S. Fujiwara, A. Takeda, and T. Kanamori. DC algorithm for extended robust support vector machine. *Neural Computation*, 29(5):1406–1438, 2017.

[61] G. Fung, O. L. Mangasarian, and J. W. Shavlik. Knowledge-based support vector machine classifiers. In *NIPS*, pages 521–528. Citeseer, 2003.

[62] G. Gallego, G. Iyengar, R. Phillips, and A. Dubey. Managing flexible products on a network. *SSRN:3567371*, 2004.

[63] G. Gallego, R. Ratliff, and S. Shebalov. A general attraction model and sales-based linear program for network revenue management under customer choice. *Operations Research*, 63(1):212–232, 2014.

[64] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal. Pattern classification with missing data: A review. *Neural Computing and Applications*, 19(2):263–282, 2010.

[65] L. E. Ghaoui, M. Oks, and F. Oustry. Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Operations Research*, 51(4):543–556, 2003.

[66] A. Globerson and S. Roweis. Nightmare at test time: Robust learning by feature deletion. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 353–360. ACM, 2006.

[67] J. Goh and M. Sim. Distributionally robust optimization and its tractable approximations. *Operations Research*, 58(4):902–917, 2010.

[68] D. Goldfarb and G. Iyengar. Robust convex quadratically constrained programs. *Mathematical Programming*, 97(3):495–515, 2003.

[69] B. L. Gorissen, İ. Yanıkoğlu, and D. den Hertog. A practical guide to robust optimization. *Omega*, 53:124–137, 2015.

[70] J. Gotoh, A. Takeda, and R. Yamamoto. Interaction between financial risk measures and machine learning methods. *Computational Management Science*, 11(4):365–402, 2014.

[71] J. Gotoh and S. Uryasev. Support vector machines based on convex risk functions and general norms. *Annals of Operations Research*, 249:301–328, 2017.

[72] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

[73] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417, 1933.

[74] J. Huang, K. Zhou, and Y. Guan. A study of distributionally robust multistage stochastic optimization. *arXiv:1708.07930*, 2017.

[75] A. H. Hübner and H. Kuhn. Retail category management: State-of-the-art review of quantitative research and software applications in assortment and shelf space management. *Omega*, 40(2):199–209, 2012.

[76] S. Jagabathula and P. Rusmevichientong. The limit of rationality in choice modeling: Formulation, computation, and implications. *Management Science*, 65(5):2196–2215, 2018.

[77] R. Jiang and Y. Guan. Risk-averse two-stage stochastic program with distributional ambiguity. *Operations Research*, 66(5):1390–1405, 2018.

[78] P. Kall, S. W. Wallace, and P. Kall. *Stochastic programming*. Springer, 1994.

[79] T. Kanamori, S. Fujiwara, and A. Takeda. Breakdown point of robust support vector machines. *Entropy*, 19(2):83–109, 2017.

[80] S. Katsumata and A. Takeda. Robust cost sensitive support vector machine. In *Artificial Intelligence and Statistics*, pages 434–443. PMLR, 2015.

[81] R. Khemchandani, S. Chandra, et al. Twin support vector machines for pattern classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):905–910, 2007.

[82] A. G. Kök, M. L. Fisher, and R. Vaidyanathan. Assortment planning: Review of literature and industry practice. In *Retail Supply Chain Management*, pages 99–153. Springer, 2008.

[83] D. Kuhn. *Generalized bounds for convex multistage stochastic programs*. Springer, 2005.

[84] D. Kuhn. Aggregation and discretization in multistage stochastic programming. *Mathematical Programming*, 113(1):61–94, 2008.

[85] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS, 2019.

[86] R. Kumari and S. K. Srivastava. Machine learning: A review on binary classification. *International Journal of Computer Applications*, 160(7):11–15, 2017.

[87] G. R. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. I. Jordan. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3:555–582, 2002.

[88] T. Le, D. Tran, W. Ma, T. Pham, P. Duong, and M. Nguyen. Robust support vector machine. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 4137–4144. IEEE, 2014.

[89] C. Lee and S. Mehrotra. A distributionally-robust approach for finding support vector machines. *Available from Optimization Online*, 2015.

[90] J. Lee, V. Gaur, S. Muthulingam, and G. F. Swisher. Stockout-based substitution and inventory planning in textbook retailing. *Manufacturing & Service Operations Management*, 18(1):104–121, 2016.

[91] Y.-J. Lee and O. L. Mangasarian. Ssvm: A smooth support vector machine for classification. *Computational Optimization and Applications*, 20(1):5–22, 2001.

[92] C.-N. Li, Y.-H. Shao, and N.-Y. Deng. Robust L1-norm non-parallel proximal support vector machine. *Optimization*, 65(1):169–183, 2016.

[93] J. Li, C. Chen, and A. M.-C. So. Fast epigraphical projection-based incremental algorithms for wasserstein distributionally robust support vector machine. *arXiv:2010.12865*, 2020.

[94] Z. Li and C. A. Floudas. Robust counterpart optimization: Uncertainty sets, formulations and probabilistic guarantees. In *Proceedings of the 6th Conference on Foundations of Computer-Aided Process Operations, Savannah (Georgia)*, 2012.

[95] Q. Liu and G. Van Ryzin. On the choice-based linear programming model for network revenue management. *Manufacturing & Service Operations Management*, 10(2):288–310, 2008.

[96] X. Liu and F. A. Potra. Pattern separation and prediction via linear and semidefinite programming. *Studies in Informatics and Control*, 18(1):71–82, 2009.

[97] Y. Liu, B. Zhang, B. Chen, and Y. Yang. Robust solutions to fuzzy one-class support vector machine. *Pattern Recognition Letters*, 71:73–77, 2016.

[98] R. Livni, K. Crammer, and A. Globerson. A simple geometric interpretation of svm using stochastic adversaries. In *Artificial Intelligence and Statistics*, pages 722–730. PMLR, 2012.

[99] J. López, S. Maldonado, and M. Carrasco. A robust formulation for twin multiclass support vector machine. *Applied Intelligence*, 47(4):1031–1043, 2017.

[100] W. Ma and M. A. Lejeune. A distributionally robust area under curve maximization model. *Operations Research Letters*, 48(4):460–466, 2020.

[101] Y. Ma and G. Guo. *Support vector machines applications*. Springer, 2014.

[102] F. Maggioni, E. Allevi, and M. Bertocchi. Bounds in multistage linear stochastic programming. *Journal of Optimization Theory and Applications*, 163(1):200–229, 2014.

[103] F. Maggioni, E. Allevi, and M. Bertocchi. Monotonic bounds in multistage mixed-integer stochastic programming. *Computational Management Science*, 13(3):423–457, 2016.

[104] F. Maggioni and G. C. Pflug. Bounds and approximations for multistage stochastic programs. *SIAM Journal on Optimization*, 26(1):831–855, 2016.

[105] F. Maggioni and G. C. Pflug. Guaranteed bounds for general nondiscrete multistage risk-averse stochastic optimization programs. *SIAM Journal on Optimization*, 29(1):454–483, 2019.

[106] A. İ. Mahmutoğulları, Ö. Çavuş, and M. S. Aktürk. Bounds on risk-averse mixed-integer multi-stage stochastic programming problems with mean-CVaR. *European Journal of Operational Research*, 266(2):595–608, 2018.

[107] A. İ. Mahmutoğulları, Ö. Çavuş, and M. S. Aktürk. An exact solution approach for risk-averse mixed-integer multi-stage stochastic programming problems. *Annals of Operations Research*, pages 1–22, 2019.

[108] S. Maldonado, J. López, and M. Carrasco. A second-order cone programming formulation for twin support vector machines. *Applied Intelligence*, 45(2):265–276, 2016.

[109] A. W. Marshall and I. Olkin. Multivariate Chebyshev inequalities. *The Annals of Mathematical Statistics*, 31(4):1001–1014, 1960.

[110] J. I. McGill and G. J. Van Ryzin. Revenue management: Research overview and prospects. *Transportation Science*, 33(2):233–256, 1999.

[111] M. M. Moya and D. R. Hush. Network constraints and multi-objective optimization for one-class classification. *Neural Networks*, 9(3):463–474, 1996.

[112] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, pages 1196–1204, 2013.

[113] C. Ning and F. You. Data-driven adaptive nested robust optimization: general modeling framework and efficient computational algorithm for decision making under uncertainty. *AIChE Journal*, 63(9):3790–3817, 2017.

[114] J. Nocedal and S. J. Wright. Line search methods. In *Numerical Optimization*, pages 30–65. Springer, 2006.

[115] N. Noyan, G. Rudolf, and M. Lejeune. Distributionally robust optimization with decision-dependent ambiguity set. *Available from Optimization Online*, 2018.

[116] R. Pant, T. B. Trafalis, and K. Barker. Support vector machine classification of uncertain and imbalanced data using robust optimization. In *Proceedings of the 15th WSEAS International Conference on Computers*, pages 369–374. WSEAS, 2011.

[117] J. Park and G. Bayraksan. A multistage distributionally robust optimization approach to water allocation under climate uncertainty. *arXiv:2005.07811*, 2020.

[118] M. Pellegrini, R. De Leone, P. Maponi, and M. Ferretti. Reducing power consumption in hydrometric level sensor networks using support vector machines. In *Pervasive and Embedded Computing and Communication Systems*, pages 229–232, 2013.

[119] M. Pellegrini, R. De Leone, P. Maponi, and C. Rossi. Adaptive sampling for embedded software systems using SVM: Application to water level sensors. In *Proceedings of the CTW 2012 11th Cologne-Twente Workshop on Graph and Combinatorial Optimization*, pages 100–103. COMTESSA, 2012.

[120] M. V. Pereira and L. M. Pinto. Multi-stage stochastic optimization applied to energy planning. *Mathematical Programming*, 52(1-3):359–375, 1991.

[121] G. C. Pflug and A. Pichler. *The problem of ambiguity in stochastic optimization*, pages 229–255. Springer International Publishing, 2014.

[122] A. B. Philpott, V. L. de Matos, and L. Kapelevich. Distributionally robust SDDP. *Computational Management Science*, 15(3-4):431–454, 2018.

[123] A. Pichler and A. Shapiro. Mathematical foundations of distributionally robust multistage optimization. *arXiv:2101.02498*, 2021.

[124] P. Pierre-Louis, D. Morton, and G. Bayraksan. A combined deterministic and sampling-based sequential bounding method for stochastic programming. In *Proceedings 2011 Winter Simulation Conference*, pages 4172–4183, 2011.

[125] J. Pinter. Deterministic approximations of probability inequalities. *Zeitschrift für Operations-Research*, 33(4):219–239, 1989.

[126] Z. Qi, Y. Tian, and Y. Shi. Robust twin support vector machine for pattern classification. *Pattern Recognition*, 46(1):305–316, 2013.

[127] H. Rahimian, G. Bayraksan, and T. Homem-de Mello. Identifying effective scenarios in distributionally robust stochastic programs with total variation distance. *Mathematical Programming*, 173(1-2):393–430, 2019.

[128] H. Rahimian and S. Mehrotra. Distributionally robust optimization: A review. *arXiv:1908.05659*, 2019.

[129] P. Ray. Independence of irrelevant alternatives. *Econometrica: Journal of the Econometric Society*, 41(5):987–991, 1973.

[130] R. T. Rockafellar. Coherent approaches to risk in optimization under uncertainty. In *OR Tools and Applications: Glimpses of Future Technologies*, pages 38–61. INFORMS, 2007.

[131] P. Rusmevichientong, Z.-J. M. Shen, and D. B. Shmoys. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations Research*, 58(6):1666–1680, 2010.

[132] P. Rusmevichientong, D. Shmoys, C. Tong, and H. Topaloglu. Assortment optimization under the multinomial logit model with random choice parameters. *Production and Operations Management*, 23(11):2023–2039, 2014.

[133] P. Rusmevichientong and H. Topaloglu. Robust assortment optimization in revenue management under the multinomial logit choice model. *Operations Research*, 60(4):865–882, 2012.

[134] A. Ruszczyński and A. Shapiro. Conditional risk mappings. *Mathematics of Operations Research*, 31(3):544–561, 2006.

[135] A. Ruszczyński and A. Shapiro. Optimization of convex risk functions. *Mathematics of Operations Research*, 31(3):433–452, 2006.

[136] S. R. Safavian and D. Landgrebe. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3):660–674, 1991.

[137] B. Sandıkçı, N. Kong, and A. J. Schaefer. A hierarchy of bounds for stochastic mixed-integer programs. *Mathematical Programming*, 138(1):253–272, 2013.

[138] B. Sandikçi and O. Y. Özaltin. A scalable bounding method for multistage stochastic programs. *SIAM Journal on Optimization*, 27(3):1772–1800, 2017.

[139] H. Scarf. A min-max solution of an inventory problem. In *Studies in the Mathematical Theory of Inventory and Production*. Stanford University Press, 1958.

[140] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 2000.

[141] U. Shaham, Y. Yamada, and S. Negahban. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing*, 307:195–204, 2018.

[142] C. Shang, X. Huang, and F. You. Data-driven robust optimization based on kernel learning. *Computers & Chemical Engineering*, 106:464–479, 2017.

[143] C. Shang and F. You. Distributionally robust optimization for planning and scheduling under uncertainty. *Computers & Chemical Engineering*, 110:53–68, 2018.

[144] A. Shapiro. Tutorial on risk neutral, distributionally robust and risk averse multistage stochastic programming. *European Journal of Operational Research*, 288:1–13, 2021.

[145] A. Shapiro and S. Ahmed. On a class of minimax stochastic programs. *SIAM Journal on Optimization*, 14(4):1237–1249, 2004.

[146] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on stochastic programming: Modeling and theory*. SIAM, 2021.

[147] A. Shapiro and A. Kleywegt. Minimax analysis of stochastic problems. *Optimization Methods and Software*, 17(3):523–542, 2002.

[148] A. Shapiro and L. Xin. Technical note - time inconsistency of optimal policies of distributionally robust inventory models. *Operations Research*, 68(5):1576–1584, 2020.

[149] K. Shen, Y. Ping, T. Sun, and Y. Zhou. Robust chance constrained optimization with Pearson divergence. In *DEStech Transactions on Engineering and Technology Research*, pages 122–125, 2020.

[150] P. K. Shivaswamy, C. Bhattacharyya, and A. J. Smola. Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research*, 7:1283–1314, 2006.

[151] S. Silvi, M. C. Verdenelli, C. Cecchini, M. M. Coman, M. S. Bernabei, J. Rosati, R. De Leone, C. Orpianesi, and A. Cresci. Probiotic-enriched foods and dietary supplement containing synbio positively affects bowel habits in healthy adults: An assessment using standard statistical analysis and support vector machines. *International Journal of Food Sciences & Nutrition*, 65(8):994–1002, 2014.

[152] M. Singla, D. Ghosh, and K. Shukla. A survey of robust optimization based machine learning with special reference to support vector machines. *International Journal of Machine Learning and Cybernetics*, 11(7):1359–1385, 2020.

[153] S. Sra, S. Nowozin, and S. J. Wright. *Optimization for machine learning*. Mit Press, 2012.

[154] M. Staib and S. Jegelka. Distributionally robust optimization and generalization in kernel methods. *Advances in Neural Information Processing Systems*, 32:9134–9144, 2019.

[155] J. M. Steele. *The Cauchy-Schwarz master class: An introduction to the art of mathematical inequalities*. Cambridge University Press, 2004.

[156] G. Stempfel and L. Ralaivola. Learning svms from sloppily labeled data. In *International Conference on Artificial Neural Networks*, pages 884–893. Springer, 2009.

[157] A. Takeda and T. Kanamori. A robust approach based on conditional value-at-risk measure to statistical learning problems. *European Journal of Operational Research*, 198(1):287–296, 2009.

[158] A. Takeda and M. Sugiyama. ν-support vector machine as conditional value-at-risk minimization. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1056–1063, 2008.

[159] K. Talluri and G. Van Ryzin. Revenue management under a general discrete choice model of consumer behavior. *Management Science*, 50(1):15–33, 2004.

[160] B. Taskesen, V. A. Nguyen, D. Kuhn, and J. Blanchet. A distributionally robust approach to fair classification. *arXiv:2007.09530*, 2020.

[161] T. B. Trafalis and S. A. Alwazzi. Support vector machine classification with noisy data: A second order cone programming approach. *International Journal of General Systems*, 39(7):757–781, 2010.

[162] T. B. Trafalis and R. C. Gilbert. Robust classification and regression using support vector machines. *European Journal of Operational Research*, 173(3):893–909, 2006.

[163] T. B. Trafalis and R. C. Gilbert. Robust support vector machines for classification and computational issues. *Optimization Methods and Software*, 22(1):187–198, 2007.

[164] P. Tsyurmasto, J. Gotoh, and S. Uryasev. Support vector classification with positive homogeneous risk functionals. Technical report, University of Florida, 2013.

[165] L. V. Utkin and A. I. Chekh. A new robust model of one-class classification by interval-valued training data using the triangular kernel. *Neural Networks*, 69:99–110, 2015.

[166] L. V. Utkin, A. I. Chekh, and Y. A. Zhuk. Binary classification svm-based algorithms with interval-valued training data using triangular and epanechnikov kernels. *Neural Networks*, 80:53–66, 2016.

[167] L. V. Utkin and Y. A. Zhuk. An one-class classification support vector machine model by interval-valued training data. *Knowledge-Based Systems*, 120:43–56, 2017.

[168] G. J. van Ryzin and K. T. Talluri. An introduction to revenue management. In *Emerging Theory, Methods, and Applications*, pages 142–194. INFORMS, 2005.

[169] V. Vapnik and A. Chervonenkis. *Theory of pattern recognition*. Nauka, Moscow, 1974.

[170] S. Vishwanathan and M. N. Murty. Ssvm: a simple svm algorithm. In *Proceedings of the 2002 International Joint Conference on Neural Networks*, volume 3, pages 2393–2398. IEEE, 2002.

[171] C. A. Vitt, D. Dentcheva, and H. Xiong. Risk-averse classification. *Annals of Operations Research*, pages 1–35, 2019.

[172] G. Vulcano, G. Van Ryzin, and W. Chaar. Om practice - choice-based revenue management: An empirical study of estimation and optimization. *Manufacturing & Service Operations Management*, 12(3):371–392, 2010.

[173] S. W. Wallace and S.-E. Fleten. Stochastic programming models in energy. In *Stochastic Programming, Handbooks in Operations Research and Management Science*, pages 637–677. Elsevier, 2003.

[174] X. Wang, N. Fan, and P. M. Pardalos. Stochastic subgradient descent method for large-scale robust chance-constrained support vector machines. *Optimization Letters*, 11(5):1013–1024, 2017.

[175] X. Wang, N. Fan, and P. M. Pardalos. Robust chance-constrained support vector machines with second-order moment information. *Annals of Operations Research*, 263(1):45–68, 2018.

[176] Y. Wang. Robust ν-support vector machine based on worst-case conditional value-at-risk minimization. *Optimization Methods and Software*, 27(6):1025–1038, 2012.

[177] Y. Wang, V. A. Nguyen, and G. A. Hanasusanto. Wasserstein robust support vector machines with fairness constraints. *arXiv:2103.06828*, 2021.

[178] W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

[179] Y. Wu and Y. Liu. Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 102(479):974–983, 2007.

[180] H. Xiao, B. Biggio, B. Nelson, H. Xiao, C. Eckert, and F. Roli. Support vector machines under adversarial label contamination. *Neurocomputing*, 160:53–62, 2015.

[181] H. Xu, C. Caramanis, and S. Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10:1485–1510, 2009.

[182] H. Xu, C. Caramanis, S. Mannor, and S. Yun. Risk sensitive robust support vector machines. In *Proceedings of the 48h IEEE Conference on Decision and Control held jointly with 2009 28th Chinese Control Conference*, pages 4655–4661. IEEE, 2009.

[183] L. Xu, K. Crammer, and D. Schuurmans. Robust support vector machine training via convex outlier ablation. In *Association for the Advancement of Artificial Intelligence*, volume 6, pages 536–542, 2006.

[184] X. Yu and S. Shen. Multistage distributionally robust mixed-integer programming with decision-dependent moment-based ambiguity sets. *Mathematical Programming*, pages 1–40, 2020.

[185] J. Yue, B. Chen, and M.-C. Wang. Expected value of distribution information for the newsvendor problem. *Operations Research*, 54(6):1128–1136, 2006.

[186] T. H. Yunes, D. Napolitano, A. Scheller-Wolf, and S. Tayur. Building efficient product portfolios at John Deere & Company. *Operations Research*, 55(4):615–629, 2007.

[187] J. Žáčková. On minimax solutions of stochastic linear programming problems. *Časopis pro Pěstování Matematiky*, 91(4):423–430, 1966.

[188] D. Zhang and D. Adelman. An approximate dynamic programming approach to network revenue management with customer choice. *Transportation Science*, 43(3):381–394, 2009.

[189] W. Zhang, H. Rahimian, and G. Bayraksan. Decomposition algorithms for risk-averse multistage stochastic programs with application to water allocation under uncertainty. *INFORMS Journal on Computing*, 28(3):385–404, 2016.

[190] L. Zhou, L. Wang, L. Liu, P. Ogunbona, and D. Shen. Support vector machines for neuroimage analysis: Interpretation from discrimination. In *Support Vector Machines Applications*, pages 191–220. Springer, 2014.

[191] Y. Zhou, M. Kantarcioglu, B. Thuraisingham, and B. Xi. Adversarial support vector machine learning. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1059–1067, 2012.

[192] S. Zymler, D. Kuhn, and B. Rustem. Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, 137(1):167–198, 2013.