Theory and Algorithms of Robust Chance Constraints

by

Haoming Shen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in the University of Michigan
2023

Doctoral Committee:

Professor Ruiwei Jiang, Chair
Professor Jon Lee
Professor Siqian Shen
Professor Cong Shi
Professor Yafeng Yin

Haoming Shen
hmshen@umich.edu
ORCID iD: 0009-0003-7965-3302

© Haoming Shen 2023

DEDICATION

This thesis is dedicated to my parents, whose infinite love, unwavering support and constant encouragement have sustained me throughout my educational pursuits.

ACKNOWLEDGMENTS

I would like to express my foremost gratitude to my advisor Dr. Ruiwei Jiang, who, as my guide and friend in the past five years, is always able to provide me unwavering support and constructive feedback, for which I am forever indebted. Above all, I have been greatly inspired by and benefited from his tireless dedication to scientific inquiry, his astute understanding of fundamental concepts, and his resourceful and pragmatic problem-solving approaches.

My appreciation also goes to my dissertation committee members. I would like to express my sincere appreciation to Dr. Siqian Shen for introducing me to the field of operations research and for her continuous support throughout my academic journey. I am also grateful to Dr. Jon Lee, Dr. Cong Shi, and Dr. Yafeng Yin for their valuable feedback and insightful suggestions in my preliminary exam and final defense. Their expertise and guidance have significantly contributed to the quality and rigor of this thesis.

I am fortunate to have found a welcoming home in the Department of Industrial and Operations Engineering at UM. I am grateful for the opportunity to have learned from my exceptional lab mates, Xubo Yue, Seokhyun Chung, Rohan Ghuge, and graduate student colleagues, Yiling Zhang, Huiwen Jia, Xian Yu, Luze Xu, and many others. I would also like to express my sincere gratitude to my friends outside the department, including Qing Jin, Nan Jiang, Ransheng Guan, Yongzhao Wang. Their friendship, support, and camaraderie have made this academic journey both rewarding and enjoyable.

Finally, I would like to express my heartfelt gratitude to my parents, who have always encouraged and supported me to pursue my own life. Their unconditional support, guidance, and sacrifices have been the bedrock of my academic and personal accomplishments, and I will always be grateful to them.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGMENTS iii

LIST OF FIGURES vii

LIST OF TABLES viii

LIST OF APPENDICES ix

ABSTRACT x

CHAPTER

1 Introduction **1**

2 Convex Wasserstein Robust Chance Constraints **5**
2.1 Problem Description and Literature Review 5
2.1.1 Key Definitions and Examples 9
2.2 Pessimistic Chance Constraint 11
2.2.1 Joint (**P-CC**) with RHS Uncertainty 15
2.2.2 Two-Sided (**P-CC**) with LHS Uncertainty 18
2.3 Solution Approaches for Pessimistic Chance Constraint 24
2.3.1 Block Coordinate Ascent Algorithm 24
2.3.2 SOC Inner Approximation 28
2.3.2.1 Inner Approximations for $\mathcal{C}_\delta$ and $\mathcal{X}_{\mathsf{T}}^p$ 28
2.3.2.2 Approximation Error and Asymptotic Exactness of $\widehat{\mathcal{X}}_{\mathsf{T}}^p$ . . . 30
2.4 Optimistic Chance Constraint 32
2.4.1 Joint (**O-CC**) with RHS Uncertainty 32
2.4.2 Two-Sided (**O-CC**) with LHS Uncertainty 34
2.5 Numerical Experiments 38
2.5.1 Production Planning 39
2.5.1.1 Out-of-Sample Performance 41
2.5.2 Robust Linear Regression 42
2.5.2.1 Out-of-Sample Performance 42

2.5.2.2 Scalability of the Inner Approximation Approach 43
2.6 Conclusion 44

3 Wasserstein Robust Chance-Constrained Set Covering **45**
3.1 Problem Description and Literature Review 45
3.1.1 Literature Review 47
3.1.2 Contributions 49
3.2 Two-Stage Reformulation 50
3.2.1 Conditional Value-at-Risk Reformulation 50
3.2.2 NP-hardness 52
3.2.3 Two-Stage Reformulation 54
3.2.3.1 Polyhedral Results for Submodular Functions 54
3.2.3.2 Reformulation 55
3.3 Valid Inequalities 59
3.3.1 Valid inequalities from a single scenario 59
3.3.2 Valid inequalities from cross scenarios 64
3.4 Extension to Knapsack Chance Constraint 68
3.5 Numerical Experiments 69
3.5.1 Independent uncertain parameters 69
3.5.1.1 Impact of the radius δ on out-of-sample performance 71
3.5.1.2 Impact of the data size N on out-of-sample performance . . 71
3.5.2 Dependent uncertain parameters 73
3.5.3 Value of Incorporating Support Information 75
3.5.4 Strength of the two-stage reformulation and valid inequalities 78
3.6 Conclusion 79

4 Locating Charging Stations for Battery Electric Buses **83**
4.1 Problem Description and Literature Review 83
4.2 Distributionally Robust Chance Constraint Model 86
4.2.1 Deterministic model 86
4.2.2 DRCC model 88
4.3 Solution Methods 95
4.3.1 Delayed constraint generation 96
4.3.1.1 Low-rank approximation of SOC constraints 99
4.3.2 Dual decomposition 100
4.3.3 Valid inequalities 103
4.4 Numerical Experiments 105
4.4.1 The selection of radius ρ 106
4.4.2 Benchmark between **(DET)**, (**CC**) and (**DRCC**) 107
4.4.3 Impact of radius ρ and sample size N on reliability 107
4.4.4 Comparision of optimal solutions of **(DET)**, (**DRCC**) and (**CC**) 108
4.4.5 Computational performance 111
4.5 Conclusion 114

5 Conclusions **115**

References **116**

APPENDICES 130

A.1 Preliminary Results 130
A.2 Proof of Lemma 1 135
A.3 Proof of Lemma 2 136
A.4 A Generalized Theorem 1 For Quasi-Concave Inequalities 136
A.5 Proof of Lemma 3 138
A.6 Proof of Lemma 4 138
A.7 Proof of Example 5 139
A.8 Proof of A Fact Connecting $\mathcal{X}_{\mathsf{T}}^{p}$ and $\mathcal{C}_{\delta}$ 141
A.9 Proof of Corollary 2 142
A.10 Proof of Proposition 4 143
A.11 Proofs of Preparatory Lemmas 12, 13, and 14 144
A.12 Proof of Theorem 4 149
A.13 Proof of Theorem 5 157
A.14 Proof of Lemma 5 158
A.15 A Generalized Theorem 6 For Quasi-Concave Inequalities 159
B.1 Proof of Proposition 8 160
B.2 Proof of Lemma 9 162
B.3 Proof of Theorem 9 163
B.3.1 Proofs of Propositions 34, 35, and 36 168
B.4 Proof of Theorem 11 177
C.1 Nomenclature of Chapter 4 180

LIST OF FIGURES

FIGURE

1.1 Bad Out-of-Sample Performance of $\widehat{\mathbb{P}}_N$ 2

2.1 Contours of $g_\epsilon(\ell, u)$ with varying δ and a polyhedral inner approximation $\widehat{\mathcal{C}}_N$ of $\mathcal{C}_\delta$ with $N = 5$ and $\delta = 0.050$ 19
2.2 Visualization of random variables ζ_1 and ζ_2 in Examples 6–7 22
2.3 Visualization of γ_τ and $(\bar{g}_\epsilon)^{-1}(s)$ 31
2.4 Convergence of Algorithm 1 on Production Planning Instances; solid line = average of the difference $\phi(x_k, y_k) - \phi^*$ across five runs, error bar = standard deviation of the difference 40
2.5 Risk envelopes under different risk thresholds 40
2.6 Out-of-sample performance of (**P-CC**) and (**CC**) on the production planning problem 41
2.7 Out-of-sample performance of (**P-CC**) and (**CC**) on robust linear regression 43
2.8 Runtime (in seconds) under different problem and training data sizes 44

3.1 A numerical example of the shifted function $f((z - \beta)^+)$ and $h_{(\ell,\rho)}(z)$. 62
3.2 Out-Of-Sample Performance of (DRC) and (SAA) as a function of δ. 71
3.3 Out-of-sample Performance of (DRC) and (SAA) as a function of N. 72
3.4 Network of a medical facility location case study. 72
3.5 Out-of-sample performance of (DRC) as a function of N. 74
3.6 Out-of-sample performance of (DRC) as a function of δ. 74

4.1 A toy BEB system with three stations. 93
4.2 Eigenvalues of the matrix $\boldsymbol{H}_{kmn}\Sigma\boldsymbol{H}_{kmn}^{\mathsf{T}}$ based on the SLC data 98
4.3 An overview of SLC transportation system 106
4.4 Benchmark between (**CC**) and (**DRCC**) over different training sizes N_{train}, $\alpha = 0.1$ 108
4.5 Reliability of (**DRC**) and (CC) as a function of δ. 109
4.6 Reliability of (**DRCC**) and (**CC**) as a function of N_{train}. 110
4.7 Comparision between optimal solutions of (**CC**) and (**DRCC**) 112

LIST OF TABLES

TABLE

3.1 Comparison between optimizing over Z_0 and Z 77
3.2 Benchmark between different models on synthetic data ($N = 50$) 80
3.3 Benchmark between different models on synthetic data ($N = 100$) 81
3.4 Benchmark between different models on synthetic data ($N = 200$) 82

4.1 Examples of ϕ-divergence 92
4.2 SLC data parameters 105
4.3 Optimal solutions of different models 111
4.4 Computational benchmark between three formulations 113

LIST OF APPENDICES

A Convex Wasserstein Robust Chance Constraints **130**

B Wasserstein Robust Chance-Constrained Set Covering **160**

C Locating Charging Stations for Battery Electric Buses **180**

ABSTRACT

With limited knowledge or data, Robust Chance Constraints (RCCs), also known as Distributionally Robust Chance Constraints in the literature, provide a natural and interpretable way to obtain probabilistic safety guarantees for a variety of complex uncertain systems, examples of which include but not limited to power grids, transportation systems, scheduling systems, and financial markets. Despite its good out-of-sample performance and wide range of applications, RCCs usually produce non-convex feasible regions in general, giving rise to computational intractability. To alleviate such challenge, this dissertation aims to answer two main questions: (1) When are RCCs convex and/or tractable? (2) How can we solve RCCs efficiently? Firstly, we uncover a set of sufficient conditions under which RCCs produce convex feasible regions and we accordingly design efficient algorithms for it, making it possible to solve ranges of important problems arising in motion planning, electric vehicle operations, and optimal power flow. Secondly, we consider a class of RCCs that contains a set covering structure, which arise frequently in facility location, production planning, and scheduling. However, such covering RCCs are NP-hard to solve in general and do not satisfy our proposed sufficient conditions. Therefore we improve the computational performance by constructing strong valid inequalities for a mixed integer set appearing in the reformulation of RCCs and demonstrate its effectiveness using a emergency facility location problem. In addition, we also apply our results on RCCs to locate the charging stations for battery electric buses in a transit system, and propose two decomposition algorithms, including delayed constraint generation and dual decomposition, which demonstrate significant speed-ups as compared to commercial solvers. A real-world case study based on the Salt Lake City bus transit system demonstrates the effectiveness of our RCC model and the solution efficacy of the decomposition algorithms.

CHAPTER 1

Introduction

Uncertainties, inherent to many real world systems, often pose great challenges to decision-making processes. For example, rare and extreme cold temperatures greatly challenged Texas's state energy supply in February 2021, uncertain safety-critical events in high-dimensional driving environment impede the development of autonomous vehicle, and the sudden outbreak of COVID-19 has caused unprecedented and widespread disruption to people's daily lives. An intuitive and powerful approach to deal with the underlying uncertainty is chance (also called probabilistic) constraints (CCs), which explicitly require all feasible decisions to satisfy safety constraints with a high probability, thereby providing a probabilistic guarantee to uncertain systems. More precisely, CCs have the following generic form:

$$\mathbb{P}_{\text{true}}\Big[\ \mathtt{sys}(\widetilde{\xi}, x) \text{ is safe}\ \Big] \geq (1-\epsilon),$$

where $\widetilde{\xi}$ is a random vector representing the intrinsic uncertainty in the system, x denotes the decisions, $\mathtt{sys}(\widetilde{\xi}, x)$ is an uncertain system that is affected by $\widetilde{\xi}$ and controlled by x, $\mathbb{P}_{\text{true}}$ denotes the unknown *true* distribution of $\widetilde{\xi}$, $(1-\epsilon)$ is a risk level chosen by the decision maker (DM), and the safety constraints are usually described by a system of inequalities.

Despite its natural interpretations and wide range of applications in, e.g., power system, vehicle routing, scheduling, portfolio management, and facility location, the ambiguous probability distribution of the uncertainty, $\mathbb{P}_{\text{true}}$, greatly hinders the practical applications of CC. In many cases, DMs do not have the true knowledge about the uncertainty involved, and usually they only have a series of historical data, making it impractical to obtain an accurate description of the underlying uncertainty. In such situations, one may use the empirical distribution generated from the historical samples instead and obtain "optimal" decisions by optimizing the associated CCs accordingly. However, uncertain parameters in real-world are often high-dimensional in nature and CCs defined using crude approximations (e.g., empirical distributions) of the uncertainty may not produce decisions that could achieve the required probabilistic safety guarantee under the unknown *true*

distribution $\mathbb{P}_{\text{true}}$. To provide a numerical illustration, let us consider chance-constrained optimal power flow (CC-OPF) problems in power systems with significant renewable penetration. In such problems, the system operator needs to determine the optimal power generation plan of all thermal units so that, together with the renewable generators, they could satisfy electricity demands and system safety constraints with a high probability. In the following numerical example, we assume that the unknown true distribution $\mathbb{P}_{\text{true}}$ of renewable generation follows a Weibull distribution, and through CCs, we enforce feasible power generation plans to satisfy all safety constraints and electricity demands with probability at least 95%. In order to simulate real-world situations, we first create synthetic datasets and generate the associated empirical distributions $\widehat{\mathbb{P}}_N$ by collecting $N \in \{5, 10, 100, 200, 500\}$ indepedent and identically distributed (i.i.d.) samples of $\mathbb{P}_{\text{true}}$. Then, for each problem instance we optimize over the CC defined using $\widehat{\mathbb{P}}_N$, find the optimal solution, evaluate its out-of-sample (OOS) performance, that is, the probability of safety constraints being satisfied under the unknown *true* distribution $\mathbb{P}_{\text{true}}$, i.e. Weibull, distribution, and check whether it is above the require risk level of 95%. For each sample

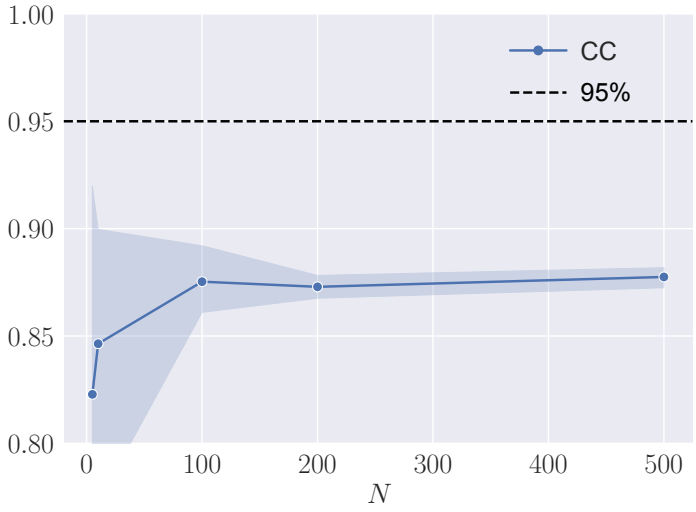


Figure 1.1: Bad Out-of-Sample Performance of $\widehat{\mathbb{P}}_N$

size N, Figure 1.1 shows the average (resp. the 90% confidence interval) of OOS over 5 random instances using the blue solid lines (resp. the blue shaded region), from which we observe that the OOS performance gradually increases and stabilizes as N becomes larger, but it consistently fails to reach the required risk level of 95% (the dashed black line) even when 500 samples are provided. This demonstrates that the solution obtained from $\widehat{\mathbb{P}}_N$ may not be reliable, i.e. it may not satisfy the ambiguous CC defined using the

unknown $\mathbb{P}_{\text{true}}$, in which case, putting the uncertain system at risk.

A popular and powerful framework for dealing with the ambiguity in $\mathbb{P}_{\text{true}}$ is Distributionally Robust Optimization, (DRO), which acknowledges the inherent distributional ambiguity of the uncertainty and attempts to capture that using an ambiguity set. This ambiguity set represents a family of plausible probability distributions that are close to, or centered around, a given reference distribution, which is usually the empirical one. By seeking optimal solutions that hedge against all distributions in the ambiguity set, DRO is able to help DMs make more informed and reliable decisions in the face of uncertainty. Applying this idea to CC, we obtain robust chance constraints (RCC), also known as distributionally robust chance constraints in the literature, where we now enforce CC for *all* distributions in the ambiguity set, rather than solely relying on the single CC associated with $\widehat{\mathbb{P}}_N$. If the ambiguity set is properly chosen, in the sense that it contains $\mathbb{P}_{\text{true}}$ with a high probability, then the resulting decision will be able to satisfy the ambiguous CC defined using $\mathbb{P}_{\text{true}}$ with high probability, thereby enhancing the OOS performance significantly. As a summary, RCC provides an interpretable and widely-applicable way to enforce safety constraints for uncertain systems, and compared to CC, it alleviates the ambiguity in the unknown true distribution and generalizes better when previously unseen samples arise.

Although RCCs, as a special case of DRO formulations, are known to have good out-of-sample performance, they are generally non-convex and, as a result, not scalable. This is due to the fact that CC, which is notorious difficult to optimize over and compute, is actually a special case of RCC when the ambiguity set is a singleton. Therefore, to address such computational burden, this dissertation aims to answer the following two questions: (1) When are RCCs convex and/or tractable? (2) How can we solve RCCs efficiently?

To answer the first question, in Chapter 2 we identify sufficient conditions under which RCCs have *convex* feasible regions. Specifically, we focus on two commonly encountered structures in practice: (a) *joint* chance constraints (JCC), where the "safe" region defined in the space of decisions and uncertainty is a closed convex set; (b) *two-sided* chance constraints, where the "safe" region in the space of uncertainty is a strip region parametrized by decisions. For JCC in (a), we show that the the ensuing feasible region is convex if the ambiguity set is centered around a log-concave distribution and decisions and the uncertainty are decoupled. Furthermore, we design a block coordinate ascent algorithm for solving robust JCC and prove its convergence to global optimum. In addition, for two-sided CC in (b), we show its convexity if the ambiguity set is centered around an elliptical and star-unimodal distribution. Moreover, we propose a family of second-order conic inner approximations, bound their approximation error, and prove their asymptotic exactness.

Motivated by emergency facility location problems, in Chapter 3 we study RCCs with a special set covering structure where both decisions and uncertain parameters are binary-valued. However, due to the inherent binary nature in the uncertainty, the sufficient conditions for convexity we identified in Chapter 2 do not apply, and we show that RCC with such covering structures are NP-hard to solve in general. Therefore, we aim to improve the computational efficiency by exploiting the special covering structure and designing better solution approaches. To this end, we recast the RCC as a two-stage stochastic program, which facilitates decomposition algorithms. Furthermore, we derive two families of valid inequalities. The first family targets the hypograph of a "shifted" submodular function, which is associated with each scenario of the two-stage reformulation. We show that the valid inequalities give a complete description of the convex hull of the hypograph. The second family mixes inequalities across multiple scenarios and gains further strength via lifting. Our numerical experiments on emergency medical facility location problems and randomly generated test instances demonstrate the out-of-sample performance of the RCC model and the effectiveness of our proposed reformulation and valid inequalities.

In Chapter 4, we apply RCC models studied in previous chapters to a transit system design problem for battery electric buses (BEBs). This is inspired by the increasing trend of emission reduction and fuel economy, among which transit agencies are replacing traditional diesel-powered buses with BEBs. Nontheless, the limited driving range, long charging time, as well as the uncertain traffic conditions and passenger demands pose challenges to the design of a BEB transit system. To account for such uncertainty, we propose a data-driven optimization approach to locate fast, enroute BEB charging stations in a transit system, with an objective of protecting BEB from running out of power when the random battery energy consumption and the available charging time follow an ambiguous probability distribution. In particular, we consider a special case of two-sided RCC called *individual* RCC, whose safety region in the space of decisions and the uncertainty is a half-space, and we reformulate it as a tractable second-order conic constraint. Additionally, we study two decomposition algorithms, including delayed constraint generation and dual decomposition, which demonstrate significant speed-ups as compared to commercial solvers. Finally, we conduct a real-world case study based on the Salt Lake City bus transit system to demonstrate the effectiveness of our model and the solution efficacy of the decomposition algorithms.

CHAPTER 2

Convex Wasserstein Robust Chance Constraints

2.1 Problem Description and Literature Review

Many optimization models include safety principles taking the form

$$A(x)\xi \leq b(x),$$

where $x \in \mathbb{R}^n$ represents decision variables, $\xi \in \mathbb{R}^m$ represents model parameters, and $A(x) \in \mathbb{R}^{m\times n}$ and $b(x) \in \mathbb{R}^m$ are affine functions of x. When ξ is subject to uncertainty and follows a probability distribution $\mathbb{P}_{\text{true}}$, a convenient way of protecting these safety principles is to use chance constraint

$$\mathbb{P}_{\text{true}}\Big[A(x)\xi \leq b(x)\Big] \geq 1-\epsilon, \tag{CC}$$

where $1-\epsilon \in (0,1)$ represents a pre-specified risk threshold. (**CC**) requires to satisfy all safety principles with high probability (i.e., $1-\epsilon$ is usually close to one, e.g., 0.95). (**CC**) was first studied in the 1950s [CC59; CCS58; MW65; Pré70] and finds a wide range of applications in, e.g., power system [WGW11], vehicle routing [SG83], scheduling [DS16], portfolio management [Li95], and facility location [MG06]. We mention two examples.

Example 1. (Production Planning) Suppose that we produce certain commodity at n facilities to serve m demand locations. If x_j denotes the production capacity of facility j and T_{ij} denotes the service coverage of facility j for location i (i.e., $T_{ij}=1$ if facility j can serve location i and $T_{ij}=0$ otherwise) for all $i \in [m]$ and $j \in [n]$, then chance constraint

$$\mathbb{P}_{\text{true}}\big[Tx \geq \xi\big] \geq 1-\epsilon \tag{PP}$$

assures that the production capacities are able to satisfy the demands ξ at all locations. Here, $A(x)$ in (**CC**) equals the $m \times m$ identify matrix and $b(x)$ equals Tx.

Example 2. (Linear Regression) Suppose that we are given a set of historical data from a pair of predictor-response variables $(X, y) \in \mathbb{R}^{n+1}$ and seek to find $(\beta, \beta_0) \in \mathbb{R}^{n+1}$ such that the affine relationship $y = X^\mathsf{T}\beta + \beta_0$ best explains the data. Then, the chance constraint

$$\mathbb{P}_{\text{true}}\big[-\gamma \leq y - X^\mathsf{T}\beta - \beta_0 \leq \gamma\big] \geq 1 - \epsilon \tag{\textbf{LR}}$$

assures that the residual of the regression is at most γ with high probability [Moh21]. Here, the random variables are $\xi = (X, y)$ and the decision variables are $x = (\beta_0, \beta, \gamma)$. Accordingly, $A(x) = [-\beta^\mathsf{T}, 1;\ \beta^\mathsf{T}, -1]$ and $b(x) = [\gamma + \beta_0;\ \gamma - \beta_0]$.

In (**PP**), the random vector ξ is decoupled from the decision variables x because, in this example, $A(x)$ is independent of x. For such chance constraints with $A(x) \equiv A$, we follow the convention in the literature and refer to them as chance constraints with right-hand side (RHS) uncertainty. In contrast, ξ and x are multiplied in (**LR**). To distinguish chance constraints in this form from those with RHS uncertainty, we call them chance constraints with left-hand side (LHS) uncertainty. In addition, we say a chance constraint is *individual* if $m = 1$, *two-sided* if $m = 2$ and the two rows of $A(x)$ are opposite (such as in (**LR**)), and *joint* if $m \geq 2$ (such as in (**PP**)).

Although (**CC**) provides an intuitive way to model uncertainty in safety principles, it produces a non-convex feasible region in general, giving rise to concerns of challenging computation. To this end, a stream of prior work proposed effective mixed-integer programming (MIP) approaches based on the notions of, e.g., sample average approximation [LA08; LAN08] and p-efficient points [Pré90; BR02], and derived valid inequalities to strengthen the ensuing MIP formulations (see, e.g., [Küç12; Lue14] and a recent survey [KJ21]). Another stream of prior work identified sufficient conditions for (**CC**) to produce a convex feasible region. For individual (**CC**), [PP63] derived a second-order conic (SOC) representation when ξ follows a Gaussian distribution, and [LLS01] and [CE06] further extended this result when ξ follows an elliptical log-concave distribution (see Definition 2). Two-sided (**CC**) was first studied by [LBV15], who assumed a Gaussian $\mathbb{P}_{\text{true}}$ and proved the convexity of the ensuing feasible region. Later, [Fat+21] generalized the study to a Gaussian mixture model. For joint (**CC**) with RHS uncertainty, [Pré13] (see his Theorem 10.2) proved the convexity of the ensuing feasible region when ξ follows a log-concave distribution, examples of which include Gaussian, exponential, beta (if both shape parameters are at least 1), uniform on convex support, etc. Furthermore, [SDR09] generalized this result to α-concave distributions (see Definition 2).

In most practical problems, the (true) distribution $\mathbb{P}_{\text{true}}$ of the random parameters ξ is unknown or ambiguous to the modeler, who often replaces $\mathbb{P}_{\text{true}}$ in (**CC**) with a crude

estimate, denoted by $\mathbb{P}$. Candidates of $\mathbb{P}$ includes the empirical distribution based on past observations of ξ and Gaussian distribution, whose mean and covariance matrix can be estimated based on these past observations. Since $\mathbb{P}$ may not perfectly model the uncertainty of ξ, it is reasonable to take into account its neighborhood, or more formally, an ambiguity set $\mathcal{P}$ around $\mathbb{P}$. In this chapter, we adopt a Wasserstein ambiguity set defined as

$$\mathcal{P} := \left\{ \mathbb{Q} \in \mathcal{P}_0 \colon d_W(\mathbb{Q}, \mathbb{P}) \leq \delta \right\},$$

where $\mathcal{P}_0$ is the set of all probability distributions, $\delta > 0$ is a pre-specified radius of $\mathcal{P}$, and $d_W(\cdot, \cdot)$ represents the Wasserstein distance (see, e.g., [MK18a; Kuh+19]). Specifically, the Wasserstein distance between two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ is defined through

$$d_W(\mathbb{P}_1, \mathbb{P}_2) := \inf_{\mathbb{P}_0 \sim (\mathbb{P}_1, \mathbb{P}_2)} \mathbb{E}_{\mathbb{P}_0} \Big[\|X_1 - X_2\| \Big], \tag{2.1}$$

where X_1, X_2 are two random variables following distributions $\mathbb{P}_1$, $\mathbb{P}_2$ respectively, $\mathbb{P}_0$ is the coupling of $\mathbb{P}_1$ and $\mathbb{P}_2$, and $\|\cdot\|$ represents a norm. $d_W(\mathbb{P}_1, \mathbb{P}_2)$ can be interpreted as the minimum cost, with respect to $\|\cdot\|$, of transporting the probability masses of $\mathbb{P}_1$ to recover $\mathbb{P}_2$. Hence, the Wasserstein ambiguity set $\mathcal{P}$ is a ball (in the space of probability distributions) centered around $\mathbb{P}$, which for this reason is referred to as the reference distribution. Additionally, $\mathcal{P}$ may include the true distribution $\mathbb{P}_{\text{true}}$, i.e., $\mathbb{P}_{\text{true}} \in \mathcal{P}$, when the radius δ is large enough. As a result, the pessimistic counterpart

$$\inf_{\mathbb{Q} \in \mathcal{P}} \mathbb{Q} \left[A(x)\xi \leq b(x) \right] \geq 1 - \epsilon \tag{**P-CC**}$$

implies (**CC**) because it requires that (**CC**) holds with respect to all distributions in $\mathcal{P}$. In contrast, an optimistic modeler may be satisfied as long as there exists some distribution in $\mathcal{P}$, with respect to which (**CC**) holds. This gives rise to the following optimistic counterpart of (**CC**):

$$\sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{Q} \left[A(x)\xi \leq b(x) \right] \geq 1 - \epsilon. \tag{**O-CC**}$$

(**O-CC**) finds applications in portfolio management to quantify the profit opportunities in stock market [SZ21]. In addition, when $\mathbb{P}_{\text{true}}$ is ambiguous, it becomes impossible to solve an optimization model with (**CC**) directly. Nevertheless, replacing (**CC**) with (**O-CC**) and (**P-CC**), respectively, produces a confidence interval for the (unknown) optimal value. Besides, since (**O-CC**) provides a relaxation of (**CC**), any valid inequality for (**O-CC**)

remains valid for computing (**CC**).

In the existing literature, (**P-CC**) is also known as distributionally robust chance constraint and, depending on the value of m and the ambiguity set $\mathcal{P}$, the feasible region of (**P-CC**) may be convex or non-convex. For individual (**P-CC**) (i.e., $m = 1$), convex representations have been derived when $\mathcal{P}$ is Chebyshev, i.e., when $\mathcal{P}$ consists of all distributions sharing the same mean and covariance matrix of ξ. Specifically, [EOO03; CE06] derived semidefinite and SOC representations of (**P-CC**) with a Chebyshev $\mathcal{P}$. With the same ambiguity set, [ZKR11] showed that (**P-CC**) is equivalent to its approximation based on conditional Value-at-Risk (CVaR) even when the safety principle becomes nonlinear in ξ. Additionally, [Han+15] and [LJM19] incorporated structural information (e.g., unimodality) into the Chebyshev $\mathcal{P}$ and derived semidefinite and SOC representations of (**P-CC**), respectively. For joint (**P-CC**) (i.e., $m \geq 2$), however, convexity results become scarce. [Han+17] characterized $\mathcal{P}$ by a conic support, the mean, and a positively homogeneous dispersion measure of ξ, and showed that (**P-CC**) with RHS uncertainty is conic representable. In addition, they showed that this result falls apart if one relaxes these conditions even in a mildest possible manner. More recently, [XA16] extended the convexity result when the safety principles depend on ξ nonlinearly and $\mathcal{P}$ is characterized by a single moment constraint of ξ. In this chapter, we study (**P-CC**) and (**O-CC**) with $\mathcal{P}$ being a Wasserstein ambiguity set.

To the best of our knowledge, the convexity results for either (**P-CC**) or (**O-CC**) with Wasserstein ambiguity do not exist in the existing literature to date. This is not surprising because [XA20] showed that it is strongly NP-hard to optimize over the feasible region of (**P-CC**), if $\mathcal{P}$ is centered around an empirical distribution of ξ. In addition, for the same setting [Xie19; CKW18; JL20] derived mixed-integer conic representations for (**P-CC**), implying a non-convex feasible region. This chapter seeks to revise the choice of the reference distribution $\mathbb{P}$, with regard to which (**P-CC**) and (**O-CC**) with Wasserstein ambiguity produce convex feasible regions. Our main results include:

1. For joint (**P-CC**) with RHS uncertainty, we prove that the ensuing feasible region is convex if the reference distribution $\mathbb{P}$ is log-concave. More generally, this result holds when $\mathbb{P}$ is α-concave with $\alpha \geq -1$. Furthermore, we derive a block coordinate ascent algorithm for optimization models involving (**P-CC**) and prove its convergence to global optimum.

2. For two-sided (**P-CC**) with LHS uncertainty, we prove its convexity when the reference distribution $\mathbb{P}$ is elliptical and star-unimodal. These conditions are tight in the sense that dropping either of them fails the convexity. Furthermore, we derive a

family of second-order conic inner approximations for two-sided (**P-CC**), bound their approximation error, and prove their asymptotic exactness.

3. We extend the aforementioned convexity results for joint (**P-CC**) with RHS uncertainty and two-sided (**P-CC**) with LHS uncertainty to their optimistic counterparts (**O-CC**).

In addition, we summarize the main convexity results in the following table.

	(**P-CC**)	(**O-CC**)
LHS Uncertainty	Theorem 2	Theorem 7
RHS Uncertainty	Theorem 1	Theorem 6

The remainder of this chapter is organized as follows. Section 2.1.1 reviews key definitions. Sections 2.2 and 2.3 study convexity and solution approaches for (**P-CC**), respectively. Section 2.4 extends the convexity results to (**O-CC**). Section 2.5 demonstrates (**P-CC**) and (**O-CC**) through two numerical experiments, and Section 2.6 concludes this chapter.

Notation: We use $\mathcal{X}^p$ and $\mathcal{X}^o$ to denote the feasible region of (**P-CC**) and (**O-CC**), respectively. We denote the n-dimensional extended real system by $\overline{\mathbb{R}}^n$. For a given decision x, we denote by $\mathcal{S}(x)$ the event $\{\, \xi \colon A(x)\xi \leq b(x) \,\}$ and by $\mathcal{S}^c(x)$ its complement. For $a, b \in \mathbb{R}$, $(a)^+ := \max\{a, 0\}$, $(a)^- := \min\{a, 0\}$, and $a \vee b := \max\{a, b\}$. For a norm $\|\cdot\|$, $\|\cdot\|_*$ denotes its dual norm. $\|\cdot\|_2$ represents the 2-norm, i.e., for $a \in \mathbb{R}^n$, $\|a\|_2 = \sqrt{\sum_{i=1}^n a_i^2}$. I_n denotes the $n \times n$ identity matrix, $\mathbf{Leb}(\cdot)$ denotes the Lebesgue measure, and the indicator $\mathbb{1}\{\, x \in \Omega \,\}$ equals one if $x \in \Omega$ and zero if $x \notin \Omega$. For two random variables X_1 and X_2, $X_1 \overset{d}{=} X_2$ means that X_1 and X_2 are identically distributed.

2.1.1 Key Definitions and Examples

We review definitions frequently used in subsequent discussions.

Definitions 1. A nonnegative function f defined on a convex subset of $\mathbb{R}^n$ is said to be α-concave with $\alpha \in \overline{\mathbb{R}}$ if for all $x, y \in$ **dom** f and $\theta \in [0, 1]$

$$f(\theta x + (1-\theta)y) \geq m_\alpha(f(x), f(y); \theta),$$

where $m_\alpha \colon \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ is defined as

$$m_\alpha(a, b; \theta) := 0 \quad \text{if } ab = 0,$$

and if $a > 0, b > 0, \theta \in [0, 1]$, then

$$m_\alpha(a, b; \theta) := \begin{cases} a^\theta b^{(1-\theta)} & \text{if } \alpha = 0, \\ \max\{a, b\} & \text{if } \alpha = +\infty, \\ \min\{a, b\} & \text{if } \alpha = -\infty, \\ (\theta a^\alpha + (1-\theta) b^\alpha)^{1/\alpha} & \text{otherwise.} \end{cases}$$

When $\alpha = 0$ or $\alpha = -\infty$, we say f is log-concave or quasi-concave, respectively.

The Minkowski sum of two Borel measurable subsets $A, B \subset \mathbb{R}^n$ is Borel measurable. Let $\theta \in [0, 1]$, then the convex combination of A, B is defined through

$$\theta A + (1-\theta) B := \left\{ \theta x + (1-\theta) y \colon x \in A, y \in B \right\}.$$

Definitions 2. A probability measure $\mathbb{P}$ defined on the Lebesgue subsets of a convex subset $\Omega \subseteq \mathbb{R}^n$ is said to be α-concave if for any Borel measurable sets $A, B \subseteq \Omega$ and for all $\theta \in [0, 1]$,

$$\mathbb{P}(\theta A + (1-\theta) B) \geq m_\alpha(\mathbb{P}(A), \mathbb{P}(B); \theta).$$

For a random variable ξ supported on $\mathbb{R}^n$, we say it is α-concave if the probability measure induced by ξ is α-concave. In particular, ξ is log-concave if it induces a 0-concave distribution.

Example 3. The PDF of an n-dimensional nondegenerate Gaussian is

$$f(x) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left[-\frac{1}{2} \|\Sigma^{-1/2}(x - \mu)\|_2 \right],$$

where μ and Σ represent its mean and covariance, respectively. Since $\ln f$ is concave, f is a log-concave function and Gaussian random variables are log-concave.

Example 4. The PDF of a uniform distribution defined on a bounded convex subset $\Omega \subset \mathbb{R}^n$ is

$$f(x) = \frac{1}{\textbf{vol}(\Omega)} \mathbb{1}\left\{ x \in \Omega \right\},$$

where **vol**(Ω) represents the volume of Ω. f is $+\infty$-concave on Ω. Therefore, n-dimensional uniform distributions over a bounded convex subset are $(1/n)$-concave.

Definitions 3. An n-dimensional random vector X is said to be elliptical and denoted by $\mathcal{E}_n(\mu, \Sigma, \phi)$ if and only if there exist a vector $\mu \in \mathbb{R}^n$, a positive semidefinite matrix $\Sigma \in \mathbb{R}^{n\times n}$, and a function $\phi\colon \mathbb{R}_+ \to \mathbb{R}$ such that the characteristic function $t \mapsto \varphi_{X-\mu}(t)$ of $X - \mu$ corresponds to $t \mapsto \phi(t^\mathsf{T}\Sigma t)$ for $t \in \mathbb{R}^n$.

Examples of elliptical distributions include uniform distribution in a ball, Gaussian, t-distribution, symmetric stable distribution, symmetric Laplace distribution, logistic distribution, and Cauchy distribution.

Definitions 4. A set $S \subseteq \mathbb{R}^n$ is called star-shaped if, for all $\xi \in S$, the line segment connecting 0 and ξ is completely contained in S. A distribution on $\mathbb{R}^n$ is called star-unimodal if it belongs to the closed convex hull of the set of uniform distributions on sets in $\mathbb{R}^n$ which are star-shaped.

The above definitions of star-shapedness and star-unimodality assume that the mode is 0, which can be achieved without loss of generality by shifting a star-unimodal random variable by its mode. Intuitively, if a star-unimodal distribution admits a density function f_ξ, then $f_\xi(zd)$ is nonincreasing in $z > 0$ for all $d \in \mathbb{R}^n$ and $d \neq 0$. That is, the density function is nonincreasing along any ray emanating from the origin. Examples of star-unimodal distributions include uniform distribution in a ball, Gaussian, t-distribution, logistic distribution, and Cauchy distribution.

We review properties of α-concave functions, α-concave probability measures, as well as elliptical and star-unimodal distributions in Appendix A.1.

2.2 Pessimistic Chance Constraint

We first review the definitions of value-at-risk (VaR) and CVaR [RU99], as well as the CVaR reformulation of $\mathcal{X}^p$ derived by [Xie19]. Then, we derive a new reformulation of $\mathcal{X}^p$ for α-concave reference distribution $\mathbb{P}$. The new reformulation leads to convexity proofs for joint (**P-CC**) with RHS uncertainty and two-sided (**P-CC**) with LHS uncertainty in Sections 2.2.1 and 2.2.2, respectively.

Definitions 5. Let X be a random variable, inducing probability distribution $\mathbb{P}_X$. The $(1-\epsilon)$-VaR of X is defined through

$$\mathsf{VaR}_{1-\epsilon}(X) := \inf\left\{\, x : \mathbb{P}_X\left[X \leq x\right] \geq 1-\epsilon \,\right\},$$

and its $(1-\epsilon)$-CVaR is defined through

$$\mathsf{CVaR}_{1-\epsilon}(X) := \min_{\gamma} \left\{ \gamma + \frac{1}{\epsilon}\mathbb{E}\Big[(X-\gamma)^+ \Big] \right\}.$$

Proposition 1 (Adapted from Theorem 1 in [Xie19])**.** For $\delta > 0$, it holds that

$$\mathcal{X}^p = \left\{ x \in \mathbb{R}^n \colon \frac{\delta}{\epsilon} + \mathsf{CVaR}_{1-\epsilon}\Big(-\mathbf{d}\left(\zeta, \mathcal{S}^c(x)\right) \Big) \leq 0 \right\}. \tag{2.2}$$

Here, random variable ζ follows the reference distribution $\mathbb{P}$ and $\mathbf{d}\left(\zeta, \mathcal{S}^c(x)\right)$ represents the distance from ζ to the "unsafe" set $\mathcal{S}^c(x)$ [CKW18],

$$\mathbf{d}\left(\zeta, \mathcal{S}^c(x)\right) := \inf_{\xi \in \Xi} \left\{ \, \|\zeta - \xi\| \colon A(x)\xi \not\leq b(x) \, \right\},$$

and Ξ is the support of ξ.

For all $x \in \mathcal{X}^p$, it holds that

$$a_i(x) = 0 \;\Rightarrow\; b_i(x) \geq 0 \quad \forall i \in [m],$$

where $a_i(x)^\mathsf{T}$ represents row i of matrix $A(x)$ and $b_i(x)$ represents entry i of vector $b(x)$, because otherwise $\mathbb{P}[A(x)\zeta \leq b(x)] = 0$ and $x \notin \mathcal{X}^p$. Assuming the above implication without loss of generality, we define function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$,

$$f(x,\zeta) := \min_{i \in [m] \setminus I(x)} \left\{ \frac{b_i(x) - a_i(x)^\mathsf{T}\zeta}{\|a_i(x)\|_*} \right\},$$

where $I(x) := \{i \in [m] : a_i(x) = 0\}$. Then, it follows from [Xie19; CKW18] that

$$\mathbf{d}\left(\zeta, \mathcal{S}^c(x)\right) = \Big(f(x,\zeta)\Big)^+.$$

In what follows, we derive new reformulations of $\mathcal{X}^p$ based on $f(x,\zeta)$. To this end, we need the following lemma to relate the the CVaR of $f(x,\zeta)$ to that of $-\mathbf{d}\left(\zeta, \mathcal{S}^c(x)\right)$ in (2.2).

Lemma 1. Let X be a random variable, then

$$\mathsf{CVaR}_{1-\epsilon}(X^-) = \mathbb{1}\left\{ \, 0 \geq \mathsf{VaR}_{1-\epsilon}(X) \, \right\} \cdot \left[\mathsf{CVaR}_{1-\epsilon}(X) - \frac{1}{\epsilon}\mathbb{E}[X^+] \right].$$

Proof. See Appendix A.2. □

Combining Proposition 1 and Lemma 1 leads to the following reformulation of $\mathcal{X}^p$.

Corollary 1. For $\delta > 0$, $\mathcal{X}^p$ can be recast into

$$\frac{\delta}{\epsilon} + \mathsf{CVaR}_{1-\epsilon}\left(-f(x,\zeta)\right) \leq \frac{1}{\epsilon}\mathbb{E}\left[\left(-f(x,\zeta)\right)^+\right] \tag{2.3}$$

$$0 \geq \mathsf{VaR}_{1-\epsilon}\left(-f(x,\zeta)\right) \tag{2.4}$$

In this chapter, we focus on cases in which $\mathbb{P}$ is α-concave. The next lemma shows that an α-concave $\mathbb{P}$ yields atomless $\mathbf{d}\left(\zeta, \mathcal{S}^c(x)\right)$ and $f(x,\zeta)$, which lead to a further reformulation of $\mathcal{X}^p$.

Lemma 2. If the reference distribution $\mathbb{P}$ is α-concave, then for all x, $\mathbb{P}\big[\mathbf{d}\left(\zeta, \mathcal{S}^c(x)\right) = y\big] = 0$ for all $y > 0$ and $\mathbb{P}\big[f(x,\zeta) = y\big] = 0$ for all $y \in \mathbb{R}$.

Proof. See Appendix A.3. □

Proposition 2. Suppose that $\mathbb{P}$ is α-concave. Then, for $\delta > 0$, $\mathcal{X}^p$ can be recast into

$$\delta \leq \int_0^{\mathsf{VaR}_\epsilon\left(f(x,\zeta)\right)} \Big(\mathbb{P}\left[f(x,\zeta) \geq t\right] - (1-\epsilon)\Big)\,\mathrm{d}t \tag{2.5}$$

$$\mathbb{P}\left[A(x)\zeta \leq b(x)\right] \geq 1 - \epsilon \tag{2.6}$$

Proof. First, moving the CVaR term to the RHS of (2.3) yields

$$\begin{aligned}
\delta &\leq \mathbb{E}\Big[f(x,\zeta)\cdot \mathbb{1}\Big\{-f(x,\zeta) \geq \mathsf{VaR}_{1-\epsilon}\left(-f(x,\zeta)\right)\Big\}\Big] - \mathbb{E}\Big[f(x,\zeta)\cdot \mathbb{1}\left\{-f(x,\zeta) \geq 0\right\}\Big] \\
&= \mathbb{E}\Big[f(x,\zeta)\cdot \mathbb{1}\Big\{\mathsf{VaR}_{1-\epsilon}\left(-f(x,\zeta)\right) \leq -f(x,\zeta) \leq 0\Big\}\Big] \\
&= \mathbb{E}\Big[f(x,\zeta)\cdot \mathbb{1}\Big\{0 \leq f(x,\zeta) \leq \mathsf{VaR}_\epsilon\left(f(x,\zeta)\right)\Big\}\Big],
\end{aligned} \tag{2.7}$$

where the first equality is because $f(x,\zeta)$ is atomless and the second equality is because $\mathsf{VaR}_{1-\epsilon}(-X) = -\mathsf{VaR}_\epsilon(X)$. Now, we use the wedding lake representation of nonnegative

integrable functions to further recast the RHS of (2.7) as

$$
\begin{aligned}
&\mathbb{E}\Big[f(x,\zeta) \cdot \mathbb{1}\left\{ 0 \le f(x,\zeta) \le \mathsf{VaR}_\epsilon\left(f(x,\zeta)\right) \right\} \Big] \\
&= \int_{\Xi} f(x,\zeta) \cdot \mathbb{1}\left\{ 0 \le f(x,\zeta) \le \mathsf{VaR}_\epsilon\left(f(x,\zeta)\right) \right\} \mathrm{d}\mathbb{P}(\zeta) \\
&= \int_{\Xi} \int_{\mathbb{R}_+} \mathbb{1}\left\{ t \le f(x,\zeta) \cdot \mathbb{1}\left\{ 0 \le f(x,\zeta) \le \mathsf{VaR}_\epsilon\left(f(x,\zeta)\right) \right\} \right\} \mathrm{d}t \ \mathrm{d}\mathbb{P}(\zeta) \\
&= \int_{\Xi} \int_{\mathbb{R}_+} \mathbb{1}\left\{ t \le f(x,\zeta) \le \mathsf{VaR}_\epsilon\left(f(x,\zeta)\right) \right\} \mathrm{d}t \ \mathrm{d}\mathbb{P}(\zeta) \\
&= \int_{\mathbb{R}_+} \mathbb{P}\Big[t \le f(x,\zeta) \le \mathsf{VaR}_\epsilon\left(f(x,\zeta)\right) \Big] \mathrm{d}t, \qquad \text{(by the Tonelli's Theorem)} \\
&= \int_0^{\mathsf{VaR}_\epsilon\left(f(x,\zeta)\right)} \Big(\mathbb{P}\left[f(x,\zeta) \ge t \right] - (1-\epsilon) \Big) \mathrm{d}t,
\end{aligned}
$$

where the first two equalities are by definitions of expectation and wedding lake representation, respectively. We justify the third equality by arguing that, for any $x \in \mathcal{X}^p$ and $\zeta \in \Xi$,

$$
\mathbb{1}\left\{ t \le f(x,\zeta) \cdot \mathbb{1}\left\{ 0 \le f(x,\zeta) \le \mathsf{VaR}_\epsilon\left(f(x,\zeta)\right) \right\} \right\} = \mathbb{1}\left\{ t \le f(x,\zeta) \le \mathsf{VaR}_\epsilon\left(f(x,\zeta)\right) \right\} \tag{2.8}
$$

holds Lebesgue-almost everywhere for $t \in \mathbb{R}_+$. We discuss the following three cases:

(i) If ζ makes $f(x,\zeta) < 0$, then the LHS of (2.8) simplifies to $\mathbb{1}\{ t \le 0 \}$, which coincides with the RHS.

(ii) If ζ makes $f(x,\zeta) \in [0, \mathsf{VaR}_\epsilon(f(x,\zeta))]$, then the LHS of (2.8) simplifies to $\mathbb{1}\{ t \le f(x,\zeta) \}$, coinciding with the RHS.

(iii) If ζ makes $f(x,\zeta) > \mathsf{VaR}_\epsilon(f(x,\zeta))$, then the LHS and RHS of (2.8) simplify to $\mathbb{1}\{ t \le 0 \}$ and 0, respectively, which differ only at $t = 0$ for $t \in \mathbb{R}_+$.

The last equality is because

$$
\mathbb{P}\Big[t \le f(x,\zeta) \le \mathsf{VaR}_\epsilon\left(f(x,\zeta)\right) \Big] = \mathbb{P}\Big[t \le f(x,\zeta) \Big] - \mathbb{P}\Big[t \ge \mathsf{VaR}_\epsilon\left(f(x,\zeta)\right) \Big]
$$

when $t \in [0, \mathsf{VaR}_\epsilon f(x,\zeta)]$. This recasts (2.3) into (2.5).

Second, constraint (2.4) is equivalent to $\mathbb{P}\big[f(x,\zeta) \geq 0\big] \geq 1-\epsilon$ by definition of VaR, which can be further recast as

$$\mathbb{P}\Big[a_i(x)^\mathsf{T}\zeta \leq b_i(x),\ \forall i \in [m] \setminus I(x)\Big] \geq 1-\epsilon$$

by definition of $f(x,\zeta)$. For all $x \in \mathcal{X}^p$ and $i \in [m]$, we assume without loss of generality that $b_i(x) \geq 0$ whenever $a_i(x) = 0$ (because otherwise $\mathbb{P}[A(x)\zeta \leq b(x)] = 0$), and it holds that $a_i(x)^\mathsf{T}\zeta \leq b_i(x)$ for all $i \in I(x)$. It follows that (2.4) is equivalent to (2.6), which completes the proof. □

Remark 1. We notice that constraint (2.6) is simply (**CC**) with respect to the reference distribution $\mathbb{P}$ of the Wasserstein ball $\mathcal{P}$. In addition, constraint (2.5) encodes a robust guarantee. Intuitively, the RHS of (2.5) evaluates the budget needed to shift the probability masses of $\mathbb{P}$ so that the corresponding (**CC**) can be violated. Constraint (2.5) makes sure that this budget is beyond the radius of $\mathcal{P}$, i.e., (**CC**) will not be violated as long as the shifted distribution lies within $\mathcal{P}$.

2.2.1 Joint (P-CC) with RHS Uncertainty

For (**CC**) with RHS uncertainty, it is well celebrated that the ensuing feasible region is convex when ξ has an α-concave distribution (particularly, ξ is log-concave when $\alpha = 0$) [Pré13; SDR09].

Proposition 3 (Theorem 4.39 and Corollary 4.41 in [SDR09])**.** If $\xi \in \mathbb{R}^m$ follows an α-concave probability distribution, then $H(x) := \mathbb{P}\left[A\xi \leq b(x)\right]$ is α-concave on the set $\mathcal{D} := \{x \in \mathbb{R}^n \colon \exists\, \xi \text{ such that } A\xi \leq b(x)\}$ and the following set is convex and closed:

$$\mathcal{X} := \Big\{\, x \in \mathbb{R}^n \colon \mathbb{P}\left[A\xi \leq b(x)\right] \geq 1-\epsilon \,\Big\}.$$

In this subsection, we seek to extend this result to (**P-CC**).

Theorem 1. Suppose that the reference distribution $\mathbb{P}$ of $\mathcal{P}$ is α-concave with $\alpha \geq -1$. Then, for $\delta > 0$ the set

$$\mathcal{X}^p_{\mathsf{R}} := \left\{ x \in \mathbb{R}^n \colon \inf_{\mathbb{Q}\in\mathcal{P}} \mathbb{Q}\left[A\xi \leq b(x)\right] \geq 1-\epsilon \right\}$$

is convex and closed.

Although Theorem 1 pertains to (**P-CC**) with linear inequalities, the convexity result extends to (**P-CC**) with *quasi-concave* inequalities. We present a detailed description and a proof for this generalization in Appendix A.4. Before presenting a proof of Theorem 1, we present some useful lemmas. Without loss of generality, we assume that each row of matrix A, denoted by a_i^T for all $i \in [m]$, satisfies

(i) $a_i \neq 0$, because otherwise we can add a deterministic constraint $b_i(x) \geq 0$ to $\mathcal{X}_\mathsf{R}^p$ and eliminate inequality i from (**P-CC**);

(ii) $\|a_i\|_* = 1$, because otherwise we can divide both sides of inequality i by $\|a_i\|_*$ and set $a_i \leftarrow a_i/\|a_i\|_*$, $b_i(x) \leftarrow b_i(x)/\|a_i\|_*$.

Recall that for $\zeta \in \mathbb{R}^m$ the distance $\mathbf{d}\big(\zeta, \mathcal{S}^c(x)\big)$ to the unsafe set satisfies $\mathbf{d}\big(\zeta, \mathcal{S}^c(x)\big) = \big(f(x,\zeta)\big)^+$ with

$$f(x,\zeta) = \min_{i\in[m]} \Big\{ b_i(x) - a_i^\mathsf{T}\zeta \Big\}$$

and $f(x,\zeta)$ is jointly concave in (x,ζ).

Lemma 3. For all $\epsilon \in (0,1)$, if ζ has an α-concave distribution with $\alpha \geq -1$, then $\mathsf{VaR}_{1-\epsilon}\big(f(x,\zeta)\big)$ is concave in x on $\mathbb{R}^n$.

Proof. See Appendix A.5. □

Lemma 4. Suppose that $f(\cdot,\cdot)\colon \mathbb{R}^n \times \Xi \to \mathbb{R}$ is a continuous function, ζ follows an α-concave distribution $\mathbb{P}$, and $f(x,\zeta)$ is atomless for any $x \in \mathbb{R}^n$. Then,

$$\psi(x,t) := \mathbb{P}\big[f(x,\zeta) \geq t\big] - (1-\epsilon) \qquad \text{and} \qquad \phi(x,y) := \int_0^y \psi(x,t)\,\mathrm{d}t$$

are both continuous on $\mathbb{R}^n \times \mathbb{R}_+$.

Proof. See Appendix A.6. □

Now we are ready to prove Theorem 1.

Proof of Theorem 1. First, recall that by Proposition 2 we recast $\mathcal{X}_\mathsf{R}^p$ as constraints (2.5)–(2.6). For ease of exposition, we denote by $G(x)$ the RHS of (2.5).

Second, to show that $\mathcal{X}_\mathsf{R}^p$ is closed, it suffices to prove the closedness of the feasible region of (2.5) because that of (2.6) follows from Proposition 3. To this end, we notice that $\mathsf{VaR}_\epsilon\big(f(x,\zeta)\big)$ is continuous in x due to its concavity. Then, by Lemma 4 the mapping

$$x \mapsto \int_0^{\mathsf{VaR}_\epsilon(f(x,\zeta))} \mathbb{P}\Big[f(x,\zeta) \geq t\Big]\,\mathrm{d}t$$

is continuous. It follows that $G(x)$ is continuous and the feasible region of (2.5) is closed.

Third, to show that $\mathcal{X}_{\mathsf{R}}^{p}$ is convex, it suffices to prove the convexity of the feasible region of (2.5) because that of (2.6) follows from Proposition 3. To that end, by Proposition 3 and Lemma 11, ψ is α-concave in (x,t) on $\mathbf{dom}\,\psi := \left\{\, (x,t)\colon \psi(x,t) \geq 0 \,\right\} = \left\{\, (x,t)\colon t \leq \mathsf{VaR}_{\epsilon}(f(x,\zeta)) \,\right\}$, which is convex by Lemma 3. Then, for any $x_0, x_1 \in \mathcal{X}_{\mathsf{R}}^{p}$ and any $t_0 \in S_0 := [0, \mathsf{VaR}_{\epsilon}(f(x_0,\zeta))]$, $t_1 \in S_1 := [0, \mathsf{VaR}_{\epsilon}(f(x_1,\zeta))]$ it holds that

$$\psi(x_{1/2}, t_{1/2}) \geq m_{\alpha}\left[\psi(x_0,t_0), \psi(x_1,t_1); \frac{1}{2}\right],$$

where $x_{1/2} = (x_0 + x_1)/2$ and $t_{1/2} = (t_0 + t_1)/2$. It follows that

$$\begin{aligned} m_{\alpha_1^*}\left[\int_{S_0} \psi(x_0,t)\,\mathrm{d}t, \int_{S_1} \psi(x_1,t)\,\mathrm{d}t; \frac{1}{2}\right] &\leq \int_{\frac{1}{2}S_0 + \frac{1}{2}S_1} \psi(x_{1/2},t)\,\mathrm{d}t \\ &\leq \int_{S_{1/2}} \psi(x_{1/2},t)\,\mathrm{d}t \end{aligned}$$

where the first inequality is due to Proposition 24 and $\alpha_1^* \geq -\infty$ is a function of α (see Proposition 24), and the second inequality is because $\frac{1}{2}S_0 + \frac{1}{2}S_1 \subseteq S_{1/2} := [0, \mathsf{VaR}_{\epsilon}(f(x_{1/2},\zeta))]$. In other words, we obtain that

$$m_{\alpha_1^*}\left[G(x_0), G(x_1); \frac{1}{2}\right] \leq G\left(x_{1/2}\right)$$

and $G(x)$ is midpoint α_1^*-concave, and particularly, midpoint quasi-concave. Then, its continuity implies that $G(x)$ is quasi-concave and constraint (2.5) yields a convex feasible region. This finishes the proof. □

We close this section by commenting on the worst-case distribution with respect to (**P-CC**). For expectation-oriented optimization, it has been observed that if the Wasserstein ball is centered around a Gaussian reference distribution, then the worst-case probability distribution is also Gaussian [Kuh+19]. In contrast, the following example demonstrates that this is not the case for (**P-CC**). We present a detailed proof for this example in Appendix A.7.

Example 5. Consider the (**P-CC**)

$$\inf_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left[\xi \leq x\right] \geq 1 - \epsilon, \tag{Ex}$$

where $\epsilon \in (0, 1/2)$, and the Wasserstein ball $\mathcal{P}$ is centered around the 1-dimensional

standard Gaussian distribution and has a radius $\delta > 0$. Then, there does not exist a Gaussian distribution $\nu \in \mathcal{P}$ such that $\nu[\xi \leq x] = \inf_{\mathbb{P} \in \mathcal{P}} \mathbb{P}[\xi \leq x]$.

2.2.2 Two-Sided (P-CC) with LHS Uncertainty

We move on to two-sided (**P-CC**) with LHS uncertainty, defined through

$$\mathcal{X}_{\mathsf{T}}^{p} := \left\{ (x, \ell, u) \in \mathbb{R}^{n+2} \colon \inf_{\mathbb{Q} \in \mathcal{P}} \mathbb{Q}\left[\ell \leq x^{\mathsf{T}}\xi \leq u\right] \geq 1 - \epsilon \right\}.$$

To study the convexity of $\mathcal{X}_{\mathsf{T}}^{p}$, we make the following two assumptions about the Wasserstein ball.

Assumption 1. The Wasserstein ball $\mathcal{P}$ is such that (i) the reference distribution $\mathbb{P}$ is elliptical, particularly $\mathcal{E}_n(0, \Sigma, \phi)$ with $\Sigma \succ 0$. (ii) the norm $\|\cdot\|$ in d_W is an ellipsoidal norm with regard to $\Sigma^{1/2}$, i.e., $\|\cdot\| = \|\Sigma^{-1/2}(\cdot)\|_2$ (or equivalently, $\|\cdot\|_* = \|\Sigma^{1/2}(\cdot)\|_2$).

Assumption 2. The reference distribution $\mathbb{P}$ of $\mathcal{P}$ is star-unimodal.

Examples of $\mathbb{P}$ satisfying both Assumptions 1–2 include uniform distribution in a ball, Gaussian, t-distribution, logistic distribution, and Cauchy distribution. We remark that Assumption 2 implies that the mode of $\mathbb{P}$ is the origin, which can be achieved without loss of generality by shifting ξ by its mode, if different from 0. In addition, since $\mathbb{P}$ is elliptical, it is identically distributed as $R \cdot \Sigma^{1/2} U_n$ for a nonnegative random variable R and an n-dimensional random vector U_n uniformly distributed on the unit sphere S^{n-1} of $\mathbb{R}^n$ (see Remark 7 in Appendix A.1). We denote by $\mathbb{P}_0$ the probability measure induced by $R \cdot e_1^{\mathsf{T}} U_n$ and establish the convexity of $\mathcal{X}_{\mathsf{T}}^{p}$ as follows.

Theorem 2. Suppose that Assumptions 1 and 2 hold, $\epsilon \in (0, \frac{1}{2})$, and $\delta > 0$. Define

$$g_\epsilon(\ell, u) := \int_0^{+\infty} \Big[\Phi(u - t) - \Phi(\ell + t) - (1 - \epsilon)\Big]^+ \mathrm{d}t$$
$$\text{and} \qquad \mathcal{C}_\delta := \left\{ (\ell, u) \in \mathbb{R}^2 \colon \delta \leq g_\epsilon(\ell, u) \right\},$$

where $\Phi \colon \mathbb{R} \to [0, 1]$ denotes the cumulative distribution function of $\mathbb{P}_0$. Then,

$$\mathcal{X}_{\mathsf{T}}^{p} = \left\{ (x, \ell, u) \in \mathbb{R}^{n+2} \colon \exists\, s \geq 0 \text{ such that } \|x\|_* \leq s,\ (\ell, u, s) \in \mathsf{co}(\mathcal{C}_\delta) \right\},$$

where $\mathsf{co}(\mathcal{C}_\delta) := \mathbf{cl}\left(\left\{ (\ell, u, s) \in \mathbb{R}^3 \colon s > 0, (\ell/s, u/s) \in \mathcal{C}_\delta \right\}\right)$ is the cone induced by $\mathcal{C}_\delta$ and $\mathbf{cl}\,(\cdot)$ denotes the closure operator. Furthermore, $\mathcal{X}_{\mathsf{T}}^{p}$ is convex and closed.

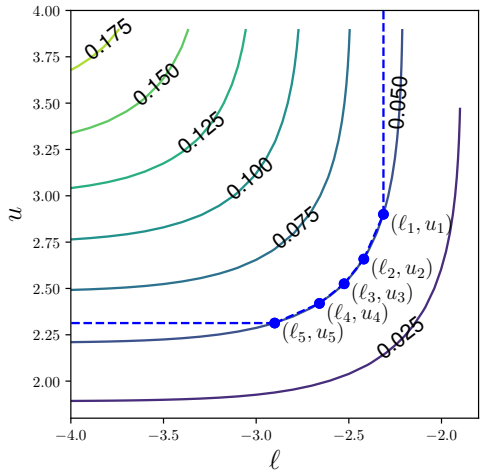


Figure 2.1: Contours of $g_\epsilon(\ell, u)$ with varying δ and a polyhedral inner approximation $\widehat{\mathcal{C}}_N$ of $\mathcal{C}_\delta$ with $N = 5$ and $\delta = 0.050$

Before presenting a proof of Theorem 2, we visualize the function $g_\epsilon(\ell, u)$ and set $\mathcal{C}_\delta$. Figure 2.1 depicts the contour of $g_\epsilon(\ell, u)$, which is symmetric with respect to the line $\ell + u = 0$. This is because $\mathbb{P}_0$ is elliptical and so $\Phi(u) - \Phi(\ell)$ is symmetric with respect to $\ell + u = 0$ on $\mathbb{R}_- \times \mathbb{R}_+$, i.e.,

$$\Phi(u) - \Phi(\ell) = \Phi(-\ell) - \Phi(-u), \quad \forall (\ell, u) \in \mathbb{R}_- \times \mathbb{R}_+,$$

which implies that

$$g_\epsilon(\ell, u) = g_\epsilon(-u, -\ell), \quad \forall (\ell, u) \in \mathbb{R}_- \times \mathbb{R}_+.$$

In addition, from the same figure, we observe that the superlevel set $\mathcal{C}_\delta$ of $g_\epsilon(\ell, u)$ takes a convex shape, which explains (intuitively) the convexity of $\mathcal{X}_\mathsf{T}^p$. We now make this observation concrete.

Proof of Theorem 2. First, we present a technical fact that connects $\mathcal{X}_\mathsf{T}^p$ with $\mathcal{C}_\delta$ and provide a proof in Appendix A.8.

Fact. For any $x \neq 0$, $(x, \ell, u) \in \mathcal{X}_\mathsf{T}^p$ if and only if $\left(\frac{\ell}{\|x\|_*}, \frac{u}{\|x\|_*}\right) \in \mathcal{C}_\delta$.

Second, we show that $\mathcal{C}_\delta$ is convex. Since $\mathbb{P}_0 \overset{d}{=} R \cdot e_1^\mathsf{T} U_n$ is unimodal, its distribution function Φ is convex on $(-\infty, 0)$ and concave on $(0, +\infty)$. In addition, $\mathcal{C}_\delta \subseteq \mathbb{R}_- \times \mathbb{R}_+$ because $\epsilon < \frac{1}{2}$. Then, $\Phi(u - t) - \Phi(\ell - t)$ is jointly concave in (u, ℓ, t), implying that $g_\epsilon(\ell, u)$ is log-concave. It follows that $\mathcal{C}_\delta$ is convex.

Therefore, to prove that $\mathcal{X}_{\mathsf{T}}^p$ is convex, it remains to show that $(x, \ell, u) \in \mathcal{X}_{\mathsf{T}}^p$ if and only if there exists an $s \geq \|x\|_*$ such that $(\ell, u, s) \in \mathsf{co}(\mathcal{C}_\delta)$. To this end, we discuss the following two cases:

1. Suppose that $x = 0$. For any $(0, \ell, u) \in \mathcal{X}_{\mathsf{T}}^p$, we have $\ell \leq 0 \leq u$ because otherwise $\mathbb{P}[\ell \leq 0 \leq u] < 1/2 < 1-\epsilon$, violating the assumption that $(0, \ell, u) \in \mathcal{X}_{\mathsf{T}}^p$. Then, $s := 1/n$ for a sufficiently large integer n ensures that $(\ell/s, u/s) \in \mathcal{C}_\delta$ and so $(\ell, u, s) \in \mathsf{co}(\mathcal{C}_\delta)$. On the contrary, for any $(0, \ell, u) \in \mathbb{R}^{n+2}$ such that there exists an $s \geq 0$ with $(\ell, u, s) \in \mathsf{co}(\mathcal{C}_\delta)$, by definition of $\mathsf{co}(\mathcal{C}_\delta)$ there exists a sequence $\{(\ell_n, u_n, s_n)\}_{n=1}^\infty$ converging to (ℓ, u, s) such that $s_n > 0$ and $g_\epsilon(\ell_n/s_n, u_n/s_n) \geq \delta$ for all n. Then, $\ell_n < 0$ and $u_n > 0$ for all n because otherwise $g_\epsilon(\ell_n/s_n, u_n/s_n) = 0 < \delta$. Driving n to infinity yields that $\ell \leq 0$ and $u \geq 0$. Hence, $(0, \ell, u) \in \mathcal{X}_{\mathsf{T}}^p$.

2. Suppose that $x \neq 0$. Pick any $(x, \ell, u) \in \mathcal{X}_{\mathsf{T}}^p$, then the above fact implies that $\left(\frac{\ell}{\|x\|_*}, \frac{u}{\|x\|_*}\right) \in \mathcal{C}_\delta$. Hence, $s := \|x\|_* > 0$ ensures that $(\ell, u, s) \in \mathsf{co}(\mathcal{C}_\delta)$. On the contrary, pick any $(x, \ell, u) \in \mathbb{R}^{n+2}$ such that $x \neq 0$ and there exists an $s \geq \|x\|_* > 0$ with $(\ell, u, s) \in \mathsf{co}(\mathcal{C}_\delta)$. By definition of $\mathsf{co}(\mathcal{C}_\delta)$, there exists a sequence $\{(\ell_n, u_n, s_n)\}_{n=1}^\infty$ converging to (ℓ, u, s) such that $s_n > 0$ and $g_\epsilon(\ell_n/s_n, u_n/s_n) \geq \delta$ for all n. Then,

$$g_\epsilon\left(\frac{\ell}{\|x\|_*}, \frac{u}{\|x\|_*}\right) \geq g_\epsilon\left(\frac{\ell}{s}, \frac{u}{s}\right) = \lim_{n\to\infty} g_\epsilon\left(\frac{\ell_n}{s_n}, \frac{u_n}{s_n}\right) \geq \delta,$$

where the first inequality is because the function $g_\epsilon(\ell, u)$ is nonincreasing in ℓ and nondecreasing in u, and the equality is due to the dominated convergence theorem (or equivalently, the continuity of g_ϵ). It follows that $\left(\frac{\ell}{\|x\|_*}, \frac{u}{\|x\|_*}\right) \in \mathcal{C}_\delta$ and so $(x, \ell, u) \in \mathcal{X}_{\mathsf{T}}^p$ by the above fact. This completes the proof.

□

Assumptions 1–2 are not only sufficient for the convexity of $\mathcal{X}_{\mathsf{T}}^p$, but also tight in the sense that Theorem 2 ceases to hold for the lack of either assumption. We demonstrate through the following two examples.

Example 6 (Loss of Convexity Without Assumption 2)**.** Consider an example of $\mathcal{X}_{\mathsf{T}}^p$ with $n = 1$ and the reference distribution $\mathbb{P}$ of the Wasserstein ball is identical to that of a random variable $\zeta_1 := R_c \cdot U_1$, where R_c and U_1 are independent random variables following the arcsine distribution and the uniform distribution on $\{-1, 1\}$, respectively. Hence, $\mathbb{P}$ satisfies Assumption 1 but violates Assumption 2 (see Figure 2.2a). By Proposition 25, we

have

$$\mathcal{X}_{\mathsf{T}}^{p} \cap \{\, (x,\ell,u) \in \mathbb{R}^3 \colon x = 1 \,\} = \left\{\, (1,\ell,u) \in \mathbb{R}^3 \colon \inf_{\mathbb{Q}\in\mathcal{P}} \mathbb{Q}\left[\ell \le \xi \le u\right] \ge (1-\epsilon) \,\right\}$$
$$= \left\{\, (1,\ell,u) \in \mathbb{R}^3 \colon g_\epsilon(\ell,u) \ge \delta \,\right\}.$$

In addition, the cumulative distribution function of ζ_1 satisfies

$$\Phi(t) = \frac{1}{2}\mathbb{P}\left[R_c \le t\right] + \frac{1}{2}\mathbb{P}\left[-R_c \le t\right] = \frac{1}{2}(F_{R_c}(t)+1)\cdot \mathbb{1}\left\{\, t \ge 0 \,\right\} + \frac{1}{2}(1-F_{R_c}(-t))\cdot \mathbb{1}\left\{\, t < 0 \,\right\},$$

where $F_{R_c}(\cdot)$ denotes the distribution function of R_c. Next, we show that, when restricted to the line segment $L_a := \{\, (\ell,u) \in (-1,0)\times(0,1) \colon u - \ell = a \,\}$ with $a \in (1,2)$, $\Phi(u) - \Phi(\ell)$ is a strictly convex function, which fails the midpoint concavity of the function $g_\epsilon(\ell,u)$ and shows that $\mathcal{X}_{\mathsf{T}}^{p} \cap \{\, (x,\ell,u) \in \mathbb{R}^3 \colon x = 1 \,\}$ is non-convex.

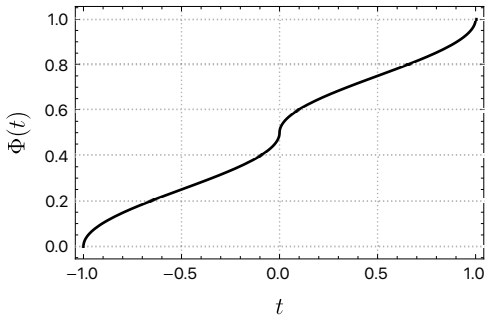
To this end, we define a function $\Psi_a : [a-1,1] \mapsto \mathbb{R}_+$ with $\Psi_a(u) := \Phi(u) - \Phi(u-a)$. A simple calculation shows that $\Psi_a''(u) = \frac{1}{4\pi}\left(f_{\Psi_a''}(u) + f_{\Psi_a''}(a-u)\right)$, where $f_{\Psi_a''}(u) := (2u-1)/(u(1-u))^{3/2}$. Because $f_{\Psi_a''}$ is strictly increasing on $(0,1)$ (see Figure 2.2b), we have

$$\Psi_a''(u) > \frac{1}{4\pi}\left(f_{\Psi_a''}(u) + f_{\Psi_a''}(1-u)\right) = 0, \quad \forall u \in (0,1)$$

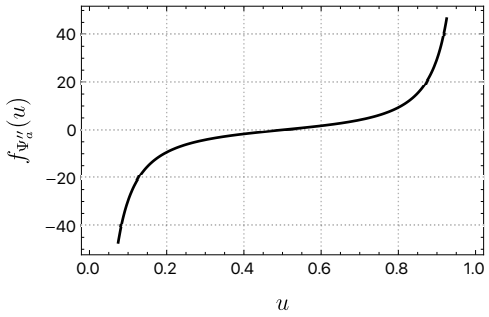
implying that $\Psi_a(u)$ is strictly convex. To see how this fails the midpoint concavity of $g_\epsilon(\ell,u)$, we define $u_1 := a-1$ and $u_2 := 1$. Then, for any $a \in (1,2)$, we have

$$\begin{aligned}
g_\epsilon\left(-\frac{a}{2},\frac{a}{2}\right) &= \int_0^{+\infty}\left[\Phi\left(\frac{a}{2}-t\right) - \Phi\left(-\frac{a}{2}+t\right) - (1-\epsilon)\right]^+ \mathrm{d}t \\
&= \int_0^{+\infty}\left[\Psi_{a-2t}\left(\frac{a}{2}-t\right) - (1-\epsilon)\right]^+ \mathrm{d}t = \int_0^{+\infty}\left[\Psi_{a-2t}\left(\frac{1}{2}(u_1+u_2)-t\right) - (1-\epsilon)\right]^+ \mathrm{d}t \\
&< \frac{1}{2}\int_0^{+\infty}\left[\Psi_{a-2t}\left(u_1-t\right) - (1-\epsilon)\right]^+ \mathrm{d}t + \frac{1}{2}\int_0^{+\infty}\left[\Psi_{a-2t}\left(u_2-t\right) - (1-\epsilon)\right]^+ \mathrm{d}t \\
&= \frac{1}{2}\left[g_\epsilon(u_1-a,u_1) + g_\epsilon(u_2-a,u_2)\right] = g_\epsilon(-1,a-1),
\end{aligned}$$

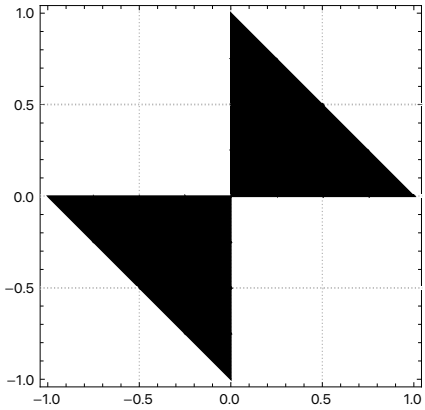
where the inequality is due to the strict convexity of Ψ_a, and the last equality uses the symmetry of g_ϵ. Let $\delta_{\epsilon,a} := g_\epsilon(-1,a_1)$, then we see that $(-1,a-1)$ and $(1-a,1)$ are in the $\delta_{\epsilon,a}$-superlevel set of g_ϵ, while their midpoint, $(-a/2,a/2)$, falls out of the $\delta_{\epsilon,a}$-superlevel set of g_ϵ. This implies that $\mathcal{X}_{\mathsf{T}}^{p}$, and particularly the intersection $\mathcal{X}_{\mathsf{T}}^{p} \cap \{\, (x,\ell,u) \in \mathbb{R}^3 \colon x = 1 \,\}$, are non-convex.



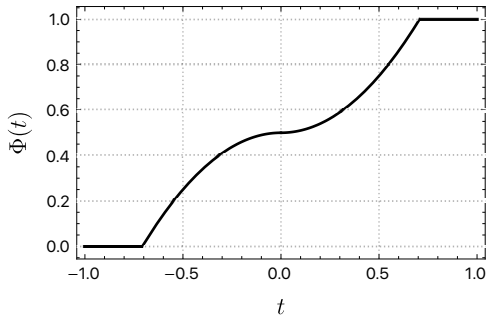
(a) Distribution function of ζ_1



(b) Graph of $f_{\Psi_a''}(u)$ on $(0, 1)$



(c) Density of ζ_2.



(d) Density of $\zeta_2^{\mathsf{T}}[1; 1]/\sqrt{2}$.

Figure 2.2: Visualization of random variables ζ_1 and ζ_2 in Examples 6–7

Example 7 (Loss of Convexity Without Assumption 1)**.** Consider an example of $\mathcal{X}_{\mathsf{T}}^p$ with $n = 2$ and the reference distribution $\mathbb{P}$ is a uniform distribution on the star-shaped set depicted in Figure 2.2c, where we let $\zeta_2 \in \mathbb{R}^2$ denote the random vector following distribution $\mathbb{P}$. Then, $\mathbb{P}$ satisfies Assumption 2 but violates Assumption 1. Proposition 25 yields

$$\begin{aligned}
&\mathcal{X}_{\mathsf{T}}^p \cap \{\, (x,\ell,u) \in \mathbb{R}^4 \colon x^{\mathsf{T}} = [1/\sqrt{2}, 1/\sqrt{2}] \,\} \\
&= \left\{ (1/\sqrt{2}, 1/\sqrt{2}, \ell, u) \in \mathbb{R}^4 \colon \inf_{\mathbb{Q}\in\mathcal{P}} \mathbb{Q}\left[\ell \le \xi^{\mathsf{T}}[1;\ 1]/\sqrt{2} \le u\right] \ge (1-\epsilon) \right\} \\
&= \left\{ (1/\sqrt{2}, 1/\sqrt{2}, \ell, u) \in \mathbb{R}^4 \colon g_\epsilon(\ell, u) \ge \delta \right\}.
\end{aligned}$$

In addition, the cumulative distribution function of $\zeta_2^{\mathsf{T}}[1;\ 1]/\sqrt{2}$ is

$$\Phi(t) = \left(\frac{1}{2} - t^2\right) \cdot \mathbb{1}\left\{ -\frac{1}{\sqrt{2}} \le t \le 0 \right\} + \left(\frac{1}{2} + t^2\right) \cdot \mathbb{1}\left\{ 0 \le t \le \frac{1}{\sqrt{2}} \right\},$$

whose graph is depicted in Figure 2.2d. Since $\Phi(t)$ is strictly convex on $[0, 1/\sqrt{2}]$, $\Phi(u) - \Phi(\ell)$ is strictly convex on the set $\{\, (\ell, u) \in [-1/\sqrt{2}, 0] \times [0, 1/\sqrt{2}] \,\}$, and so is its restriction to the line segment $\{\, (\ell, u) \in (-1, 0) \times (0, 1) \colon u - \ell = a \,\}$. Hence, following a similar argument as in Example 6, we can show that $\mathcal{X}_{\mathsf{T}}^p$, and particularly the intersection $\mathcal{X}_{\mathsf{T}}^p \cap \{\, (x,\ell,u) \in \mathbb{R}^4 \colon x^{\mathsf{T}} = [1/\sqrt{2}, 1/\sqrt{2}] \,\}$, are non-convex.

We end this section by mentioning two special cases of the two-sided (**P-CC**) with LHS uncertainty that often arise in practice and admit second-order conic representations. The first case considers symmetric bounds, i.e., $\ell = -u$ for $u \ge 0$.

Corollary 2. Suppose that Assumption 1 holds, $\epsilon \in (0, \frac{1}{2})$, and $\delta > 0$. Define

$$\mathcal{X}_{\mathsf{TS}}^p := \left\{ (x,u) \in \mathbb{R}^n \times \mathbb{R}_+ \colon \inf_{\mathbb{Q}\in\mathcal{P}} \mathbb{Q}\left[-u \le x^{\mathsf{T}}\xi \le u\right] \ge 1-\epsilon \right\}$$

$$\text{and} \qquad g_\epsilon^s(r) := \int_0^{+\infty} \Big[\Phi(r-t) - \Phi(-r+t) - (1-\epsilon)\Big]^+ \mathrm{d}t.$$

Then, it holds that

$$\mathcal{X}_{\mathsf{TS}}^p = \left\{ (x,u) \in \mathbb{R}^n \times \mathbb{R}_+ \colon u \ge \|x\|_* \cdot \inf_r \{\, r \ge 0 \colon g_\epsilon^s(r) \ge \delta \,\} \right\}.$$

Proof. See Appendix A.9. □

The second case considers individual (**P-CC**), which can be obtained by driving ℓ to

$-\infty$ in the two-sided (**P-CC**). The proof of this case is similar to that of Corollary 2 and so omitted.

Corollary 3. Suppose that Assumption 1 holds, $\epsilon \in (0, \frac{1}{2})$, and $\delta > 0$. Define

$$\mathcal{X}_{\mathsf{I}}^{p} := \left\{ (x,u) \in \mathbb{R}^n \times \mathbb{R}_+ \colon \inf_{\mathbb{Q}\in\mathcal{P}} \mathbb{Q}\left[x^{\mathsf{T}}\xi \le u\right] \ge 1-\epsilon \right\},$$

$$\overline{g}_\epsilon(r) := \int_0^{+\infty} \Big[\Phi(r-t) - (1-\epsilon)\Big]^+ \mathrm{d}t, \quad \text{and} \quad (\overline{g})_\epsilon^{-1}(s) := \inf_{r\ge 0} \{\, r \colon \overline{g}_\epsilon(r) \ge s \,\}.$$

Then, it holds that

$$\mathcal{X}_{\mathsf{I}}^{p} = \left\{ (x,u) \in \mathbb{R}^n \times \mathbb{R}_+ \colon u \ge \|x\|_* \cdot (\overline{g})_\epsilon^{-1}(\delta) \right\}.$$

The same representation in Corollary 3 has been derived in [CX21]. In fact, using integration by part, one can verify that the coefficient $(\overline{g})_\epsilon^{-1}(\delta)$ in the above representation equals the η^* in Theorem 4.8 of [CX21].

2.3 Solution Approaches for Pessimistic Chance Constraint

The convexity results in the previous section inspires us to study solution approaches for solving (**P-CC**) based on convex/continuous optimization. We study a block coordinate ascent algorithm for (**P-CC**) with RHS uncertainty in Section 2.3.1 and a second-order conic inner approximation approach for two-sided (**P-CC**) with LHS uncertainty in Section 2.3.2.

2.3.1 Block Coordinate Ascent Algorithm

We focus on a model with joint (**P-CC**) and RHS uncertainty: $\min_{x\in X}\{c^{\mathsf{T}}x : (\textbf{P-CC})\}$, where $c \in \mathbb{R}^n$ represents cost coefficients and $X \subseteq \mathbb{R}^n$ represents a set that is deterministic, compact, and convex. By Proposition 2, this model is equivalent to

$$\min_{x\in X} \; c^{\mathsf{T}}x \tag{2.9a}$$

$$\text{s.t.} \;\; \delta \le \int_0^{\mathsf{VaR}_\epsilon(f(x,\zeta))} \Big(\mathbb{P}\big[f(x,\zeta) \ge t\big] - (1-\epsilon)\Big)\,\mathrm{d}t, \tag{2.9b}$$

$$\mathsf{VaR}_\epsilon\big(f(x,\zeta)\big) \ge 0, \tag{2.9c}$$

where $f(x,\zeta) = \min_{i\in[m]}\{b_i(x) - a_i^{\mathsf{T}}\zeta\}$. Here, constraint (2.9b) appears challenging because its RHS involves an integral with upper limit $\mathsf{VaR}_\epsilon(f(x,\zeta))$. To make the model computable, we define a new variable $y \geq 0$ to represent $\mathsf{VaR}_\epsilon(f(x,\zeta))$.

Proposition 4. For $y \geq 0$, define

$$\phi(x,y) := \int_0^y \left(\mathbb{P}\Big[f(x,\zeta) \geq t\Big] - (1-\epsilon)\right) \mathrm{d}t.$$

If $\mathbb{P}$ is α-concave with $\alpha \geq -1$, then $\phi(x,y)$ is α_1^*-concave on

$$\mathbf{dom}\,\phi := \left\{\, (x,y) \in X \times \mathbb{R}_+ \colon \mathbb{P}\left[f(x,\zeta) \geq y\right] \geq (1-\epsilon) \,\right\},$$

where α_1^* is defined in Proposition 24. In addition, $\mathbf{dom}\,\phi$ is closed and constraints (2.9b)–(2.9c) is equivalent to

$$\delta \leq \max_{y\geq 0} \phi(x,y). \tag{2.9d}$$

Proof. See Appendix A.10. □

By Proposition 4, formulation (2.9a)–(2.9c) is equivalent to $\min_{x\in X}\{c^{\mathsf{T}}x : (2.9d)\}$. To address the integral arising from the RHS of constraint (2.9d), we switch the objective function with the constraint to obtain

$$\rho(u) := \sup_{x\in X, y\geq 0} \left\{\, \phi(x,y) \colon c^{\mathsf{T}}x \leq u \,\right\}, \tag{2.10}$$

where u represents a budget limit on the (original) objective function. We notice that $\rho(u)$ is non-decreasing in u, and hence $u^* \in \mathbb{R}$ is the optimal value of (2.9a)–(2.9c) if and only if u^* is the smallest number such that $\rho(u^*)$ exceeds δ. It follows that we can solve (2.9a)–(2.9c) by searching for the intersection of the function $\rho(u)$ with the constant δ, which can be done by a bisection line search on u and iteratively solving (2.10) to evaluate $\rho(u)$. In addition, $\rho(u)$ may be interesting in its own right because it represents the largest Wasserstein radius δ that allows us to find a solution x that satisfies (**P-CC**) and incurs a cost no more than u. Hence, the graph of $\rho(u)$ depicts a risk envelope that interprets the trade-off between the robustness and the cost effectiveness of (**P-CC**). We demonstrate the risk envelope numerically in Section 2.5.1.

Evaluating $\rho(u)$ is equivalent to maximizing $\phi(x,y)$ over the intersection of

$$\left\{\, (x,y) \in X \times \mathbb{R}_+ : c^{\mathsf{T}}x \leq u \,\right\}$$

and **dom** ϕ. Unfortunately, projecting onto **dom** ϕ may be inefficient since it is the feasible region of (**CC**). To avoid projection, we propose a block coordinate ascent algorithm (Algorithm 1; see, e.g., [Aus76; LT93; Ber97; GS99; BT13; Bec15]). This algorithm iteratively maximizes over y with x fixed and then maximizes over x with y fixed. Here, for fixed x with $\mathbb{P}[A\zeta \leq b(x)] \geq 1-\epsilon$, i.e., when x satisfies (**CC**), the maximization over y admits a closed-form solution $y = \mathsf{VaR}_\epsilon\big(f(x,\zeta)\big)$, that is,

$$\max_{y\geq 0} \phi(x,y) = \phi\left(x, \mathsf{VaR}_\epsilon\Big(f(x,\zeta)\Big)\right)$$

because $\phi(x,y)$ is increasing in y on the interval $\big[0, \mathsf{VaR}_\epsilon\big(f(x,\zeta)\big)\big]$ and it becomes decreasing in y when $y > \mathsf{VaR}_\epsilon\big(f(x,\zeta)\big)$. On the other hand, for fixed y, we seek to maximize $\phi(x,y)$, which appears challenging as it is an integral. Fortunately, we can recast $\phi(x,y)$ as

$$\begin{aligned}
\phi(x,y) &= \int_0^y \mathbb{P}\left[f(x,\zeta) \geq t\right] \mathrm{d}t - y\cdot(1-\epsilon) \\
&= y\int_0^1 \mathbb{P}\left[f(x,\zeta) \geq sy\right] \mathrm{d}s - y\cdot(1-\epsilon) \\
&= y\int_{\Xi}\int_{\mathbb{R}} \mathbb{1}\left\{\, (\zeta,s)\colon f(x,\zeta) \geq sy \,\right\} \cdot \mathbb{1}_{[0,1]} s \,\mathrm{d}s \,\mathrm{d}\mathbb{P}(\zeta) - y\cdot(1-\epsilon), \\
&= y\int_{\Xi\times\mathbb{R}} \mathbb{1}\left\{\, (\zeta,s)\colon f(x,\zeta) \geq sy \,\right\} \mathrm{d}\widehat{\mathbb{P}}\zeta, s - y\cdot(1-\epsilon), \\
&= y\cdot\widehat{\mathbb{P}}\left[f(x,\zeta) \geq sy\right] - y\cdot(1-\epsilon),
\end{aligned}$$

where the third equality is by Tonelli's theorem and $\widehat{\mathbb{P}}$ represents the product measure of $\mathbb{P}$ and the uniform distribution on $[0,1]$. Since these two distributions are log-concave on Ξ and $[0,1]$, respectively, $\widehat{\mathbb{P}}$ is log-concave on $\Xi\times[0,1]$. As a result, the problem simplifies to the P-model of (**CC**) with respect to a log-concave distribution, which has been well studied in [Pré13; Nor93]. As a result, Algorithm 1 uses the existing solution approach as a building block and assumes that there exists an oracle, denoted by $\mathcal{O}_u(y,\varepsilon)$, which for given y and $\varepsilon > 0$ returns an ε-optimal solution $\widehat{x} \in \left\{\, x \in X : c^\mathsf{T} x \leq u \,\right\}$ such that

$$\widehat{\mathbb{P}}\left[f(\widehat{x},\zeta) \geq sy\right] \geq \max_{x\in X\colon\, c^\mathsf{T}x\leq u} \left\{\, \widehat{\mathbb{P}}\left[f(x,\zeta) \geq sy\right] \,\right\} - \varepsilon.$$

We are now ready to present Algorithm 1.

Algorithm 1 needs an starting point (x_1, y_1) such that $\mathsf{VaR}_\epsilon(f(x_1,\zeta)) > 0$. This can be

Algorithm 1: Evaluation of $\rho(u)$

Inputs : budget u, risk level ϵ, oracle $\mathcal{O}_u$, a diminishing sequence $\{\,\varepsilon_k\,\}_k$, and an x_1 such that $y_1 := \mathsf{VaR}_\epsilon\big(f(x_1,\zeta)\big) > 0$.

for $k = 1, 2, \ldots$ **do**
- $x_{k+1} \leftarrow \mathcal{O}_u(y_k, \varepsilon_k)$;
- $y_{k+1} \leftarrow \mathsf{VaR}_\epsilon\big(f(x_{k+1},\zeta)\big)$;
- **if** *stopping criterion is satisfied* **then**
  - **return** $\phi(x_{k+1}, y_{k+1})$.

obtained by solving a (**CC**) feasibility problem,

$$\min_{x\in X}\Big\{\, 0\colon \mathbb{P}\big[f(x,\zeta)\geq \varepsilon_0\big] \geq 1-\epsilon,\ c^{\mathsf{T}}x \leq u \,\Big\}, \tag{2.11}$$

where ε_0 is a small positive constant. If formulation (2.11) is infeasible for all $\varepsilon > 0$, then $\rho(u) = 0$ because $\mathsf{VaR}_\epsilon\big(f(x,\zeta)\big)$ always remains non-positive. Numerically, one can solve (2.11) for a sequence of diminishing ε_0's to find a valid starting point. We close this section by showing that Algorithm 1 achieves global optimum.

Theorem 3. Let $\{\,(x_k, y_k)\,\}_k$ be an infinite sequence of iterates produced by Algorithm 1. Suppose that $\mathbb{P}$ is log-concave and, for all $k \geq 2$, x_k and y_k are ε_k-optimal, i.e.,

$$\max_x \phi(x, y_{k-1}) - \varepsilon_k \leq \phi(x_k, y_{k-1}) \leq \max_x \phi(x, y_{k-1}) \quad \text{and} \quad \Big| y_k - \mathsf{VaR}_\epsilon\big(f(x_k,\zeta)\big)\Big| \leq \varepsilon_k$$

with $\lim_{k\to\infty}\varepsilon_k = 0$. Then, any limit point of $\{\,(x_k, y_k)\,\}_k$ is a global optimal solution to (2.10).

Proof. The proof relies on preparatory Lemmas 12, 13, and 14, whose proofs are provided in Appendix A.11.

First, we define set $S := \mathbf{dom}\,\phi \cap \{(x,y) \in X \times \mathbb{R}_+ : c^{\mathsf{T}}x \leq u\}$. Then, by compactness of X and closedness of $\mathbf{dom}\,\phi$ (see Proposition 4), S is compact. Since all iterates (x_k, y_k) lives in S (see Lemma 12), $\{(x_k, y_k)\}_k$ has a limit point $(x^*, y^*) \in S$.

Second, we show that (x^*, y^*) is a first-order local optimal solution to (2.10), which implies its global optimality due to the log-concavity of $\phi(x, y)$. To this end, let $\Delta := (d_x, d_y)$ be an arbitrary tangent direction of S at (x^*, y^*). Then, by definition there exists a sequence $\{(x_\ell, y_\ell)\}_\ell$ in S converging to (x^*, y^*) and $t_\ell \searrow 0$ such that

$$\Delta = \lim_{\ell\to\infty} \frac{(x_\ell, y_\ell) - (x^*, y^*)}{t_\ell}.$$

Then, we examine the directional derivative of $\phi(x, y)$ along direction Δ to obtain

$$\begin{aligned}\phi'(x^*, y^*; \Delta) &= \phi'\left(x^*, y^*; \lim_{\ell\to\infty} \frac{1}{t_\ell}\Big[(x_\ell, y_\ell) - (x^*, y^*)\Big]\right) \\ &= \lim_{\ell\to\infty} \phi'\left(x^*, y^*; \frac{1}{t_\ell}\Big[(x_\ell, y_\ell) - (x^*, y^*)\Big]\right) \\ &= \lim_{\ell\to\infty} \frac{1}{t_\ell}\phi'\Big(x^*, y^*; (x_\ell, y_\ell) - (x^*, y^*)\Big) \leq 0,\end{aligned}$$

where the second and third equalities follow from the continuity and positive homogeneity of $\phi'(x^*, y^*; \Delta)$ in Δ, respectively (see Lemma 14), and the inequality follows from Lemma 13 because $(x^*, y^*) + (x_\ell, y_\ell) - (x^*, y^*) = (x_\ell, y_\ell) \in S$. This completes the proof. □

Remark 2. The convergence of block alternating minimization algorithms for convex programs [Aus76; LT93; Ber97; GS99; BT13; Bec15] requires stronger sense of convexity [Aus76; LT93], a unique minimizer with respect to each block [Ber97], or a Lipschitz gradient [BT13; Bec15]. In contrast, our Algorithm 1 does not require the objective function ϕ to be differentiable, and it even allows the "half-iterates" (x_{k+1}, y_k) to leave **dom** ϕ. This is because all the major iterates (x_k, y_k) always lie in **dom** ϕ due to the special subproblem structure of (**P-CC**). In addition, if $\ln \phi(x, y)$ has a Lipschitz continuous gradient with respect to x, then it follows from Theorem 3.7 in [Bec15] that Algorithm 1 admits an $O(1/k)$ rate of convergence. In other words, it takes $O(1/\varepsilon)$ iterations for the algorithm to achieve an ε-optimal solution to (**P-CC**).

2.3.2 SOC Inner Approximation

We focus on the two-sided (**P-CC**) with LHS uncertainty and its feasible region $\mathcal{X}_{\mathsf{T}}^p$. Although Theorem 2 provides a convex representation of $\mathcal{X}_{\mathsf{T}}^p$, it is not computable (say, in a commercial solver) because the function $g_\epsilon(\ell, u)$ is defined through an integral. This section derives SOC inner approximations of $\mathcal{X}_{\mathsf{T}}^p$, which can be directly and efficiently computed by commercial solvers.

2.3.2.1 Inner Approximations for $\mathcal{C}_\delta$ and $\mathcal{X}_{\mathsf{T}}^p$

To illustrate the basic idea, we plot the boundary of $\mathcal{C}_\delta$, i.e., the contour of the function $g_\epsilon(\ell, u)$, in Figure 2.1. Since $\mathcal{C}_\delta$ is convex, we can obtain a polyhedral inner approximation using two extreme rays of $\mathcal{C}_\delta$ and a set of points on its boundary, denoted by **bd** $(\mathcal{C}_\delta)$ (see the dotted line in Figure 2.1 for an illustration of this inner approximation). We now formalize this idea.

Algorithm 2: Construction of $\widehat{\mathcal{C}}_N$

Inputs : $\epsilon \in (0, \frac{1}{2}), \delta > 0$, and a (small) error threshold $\tau > 0$.

Initialize the set of points $\texttt{PT} = \emptyset$.

Find a $u \leftarrow (\overline{g}_\epsilon)^{-1}(\delta + \tau)$ and an ℓ such that $g_\epsilon(\ell, u) = \delta$.

if $\ell + u > 0$ **then** Replace $(\ell, u) \leftarrow (-u, -\ell)$.

$\texttt{PT} \leftarrow \texttt{PT} \cup \{ (\ell, u) \}$.

while $\ell + u \leq 0$ **do**

  Find an (ℓ', u') on **bd** $(\mathcal{C}_\delta)$ such that

  1. $(\ell', u') \geq (\ell, u)$, and
  2. the line connecting (ℓ, u) and (ℓ', u') supports $\mathcal{C}_{\delta+\tau}$.

  $\texttt{PT} \leftarrow \texttt{PT} \cup \{ (\ell', u') \}$.

  $(\ell, u) \leftarrow (\ell', u')$.

for (ℓ, u) *in* PT **do**

  $\texttt{PT} \leftarrow \texttt{PT} \cup \{ (-u, -\ell) \}$.

Sort and label all points in PT from 1 to N such that $\ell_1 > \ell_2 > \cdots > \ell_N$.

return PT

Definitions 6. Given N points $\{ (\ell_1, u_1), \ldots, (\ell_N, u_N) \}$ on **bd** (C_δ) with $\ell_1 > \ell_2 > \cdots > \ell_N$, define a polyhedron $\widehat{\mathcal{C}}_N \subseteq \mathbb{R}_- \times \mathbb{R}_+$ described by

$$\ell \leq \ell_1 \tag{2.12a}$$

$$(u - u_i)(\ell_i - \ell_{i+1}) \geq (u_i - u_{i+1})(\ell - \ell_i), \quad \forall i \in [N-1] \tag{2.12b}$$

$$u \geq u_N. \tag{2.12c}$$

In this definition, inequality (2.12a) (resp. (2.12c)) is the vertical (resp. horizontal) ray emitting from (ℓ_1, u_1) (resp. (ℓ_N, u_N)) and inequalities (2.12b) are the line segments connecting two neighboring points (ℓ_i, u_i) and (ℓ_{i+1}, u_{i+1}). Then, $\widehat{\mathcal{C}}_N$ constructs an inner approximation for $\mathcal{X}_\mathsf{T}^p$.

Proposition 5. Given $\widehat{\mathcal{C}}_N$ in Definition 6, it holds that $\widehat{\mathcal{X}}_\mathsf{T}^p \subseteq \mathcal{X}_\mathsf{T}^p$, where

$$\widehat{\mathcal{X}}_\mathsf{T}^p := \left\{ (x, \ell, u) \in \mathbb{R}^{n+2} : \begin{array}{c} \exists s \in \mathbb{R} : \|x\|_* \leq s \\ \ell - x^\mathsf{T}\mu \leq \ell_1 s \\ u - x^\mathsf{T}\mu \geq u_N s \\ \left(\dfrac{\ell_i - \ell_{i+1}}{u_i - u_{i+1}} \right) \left(u - x^\mathsf{T}\mu - u_i s \right) \geq \ell - x^\mathsf{T}\mu - \ell_i s, \quad \forall i \in [N-1] \end{array} \right\}.$$

We highlight that $\widehat{\mathcal{X}}_{\mathsf{T}}^p$ is SOC and so can be efficiently computed by commercial solvers. In view that $\mathcal{C}_\delta$ and $g_\epsilon(\ell, u)$ are symmetric with respect to $\ell + u = 0$, we also construct $\widehat{\mathcal{C}}_N$ to be symmetric; that is, we pick the N points such that $\ell_i + u_{N-i+1} = 0$ for all $i \in [N]$. We propose Algorithm 2 to find the points and construct a symmetric $\widehat{\mathcal{C}}_N$. Specifically, Algorithm 2 receives a (small) error threshold $\tau > 0$ and returns a set of N points on **bd** $(\mathcal{C}_\delta)$ such that the ensuing inner polyhedral approximation $\widehat{\mathcal{C}}_N$ satisfies $g_\epsilon(\ell, u) \in [\delta, \delta + \tau]$ for all $(\ell, u) \in$ **bd** $(\widehat{\mathcal{C}}_N)$. In other words,

$$\mathcal{C}_\delta \subseteq \widehat{\mathcal{C}}_N \subseteq \mathcal{C}_{\delta+\tau}.$$

In Step 2 of Algorithm 2, we search for $u = (\overline{g}_\epsilon)^{-1}(\delta + \tau)$, which is defined in Corollary 3 and visualized in Figure 2.3b. This can be done by running a root-finding algorithm on the function $\overline{g}_\epsilon(\cdot)$, e.g., a bisection line search. The same approach can be applied in Step 2 to find an ℓ such that $g_\epsilon(\ell, u) = \delta$, and in Step 6 to find an (ℓ', u') such that the line connecting (ℓ, u) and (ℓ', u') supports $\mathcal{C}_{\delta+\tau}$. Since the domains of functions $\overline{g}_\epsilon(\cdot)$ and $g_\epsilon(\cdot, \cdot)$ are 1-dimensional and 2-dimensional, respectively, running a bisection line search on them is efficient. As a result, the runtime of Algorithm 2 is usually negligible, even when we choose a small error threshold τ (see Section 2.5.2.2 for a numerical demonstration).

2.3.2.2 Approximation Error and Asymptotic Exactness of $\widehat{\mathcal{X}}_{\mathsf{T}}^p$

We show that $\widehat{\mathcal{X}}_{\mathsf{T}}^p$ is asymptotically exact, i.e., $\widehat{\mathcal{X}}_{\mathsf{T}}^p$ asymptotically recovers $\mathcal{X}_{\mathsf{T}}^p$, as the error threshold τ in Algorithm 2 decreases to zero and the points $\{(\ell_i, u_i)\}_{i=1}^N$ become dense. More generally, we quantify the error of $\widehat{\mathcal{X}}_{\mathsf{T}}^p$ in approximating $\mathcal{X}_{\mathsf{T}}^p$ with a positive τ.

To this end, we study the error of $\widehat{\mathcal{C}}_N$ in approximating $\mathcal{C}_\delta$ and the same approximation guarantee carries over to $\widehat{\mathcal{X}}_{\mathsf{T}}^p$ by construction. We shall show that $\mathcal{C}_\delta$ becomes a subset of $\widehat{\mathcal{C}}_N$ if we slightly expand the latter. But since $0 \notin \mathcal{C}_\delta$, we need to define such expansion with respect to a new origin within $\mathcal{C}_\delta$, giving rise to the next definition.

Definitions 7. Given a set $C \subseteq \mathbb{R}^2$, a point $(\ell_0, u_0) \in C$, and a positive scalar $\gamma > 0$, define

$$\gamma \cdot C := (\ell_0, u_0) + \Big\{\gamma(\ell - \ell_0, u - u_0) : \ (\ell, u) \in C\Big\}.$$

The new origin (ℓ_0, u_0) partitions $\mathbb{R}^2$ into four (shifted) orthants $\{\, \mathcal{O}_i, i \in [4] \,\}$, and hence $\mathcal{C}_\delta$ and $\widehat{\mathcal{C}}_N$ into four disjoint subsets:

$$\mathcal{C}_\delta = \bigcup_{i=1}^4 \Big(\mathcal{C}_\delta \cap \mathcal{O}_i\Big), \quad \widehat{\mathcal{C}}_N = \bigcup_{i=1}^4 \Big(\widehat{\mathcal{C}}_N \cap \mathcal{O}_i\Big).$$

Our approach is to find constants γ_i such that $\mathcal{C}_\delta \cap \mathcal{O}_i \subseteq \gamma_i \cdot (\widehat{\mathcal{C}}_N \cap \mathcal{O}_i)$ for all $i \in [4]$. Then, $\max_{i\in[4]}\{\gamma_i\}$ gives the approximation error of $\widehat{\mathcal{C}}_N$. We present the main result of this section as follows.

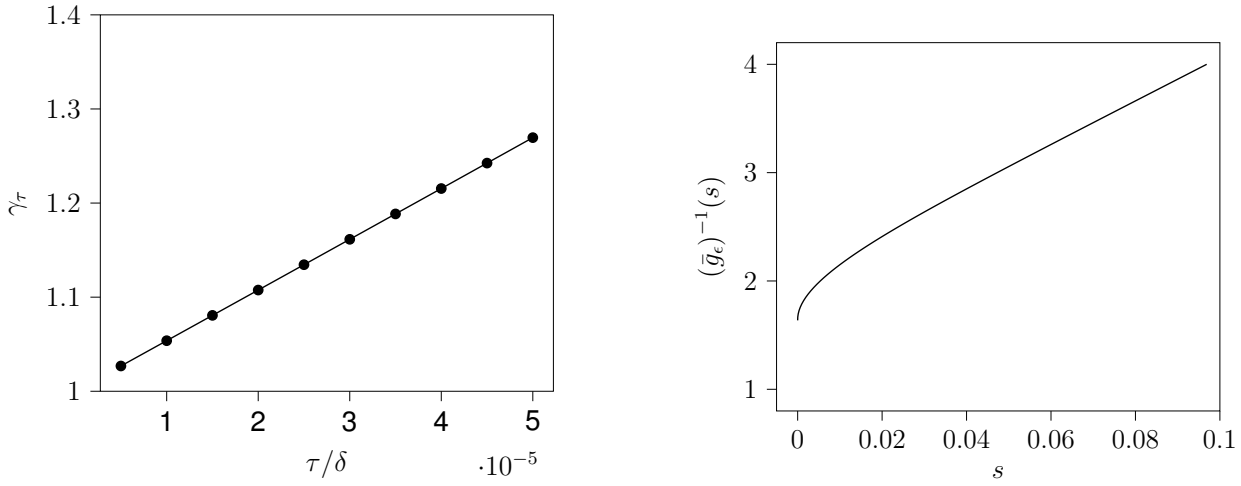
Theorem 4. Under Assumptions 1 and 2, suppose that $\epsilon \in (0, 1/2)$, $\mathbb{P}_0$ has a CDF Φ and a density function Φ', and $\widehat{\mathcal{C}}_N$ is obtained from Algorithm 2 with an error threshold $\tau > 0$. Then, for $(\ell_0, u_0) := (\ell_N, -\ell_N)$ it holds that

$$\widehat{\mathcal{C}}_N \subseteq \mathcal{C}_\delta \subseteq \gamma_\tau \cdot \widehat{\mathcal{C}}_N,$$

$$\text{where} \qquad \gamma_\tau := \max\left\{ \frac{u_0 - (\overline{g}_\epsilon)^{-1}(\delta)}{u_0 - (\overline{g}_\epsilon)^{-1}(\delta+\tau)}, \quad 1 + \frac{\sqrt{2}\ln\big((\delta+\tau)/\delta\big)}{\underline{D}\cdot\underline{\rho}} \right\}.$$

Here, $\underline{D} := \Phi'(u_0)/\big(\Phi(u_0) - \Phi(\ell_0) - (1-\epsilon)\big)$ and $\underline{\rho} := 1/\|(1,0)\|_{\mathcal{C}_{\delta+\tau}}$, where $\|\cdot\|_{\mathcal{C}_{\delta+\tau}}$ is defined through $\|x\|_{\mathcal{C}_{\delta+\tau}} := \inf\{\, r > 0 \colon x \in r \cdot \mathcal{C}_{\delta+\tau} \,\}$. Furthermore, $\gamma_\tau \to 1$ as $\tau \to 0$.

Proof. See Appendix A.12. □



(a) Approximation error bound γ_τ

(b) Plot of $(\bar{g}_\epsilon)^{-1}(s)$

Figure 2.3: Visualization of γ_τ and $(\bar{g}_\epsilon)^{-1}(s)$

We depict the approximation error bound γ_τ as a function of τ in Figure 2.3a, using a Gaussian $\mathbb{P}_0$. From this figure, we observe that γ_τ is close to 1 for small τ, suggesting that $\widehat{\mathcal{C}}_N$ is a tight inner approximation of $\mathcal{C}_\delta$. Accordingly, the ensuing $\widehat{\mathcal{X}}^p_{\mathsf{T}}$ is a tight inner approximation of $\mathcal{X}^p_{\mathsf{T}}$.

2.4 Optimistic Chance Constraint

This section extends the convexity results for (**P-CC**) in Section 2.2 to (**O-CC**). We first present a CVaR reformulation for $\mathcal{X}^o$ by adapting Theorem 1 in [Xie19]. Then, we study joint (**O-CC**) with RHS uncertainty and two-sided (**O-CC**) with LHS uncertainty in Sections 2.4.1 and 2.4.2, respectively.

Theorem 5. For $\delta > 0$ it holds that

$$\mathcal{X}^o = \left\{ x \in \mathbb{R}^n \colon \mathsf{CVaR}_\epsilon\Big(-\mathbf{d}\left(\zeta, \mathcal{S}(x)\right)\Big) + \frac{\delta}{1-\epsilon} \geq 0 \right\},$$

where the CVaR is with respect to the reference distribution $\mathbb{P}$ and $\mathbf{d}\left(\zeta, \mathcal{S}(x)\right)$ is the distance from $\zeta \in \mathbb{R}^m$ to the safe set $\mathcal{S}(x)$,

$$\mathbf{d}\left(\zeta, \mathcal{S}(x)\right) := \inf_{\xi \in \Xi} \left\{ \, \|\zeta - \xi\| \colon A(x)\,\xi \leq b(x) \, \right\}.$$

Proof. See Appendix A.13. □

2.4.1 Joint (O-CC) with RHS Uncertainty

When ξ arises from the RHS, we recall the formulation of (**O-CC**):

$$\mathcal{X}^o_{\mathsf{R}} := \left\{ x \in \mathbb{R}^n \colon \sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{Q}\Big[A\,\xi \leq b(x)\Big] \geq 1 - \epsilon \right\}.$$

As a preparation, we show that the distance $\mathbf{d}\left(\zeta, \mathcal{S}(x)\right)$ from $\zeta \in \mathbb{R}^m$ to the safe set $\mathcal{S}(x)$ is convex.

Lemma 5. $\mathbf{d}\left(\zeta, \mathcal{S}(x)\right) \equiv \min_{\xi \in \Xi} \Big\{ \|\xi - \zeta\| : A\,\xi \leq b(x) \Big\}$ is jointly convex in (ζ, x) on $\Xi \times \mathbb{R}^n$.

Proof. See Appendix A.14. □

Now we are ready to present the main result of this subsection.

Theorem 6. Suppose that the reference distribution $\mathbb{P}$ of $\mathcal{P}$ is α-concave with $0 \leq \alpha \leq 1/m$. Then, $\mathcal{X}^o_{\mathsf{R}}$ is convex and closed for $\delta > 0$.

Proof. By Theorem 5, (**O-CC**) admits the following reformulations:

$$\begin{aligned}
&\mathsf{CVaR}_\epsilon\Big\{ -\mathbf{d}\big(\zeta, \mathcal{S}(x)\big) \Big\} \geq -\frac{\delta}{1-\epsilon} \\
\iff &\inf_{\gamma\in\mathbb{R}} \left\{ \gamma + \frac{1}{1-\epsilon}\mathbb{E}_{\mathbb{P}}\left\{ \Big[-\mathbf{d}\big(\zeta, \mathcal{S}(x)\big) - \gamma \Big]^+ \right\} \right\} \geq -\frac{\delta}{1-\epsilon} \\
\iff &\gamma + \frac{1}{1-\epsilon}\mathbb{E}_{\mathbb{P}}\left\{ \Big[-\mathbf{d}\big(\zeta, \mathcal{S}(x)\big) - \gamma \Big]^+ \right\} \geq -\frac{\delta}{1-\epsilon} \qquad \forall \gamma \in \mathbb{R}.
\end{aligned}$$

In what follows, we prove that the LHS of the last reformulation is log-concave in x for any fixed γ. Since log-concave functions are quasi-concave and continuous (see Lemma 2.4 in [Nor93]), the convexity and closedness of $\mathcal{X}_{\mathsf{R}}^o$ follows from their preservation under intersection. To this end, we notice that

$$\mathbb{E}_{\mathbb{P}}\Big[\phi(\zeta, x)\Big] = \int_{\Xi} \phi(x,\zeta)\cdot f_\zeta(\zeta)\,\mathrm{d}\zeta,$$

where $\phi(x,\zeta) := \big[-\mathbf{d}\big(\zeta, \mathcal{S}(x)\big) - \gamma\big]^+$ and f_ζ represents the probability density function of ζ. It suffices to show that $\phi(x,\zeta)\cdot f_\zeta(\zeta)$ is jointly log-concave in (x,ζ) because log-concavity preserves under marginalization (see Theorem 3.3 in [SW14]). In view that log-concavity also preserves under multiplication, we complete the proof by showing that $f_\zeta(\zeta)$ is log-concave in ζ and $\phi(x,\zeta)$ is jointly log-concave in (x,ζ).

1. Since $\mathbb{P}$ is α-concave, its density function f_ζ is α'-concave by Proposition 23, where

$$\alpha' = \begin{cases} \frac{\alpha}{1-m\alpha} & \text{if } \alpha \in [0, 1/m) \\ +\infty & \text{if } \alpha = 1/m \end{cases}$$

   and $\alpha' \geq 0$. Hence, f_ζ is log-concave by Lemma 10.

2. For any pair of $(x_1,\zeta_1), (x_2,\zeta_2) \in \mathbb{R}^n \times \Xi$ and any $\theta \in [0,1]$, define $(x_\theta, \zeta_\theta) := \theta(x_1,\zeta_1) + (1-\theta)(x_2,\zeta_2)$. Then, it holds that

$$\begin{aligned}
\phi(x_\theta,\zeta_\theta) = \Big(-\mathbf{d}\big(\zeta_\theta, \mathcal{S}(x_\theta)\big) - \gamma\Big)^+ &\geq \Big(m_1\big(-\mathbf{d}\big(\zeta_1, \mathcal{S}(x_1)\big) - \gamma, -\mathbf{d}\big(\zeta_2, \mathcal{S}(x_2)\big) - \gamma; \theta\big)\Big)^+ \\
&\geq m_0\Big(\phi(x_1,\zeta_1), \phi(x_2,\zeta_2); \theta\Big),
\end{aligned}$$

   where the first inequality is because $\mathbf{d}\big(\zeta, \mathcal{S}(x)\big)$ is jointly convex in (x,ζ). To see the second inequality, we discuss the following two cases.

   (i) If either $\phi(x_1,\zeta_1)$ or $\phi(x_2,\zeta_2)$ equals zero, then $m_0\big(\phi(x_1,\zeta_1), \phi(x_2,\zeta_2); \theta\big)$ equals

zero by definition.

(ii) If both $\phi(x_1, \zeta_1)$ and $\phi(x_2, \zeta_2)$ are strictly positive, then

$$\begin{aligned}\Big(m_1\big(-\mathbf{d}\left(\zeta_1, \mathcal{S}(x_1)\right) - \gamma, -\mathbf{d}\left(\zeta_2, \mathcal{S}(x_2)\right) - \gamma; \theta\big)\Big)^+ &= m_1\Big(\phi(x_1, \zeta_1), \phi(x_2, \zeta_2); \theta\Big) \\ &\geq m_0\Big(\phi(x_1, \zeta_1), \phi(x_2, \zeta_2); \theta\Big),\end{aligned}$$

where the inequality follows from Lemma 10.

□

Although Theorem 6 pertains to (**O-CC**) with linear inequalities, the convexity result extends to (**O-CC**) with *quasi-concave* inequalities. We present this generalization in Appendix A.15.

2.4.2 Two-Sided (O-CC) with LHS Uncertainty

We extend the convexity result for two-sided (**P-CC**) in Section 2.2.2 to the optimistic setting. Specifically, define

$$\mathcal{X}_{\mathsf{T}}^o := \left\{ (x, \ell, u) \in \mathbb{R}^n \times \mathbb{R}_- \times \mathbb{R}_+ \colon \sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{Q}\left[\ell \leq x^{\mathsf{T}}\xi \leq u\right] \geq 1 - \epsilon \right\}.$$

When the Wasserstein ball $\mathcal{P}$ has an elliptical reference distribution $\mathbb{P}$, we can "project" $\mathcal{P}$ onto a Wasserstein ball for 1-dimensional distributions centered around $\mathbb{P}_0$, which is induced by $R \cdot e_1^{\mathsf{T}} U_n$ with $\mathbb{P} \overset{d}{=} R \cdot \Sigma^{1/2} U_n$ (see Remark 7 in Appendix A.1). This relates $\mathcal{X}_{\mathsf{T}}^o$ to a primitive set

$$\mathcal{X}_{\mathsf{T}_0}^o := \left\{ (\ell, u) \in \mathbb{R}_- \times \mathbb{R}_+ \colon \sup_{\mathbb{Q} \in \mathcal{P}_0} \mathbb{Q}\left[\ell \leq \zeta \leq u\right] \geq 1 - \epsilon \right\},$$

where $\mathcal{P}_0$ is a Wasserstein ball centered around $\mathbb{P}_0$ with the same radius δ as in $\mathcal{P}$.

Lemma 6. Suppose that Assumption 1 holds and $\epsilon \in (0, 1/2)$. Then, for any $x \neq 0$, $(x, \ell, u) \in \mathcal{X}_{\mathsf{T}}^o$ if and only if $\left(\frac{\ell}{\|x\|_*}, \frac{u}{\|x\|_*}\right) \in \mathcal{X}_{\mathsf{T}_0}^o$.

Proof. Theorem 5 yields

$$\mathcal{X}_{\mathsf{T}}^o = \left\{ x \in \mathbb{R}^n \colon \mathsf{CVaR}_\epsilon\left(-\mathbf{d}\left(\xi, \mathcal{S}(x)\right)\right) + \frac{\delta}{1-\epsilon} \geq 0 \right\},$$

where the distance to the safe set $\mathcal{S}(x)$ is

$$\mathbf{d}\left(\xi, \mathcal{S}(x)\right) = \inf_{\eta}\left\{ \|\eta - \xi\| \colon \ell \le x^{\mathsf{T}}\eta \le u \right\} = \frac{1}{\|x\|_*}\left[f(\ell, u, \xi)\right]^+$$

and $f(\ell, u, \xi) := (x^{\mathsf{T}}\xi - u) \vee (\ell - x^{\mathsf{T}}\xi)$. Similarly, we recast $\mathcal{X}^o_{\mathsf{T}_0}$ as

$$\mathcal{X}^o_{\mathsf{T}_0} = \left\{ (\ell, u) \in \mathbb{R}_- \times \mathbb{R}_+ \colon \mathsf{CVaR}_\epsilon(-(f_0(\ell, u, \zeta))^+) + \frac{\delta}{1-\epsilon} \ge 0 \right\},$$

where $f_0(\ell, u, \zeta) := (\zeta - u) \vee (\ell - \zeta)$ and the CVaR is with respect to $\mathbb{P}_0$. But for $\zeta \sim \mathbb{P}_0$ and $\xi \sim \mathbb{P}$, we have $\zeta \overset{d}{=} x^{\mathsf{T}}\xi/\|x\|_*$ and so

$$\frac{1}{\|x\|_*}\left[f(\ell, u, \xi)\right]^+ = \max\left\{\frac{x^{\mathsf{T}}\xi - u}{\|x\|_*}, \frac{\ell - x^{\mathsf{T}}\xi}{\|x\|_*}, 0\right\} \overset{d}{=} \max\left\{\zeta - \frac{u}{\|x\|_*}, \frac{\ell}{\|x\|_*} - \zeta, 0\right\} = (f_0(\ell, u, \zeta))^+.$$

The conclusion follows. □

We are now ready to present the convexity result for $\mathcal{X}^o_{\mathsf{T}}$.

Theorem 7. Suppose that Assumptions 1 and 2 hold, $\epsilon \in (0, \frac{1}{2})$, and $\delta > 0$. Define

$$h_\epsilon(\ell, u) := \int_0^{+\infty} \Big[(1-\epsilon) - (\Phi(u+t) - \Phi(\ell - t))\Big]^+ \mathrm{d}t.$$

Then, it holds that $\mathcal{X}^o_{\mathsf{T}_0} = \left\{ (\ell, u) \in \mathbb{R}_- \times \mathbb{R}_+ \colon h_\epsilon(\ell, u) \le \delta \right\}$ and

$$\mathcal{X}^o_{\mathsf{T}} = \left\{ (x, \ell, u) \in \mathbb{R}^n \times \mathbb{R}_- \times \mathbb{R}_+ \colon \exists\, s \ge 0 \text{ such that } \|x\|_* \le s,\ (\ell, u, s) \in \mathsf{co}\left(\mathcal{X}^o_{\mathsf{T}_0}\right) \right\},$$

where $\mathsf{co}\left(\mathcal{X}^o_{\mathsf{T}_0}\right) := \mathbf{cl}\left(\left\{ (\ell, u, s) \in \mathbb{R}_- \times \mathbb{R}_+ \times \mathbb{R} \colon s > 0, (\ell/s, u/s) \in \mathcal{X}^o_{\mathsf{T}_0} \right\}\right)$ represents the cone generated by $\mathcal{X}^o_{\mathsf{T}_0}$. Furthermore, $\mathcal{X}^o_{\mathsf{T}}$ is convex and closed.

Proof. We first show the reformulation of $\mathcal{X}^o_{\mathsf{T}_0}$. By Theorem 5 and Lemma 1, we recast $\mathcal{X}^o_{\mathsf{T}_0}$ as

$$\frac{\delta}{1-\epsilon} + \mathbb{1}\left\{ 0 \ge \mathsf{VaR}_\epsilon(-f_0(\ell, u, \zeta)) \right\} \cdot \left(\mathsf{CVaR}_\epsilon(-f_0(\ell, u, \zeta)) - \frac{1}{1-\epsilon}\mathbb{E}\left[(-f_0(\ell, u, \zeta))^+\right]\right) \ge 0,$$

where $f_0(\ell, u, \zeta) \equiv (\zeta - u) \vee (\ell - \zeta)$ and $\zeta \sim \mathbb{P}_0$. Then, we break down the indicator function

to obtain

$$\mathcal{X}^o_{\mathsf{T}_0} = \mathcal{X}^o_{\mathsf{T}_1} \cup \big((\mathbb{R}_- \times \mathbb{R}_+ \setminus \mathcal{X}^o_{\mathsf{T}_1}) \cap \mathcal{X}^o_{\mathsf{T}_2}\big),$$

$$\text{where} \quad \mathcal{X}^o_{\mathsf{T}_1} = \left\{ (\ell, u) \in \mathbb{R}_- \times \mathbb{R}_+ \colon 0 < \mathsf{VaR}_\epsilon(-f_0(\ell, u, \zeta)) \right\},$$

$$\text{and} \quad \mathcal{X}^o_{\mathsf{T}_2} = \left\{ (\ell, u) \in \mathbb{R}_- \times \mathbb{R}_+ \colon \frac{\delta}{1-\epsilon} + \mathsf{CVaR}_\epsilon(-f_0(\ell, u, \zeta)) \geq \frac{1}{1-\epsilon}\mathbb{E}\left[(-f_0(\ell, u, \zeta))^+\right] \right\}.$$

For $\mathcal{X}^o_{\mathsf{T}_1}$, we have

$$\begin{aligned} 0 < \mathsf{VaR}_\epsilon(-f_0(\ell, u, \zeta)) &\iff \mathbb{P}_0\left[-f_0(\ell, u, \zeta) \leq 0\right] < \epsilon \\ &\iff \mathbb{P}_0\left[\zeta \geq u \text{ or } \ell \geq \zeta\right] < \epsilon \\ &\iff \mathbb{P}_0\left[\ell \leq \zeta \leq u\right] > 1 - \epsilon, \\ \text{or equivalently:} \quad 0 \geq \mathsf{VaR}_\epsilon(-f_0(\ell, u, \zeta)) &\iff \mathbb{P}_0\left[\ell \leq \zeta \leq u\right] \leq (1 - \epsilon). \end{aligned}$$

For $((\mathbb{R}_- \times \mathbb{R}_+) \setminus \mathcal{X}^o_{T_1}) \cap \mathcal{X}^o_{\mathsf{T}_2}$, we have

$$\begin{aligned} & \delta + (1-\epsilon)\mathsf{CVaR}_\epsilon(-f_0(\ell, u, \zeta)) \geq \mathbb{E}\left[(-f_0(\ell, u, \zeta))^+\right] \\ \iff & \delta + \mathbb{E}\left[-f_0(\ell, u, \zeta) \cdot \mathbb{1}\left\{ -f_0(\ell, u, \zeta) \geq \mathsf{VaR}_\epsilon(-f_0(\ell, u, \zeta)) \right\}\right] \\ & \geq \mathbb{E}\left[-f_0(\ell, u, \zeta) \cdot \mathbb{1}\left\{ -f_0(\ell, u, \zeta) \geq 0 \right\}\right] \\ \iff & \delta + \mathbb{E}\left[-f_0(\ell, u, \zeta) \cdot \mathbb{1}\left\{ \mathsf{VaR}_\epsilon(-f_0(\ell, u, \zeta)) \leq -f_0(\ell, u, \zeta) \leq 0 \right\}\right] \geq 0 \\ \iff & \delta \geq \mathbb{E}\left[f_0(\ell, u, \zeta) \cdot \mathbb{1}\left\{ 0 \leq f_0(\ell, u, \zeta) \leq \mathsf{VaR}_{1-\epsilon}(f_0(\ell, u, \zeta)) \right\}\right]. \end{aligned}$$

Plugging the definition of $f_0(\ell, u, \zeta)$ into the RHS yields

$$\begin{aligned} & \mathbb{E}\left[f_0(\ell, u, \zeta) \cdot \mathbb{1}\left\{ 0 \leq f_0(\ell, u, \zeta) \leq \mathsf{VaR}_{1-\epsilon}(f_0(\ell, u, \zeta)) \right\}\right] \\ = & \int_\Xi (\zeta - u) \vee (\ell - \zeta) \cdot \mathbb{1}\left\{ 0 \leq (\zeta - u) \vee (\ell - \zeta) \leq \mathsf{VaR}_{1-\epsilon}(f_0(\ell, u, \zeta)) \right\} \mathrm{d}\mathbb{P}(\zeta) \\ = & \int_\Xi \int_0^{+\infty} \mathbb{1}\left\{ t \leq (\zeta - u) \vee (\ell - \zeta) \leq \mathsf{VaR}_{1-\epsilon}(f_0(\ell, u, \zeta)) \right\} \mathrm{d}t \, \mathrm{d}\mathbb{P}(\zeta) \\ = & \int_0^{+\infty} \mathbb{P}\left[t \leq (\zeta - u) \vee (\ell - \zeta) \leq \mathsf{VaR}_{1-\epsilon}((\zeta - u) \vee (\ell - \zeta))\right] \mathrm{d}t \\ = & \int_0^{+\infty} \Big((1-\epsilon) - \mathbb{P}\left[(\zeta - u) \vee (\ell - \zeta) \leq t\right]\Big)^+ \mathrm{d}t = \int_0^{+\infty} \big((1-\epsilon) - \mathbb{P}\left[\ell - t \leq \zeta \leq u + t\right]\big)^+ \mathrm{d}t. \end{aligned}$$

Therefore,

$$\left((\mathbb{R}_- \times \mathbb{R}_+) \setminus \mathcal{X}^o_{T_1}\right) \cap \mathcal{X}^o_{\mathsf{T}_2} = \left\{ (\ell, u) \in \mathbb{R}_- \times \mathbb{R}_+ : \begin{array}{c} 0 \geq \mathsf{VaR}_\epsilon\left(-f_0(\ell, u, \zeta)\right) \\ \delta \geq \int_0^{+\infty} \left((1-\epsilon) - \mathbb{P}\left[\ell - t \leq \zeta \leq u + t\right]\right)^+ \mathrm{d}t. \end{array} \right\}$$

It follows that

$$\begin{aligned}
\mathcal{X}^o_{\mathsf{T}_0} &= \mathcal{X}^o_{\mathsf{T}_1} \cup \left((\mathbb{R}_- \times \mathbb{R}_+ \setminus \mathcal{X}^o_{\mathsf{T}_1}) \cap \mathcal{X}^o_{\mathsf{T}_2}\right) \\
&= \left\{ (\ell, u) \in \mathbb{R}_- \times \mathbb{R}_+ : 0 < \mathsf{VaR}_\epsilon(-f_0(\ell, u, \zeta)) \right\} \\
&\quad \cup \left\{ (\ell, u) \in \mathbb{R}_- \times \mathbb{R}_+ : \begin{array}{c} 0 \geq \mathsf{VaR}_\epsilon\left(-f_0(\ell, u, \zeta)\right) \\ \delta \geq \int_0^{+\infty} \left((1-\epsilon) - \mathbb{P}_0\left[\ell - t \leq \zeta \leq u + t\right]\right)^+ \mathrm{d}t \end{array} \right\} \\
&= (\mathbb{R}_- \times \mathbb{R}_+) \cap \left\{ (\ell, u) \in \mathbb{R}_- \times \mathbb{R}_+ : \begin{array}{c} \mathbb{P}_0\left[\ell \leq \zeta \leq u\right] > (1-\epsilon), \text{ or} \\ \delta \geq \int_0^{+\infty} \left((1-\epsilon) - \mathbb{P}_0\left[\ell - t \leq \zeta \leq u + t\right]\right)^+ \mathrm{d}t \end{array} \right\} \\
&= \left\{ (\ell, u) \in \mathbb{R}_- \times \mathbb{R}_+ : \delta \geq \int_0^{+\infty} \left((1-\epsilon) - \mathbb{P}_0\left[\ell - t \leq \zeta \leq u + t\right]\right)^+ \mathrm{d}t \right\} \\
&= \left\{ (\ell, u) \in \mathbb{R}_- \times \mathbb{R}_+ : \delta \geq h_\epsilon(\ell, u) \right\}
\end{aligned}$$

where the second to the last equality is because for all $(\ell_1, u_1) \in \mathbb{R}_- \times \mathbb{R}_+$ such that $\mathbb{P}_0\left[\ell_1 \leq \zeta \leq u_1\right] \geq (1-\epsilon)$, we have

$$(1-\epsilon) - \mathbb{P}_0\left[\ell_1 - t \leq \zeta \leq u + t\right] \leq (1-\epsilon) - \mathbb{P}_0\left[\ell_1 \leq \zeta \leq u\right] \leq 0, \quad \forall t \geq 0,$$

implying

$$\int_0^{+\infty} \left((1-\epsilon) - \mathbb{P}_0\left[\ell_1 - t \leq \zeta \leq u_1 + t\right]\right)^+ \mathrm{d}t = 0 \leq \delta.$$

Second, we show that $\mathcal{X}^o_{\mathsf{T}_0}$ is convex. By assumption, $\mathbb{P}_0 \overset{d}{=} R \cdot e_1^\mathsf{T} U_n$ is unimodal on $\mathbb{R}$ and its distribution function Φ is concave on $(0, +\infty)$ and convex on $(-\infty, 0)$. Hence,

$\Phi(u+t)-\Phi(\ell-t)$ is jointly concave in (ℓ,u,t) on $\mathbb{R}_-\times\mathbb{R}_+^2$. It follows that the integrand of h_ϵ is jointly convex in (ℓ,u,t), and so h_ϵ is convex in $(\ell,u)\in\mathbb{R}_-\times\mathbb{R}_+$ because partial integration of a convex function preserves its convexity.

Finally, to prove that $\mathcal{X}_{\mathsf{T}}^o$ is convex, it remains to show that $(x,\ell,u)\in\mathcal{X}_{\mathsf{T}}^o$ if and only if there exists an $s\geq\|x\|_*$ such that $(\ell,u,s)\in\text{co}\left(\mathcal{X}_{\mathsf{T}_0}^o\right)$. To this end, we discuss the following two cases:

1. $x=0$: Suppose that $(0,\ell,u)\in\mathcal{X}_{\mathsf{T}_0}^o$, then $\ell\leq 0\leq u$ and for $s_n:=1/n$ we have

$$h_\epsilon(\ell/s_n,u/s_n)=\int_0^{+\infty}\Big[(1-\epsilon)-(\Phi(n\cdot u+t)-\Phi(n\cdot\ell-t))\Big]^+\,\mathrm{d}t\to 0\text{ as }n\to\infty.$$

   Therefore, there exists an n such that $(\ell,u,1/n)\in\mathcal{X}_{\mathsf{T}_0}^o$. On the contrary, if there exists an $s>0$ such that $(\ell,u,s)\in\text{co}(\mathcal{X}_{\mathsf{T}_0}^o)$, then it is the limit point of a sequence $\{(\ell_n,u_n,s_n)\}_{n=1}^{+\infty}\subseteq\mathcal{X}_{\mathsf{T}_0}^o$ satisfying $\ell_n\leq u_n$ for all n. Then, $\ell\leq u$ as well, implying that $(0,\ell,u)\in\mathcal{X}_{\mathsf{T}}^o$.

2. $x\neq 0$: Suppose that $(x,\ell,u)\in\mathcal{X}_{\mathsf{T}}^o$. Then, Lemma 6 implies that $\left(\frac{\ell}{\|x\|_*},\frac{u}{\|x\|_*}\right)\in\mathcal{X}_{\mathsf{T}_0}^o$, i.e., $(\ell,u,\|x\|_*)\in\text{co}\left(\mathcal{X}_{\mathsf{T}_0}^o\right)$. On the contrary, suppose that there exists an $s\geq\|x\|_*>0$ such that $(\ell,u,s)\in\text{co}\left(\mathcal{X}_{\mathsf{T}_0}^o\right)$, then it is the limit point of a sequence $\{(\ell_n,u_n,s_n)\}_{n=1}^{+\infty}\subseteq\mathcal{X}_{\mathsf{T}_0}^o$. Observe that

$$h_\epsilon\left(\frac{\ell}{\|x\|_*},\frac{u}{\|x\|_*}\right)\leq h_\epsilon\left(\frac{\ell}{s},\frac{u}{s}\right)=\lim_{n\to\infty}h_\epsilon\left(\frac{\ell_n}{s_n},\frac{u_n}{s_n}\right)\leq\delta,$$

   where the first inequality is because $h_\epsilon(\ell,u)$ is increasing in ℓ and decreasing in u, and the equality is due to the continuity of h_ϵ. Therefore, $\left(\frac{\ell}{\|x\|_*},\frac{u}{\|x\|_*}\right)\in\mathcal{X}_{\mathsf{T}_0}^o$ and so $(x,\ell,u)\in\mathcal{X}_{\mathsf{T}}^o$ by Lemma 6.

□

2.5 Numerical Experiments

We demonstrate the theoretical results through two numerical experiments: a (**PP**) model using joint (**P-CC**) in Section 2.5.1 and an (**LR**) model using two-sided (**P-CC**) in Section 2.5.2.

2.5.1 Production Planning

We consider a (**PP**) model that seeks to procure production capacity so that all demands can be satisfied with high probability and a minimal procurement cost (see Example 1). Specifically, we consider the following formulation with (**P-CC**):

$$\begin{aligned} &\min\ c^{\mathsf{T}}x, \\ &\text{s.t.}\ \inf_{\mathbb{Q}\in\mathcal{P}} \mathbb{Q}\left[Tx \geq \xi\right] \geq 1-\epsilon, \\ &\qquad 0 \leq x_i \leq U, \forall i \in [n], \end{aligned}$$

where c represents the procurement costs, U represents a homogeneous upper bound of production capacity for all facilities, and the reference distribution $\mathbb{P}$ of the Wasserstein ball $\mathcal{P}$ is assumed to be pairwise independent and Gaussian. To apply Algorithm 1, we switch the objective function with (**P-CC**) to obtain

$$\begin{aligned} \rho(u) = \max_{x\in\mathbb{R}^n_+, y\in\mathbb{R}_+} \phi(x,y) \equiv \int_0^y \left(\mathbb{P}\left[\min_{t\in[m]}(T_i x - \zeta_i) \geq t\right] - (1-\epsilon)\right) \mathrm{d}t \\ \text{s.t.}\ c^{\mathsf{T}}x \leq u, \\ 0 \leq x_i \leq U, \quad \forall i \in [n], \end{aligned}$$

where we adjust the procurement budget u and apply the algorithm with various u to obtain a risk envelope. In addition, when applying Algorithm 1, we employ the stochastic approach described in [Nor93] to be the oracle $\mathcal{O}_u(y_k, \varepsilon_k)$ in Step 5 and terminate the algorithm whenever the change in y_k becomes sufficiently small, specifically, when $|y_k - y_{k+1}| \leq 10^{-6}$.

We demonstrate the convergence of Algorithm 1 in Figure 2.4, which is obtained by running the algorithm for five times on an instance with $n = 10$, $m = 5$, $U = 200$, c randomly drawn from the set $\{1, \ldots, 10\}$, and $\mathbb{E}_{\mathbb{P}}[\zeta_i]$ randomly drawn from the interval $[10, 51]$. In this figure, the solid line represents the difference between each iterate $\phi(x_k, y_k)$ and the final iterate ϕ^*, averaged across the five runs, and the error bar represents the standard deviation of the difference. From this figure, we observe that Algorithm 1 converges at a linear rate in only a few iterations.

We demonstrate the trade-off between the robustness and the budget in Figure 2.5, which is obtained by solving instances with $\epsilon \in \{0.1, 0.15, 0.2\}$, $n \in \{10, 30\}$, and $m \in \{5, 20\}$. The vertical axis of this figure represents $\rho(u)$, i.e., the largest Wasserstein radius δ that allows (**P-CC**) to be satisfied. From this figure, we observe that, for fixed ϵ, the largest

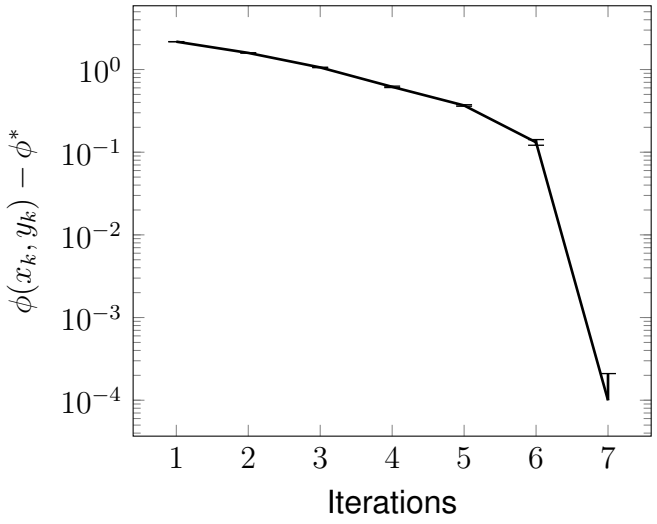
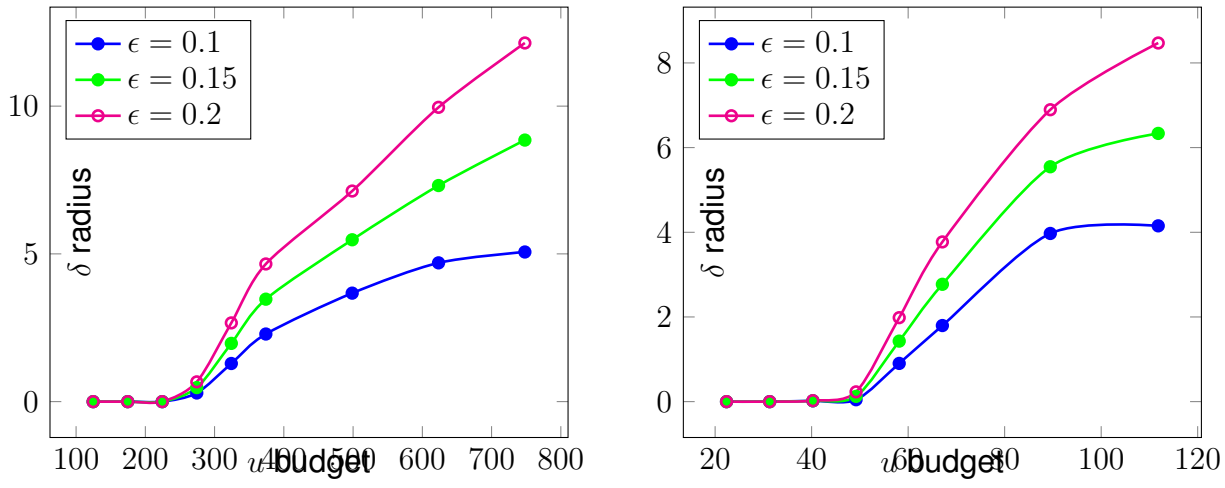


Figure 2.4: Convergence of Algorithm 1 on Production Planning Instances; solid line = average of the difference $\phi(x_k, y_k) - \phi^*$ across five runs, error bar = standard deviation of the difference



(a) $n = 10$ and $m = 5$

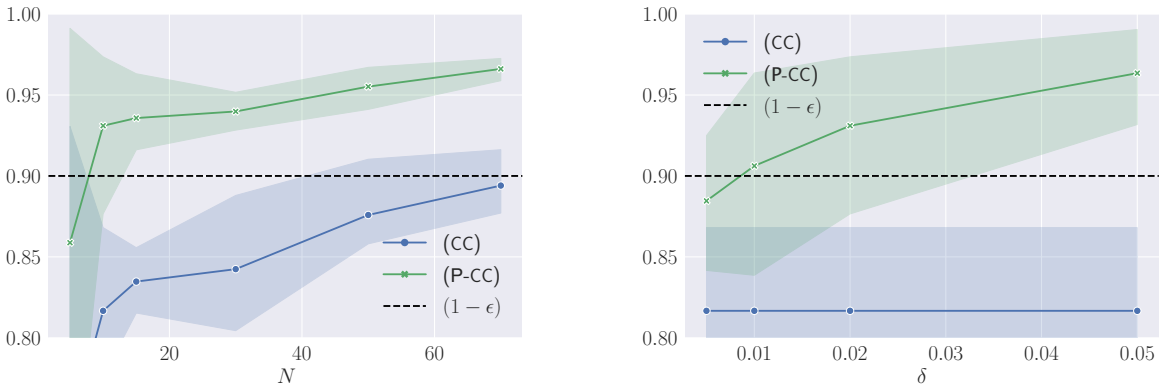
(b) $n = 30$ and $m = 20$

Figure 2.5: Risk envelopes under different risk thresholds

allowable δ is an S-shaped function of the the budget u. That is, δ remains at zero for small u, and then δ increases with a diminishing momentum as u becomes larger. In addition, for fixed δ, it needs a larger budget u to keep (**P-CC**) satisfied as ϵ decreases.

2.5.1.1 Out-of-Sample Performance

We compare the out-of-sample performance of (**P-CC**) and (**CC**) on production planning problems with $\delta \in \{0.005, 0.01, 0.02, 0.05\}$, $N \in \{5, 10, 15, 30, 50, 70\}$, $\epsilon = 0.1$, $m = 5$, and $n = 10$. Specifically, we assume that $\mathbb{P}_{\text{true}}$ is a multivariate Gaussian distribution with known mean and covariance and draw a set of N training data samples from $\mathbb{P}_{\text{true}}$. Then, we center $\mathcal{P}$ around the Gaussian distribution with empirical mean and covariance matrix, which we estimate from the training data. After obtaining the optimal solutions to (**P-CC**) and (**CC**) models, we compare their out-of-sample performance by the probability of fully satisfying the demands with respect to $\mathbb{P}_{\text{true}}$ and report the results in Figure 2.6. In this figure, the solid line and dots are the average out-of-sample performance across three randomly generated instances and the shaded region around them are the 95% confidence interval. From Figure 2.6a, we observe that the out-of-sample performance of both models



(a) Radius $\delta = 0.02$

(b) Data size $N = 10$

Figure 2.6: Out-of-sample performance of (**P-CC**) and (**CC**) on the production planning problem

improve as we obtain more training data. However, there is a significant difference in their sensitivity to having more data: (**P-CC**) achieves the target reliability level of 90% with only $N = 10$ data, whereas (**CC**) fails to achieve the same level even when 70 samples are provided. In addition, we observe from Figure 2.6b that, with as few as $N = 10$ data, the reliability of (**P-CC**) quickly increases and achieves the target reliability level as soon

as δ exceeds 0.01. Intuitively, a small degree of pessimism suffices to improve reliability drastically.

2.5.2 Robust Linear Regression

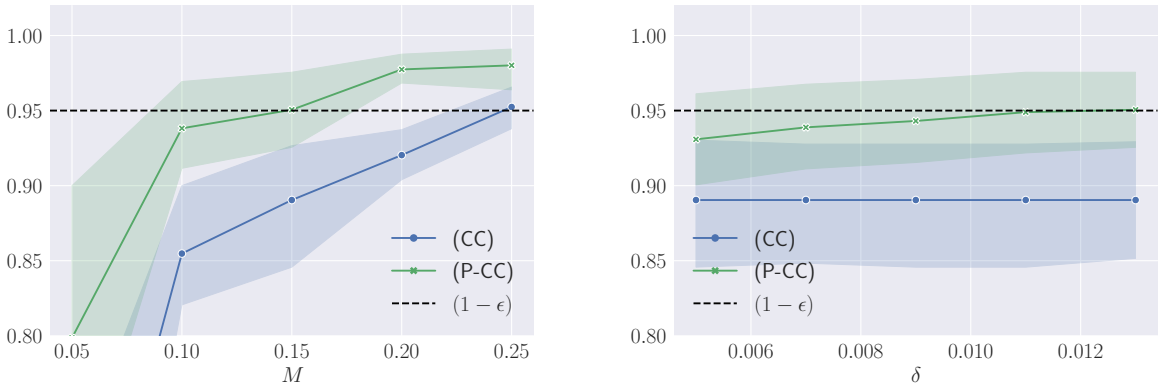
We consider the following (**LR**) model that controls the chance of having large residuals in linear regression (see Example 2):

$$\begin{aligned} \min_{\gamma,\beta,\beta_0} \ & \gamma, \\ \text{s.t.} \ & \inf_{\mathbb{Q}\in\mathcal{P}} \mathbb{Q}\left[-\gamma \le y - X^\mathsf{T}\beta - \beta_0 \le \gamma\right] \ge 1-\epsilon, \\ & \gamma \ge 0, \beta \in \mathbb{R}^n, \beta_0 \in \mathbb{R}. \end{aligned}$$

We compute the above two-sided (**P-CC**) using the inner approximations through Algorithm 2. To demonstrate the out-of-sample performance of (**P-CC**) and scalability of the solution approach, we use the Diabetes dataset, which consists of 442 patients' measurements of ten baseline variables (including age, gender, body mass index, average blood pressure, etc.) and their quantitative measurements of disease progression. We randomly split the 442 data samples into two subsets: (i) the training set consisting of $442 \cdot M$ data with $M \in (0,1)$ and (ii) the testing set consisting of the remaining data. For both (**P-CC**) and (**CC**), the latter of which ignores ambiguity in the two-sided chance constraint, we obtain an optimal solution (β, β_0, γ) by using the training data only. Then, we expose these solutions to the testing data and evaluate their out-of-sample performance, which is the empirical probability of having the residual not exceeding γ. We repeat this procedure for risk level $\epsilon = 0.05$, training data size $M \in \{\,0.05, 0.1, 0.15, 0.2, 0.25\,\}$, and radius $\delta \in \{\,0.005, 0.007, 0.009, 0.011, 0.013\,\}$.

2.5.2.1 Out-of-Sample Performance

We report the out-of-sample performance of (**P-CC**) and (**CC**) in Figure 2.7, where the solid lines and dots are the average value over five runs and the shaded regions represent the associated 95% confidence intervals. In particular, Figure 2.7a depicts the out-of-sample performance as the training data size M increases. From this figure, we observe that the out-of-sample performance of both models improve as we have more training data, but there is a significant difference in the effectiveness of using the data. For example, (**P-CC**) is able to achieve the target reliability of 95% using only 15% of the data on hand, while (**CC**) needs 25% of the data to achieve the same target. This demonstrates



(a) Radius $\delta = 0.013$

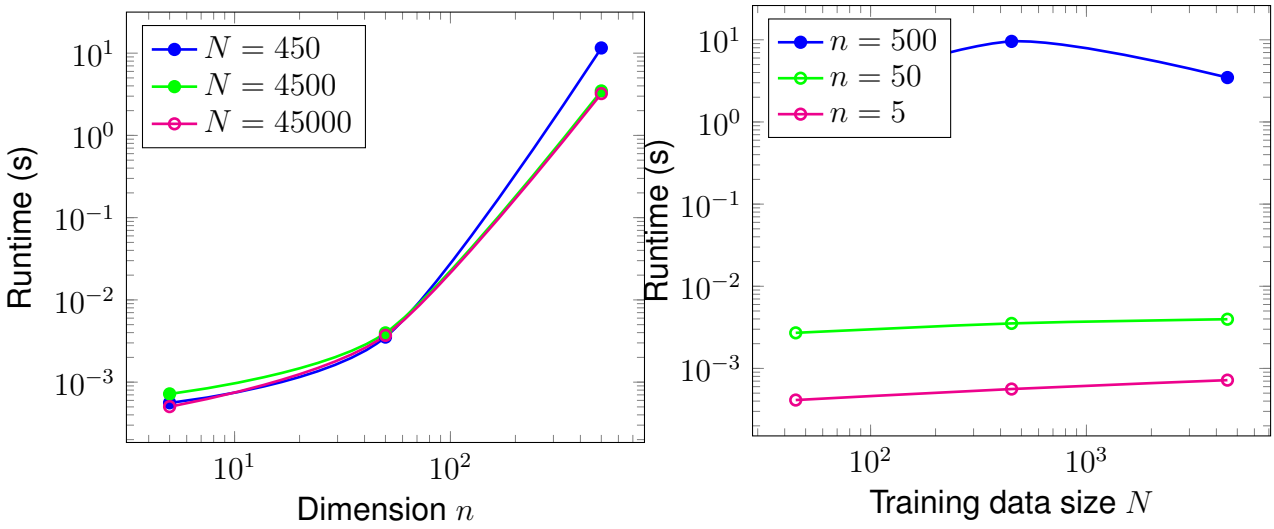
(b) Training data size $M = 0.15$

Figure 2.7: Out-of-sample performance of (**P-CC**) and (**CC**) on robust linear regression

that the (**P-CC**) model can utilize the data more effectively in a data-driven context. In addition, Figure 2.7b depicts the improvement of the (**P-CC**) out-of-sample performance as we increase the radius δ of the Wasserstein ball. From this figure, we observe that the reliability of (**P-CC**) achieves the target level of 95% as soon as δ exceeds 0.01.

2.5.2.2 Scalability of the Inner Approximation Approach

In order to demonstrate the scalability of the inner approximation approach for solving two-sided (**P-CC**), we employ the popular scikit-learn package (specifically, the function `make_regression()`) to generate synthetic regression data samples $\{ (X_i, y_i) \}_{i=1}^{N}$ with problem size $n \in \{ 5, 50, 500 \}$ and training data size $N \in \{ 450, 4500, 45000 \}$. Then, we solve the corresponding (**LR**) models and report the runtime as a function of n and N in Figures 2.8a–2.8b, respectively. In particular, from Figure 2.8a, we observe that the runtime increases polynomially as the problem size n increases, and the increasing trend is independent of the training data size N. In addition, all test instances are solved within 15 seconds, even when we increase n to 500 and N to $45,000$. From Figure 2.8b, we observe that the runtime is insensitive of the training data size N. These demonstrate the scalability of our inner approximation approach for solving two-sided (**P-CC**), especially with a large volume of training data.



(a) Scalability in problem size

(b) Scalability in training data size

Figure 2.8: Runtime (in seconds) under different problem and training data sizes

2.6 Conclusion

In this chapter, we uncover a family of sufficient conditions under which (**P-CC**)s and (**O-CC**)s admit convex feasible regions, making it possible to solve such problems efficiently. In addition, we design a coordinate ascent algorithm that could find global optimal solutions of joint (**P-CC**)s and demonstrate the effectiveness and efficiency of our proposed approaches on production planning problems and robust linear regression problems.

CHAPTER 3

Wasserstein Robust Chance-Constrained Set Covering

3.1 Problem Description and Literature Review

In this chapter, we consider a set covering model, in which one selects a subset of n elements to cover I targets with a minimum cost. If we employ $x \in \{0,1\}^n$ to denote a binary decision vector such that, for all $j \in [n] := \{1,\ldots,n\}$, $x_j = 1$ if element j is selected and $x_j = 0$ otherwise, then we can formulate such coverage requirement as a set of linear inequalities:

$$x^{\mathsf{T}}\xi_i \geq 1, \quad \forall i \in [I], \tag{3.1}$$

where $\xi_i := [\xi_{i1},\ldots,\xi_{in}]^{\mathsf{T}}$ denotes a vector of binary parameters such that $\xi_{ij} = 1$ if element j is able to cover target i and $\xi_{ij} = 0$ otherwise. Set covering models find a variety of real-world applications, including scheduling [SW88], production planning [VWS89], facility location [Gun82], and vehicle routing [BS97], etc. Consider, for example, we wish to build medical facilities among n locations to cover I target residential regions. In this context, ξ_i describes the connection between target i and all candidate locations and it may depend on their distances, e.g., $\xi_{ij} = 1$ if target i is near location j. Accordingly, constraints (3.1) ensure that every target is within the neighborhood of some open facilities. This model can be generalized to incorporate backup coverage (see [YL05]). For example, crucial targets with higher priority (e.g., nursing homes) may be covered by multiple open facilities. This generalizes constraints (3.1) to $x^{\mathsf{T}}\xi_i \geq v_i, \forall i \in [I]$, where $v_i \geq 1$ denotes an integer constant that represents the coverage level of target i.

In emergency (e.g., natural disasters), the connection between target i and open facilities may be randomly disrupted. In that case, it is convenient to model ξ_i as Bernoulli random variables, denoted by $\widetilde{\xi}_i, i \in [I]$, and formulate the generalized set covering constraints

probabilistically:

$$\mathbb{P}_{\text{true}}\left\{ x^{\mathsf{T}}\widetilde{\xi}_i \geq v_i, \forall i \in [I] \right\} \geq 1-\epsilon, \tag{3.2}$$

where $\mathbb{P}_{\text{true}}$ denotes the joint probability distribution of $\widetilde{\xi}$ and ϵ, often chosen to be small such as 0.1 and 0.05, denotes a pre-specified risk level. Intuitively, a selection x that satisfies (3.2) can maintain the desired coverage level of *all* targets with high probability. Chance constraint (3.2) is called single if $I = 1$ and joint if $I \geq 2$.

In reality, our knowledge on $\mathbb{P}_{\text{true}}$ is usually ambiguous. For example, the historical data of $\widetilde{\xi}$ may be limited because of the infrequency of natural disasters. Due to such ambiguity on $\mathbb{P}_{\text{true}}$, we adopt a distributionally robust perspective. Specifically, we assume the access to a set of N independent and identically distributed (i.i.d.) samples drawn from $\mathbb{P}_{\text{true}}$, denoted as $\{\widehat{\xi}^j\}_{j\in[N]}$. This gives rise to an empirical distribution $\widehat{\mathbb{P}}_{\widetilde{\xi}} := (1/N)\sum_{j=1}^{N}\Delta_{\widehat{\xi}^j}$ and provides an approximation to $\mathbb{P}_{\text{true}}$, where $\Delta_{\widehat{\xi}^j}$ denotes the Dirac measure on the singleton $\{\widehat{\xi}^j\}$. Unfortunately, as we will demonstrate in Section 3.5.1, simply replacing $\mathbb{P}_{\text{true}}$ with $\widehat{\mathbb{P}}_{\widetilde{\xi}}$ in chance constraint (3.2) may not produce a feasible solution. That is, a solution to such empirical approximation has low confidence of satisfying the actual chance constraint (3.2) under $\mathbb{P}_{\text{true}}$. As an alternative, distributionally robust approaches consider all distributions lying in a neighborhood of $\widehat{\mathbb{P}}_{\widetilde{\xi}}$. We denote this neighborhood as an ambiguity set $\mathcal{P}$, which can be defined based on various discrepancy measures between probability distributions, e.g., Hellinger distance, Prokhorov metric, Wasserstein distance, etc. In this chapter, we adopt the Wasserstein distance (see, e.g., [MK18a]), which measures the discrepancy between distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ by

$$d_W(\mathbb{P}_1,\mathbb{P}_2) := \inf_{\mathbb{Q}\sim(\mathbb{P}_1,\mathbb{P}_2)} \mathbb{E}_{\mathbb{Q}}\left[\|\widetilde{X}_1 - \widetilde{X}_2\|_p\right],$$

where $\widetilde{X}_1$, $\widetilde{X}_2$ are random variables following distributions $\mathbb{P}_1$, $\mathbb{P}_2$ respectively, $\mathbb{Q}\sim(\mathbb{P}_1,\mathbb{P}_2)$ denotes that $\mathbb{Q}$ is a joint distribution of $\widetilde{X}_1$ and $\widetilde{X}_2$ with marginals $\mathbb{P}_1$ and $\mathbb{P}_2$, and $\|\cdot\|_p$ denotes the p-norm. Intuitively, $\mathbb{Q}$ is a plan of transporting the probability masses of $\mathbb{P}_1$ to make it coincide with $\mathbb{P}_2$, and $d_W(\mathbb{P}_1,\mathbb{P}_2)$ equals the minimum transportation cost evaluated under the p-norm. Accordingly, we consider the following Wasserstein ambiguity set:

$$\mathcal{P} := \left\{ \mathbb{P}\in\mathcal{P}_0(\{0,1\}^{I\times n})\colon d_W(\mathbb{P},\widehat{\mathbb{P}}_{\widetilde{\xi}}) \leq \delta \right\}, \tag{3.3}$$

where $\mathcal{P}_0(\{0,1\}^{I\times n})$ denotes the set of all distributions supported on $\{0,1\}^{I\times n}$ and $\delta>0$ denotes the radius of the ambiguity set. This leads to the following generalized distribution-

ally robust chance-constrained set covering problem:

(**DRC**)
$$\begin{aligned} \min_{x} \;\; & c^{\mathsf{T}} x, \\ \text{s.t.} \;\; & \inf_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left\{ x^{\mathsf{T}}\widetilde{\xi}_i \geq v_i, \forall i \in [I] \right\} \geq 1-\epsilon, \\ & x \in \{0,1\}^n, \end{aligned} \tag{3.4}$$

where $c \in \mathbb{R}^n$ denotes a deterministic cost vector and constraint (3.4) requires to satisfy the chance constraint under all distributions within $\mathcal{P}$. If δ is sufficiently large that $\mathbb{P}_{\text{true}} \in \mathcal{P}$ then any feasible solution x to (DRC) satisfies the "true" chance constraint (3.2). In a data-driven context, it has been shown that as the data size N increases, the confidence of $\mathbb{P}_{\text{true}} \in \mathcal{P}$ increases to one exponentially fast (see, e.g., [FG14] and [GS02]).

We remark that the Wasserstein distance is equivalent to various other discrepancy measures, including Hellinger distance and Prokhorov metric. For example, one can bound the Hellinger distance between $\mathbb{P}$ and $\widehat{\mathbb{P}}_{\widetilde{\xi}}$, from both above and below, by $d_W(\mathbb{P}, \widehat{\mathbb{P}}_{\widetilde{\xi}})$ (see, e.g., [GS02]). This implies that we can employ constraint (3.4) to approximate distributionally robust chance constraints defined under other discrepancy measures.

3.1.1 Literature Review

Chance-constrained programming (**CC**) arises in a wide range of applications including power systems [ZL10], transportation [CY04], facility location [MG06], and wireless communication [HKB05]. Dating back to [CCS58; MW65; Pré70], **CC**s are considered very challenging to solve because (i) checking the feasibility of a given solution x demands multivariate integral, which is difficult to calculate, and (ii) **CC** produces a nonconvex and even disconnected feasible region in general. To mitigate this challenge, prior work has studied convex conservative approximation [NS07; BEN09] and sample average approximation (SAA) [LA08; CC04], both of which can efficiently search for feasible solutions with guarantee. Nevertheless, the former may remain challenging to solve if decision variables are discrete (like in this chapter), and the latter often demands the capability of drawing as many samples from $\mathbb{P}_{\text{true}}$ as needed. In addition, distributionally robust chance-constrained models (**DRC**) have received increasing attention in recent years [EOO03; CE06; VBC07; NS07; ZKR11; XCM12; JG15; Dua+18; CKW18; Yan20; LJM19]. There are many successful developments on the tractability of single and joint chance constraints with moment ambiguity sets, which characterize $\mathcal{P}$ based on moment information of $\mathbb{P}_{\text{true}}$ [CE06; ZKR11; YX14; Han+15; XA16; Han+17; LJM19]. Nevertheless, moment ambiguity sets become more conservative than their counterparts based on discrepancy measures (e.g., a Wasser-

stein ambiguity set) when more data samples are available. On the other hand, **DRC** with a Wasserstein ambiguity set is not polynomially solvable in general [XA20]. Prior work has developed exact reformulations and valid inequalities [Xie19; JL20; CKW18; Ho-+21; Ho-+20] for solving this model. For example, [Xie19; JL20; CKW18] derived mixed-integer linear or conic reformulations. When decision variables are purely binary, [Xie19] exploited the submodularity of the reformulation to produce extended polymatroid inequalities. In addition, [JL20] derived precedence valid inequalities among the binary variables they employed to indicate constraint satisfaction in each scenario. [Ho-+21] and [Ho-+20] focused on **DRC** with right-hand-side (RHS) and left-hand-side (LHS) Wasserstein ambiguity, respectively. By exploring the connection between **DRC** and the SAA formulation of **CC**, they employed the mixing scheme to produce valid inequalities (see, e.g, [GP01; Küç12; Lue14]). Nevertheless, prior work has paid less attention to problems with binary decision variables and discrete uncertainty. In this chapter, we study joint **DRC** with LHS Wasserstein ambiguity in generalized set covering, where both decision and random variables are purely binary.

Chance-constrained integer programs [BB09; SL13; Lue14; SLK14; WK19; WLM19] are stochastic variants of combinatorial optimization problems when uncertainty arises. In [Lue14], the authors developed a general decomposition framework for solving chance-constrained programs and derived strong mixed valid inequalities by combining "base" inequalities from each scenario. In addition, [SLK14] proposed an efficient coefficient strengthening method and lifted probabilistic cover inequalities for chance-constrained bin packing problems. Recently, [WLM19] studied a chance-constrained assignment problem and its **DRC** variant, for which they derived strong lifted cover inequalities with efficient separation heuristics. In this work, we exploit the special structures of set covering and Wasserstein ambiguity to derive two families of valid inequalities. The first family produces the convex hull of the hypograph of a "shifted" submodular function. To the best of our knowledge, such a convex hull was not discovered in prior work. The second family mixes inequalities across multiple scenarios, bearing a resemblance to the derivation in [Lue14]. Nevertheless, we take advantage of the binary nature of our decision and random variables and strengthen the mixed inequalities further via lifting.

Chance-constrained set covering models fall into two main categories. In the first category, uncertainty arises on the RHS [BR02; SGL10]. In [BR02], the authors developed a specialized branch-and-bound algorithm based on the enumeration of p-efficient points, which were initially introduced by [Pré90]. Later, [SGL10] simplified the enumeration approach and derived polarity cuts to improve the computational performance. The second category models LHS uncertainty [FM12; AP13; WK19], as in this chapter. For

example, [FM12] studied single chance constraints, where all components of the Bernoulli random vector $\widetilde{\xi}_i$ are independent, and developed efficient cutting plane approaches. In addition, [WK19] proposed an exact approach for solving **CC**s when there exists an oracle to retrieve the probability of any events under $\mathbb{P}_{\text{true}}$. They demonstrated this approach on chance-constrained partial set covering problems with either independence or linear threshold assumptions. To the best of our knowledge, [AP13] is the only prior work on **DRC** for set covering, but they studied single chance constraints under a moment ambiguity set. They derived a compact equivalent reformulation and exploited its supermodularity to derive strong valid inequalities. Different from [AP13], we study joint **DRC** under Wasserstein ambiguity. In addition, the supermodularity of their reformulation stems from the correlation among $\widetilde{\xi}_{ij}$'s, while our "shifted" submodular function is a result of the generalized set covering and the Wasserstein ambiguity.

3.1.2 Contributions

We derive exact reformulations and two families of valid inequalities for (DRC). Our main contributions include

1. We establish the NP-hardness of (DRC) and show that (DRC) admits a deterministic two-stage reformulation, which can be solved efficiently by decomposition algorithms.
2. We derive a complete description of the convex hull of a basic mixed-integer set, which stems from the hypograph of a "shifted" submodular function and arises in each scenario of the two-stage reformulation.
3. We derive a family of cross-scenario inequalities by lifting the mixed valid inequalities obtained from multiple scenarios.
4. We conduct extensive numerical experiments to demonstrate (i) the out-of-sample performance of (DRC) and (ii) the effectiveness of our two-stage reformulation and valid inequalities.

The rest of this chapter is organized as follows. Section 3.2 shows the NP-hardness and develops an exact two-stage reformulation for (DRC). Section 3.3 derives the single- and cross-scenario valid inequalities. Section 3.4 extends the reformulation and valid inequalities to a distributionally robust knapsack chance constraint. Section 3.5 demonstrates the effectiveness of the proposed model and solution approaches, and finally Section 3.6 concludes this chapter.

Notation: $\mathbb{Z}_+$ denotes the set of nonnegative integers. For integers m and n, $[n] := \{1, \ldots, n\}$, $\mathbf{1}_{m\times n}$ denotes an $m \times n$ matrix of all ones, $\mathbf{1}$ denotes a vector of all ones with suitable dimension, and $\mathbf{e}_m$ denotes the mth standard basis vector with suitable dimension. For $x \in \mathbb{R}$, $(x)^+ := \max\{x, 0\}$.

For set E, $|E|$ denotes its cardinality, **conv**(E) denotes its convex hull, and the indicator function $\mathbb{1}\{x \in E\} := 1$ if $x \in E$ and $\mathbb{1}\{x \in E\} := 0$ otherwise.

3.2 Two-Stage Reformulation

We derive a reformulation of (DRC) based on the conditional value-at-risk (CVaR) in Section 3.2.1 and show its NP-hardness in Section 3.2.2. Then, in Section 3.2.3, we further recast it as a two-stage stochastic program.

3.2.1 Conditional Value-at-Risk Reformulation

We recall the definitions of value-at-risk (VaR) and CVaR [RU99]. For a random variable $\widetilde{X}$ following its induced probability distribution $\mathbb{P}_{\widetilde{X}}$, the $(1-\epsilon)$-VaR of $\widetilde{X}$ is defined as

$$\mathsf{VaR}_{1-\epsilon}(\widetilde{X}) := \inf\left\{ x : \mathbb{P}_{\widetilde{X}}\{\widetilde{X} \le x\} \ge 1-\epsilon \right\},$$

and its $(1-\epsilon)$-CVaR is defined as:

$$\mathsf{CVaR}_{1-\epsilon}(\widetilde{X}) = \min_{\gamma}\left\{ \gamma + \frac{1}{\epsilon}\mathbb{E}\left(\widetilde{X} - \gamma\right)^+ \right\}. \tag{3.5}$$

Previously, [Xie19] derived a CVaR reformulation for **DRC** when $\widetilde{\xi}$ is supported in a normed vector space. Since our $\widetilde{\xi}$ is binary-valued, we adapt the framework of [Xie19] to obtain a slightly different reformulation, which also admits a CVaR interpretation.

Proposition 6 (Adapted from Theorem 1 in [Xie19])**.** Let Z represent the feasible region produced by constraint (3.4) in (DRC). Then, it holds that

$$Z = \left\{ x \in \{0,1\}^n : \begin{array}{c} \exists\, \gamma \in \mathbb{R}_+, z \in \mathbb{R}_-^N : \\ \delta - \gamma\epsilon \le \dfrac{1}{N}\displaystyle\sum_{j\in[N]} z_j, \\ z_j + \gamma \le \min\limits_{i\in[I]}\left(\left(x^\intercal \widehat{\xi}_i^j - v_i + 1\right)^+\right)^{1/p}, \forall j \in [N] \end{array} \right\}. \tag{3.6}$$

In addition, Z admits the following CVaR interpretation:

$$Z = \left\{ x \in \{0,1\}^n : \frac{\delta}{\epsilon} + \mathbf{CVaR}_{1-\epsilon}\left[-g(x,\widehat{\xi})\right] \leq 0 \right\}, \tag{3.7}$$
$$\text{where} \quad g(x,\xi) = \min_{i\in[I]} \left(\left(x^{\mathsf{T}}\xi_i - v_i + 1 \right)^+ \right)^{1/p}.$$

Proof (see Theorem 1 in [Xie19] and Theorem 3 in [BM19]). We first rewrite constraint (3.4) using complement:

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left\{ \exists\, i \in [I] :\ x^{\mathsf{T}}\widetilde{\xi}_i \leq v_i - 1 \right\} \leq \epsilon.$$

By Theorem 3 in [BM19], we expand the above supremum to obtain

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left\{\exists\, i \in [I] \colon x^{\mathsf{T}}\widetilde{\xi}_i \leq v_i - 1\right\} = \min_{\lambda\geq 0} \left\{ \lambda\delta + \frac{1}{N}\sum_{j\in[N]} \left(1 - \lambda\cdot g(x,\widehat{\xi}^j)\right)^+ \right\},$$

where

$$\begin{aligned} g(x,\widehat{\xi}^j) &:= \inf_{\xi\in\Xi} \left\{ \|\xi - \widehat{\xi}^j\|_p \colon \exists\, i \in [I] \text{ s.t. } x^{\mathsf{T}}\xi_i \leq v_i - 1 \right\} \\ &= \min_{i\in[I]} \inf_{\xi\in\Xi} \left\{ \|\xi - \widehat{\xi}^j\|_p \colon x^{\mathsf{T}}\xi_i \leq v_i - 1 \right\}. \end{aligned}$$

Then, the **CVaR** interpretation follows from Corollary 1 in [Xie19]. In what follows, we recast $g(x,\widehat{\xi}^j)$. To this end, let $\mathcal{T}$ be the index set of x, i.e., $\mathcal{T} := \{k \in [n] : x_k = 1\}$. For any $i \in [I]$, we have

$$\begin{aligned} \min_{x^{\mathsf{T}}\xi_i \leq v_i - 1} \|\xi - \widehat{\xi}^j\|_p &= \min_{x^{\mathsf{T}}\xi_i \leq v_i - 1} \left(\sum_{k=1}^{n} \left| \xi_{ik} - \widehat{\xi}^j_{ik} \right|^p \right)^{1/p} \\ &= \min_{x^{\mathsf{T}}\xi_i \leq v_i - 1} \left(\sum_{k\notin\mathcal{T}} \left| \xi_{ik} - \widehat{\xi}^j_{ik} \right|^p + \sum_{k\in\mathcal{T}} \left| \xi_{ik} - \widehat{\xi}^j_{ik} \right|^p \right)^{1/p} \\ &= \min_{x^{\mathsf{T}}\xi_i \leq v_i - 1} \left(\sum_{k\in\mathcal{T}} \left| \xi_{ik} - \widehat{\xi}^j_{ik} \right|^p \right)^{1/p} \\ &= \left(\left(x^{\mathsf{T}}\widehat{\xi}^j_i - (v_i - 1) \right)^+ \right)^{1/p}. \end{aligned}$$

It follows that $g(x, \widehat{\xi}^j) = \min_{i\in[I]} \left((x^\mathsf{T}\widehat{\xi}_i^j - v_i + 1)^+\right)^{1/p}$. This completes the proof.

□

Although (DRC) admits the same CVaR interpretation as in Xie [Xie19], the reformulation Z of (DRC) is inherently different from that of [Xie19]. This is because the uncertain parameters $\widetilde{\xi}$ are binary-valued in (DRC) while those in [Xie19] are supported in a vector space. As a result, projecting a vector onto a half-space, which is a key step of reformulating Wasserstein chance constraints, differs in discrete and vector spaces. Specifically, in the proof of Proposition 6,

$$\min_{x^\mathsf{T}\xi_i \le v_i - 1} \|\xi - \widehat{\xi}^j\|_p = \begin{cases} \left((x^\mathsf{T}\widehat{\xi}_i^j - v_i + 1)^+\right)^{1/p} & \text{if } \Xi = \{0,1\}^n \text{ (as in the proof of Proposition 6),} \\ \frac{(x^\mathsf{T}\widehat{\xi}_i^j - v_i + 1)^+}{\|x\|_{p/(p-1)}} & \text{if } \Xi = \mathbb{R}^n \text{ (as in [Xie19]).} \end{cases}$$

Remark 3. The reformulation of Z in (3.6) is of independent interest because it remains valid beyond the current setting of set covering. For example, (DRC) admits the same reformulation when $x \in \{0,1\}^n$, $\Xi = \mathbb{Z}_+^{I\times n}$, and $p = 1$. In addition, the reformulation remains valid when some target cannot be covered by certain elements. Formally, suppose that ξ is supported on the set

$$\left\{\xi \in \{0,1\}^{I\times n} : \xi_{ik} = 0 \quad \forall (i,k) \in S\right\}$$

for a subset $S \subseteq [I] \times [n]$, and accordingly $\widehat{\xi}_{ik}^j = 0$ for all $(i,k) \in S$. Then, the reformulation of Z in (3.6) remains valid.

3.2.2 NP-hardness

We establish the NP-hardness of solving (DRC) in the following proposition.

Proposition 7. (DRC) is NP-hard to solve for any given fixed risk level $\epsilon \in (0,1)$.

Proof. We show that (DRC) is equivalent to the following problem when $N = 1$ and $v_i = 1$

for all $i \in [I]$:

$$\begin{aligned} \min_{x} \;& c^{\mathsf{T}} x, \\ \text{s.t.} \;& \frac{\delta}{\epsilon} \leq \min_{i \in [I]} \left(x^{\mathsf{T}} \widehat{\xi}_i^1 \right)^{1/p}, \\ & x \in \{ 0, 1 \}^n . \end{aligned} \tag{DRC'}$$

Let $(x_1, \gamma_1, z_1) \in Z$ be a feasible solution to (DRC), then x_1 is feasible to (DRC') because

$$\frac{\delta}{\epsilon} \leq \frac{z_1}{\epsilon} + \gamma_1 \leq z_1 + \gamma_1 \leq \min_{i \in [I]} \left(x^{\mathsf{T}} \widehat{\xi}_i^1 \right)^{1/p}.$$

On the other hand, if x_2 is feasible to (DRC'), then together with $\gamma_2 := \delta/\epsilon, z_2 := 0$ they are feasible to (DRC) because

$$\begin{aligned} & \delta - \gamma_2 \epsilon = 0 \leq z_2, \\ & z_2 + \gamma_2 = \delta/\epsilon \leq \min_{i \in [I]} \left(x_2^{\mathsf{T}} \widehat{\xi}_i^1 \right)^{1/p}. \end{aligned}$$

Therefore, (DRC) and (DRC') are equivalent. Since restricting δ to be ϵ in (DRC') recovers the set cover problem, we conclude that (DRC) is NP-hard to solve. □

One may be tempted to obtain the convex hull of the following mixed-integer set $\mathcal{Q}$ arising from the reformulation (3.6),

$$\mathcal{Q} := \left\{ (\theta, x) \in \mathbb{R} \times \{ 0, 1 \}^n : \theta \leq \min_{i \in [I]} \left(x^{\mathsf{T}} \xi_i \right)^{1/p} \right\}$$

for given $\xi_i \in \{ 0, 1 \}^n$, $i \in [I]$. Nonetheless, we show that optimizing a linear function over $\mathcal{Q}$ is NP-hard as well. The proof of this proposition is given in Appendix B.1.

Proposition 8. The following problem is NP-hard to solve for any given fixed $p \in \mathbb{R}, p \geq 1$:

$$\begin{aligned} \min_{x, \theta} \;& c^{\mathsf{T}} x - \theta^{1/p} \\ \text{s.t.} \;& \theta \leq x^{\mathsf{T}} \xi_i, \forall i \in [I], \\ & \theta \in \mathbb{R}, x \in \{ 0, 1 \}^n . \end{aligned} \tag{3.8}$$

3.2.3 Two-Stage Reformulation

The reformulation (3.6) of Z is potentially challenging to optimize over, particularly because function $g(x, \xi)$ defined in Proposition 6 is non-concave in x.

Although one can linearize g with the help of auxiliary binary variables and big-M coefficients, the formulation thus obtained is computationally ineffective. Another linearized formulation without big-M coefficients can be obtained by following [Xie19] if $g(x, \xi)$ is supermodular in x. Unfortunately, this is not the case even when $I = 1$ (see Section 3.3.1). As an alternative, we exploit the (hidden) submodularity of g to obtain a two-stage reformulation without additional binary variables or big-M coefficients. We review key concepts of submodularity in Section 3.2.3.1 and present the reformulation in Section 3.2.3.2.

3.2.3.1 Polyhedral Results for Submodular Functions

Definitions 8. A function $\phi \colon 2^{[n]} \to \mathbb{R}$ is submodular if for any $\mathcal{R}, \mathcal{T} \subseteq [n]$, we have

$$\phi(\mathcal{R}) + \phi(\mathcal{T}) \geq \phi(\mathcal{R} \cup \mathcal{T}) + \phi(\mathcal{R} \cap \mathcal{T}).$$

In addition, ϕ is supermodular if $-\phi$ is submodular. □

For ease of exposition, we use $\phi(\mathcal{T})$ and $\phi(x_{\mathcal{T}})$ interchangeably, where $x_{\mathcal{T}} \in \{ 0, 1 \}^n$ is the indicating vector of $\mathcal{T}$ such that $x_k = 1$ if and only if $k \in \mathcal{T}$. An example of submodular function follows.

Lemma 7 ([Top78; Xie19])**.** For fixed $\alpha \in \mathbb{R}^n_+, \alpha_o \in \mathbb{R}$ function $\phi(x) := \min\{-\alpha^{\mathsf{T}} x + \alpha_o, 0\}$ is submodular.

Next, consider the epigraph **epi**(ϕ) of a submodular function ϕ:

$$\textbf{epi}(\phi) := \left\{ (\theta, x) \in \mathbb{R} \times \{ 0, 1 \}^n : \phi(x) \leq \theta \right\}.$$

Then, the convex hull of **epi**(ϕ) is fully described by the extended polymatroid inequalities (EPI) [Sch03], i.e.,

$$\textbf{conv}\left(\textbf{epi}(\phi)\right) := \left\{ (\theta, x) \in \mathbb{R} \times [0, 1]^n : \phi(\varnothing) + \sum_{k=1}^{n} \left[\phi(\mathcal{T}_k) - \phi(\mathcal{T}_{k-1}) \right] x_{\sigma_k} \leq \theta, \forall \sigma \in \Sigma \right\},$$

where Σ denotes all permutations of $[n]$ and $\mathcal{T}_k := \{ \sigma_1, \ldots, \sigma_k \}, \mathcal{T}_0 := \varnothing$. The separation of **conv** $(\textbf{epi}(\phi))$ is very efficient even though it involves $n!$ number of constraints.

Theorem 8 (Proposition 1 in [AN08], Theorem 44.3 in [Sch03], Section 3 in [Lov83])**.** If $(\widehat{\theta}, \widehat{x}) \notin$ **conv** (**epi**(ϕ)), then it violates constraint

$$\phi(\varnothing) + \sum_{k=1}^{n} \left[\phi(\mathcal{T}_k) - \phi(\mathcal{T}_{k-1})\right] x_{\sigma_k} \leq \theta,$$

where $\sigma \in \Sigma$ is the permutation such that $\widehat{x}_{\sigma_1} \geq \widehat{x}_{\sigma_2} \geq \cdots \geq \widehat{x}_{\sigma_n}$.

3.2.3.2 Reformulation

We first exchange the order of applying $(\cdot)^+$ and $(\cdot)^{1/p}$ in the definition of g. Since $x^\mathsf{T}\widehat{\xi}^j - v_i + 1$ can be negative, we extend the domain of function $(\cdot)^{1/p}$ in the following.

Proposition 9. For $z \in \mathbb{Z}$, define $\bar{f}(z) := z^{1/p} \cdot \mathbb{1}\{ z \geq 0 \} + z \cdot \mathbb{1}\{ z \leq -1 \}$. Then,

$$g(x, \xi) \equiv \min_{i \in [I]} \left(\left(x^\mathsf{T}\widehat{\xi}_i^j - v_i + 1 \right)^+ \right)^{1/p} = \left(\bar{f} \left(\min_{i \in [I]} \left\{ x^\mathsf{T}\widehat{\xi}_i^j - v_i + 1 \right\} \right) \right)^+ .$$

Proof. For $x \in \{ 0, 1 \}^n$, we have

$$\begin{aligned} \min_{i \in [I]} \left(\left(x^\mathsf{T}\widehat{\xi}_i^j - v_i + 1 \right)^+ \right)^{1/p} &= \left(\min_{i \in [I]} \left(x^\mathsf{T}\widehat{\xi}_i^j - v_i + 1 \right)^+ \right)^{1/p} \\ &= \left(\left(\min_{i \in [I]} \left\{ x^\mathsf{T}\widehat{\xi}_i^j - v_i + 1 \right\} \right)^+ \right)^{1/p} \\ &= \left(\bar{f} \left(\min_{i \in [I]} \left\{ x^\mathsf{T}\widehat{\xi}_i^j - v_i + 1 \right\} \right) \right)^+ , \end{aligned}$$

where the first equality is because $(\cdot)^{1/p}$ is monotone, the second equality is because $(\cdot)^+$ is monotone, and the third equality is because $\bar{f}(z) \geq 0$ if and only if $z \geq 0$. □

Although $\bar{f}$ is defined on $\mathbb{Z}$, the conclusion of Proposition 9 holds even when $x^\mathsf{T}\widehat{\xi}_i^j$ takes a fractional value. Next, we linearize the nonlinear function $\bar{f}$. In what follows, Propositions 10–12 utilize the fact that $x^\mathsf{T}\widehat{\xi}_i^j$ is an integer.

Proposition 10. For any $x \in \{0,1\}^n$, it holds that

$$\begin{aligned}\bar{f}\left(\min_{i\in[I]}\left\{x^{\mathsf{T}}\widehat{\xi}_i^j - v_i + 1\right\}\right) = \max_{y}\ & c_p^{\mathsf{T}} y + \bar{f}(1-v_m) \\ \text{s.t.}\ & \mathbf{1}^{\mathsf{T}} y \le \min_{i\in[I]}\left\{x^{\mathsf{T}}\widehat{\xi}_i^j - v_i + v_m\right\}, \\ & y \in [0,1]^n,\end{aligned}$$

where $v_m := \max\{v_i \colon i \in [I]\}$ and $c_p := [\bar{f}(2-v_m)-\bar{f}(1-v_m), \bar{f}(3-v_m)-\bar{f}(2-v_m), \ldots, \bar{f}(n+1-v_m)-\bar{f}(n-v_m)]^{\mathsf{T}}$.

Proof. For fixed $x \in \{0,1\}^n$, define $\bar{v} := \min\{x^{\mathsf{T}}\widehat{\xi}_i^j - v_i + 1 \colon i \in [I]\}$. Since the above linear program is a continuous knapsack problem with an integer knapsack capacity and $c_p \ge 0$, there exists an optimal solution $y^* \in \{0,1\}^n$ such that $\mathbf{1}^{\mathsf{T}} y^* = \bar{v} + v_m - 1$. Also note that vector c_p has non-increasing entries because $\bar{f}$ is a concave increasing function by construction. Hence, without loss of optimality, the first $\bar{v} + v_m - 1$ entries of y^* equal one and the remaining entries equal zero. This yields an optimal objective value

$$\begin{aligned}c_p^{\mathsf{T}} y^* &= \bar{f}(2-v_m) - \bar{f}(1-v_m) + \bar{f}(3-v_m) - f(2-v_m) + \cdots + \bar{f}(\bar{v}) - \bar{f}(\bar{v}-1) + \bar{f}(1-v_m) \\ &= \bar{f}(\bar{v}) = \bar{f}\left(\min_{i\in[I]}\left\{x^{\mathsf{T}}\widehat{\xi}_i^j - v_i + 1\right\}\right).\end{aligned}$$

□

We are now ready to present the main result of this section.

Proposition 11. (DRC) admits the following reformulation:

$$\begin{aligned}(\mathbf{MP}) \quad \min_{x,\gamma,z}\ & c^{\mathsf{T}} x \\ \text{s.t.}\ & \delta - \gamma\epsilon \le \frac{1}{N}\sum_{j\in[N]} z_j, \\ & -z_j - \gamma \ge \min\{Q_j(\mu,x), 0\}, \quad \forall \mu \in \mathcal{M}, \forall j \in [N], \qquad (3.9) \\ & x \in \{0,1\}^n, \gamma \in \mathbb{R}_+, z \in \mathbb{R}_-^N,\end{aligned}$$

where $Q_j(\mu,x) := \sum_{i\in[I]} \mu_{1i}\left(x^{\mathsf{T}}\widehat{\xi}_i^j - v_i + v_m\right) + \mathbf{1}^{\mathsf{T}}\mu_2 - \bar{f}(1-v_m)$ and $\mathcal{M} := \{(\mu_1,\mu_2) \in \mathbb{R}_-^I \times \mathbb{R}_-^n : \mu_1^{\mathsf{T}}\mathbf{1}_{I\times n} + \mu_2^{\mathsf{T}} \le -c_p^{\mathsf{T}}\}$.

Proof. By Propositions 6 and 10, (DRC) is equivalent to:

$$\begin{aligned}
\min_{x,\gamma,z}\ & c^{\mathsf{T}}x \\
\text{s.t.}\ & \delta - \gamma\epsilon \le \frac{1}{N}\sum_{j\in[N]} z_j, \\
& -z_j - \gamma \ge \min\left\{\, Q'_j(x), 0 \,\right\}, \quad \forall j \in [N], \\
& x \in \{\,0,1\,\}^n, \gamma \in \mathbb{R}_+, z \in \mathbb{R}^N_-,
\end{aligned}$$

where $Q'_j(x)$ is defined as:

$$\begin{aligned}
(\mathbf{SP}_j) \qquad Q'_j(x) := \min_{y_j}\ & -c_p^{\mathsf{T}}y - \bar{f}(1-v_m) \\
\text{s.t.}\ & \mathbf{1}^{\mathsf{T}}y_j \le x^{\mathsf{T}}\widehat{\xi}^j_i - v_i + v_m, \quad \forall i \in [I], && \text{(3.10a)} \\
& y_j \in [0,1]^n. && \text{(3.10b)}
\end{aligned}$$

Taking dual of the above linear program yields:

$$\begin{aligned}
(\mathbf{DSP}_j) \qquad Q'_j(x) = \max_{\mu}\ & Q_j(\mu, x) \equiv \sum_{i\in[I]} \mu_{1i}\left(x^{\mathsf{T}}\widehat{\xi}^j_i - v_i + v_m\right) + \mathbf{1}^{\mathsf{T}}\mu_2 - \bar{f}(1-v_m) \\
\text{s.t.}\ & \mu \in \mathcal{M},
\end{aligned}$$

where dual variables μ_1, μ_2 are associated with constraints (3.10a) and (3.10b), respectively. Then, the reformulation follows by noting that

$$\min\left\{\, \max_{\mu\in\mathcal{M}} Q_j(\mu, x), 0 \,\right\} = \max_{\mu\in\mathcal{M}} \min\left\{\, Q_j(\mu, x), 0 \,\right\}. \tag{3.11}$$

□

Remark 4. By Lemma 7, the RHS of constraints (3.9) is submodular in x. As a result, we can replace (3.9) with the EPI with respect to this submodular function. This implies that (DRC) admits a mixed-integer *linear* reformulation.

Proposition 11 suggests a Benders decomposition (BD) algorithm of solving (DRC) in an iterative manner. In each iteration, BD solves (MP) with constraints (3.9) replaced by a relaxation, and sends an optimal solution $(\widehat{x}, \widehat{\gamma}, \widehat{z})$ to every (DSP_j) to check feasibility. If feasible, then $(\widehat{x}, \widehat{\gamma}, \widehat{z})$ is optimal to (DRC); otherwise, we identify a $\widehat{\mu} \in \mathcal{M}$ such that $-\widehat{z}_j - \widehat{\gamma} < \min\{Q_j(\widehat{\mu}, \widehat{x}), 0\}$. In the latter case, we obtain a violated EPI with respect to $\min\{Q_j(\widehat{\mu}, x), 0\}$ and add it as a Benders feasibility cut back to (MP). Naturally, the efficacy

of this BD algorithm depends on that of solving (DSP_j). The next proposition gives a closed-form solution to (DSP_j).

Proposition 12. For fixed x, an optimal solution $(\widehat{\mu}_1, \widehat{\mu}_2)$ to (DSP_j) satisfies

$$\begin{aligned} \widehat{\mu}_1 &= -\mathbf{e}_{i^*}(c_p)_{v^*}, \\ (\widehat{\mu}_2)_k &= \min\Big\{ -(c_p)_k - \mathbf{1}^\mathsf{T}\widehat{\mu}_1, 0 \Big\}, \quad \forall k \in [n], \end{aligned} \tag{3.12}$$

where $v_m = \max\{\, v_i \colon i \in [I] \,\}$, $i^* \in \operatorname{argmin}\{x^\mathsf{T}\widehat{\xi}_i^j - v_i + v_m \,:\, i \in [I]\}$, $\bar{v} := \min\{\, x^\mathsf{T}\widehat{\xi}_i^j - v_i + v_m \colon i \in [I] \,\}$, and $v^* \in \{\, \lceil \bar{v} \rceil, \lfloor \bar{v} \rfloor \,\}$.

Proof. To maximize $Q_j(\mu, x)$, every component in μ_2 should attain its upper bound at optimality; and hence (3.12) follows. Since (i) $\mu_1 \le 0$, (ii) $x^\mathsf{T}\widehat{\xi}_i^j - v_i + v_m \ge 0$ for all $i \in [I]$, and (iii) $\sum_{k\in[n]} \big((c_p)_k + \mathbf{1}^\mathsf{T}\mu_1\big)^+$ relies solely on $\mathbf{1}^\mathsf{T}\mu_1$, the support of μ_1 is a singleton at optimum, i.e., $\mu_1 = s \cdot \mathbf{e}_{i^*}$ for a nonpositive scalar s. Accordingly, (DSP_j) reduces to a one-dimensional problem

$$\max_{\mu\in\mathcal{M}} Q_j(\mu, x) = \max_{s\in\mathbb{R}_-} \left\{ s \cdot \bar{v} - \sum_{k\in[n]} \big((c_p)_k + s\big)^+ \right\} - \bar{f}(1 - v_m),$$

and $\mathbf{1}^\mathsf{T}\mu_1 = s$. Since the entries of c_p are descending, a maximizer to the above problem equals either $-(c_p)_{\lfloor \bar{v} \rfloor}$ or $-(c_p)_{\lceil \bar{v} \rceil}$. □

Therefore, by Theorem 8 the Benders feasibility cut takes the following form:

$$-z_j - \gamma \ge \phi^j(\varnothing) + \sum_{k=1}^{n} \Big[\phi^j(\mathcal{T}_k) - \phi^j(\mathcal{T}_{k-1})\Big] x_{\sigma_k}, \tag{3.13}$$

$$\text{where} \quad \phi^j(x) := \min\left\{ \mathbf{1}^\mathsf{T}\widehat{\mu}_1 \left(x^\mathsf{T}\widehat{\xi}_{i^*}^j - v_{i^*} + v_m \right) + \mathbf{1}^\mathsf{T}\widehat{\mu}_2 - \bar{f}(1 - v_m),\ 0 \right\}$$

$$\text{and} \quad \phi^j(\mathcal{T}_k) - \phi^j(\mathcal{T}_{k-1}) = \begin{cases} \mathbf{1}^\mathsf{T}\widehat{\mu}_1 & \text{if } \phi^j(\mathcal{T}_{k-1}) < 0 \\ \phi^j(\mathcal{T}_k) & \text{if } \phi^j(\mathcal{T}_{k-1}) = 0 \text{ and } \phi^j(\mathcal{T}_k) < 0 \\ 0 & \text{if } \phi^j(\mathcal{T}_k) = 0. \end{cases}$$

We compare our two-stage reformulation of (DRC) with that of [Xie19] when decision variables are binary. As pointed out after Proposition 6, our setting is inherently different from that of [Xie19], and as a result, the technique Xie [Xie19] used to exploit the supermodularity does not apply to our case. To be concrete, suppose that $\widetilde{\xi}$ is supported in a vector space as in [Xie19]. Then, applying Theorem 1 and Proposition 1 of [Xie19] to our

set covering model yields the following reformulation of (DRC):

$$Z_0 = \left\{ x \in \{0,1\}^n : \begin{array}{l} \exists\, \gamma \in \mathbb{R}_+, \nu \in \mathbb{R}_+, z \in \mathbb{R}_-^N : \\ \delta\nu - \gamma\epsilon \le \dfrac{1}{N}\displaystyle\sum_{j\in[N]} z_j, \\ \|x\|_{p/(p-1)} \le \nu, \\ z_j + \gamma \le \left(x^\mathsf{T}\widehat{\xi}_i^j - v_i\right)^+, \forall i \in [I], j \in [N] \end{array} \right\}.$$

In the last constraints of Z_0, the supermodularity of $(x^\mathsf{T}\widehat{\xi}_i^j - v_i)^+$ in x appears naturally. In contrast, in our case the term $\left((x^\mathsf{T}\widehat{\xi}_i^j - v_i + 1)^+\right)^{1/p}$ in (3.6) is neither supermodular in x nor in its complements (i.e., replacing some entries x_k with $1 - x_k$) because the root function $(\cdot)^{1/p}$ is strictly concave for all $p > 1$. To exploit supermodularity, we need to recast the root function as a linear program (i.e., Proposition 10) and then take dual.

3.3 Valid Inequalities

We derive two families of valid inequalities for the reformulation (3.6) of Z. Section 3.3.1 provides inequalities based on a single scenario $j \in [N]$, Section B.3.1 consists of detailed proofs for establishing a convex-hull result of these inequalities, and Section 3.3.2 provides cross-scenario inequalities.

3.3.1 Valid inequalities from a single scenario

We focus on a nonlinear inequality $z_j + \gamma \le ((x^\mathsf{T}\widehat{\xi}_i^j - v_i + 1)^+)^{1/p}$ in the reformulation (3.6), for a single scenario $j \in [N]$ and a single covering $i \in [I]$. If $v_i = 1$ then the RHS of this inequality becomes a submodular function in x because $p \ge 1$ and $\widehat{\xi}_i^j \ge 0$. For general v_i, the RHS can be viewed as the submodular function being "shifted" by $(v_i - 1)$. However, after such a shift the RHS becomes neither submodular nor supermodular, preventing us from using existing valid inequalities (e.g., [AA09; Sch03]). In this section, we generalize this inequality and consider the following hypograph of a shifted submodular function:

$$\mathcal{X} := \left\{ (\theta, x) \in \mathbb{R} \times \{0,1\}^n : \theta \le f\left(\left(x^\mathsf{T}\xi - \beta\right)^+\right) \right\},$$

where $f : \mathbb{R}_+ \to \mathbb{R}$ is a strictly concave increasing function with $f(0) = 0$, $\xi \in \{0,1\}^n$ is a given binary vector, and $\beta \in \mathbb{Z}_+$ is a given integer. Let $\mathcal{Z}$ be the set of indices where

ξ takes 1. If $\beta \geq |\mathcal{Z}| - 1$ then $f((x^\mathsf{T}\xi - \beta)^+) \equiv f(1)(x^\mathsf{T}\xi - \beta)^+$ is supermodular in x. In this case, **conv**$(\mathcal{X})$ can be fully described by the EPI. Hence, we assume $1 \leq \beta \leq |\mathcal{Z}| - 2$ without loss of generality (w.l.o.g.). Additionally, since ξ has at most $|\mathcal{Z}|$ numbers of ones, all possible values for $(x^\mathsf{T}\xi - \beta)^+$ lie in the set

$$L := \Big[|\mathcal{Z}| - \beta\Big].$$

We replace f with its inner piecewise linear approximation with integer breakpoints in L in the definition of $\mathcal{X}$. The following lemma formalizes this.

Lemma 8. For $\ell \in L$, let φ_ℓ be a linear function defined as $\varphi_\ell(z) := a_\ell z + b_\ell$ where $a_\ell := f(\ell) - f(\ell - 1)$ and $b_\ell := f(\ell) - a_\ell \cdot \ell$. To avoid clutter, we also define $a_{|\mathcal{Z}|-\beta+1} := 0$ and $b_{|\mathcal{Z}|-\beta+1} := f(|\mathcal{Z}| - \beta)$. Then, we have $f(z) = \min\limits_{\ell \in L} \{\, \varphi_\ell(z) \,\}$ for any $z \in L$. In particular, $f(\ell) = \varphi_\ell(\ell) = \varphi_{\ell+1}(\ell)$ for all $\ell \in L$.

Proof. It is straightforward that $f(\ell) = \varphi_\ell(\ell) = \varphi_{\ell+1}(\ell)$ for all $\ell \in L$. To show that $f(z) = \min_{\ell \in L} \{\, \varphi_\ell(z) \,\}$ for any $z \in L$, it is sufficient to argue that $\varphi_i(\ell) \geq \varphi_\ell(\ell)$ for every $i \in L$. When $i < \ell$, $\varphi_i(\ell) \geq \varphi_\ell(\ell)$ is equivalent to $f(i) - f(i-1) \geq (f(\ell) - f(i))/(\ell - i)$, which holds due to the concavity of f. When $i > \ell$, $\varphi_i(\ell) \geq \varphi_\ell(\ell)$ is equivalent to $(f(i) - f(\ell))/(i - \ell) \geq f(i) - f(i-1)$, which holds due to the concavity of f. □

We mention properties about a_ℓ and b_ℓ in the following remark.

Remark 5. For a_ℓ and b_ℓ, it holds that

(i) a_ℓ are strictly positive and decreasing in ℓ. This is because f is strictly concave and increasing.

(ii) $b_\ell \geq 0$ for all $\ell \in L$. This is because

$$b_\ell = \sum_{k=1}^{\ell} \big(f(k) - f(k-1)\big) - \ell \cdot \big(f(\ell) - f(\ell-1)\big) = \sum_{k=1}^{\ell} (a_k - a_\ell) \geq 0,$$

where the last inequality is because a_ℓ are decreasing in ℓ.

(iii) b_ℓ / a_ℓ are strictly increasing in ℓ. This is because a_ℓ are strictly decreasing and b_ℓ are increasing since

$$b_{\ell+1} - b_\ell = \sum_{k=1}^{\ell+1} (a_k - a_\ell) - \sum_{k=1}^{\ell} (a_k - a_\ell) = \ell(a_\ell - a_{\ell+1}) > 0.$$

In addition, for ease of presenting the valid inequalities we define

$$L_1 := \bigcup_{\ell \in L} \left\{ (\ell, \rho) \colon \rho = \frac{b_\ell}{a_\ell} \right\},$$
$$L_2 := \bigcup_{\ell \in L} \left\{ (\ell, \rho) \colon \rho \in [\beta] \cap \left(\frac{b_\ell}{a_\ell}, \frac{b_{\ell+1}}{a_{\ell+1}} \right) \right\},$$

and $h_{(\ell,\rho)} : \mathbb{R}_+ \to \mathbb{R}$,

$$h_{(\ell,\rho)}(z) := \left(\frac{f(\ell)}{(\ell + \rho)} \big(z - (\beta - \rho) \big) \right)^+, \quad \forall (\ell, \rho) \in L_1 \cup L_2.$$

Remark 6. For $(\ell, \rho) \in L_1$, we can simplify $h_{(\ell,\rho)}$ as

$$h_{(\ell,\rho)}(z) = \big(a_\ell \, (z - \beta) + b_\ell \big)^+ = \big(\varphi_\ell \, (z - \beta) \big)^+ ,$$

from which $h_{(\ell,\rho)}(z)$ coincides with $f((z-\beta)^+)$ for all $(\ell, \rho) \in L_1$ when $z \in \{\ell + \beta - 1, \ell + \beta\}$. For ease of exposition, we use $(\varphi_\ell)^+$ and $h_{(\ell,\rho)}$ interchangeably if $(\ell, \rho) \in L_1$. For $(\ell, \rho) \in L_2$, $h_{(\ell,\rho)}(z) = 0$ if $z \leq \beta - \rho$ and $h_{(\ell,\rho)}(z) > 0$ if $z > \beta - \rho$. In particular, $h_{(\ell,\rho)}(z) = f(\ell)$ when $z = \ell + \beta$.

Example 8. Suppose that $|\mathcal{Z}| = 4$, $\beta = 2$, and function $f(z) = \sqrt{z}$. Then, $L = [2]$, $L_1 = \{ (1, 0), (2, \sqrt{2}) \}$, $L_2 = \{ (1, 1), (2, 2) \}$, and the shifted function $f((z-\beta)^+) = \sqrt{(z-2)^+}$. Accordingly, for $(\ell, \rho) \in L_1$, the function $h_{(\ell,\rho)}$ takes the following two forms:

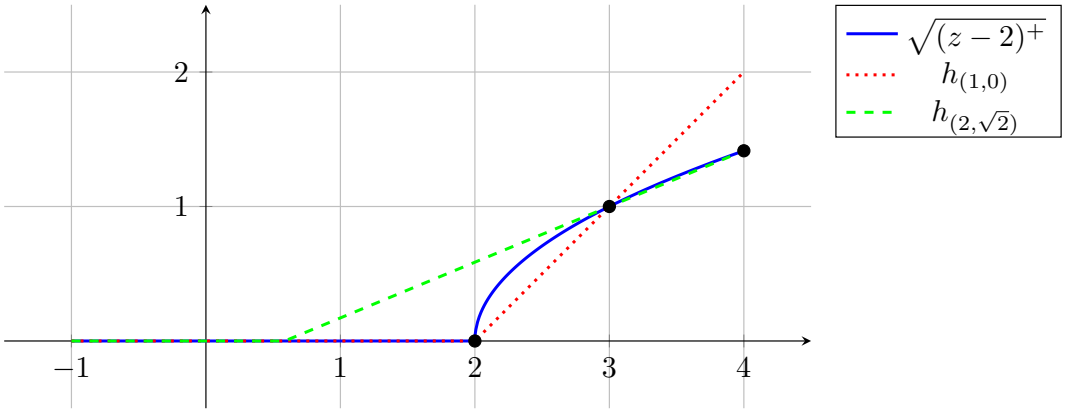
$$h_{(1,0)}(z) = (z - 2)^+, \quad h_{(2,\sqrt{2})}(z) = \big(\sqrt{2} - 1 \big) \Big(z - (2 - \sqrt{2}) \Big)^+ ;$$

and for $(\ell, \rho) \in L_2$, it takes the following two forms:

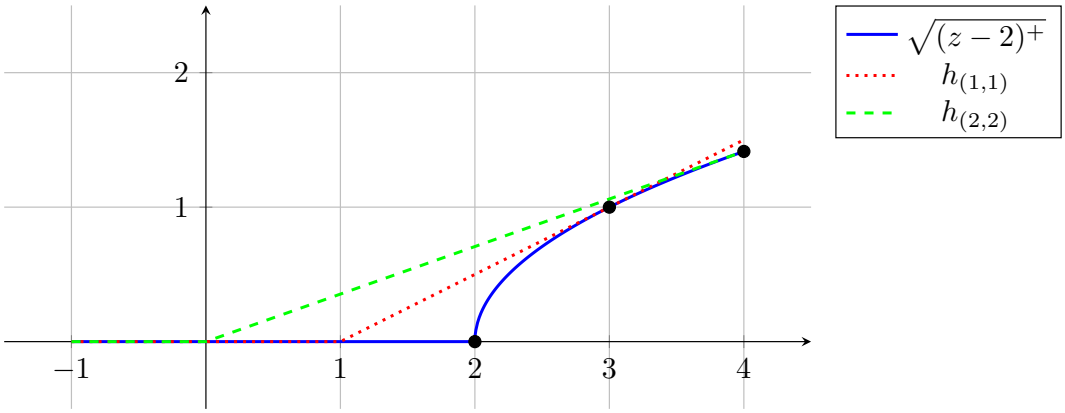
$$h_{(1,1)}(z) = \frac{1}{2} (z - 1)^+, \quad h_{(2,2)}(z) = \frac{\sqrt{2}}{4} \, (z)^+ .$$

We depict the shifted function $f((z-\beta)^+)$ and $h_{(\ell,\rho)}(z)$ in Fig. 3.1. In particular, from Fig. 3.1a, we observe that, as expected, each $h_{(\ell,\rho)}(z)$, $(\ell, \rho) \in L_1$ coincides with $f((z-\beta)^+)$ when $z \in \{\ell + \beta - 1, \ell + \beta\}$. From Fig. 3.1b, we observe that each $h_{(\ell,\rho)}(z)$, $(\ell, \rho) \in L_2$ has an integer z-intercept and it coincides with $f((z-\beta)^+)$ only at the point $(\ell + \beta, f(\ell))$. □

In what follows, we study the hypograph of $h_{(\ell,\rho)}$ and its relationship with $\mathcal{X}$. For



(a) $f((z-\beta)^+)$ and $h_{(\ell,\rho)}(z), (\ell,\rho) \in L_1$



(b) $f((z-\beta)^+)$ and $h_{(\ell,\rho)}(z), (\ell,\rho) \in L_2$

Figure 3.1: A numerical example of the shifted function $f((z-\beta)^+)$ and $h_{(\ell,\rho)}(z)$.

$(\ell, \rho) \in L_1 \cup L_2$, define

$$\mathcal{X}_{(\ell,\rho)} := \left\{ (\theta, x) \in \mathbb{R} \times \{ 0,1 \}^n : \theta \leq h_{(\ell,\rho)}(x^{\mathsf{T}}\xi) \right\}.$$

First, we observe that $\{\mathcal{X}_{(\ell,\rho)} : (\ell,\rho) \in L_1\}$ provides a reformulation for $\mathcal{X}$, while $\{\mathcal{X}_{(\ell,\rho)} : (\ell,\rho) \in L_2\}$ provides a set of valid inequalities.

Lemma 9. For a given $x \in \{ 0,1 \}^n$, we have

$$f\left(\left(x^{\mathsf{T}}\xi - \beta\right)^+\right) = \min_{(\ell,\rho)\in L_1} h_{(\ell,\rho)}(x^{\mathsf{T}}\xi) \leq \min_{(\ell,\rho)\in L_2} h_{(\ell,\rho)}(x^{\mathsf{T}}\xi). \tag{3.14}$$

Proof. See Appendix B.2 □

The previous Lemma implies

$$f\left(\left(x^{\mathsf{T}}\xi - \beta\right)^+\right) = \min_{(\ell,\rho)\in L_1} h_{(\ell,\rho)}(x^{\mathsf{T}}\xi) = \min_{(\ell,\rho)\in L_1\cup L_2} h_{(\ell,\rho)}(x^{\mathsf{T}}\xi).$$

Second, we notice that the EPIs arising from $\{ \mathcal{X}_{(\ell,\rho)} : (\ell,\rho) \in L_2 \}$ and $\{ \mathcal{X}_{(\ell,\rho)} : (\ell,\rho) \in L_1 \}$ are sufficient to describe $\mathbf{conv}\,(\mathcal{X})$.

Theorem 9. It holds that

$$\mathbf{conv}(\mathcal{X}) = \bigcap_{(\ell,\rho)\in L_1\cup L_2} \mathbf{conv}(\mathcal{X}_{(\ell,\rho)}).$$

We defer the proof of Theorem 9 to Appendix B.3 and discuss its implication on the separation of $\mathbf{conv}(\mathcal{X})$. To this end, by Theorem 9, we only need to separate from $\mathbf{conv}(\mathcal{X}_{(\ell,\rho)})$ for all $(\ell,\rho) \in L_1 \cup L_2$, where $|L_1| = |\mathcal{Z}| - \beta$, $|L_2| \leq \beta$ and so $|L_1| + |L_2| \leq |\mathcal{Z}| \leq n$. In addition, for each (ℓ,ρ), $h_{(\ell,\rho)}(x^{\mathsf{T}}\xi)$ is supermodular in x by Lemma 7. Hence, by Theorem 8, we can separate from each $\mathbf{conv}(\mathcal{X}_{(\ell,\rho)})$ by running a sorting algorithm among the n entries of a given solution $\widehat{x}$. In addition, the resulting order is valid for all (ℓ,ρ). Therefore, we can separate from $\mathbf{conv}(\mathcal{X})$ by running a *single* sorting and checking the violation of at most n inequalities.

When applying the BD algorithm to solve (DRC), we add a Benders feasibility cut (3.13) if the incumbent solution $(\widehat{x}, \widehat{\gamma}, \widehat{z})$ is infeasible to (DSP_j). In the meantime, we add more inequalities by separating $(\widehat{x}, \widehat{\gamma}, \widehat{z})$ from $\mathbf{conv}(\mathcal{X}^j_{i^*})$ with

$$\mathcal{X}^j_{i^*} := \left\{ (x, \gamma, z_j) \in \{ 0,1 \}^n \times \mathbb{R}^2 : z_j + \gamma \leq \left(\left(x^{\mathsf{T}}\widehat{\xi}^j_{i^*} - v_{i^*} + 1\right)^+\right)^{1/p} \right\},$$

where i^* is identified in Proposition 12. Since $(\cdot)^{1/p}$ is strictly concave and increasing, $\mathcal{X}^j_{i^*}$ is isomorphic to $\mathcal{X}$ with $z_j + \gamma$ and $v_{i^*} - 1$ playing the roles of θ and β, respectively. As a consequence, separating from $\mathbf{conv}(\mathcal{X}^j_{i^*})$ is efficient. In addition, with regard to any (ℓ, ρ), a valid inequality takes the following form:

$$-z_j - \gamma \geq \Phi^j(\varnothing) + \sum_{k=1}^{n} \left[\Phi^j(\mathcal{T}_k) - \Phi^j(\mathcal{T}_{k-1})\right] x_{\sigma_k}, \tag{3.15}$$

$$\text{where} \quad \Phi^j(x) := \frac{\ell^{1/p}}{\ell + \rho} \min\left\{-x^\mathsf{T}\widehat{\xi}^j_{i^*} - v_{i^*} - \rho + 1,\ 0\right\},$$

$$\Phi^j(\mathcal{T}_k) - \Phi^j(\mathcal{T}_{k-1}) = \begin{cases} -\frac{\ell^{1/p}}{\ell+\rho} & \text{if } \Phi^j(\mathcal{T}_{k-1}) < 0 \\ \Phi^j(\mathcal{T}_k) & \text{if } \Phi^j(\mathcal{T}_{k-1}) = 0 \text{ and } \Phi^j(\mathcal{T}_k) < 0 \\ 0 & \text{if } \Phi^j(\mathcal{T}_k) = 0, \end{cases}$$

and σ is a permutation of $[n]$ such that $\widehat{x}_{\sigma_1} \geq \widehat{x}_{\sigma_2} \geq \cdots \geq \widehat{x}_{\sigma_n}$. In our implementation, we perform this separation via the lazy callback.

3.3.2 Valid inequalities from cross scenarios

Using the fact that $x^\mathsf{T}\xi$ is an integer, we study valid inequalities arising from multiple scenarios $j \in [N]$ and multiple coverings $i \in [I]$. To this end, we focus on a set $\mathcal{S}$ consisting of J "base" inequalities and other domain restrictions:

$$\mathcal{S} := \left\{ (x, \gamma, z) \in \{0,1\}^n \times \mathbb{R}^{1+J} : \begin{array}{c} -z_j - \gamma \geq d^j\left(x^\mathsf{T}\bar{\xi}^j\right) + d^j_0, \quad \forall j \in [J] \\ z_j \leq 0, \quad \forall j \in [J] \end{array} \right\}, \tag{3.16}$$

where $d^j_0 \in \mathbb{R}$ and $d^j < 0$ represent constant parameters. The base inequalities can come from two sources. First, the Benders feasibility cut (3.13) implies that

$$\begin{aligned} -z_j - \gamma &\geq \phi^j(\varnothing) + \sum_{k=1}^{n} \left[\phi^j(\mathcal{T}_k) - \phi^j(\mathcal{T}_{k-1})\right] x_{\sigma_k} \\ &= \phi^j(\varnothing) + \sum_{k:\phi^j(\mathcal{T}_{k-1})<0} (\mathbf{1}^\mathsf{T}\widehat{\mu}_1) x_{\sigma_k} + \sum_{\substack{k:\phi^j(\mathcal{T}_{k-1})=0, \\ \phi^j(\mathcal{T}_k)<0}} \phi^j(\mathcal{T}_k) x_{\sigma_k} + \sum_{k:\phi^j(\mathcal{T}_k)=0} 0 \cdot x_{\sigma_k} \\ &\geq \phi^j(\varnothing) + (\mathbf{1}^\mathsf{T}\widehat{\mu}_1) x^\mathsf{T}\bar{\xi}^j, \end{aligned}$$

where $\bar{\xi}^j$ denotes a binary vector, in which $\bar{\xi}^j_{\sigma_k} = 1$ if and only if $\phi^j(\mathcal{T}_k) < 0$, and the last inequality is because $\phi^j(\mathcal{T}_k) \geq \mathbf{1}^\mathsf{T}\widehat{\mu}_1$ if $\phi^j(\mathcal{T}_{k-1}) = 0$ and $\phi^j(\mathcal{T}_k) < 0$. As a result,

(3.13) generates a base inequality with $d_0^j = \phi^j(\varnothing)$ and $d^j = \mathbf{1}^\mathsf{T}\widehat{\mu}_1 < 0$. Second, the single-scenario valid inequality (3.15) implies

$$\begin{aligned}
-z_j - \gamma &\geq \Phi^j(\varnothing) + \sum_{k=1}^{n} \left[\Phi^j(\mathcal{T}_k) - \Phi^j(\mathcal{T}_{k-1})\right] x_{\sigma_k} \\
&= \Phi^j(\varnothing) + \sum_{k:\Phi^j(\mathcal{T}_{k-1})<0} \left(-\frac{\ell^{1/p}}{\ell+\rho}\right) x_{\sigma_k} + \sum_{\substack{k:\Phi^j(\mathcal{T}_{k-1})=0,\\ \phi^j(\mathcal{T}_k)<0}} \Phi^j(\mathcal{T}_k) x_{\sigma_k} + \sum_{k:\Phi^j(\mathcal{T}_k)=0} 0 \cdot x_{\sigma_k} \\
&\geq \Phi^j(\varnothing) + \left(-\frac{\ell^{1/p}}{\ell+\rho}\right) x^\mathsf{T}\bar{\xi}^j,
\end{aligned}$$

where $\bar{\xi}^j$ denotes a binary vector, in which $\bar{\xi}^j_{\sigma_k} = 1$ if and only if $\Phi^j(\mathcal{T}_k) < 0$. As a result, (3.15) generates a base inequality with $d_0^j = \Phi^j(\varnothing)$ and $d^j = -\ell^{1/p}/(\ell+\rho) < 0$. We remark that the number of base inequalities, J, need not to be the same as the number of scenarios, N. This is because a single scenario can generate multiple base inequalities with regard to different coverings.

We employ the mixing scheme [GP01] to derive valid inequalities for $\mathcal{S}$. This scheme produces a new inequality by combining existing ones. Specifically, consider the following mixing set $\mathcal{MIX}$ consisting of J base inequalities:

$$\mathcal{MIX} := \left\{ x \in \{0,1\}^n : \begin{array}{ll} f^j(x) + Bg^j(x) \geq u^j, & \forall j \in [J] \\ f^j(x) \geq 0, & \forall j \in [J] \end{array} \right\},$$

where $B \in \mathbb{R}_+, u^j \in \mathbb{R}$, and $g^j(x) \in \mathbb{Z}$. Then, [GP01] derived the following mixed inequality.

Theorem 10 (Theorem 2 in [GP01])**.** The following mixed inequality is valid for $\mathcal{MIX}$:

$$\max_{j\in[J]}\{f^j(x)\} \geq \sum_{j=1}^{J} (\nu^j - \nu^{j-1})(\tau^j - g^j(x)),$$

where $\nu^0 := 0$, $\tau^j := \lceil u^j/B \rceil$, $\nu^j := u^j - B(\tau^j - 1)$ for all $j \in [J]$, and ν^j's are sorted so that $\nu^j \geq \nu^{j-1}$ for all $j \in [J]$. □

To apply the mixing scheme, we divide both sides of the base inequalities in (3.16) by $-d^j$ to obtain

$$\frac{z_j}{d^j} + x^\mathsf{T}\bar{\xi}^j \geq -\frac{d_0^j + \gamma}{d^j}, \quad \forall j \in [J].$$

This would obey the structure of $\mathcal{MIX}$ if $-(d_0^j + \gamma)/d^j$ (counterpart of u^j) were a constant, or equivalently, if variable γ were a constant. Although this is not true in general, we

observe that, without loss of optimality, γ can take only a finite set of values. It is worth mentioning that a similar result appeared in [Ho-+21, Lemma 1].

Proposition 13. There exists an optimal solution (x^*, γ^*, z^*) to (DRC), using the reformulation Z defined in (3.6), such that

$$\begin{aligned} \gamma^* &= -\mathsf{VaR}_{1-\epsilon}\left[-g(x^*, \widehat{\xi})\right] \\ \text{and} \quad z_j^* &= \min\left\{\, 0,\, g(x^*, \widehat{\xi}^j) - \gamma^* \,\right\}, \quad \forall j \in [N], \end{aligned}$$

where $g(x, \xi) = \min_{i\in[I]}\Big((x^\mathsf{T}\xi_i - v_i + 1)^+\Big)^{1/p}$.

Proof. By Proposition 6 and the definition (3.5) of CVaR, we recast the set Z as

$$Z = \left\{ x \in \{\,0,1\,\}^n : \frac{\delta}{\epsilon} + \min_{\gamma'}\Big\{\gamma' + \frac{1}{N\epsilon}\sum_{j\in[N]}\Big(-g(x,\widehat{\xi}^j) - \gamma'\Big)^+\Big\} \le 0 \right\}.$$

Then, by the definition (3.5) of CVaR, γ' equals $\mathsf{VaR}_{1-\epsilon}\Big[-g(x^*, \widehat{\xi})\Big]$ at optimality. The conclusion follows. □

By Proposition 13, we can restrict the value of γ to be in a discrete set $\{r \in \mathbb{R} : r^p \in \{0\}\cup[n]\}$ because $(x^*)^\mathsf{T}\widehat{\xi}_i$ and v_i are integers for all $i \in [I]$. In addition, since $g(x, \widehat{\xi}) \le g(\mathbf{1}, \widehat{\xi})$ almost surely for all $x \in \{\,0,1\,\}^n$ and (DRC) requires that $\mathsf{CVaR}_{1-\epsilon}\big[-g(x,\widehat{\xi})\big] \le -\delta/\epsilon$, we can further restrict the value of γ to be that

$$\gamma \in \left\{ r \colon \lceil \left(\frac{\delta}{\epsilon}\right)^p \rceil^{1/p} \le r \le -\mathsf{VaR}_{1-\epsilon}(-g(\mathbf{1},\widehat{\xi})),\ r^p \in [n] \right\}.$$

For ease of exposition, we denote this set as $\{r_k : k \in \Gamma\}$, where Γ is the set of indices, and let r_1 be the smallest element. It follows that $\gamma \ge r_1$ at optimality and the base inequalities in (3.16) imply

$$\frac{z_j}{d^j} + x^\mathsf{T}\bar{\xi}^j \ge -\frac{d_0^j + r_1}{d^j}, \quad \forall j \in [J].$$

Then, applying the mixing scheme in Theorem 10 produces a mixed inequality for $\mathcal{S}$:

$$\max_{j\in[J]}\left\{\, \frac{z_j}{d^j} \,\right\} \ge \sum_{j=1}^{J}(\nu_1^j - \nu_1^{j-1})(\tau_1^j - x^\mathsf{T}\bar{\xi}^j), \tag{3.17}$$

where $\tau_1^j := \lceil -\frac{r_1 + d_0^j}{d^j} \rceil$, $\nu_1^0 := 0$, and $\nu_1^j := -\frac{r_1 + d_0^j}{d^j} - (\tau_1^j - 1)$ such that $\nu_1^j \ge \nu_1^{j-1}$ for

all $j \in [J]$. Nevertheless, the mixed inequality (3.17) is potentially weak because it simply relaxes γ to be r_1 and ignores the possibility of γ taking other values in $\{r_k : k \in \Gamma\}$. To strengthen it, we notice that γ, which equals one and only one r_k for $k \in \Gamma$, admits a binary encoding:

$$\gamma = \sum_{k\in\Gamma} r_k \gamma_k, \quad \sum_{k\in\Gamma} \gamma_k = 1,$$

where $\{\gamma_k \in \{0,1\} : k \in \Gamma\}$ are (auxiliary) binary variables that determine the value of γ. Accordingly, we recast set $\mathcal{S}$ as

$$\mathcal{S}_{\mathbb{B}} := \left\{ (x,\gamma,z) \in \{0,1\}^{n+|\Gamma|} \times \mathbb{R}^J_- : \begin{array}{l} -z_j - \sum_{k\in\Gamma} r_k\gamma_k \geq d^j \left(x^{\mathsf{T}} \bar{\xi}^j\right) + d^j_0, \quad \forall j \in [J] \\ \sum_{k\in\Gamma} \gamma_k = 1 \end{array} \right\}.$$

As a consequence, the mixed inequality (3.17) is valid for a restriction of $\mathcal{S}_{\mathbb{B}}$ when fixing $\gamma_1 = 1$, i.e., $\mathcal{S}_{\mathbb{B}} \cap \{(x,\gamma,z) : \gamma_1 = 1\}$. Hence, we can view (3.17) as a *seed* inequality and strengthen it by *lifting* variables $\gamma_2, \ldots, \gamma_{|\Gamma|}$ back (see, e.g., [WN14; GNS00] for more on the lifting scheme).

Theorem 11. The following lifted inequality is valid for $\mathcal{S}_{\mathbb{B}}$:

$$\max_{j\in[J]} \left\{ \frac{z_j}{d^j} \right\} \geq \sum_{j=1}^{J} (\nu_1^j - \nu_1^{j-1})(\tau_1^j - x^{\mathsf{T}}\bar{\xi}^j) + \sum_{k\in\Gamma} \alpha_k \gamma_k, \tag{3.18}$$

$$\text{where} \quad \alpha_k := \left(\min_{j\in[J]} \min\{\nu_k^j - \nu_1^J, 0\} + \sum_{j=1}^{J} (\nu_1^j - \nu_1^{j-1})(\tau_k^j - \tau_1^j) \right)^+, \quad \forall k \in \Gamma,$$

$\tau_k^j := \lceil -\frac{r_k + d_0^j}{d^j} \rceil$, $\nu_1^0 = 0$, and $\nu_k^j := -\frac{r_k + d_0^j}{d^j} - (\tau_k^j - 1)$ such that $\nu_1^j \geq \nu_1^{j-1}$ for all $j \in [J]$.

Proof. Please see Appendix B.4. □

We remark that, thanks to constraint $\sum_{k\in\Gamma} \gamma_k = 1$, the lifted inequality in Theorem 11 is sequence-*independent* (cf. [GNS00]). In our implementation, we generate the base inequalities in (3.16) from the single-scenario valid inequality (3.15). We add the lifted inequalities via the lazy callback.

3.4 Extension to Knapsack Chance Constraint

In this section, we extend the reformulation and valid inequalities for (DRC) to knapsack chance constraints with random, binary-valued coefficients. Specifically, consider the following ambiguous knapsack chance constraint

$$\inf_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left\{ x^{\mathsf{T}}\widetilde{\xi}_i \le v_i,\ \forall i \in [I] \right\} \ge 1-\epsilon, \tag{KCC}$$

where the Wasserstein ambiguity set $\mathcal{P}$ is defined in (3.3), $\widetilde{\xi} \in \Xi = \{0,1\}^{I\times n}$, and $v_i \in [n-1]$ for all $i \in [I]$. In this case, we can follow the proof of Proposition 6 to show that

$$\min_{\xi:\ x^{\mathsf{T}}\xi_i \ge v_i+1} \|\xi - \widehat{\xi}^j\|_p = \left(\left(v_i + 1 - x^{\mathsf{T}}\widehat{\xi}_i^j\right)^+\right)^{1/p} + \chi\left(x^{\mathsf{T}}\mathbf{1} \ge v_i+1\right),$$

where $\chi(\cdot)$ denotes a characteristic function, which takes value 0 if $x^{\mathsf{T}}\mathbf{1} \ge v_i+1$ and takes value $+\infty$ otherwise. Hence, if we assume $v_1 \le v_2 \le \cdots \le v_I$ w.l.o.g. then (KCC) is equivalent to $\bigcup_{k=0}^{I} Z_k$, where $Z_0 := \{x \in \{0,1\}^n : x^{\mathsf{T}}\mathbf{1} \le v_1\}$ and, for all $k \in [I]$, Z_k resembles the reformulation Z for (DRC) in Proposition 6. That is,

$$Z_k := \left\{x\colon v_k + 1 \le x^{\mathsf{T}}\mathbf{1} \le v_{k+1}\right\} \cap \left\{ x\colon \begin{array}{l} \exists\, \gamma \in \mathbb{R}_+, z \in \mathbb{R}_-^N : \\ \delta - \gamma\epsilon \le \dfrac{1}{N}\displaystyle\sum_{j\in[N]} z_j, \\ z_j + \gamma \le \min\limits_{i\in[k]} \left(\left(v_i + 1 - x^{\mathsf{T}}\widehat{\xi}_i^j\right)^+\right)^{1/p}, \forall j \in [N] \end{array} \right\},$$

where $v_{I+1} := n$. As a consequence, we can readily adapt the valid inequalities in Section 3.3 to solve optimization problems involving (KCC). For example, under the complement $y := \mathbf{1} - x$, the final constraints in the definition of Z_k can be rewritten as

$$z_j + \gamma \le \left(\left(y^{\mathsf{T}}\widehat{\xi}_i^j - \mathbf{1}^{\mathsf{T}}\widehat{\xi}_i^j + v_i + 1\right)^+\right)^{1/p},$$

which is precisely the setting studied in Section 3.3 if $(\mathbf{1}^{\mathsf{T}}\widehat{\xi}_i^j - v_i - 1) \ge 1$. Hence, the valid inequalities (3.15) and (3.18) can be adapted to compute (KCC) more effectively. On the other hand, when $\beta' := (\mathbf{1}^{\mathsf{T}}\widehat{\xi}_i^j - v_i - 1) \le 0$, it can be shown that the convex hull of the mixed-integer set

$$\mathcal{X}' := \left\{ (\theta, x) \in \mathbb{R} \times \{0,1\}^n : \theta \le f(x^{\mathsf{T}}\xi - \beta') \right\}$$

coincides with the continuous relaxation of its piecewise linear reformulation, i.e.,

$$\textbf{conv}(\mathcal{X}') = \left\{ (\theta, x) \in \mathbb{R} \times [0,1]^n \colon \theta \leq a_\ell(x^\mathsf{T}\xi - \beta') + b_\ell,\ \forall \ell \in \{ 1-\beta', \ldots, n-\beta' \} \right\}.$$

Finally, we notice that a special case of (KCC) is an ambiguous set packing chance constraint, in which $v_i = 1$ for all $i \in [I]$. Hence, our reformulation and valid inequalities also apply to set packing chance constraints with random set membership.

3.5 Numerical Experiments

We demonstrate the out-of-sample performance of (DRC) and the effectiveness of our two-stage reformulation and valid inequalities in the following four numerical experiments:

1. Section 3.5.1 reports the out-of-sample performance of (DRC) with various radii δ and data sizes N when the entries of $\widetilde{\xi}$ are i.i.d.;
2. Section 3.5.2 examines correlated $\widetilde{\xi}$ entries, using a context of medical facility location under natural disasters;
3. Section 3.5.3 evaluates the value of incorporating the support information $\Xi = \{0,1\}^{I\times n}$ into (DRC) as opposed to relaxing it to be $\Xi = \mathbb{R}^{I\times n}$;
4. Section 3.5.4 demonstrates the strength of the valid inequalities derived in Section 3.3.

All experiments in this section are conducted using the Python API of Gurobi $9.0.2$ on a single core of an Intel Xeon Gold 6154 Processor provided by the UM Great Lakes cluster.

3.5.1 Independent uncertain parameters

In this experiment, we generate *testing* data of $\widetilde{\xi}$ from the Bernoulli distribution with i.i.d. entries, and the probability of success $q_i := \mathbb{P}_{\text{true}}\left\{ \widetilde{\xi}_{ij} = 1 \right\}$ is uniformly sampled from the interval $[0.4, 0.8]$. In the out-of-sample evaluation, this specific choice of $\mathbb{P}_{\text{true}}$ facilitates the

feasibility checking of a given solution $x \in \{0,1\}^n$ because

$$\begin{aligned}\mathbb{P}_{\text{true}}\left\{ x^{\mathsf{T}}\widetilde{\xi}_i \geq v_i,\, \forall i \in [I] \right\} &= \prod_{i=1}^{I} \mathbb{P}_{\text{true}}\left\{ x^{\mathsf{T}}\widetilde{\xi}_i \geq v_i \right\} \\ &= \prod_{i=1}^{I}\sum_{s=v_i}^{n} \mathbb{P}_{\text{true}}\left\{ \sum_{k:\, x_k=1} \widetilde{\xi}_{ik} = s \right\} \\ &= \prod_{i=1}^{I}\sum_{s=v_i}^{n} \binom{\mathbf{1}^{\mathsf{T}}x}{s} q_i^s (1-q_i)^{(\mathbf{1}^{\mathsf{T}}x)-s} \end{aligned} \tag{3.19}$$

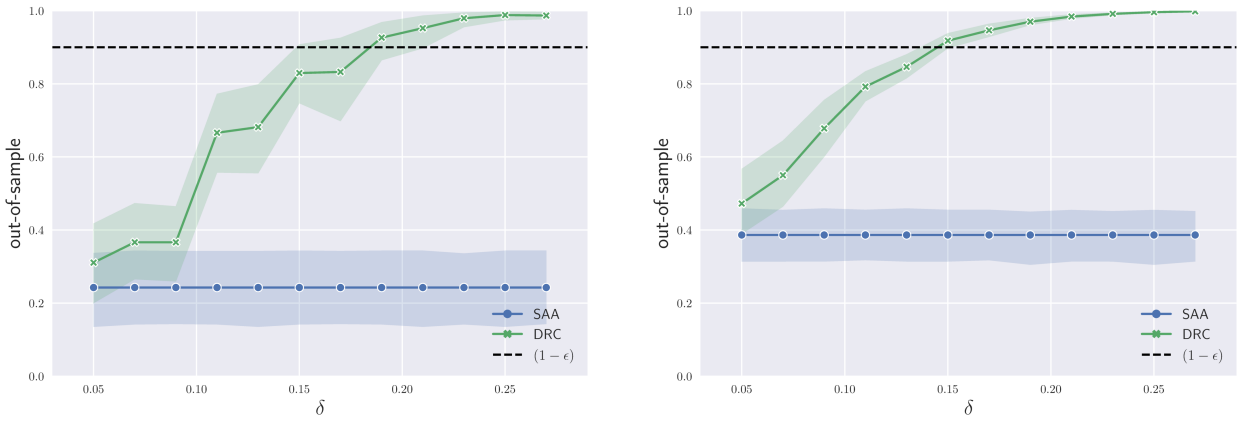
admits a closed-form expression. In addition, we generate the *training* historical data $\{\widehat{\xi}_j\}_{j=1}^N$ by adding Gaussian noise to $\widetilde{\xi}$:

$$\widehat{\xi}_{ik} := \text{round}(\widetilde{\xi}_{ik} + 0.25 \cdot \widetilde{e}), \quad \forall i \in [I], k \in [n],$$

where $\widetilde{e}$ is a standard Gaussian random variable and round$(\cdot)$ rounds real numbers to $\{0,1\}$. For a given solution $x \in \{0,1\}^n$, checking whether it is feasible for chance constraint (3.2) simplifies to computing its out-of-sample performance using (3.19) and comparing with the risk level $1-\epsilon$. We compare the out-of-sample performance of (DRC) with the sample average approximation formulation:

$$\begin{aligned} \min\ & c^{\mathsf{T}}x, \\ \text{s.t.}\ & \frac{1}{N}\sum_{j=1}^{N} z_j \geq 1-\epsilon, \\ & x^{\mathsf{T}}\widehat{\xi}_i^j - v_i \geq v_m(z_j - 1), \quad \forall i \in [I], j \in [N], \\ & x \in \{0,1\}^n, z \in \{0,1\}^N. \end{aligned} \tag{SAA}$$

Random test instances with $n = 30, I = 10$ are generated in the aforementioned way. For each instance, $N \in \{100, 200, 300, 400, 500\}$ training data are generated, the RHS $v_i, i \in [I]$ are all 1, coefficients c in the objective function are uniformly sampled from integers between 1 and 100, the radius δ of the Wasserstein ball enumerates $\{0.05, 0.07, 0.09, 0.11, 0.13, 0.15, 0.17, 0.19, 0.21, 0.23, 0.25, 0.27\}$, and the risk level $\epsilon = 0.1$. For each parameter setting, we compute the average of out-of-sample performance, as well as its 90% confidence interval, for the returned solution over 5 different random instances and study how the radius δ and the sample size N affect the out-of-sample performance of (DRC) and (SAA).



(a) $\epsilon = 0.1, N = 20$ (b) $\epsilon = 0.1, N = 300$

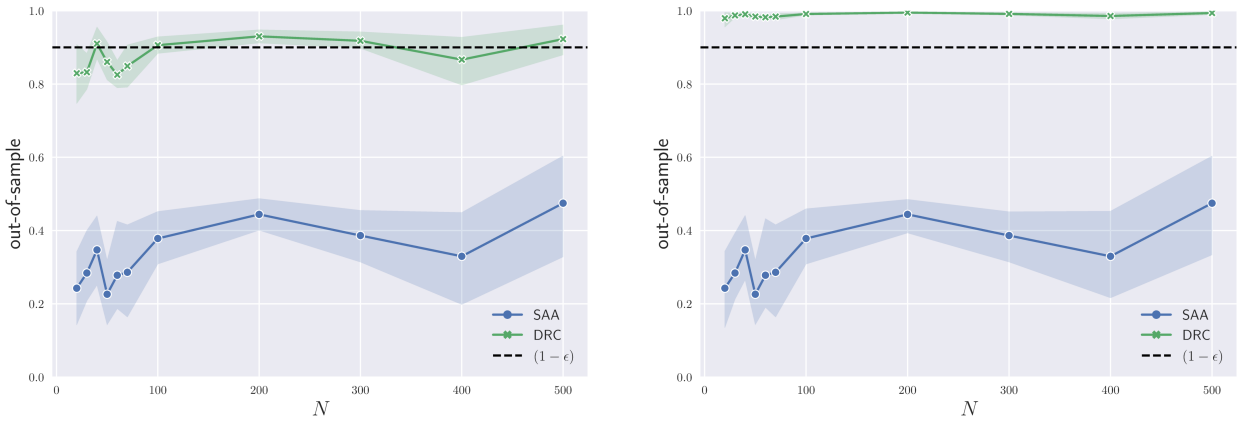
Figure 3.2: Out-Of-Sample Performance of (DRC) and (SAA) as a function of δ.

3.5.1.1 Impact of the radius δ on out-of-sample performance

Fig. 4.5 displays the impact of δ on out-of-sample performance of the (DRC) and (SAA) optimal solutions. We notice that the out-of-sample performance of (DRC) increases as δ becomes larger. This makes sense because the larger the radius δ is, the more distributions the Wasserstein ball $\mathcal{P}$ includes, and accordingly the more likely (DRC) generates a solution feasible with respect to $\mathbb{P}_{\text{true}}$. The (SAA) model, however, yields a low out-of-sample performance. Specifically, Fig. 4.5a shows that, with a proper choice of δ, (DRC) is able to provide a reliable solution satisfying the chance constraint even when the training data is very limited. In contrast, (SAA) does not return a feasible solution even if the size of training data has a tenfold increase.

3.5.1.2 Impact of the data size N on out-of-sample performance

Fig. 4.6 displays how N affects out-of-sample performance of the (DRC) and (SAA) optimal solutions. We observe that the out-of-sample performance of both models shows an increasing trend as the data size becomes larger. Specifically, (DRC) can return a high-quality solution with a proper choice of radius δ depending on much less data. In contrast, (SAA) does not produce a feasible solution even when 500 training data are provided. In addition, (DRC) is significantly more stable than (SAA), as (DRC) generates a much narrower 90% confidence interval (the shaded area in Fig. 4.6) than (SAA). This demonstrates that (DRC) is capable of generating reliable and stable solutions even with limited training data.



(a) $\epsilon = 0.1, \delta = 0.15$ (b) $\epsilon = 0.1, \delta = 0.23$

Figure 3.3: Out-of-sample Performance of (DRC) and (SAA) as a function of N.

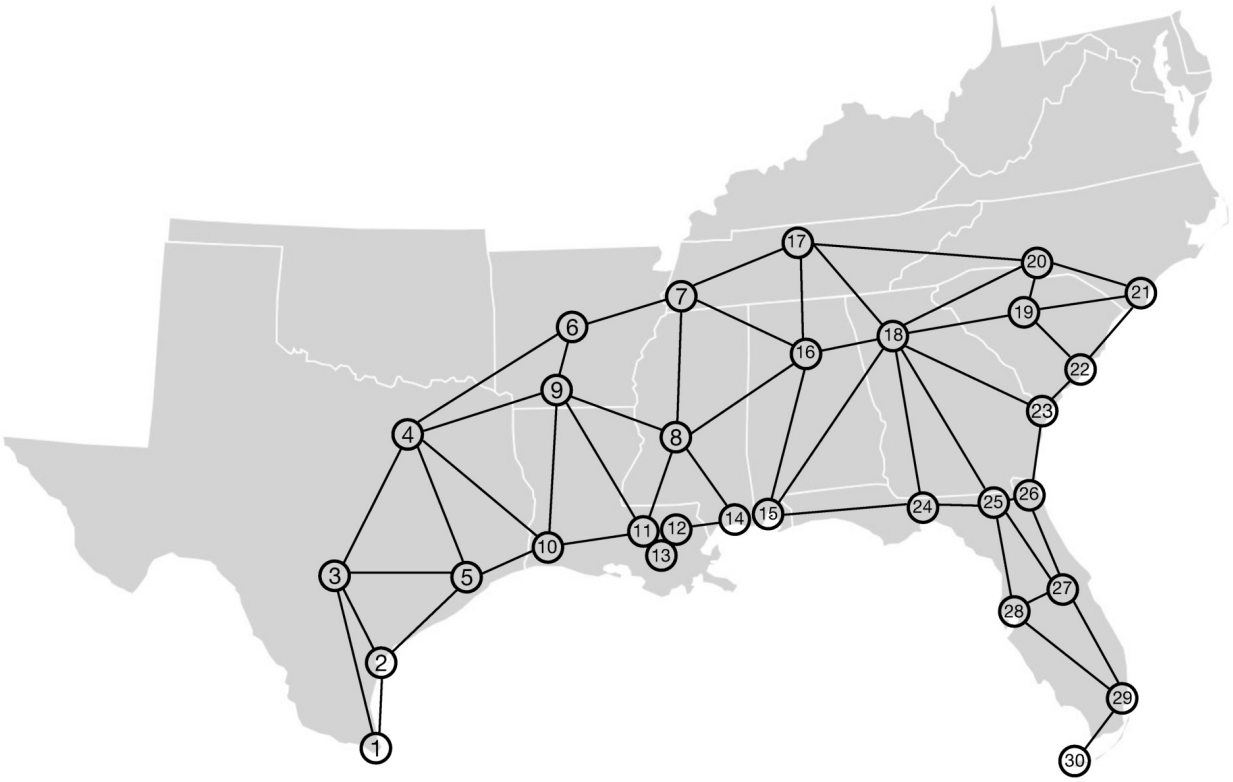


Figure 3.4: Network of a medical facility location case study.

3.5.2 Dependent uncertain parameters

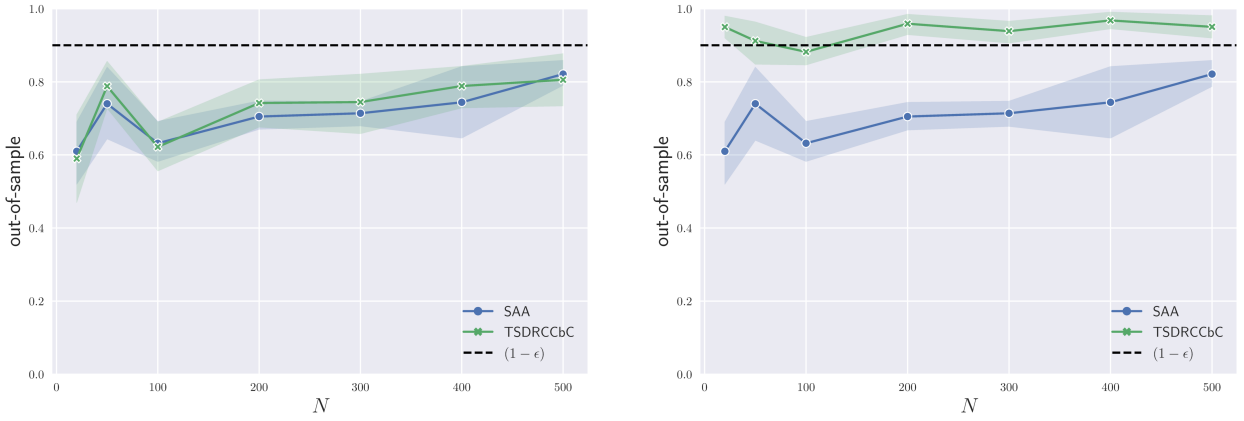
In this experiment, we test (DRC) with dependent uncertain parameters based on a medical facility location problem with natural disasters. Specifically, we consider an undirected network with I nodes and a set $\mathcal{E} \subseteq [I] \times [I]$ of roads connecting these nodes. As a special case of (DRC) with $n = I$, this problem establishes medical facilities among these I nodes in order to cover the medical needs from the same set of nodes under coverage uncertainty. The binary random variables $\widetilde{\xi}_{ij}$, $i, j \in [I]$ indicate whether a facility in node j can cover the need from node i. To generate correlated $\widetilde{\xi}_{ij}$, we consider natural disasters and let $\widetilde{H}_j$ denote a binary random variable such that $\widetilde{H}_j = 1$ indicates a disaster in node j. In addition, we partition the nodes into subsets $\{S_k\}_{k=1}^K$ according to geographic proximity and designate that if a node in a subset is affected by the disaster then all nodes in the same subset are affected, i.e., $\widetilde{H}_j = 1$ for all $j \in S_k$ whenever $\widetilde{H}_i = 1$ for an $i \in S_k$. Furthermore, we consider a random travel time $\widetilde{T}_e$ on each road $e \in \mathcal{E}$ and assume that $\widetilde{T}_e$ depends on the two nodes road e connects: if neither of the nodes is affected by the disaster then $\widetilde{T}_e$ takes a constant value T_e^0; and if either node is affected then $\widetilde{T}_e$ follows a Gaussian distribution $N(4 \cdot T_e^0, 1)$. Accordingly, we define $\widetilde{\xi}_{ij}$ as

$$\widetilde{\xi}_{ij} := (1 - \widetilde{H}_j) \times \mathbb{1} \left\{ \sum_{e \in \textsf{PATH}(j,i)} \widetilde{T}_e \leq T_{\textsf{max}} \right\}, \tag{3.20}$$

where $\textsf{PATH}(j, i)$ represents the shortest path from node j to node i after the disaster and $T_{\textsf{max}}$ denotes the longest time allowed to provide medical service to a disaster node. That is, $\widetilde{\xi}_{ij} = 1$ only when node j is not affected by the disaster and the shortest path from j to i is shorter than $T_{\textsf{max}}$.

Our case study is based on a network of southern US with $n = I = 30$ (cf. [ST20; RT10; VMÖ20]; see Figure 3.4). The partition $\{S_k\}_{k=1}^K$ of these nodes includes subsets $\{\,11, 12, 13\,\}$, $\{\,14, 15\,\}$, $\{\,19, 20\,\}$, $\{\,22, 23\,\}$, $\{\,24, 25, 26\,\}$, $\{\,27, 28\,\}$, and all the remaining subsets are singletons.

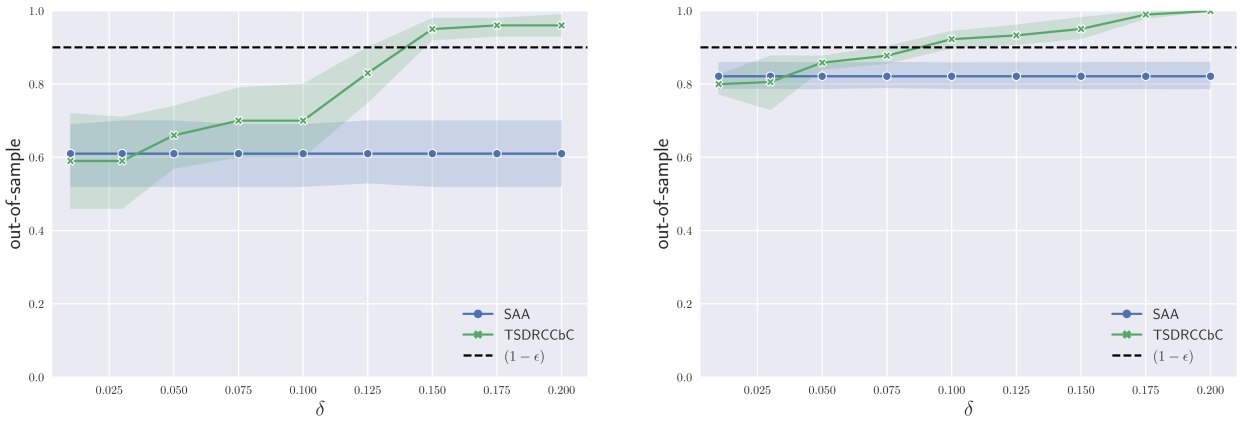
First, to generate $\widetilde{H}_j$, we assign all nodes into three categories: $\mathcal{C}_1 := \{\,2, 5, 11, 13, 14, 15, 29, 30, 21, 22\,\}$, $\mathcal{C}_2 := \{\,1, 3, 10, 12, 25, 26, 27, 28, 23\,\}$, and $\mathcal{C}_3 := [I] \setminus (\mathcal{C}_1 \cup \mathcal{C}_2)$. For each node j, we independently draw a Gaussian random number centered around the category index of j with a standard deviation of 1. For example, we sample from $N(2, 1)$ for node 1 and from $N(1, 1)$ for node 2. Then, we set $\widetilde{H}_j = 1$ if node j generates the smallest Gaussian random number, and we broadcast $\widetilde{H}_j = 1$ in the subset j belongs with. For example, if $j = 1$ then $\widetilde{H}_1 = 1$ and $\widetilde{H}_i = 0$ for all $i \in [I] \setminus \{1\}$; and if $j = 11$ then $\widetilde{H}_{11} = \widetilde{H}_{12} = \widetilde{H}_{13} = 1$ and $\widetilde{H}_i = 0$ for all $i \in [I] \setminus \{11, 12, 13\}$.



(a) $\epsilon = 0.1, \delta = 0.03$ (b) $\epsilon = 0.1, \delta = 0.15$

Figure 3.5: Out-of-sample performance of (DRC) as a function of N.

Second, we generate the random travel times $\widetilde{T}_e$ for all $e \in \mathcal{E}$ based on $\{\widetilde{H}_j\}_{j=1}^I$ as described above and the random coverage $\widetilde{\xi}$ using (3.20) and $T_{\mathsf{max}} := 12$. For each node $i \in [I]$, the coverage level v_i is set to be 4 minus the category index of i. For example, $v_1 = 2$ and $v_2 = 3$.



(a) $\epsilon = 0.1, N = 20$ (b) $\epsilon = 0.1, N = 500$

Figure 3.6: Out-of-sample performance of (DRC) as a function of δ.

Third, to evaluate the out-of-sample performance of (DRC), we generate $N \in \{20, 50, 100, 200, 300, 400, 500\}$ samples as training data, the radius δ of the Wasserstein ball are taken from $\{0.01, 0.03, 0.05, 0.075, 0.1, 0.125, 0.15, 0.175, 0.2\}$, and the risk level $\epsilon = 0.1$. For each parameter setting, we generate 5 different random instances, evaluate the out-of-sample performance using $100,000$ random coverage samples generated in the same way

as the training data, and report the average performance along with the 90% confidence interval in Figures 3.5 and 3.6. Figure 3.5 depicts the out-of-sample performance as a function of the data size N. From this figure, we observe that when δ is small, the performance of (DRC) and (SAA) are similar. As δ increases, e.g., when $\delta = 0.15$, we observe a large improvement in the out-of-sample performance of (DRC) as it requires feasible solutions to satisfy chance constraints with respect to more distributions. In Figure 3.6(a), even with a small training data size, e.g., $N = 20$, (DRC) can return a reliable solution when δ is properly chosen, say $\delta = 0.15$. In contrast, (SAA) did not produce a feasible solution in the sense of out-of-sample performance, even when $N = 500$ as displayed in Figure 3.6(b).

3.5.3 Value of Incorporating Support Information

A key difference between this chapter and Xie [Xie19] lies in the support Ξ of uncertainty parameters. We derive a reformulation Z of (DRC) in (3.6) by assuming a binary-valued support $\Xi = \{0,1\}^{I\times n}$. If we relax it to be $\Xi = \mathbb{R}^{I\times n}$, a different reformulation of the feasible region produced by chance constraint (3.4) follows from Theorem 1 and Proposition 1 of [Xie19]:

$$Z_0 = \left\{ x \in \{0,1\}^n : \begin{array}{l} \exists\, \gamma \in \mathbb{R}_+, \nu \in \mathbb{R}_+, z \in \mathbb{R}_-^N : \\ \delta\nu - \gamma\epsilon \le \dfrac{1}{N}\displaystyle\sum_{j\in[N]} z_j, \\ \|x\|_{p/(p-1)} \le \nu, \\ z_j + \gamma \le \left(x^\intercal \widehat{\xi}_i^j - v_i\right)^+, \forall i \in [I], j \in [N] \end{array} \right\}. \tag{3.21}$$

Intuitively, as Z relies on a more restricted support, it leads to a larger feasible region and accordingly (DRC) becomes less conservative. The following proposition formalizes this intuition.

Proposition 14. For fixed δ and ϵ, it holds that $Z_0 \subseteq Z$.

Proof. Since $v_i \ge 1$ for all $i \in [I]$, $x = 0$ is not feasible to (DRC). Then, we recast Z_0 in

(3.21) as follows by projecting out variable ν:

$$Z_0 = \left\{ x \in \{0,1\}^n : \begin{array}{c} \exists \gamma \in \mathbb{R}_+, z \in \mathbb{R}_-^N : \\ \delta - \gamma\epsilon \le \frac{1}{N}\sum_{j\in[N]} z_j, \\ z_j + \gamma \le \min_{i\in[I]} \frac{\left(x^\mathsf{T}\widehat{\xi}_i^j - v_i\right)^+}{\|x\|_*}, \forall j \in [N] \end{array} \right\},$$

where $\|\cdot\|_* := \|\cdot\|_{p/(p-1)}$ represents the dual norm of $\|\cdot\|_p$. Hence, it suffices to show

$$\frac{\left(x^\mathsf{T}\xi - v\right)^+}{\|x\|_*} \le \left(\left(x^\mathsf{T}\xi - v + 1\right)^+\right)^{1/p}$$

for any $0 \neq x \in \{0,1\}^n$, $\xi \in \{0,1\}^n$, and $v \in [n]$. To this end, we discuss the following two cases.

1. When $x^\mathsf{T}\xi \le v$, we have

$$\frac{\left(x^\mathsf{T}\xi - v\right)^+}{\|x\|_*} = 0 \le \left(\left(x^\mathsf{T}\xi - v + 1\right)^+\right)^{1/p}.$$

2. When $x^\mathsf{T}\xi \ge v + 1$, let $\mathcal{T}$ be the index set of $x \wedge \xi$, i.e., $\mathcal{T} := \{i \in [n] : \xi_i = x_i = 1\}$, then $|\mathcal{T}| \ge v + 1$. Let $\mathcal{T}'$ be $\mathcal{T}$ with arbitrary v elements removed and ξ' be the indicating vector of $\mathcal{T}'$. That is, $\xi' \in \{0,1\}^n$ and, for all $i \in [n]$, $\xi'_i = 1$ if and only if $i \in \mathcal{T}'$. Then, $x^\mathsf{T}\xi' = \mathbf{1}^\mathsf{T}\xi' = x^\mathsf{T}\xi - v$ and $\|\xi'\|_p = (\mathbf{1}^\mathsf{T}\xi')^{1/p}$. We finish the proof by noticing that:

$$\begin{aligned} (x^\mathsf{T}\xi - v) &= x^\mathsf{T}\xi' \\ &\le \|\xi'\|_p \|x\|_* \\ &= \left(\mathbf{1}^\mathsf{T}\xi'\right)^{1/p} \|x\|_* \\ &= \left(x^\mathsf{T}\xi - v\right)^{1/p} \|x\|_* \le \left(x^\mathsf{T}\xi - v + 1\right)^{1/p} \|x\|_*, \end{aligned}$$

where the first inequality is by Cauchy–Schwarz inequality and the last inequality is because $x^\mathsf{T}\xi - v \ge 0$.

□

Next, we demonstrate Proposition 14 numerically using a set of random test instances, generated in the same way as in Section 3.5.1, with various n, I, N, δ and the same risk

				bin			cont		
n	I	N	δ	Optval	Time	OOS	Optval	Time	OOS
20	10	50	0.01	136.0	1.31	0.82	270.4	8.13	0.98
			0.05	159.4	0.94	0.88	557.4	7.02	0.98
			0.1	203.4	0.62	0.90	INF	0.55	N/A
		100	0.01	115.4	2.38	0.88	239.8	15.58	0.98
			0.05	128.8	1.39	0.91	552.4	18.00	0.99
			0.1	210.6	1.16	0.94	1,020.0	1.66	1.00
	25	50	0.01	200.2	2.32	0.84	365.2	15.99	0.98
			0.05	216.2	1.50	0.87	615.7	9.96	0.99
			0.1	258.0	0.99	0.89	INF	0.67	N/A
		100	0.01	213.8	5.50	0.85	409.2	49.20	0.98
			0.05	233.8	2.98	0.90	841.7	8.17	0.99
			0.1	304.8	1.83	0.95	INF	1.16	N/A
30	10	50	0.01	150.2	11.24	0.83	236.6	109.86	0.96
			0.05	161.6	5.83	0.87	404.0	109.56	0.99
			0.1	204.2	1.63	0.93	752.6	44.02	0.99
		100	0.01	147.0	13.62	0.90	255.2	191.69	0.97
			0.05	162.6	5.58	0.91	459.2	108.07	0.99
			0.1	223.6	2.32	0.95	959.2	33.22	1.00
	25	50	0.01	205.6	19.16	0.85	322.6	177.63	0.96
			0.05	226.2	8.46	0.88	544.8	96.73	0.99
			0.1	252.8	3.33	0.91	1,253.6	7.09	0.99
		100	0.01	173.2	23.38	0.88	305.8	588.23	0.98
			0.05	199.4	10.16	0.91	579.8	282.40	0.99
			0.1	261.4	3.68	0.96	1,201.3	24.23	1.00
Average				197.84	5.47	0.89	578.41	79.53	0.99

Table 3.1: Comparison between optimizing over Z_0 and Z

level $\epsilon = 0.1$. For each problem size, we solve our two-stage reformulation in Proposition 11 (denoted as `bin`) and Xie [Xie19]'s reformulation $\min_{x \in Z_0}\{c^\mathsf{T} x\}$ (denoted as `cont`). We solve five random instances for each problem size and report the average optimal value, out-of-sample constraint satisfaction probability (OOS), and CPU time in Table 3.1. An "INF" is placed in the column "Optval" if all five random instances are infeasible, and "N/A" represents "not available". From Table 3.1, we observe that for the same problem size, Xie [Xie19]'s reformulation always returns a higher optimal value than ours. For example, for $n = 30, I = 25, N = 100, \delta = 0.01$, `bin` returns an optimal value of 173.2, while `cont`'s optimal value 305.8 is almost twice as large. Additionally, for $n = 20, I = 25, N = 100, \delta = 0.1$, `cont` is infeasible in all five instances while `bin` always remains feasible and its average optimal value 304.8 is even lower than `cont`'s with a smaller radius $\delta = 0.01$. Besides, from Table 3.1 we notice that the out-of-sample constraint satisfaction probability of `cont` is nearly 1.00 across all instances, while that of `bin` is closer to the target probability (i.e., 0.90). These computational results align with the fact $Z_0 \subseteq Z$ and demonstrate that incorporating support information can make (DRC) significantly less conservative.

3.5.4 Strength of the two-stage reformulation and valid inequalities

We demonstrate the strength of our two-stage reformulation, and the single- and cross-scenario valid inequalities. Random test instances with $n \in \{60, 80\}$, $I \in \{70, 90\}$ are generated, each of which is paired with $N \in \{\,50, 100, 200\,\}$, $\delta \in \{0.05, 0.1, 0.2, 0.3\}$, and $\epsilon \in \{0.05, 0.1\}$. The numerical results are reported in Tables 3.2–3.4, where we use `2-Stg`, `+single`, and `+cross` to denote using the two-stage reformulation in Proposition 11 only, two-stage reformulation with single-scenario inequalities, and two-stage reformulation with both single- and cross-scenario inequalities, respectively. In addition, we report average CPU time (in seconds) and average optimality gap over 5 random test instances, after they are terminated due to either a proof of optimality/infeasibility (INF) or a timelimit of 3,600 seconds, whichever occurs first. The number in the parenthesis following the gap, if displayed, denotes the number of instances we fail to prove optimality/INF within the timelimit. Finally, an "INF" is placed in the column "Gap (%)" if all five random instances are infeasible.

From Tables 3.2–3.4, we observe that `2-Stg` generally took shorter time to solve the easier instances than `+single`, while the strength of single-scenario inequalities began to reveal in the harder instances. For instance, when $n = 60, I = 90, \delta = 0.05, \epsilon = 0.05, N = 50$, `2-Stg` only solved 2 out of 5 random instances and took 2333.7 seconds on average, whereas `+single` solved all 5 instances and only took 70.1 seconds on average. In Table 3.2,

single-scenario inequalities on average shortened the CPU time by approximately 20% and shrunk the optimality gap significantly (around 48%). Specifically, they helped a lot on problems where the ratios of δ/ϵ is larger, e.g., when $(\delta, \epsilon) = (0.05, 0.05)$ or $(\delta, \epsilon) = (0.1, 0.1)$. This demonstrates the effectiveness of the single-scenario inequalities. In Table 3.4, when the training data size N became larger ($N = 200$), the speedup brought by single-scenario inequalities remained considerable, even though its average CPU time and final gap were on par with 2-Stg. For instance, when $n = 60, I = 70, \delta = 0.1, \epsilon = 0.1, N = 200$, 2-Stg solved 4 out of 5 instances and spent 1731.0 seconds on average, while +single solved all the instances and took only 45% of the time on average. By comparing the performance of +cross and +single, we observe that cross-scenario inequalities can further improve the computational performance. For example, in Table 3.2, where $n = 80, I = 90, \delta = 0.05, \epsilon = 0.05, N = 50$, +cross solved 4 instances to optimality within the timelimit, while +single solved only 2, let alone 2-Stg, which failed to solve any of the 5 instances. As for the most challenging instances, in which all three approaches failed to prove optimality within the timelimit, +cross proved significantly smaller optimality gaps than 2-Stg and +single. This demonstrates the strength of the cross-scenario inequalities on capturing the intersections of the feasible regions arising from multiple scenarios or coverings.

3.6 Conclusion

In this chapter, we study RCCs with a set covering structure, i.e. (DRC). Specifically, we (1) demonstrate the value of incorporating the binary support information into the model, (2) prove the NP-hardness of solving (DRC), (3) derive a complete description of the convex hull of a basic mixed-integer set arising in the deterministic reformulation of (DRC), and (4) demonstrate the OOS performance and the strength of our valid inequalities on randomly generated test instances and emergency medical facility location problems.

Table 3.2: Benchmark between different models on synthetic data ($N = 50$)

				2-Stg		+single		+cross	
n	I	δ	ϵ	Time	Gap (%)	Time	Gap (%)	Time	Gap (%)
60	70	0.05	0.05	1498.0	2 (1)	63.5	0	38.0	0
			0.10	3057.4	30.1 (4)	3009.7	19.5 (4)	1447.3	0.6 (1)
		0.10	0.05	5.3	0	11.8	0	8.7	0
			0.10	836.3	2.9 (1)	722.8	0	188.8	0
		0.20	0.05	1.6	INF	9.1	INF	6.4	INF
			0.10	5.7	0	12.1	0	8.6	0
		0.30	0.05	1.1	INF	7.4	INF	0.0	INF
			0.10	3.8	0	11.7	0	8.7	0
	90	0.05	0.05	2333.7	8.2 (3)	70.1	0	31.0	0
			0.10	3600.3	44 (5)	3600.0	30.3 (5)	1829.0	0
		0.10	0.05	10.4	0	14.2	0	10.6	0
			0.10	3328.5	9 (4)	905.8	0	184.5	0
		0.20	0.05	1.8	INF	9.0	INF	6.6	INF
			0.10	9.1	0	15.1	0	11.0	0
		0.30	0.05	1.2	INF	7.2	INF	0.0	INF
			0.10	4.5	0	11.9	0	9.2	0
80	70	0.05	0.05	3600.0	42.7 (5)	3088.9	10.8 (4)	2333.8	6.6 (2)
			0.10	3600.0	65.2 (5)	3600.1	51.2 (5)	3600.0	23.4 (5)
		0.10	0.05	27.1	0	27.9	0	18.6	0
			0.10	3600.2	35 (5)	3431.0	10.9 (4)	2622.2	7.1 (3)
		0.20	0.05	7.2	0	17.8	0	12.7	0
			0.10	23.7	0	31.4	0	20.2	0
		0.30	0.05	1.7	INF	14.1	INF	0.0	INF
			0.10	7.5	0	20.6	0	13.9	0
	90	0.05	0.05	3600.1	39.6 (5)	3126.5	5.2 (3)	2203.9	1.1 (1)
			0.10	3600.0	64.1 (5)	3600.1	52.1 (5)	3600.0	22.5 (5)
		0.10	0.05	57.0	0	30.3	0	19.3	0
			0.10	3600.1	35.5 (5)	3600.0	16.1 (5)	3255.6	8.2 (4)
		0.20	0.05	6.5	0	17.0	0	12.6	0
			0.10	44.0	0	32.6	0	21.6	0
		0.30	0.05	2.0	INF	14.9	INF	0.0	INF
			0.10	9.2	0	19.5	0	14.7	0
Average*				1140.2	14.6	911.1	7.5	673.0	2.7

* INF instances are excluded when calculating average gap.

Table 3.3: Benchmark between different models on synthetic data ($N = 100$)

				2-Stg		+single		+cross	
n	I	δ	ϵ	Time	Gap (%)	Time	Gap (%)	Time	Gap (%)
60	70	0.05	0.05	968.7	2 (1)	671.5	0	247.8	0
			0.10	3600.0	43.6 (5)	3600.0	29.9 (5)	2103.2	6.5 (2)
		0.10	0.05	10.8	0	25.6	0	18.7	0
			0.10	1903.4	2.9 (2)	990.7	0.7 (1)	264.7	0
		0.20	0.05	2.5	INF	18.4	INF	13.9	INF
			0.10	13.5	0	32.2	0	18.5	0
		0.30	0.05	2.0	INF	14.8	INF	0.0	INF
			0.10	8.0	0	26.6	0	17.4	0
	90	0.05	0.05	2696.9	8 (3)	1159.9	2.1 (1)	307.0	0
			0.10	3600.0	48.1 (5)	3600.1	35 (5)	2594.6	8.5 (3)
		0.10	0.05	18.0	0	34.0	0	22.3	0
			0.10	2440.3	4.8 (3)	1717.0	1.7 (1)	544.9	0.7 (1)
		0.20	0.05	2.7	INF	18.1	INF	13.9	INF
			0.10	24.0	0	53.4	0	27.4	0
		0.30	0.05	2.1	INF	15.0	INF	0.0	INF
			0.10	11.6	0	33.1	0	20.9	0
80	70	0.05	0.05	3600.0	30.6 (5)	3305.9	9.8 (4)	2528.1	6.1 (3)
			0.10	3600.0	65 (5)	3600.2	51.1 (5)	3600.0	26.1 (5)
		0.10	0.05	225.6	0	102.0	0	53.8	0
			0.10	3600.1	29.8 (5)	3600.0	16.6 (5)	2941.8	9.2 (3)
		0.20	0.05	7.4	0	33.4	0	24.6	0
			0.10	415.2	0	234.7	0	60.2	0
		0.30	0.05	2.9	INF	29.3	INF	0.0	INF
			0.10	23.8	0	60.8	0	35.5	0
	90	0.05	0.05	3600.0	30.8 (5)	3579.4	10.4 (4)	3197.5	7.4 (4)
			0.10	3600.3	65.5 (5)	3601.0	52 (5)	3600.3	27.1 (5)
		0.10	0.05	46.6	0	50.8	0	35.1	0
			0.10	3600.0	27.1 (5)	3600.0	15.2 (5)	3241.2	7.5 (3)
		0.20	0.05	7.2	0	31.9	0	23.9	0
			0.10	95.5	0	70.5	0	38.1	0
		0.30	0.05	3.1	INF	28.7	INF	0.0	INF
			0.10	24.4	0	59.2	0	34.1	0
Average*				1179.9	13.8	1062.4	8.6	800.1	3.8

* INF instances are excluded when calculating average gap.

Table 3.4: Benchmark between different models on synthetic data ($N = 200$)

				2-Stg		+single		+cross	
n	I	δ	ϵ	Time	Gap (%)	Time	Gap (%)	Time	Gap (%)
60	70	0.05	0.05	432.4	0	326.5	0	81.6	0
			0.10	3600.0	44.7 (5)	3600.0	32.8 (5)	3428.5	7.7 (4)
		0.10	0.05	14.2	0	47.8	0	34.2	0
			0.10	1731.0	0.6 (1)	745.4	0	152.7	0
		0.20	0.05	4.3	INF	35.5	INF	31.5	INF
			0.10	25.2	0	67.3	0	38.1	0
		0.30	0.05	3.4	INF	30.0	INF	0.0	INF
			0.10	13.4	0	50.1	0	39.5	0
	90	0.05	0.05	2044.0	2.1 (2)	809.0	0	105.8	0
			0.10	3600.0	48.8 (5)	3600.9	36.4 (5)	3566.4	13 (4)
		0.10	0.05	22.9	0	63.3	0	41.1	0
			0.10	2930.7	6.3 (4)	2457.6	2.8 (3)	461.4	0
		0.20	0.05	5.0	INF	36.6	INF	26.9	INF
			0.10	48.5	0	95.8	0	50.0	0
		0.30	0.05	3.7	INF	29.1	INF	0.0	INF
			0.10	23.3	0	74.4	0	46.3	0
80	70	0.05	0.05	3600.0	31.3 (5)	3600.0	21.2 (5)	3600.0	12.9 (5)
			0.10	3600.1	75.2 (5)	3600.4	54.3 (5)	3600.3	35.4 (5)
		0.10	0.05	193.8	0	165.2	0	78.3	0
			0.10	3600.0	31.3 (5)	3600.2	22.1 (5)	3600.0	12.9 (5)
		0.20	0.05	15.2	0	71.2	0	56.2	0
			0.10	180.7	0	200.7	0	97.5	0
		0.30	0.05	5.8	INF	58.6	INF	0.0	INF
			0.10	46.0	0	119.7	0	83.9	0
	90	0.05	0.05	3600.0	23.3 (5)	3600.0	14.5 (5)	2141.8	3.2 (2)
			0.10	3600.2	70.2 (5)	3601.2	51.7 (5)	3600.0	30.1 (5)
		0.10	0.05	110.8	0	163.7	0	98.2	0
			0.10	3600.0	30.3 (5)	3600.0	17.6 (5)	3600.0	6.3 (5)
		0.20	0.05	14.3	0	69.9	0	47.9	0
			0.10	856.0	0.1	787.7	0	136.4	0
		0.30	0.05	6.0	INF	56.6	INF	0.0	INF
			0.10	123.2	0	278.2	0	110.7	0
Average*				1176.7	9.1	1113.8	9.7	904.9	4.7

* INF instances are excluded when calculating average gap.

CHAPTER 4

Locating Charging Stations for Battery Electric Buses

4.1 Problem Description and Literature Review

Due to the benefits in reducing emissions, improving fuel economy, lowering noise levels, and cutting down maintenance and operation costs, battery electric buses (BEBs) have enjoyed fast-growing adoption by transit agencies around the world in recent years (e.g., [Laj14]; [Mah+16]; [Xyl+17]; [LS17]; [NAS18]; [Du+19]; [TLH19]). By the end of 2018, out of the world's 3 million buses some 425,000 were BEBs [Eck19]. Moreover, a recent study by Bloomberg New Energy Finance predicted that 35% of the global bus fleet will be electrified by 2025 [Eck19]. China is leading the way of adopting BEBs. Shenzhen, a megacity in China with 16,359 buses in operation, has achieved a 100% BEB fleet by the end of 2017 [Lin+19]. Other big cities in China are also accelerating their bus fleet electrification process. Other regions have also been conducting smaller-scale BEB pilot programs. In the United Kingdom, BEB fleets have been deployed and tested in the cities of London, Nottingham, and Liverpool [Ayr18]; and in the United States, BEB pilot programs or plans have been established in many metropolises, such as New York City [New18], Chicago [Chi14], Los Angeles [Kin17], and Salt Lake City [Uta16].

Despite their many potential benefits, the deployment of BEBs also brings challenges to transit agencies. First, the on-board batteries of BEBs are costly and have limited driving range. This is because the energy density of batteries is lower than diesel for conventional diesel buses. To reach a comparable driving range, a BEB needs to carry large and expensive battery packs. For example, a 40-ft bus needs to carry a 660-kWh battery pack in order to have a nominal driving range of 328 miles [Pro19]. Such a battery pack alone costs about $231,000 [Asi18], which is nearly 50% of the price of a diesel bus [Cal18]. Second, the charging needs of BEBs make it difficult to maintain existing

bus schedules. According to a recent survey of 18 transit agencies deploying BEBs, 11 of them reported adjusting their schedules to accommodate the unique operational needs of BEBs [NAS18]. However, to achieve smooth transition, maintaining existing schedules is important for many transit agencies [Wan+17]. Consequently, for successful deployment of BEBs, transit agencies need to strategically plan the charging infrastructure, in order to ensure the normal operation of the bus system with a minimal impact on its existing schedule [NAS18]. The main focus of this study is to investigate the strategic planning of charging facilities for BEBs.

The charging infrastructure for BEBs includes battery-swapping stations, wireless charging lanes, and (plug-in, overhead conductive, or stationary wireless) charging stations [CYS18; Lin+19]. Battery swapping stations can replace depleted batteries of BEBs with fully charged ones, reducing the charging times. However, they demand large lands and high capital investments for both construction and building battery inventories, prohibiting a wide adoption [Li16; Lin+19]. Wireless charging lanes can charge BEBs while they are in motion, and so BEBs can remain in service while charging at the same time [LS17; Jan18; CYS18]. Unfortunately, due to the difficulty and high cost associated with their construction, wireless charging lanes remain in the early demonstration stage and are not ready for deployment. In contrast to these options, charging stations, typically installed at depots, en-route, or transit centers, dominate today's BEB charging infrastructure. For instance, Shenzhen deploys plug-in fast charging stations to charge its fleet of 16,359 BEBs. In addition, according to a recent survey NASEM [NAS18], more than half of the U.S. transit agencies deploying BEBs are using en-route overhead/wireless charging stations. In this chapter, we focus on charging stations as the primary charging method for BEBs.

Deterministic models have been proposed for the strategic planning of BEB charging stations. Chen et al. [Che+13] developed a mixed-integer linear programming model to determine the charging station deployment and battery capacity for a single BEB bus line. Ke, Chung, and Chen [KCC16] employed a simulation-based optimization approach to determine the charging schedule of a BEB bus system. Kunith, Mendelevitch, and Goehlich [KMG17] proposed a cost-optimization model to locate the charging stations and determine the battery size for a system of multiple BEB bus lines. Xylia et al. [Xyl+17] developed a model to optimize the distribution of charging stations for BEBs in Stockholm, Sweden. These works assume that the energy consumption and the dwell time of BEBs at stops are deterministic. In reality, however, these parameters are usually stochastic, e.g., the energy consumption depends on the weather and traffic conditions and the dwell times depend on the volume of passengers. Without accounting for these uncertainties, a planning model may produce a suboptimal or even infeasible design, in which BEBs run

out of power before reaching the next charging station.

To the best of our knowledge, only two studies in the literature have considered the parameter uncertainty in the planning of BEB charging stations. In particular, An [An20] considered random bus charging demand and proposed a stochastic program (SP) to jointly optimize the charging station locations and the bus fleet size. Differently, Liu, Song, and He [LSH18] proposed a robust optimization (RO) model to locate the charging stations and determine the battery size. Comparing these two works, we observe that the SP assumes a full access to the probability distribution of the charging demand, while the RO model on the contrary assumes nothing about the distribution except for the support of the uncertainty. In this chapter, we adopt a distributionally robust optimization (DRO) approach in order to extract more information from the available data of the random parameters (e.g., energy consumption and dwell time). In DRO, we characterize an *ambiguity set* of probability distributions using the data, and optimize the expected performance of the system with respect to a distribution chosen adversarially from the ambiguity set. In this sense, DRO can be seen as a compromise of SP and RO, in the sense that DRO asymptotically reduces to SP and RO when the ambiguity set shrinks and enlarges, respectively. We refer to Rahimian and Mehrotra [RM19] and Kuhn et al. [Kuh+19]. Recently, DRO has found various transportation applications, including vehicle routing [JQS16], electrified transportation planning [Haj+20], humanitarian relief network design [Zha+20], traffic sensor deployment [ZFM18], and connected and automated vehicle platooning [ZZ20].

In this chapter, we propose a DRO model for jointly optimizing the location of BEB charging stations and the size of battery pack. We employ chance constraints to ensure that BEBs can normally operate with a high probability under random energy consumption and dwell time. Moreover, we characterize the ambiguous distribution of the random parameters using real-world data and a ϕ-divergence ambiguity set. We summarize the main contributions as follows.

1. We derive a distributionally robust chance constraint (DRCC) model to account for the random energy consumption and dwell time. We recast this model as a mixed-integer second-order conic program (MISOCP), which can be readily solved by commercial solvers.

2. We design two decomposition algorithms to speed up the solution of the MISOCP reformulation. The first algorithm employs delayed constraint generation and replaces most SOC constraints by their low-rank, linear approximation. The second algorithm employs dual decomposition to break the MISOCP formulation down to smaller subproblems pertaining to individual bus lines.

3. We conduct extensive numerical experiments based on the Salk Lake City (SLC) bus transit system. We demonstrate the effectiveness of the DRCC model and that both decomposition algorithms achieve a significant speed-up when compared to commercial solvers.

The remainder of the chapter is structured as follows. Section 4.2 presents the DRCC model and reformulates it as a MISOCP. Section 4.3 studies the two decomposition algorithms. Section 4.4 reports the numerical results. Finally, Section 4.5 concludes this chapter.

Notation. For $n \in \mathbb{N}_+$, we define $[n] := \{1, \ldots, n\}$. We let $\mathbf{1}$ denote a vector of all ones, whose dimension is clear from the context. For square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ in $\mathbb{R}^{n\times n}$, we define $\mathbf{tr}\,(\boldsymbol{A}) := \sum_{i=1}^{n} \boldsymbol{A}_{ii}$ as its trace, $\mathrm{vec}\,(\boldsymbol{A}) := [\boldsymbol{A}_{11}, \boldsymbol{A}_{12}, \ldots, \boldsymbol{A}_{nn}]^\mathsf{T}$ as its vectorization, $\boldsymbol{A} \otimes \boldsymbol{B}$ represents the Kronecker product of $\boldsymbol{A}$ and $\boldsymbol{B}$, and $\boldsymbol{I}_n$ denotes an $n \times n$ identity matrix.

4.2 Distributionally Robust Chance Constraint Model

We first review a deterministic model for the strategic planning of BEB charging stations in Section 4.2.1 and provide an equivalent, alternative formulation that is more tractable for a chance-constraint extension in Section 4.2.2. Then, based on the alternative deterministic formulation, we propose the DRCC model and recast it as a MISOCP.

4.2.1 Deterministic model

We start by describing the deterministic model introduced in Kunith, Mendelevitch, and Goehlich [KMG17] and Liu, Song, and He [LSH18]. The model seeks to jointly optimize the location of BEB charging stations and the size of BEB battery size, with the objective of minimizing the the total investment while ensuring the normal operations of BEBs. To this end, the model assumes that

(A) Each bus line operates on a fixed route.

(B) Each bus line has a base station, where all buses on this line begin and end their service loop (different bus lines may share a base station).

(C) After finishing a service loop, a BEB will be fully charged at the base station before it starts another loop.

These assumptions usually hold in reality and we refer to Liu, Song, and He [LSH18] for a detailed discussion. We consider a network consisting of K bus lines and each line k starts its service from the base node o_k, successively visits nodes $n_1, n_2, \ldots, n_{N_k-1}$ before returning to o_k, where $\mathcal{N}_k := \{o_k, n_1, \ldots, n_{N_k-1}\}$ denotes the set of nodes bus line k stops, $\mathcal{N} := \cup_{k\in[K]}\mathcal{N}_k$ denotes the set of all nodes in the network, and $N_k := |\mathcal{N}_k|$ denotes the number of bus stops. Accordingly, bus line k passes the directed edges

$$\mathcal{L}_k := \left\{ (n_i, n_{i+1}) \colon i \in [N_k - 2] \right\} \cup \left\{ (o_k, n_1), (n_{N_k-1}, o_k) \right\}.$$

For each bus station $i \in \mathcal{N}$, there are T types of charging facilities we can choose from, each with different power and investment cost. Accordingly, we let x_{it} denote a binary decision variable for building a type-t fast-charging station at bus station i for all $t \in [T]$ (i.e., $x_{it} = 1$ if build and $x_{it} = 0$ otherwise), and $e_k^{\max}$ denote a continuous variable for the battery size for the BEBs of line $k \in [K]$. Using these decision variables, we formulate the deterministic model as follows.

(**DET**)
$$\begin{aligned}
\min_{x,e,e^{\max}} \quad & \sum_{i\in\mathcal{N}}\sum_{t\in[T]} a_{it}^{\mathsf{chgr}} x_{it} + \sum_{k\in[K]} a^{\mathsf{bat}} \zeta_k e_k^{\max}, \\
\text{s.t.} \quad & \sum_{t\in[T]} x_{it} \le 1 \quad \forall i \in \mathcal{N}, && \text{(4.1a)} \\
& e_{k,o_k} = \bar{\epsilon} \cdot e_k^{\max} \quad \forall k \in [K], && \text{(4.1b)} \\
& e_{kj} \le e_{ki} - c_{kij} + \sum_{t\in[T]} p_t x_{jt} \tau_{kj} \quad \forall k \in [K], (i,j) \in \mathcal{L}_k : j \ne o_k, && \text{(4.1c)} \\
& e_{ki} \le \bar{\epsilon} \cdot e_k^{\max} \quad \forall k \in [K], i \in \mathcal{N}_k, && \text{(4.1d)} \\
& e_{ki} - c_{kij} \ge \underline{\epsilon} \cdot e_k^{\max} \quad \forall k \in [K], (i,j) \in \mathcal{L}_k, && \text{(4.1e)} \\
& x_{it} \in \{0,1\} \quad \forall i \in \mathcal{N}, t \in [T], \quad e_k^{\max} \ge 0 \quad \forall k \in [K]. && \text{(4.1f)}
\end{aligned}$$

For the ease of exposition, we incorporate a nomenclature of the notations in Appendix C.1. The objective function of (DET) seeks to minimize the total investment, including the cost of building charging stations and the cost of procuring battery packs. Constraints (4.1a) describe that each bus station can install at most one type of charging facility. Constraints (4.1b) specify that a BEB is fully charged when it begins its service from the base station, where e_{ki} denotes the remaining battery power of a line-k BEB when it leaves station i. Constraints (4.1c) describe the change in battery power of a line-k BEB due to its energy consumption c_{kij} on link (i,j) and energy charging (specifically, in the amount of $\sum_{t\in[T]} p_t x_{jt} \tau_{kj}$) at station j. Here, p_t denotes the power of a type-t charging facility and τ_{kj}

denotes the dwell time (i.e., the charging time) of a line-k BEB at station i. We exclude the constraint with respect to the last link of the loop (when $j = o_k$) in view of Assumption (C). Finally, constraints (4.1d) and (4.1e) specify the maximal and minimal battery power of a BEB respectively, and in particular, constraints (4.1e) ensure that the BEB has at least the minimal power and so runs normally before it arrives at the next station.

We remark that (DET) can be extended to relax Assumption (C) above and allow the BEB to get a full recharge at the base station every m service loops, where m is an integer and $m \geq 2$. Specifically, we can make m copies of all the nodes in $\mathcal{N}_k$ and concatenate these copies together to enlengthen the original service loop for m times. Accordingly, if we fully recharge the BEB at only one of the m copies of the base station, then it is equivalent that the BEB is fully recharged every m service loops (in its original loop).

4.2.2 DRCC model

Despite its generality, it is intractable to extend (DET) to account for uncertainties. To see this, suppose that the energy consumption c_{kij} and dwell time τ_{kj} are random. Then, the battery power e_{ki} of a BEB after leaving station i also becomes random because it depends on c_{kij} and τ_{kj}, that is, $e_{ki} \equiv e_{ki}(c_{kij}, \tau_{kj})$. Consequently, a chance constraint that intends to protect the BEB from running out of power (that is, to keep $e_{ki}(c_{kij}, \tau_{kj})$ above the minimal power $\underline{e}e^{\max}$ at all time) would need to compute $e_{ki}(c_{kij}, \tau_{kj})$ for each realization of (c_{kij}, τ_{kj}) in an auxiliary optimization problem. This gives rise to a *chance constraint with recourse*, which is much more challenging to compute than chance constraints *without* recourse.

In what follows, we shall recast (DET) by projecting out the variables e_{ki}. The ensuing model, which we denote as (EQU), is equivalent to (DET) but only relies on decision variables x_{it} and $e_k^{\max}$, which do not depend on (c_{kij}, τ_{kj}) because they need to be determined before any BEB runs. Then, we formulate a tractable DRCC model *without* recourse based on (EQU).

To this end, for all bus line $k \in [K]$ and stations $m, n \in \mathcal{N}_k$, we denote $m <_k n$ if a line-k bus leaving o_k passes m before it passes n. For example, for line k, we have $o_k <_k n_1 <_k n_2 <_k \cdots <_k n_{N_k} <_k o_k$. For ease of exposition, we also define $o_k <_k o_k$. For all $k \in [K]$ and $m, n \in \mathcal{N}_k$ with $m <_k n$, we let $\mathcal{N}_k^{mn}$ denote the stations strictly between nodes m and n on bus line k, i.e., the stations a line-k bus passes when it travels from m to n, with m and n excluded. Accordingly, we let $\mathcal{L}_k^{mn}$ denote the links between stations m and n on bus line k, i.e., the links a line-k bus goes through when it travels from m to n. In the following theorem, we provide an equivalent reformulation of constraints (4.1b)–(4.1e)

and give the formulation (EQU).

Theorem 12. Constraints (4.1b)–(4.1e) are equivalent to (4.2) defined below. In other words, for any (x, e^{max}) satisfying (4.1a) and (4.1f), there exists an e that satisfies constraints (4.1b)–(4.1e) if and only if (x, e^{max}) satisfies the following constraints:

$$(\bar{\epsilon} - \underline{\epsilon}) \cdot e_k^{\mathsf{max}} \geq \sum_{(i,j)\in\mathcal{L}_k^{mn}} c_{kij} - \sum_{j\in\mathcal{N}_k^{mn}} \sum_{t\in[T]} p_t x_{jt} \tau_{kj} \quad \forall k \in [K],\ m, n \in \mathcal{N}_k : m <_k n. \tag{4.2}$$

Therefore, (DET) yields the same optimal value as (EQU), defined through

(**EQU**)
$$\begin{aligned} \min_{x, e^{\mathsf{max}}} \ & \sum_{i\in\mathcal{N}} \sum_{t\in[T]} a_{it}^{\mathsf{chgr}} x_{it} + \sum_{k\in[K]} a^{\mathsf{bat}} \zeta_k e_k^{\mathsf{max}} \\ \text{s.t.} \ & \text{(4.1a), (4.1f), (4.2).} \end{aligned}$$

Proof. For ease of exposition, this proof focuses on an (arbitrary) bus line k and we concatenate decision variables $\{ x_{jt} \colon t \in [T] \}$ and $\{ p_t \colon t \in [T] \}$ into column vectors $x_j := \Big[x_{jt}\Big]_{t\in[T]}$ and $p := \Big[p_t\Big]_{t\in[T]}$, respectively. In addition, we let j^- denote the (immediately) precedent station of j, i.e., $(j^-, j) \in \mathcal{L}_k$. Specifically, o_k^- denotes the last station on line k before returning to the base station.

First, suppose that x, e^{max} satisfies constraints (4.2). We construct an e that satisfies constraints (4.1b)–(4.1e). Specifically, we define

$$\begin{aligned} e_{k,o_k} &:= \bar{\epsilon} \cdot e_k^{\mathsf{max}}, \\ e_{kj} &:= \min \Big\{ \bar{\epsilon} \cdot e_k^{\mathsf{max}},\ e_{k,j^-} - c_{k,j^-,j} + (p^{\mathsf{T}} \boldsymbol{x}_j) \tau_{kj} \Big\}, \quad \forall j \in \mathcal{N}_k : j \neq o_k. \end{aligned}$$

Then, e satisfies constraints (4.1b)– (4.1d) by construction. To verify constraints (4.1e), for any $(i, j) \in \mathcal{L}_k$, we let $\widehat{i} \in \mathcal{N}_k$ denote the last station before i, where the BEB is fully charged. The existence of $\widehat{i}$ is guaranteed because the BEB is fully charged at the base station o_k. Then, it holds that

$$\begin{aligned} e_{ki} - c_{kij} &= e_{k\widehat{i}} - \sum_{(m,n)\in\mathcal{L}_k^{\widehat{i},i}} c_{kmn} + \sum_{\ell\in\mathcal{N}_k^{\widehat{i},i}} \Big(p^{\mathsf{T}} x_\ell\Big) \tau_{k\ell} - c_{kij} \\ &= \bar{\epsilon} e_k^{\mathsf{max}} - \sum_{(m,n)\in\mathcal{L}_k^{\widehat{i},j}} c_{kmn} + \sum_{\ell\in\mathcal{N}_k^{\widehat{i},i}} \Big(p^{\mathsf{T}} x_\ell\Big) \tau_{k\ell} \\ &\geq \underline{\epsilon} e_k^{\mathsf{max}}, \end{aligned}$$

where the first equality follows from the definitions of $\widehat{i}$ and e_{kj}, the second equality follows

from $e_{k\widehat{i}} = e_k^{\max}$, and the inequality follows from constraint (4.2). It follows that e satisfies constraints (4.1e) as well.

Second, suppose that $(x, e, e^{\max})$ satisfies constraints (4.1b)–(4.1e). Then, for any $m, n \in \mathcal{N}_k$ with $m <_k n$, it holds that

$$\begin{aligned} e_{km} &\leq \bar{\epsilon} \cdot e_k^{\max}, \\ -e_{kn^-} &\leq -c_{k,n^-,n} - \underline{\epsilon} \cdot e_k^{\max}, \\ e_{kj} - e_{ki} &\leq -c_{kij} + (p^\mathsf{T} x_j)\tau_{kj}, \quad \forall (i,j) \in \mathcal{L}_k^{m,n^-}. \end{aligned}$$

Adding up these inequalities yields constraint (4.2). This completes the proof. □

For ease of exposition, we simplify constraints (4.2). For $m, n \in \mathcal{N}_k^{m,o_k}$, we define $N_k^{mn} := |\mathcal{N}_k^{mn}|$ and $L_k^{mn} := |\mathcal{L}_k^{mn}|$ to be the cardinality of sets $\mathcal{N}_k^{mn}$ and $\mathcal{L}_k^{mn}$, respectively. In addition, we define matrices $\boldsymbol{X}_k^{mn} := \left[x_1, \ldots, x_{N_k^{mn}}\right]$, $\boldsymbol{C}_k^{mn} := \left[c_{kij}\right]_{(i,j)\in\mathcal{L}_k^{mn}}$, and vector $\boldsymbol{\tau}_k := \left[\tau_{1k}, \ldots, \tau_{N_k^{mn},k}\right]$. Then,

$$\begin{aligned} \sum_{(i,j)\in\mathcal{L}_k^{mn}} c_{kij} - \sum_{j\in\mathcal{N}_k^{mn}} \sum_{t\in\mathcal{T}} p_t x_{jt} \tau_{kj} &= \mathbf{1}^\mathsf{T} \operatorname{vec}\left(\boldsymbol{C}_k^{mn}\right) - p^\mathsf{T} \boldsymbol{X}_k^{mn} \boldsymbol{\tau}_k \\ &= \mathbf{1}^\mathsf{T} \operatorname{vec}\left(\boldsymbol{C}_k^{m,n}\right) - \operatorname{vec}\left((\boldsymbol{X}_k^{mn})^\mathsf{T}\right)^\mathsf{T} \left(p \otimes \boldsymbol{I}_{N_k^{mn}}\right) \operatorname{vec}\left(\boldsymbol{\tau}_k\right) \\ &= \boldsymbol{z}_{kmn}^\mathsf{T} \boldsymbol{H}_{kmn} \boldsymbol{\xi}_{kmn}, \end{aligned}$$

where

$$\boldsymbol{z}_{kmn} := \left[\operatorname{vec}\left((\boldsymbol{X}_k^{mn})^\mathsf{T}\right)^\mathsf{T} \quad \mathbf{1}^\mathsf{T}\right]^\mathsf{T}, \quad \boldsymbol{H}_{kmn} := \begin{bmatrix} -p \otimes \boldsymbol{I}_{N_k^{mn}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{L_k^{mn}} \end{bmatrix}, \quad \boldsymbol{\xi}_{kmn} := \left[\boldsymbol{\tau}_k^\mathsf{T} \quad \operatorname{vec}\left(\boldsymbol{C}_k\right)^\mathsf{T}\right]^\mathsf{T}.$$

We note that vector $\boldsymbol{z}_{kmn}$ is an affine function of the decision variables x, vector $\boldsymbol{\xi}_{kmn}$ consists of the random parameters $\boldsymbol{\tau}_k$ and $\boldsymbol{C}_k$, while matrix $\boldsymbol{H}_{kmn}$ consists of constants. It follows that constraints (4.2) can be simplified as

$$(\bar{\epsilon} - \underline{\epsilon}) \cdot e_k^{\max} \geq \boldsymbol{z}_{kmn}^\mathsf{T} \boldsymbol{H}_{kmn} \boldsymbol{\xi}_{kmn}, \quad \forall k \in \mathcal{K},\ m, n \in \mathcal{N}_k : m <_k n. \tag{4.3}$$

We now consider the case of random energy consumption c_{kij} and random dwell time τ_{kj}. In reality, the energy consumption depends on the weather and traffic conditions of running BEBs and the dwell times depend on the volume of passengers, which is highly stochastic in different seasons of a year and in different times of a day. In this case, constraints (4.3) are subject to risk of violation. To hedge against this risk, we seek to

find a strategic plan that can satisfy (4.3) with high probability (e.g., 90% or 95%). If we are able to estimate the *joint* probability distribution $\mathbb{P}$ of $\boldsymbol{\xi}$ accurately, then the chance constraint takes the form

$$\mathbb{P}\left\{ (\bar{\epsilon} - \underline{\epsilon}) \cdot e_k^{\max} \geq \boldsymbol{z}_{kmn}^{\mathsf{T}} \boldsymbol{H}_{kmn} \boldsymbol{\xi}_{kmn} \right\} \geq 1 - \alpha, \qquad \textbf{(CC)}$$

where $1-\alpha$ denotes a pre-specific risk threshold (e.g., $1-\alpha = 90\%$ or 95%). Unfortunately, the full access to $\mathbb{P}$ is rarely available and we usually have, at best, past data of $\boldsymbol{\xi}$, which can be seen as independent samples drawn from $\mathbb{P}$. In this case, a popular proxy of $\mathbb{P}$ is the Gaussian distribution $\mathcal{N}(\mu, \Sigma)$, where the mean μ and covariance matrix Σ can be estimated from the data of $\boldsymbol{\xi}$. The benefits of this proxy are two-fold: for the one thing, Gaussian distribution are ubiquitous in reality and have been successfully applied in various applications. For another thing, (**CC**) admits a SOC representation when $\mathbb{P}$ is Gaussian, which can be efficiently computed. However, when the true $\mathbb{P}$ fails to be Gaussian, we may fail to deliver the high probability of satisfaction promised by (**CC**). In this paper, we robustify (**CC**) by considering an ambiguity set centered around $\mathcal{N}(\mu, \Sigma)$:

$$\mathcal{D} = \left\{ \mathbb{P} : d\Big(\mathbb{P},\ \mathcal{N}(\mu, \Sigma)\Big) \leq \rho \right\},$$

where $d(\cdot, \cdot)$ denotes a discrepancy measure between two probability distributions and ρ denotes a pre-specified threshold of the discrepancy. Intuitively, $\mathcal{D}$ uses $\mathbb{P}_0 := \mathcal{N}(\mu, \Sigma)$ as a reference distribution for $\boldsymbol{\xi}$ and consists of a neighborhood of the reference, which allows us to misspecify the true probability distribution of $\boldsymbol{\xi}$ up to the threshold ρ. Accordingly, we define a robust counterpart of (**CC**) through

$$\inf_{\mathbb{P} \in \mathcal{D}} \ \mathbb{P}\left\{ (\bar{\epsilon} - \underline{\epsilon}) \cdot e_k^{\max} \geq \boldsymbol{z}_{kmn}^{\mathsf{T}} \boldsymbol{H}_{kmn} \boldsymbol{\xi}_{kmn} \right\} \geq 1 - \alpha, \qquad \textbf{(DRCC)}$$

which requires to satisfy (**CC**) for all distributions in $\mathcal{D}$. In order to recast (**DRCC**) into a computable form, this chapter considers the following two popular choices for $d(\cdot, \cdot)$.

- The ϕ**-divergence** between distributions $\mathbb{P}$ and $\mathbb{P}_0$ is defined through

  $$d_\phi(\mathbb{P}, \mathbb{P}_0) := \int \phi\left(\frac{f(\boldsymbol{\xi})}{f_0(\boldsymbol{\xi})}\right) f_0(\boldsymbol{\xi}) d\boldsymbol{\xi},$$

  where f and f_0 denote the density functions of $\mathbb{P}$ and $\mathbb{P}_0$, respectively, and $\phi : \mathbb{R}_+ \to \mathbb{R}$ is a convex function such that $\phi(1) = 0$, $0\phi(x/0) := x \lim_{p \to \infty} \phi(p)/p$ if $x > 0$, and $0\phi(x/0) := 0$ if $x = 0$. Table 4.1 displays three examples of d_ϕ.

ϕ-Divergence d_ϕ	Function $\phi(x)$	Adjusted Risk Threshold $1-\beta$
Kullback-Leibler	$x\log(x)-x+1$	$\inf_{x\in(0,1)}\left\{\frac{\exp(-\rho)x^{1-\alpha}-1}{x-1}\right\}$
χ-divergence of order 2	$(x-1)^2$	$1-\alpha+\frac{\sqrt{\rho^2+4\rho(\alpha-\alpha^2)}-(1-2\alpha)\rho}{2\rho+2}$
Variation distance	$\lvert x-1\rvert$	$1-\alpha+\rho/2$

Table 4.1: Examples of ϕ-divergence

- The **Wasserstein distance** between distributions $\mathbb{P}$ and $\mathbb{P}_0$ is defined through

$$d_W(\mathbb{P},\mathbb{P}_0) := \inf_{\Pi\sim(\mathbb{P},\mathbb{P}_0)} \mathbb{E}_\Pi\Big[\|\boldsymbol{\xi}-\boldsymbol{\xi}_0\|\Big],$$

  where $\boldsymbol{\xi}$ and $\boldsymbol{\xi}_0$ denote two random vectors following distributions $\mathbb{P}$ and $\mathbb{P}_0$, respectively, Π denotes a coupling of $\mathbb{P}$ and $\mathbb{P}_0$, and $\|\cdot\|$ represents a norm. $d_W(\mathbb{P},\mathbb{P}_0)$ can be interpreted as the minimum cost of transporting the probability masses of $\mathbb{P}$ to recover $\mathbb{P}_0$.

In a data-driven context, we calibrate the value of ρ so that $\mathcal{D}$ includes the true distribution of $\boldsymbol{\xi}$ with a high confidence. We refer to [Par18] for the calibration of the ϕ-divergence and to [FG15] and [MK18b] for the calibration of the Wasserstein distance. When we model (**DRCC**) using either d_ϕ or d_W, we not only robustify the out-of-sample performance of (**CC**), but (**DRCC**) still admits a SOC representation, which maintains the computational efficacy of (**CC**). We detail in the following proposition.

Proposition 15. (**ϕ-divergence**) If $d=d_\phi$, then (**DRCC**) is equivalent to the following SOC constraint:

$$(\bar{\epsilon}-\underline{\epsilon})\cdot e_k^{\mathsf{max}} \;\geq \mu^{\mathsf{T}}\boldsymbol{H}_{kmn}\boldsymbol{z}_{kmn} + b_\phi\left\|\Sigma^{1/2}\boldsymbol{H}_{kmn}^{\mathsf{T}}\boldsymbol{z}_{kmn}\right\|_2, \tag{4.4a}$$

where $b_\phi := \Phi^{-1}(\min\{1,1-\beta\})$, Φ denotes the cumulative distribution function of the standard normal distribution, and

$$\beta := \inf_{s>0,\ \underline{q}(\phi^*)\leq s_0+s\leq\overline{q}(\phi^*)}\left\{\frac{\phi^*(s_0+s)-s_0-\alpha s+\rho}{\phi^*(s_0+s)-\phi^*(s_0)}\right\}.$$

Above, $\phi^*(t) := \sup_{x\in\mathbb{R}}\{tx-\phi(x)\}$ denotes the conjugate of ϕ, $\underline{q}(\phi^*)=\sup\{q\in\mathbb{R}:\phi^* \text{ is a finite constant on } (-\infty,q]\}$, and $\overline{q}(\phi^*)=\inf\{q\in\mathbb{R}:\phi^*(q)=+\infty\}$.

(**Wasserstein distance**) If $d=d_W$ and the norm $\|\cdot\|$ in d_W is an ellipsoidal norm with regard to $\Sigma^{1/2}$, i.e., $\|\cdot\|=\|\Sigma^{-1/2}(\cdot)\|_2$, then (**DRCC**) is equivalent to the following SOC

constraint:

$$(\bar{\epsilon} - \underline{\epsilon}) \cdot e_k^{\max} \geq \mu^{\mathsf{T}} \boldsymbol{H}_{kmn} \boldsymbol{z}_{kmn} + b_W \left\| \Sigma^{1/2} \boldsymbol{H}_{kmn}^{\mathsf{T}} \boldsymbol{z}_{kmn} \right\|_2, \tag{4.4b}$$

where $b_W := g^{-1}(\rho)$ with

$$g(x) := \int_0^{+\infty} \Big[\Phi(x - t) - (1 - \epsilon) \Big]^+ dt.$$

The SOC representation for $d = d_\phi$ follows from Theorem 1 of [JG16] and that for $d = d_W$ follows from Theorem 4.8 of [CX21] and Corollary 3 of [SJ21]. We hence omit the proof of Proposition 15 for brevity. In addition, we notice that, although the coefficients b_ϕ and b_W do not admit closed-form expressions in general (see Table 4.1 for the closed-form expression of d_ϕ in special cases), they can be efficiently found by bisection line search methods (see, e.g., Section 4 of Jiang and Guan [JG16] for b_ϕ).

We close this section with an example to elaborate the solution quality of (DET), (**CC**), and (**DRCC**).

Example 9. Figure 4.1 depicts a BEB network with one base station 0 and two intermediate stations 1 and 2. For each BEB, the cost of purchasing a battery pack is $\$1000$ per kWh,

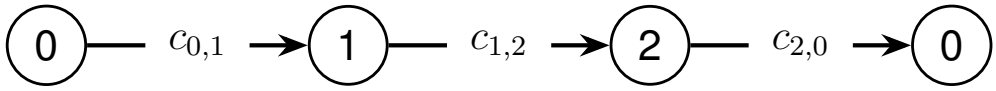


Figure 4.1: A toy BEB system with three stations.

with $\underline{\epsilon} = 0.1$ and $\bar{\epsilon} = 0.9$. For charging facility, there is only one type to choose from, which costs $\$50,000$ to install and has power 110 kW. In addition, the dwell times at stations $0, 1, 2$ are 900, 250, and 250 seconds, respectively, and they are assumed deterministic. Suppose that the energy consumption on links $(0, 1), (1, 2), (2, 0)$ follow a joint Gaussian distribution with mean $\Big[3, 12, 12\Big]^{\mathsf{T}}$ kWh and covariance matrix $\boldsymbol{I}_3$. Without knowing this true distribution, we draw 10 independent samples from it and produce a reference distribution

$\mathbb{P}_0 = \mathcal{N}(\mu, \Sigma)$, where the empirical estimates μ and Σ are

$$\mu = \begin{bmatrix} 2.44, & 11.94, & 12.24 \end{bmatrix}^{\mathsf{T}},$$
$$\Sigma = \begin{bmatrix} 0.9418 & 0 & 0 \\ 0 & 0.2691 & 0 \\ 0 & 0 & 1.1323 \end{bmatrix}.$$

First, we consider (DET) and constraint (4.2) with $m, n = 0$, that is, the BEB runs a full service loop from and back to the base station. In this case, the constraint takes the form

$$0.8e^{\max} \geq 2.44 + 11.94 + 12.24 - 110 \times \left(\frac{5}{72} x_1 + \frac{5}{72} x_2 \right). \tag{4.5}$$

The optimal solution to (DET) is $x_0 = 1, x_1 = x_2 = 0$, and $e^{\max} = 33.29$ kWh. When we incorporate the energy consumption uncertainty, however, this solution fails to satisfy the chance-constraint counterpart of (4.5), which has a risk threshold $1 - \alpha = 90\%$. Specifically,

$$\begin{aligned} \mathbb{P}\left\{ 0.8 \times 33.29 \geq c_{0,1} + c_{1,2} + c_{2,0} - 110 \times 0 \right\} &= \mathbb{P}\left\{ c_{0,1} + c_{1,2} + c_{2,0} - 27 \leq 26.632 - 27 \right\} \\ &= \Phi(-0.368) \approx 35.569\% < 90\%. \end{aligned}$$

Second, we consider (**CC**) with respect to the empirical estimate $\mathbb{P}_0 = \mathcal{N}(\mu, \Sigma)$. In this case, the optimal solution becomes $x_0 = x_2 = 1, x_1 = 0, e^{\max} = 25.73$ kWh. As compared to (DET), the (**CC**) model installs one more charging facility at station 2 and decreases the battery size. However, this solution still falls short of satisfying the chance-constraint counterpart of (4.5) with respect to the true distribution $\mathbb{P}$. Specifically,

$$\begin{aligned} &\mathbb{P}\left\{ 0.8 \times 25.73 \geq c_{0,1} + c_{1,2} + c_{2,0} - 110 \times \frac{5}{72} \right\} \\ &= \mathbb{P}\left\{ (c_{0,1} + c_{1,2} + c_{2,0} - 27) \leq (20.584 + 7.64 - 27) \right\} \\ &= \Phi(1.224) \approx 88.877\% < 90\%. \end{aligned}$$

Finally, we consider (**DRCC**). We construct the ambiguity set $\mathcal{D}$ using the Kullback-Leibler divergence (see Table 4.1) and set $\rho := 0.1$. The optimal solution becomes $x_0 = x_2 = 1, x_1 = 0, e^{\max} = 26.60$, and it satisfies the chance-constraint counterpart of (4.5)

with respect to the true distribution $\mathbb{P}$:

$$\begin{aligned}
&\mathbb{P}\left\{ 0.8 \times 27.19 \geq c_{0,1} + c_{1,2} + c_{2,0} - 110 \times \frac{5}{72} \right\} \\
&= \mathbb{P}\left\{ (c_{0,1} + c_{1,2} + c_{2,0} - 27) \leq (21.28 + 7.64 - 27) \right\} \\
&= \Phi(1.92) \approx 97.257\% > 90\%.
\end{aligned}$$

4.3 Solution Methods

By replacing constraints (4.2) with (**DRCC**) in (EQU), we obtain a mixed-integer SOC program, which can be readily solved by commercial solvers (e.g., GUROBI). However, as we shall demonstrate in Section 4.4, commercial solvers fail to produce a feasible strategic plan for BEBs in even small-scale instances of (**DRCC**), let alone solve such instances to global optimum. In this section, we propose two decomposition algorithms that can achieve this target in a reasonable amount of time. We study a delayed constraint generation (primal decomposition) algorithm in Section 4.3.1 and a dual decomposition algorithm in Section 4.3.2.

Algorithm 3: DELAYED CONSTRAINT GENERATION

Inputs : A BEB network, empirical estimates μ and Σ.
Outputs : An optimal design to the (**DRCC**) model.

1 Initialize a subset {SOC} of constraints (4.4a), which only incorporate those with $m = o_k$ and $n = o_k$ for all $k \in [K]$.;
2 Solve a relaxation of the (**DRCC**) model:

$$\begin{aligned}
\min_{x, e^{\mathsf{max}}} \quad & \sum_{i \in \mathcal{N}} \sum_{t \in [T]} a_{it}^{\mathsf{chgr}} x_{it} + \sum_{k \in [K]} a^{\mathsf{bat}} \zeta_k e_k^{\mathsf{max}} \\
\text{s.t.} \quad & \text{(4.1a), (4.1f), \{SOC\}}
\end{aligned}$$

and record an optimal solution $(\widehat{x}, \widehat{e}^{\mathsf{max}})$.;
3 Check if $(\widehat{x}, \widehat{e}^{\mathsf{max}})$ violates a constraint (4.4a) with respect to some bus line k and path m, n.;
4 **if** *(violated)* **then**
5 Add the violated constraint (4.4a) into the set {SOC}.;
6 **Go To** Step 2.;
7 **else**
8 **Return** $(\widehat{x}, \widehat{e}^{\mathsf{max}})$ *as an optimal design.*

4.3.1 Delayed constraint generation

A main challenge of solving the (**DRCC**) model efficiently is the large number of SOC constraints (4.4a)–(4.4b), which increases linearly in K and quadratically in $|\mathcal{N}_k|$. Nevertheless, the active SOC constraints, i.e., those satisfied with equality at the optimum of the (**DRCC**) model, may be much smaller in number. Motivated by this intuition, we start with a relaxation of the (**DRCC**) model with a small subset of the SOC constraints incorporated. Specifically, we start by incorporating the "longest" SOC constraints, which corresponds to originating from the base station $m = o_k$ and ending at the based station $n = o_k$ in (4.4a)–(4.4b). Then, to achieve global optimum, we iteratively add back the omitted SOC constraints only if they are violated. Hence, we call this algorithm delayed constraint generation (DCG) and summarize it in Algorithm 3. In this algorithm, we assume that $d = d_\phi$, that is, we use the ϕ-divergence ambiguity set, and the counterpart for using the Wasserstein ambiguity set is similar. Since there are a finite number of these SOC constraints, Algorithm 3 terminates finitely to the global optimum of the (**DRCC**) model by construction.

Proposition 16. Algorithm 3 terminates in a finite number of iterations and returns an optimal design to the (**DRCC**) model.

In Step 3 of Algorithm 3, we need to check if an incumbent solution $(\widehat{x}, \widehat{e}^{\max})$ violates any constraint (4.4a) (respectively, (4.4b) for $d = d_W$). This appears time-consuming because we need to enumerate all bus lines $k \in [K]$ and all paths $m, n \in \mathcal{N}_k$. Fortunately, the next proposition shows that such an enumeration is unnecessary because, for fixed $\widehat{x}$, a small subset of constraints (4.4a) (respectively, (4.4b) for $d = d_W$)"dominate" other constraints.

Proposition 17. Define $\gamma \in \mathbb{R}$ such that $\gamma := 1 - \Phi(b_\phi)$, or equivalently, $b_\phi = \Phi^{-1}(1-\gamma)$. In addition, assume that $\mathbb{P}_0(c_{kij} \geq 0) \geq 1 - \gamma$ for all bus lines $k \in [K]$ and links $(i,j) \in \mathcal{N}_k$. Consider an incumbent solution $(\widehat{x}, \widehat{e}^{\max}) \in \{0,1\}^{N\times T} \times \mathbb{R}^K_+$. If $(\widehat{x}, \widehat{e}^{\max})$ satisfies constraints (4.4a) (respectively, (4.4b)) for all $m, n \in \mathcal{N}_k$ such that $m <_k n$ and $\widehat{x}$ installs charging facilities at both stations m and n, i.e., $\sum_{t=1}^T \widehat{x}_{mt} = 1$ and $\sum_{t=1}^T \widehat{x}_{nt} = 1$, then $(\widehat{x}, \widehat{e}^{\max})$ satisfies all constraints (4.4a) (respectively, (4.4b)).

Proof. We present a proof for constraints (4.4a) and omit that for constraints (4.4b) due to the similarity. First, we recast constraint (4.4a) as a Value-at-Risk (VaR) constraint with

respect to the reference distribution $\mathbb{P}_0 \equiv \mathcal{N}(\mu, \Sigma)$. To this end, we derive

$$
\begin{aligned}
(\bar{\epsilon} - \underline{\epsilon}) \cdot e_k^{\mathsf{max}} \;&\geq\; \mu^{\mathsf{T}} \boldsymbol{H}_{kmn} \boldsymbol{z}_{kmn} + b_\phi \left\| \Sigma^{1/2} \boldsymbol{H}_{kmn}^{\mathsf{T}} \boldsymbol{z}_{kmn} \right\|_2, \\
&=\; \mathbb{E}_{\mathbb{P}_0}\left[\boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{H}_{kmn} \boldsymbol{z}_{kmn} \right] + \Phi^{-1}(1-\gamma) \mathsf{Cov}_{\mathbb{P}_0}\left[\boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{H}_{kmn} \boldsymbol{z}_{kmn} \right] \\
\iff\; \frac{(\bar{\epsilon} - \underline{\epsilon}) \cdot e_k^{\mathsf{max}} - \mathbb{E}_{\mathbb{P}_0}\left[\boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{H}_{kmn} \boldsymbol{z}_{kmn} \right]}{\mathsf{Cov}_{\mathbb{P}_0}\left[\boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{H}_{kmn} \boldsymbol{z}_{kmn} \right]} \;&\geq\; \Phi^{-1}(1-\gamma) \\
\iff\; \mathcal{N}_{(0,1)} \left\{ \frac{\boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{H}_{kmn} \boldsymbol{z}_{kmn} - \mathbb{E}_{\mathbb{P}_0}\left[\boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{H}_{kmn} \boldsymbol{z}_{kmn} \right]}{\mathsf{Cov}_{\mathbb{P}_0}\left[\boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{H}_{kmn} \boldsymbol{z}_{kmn} \right]} \leq \frac{(\bar{\epsilon} - \underline{\epsilon}) \cdot e_k^{\mathsf{max}} - \mathbb{E}_{\mathbb{P}_0}\left[\boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{H}_{kmn} \boldsymbol{z}_{kmn} \right]}{\mathsf{Cov}_{\mathbb{P}_0}\left[\boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{H}_{kmn} \boldsymbol{z}_{kmn} \right]} \right\} \\
\geq\; 1 - \gamma \\
\iff\; \mathbb{P}_0 \left\{ \boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{H}_{kmn} \boldsymbol{z}_{kmn} \leq (\bar{\epsilon} - \underline{\epsilon}) \cdot e_k^{\mathsf{max}} \right\} \;&\geq\; 1 - \gamma \\
\iff\; \mathsf{VaR}_{1-\gamma}\left(\boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{H}_{kmn} \boldsymbol{z}_{kmn} \right) \;&\leq\; (\bar{\epsilon} - \underline{\epsilon}) \cdot e_k^{\mathsf{max}},
\end{aligned}
$$

where $\mathsf{Cov}_{\mathbb{P}_0}(\cdot)$ denotes the covariance with respect to distribution $\mathbb{P}_0$, $\mathcal{N}_{(0,1)}$ denotes the standard Gaussian distribution, and $\mathsf{VaR}_{1-\gamma}(\cdot)$ denotes the Value-at-Risk with respect to $\mathbb{P}_0$ at the confidence level $1-\gamma$.

Second, consider any incumbent $(\widehat{x}, \widehat{e}^{\mathsf{max}}) \in \{0,1\}^{N \times T} \times \mathbb{R}_+^K$, in which $\widehat{x}$ does not install a charging facility at station m, i.e., $\sum_{t=1}^T \widehat{x}_{mt} = 0$, and denote $\widehat{\boldsymbol{z}}_{kmn} := \left[\mathrm{vec}((\widehat{\boldsymbol{X}}_k^{mn})^{\mathsf{T}})^{\mathsf{T}} \quad \boldsymbol{1}^{\mathsf{T}} \right]^{\mathsf{T}}$. We shall show that

$$
\mathsf{VaR}_{1-\gamma}\left(\boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{H}_{kmn} \widehat{\boldsymbol{z}}_{kmn} \right) \;\leq\; \mathsf{VaR}_{1-\gamma}\left(\boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{H}_{k,m^-,n} \widehat{\boldsymbol{z}}_{k,m^-,n} \right).
$$

If this indeed holds, then satisfying constraints (4.4a) with respect to path m^-, n implies its satisfaction with respect to path m, n. The same thing can be proved for the case for which $\widehat{x}$ does not install a charging facility at station n, establishing the claim of this proposition. To see the above inequality, we notice that

$$
\sum_{j \in \mathcal{N}_k^{mn}} \sum_{t \in \mathcal{T}} p_t \widehat{x}_{jt} \tau_{kj} = \sum_{j \in \mathcal{N}_k^{m^-,n}} \sum_{t \in \mathcal{T}} p_t \widehat{x}_{jt} \tau_{kj} \qquad (\star)
$$

because $\sum_{t=1}^{T} \widehat{x}_{mt} = 0$. It follows that

$$
\begin{aligned}
\mathsf{VaR}_{1-\gamma}\left(\boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{H}_{kmn} \widehat{\boldsymbol{z}}_{kmn}\right) &= \mathsf{VaR}_{1-\gamma}\left(\sum_{(i,j)\in\mathcal{L}_k^{mn}} c_{kij} - \sum_{j\in\mathcal{N}_k^{mn}} \sum_{t\in\mathcal{T}} p_t \widehat{x}_{jt} \tau_{kj}\right) \\
&= \mathsf{VaR}_{1-\gamma}\left(\sum_{(i,j)\in\mathcal{L}_k^{m^-,n}} c_{kij} - c_{k,m^-,m} - \sum_{j\in\mathcal{N}_k^{m^-,n}} \sum_{t\in\mathcal{T}} p_t \widehat{x}_{jt} \tau_{kj}\right) \\
&= \mathsf{VaR}_{1-\gamma}\left(\boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{H}_{k,m^-,n} \widehat{\boldsymbol{z}}_{k,m^-,n} - c_{k,m^-,m}\right) \\
&\leq \mathsf{VaR}_{1-\gamma}\left(\boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{H}_{k,m^-,n} \widehat{\boldsymbol{z}}_{k,m^-,n}\right) + \mathsf{VaR}_{1-\gamma}\left(-c_{k,m^-,m}\right) \\
&\leq \mathsf{VaR}_{1-\gamma}\left(\boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{H}_{k,m^-,n} \widehat{\boldsymbol{z}}_{k,m^-,n}\right),
\end{aligned}
$$

where the second equality uses ($\star$), the first inequality uses the subadditivity of $\mathsf{VaR}_{1-\gamma}(\cdot)$ for Gaussian random variables, and the final inequality uses the assumption $\mathbb{P}_0(c_{kij} \geq 0) \geq 1-\gamma$. This completes the proof. □

We notice that the assumption that $\mathbb{P}_0(c_{kij} \geq 0) \geq 1-\gamma$ is quite mild, because c_{kij} is the travel time of a BEB between stations i and j. When implementing Algorithm 3 in commercial solvers (e.g., GUROBI), we generate the SOC constraints (specifically, Steps 3–8 of Algorithm 3) using the lazy callback.

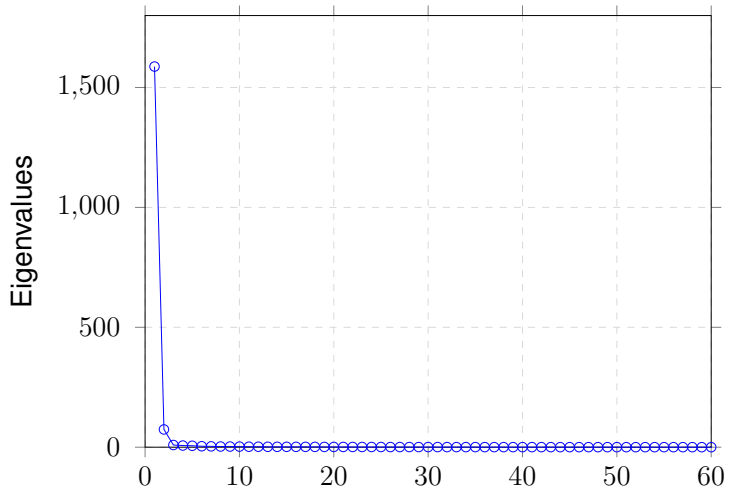


Figure 4.2: Eigenvalues of the matrix $\boldsymbol{H}_{kmn} \Sigma \boldsymbol{H}_{kmn}^{\mathsf{T}}$ based on the SLC data

4.3.1.1 Low-rank approximation of SOC constraints

In early iterations of Algorithm 3, the relaxation of the (**DRCC**) may be weak because most of the SOC constraints (4.4a)–(4.4b) are relaxed, leading to loose lower bounds and slow convergence. To strengthen the relaxation, we propose a low-rank, linear approximation of the SOC constraints.

Recall that the SOC constraints (4.4a)–(4.4b) take the form

$$(\bar{\epsilon} - \underline{\epsilon}) \cdot e_k^{\max} \;\geq\; \mu^{\mathsf{T}} \boldsymbol{H}_{kmn} \boldsymbol{z}_{kmn} + b\sqrt{\boldsymbol{z}_{kmn}^{\mathsf{T}} \Gamma \boldsymbol{z}_{kmn}},$$

where $\Gamma := \boldsymbol{H}_{kmn} \Sigma \boldsymbol{H}_{kmn}^{\mathsf{T}}$ and $b = b_\phi$ or b_W. Since the matrix Γ is positive semidefinite, it admits a singular value decomposition

$$\Gamma \;=\; \sum_{i=1}^{n_z} \sigma_i u_i u_i^{\mathsf{T}},$$

where $n_z := T \cdot N^{mn_k} + L_k^{mn}$ denotes the dimension of $\boldsymbol{z}_{kmn}$, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ denote the (sorted) eigenvalues of Γ, and u_i denotes the corresponding eigenvectors. Then, $\Gamma \geq \sigma_i u_i u_i^{\mathsf{T}}$ for all $i \in [n_z]$. Hence, the following inequality holds valid for any $(e^{\max}, \boldsymbol{z})$ satisfying the SOC constraints (4.4a)–(4.4b):

$$(\bar{\epsilon} - \underline{\epsilon}) \cdot e_k^{\max} \;\geq\; \mu^{\mathsf{T}} \boldsymbol{H}_{kmn} \boldsymbol{z}_{kmn} + b\sqrt{\sigma_i} \big| u_i^{\mathsf{T}} \boldsymbol{z}_{kmn} \big|.$$

The following proposition establishes the low-rank, linear approximation of the SOC constraints.

Proposition 18. For all $i \in [n_z]$, the following linear inequalities are valid for the (**DRCC**) model:

$$(\bar{\epsilon} - \underline{\epsilon}) \cdot e_k^{\max} \;\geq\; \mu^{\mathsf{T}} \boldsymbol{H}_{kmn} \boldsymbol{z}_{kmn} + b\sqrt{\sigma_i} u_i^{\mathsf{T}} \boldsymbol{z}_{kmn}, \tag{4.6a}$$

$$(\bar{\epsilon} - \underline{\epsilon}) \cdot e_k^{\max} \;\geq\; \mu^{\mathsf{T}} \boldsymbol{H}_{kmn} \boldsymbol{z}_{kmn} - b\sqrt{\sigma_i} u_i^{\mathsf{T}} \boldsymbol{z}_{kmn}, \tag{4.6b}$$

where $b = b_\phi$ or b_W.

In implementation of Algorithm 3, we add inequalities (4.6a)–(4.6b) to strengthen the relaxation in Step 2. In addition, we choose $i = 1$ in (4.6a)–(4.6b), that is, these inequalities are with respect to the largest eigenvalue of $\Gamma \equiv \boldsymbol{H}_{kmn} \Sigma \boldsymbol{H}_{kmn}^{\mathsf{T}}$. This is particularly motivated by the data of energy consumption and dwell time obtained from the SLC bus transit system. We depict in Figure 4.2 the distribution of the eigenvalues of Γ with $m, n = o_k$.

From this figure, we observe that the largest eigenvalue of Γ strongly dominates all the remaining eigenvalues. This implies that $\sigma_1 u_1 u_1^\mathsf{T}$ is a near-exact rank-1 approximation of Γ. Accordingly, the linear inequalities (4.6a)–(4.6b) with respect to $i = 1$ nearly recovers the SOC constraints (4.4a)–(4.4b).

4.3.2 Dual decomposition

A second main challenge of solving the (**DRCC**) model lies in the large number of bus lines K, which may be in the order of dozens for a metropolis. Nevertheless, these bus lines may only meet at several major stations or transit centers, and otherwise do not share bus stations. This motivates us to decompose the (**DRCC**) model into smaller formulations, each pertaining to an individual bus line, which can then be solved in parallel and so much more efficiently. This section makes this idea concrete by using a dual decomposition (DD) algorithm. DD was first introduced by ([CS99]) and has since been extensively studied in the literature, e.g., by ([Ahm13; Lub+13; Bol+18a; Bol+18b]), and applied in power systems ([Fd06]), supply chain design ([STA09]), and scheduling ([YY14]). To date, most DD algorithms have been developed for stochastic programs, where each formulation solved in parallel pertains to a scenario of the random parameters. In contrast, our DD algorithm is based on a physical network and each formulation solved in parallel pertains to a bus line.

First, we group the location variables according to they are shared by multiple bus lines or belong to an individual bus line. Specifically, we define

$$\bar{x} := \Big\{ x_{it} \colon \exists\, k_1, k_2 \in [K] \text{ such that } i \in \mathcal{N}_{k_1} \cap \mathcal{N}_{k_2}, t \in [T] \Big\},$$
$$\widetilde{x}_k := \Big\{ x_{it} \colon i \in \mathcal{N}_k \setminus \mathcal{N}_{k'}, \quad \forall k' \neq k, t \in [T] \Big\} \quad \forall k \in [K],$$

which represent the location variables shared by more than one bus line and those exclusively belonging to an individual line k, respectively. In addition, we create a copy of $\bar{x}$ for each bus line and denote it as $\bar{x}_k$ for all $k \in [K]$. Then, we rewrite the (**DRCC**) model as an augmented form:

(**AUG**)
$$\min_{\bar{x}_k, \widetilde{x}_k, e_k^{\max}} \sum_{k \in [K]} Q_k(\bar{x}_k, \widetilde{x}_k, e_k^{\max}),$$
$$\text{s.t.} \sum_{k \in [K]} A_k \bar{x}_k = 0, \tag{4.7a}$$
$$(\bar{x}_k, \widetilde{x}_k, e_k^{\max}) \in \mathcal{X}_k \quad \forall\, k \in [K]. \tag{4.7b}$$

In (**AUG**), constraints (4.7a) denote the "consensus" constraints among different copies of $\bar{x}$, $\bar{x}_1 = \bar{x}_2 = \cdots = \bar{x}_K$. They are also known as the nonanticipativity constraints in the literature of stochastic program (see, e.g., [CS99; Ahm13]). In addition, these constraints admit various formulations and we define them in general as in (4.7a). In constraints (4.7b), each $\mathcal{X}_k$ consists of the constraints (4.1a), (4.1f), (4.4a) for $d = d_\phi$ (or (4.4b) for $d = d_W$) pertaining to bus line k only. In the objective function, Q_k is defined through

$$Q_k(\bar{x}_k, \widetilde{x}_k, e_k^{\mathsf{max}}) := \sum_{i \in \bar{\mathcal{N}}} \sum_{t \in [T]} \frac{a_{i,t}^{\mathsf{chgr}}}{K} \bar{x}_k + \sum_{i \in \widetilde{\mathcal{N}}_k} \sum_{t \in [T]} a_{i,t}^{\mathsf{chgr}} \widetilde{x}_k + a^{\mathsf{bat}} \zeta_k e_k^{\mathsf{max}},$$

where $\bar{\mathcal{N}}$ and $\widetilde{\mathcal{N}}_k$ denote the bus stations shared by more than one bus lines and those exclusively belonging to line k only, respectively.

Second, we notice that (**AUG**) is separable for each bus line k only except for the consensus constraints (4.7a). As a result, relaxing (4.7a) in the Lagrangian manner leads to the following DD formulation:

$$\max_{\lambda} \left\{ V(\lambda) := \sum_{k \in [K]} Q_k^*(\lambda, \mathcal{X}_k) \right\}, \tag{**DD**}$$

$$\text{where} \quad Q_k^*(\lambda, \mathcal{X}_k) := \min_{\bar{x}_k, \widetilde{x}_k, e_k^{\mathsf{max}}} \left\{ Q_k(\bar{x}_k, \widetilde{x}_k, e_k^{\mathsf{max}}) + \lambda^{\mathsf{T}} A_k \bar{x}_k \colon \ (\bar{x}_k, \widetilde{x}_k, e_k^{\mathsf{max}}) \in \mathcal{X}_k \right\}.$$

In (**DD**), λ denotes the dual decision variables and, for any fixed λ, the objective function value $V(\lambda)$ can be obtained by solving K separable mixed-integer SOCPs of smaller sizes in parallel. In addition, $V(\lambda)$ is concave in λ because each $Q_k^*(\lambda, \mathcal{X}_k)$ is the pointwise minimum of affine functions of λ. As a result, (**DD**) can be solved subgradient algorithms, which provides a lower bound for the optimal value of (**AUG**) by construction. We remark that this lower bound may be strict because (**DD**) does not guarantee strong duality in general.

To overcome this, we recover a feasible solution to (**AUG**) and obtain an upper bound of its optimal value. Specifically, we denote $(\bar{x}_k^*, \widetilde{x}_k^*, e_k^{\mathsf{max}^*})$ as an optimal solution obtained from computing $Q_k^*(\lambda, \mathcal{X}_k)$. Then, the pointwise maximum of $(\bar{x}_k^*, \widetilde{x}_k^*, e_k^{\mathsf{max}^*})$ gives a feasible solution to (**AUG**). In addition, following ([Ahm13]), we exclude this feasible solution from each $\mathcal{X}_k$ to strengthen the (**DD**) formulation and increase the lower bound. Specifically, given a location decision $\widehat{x}$, we exclude it by adding the following integer cuts into each $\mathcal{X}_k$:

$$\sum_{i \colon \widehat{x}_i = 1} (1 - x_i) + \sum_{i \colon \widehat{x}_i = 0} x_i \geq 1. \tag{4.8}$$

Algorithm 4: DUAL_DECOMPOSITION

Inputs : $\mathcal{N}, \mathcal{L}, \mathcal{T}, \mathcal{K}, \mathcal{L}_k, \mathcal{N}_k, a^{\text{bat}}, a^{\text{chgr}}_{i,t}, c_{k,i,j}, \tau_{k,i}, p_t, o_k, \bar{\epsilon}, \underline{\epsilon}, \gamma$, `tol`.

Outputs : Optimal value of (**DRCC**).

Initialize lower bound and upper bound $\texttt{lb} \leftarrow -\infty$, $\texttt{ub} \leftarrow +\infty$, $S = \varnothing$;
Initialize and create $|\mathcal{K}|$ worker models for each bus line, `mod-k`;
while *lb < ub* **do**
  $\lambda \leftarrow 0$, $\Delta \leftarrow +\infty$;
  while $\Delta >$ `tol` **do**
    Solve $Q^*_k(\lambda, \mathcal{X}_k \setminus S)$ for $k \in \mathcal{K}$ in parallel;
    Obtain optimal solutions $\bar{x}^*_k$ and optimal values Q^*_k, for $k \in \mathcal{K}$;
    $\lambda \leftarrow \lambda + \gamma \cdot \sum_{k \in \mathcal{K}} A_k \bar{x}^*_k$;
  $\texttt{lb} \leftarrow \max\{ \texttt{lb}, \sum_{k \in \mathcal{K}} Q^*_k \}$;
  $S_\Delta := \{ \bar{x}^*_k \colon k \in \mathcal{K} \}$;
  $S_\Delta \leftarrow$GENERATE_FEASIBLE_SOLS(S_Δ);
  for $\bar{x}^* \in S_\Delta$ **do**
    Solve $Q^*_k(\lambda, \mathcal{X}_k \setminus S)$ for $k \in \mathcal{K}$ in parallel with with $\bar{x}$ fixed to be $\bar{x}^*$;
    $\texttt{ub} \leftarrow \min\{ \texttt{ub}, \sum_{k \in \mathcal{K}} Q^*_k(\lambda, \mathcal{X}_k \setminus S) \}$ Add integrality cut (4.8) to `mod-k` for $k \in \mathcal{K}$
  Update $S \leftarrow S \cup S_\Delta$
Return *mod.ub*.

We iteratively solve and strengthen (**DD**) to obtain a sequence of increasing lower bound, as well as obtain a sequence of decreasing upper bound by generating new feasible solution to (**AUG**). Algorithm 4 summarizes this iterative procedure, and since there is a finite number of location choices, it converges finitely to the global optimum.

Proposition 19. Algorithm 4 terminates in a finite number of iterations and returns an optimal design to the (**DRCC**) model.

In Algorithm 4, we first initialize $|\mathcal{K}|$ subproblems (one for each bus line) and variables $\texttt{lb}, \texttt{ub}, S$ for keeping track of the best lower / upper bound and the set of explored solutions. Next, we iteratively update the lower and upper bound as long as the gap between them is strictly positive. At each iteration, we start with $\lambda = 0$ and solve (**DD**) using gradient descent algorithm, which repeatedly moves λ along the gradient direction by γ until the change in objective is smaller than the given threshold Δ. Then, we collect all the optimal solutions obtained $\bar{x}^*_k, k \in \mathcal{K}$ into set S_Δ and use optimal values $Q^*_k, k \in \mathcal{K}$ to update the current best lower bound. After that, we can generate multiple feasible solutions by taking the pointwise maximum between any subset of solutions in S_Δ, all of which are used to update the upper bound accordingly. Before this iteration is completed, we cut off all explored solutions in S_Δ from the feasible region $\mathcal{X}_k, \forall k \in \mathcal{K}$ so that we can keep getting better lower bounds in the next iteration.

Although Ahmed [Ahm13]'s scheme can be applied to a wide range of applications and is guaranteed to converge after finite number of iterations (Proposition 1 in Ahmed [Ahm13]), such result is based on the enumeration of all possible binary valued feasible solutions. Therefore, when the number of shared decision variables, i.e., $\bar{\mathcal{N}} \times |\mathcal{T}|$ in our case, is large, it generally requires many iterations to converge since the improvement on the lower bound is usually small. Therefore, to further improve the convergence, we develop multiple valid inequalities for this approach, and all of them aggressively shrink the feasible region.

4.3.3 Valid inequalities

The following family of valid inequalities targets at removing infeasible solutions from the feasible region. In particular, if we know a given solution is already infeasible, then all solutions dominated by it should also be infeasible and can be discarded.

Proposition 20. If a given solution $\bar{x}_0$ is not feasible to (**AUG**), i.e., there does not exists $\widetilde{x}_k, e_k^{\max}, k \in \mathcal{K}$ such that $(\bar{x}_0, \widetilde{x}_k, e_k^{\max}, k \in \mathcal{K})$ together is feasible, then the following inequality is valid for (**AUG**):

$$\sum_{i:\ (\bar{x}_0)_i = 0} \bar{x}_i \geq 1. \tag{4.9}$$

Proof. We show that any binary $\bar{x}'$ violating the constraint (4.9) above cannot be a feasible solution. By assumption, $\sum_{i:\ (\bar{x}_0)_i = 0} \bar{x}'_i = 0$, i.e., $\bar{x}_0 \geq \bar{x}'$. Considering that $\bar{x}_0$ is infeasible, we know that for any set of $\widetilde{x}_k, k \in \mathcal{K}$ there exists some $\widehat{m} \in \mathcal{N}_{\widehat{k}}$ for some $\widehat{k} \in \mathcal{K}$ such that

$$\mathbb{P}_0 \left\{ E(\bar{\boldsymbol{x}}, \tau_{\widehat{k}}) \geq \sum_{(i,j) \in \mathcal{L}_{\widehat{k}}^{[\widehat{m}, o_{\widehat{k}}]}} c_{\widehat{k},i,j} \right\} < \min\{1, 1-\beta\},$$

$$E(\bar{\boldsymbol{x}}, \tau_{\widehat{k}}) := \sum_{j \in \mathcal{N}_k^{(\widehat{m}, o_{\widehat{k}})}} (p^{\mathsf{T}} \bar{\boldsymbol{x}}_j) \tau_{\widehat{k},j} + p^{\mathsf{T}} \bar{\boldsymbol{x}}_{o_{\widehat{k}}} \tau_{\widehat{k}, o_{\widehat{k}}},$$

where $\bar{\boldsymbol{x}}$, the concatentation of $\bar{x}_0$ and $\widetilde{x}_k, k \in \mathcal{K}$ is a solution to the entire transportation system. Also, by assumption, $\bar{\boldsymbol{x}}'$, the concatentation of $\bar{x}'$ and $\widetilde{x}_k, k \in \mathcal{K}$, is dominated by $\bar{\boldsymbol{x}}$, therefore for any specific realization of $\tau_{\widehat{k},j}, j \in \mathcal{N}_k^{\widehat{m}, o_{\widehat{k}}}$, we have

$$E(\bar{\boldsymbol{x}}, \tau_{\widehat{k}}) \geq E(\bar{\boldsymbol{x}}', \tau_{\widehat{k}}).$$

Thus,

$$\mathbb{P}_0 \left\{ E(\bar{\boldsymbol{x}}', \tau_{\widehat{k}}) \geq \sum_{(i,j) \in \mathcal{L}_{\widehat{k}}^{[\widehat{m}, o_{\widehat{k}}]}} c_{\widehat{k},i,j} \right\} < \min\{1, 1-\beta\},$$

which shows $\bar{x}'$ is infeasible. □

In practice, the dwell time at certain bus stops can usually be quite short, e.g., less than 1 minute with a probability of more than 90%. At these stops, fast charger may not be able to provide much energy even if it is installed despite its high cost. Hence, it is unlikely that the optimal solution could place chargers at such stops, and the following proposition helps us prove it while solving:

Proposition 21. Let `ub` be the current best upper bound we obtained so far and $q \in \{0, 1\}$ be a constant. If there exists an lower bound `lb` to (**AUG**) when one additional constraint $x_{i,t} = q$ is enforced, i.e.,

(**AUG-q**)
$$\begin{aligned} \min_{\bar{x}_k, \widetilde{x}_k, e_k^{\max}} \ & \sum_{k \in K} Q_k(\bar{x}_k, \widetilde{x}_k, e_k^{\max}), \\ \text{s.t. } & \text{(4.7a)–(4.7b)}, \\ & x_{i,t} = q, \end{aligned}$$

and $\texttt{lb} > \texttt{ub}$, then $x_{i,t} = (1-q)$ is a valid inequality to (**AUG**).

Proof. Let $\mathcal{X}$ and V_q be the feasible region and optimal value of (**AUG-q**). Then, for any feasible solution x' with $x'_{i,t} = q$, the objective value of (**AUG-q**) evaluated at x', V'_q satisfies:

$$\texttt{ub} < \texttt{lb} \leq V_q \leq V'_q,$$

which implies that at optimality $x^*_{i,t} \neq q$. The conclusion follows by noting that $q \in \{0, 1\}$. □

After we collect feasible solutions $S_\Delta := \{ \bar{x}^*_k \colon k \in \mathcal{K} \}$ from $|\mathcal{K}|$ subproblems, we figure out the set of decision variables that all $|\mathcal{K}|$ solutions unanimously set to be 1 and 0 and call them $\mathcal{S}_1$ and $\mathcal{S}_0$:

$$\begin{aligned} \mathcal{S}_0 &:= \left\{ (i,t) \in \bar{\mathcal{N}} \times \mathcal{T} \colon (\bar{x}^*_k)_{i,t} = 0, \forall\, k \in \mathcal{K} \right\}, \\ \mathcal{S}_1 &:= \left\{ (i,t) \in \bar{\mathcal{N}} \times \mathcal{T} \colon (\bar{x}^*_k)_{i,t} = 1, \forall\, k \in \mathcal{K} \right\}, \end{aligned}$$

Then, for variables in set $\mathcal{S}_q, q \in \{0, 1\}$, we try to cut off the subspace $x_{i,t} = 1 - q$ by comparing the current best upper bound `ub` with the lower bound of a "relaxed" (**DRCC**) model, where only domain constraints, linear constraints, and constraints (4.4a) with $m, n = o_k$ are kept. Although solving the whole system with $2K$ SOC constraints seems to be time consuming, there is actually no need to solve this relaxed problem to optimality since we only care about whether there exists a lower bound larger than our current best upper bound or not. Setting the `BestBdStop` and `BestObjStop` parameters in `Gurobi` to `ub` can reduce the overall time costs a lot.

4.4 Numerical Experiments

In this section, we demonstrate the effectiveness of (**DRCC**) model on the shuttle bus system of Salt Lake City (SLC), as shown in Figure 4.3, where the six lines it includes (line $2, 3, 6, 11, 519, 520$) have two different base stations, cover 294 bus stops in total, and involve 53 shared bus stops. Detailed parameters are shown in Table 4.2. For each line, we obtain independent samples of the dwell time at each bus stop and the energy consumption on every edge, with which the ambiguity set $\mathcal{D}_\phi$ for each single chance constraint can be constructed using the empirical mean and covariance matrices.

All experiments in this section are implemented using the Python API of Gurobi $9.0.2$, Open MPI and conducted on a single node of the Great Lakes cluster provided by University of Michigan, which contains two 3.0GHz Intel Xeon Gold 6154 CPUs.

In the following subsections, we first discuss the selection of radius ρ; next we benchmark the performance of **(DET)**, (**CC**) and (**DRCC**) on the SLC data set. Then, we show specifically how different choices of radius ρ and sample size N impact the performance of (**DRCC**) and (**CC**) models. At last, we showcase how optimal solutions of our proposed (**DRCC**) model differs from the (**CC**) model.

Table 4.2: SLC data parameters

Parameters of SLC data set	Value
Charging power of type 1 charging station (kW)	100
Charging power of type 2 charging station (kW)	200
Amortized cost of installing and maintaining a type 1 charging station ($)	21912
Amortized cost of installing and maintaining a type 2 charging station ($)	28884
Amortized cost of battery, including replacement ($/kWh)	52
Lower bound of remaining battery power (%)	20
Upper bound of remaining battery power (%)	80

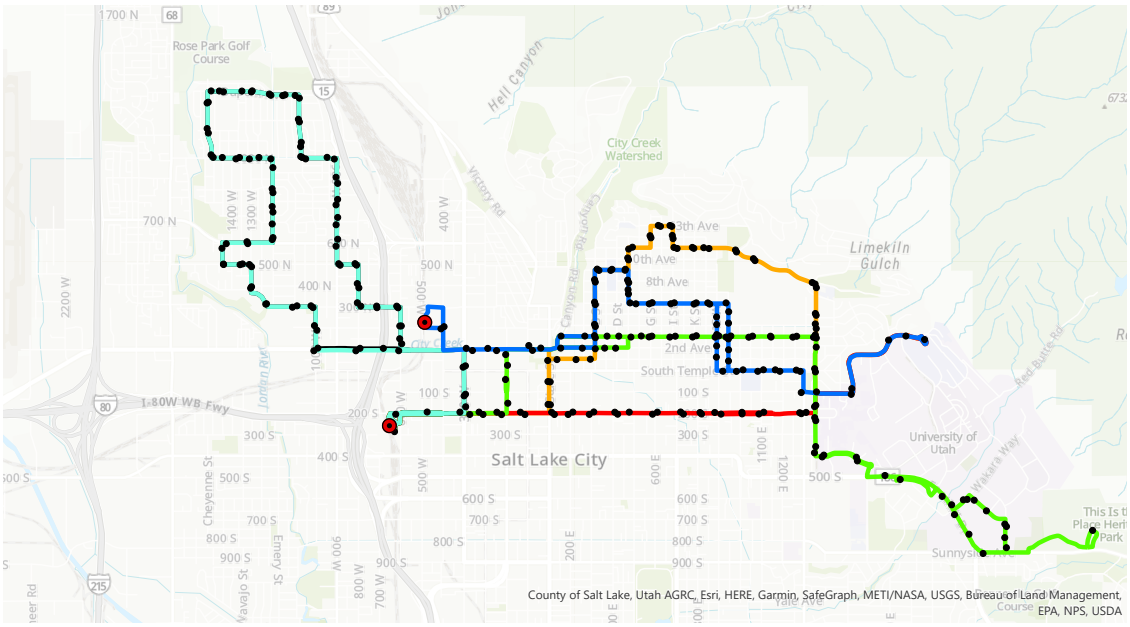


Figure 4.3: An overview of SLC transportation system

4.4.1 The selection of radius ρ

Different choices of the radius ρ reflect how conservative the decision maker is: selecting a large ρ includes more distributions in $\mathcal{D}_\phi$ and hence leads to a more reliable solution with higher costs; a smaller ρ produces a lower cost decision, which has a higher risk of violating the chance constraint under the true distribution. To balance and k-*fold cross-validation* to choose a good radius ρ_* from all the candidates $\{\,\rho_i\,\}_{i=1}^{M}, M \in \mathbb{Z}_+$. The holdout method partitions the given samples into a training set of size $N_{\text{train}} \in \mathbb{Z}_+$ and a test set of size $N_{\text{test}} = N - N_{\text{train}}$. Then, for every candidate ρ_i, it constructs an ambiguity set $\mathcal{D}_{\phi_i}$ with radius ρ_i and solve (**DRCC**) for an optimal solution, whose performance is evaluated on the test set. The radius with the best evaluation is reported as ρ_*. k-fold cross-validation partitions all samples into $k \in \mathbb{Z}_+, k \geq 2$ subsets and run holdout method k times for every candidate ρ_i. At each time, one of the k subsets is chosen to be the test set, while the other $k-1$ subsets become the training set. The candidates in $\{\,\rho_i\,\}_{i=1}^{M}$ with the higherst average performance on the k runs of holdout method is selected as ρ_*.

4.4.2 Benchmark between (DET), (CC) and (DRCC)

In this section, we benchmark the performance of **(DET)**, (**CC**) and (**DRCC**) on SLC data with $\alpha = 0.1$ and sample sizes in $\{100, 150, 200, 300, 400, 600, 800, 1000\}$. If the number of samples in SLC is not enough, we compensate for this shortage by uniformly resampling the existing samples. To evaluate the two models, we split the data into a training set with 80% of all the samples and a test set with the rest 20%. We first run 4-fold cross-validation on the training set to select the radius ρ for (**DRCC**) model from 10 evenly spaced numbers between 0.001 to 0.025. Then we solve **(DET)**, (**CC**), and (**DRCC**) on the training set, and obtain their optimal solutions $(x_{\mathsf{deter}}, e_{\mathsf{deter}}, e_{\mathsf{deter}}^{\mathsf{max}})$, $(x_{\mathsf{cc}}, e_{\mathsf{cc}}, e_{\mathsf{cc}}^{\mathsf{max}})$ and $(x_{\mathsf{dr}}, e_{\mathsf{dr}}, e_{\mathsf{dr}}^{\mathsf{max}})$, respectively. Next, we evaluate their returned solution on the test set using the following formulae:

$$
\begin{aligned}
R_1(x, e, e^{\mathsf{max}}) &:= \min_{k \in \mathcal{K}, m \in \mathcal{N}_k} \widehat{\mathbb{P}}_{\mathsf{test}} \left\{ 0 \geq \left(\boldsymbol{z}'_{k,m,o_k} \right)^{\mathsf{T}} \boldsymbol{H}'_{k,m,o_k} \boldsymbol{\xi}'_{k,m,o_k} \right\}, \\
R_2(x, e, e^{\mathsf{max}}) &:= \min_{\substack{k \in \mathcal{K}, m,n \in \mathcal{N}_k, \\ m <_k n}} \widehat{\mathbb{P}}_{\mathsf{test}} \left\{ (\bar{\epsilon} - \underline{\epsilon}) \cdot e_k^{\mathsf{max}} \geq \left(\boldsymbol{z}_{k,m,n} \right)^{\mathsf{T}} \boldsymbol{H}_{k,m,n} \boldsymbol{\xi}_{k,m,n} \right\}, \\
R(x, e, e^{\mathsf{max}}) &:= \min \left\{ R_1(x, e, e^{\mathsf{max}}), R_2(x, e, e^{\mathsf{max}}) \right\},
\end{aligned}
$$

where $R_1(x, e, e^{\mathsf{max}})$ and $R_2(x, e, e^{\mathsf{max}})$ represent the reliability of constraints (**CC**), $R(x, e, e^{\mathsf{max}})$ is the reliability of all the single chance constraints, and $\widehat{\mathbb{P}}_{\mathsf{test}}$ denotes the empirical distribution generated by the test data. Figure 4.4 shows the average reliability, as well as the 90% confidence interval, of the returned solutions, for different sample sizes N over 5 runs, each of which partitions the training set and the test set randomly. From Figure 4.4, we note that the deterministic model **(DET)** keeps failing to return a reliable solution even when 800 samples are given. In contrast, the chance constraint model (**CC**) achieves a much higher reliability all the time, but its returned optimal solutions are still not feasible when evaluated under the test data for all sample sizes. (**DRCC**) model consistently achieves a even higher average reliability and it only requires no more than 250 samples to get a feasible solution to all single chance constraints evaluated using $\widehat{\mathbb{P}}_{\mathsf{test}}$. This demonstrates the robustness of our (**DRCC**) model against the ambiguous true probability distribution $\mathbb{P}$ when the number of training samples is limited.

4.4.3 Impact of radius ρ and sample size N on reliability

To study how radius ρ and sample size N affect the reliability on the test set, we show the performance of (**CC**) and (**DRCC**) on SLC data with the same set of sample sizes and

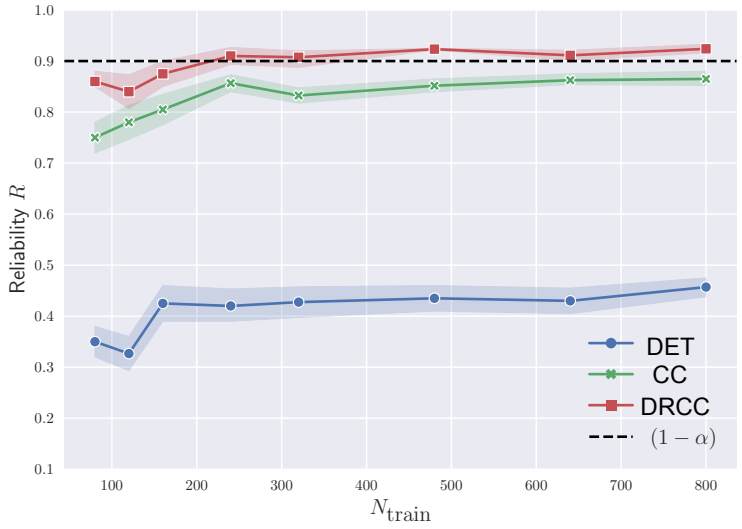
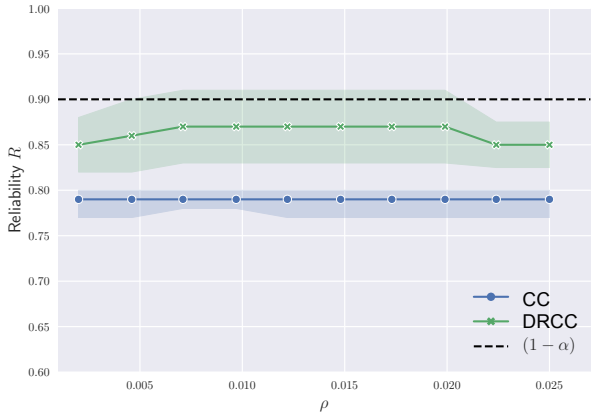


Figure 4.4: Benchmark between (**CC**) and (**DRCC**) over different training sizes N_{train}, $\alpha = 0.1$

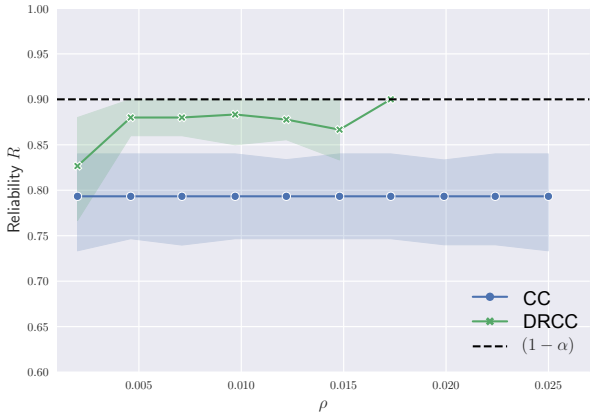
the same range of ρ's as in Section 4.4.2. Similarly, for all sample sizes, 80% is used to fit the two models, and the rest 20% is used to evaluate the reliability. Figure 4.5 shows the average reliability and its 90% confidence interval of the two models under different radius ρ's. We notice that the reliability R of (**DRCC**) model increases as ρ increases because the larger the radius ρ is, the more distributions are included in $\mathcal{D}_\phi$, and hence, $(x_{\text{dr}}, e_{\text{dr}}, e_{\text{dr}}^{\max})$ is more robust. Furthermore, Figure 4.5a shows that (**DRCC**) model can produce a solution feasible to $\widehat{\mathbb{P}}_{\text{test}}$ even when the number of training samples is very limited as long as a proper ρ is chosen. Figure 4.6 shows the change in average reliability and its 90% confidence interval when the number of samples increases. Not surprisingly, we note that both model become more reliable when the number of samples increases; specifically, when ρ is small, e.g., $\rho = 0.002$, or $\rho = 0.0045$, the performance of (**DRCC**) model almost aligns with (**CC**), however, when ρ is large, (**DRCC**) model appears to be more reliable, e.g., $\rho = 0.022$ or $\rho = 0.025$.

4.4.4 Comparision of optimal solutions of (DET), (DRCC) and (CC)

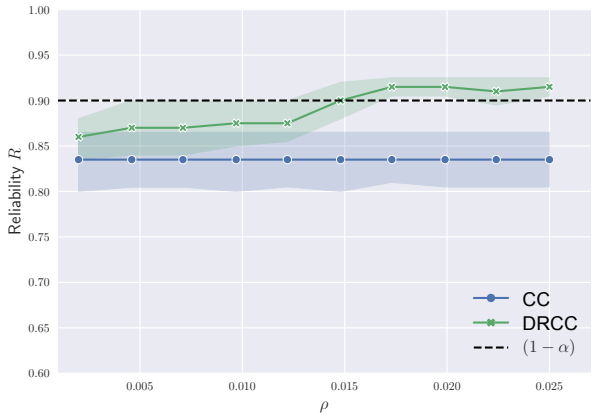
In this subsection, we compare the optimal solutions returned by **(DET)**, (**DRCC**) and (**CC**) on a SLC system with $\rho = 0.01, N = 200, \alpha = 0.1$, where 160 samples is used as training data to fit the three models, and the rest 40 samples comprises the test set, on which we evaluate their reliability. Table 4.3 shows the optimal decisions including the size of the batteries for each bus line, number of chargers of each type placed, overall



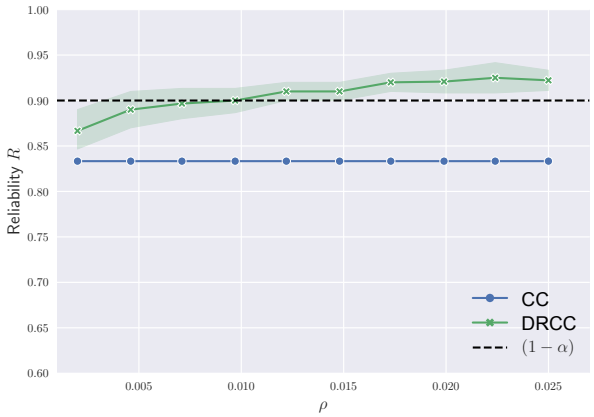
(a) $\alpha = 0.1, N_{\text{train}} = 80$



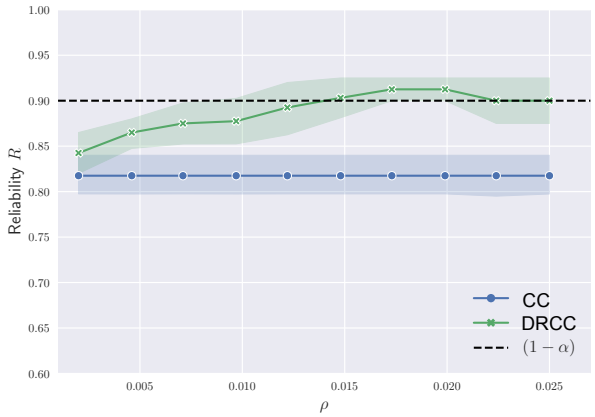
(b) $\alpha = 0.1, N_{\text{train}} = 120$



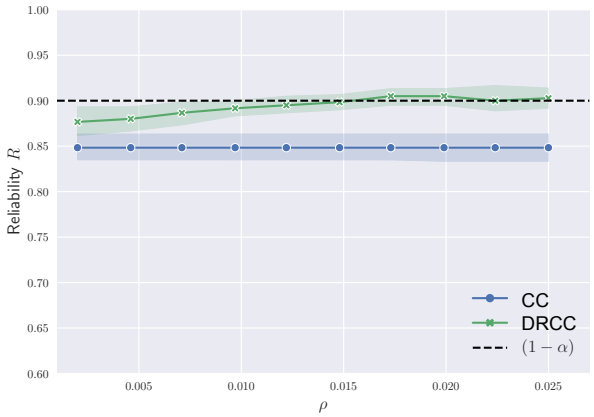
(c) $\alpha = 0.1, N_{\text{train}} = 160$



(d) $\alpha = 0.1, N_{\text{train}} = 240$

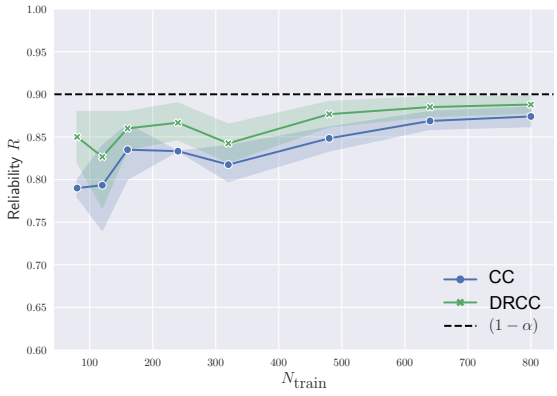


(e) $\alpha = 0.1, N_{\text{train}} = 320$

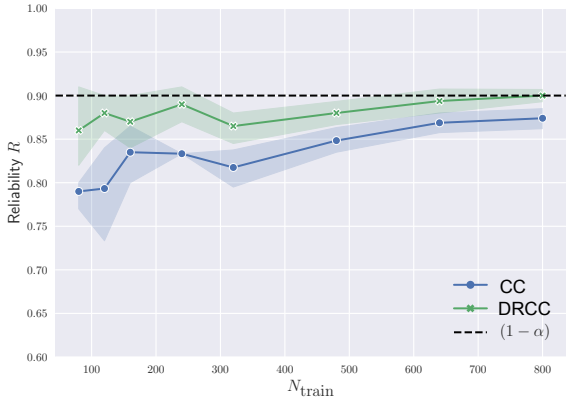


(f) $\alpha = 0.1, N_{\text{train}} = 480$

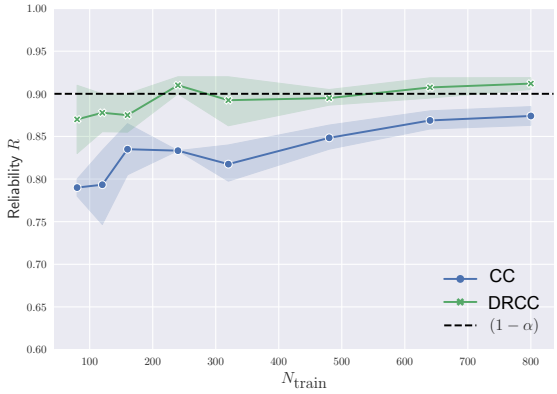
Figure 4.5: Reliability of (**DRC**) and (CC) as a function of δ.



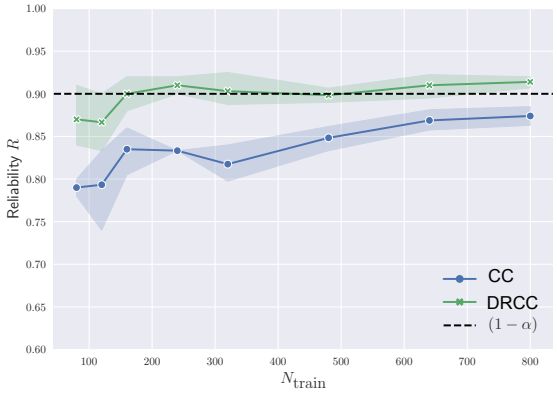
(a) $\alpha = 0.1, \rho = 0.002$



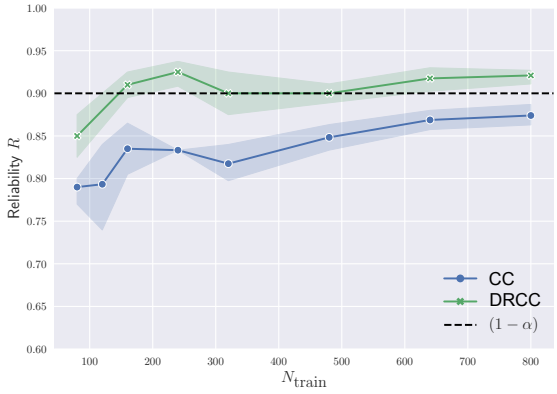
(b) $\alpha = 0.1, \rho = 0.0045$



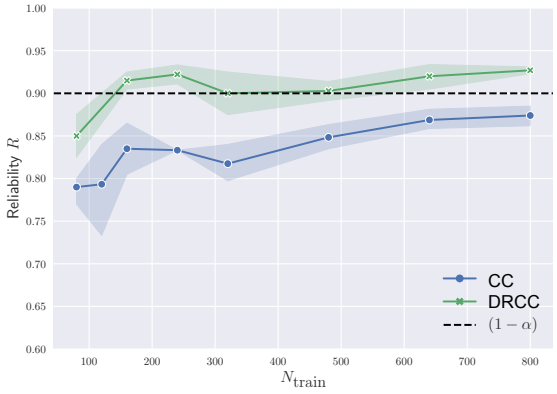
(c) $\alpha = 0.1, \rho = 0.012$



(d) $\alpha = 0.1, \rho = 0.015$



(e) $\alpha = 0.1, \rho = 0.022$



(f) $\alpha = 0.1, \rho = 0.025$

Figure 4.6: Reliability of (**DRCC**) and (**CC**) as a function of N_{train}.

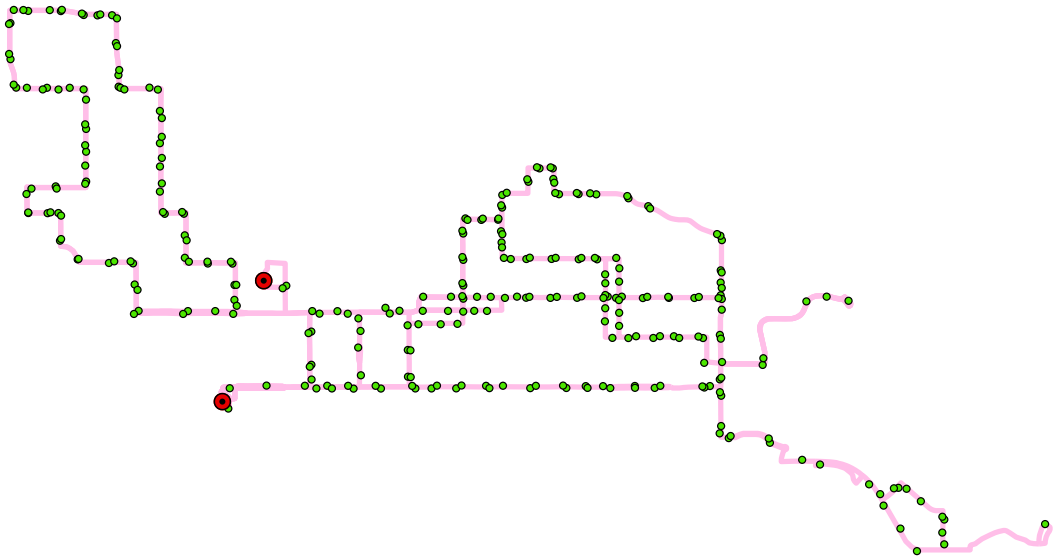
costs, and the reliability on the test set. Comparing the Column of **(DET)** and (**CC**) we observe that both of them placed two 200kW chargers at the base station but the battery sizes chosen by **(DET)** are much smaller than (**CC**), which results in a less reliable plan but cheaper. Figure 4.7 shows the optimal solutions of (**CC**) and (**DRCC**) in detail; both of them place chargers at the two base stations, but (**DRCC**) model also placed two 200kW chargers and two 100 kW chargers at four other bus stops. Although the overall cost (248,438.57) is nearly twice as (**CC**)'s (143,251.68), it provides a much reliable solution satisfying our required risk level of 0.1.

Table 4.3: Optimal solutions of different models

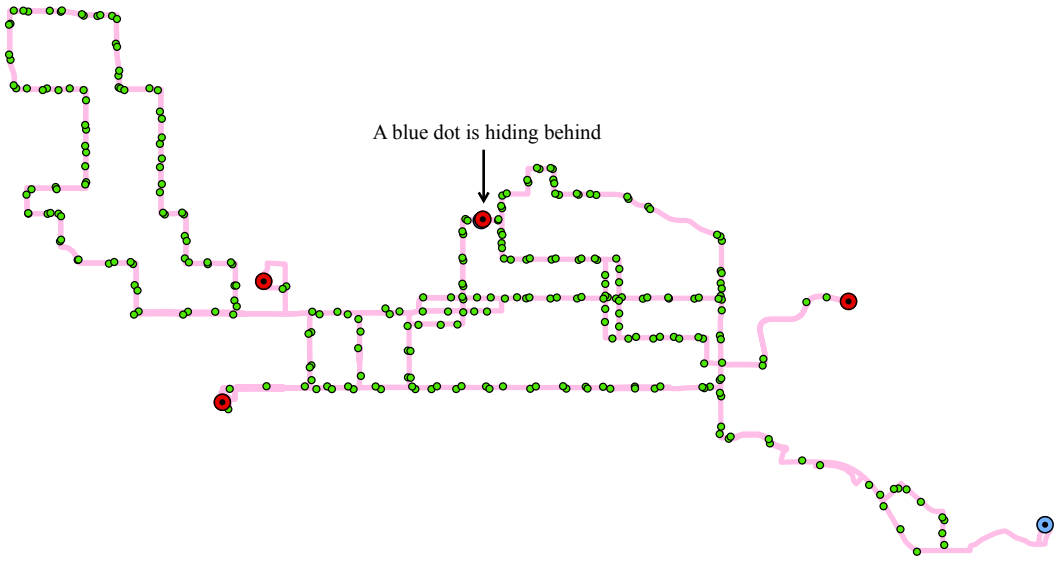
		(DET)	(**CC**)	(**DRCC**)
Battery size (kWh)	line 519	19.66	23.58	24.30
	line 520	19.98	24.00	24.73
	line 11	25.08	29.85	27.77
	line 6	25.24	30.63	27.86
	line 3	32.49	38.51	46.66
	line 2	20.49	24.40	26.83
# of chargers	100	0	0	2
	200	2	2	4
Overall costs ($)		115,293.53	143,251.68	248,438.57
Reliability (%)		48.13	87.50	91.88

4.4.5 Computational performance

In this subsection, we compare the computational performance between our monolithic reformulation in Section 4.2.2 (`Mono`), separation based algorithm, Algorithm 3 (`Sep`), and dual decomposition based algorithm, Algorithm 4 (`DD`). Specifically, we randomly test different problem sizes, $N \in \{ 100, 200, 400, 800 \}$, multiple radii $\rho \in \{ 0.05, 0.01, 0.015 \}$, and $\alpha = 0.1$. For each problem size, we randomly generate 5 different instances, each of which is solved by three algorithms to 0.1% gap with a total timelimit of two hours (7200 seconds). Since `Gurobi` can utilize parallel computing in the branch-and-bound tree, both `Mono` and `Sep` can benefit from multiple cores. Therefore, we give all three algorithms the same amount of computational resources, i.e., 36 cores on a single node of the cluster, and report the average solution time (in seconds) taken, best objective value obtained, and the final optimality gap (%) reached by each algorithm in Table 4.4, where an "inf" in Column gap and "N/A" in Column value indicate that the algorithm failed to return a feasible solution to (**DRCC**) model in any of the 5 instances. In particular, we notice that solving the entire mixed-integer SOCP together, i.e., `Mono`, usually cannot return a feasible solution within the timelimit. In contrast, `DD` can solve most of the instances to optimality



(a) Optimal solution of (**CC**) model.



(b) Optimal solution of (**DRCC**) model.

●: Bus stations, ●: Charger of type 1, ●: Charger of type 2.

Figure 4.7: Comparision between optimal solutions of (**CC**) and (**DRCC**)

and require much less time as shown in Table 4.4. For example, when $N = 100, \rho = 0.01$, DD took 1574.89 seconds on average and successfully solved all 5 instances, but Mono failed to return any feasible solution to any instance. Although DD encountered some difficulties when ρ is large, e.g., $\rho = 0.015$, it proved much smaller optimality gap on average (9.6% when $N = 100$, 4.5% when $N = 200$, and 15.5% when $N = 800$) compared to Mono, which demonstrates the effectiveness of our proposed decomposition strategy and valid inequalities. Comparing Sep and DD, we note that Sep is much faster than DD on this dataset even in these difficult instances where ρ is large: When $N = 800, \rho = 0.015$, DD could not solve all 5 instances, but Sep consistently outperformed its counterpart by solving all the instances with an average time of 313.37 seconds. Moreover, the average time it took to solve all tests was 97.39 seconds, much quicker than DD, which took 1903.31 seconds on average. Although constraints of type (4.2) are usually the hardest ones to satisfy and provide a good approximation to the entire (**DRCC**) model already, the unexpected efficiency of Sep, in our opinion, relies on the success of low rank approximation as well (Figure 4.2), as the amount of maximum violation after the first solve is usually small and the algorithm often converges quickly with only a few additional SOC constraints added. Even though DD was slower than Sep on our tests, we believe it has better generalizability especially to data sets with large number of bus lines.

Table 4.4: Computational benchmark between three formulations

		DD			Mono			Sep		
N	ρ	time	gap	value	time	gap	value	time	gap	value
100	0.005	957.51	0.0	196506.05	7221.58	inf	N/A	33.25	0.0	196429.75
	0.01	1574.89	0.0	241883.82	7221.18	inf	N/A	32.07	0.0	241799.75
	0.015	3923.77	9.6	300667.29	7220.30	inf	N/A	253.37	0.0	300403.62
200	0.005	983.20	0.0	193124.75	7314.56	inf	N/A	26.61	0.0	193124.74
	0.01	1199.95	0.0	232895.98	7228.41	inf	N/A	27.86	0.0	232895.98
	0.015	3877.30	4.5	289661.66	7431.48	inf	N/A	317.95	0.0	289661.65
400	0.005	962.73	0.0	186979.77	7217.68	inf	N/A	25.53	0.0	186979.76
	0.01	1567.68	0.0	226197.36	7222.89	inf	N/A	40.82	0.0	226147.72
	0.015	1996.25	0.0	257382.36	7221.73	inf	N/A	41.29	0.0	257297.51
800	0.005	933.00	0.0	199456.73	7229.63	inf	N/A	23.20	0.0	199456.73
	0.01	1115.15	0.0	242136.72	7219.76	inf	N/A	33.38	0.0	242126.33
	0.015	3748.26	15.5	310888.08	7221.85	inf	N/A	313.37	0.0	309522.71
Average		1903.31	2.5	239815.05	7247.59	inf	N/A	97.39	0.0	239653.86

4.5 Conclusion

In this chapter, we apply RCCs to place charging stations for BEBs. Specifically, we formulate a (DRCC) model that explicitly considers the uncertain energy consumption and dwell time on the bus routes and recast it as a MISOCP. Because naively solving this deterministic reformulation using commercial solvers is time consuming, we propose two decomposition algorithms to improve the computational performance, the first of which approximate SOC constraints by linear inequalities, and the second of which is based on dual decomposition, which decomposes the system into smaller subproblems pertaining to individual bus lines and exploits parallel computing. Numerical experiments on the real transit system in SLC demonstrate the effectiveness of our proposed DRCC model and the efficiency of our decomposition algorithms.

CHAPTER 5

Conclusions

In this dissertation, we aim to alleviate the computational burden of solving RCCs. Specifically, we first identified a family of sufficient conditions under which such RCCs are actually convex and even polynomially solvable. In addition, when the problem does not satisfies such sufficient conditions but has a set covering structure, which appear frequently in real-world applications, we exploit the binary support information intrinsic in the problem and derive the convex hull of an underlying mixed-integer set, thus greatly improving the computational performance. Besides, we apply RCCs to the planning of charging stations for BEBs and design customized solution approaches that could efficiently obtain optimal locations for charging stations on real-sized transit systems. Possible future directions include: (1) design efficient algorithms for solving large-scale convex RCCs. (2) investigate the computational complexity of solving convex RCCs. (3) study how to properly enforce RCCs in a dynamic setting.

References

[AA09] S. Ahmed and A. Atamtürk. “Maximizing a Class of Submodular Utility Functions”. In: *Mathematical Programming* 128.1-2 (2009), pp. 149–169 (cit. on p. 59).

[Ahm13] S. Ahmed. “A Scenario Decomposition Algorithm for 0-1 Stochastic Programs”. In: *Operations Research Letters* 41.6 (Nov. 2013), pp. 565–569. ISSN: 0167-6377 (cit. on pp. 100, 101, 103).

[AN08] A. Atamtürk and V. Narayanan. “Polymatroids and Mean-Risk Minimization in Discrete Optimization”. In: *Operations Research Letters* 36.5 (2008), pp. 618–622. ISSN: 0167-6377 (cit. on p. 55).

[An20] K. An. “Battery electric bus infrastructure planning under demand uncertainty”. In: *Transportation Research Part C: Emerging Technologies* 111 (2020), pp. 572–587 (cit. on p. 85).

[AP13] S. Ahmed and D. J. Papageorgiou. “Probabilistic Set Covering With Correlations”. In: *Operations Research* 61.2 (2013), pp. 438–452. ISSN: 0030364X, 15265463 (cit. on pp. 48, 49).

[Asi18] Asian Development Bank. *SUSTAINABLE TRANSPORT SOLUTIONS: Low-Carbon Buses in the People’s Republic of China*. Asian Development Bank, 2018. DOI: http://dx.doi.org/10.22617/TCS189646-2. 2018 (cit. on p. 83).

[Aus76] A. Auslender. “Optimisation”. In: *Méthodes numériques* (1976) (cit. on pp. 26, 28).

[Ayr18] J. Ayre. *New BYD ADL Electric Bus Fleet Deployed In London — Route 153 Now Fully Electric*. https://cleantechnica.com/2018/02/13/new-byd-adl-electric-bus-fleet-deployed-london-route-153-now-electric/. Accessed on March 15, 2019. 2018 (cit. on p. 83).

[BB09] P. Beraldi and M. E. Bruni. “An Exact Approach for Solving Integer Problems Under Probabilistic Constraints With Random Technology Matrix”. In: *Annals of Operations Research* 177.1 (Nov. 2009), pp. 127–137. ISSN: 1572-9338 (cit. on p. 48).

[BBV04] S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004 (cit. on p. 134).

[BC+11] H. H. Bauschke, P. L. Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*. Vol. 408. Springer, 2011 (cit. on p. 149).

[Bec15] A. Beck. “On the Convergence of Alternating Minimization for Convex Programming With Applications To Iteratively Reweighted Least Squares and Decomposition Schemes”. In: *SIAM Journal on Optimization* 25.1 (Jan. 2015), pp. 185–209. ISSN: 1095-7189 (cit. on pp. 26, 28).

[BEN09] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*. Vol. 28. Princeton University Press, 2009 (cit. on p. 47).

[Ber97] D. P. Bertsekas. “Nonlinear Programming”. In: *Journal of the Operational Research Society* 48.3 (1997), pp. 334–334 (cit. on pp. 26, 28).

[BM19] J. Blanchet and K. Murthy. “Quantifying Distributional Model Risk Via Optimal Transport”. In: *Mathematics of Operations Research* 44.2 (2019), pp. 565–600 (cit. on p. 51).

[Bol+18a] N. Boland, J. Christiansen, B. Dandurand, A. Eberhard, J. Linderoth, J. Luedtke, and F. Oliveira. “Combining Progressive Hedging With a Frank–wolfe Method To Compute Lagrangian Dual Bounds in Stochastic Mixed-Integer Programming”. In: *SIAM Journal on Optimization* 28.2 (2018), pp. 1312–1336 (cit. on p. 100).

[Bol+18b] N. Boland, J. Christiansen, B. Dandurand, A. Eberhard, and F. Oliveira. “A Parallelizable Augmented Lagrangian Method Applied To Large-Scale Non-Convex-Constrained Optimization Problems”. In: *Mathematical Programming* 175.1-2 (Mar. 2018), pp. 503–536. ISSN: 1436-4646 (cit. on p. 100).

[BR02] P. Beraldi and A. Ruszczyński. “The Probabilistic Set-Covering Problem”. In: *Operations Research* 50.6 (2002), pp. 956–967 (cit. on pp. 6, 48).

[BS97] J. Bramel and D. Simchi-Levi. “On the Effectiveness of Set Covering Formulations for the Vehicle Routing Problem With Time Windows”. In: *Operations Research* 45.2 (1997), pp. 295–301 (cit. on p. 45).

[BT13] A. Beck and L. Tetruashvili. “On the Convergence of Block Coordinate Descent Type Methods”. In: *SIAM Journal on Optimization* 23.4 (Jan. 2013), pp. 2037–2060. ISSN: 1095-7189 (cit. on pp. 26, 28).

[Bul13] P. S. Bullen. *Handbook of means and their inequalities*. Vol. 560. Springer Science & Business Media, 2013 (cit. on p. 133).

[Cal18] California Air Resources Board. *Innovative Clean Transit Regulation*. http://www.dof.ca.gov/Forecasting/Economics/Major_Regulations/Major_Regulations_Table/documents/ICT_SRIA_ARB_4-23-18.pdf. Accessed on March 10, 2020. 2018 (cit. on p. 83).

[CC04] G. Calafiore and M. Campi. “Uncertain Convex Programs: Randomized Solutions and Confidence Levels”. In: *Mathematical Programming* 102.1 (2004), pp. 25–46 (cit. on p. 47).

[CC59] A. Charnes and W. W. Cooper. “Chance-Constrained Programming”. In: *Management science* 6.1 (1959), pp. 73–79 (cit. on p. 5).

[CCS58] A. Charnes, W. W. Cooper, and G. H. Symonds. “Cost Horizons and Certainty Equivalents: an Approach To Stochastic Programming of Heating Oil”. In: *Management Science* 4.3 (1958), pp. 235–263. ISSN: 00251909, 15265501 (cit. on pp. 5, 47).

[CE06] G. C. Calafiore and L. El Ghaoui. “On Distributionally Robust Chance-Constrained Linear Programs”. In: *Journal of Optimization Theory and Applications* 130.1 (2006), pp. 1–22 (cit. on pp. 6, 8, 47).

[Che+13] J. Chen, B. Atasoy, T. Robenek, M. Bierlaire, and M. Thémans. “Planning of feeding station installment for electric urban public mass-transportation system”. In: *13th Swiss Transportation Research Conference*. CONF. 2013 (cit. on p. 84).

[Chi14] Chicago Transit Authority. *CTA Announces First Electric-Powered Buses Added to its Fleet*. https://www.transitchicago.com/cta-announces-first-electric-powered-buses-added-to-its-fleet/. Accessed on March 15, 2019. 2014 (cit. on p. 83).

[CHS81] S. Cambanis, S. Huang, and G. Simons. “On the Theory of Elliptically Contoured Distributions”. In: *Journal of Multivariate Analysis* 11.3 (1981), pp. 368–385 (cit. on p. 131).

[CKW18] Z. Chen, D. Kuhn, and W. Wiesemann. “Data-Driven Chance Constrained Programs Over Wasserstein Balls”. In: *arXiv preprint arXiv:1809.00210* (2018) (cit. on pp. 8, 12, 47, 48).

[CS99] C. C. Carøe and R. Schultz. “Dual Decomposition in Stochastic Integer Programming”. In: *Operations Research Letters* 24.1-2 (1999), pp. 37–45 (cit. on pp. 100, 101).

[CX21] Z. Chen and W. Xie. “Sharing the Value-At-Risk Under Distributional Ambiguity”. In: *Mathematical Finance* 31.1 (2021), pp. 531–559 (cit. on pp. 24, 93).

[CY04] A. Chen and C. Yang. “Stochastic Transportation Network Design Problem With Spatial Equity Constraint”. In: *Transportation Research Record* 1882.1 (2004), pp. 97–104 (cit. on p. 47).

[CYS18] Z. Chen, Y. Yin, and Z. Song. “A cost-competitiveness analysis of charging infrastructure for electric bus operations”. In: *Transportation Research Part C: Emerging Technologies* 93 (2018), pp. 351–366 (cit. on p. 84).

[DJ88] S. Dharmadhikari and K. Joag-Dev. *Unimodality, Convexity, and Applications*. Elsevier, 1988 (cit. on p. 132).

[DS16] Y. Deng and S. Shen. “Decomposition Algorithms for Optimizing Multi-Server Appointment Scheduling With Chance Constraints”. In: *Mathematical Programming* 157.1 (2016), pp. 245–276 (cit. on p. 5).

[Du+19] J. Du, F. Li, J. Li, X. Wu, Z. Song, Y. Zou, and M. Ouyang. “Evaluating the technological evolution of battery electric buses: China as a case”. In: *Energy* 176 (2019), pp. 309–319 (cit. on p. 83).

[Dua+18] C. Duan, W. Fang, L. Jiang, L. Yao, and J. Liu. “Distributionally Robust Chance-Constrained Approximate AC-OPF With Wasserstein Metric”. In: *IEEE Transactions on Power Systems* 33.5 (2018), pp. 4924–4936 (cit. on p. 47).

[Eck19] B. Eckhouse. *The U.S. Has a Fleet of 300 Electric Buses. China Has 421,000*. Bloomberg, https://www.bloomberg.com/news/articles/2019-05-15/in-shift-to-electric-bus-it-s-china-ahead-of-u-s-421-000-to-300. Accessed on September 10, 2019. 2019 (cit. on p. 83).

[EOO03] L. El Ghaoui, M. Oks, and F. Oustry. “Worst-Case Value-At-Risk and Robust Portfolio Optimization: a Conic Programming Approach”. In: *Operations Research* 51.4 (2003), pp. 543–556 (cit. on pp. 8, 47).

[Erd45] P. Erdös. “Some Remarks on the Measurability of Certain Sets”. In: *Bulletin of the American Mathematical Society* 51.10 (1945), pp. 728–731 (cit. on p. 136).

[Fat+21] A. M. Fathabad, J. Cheng, K. Pan, and B. Yang. *Tight Conic Approximations for Chance-Constrained AC Optimal Power Flow*. 2021 (cit. on p. 6).

[Fd06] E. Finardi and E. daSilva. “Solving the Hydro Unit Commitment Problem Via Dual Decomposition and Sequential Quadratic Programming”. In: *IEEE Transactions on Power Systems* 21.2 (May 2006), pp. 835–844. ISSN: 0885-8950 (cit. on p. 100).

[FG14] N. Fournier and A. Guillin. “On the Rate of Convergence in Wasserstein Distance of the Empirical Measure”. In: *Probability Theory and Related Fields* 162.3-4 (2014), pp. 707–738 (cit. on p. 47).

[FG15] N. Fournier and A. Guillin. “On the rate of convergence in Wasserstein distance of the empirical measure”. In: *Probability theory and related fields* 162.3-4 (2015), pp. 707–738 (cit. on p. 92).

[FKN18] K.-T. Fang, S. Kotz, and K. W. Ng. *Symmetric multivariate and related distributions*. Chapman and Hall/CRC, 2018 (cit. on p. 133).

[FM12] M. Fischetti and M. Monaci. “Cutting Plane Versus Compact Formulations for Uncertain (integer) Linear Programs”. In: *Mathematical Programming Computation* 4.3 (Sept. 2012), pp. 239–273. ISSN: 1867-2957 (cit. on pp. 48, 49).

[Fra04] G. Frahm. “Generalized elliptical distributions: Theory and Applications”. Universität zu Köln, 2004 (cit. on p. 131).

[GK16] R. Gao and A. J. Kleywegt. “Distributionally Robust Stochastic Optimization With Wasserstein Distance”. In: (2016). arXiv: 1604.02199 [math.OC] (cit. on p. 157).

[GNS00] Z. Gu, G. L. Nemhauser, and M. W. Savelsbergh. “Sequence Independent Lifting in Mixed Integer Programming”. In: *Journal of Combinatorial Optimization* 4.1 (Mar. 2000), pp. 109–129. ISSN: 1573-2886 (cit. on p. 67).

[GP01] O. Günlük and Y. Pochet. “Mixing Mixed-Integer Inequalities”. In: *Mathematical Programming* 90.3 (May 2001), pp. 429–457. ISSN: 1436-4646 (cit. on pp. 48, 65).

[GS02] A. L. Gibbs and F. E. Su. “On Choosing and Bounding Probability Metrics”. In: *International Statistical Review* 70.3 (2002), pp. 419–435 (cit. on p. 47).

[GS99] L. Grippof and M. Sciandrone. “Globally Convergent Block-Coordinate Techniques for Unconstrained Optimization”. In: *Optimization methods and software* 10.4 (1999), pp. 587–637 (cit. on pp. 26, 28).

[Gun82] G. Gunawardane. “Dynamic Versions of Set Covering Type Public Facility Location Problems”. In: *European Journal of Operational Research* 10.2 (June 1982), pp. 190–195. ISSN: 0377-2217 (cit. on p. 45).

[Gup80] S. D. Gupta. “Brunn-Minkowski Inequality and Its Aftermath”. In: *Journal of Multivariate Analysis* 10.3 (1980), pp. 296–318. ISSN: 0047-259X (cit. on p. 130).

[Haj+20] A. Hajebrahimi, I. Kamwa, E. Delage, and M. Abdelaziz. “Adaptive Distributionally Robust Optimization for Electricity and Electrified Transportation Planning”. In: *IEEE Transactions on Smart Grid* (2020) (cit. on p. 85).

[Han+15] G. A. Hanasusanto, V. Roitch, D. Kuhn, and W. Wiesemann. “A Distributionally Robust Perspective on Uncertainty Quantification and Chance Constrained Programming”. In: *Mathematical Programming* 151.1 (2015), pp. 35–62 (cit. on pp. 8, 47).

[Han+17] G. A. Hanasusanto, V. Roitch, D. Kuhn, and W. Wiesemann. “Ambiguous Joint Chance Constraints Under Mean and Dispersion Information”. In: *Operations Research* 65.3 (2017), pp. 751–767 (cit. on pp. 8, 47).

[HKB05] K.-L. Hsiung, S.-J. Kim, and S. Boyd. “Power control in lognormal fading wireless channels with uptime probability specifications via robust geometric programming”. In: *Proceedings of the 2005, American Control Conference, 2005.* IEEE. 2005, pp. 3955–3959 (cit. on p. 47).

[Ho-+20] N. Ho-Nguyen, F. Kılınç-Karzan, S. Küçükyavuz, and D. Lee. “Strong Formulations for Distributionally Robust Chance-Constrained Programs with Left-Hand Side Uncertainty under Wasserstein Ambiguity”. In: *arXiv preprint arXiv:2007.06750* (2020). eprint: arXiv:2007.06750 (cit. on p. 48).

[Ho-+21] N. Ho-Nguyen, F. Kılınç-Karzan, S. Küçükyavuz, and D. Lee. “Distributionally Robust Chance-Constrained Programs With Right-Hand Side Uncertainty Under Wasserstein Ambiguity”. In: *Mathematical Programming* (Feb. 2021). ISSN: 1436-4646 (cit. on pp. 48, 66).

[Jan18] Y. J. Jang. “Survey of the operation and system study on wireless charging electric vehicle systems”. In: *Transportation Research Part C: Emerging Technologies* 95 (2018), pp. 844–866 (cit. on p. 84).

[JG15] R. Jiang and Y. Guan. “Data-Driven Chance Constrained Stochastic Program”. In: *Mathematical Programming* 158.1-2 (2015), pp. 291–327 (cit. on p. 47).

[JG16] R. Jiang and Y. Guan. “Data-driven chance constrained stochastic program”. In: *Mathematical Programming* 158.1-2 (2016), pp. 291–327 (cit. on p. 93).

[JL20] R. Ji and M. A. Lejeune. “Data-Driven Distributionally Robust Chance-Constrained Optimization With Wasserstein Metric”. In: *Journal of Global Optimization* 79.4 (Nov. 2020), pp. 779–811. ISSN: 1573-2916 (cit. on pp. 8, 48).

[JQS16] P. Jaillet, J. Qi, and M. Sim. “Routing optimization under uncertainty”. In: *Operations research* 64.1 (2016), pp. 186–200 (cit. on p. 85).

[KCC16] B.-R. Ke, C.-Y. Chung, and Y.-C. Chen. “Minimizing the costs of constructing an all plug-in electric bus transportation system: A case study in Penghu”. In: *Applied energy* 177 (2016), pp. 649–660 (cit. on p. 84).

[Kin17] M. Kinman. *DASHing to a Clean Electric Bus Future for Los Angeles*. https://environmentcalifornia.org/blogs/blog/cae/dashing-clean-electric-bus-future-los-angeles. Accessed on March 15, 2019. 2017 (cit. on p. 83).

[KJ21] S. Küçükyavuz and R. Jiang. “Chance-Constrained Optimization: A Review of Mixed-Integer Conic Formulations and Applications”. In: *arXiv preprint arXiv:2101.08746* (2021) (cit. on p. 6).

[KMG17] A. Kunith, R. Mendelevitch, and D. Goehlich. “Electrification of a city bus network—An optimization model for cost-effective placing of charging infrastructure and battery sizing of fast-charging electric bus systems”. In: *International Journal of Sustainable Transportation* 11.10 (2017), pp. 707–720 (cit. on pp. 84, 86).

[Küç12] S. Küçükyavuz. “On Mixing Sets Arising in Chance-Constrained Programming”. In: *Mathematical programming* 132.1-2 (2012), pp. 31–56 (cit. on pp. 6, 48).

[Kuh+19] 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 and Management Science in the Age of Analytics* (Oct. 2019), pp. 130–166 (cit. on pp. 7, 17, 85).

[LA08] J. Luedtke and S. Ahmed. “A Sample Approximation Approach for Optimization With Probabilistic Constraints”. In: *SIAM Journal on Optimization* 19.2 (2008), pp. 674–699 (cit. on pp. 6, 47).

[Laj14] A. Lajunen. “Energy consumption and cost-benefit analysis of hybrid and electric city buses”. In: *Transportation Research Part C: Emerging Technologies* 38 (2014), pp. 1–15 (cit. on p. 83).

[LAN08] J. Luedtke, S. Ahmed, and G. L. Nemhauser. “An Integer Programming Approach for Linear Programs With Probabilistic Constraints”. In: *Mathematical Programming* 122.2 (Oct. 2008), pp. 247–272. ISSN: 1436-4646 (cit. on p. 6).

[LBV15] M. Lubin, D. Bienstock, and J. P. Vielma. “Two-Sided Linear Chance Constraints and Extensions”. In: *arXiv preprint arXiv:1507.01995* (2015) (cit. on p. 6).

[Li16] J.-Q. Li. “Battery-electric transit bus developments and operations: A review”. In: *International Journal of Sustainable Transportation* 10.3 (2016), pp. 157–169 (cit. on p. 84).

[Li95] S. Li. “An Insurance and Investment Portfolio Model Using Chance Constrained Programming”. In: *Omega* 23.5 (1995), pp. 577–585 (cit. on p. 5).

[Lin+19] Y. Lin, K. Zhang, Z.-J. M. Shen, B. Ye, and L. Miao. “Multistage large-scale charging station planning for electric buses considering transportation network and power grid”. In: *Transportation Research Part C: Emerging Technologies* 107 (2019), pp. 423–443 (cit. on pp. 83, 84).

[LJM19] B. Li, R. Jiang, and J. L. Mathieu. “Ambiguous Risk Constraints With Moment and Unimodality Information”. In: *Mathematical Programming* 173.1-2 (2019), pp. 151–192 (cit. on pp. 8, 47).

[LLS01] C. Lagoa, X. Li, and M. Sznaier. “Application of probabilistically constrained linear programs to risk-adjusted controller design”. In: *Proceedings of the 2001 American Control Conference. (Cat. No.01CH37148)*. Vol. 2. 2001, 738–743 vol.2 (cit. on p. 6).

[Lov83] L. Lovász. “Submodular Functions and Convexity”. In: *Mathematical Programming The State of the Art* (1983), pp. 235–257 (cit. on p. 55).

[LS17] Z. Liu and Z. Song. “Robust planning of dynamic wireless charging infrastructure for battery electric buses”. In: *Transportation Research Part C: Emerging Technologies* 83 (2017), pp. 77–103 (cit. on pp. 83, 84).

[LSH18] Z. Liu, Z. Song, and Y. He. “Planning of fast-charging stations for a battery electric bus system under energy consumption uncertainty”. In: *Transportation Research Record* 2672.8 (2018), pp. 96–107 (cit. on pp. 85–87).

[LT93] Z.-Q. Luo and P. Tseng. “Error Bounds and Convergence Analysis of Feasible Descent Methods: a General Approach”. In: *Annals of Operations Research* 46.1 (1993), pp. 157–178 (cit. on pp. 26, 28).

[Lub+13] M. Lubin, K. Martin, C. G. Petra, and B. Sandıkçı. “On Parallelizing Dual Decomposition in Stochastic Integer Programming”. In: *Operations Research Letters* 41.3 (May 2013), pp. 252–258. ISSN: 0167-6377 (cit. on p. 100).

[Lue14] J. Luedtke. “A Branch-And-Cut Decomposition Algorithm for Solving Chance-Constrained Mathematical Programs With Finite Support”. In: *Mathematical Programming* 146.1-2 (2014), pp. 219–244 (cit. on pp. 6, 48).

[Mah+16] M. Mahmoud, R. Garnett, M. Ferguson, and P. Kanaroglou. “Electric buses: A review of alternative powertrains”. In: *Renewable and Sustainable Energy Reviews* 62 (2016), pp. 673–684 (cit. on p. 83).

[MG06] P. A. Miranda and R. A. Garrido. “A Simultaneous Inventory Control and Facility Location Model With Stochastic Capacity Constraints”. In: *Networks and Spatial Economics* 6.1 (2006), pp. 39–53 (cit. on pp. 5, 47).

[MK18a] P. Mohajerin Esfahani and D. Kuhn. “Data-Driven Distributionally Robust Optimization Using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations”. In: *Mathematical Programming* 171.1 (Sept. 2018), pp. 115–166. ISSN: 1436-4646 (cit. on pp. 7, 46).

[MK18b] P. Mohajerin Esfahani and D. Kuhn. “Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations”. In: *Mathematical Programming* 171.1-2 (2018), pp. 115–166 (cit. on p. 92).

[Moh21] A. Mohammadi Fathabad. *Outlier Detection based on Robust Regression via Chance-Constrained Programming*. 2021 (cit. on p. 6).

[MW65] B. L. Miller and H. M. Wagner. “Chance Constrained Programming With Joint Constraints”. In: *Operations Research* 13.6 (1965), pp. 930–945 (cit. on pp. 5, 47).

[NAS18] NASEM. “Battery Electric Buses—State of the Practice”. In: (2018). Washington, DC: The National Academies Press. https://doi.org/10.17226/25061 (cit. on pp. 83, 84).

[New18] New York State Pressroom. *Governor Cuomo Announces All-Electric Bus Pilot Program to Reduce Emissions and Modernize Public Transit Fleet*. https://www.governor.ny.gov/news/governor-cuomo-announces-all-electric-bus-pilot-program-reduce-emissions-and-modernize-public. Accessed on March 15, 2019. 2018 (cit. on p. 83).

[Nor93] V. I. Norkin. “The Analysis and Optimization of Probability Functions”. In: (1993) (cit. on pp. 26, 33, 39, 136, 145).

[NS07] A. Nemirovski and A. Shapiro. “Convex Approximations of Chance Constrained Programs”. In: *SIAM Journal on Optimization* 17.4 (2007), pp. 969–996 (cit. on p. 47).

[Par18] L. Pardo. *Statistical inference based on divergence measures*. CRC press, 2018 (cit. on p. 92).

[PP63] C. van de Panne and W. Popp. “Minimum-Cost Cattle Feed Under Probabilistic Protein Constraints”. In: *Management Science* 9.3 (Apr. 1963), pp. 405–430. ISSN: 1526-5501 (cit. on p. 6).

[Pré13] A. Prékopa. *Stochastic programming*. Vol. 324. Springer Science & Business Media, 2013 (cit. on pp. 6, 15, 26).

[Pré70] A. Prékopa. “On probabilistic constrained programming”. In: *Proceedings of the Princeton Symposium on Mathematical Programming*. Vol. 113. Princeton, NJ. 1970, p. 138 (cit. on pp. 5, 47).

[Pré90] A. Prékopa. “Dual Method for the Solution of a One-Stage Stochastic Programming Problem With Random RHS Obeying a Discrete Probability Distribution”. In: *ZOR Zeitschrift für Operations Research Methods and Models of Operations Research* 34.6 (Nov. 1990), pp. 441–461. ISSN: 1432-5217 (cit. on pp. 6, 48).

[Pro19] Proterra. *CATALYST® : 40 FOOT BUS PERFORMANCE SPECIFICATIONS*. https://www.proterra.com/wp-content/uploads/2019/06/Proterra-Catalyst-40-ft-Spec-Sheet.pdf. Accessed on March 10, 2020. 2019 (cit. on p. 83).

[RM19] H. Rahimian and S. Mehrotra. “Distributionally robust optimization: A review”. In: *arXiv preprint arXiv:1908.05659* (2019) (cit. on p. 85).

[RT10] C. G. Rawls and M. A. Turnquist. “Pre-Positioning of Emergency Supplies for Disaster Response”. In: *Transportation Research Part B: Methodological* 44.4 (2010), pp. 521–534. ISSN: 0191-2615 (cit. on p. 73).

[RU99] R. T. Rockafellar and S. Uryasev. “Optimization of Conditional Value-At-Risk”. In: *Journal of Risk* (1999) (cit. on pp. 11, 50).

[Sch03] A. Schrijver. *Combinatorial Optimization: Polyhedra and Efficiency*. Vol. 24. Springer Science & Business Media, 2003 (cit. on pp. 54, 55, 59).

[SDR09] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009 (cit. on pp. 6, 15, 130).

[SG83] W. R. Stewart Jr and B. L. Golden. "Stochastic Vehicle Routing: a Comprehensive Approach". In: *European Journal of Operational Research* 14.4 (1983), pp. 371–385 (cit. on p. 5).

[SGL10] A. Saxena, V. Goyal, and M. A. Lejeune. "MIP Reformulations of the Probabilistic Set Covering Problem". In: *Mathematical Programming* 121.1 (2010), pp. 1–31 (cit. on p. 48).

[SJ21] H. Shen and R. Jiang. "Convex chance-constrained programs with Wasserstein ambiguity". In: *arXiv preprint arXiv:2111.02486* (2021) (cit. on p. 93).

[SL13] Y. Song and J. R. Luedtke. "Branch-And-Cut Approaches for Chance-Constrained Formulations of Reliable Network Design Problems". In: *Mathematical Programming Computation* 5.4 (Aug. 2013), pp. 397–432. ISSN: 1867-2957 (cit. on p. 48).

[SLK14] Y. Song, J. R. Luedtke, and S. Küçükyavuz. "Chance-Constrained Binary Packing Problems". In: *INFORMS Journal on Computing* 26.4 (2014), pp. 735–747 (cit. on p. 48).

[ST20] K. S. Shehadeh and E. L. Tucker. "A Distributionally Robust Optimization Approach for Location and Inventory Prepositioning of Disaster Relief Supplies". In: (2020). arXiv: 2012.05387 [math.OC] (cit. on p. 73).

[STA09] P. Schütz, A. Tomasgard, and S. Ahmed. "Supply Chain Design Under Uncertainty Using Sample Average Approximation and Dual Decomposition". In: *European Journal of Operational Research* 199.2 (Dec. 2009), pp. 409–419. ISSN: 0377-2217 (cit. on p. 100).

[SW14] A. Saumard and J. A. Wellner. "Log-Concavity and Strong Log-Concavity: a Review". In: *Statistics surveys* 8 (2014), p. 45 (cit. on p. 33).

[SW88] B. M. Smith and A. Wren. "A Bus Crew Scheduling System Using a Set Covering Formulation". In: *Transportation Research Part A: General* 2 (1988), pp. 97–108 (cit. on p. 45).

[SZ21] D. Singh and S. Zhang. "Distributionally Robust Profit Opportunities". In: *Operations Research Letters* 49.1 (2021), pp. 121–128. ISSN: 0167-6377 (cit. on p. 7).

[TLH19] X. Tang, X. Lin, and F. He. “Robust scheduling strategies of electric buses under stochastic traffic conditions”. In: *Transportation Research Part C: Emerging Technologies* 105 (2019), pp. 163–182 (cit. on p. 83).

[Top78] D. M. Topkis. “Minimizing a Submodular Function on a Lattice”. In: *Operations Research* 26.2 (1978), pp. 305–321. ISSN: 0030364X, 15265463 (cit. on p. 54).

[Uta16] Utah Transit Authority. *UTA Announces Plans to Add First All-Electric Buses to Fleet*. https://www.rideuta.com/news/2016/04/UTA-Announces-Plans-for-All-Electric-Buses. Accessed on March 15, 2019. 2016 (cit. on p. 83).

[VBC07] L. Vandenberghe, S. Boyd, and K. Comanor. “Generalized Chebyshev Bounds Via Semidefinite Programming”. In: *SIAM Review* 49.1 (2007), pp. 52–64 (cit. on p. 47).

[VMÖ20] G. A. Velasquez, M. E. Mayorga, and O. Y. Özaltın. “Prepositioning Disaster Relief Supplies Using Robust Optimization”. In: *IISE Transactions* 52.10 (2020), pp. 1122–1140. eprint: https://doi.org/10.1080/24725854.2020.1725692 (cit. on p. 73).

[VWS89] F. J. Vasko, F. E. Wolf, and K. L. Stott Jr. “A Set Covering Approach To Metallurgical Grade Assignment”. In: *European Journal of Operational Research* 38.1 (1989), pp. 27–34 (cit. on p. 45).

[Wan+17] Y. Wang, Y. Huang, J. Xu, and N. Barclay. “Optimal recharging scheduling for urban electric buses: A case study in Davis”. In: *Transportation Research Part E: Logistics and Transportation Review* 100 (2017), pp. 115–132 (cit. on p. 84).

[WGW11] Q. Wang, Y. Guan, and J. Wang. “A Chance-Constrained Two-Stage Stochastic Program for Unit Commitment With Uncertain Wind Power Output”. In: *IEEE transactions on power systems* 27.1 (2011), pp. 206–215 (cit. on p. 5).

[WK19] H. Wu and S. Küçükyavuz. “Probabilistic Partial Set Covering With an Oracle for Chance Constraints”. In: *SIAM Journal on Optimization* 29.1 (Jan. 2019), pp. 690–718. ISSN: 1052-6234 (cit. on pp. 48, 49).

[WLM19] S. Wang, J. Li, and S. Mehrotra. “A Solution Approach To Distributionally Robust Chance-Constrained Assignment”. In: *Optimization-Online Preprints* (Aug. 2019) (cit. on p. 48).

[WN14] L. A. Wolsey and G. L. Nemhauser. *Integer and combinatorial optimization*. John Wiley & Sons, 2014 (cit. on p. 67).

[XA16] W. Xie and S. Ahmed. “On Deterministic Reformulations of Distributionally Robust Joint Chance Constrained Optimization Problems”. In: *SIAM Journal on Optimization* 28 (Sept. 2016) (cit. on pp. 8, 47).

[XA20] W. Xie and S. Ahmed. “Bicriteria Approximation of Chance-Constrained Covering Problems”. In: *Operations Research* (Feb. 2020). ISSN: 1526-5463 (cit. on pp. 8, 48).

[XCM12] H. Xu, C. Caramanis, and S. Mannor. “Optimization Under Probabilistic Envelope Constraints”. In: *Operations Research* 60.3 (2012), pp. 682–699 (cit. on p. 47).

[Xie19] W. Xie. “On Distributionally Robust Chance Constrained Programs With Wasserstein Distance”. In: *Mathematical Programming* (2019) (cit. on pp. 8, 11, 12, 32, 48, 50–52, 54, 58, 75, 78).

[Xyl+17] M. Xylia, S. Leduc, P. Patrizio, F. Kraxner, and S. Silveira. “Locating charging infrastructure for electric buses in Stockholm”. In: *Transportation Research Part C: Emerging Technologies* 78 (2017), pp. 183–200 (cit. on pp. 83, 84).

[Yan20] I. Yang. “Wasserstein Distributionally Robust Stochastic Control: a Data-Driven Approach”. In: *IEEE Transactions on Automatic Control* (2020) (cit. on p. 47).

[YL05] J. Yang and J. Y.-T. Leung. “A Generalization of the Weighted Set Covering Problem”. In: *Naval Research Logistics* 52.2 (Mar. 2005), pp. 142–149. ISSN: 1520-6750 (cit. on p. 45).

[YX14] W. Yang and H. Xu. “Distributionally Robust Chance Constraints for Non-Linear Uncertainties”. In: *Mathematical Programming* 155.1-2 (2014), pp. 231–265 (cit. on p. 47).

[YY14] J. Yang and Z. Yang. “Optimal Scheduling of Electric Vehicle Using Dual Decomposition”. In: *The 26th Chinese Control and Decision Conference (2014 CCDC)* (May 2014) (cit. on p. 100).

[ZFM18] N. Zhu, C. Fu, and S. Ma. “Data-driven distributionally robust optimization approach for reliable travel-time-information-gain-oriented traffic sensor location model”. In: *Transportation research part B: methodological* 113 (2018), pp. 91–120 (cit. on p. 85).

[Zha+20] P. Zhang, Y. Liu, G. Yang, and G. Zhang. “A distributionally robust optimization model for designing humanitarian relief network with resource reallocation”. In: *Soft Computing* 24.4 (2020), pp. 2749–2767 (cit. on p. 85).

[ZKR11] S. Zymler, D. Kuhn, and B. Rustem. “Distributionally Robust Joint Chance Constraints With Second-Order Moment Information”. In: *Mathematical Programming* 137.1-2 (2011), pp. 167–198 (cit. on pp. 8, 47).

[ZL10] H. Zhang and P. Li. “Probabilistic Analysis for Optimal Power Flow Under Uncertainty”. In: *Generation, Transmission & Distribution, IET* 4 (June 2010), pp. 553–561 (cit. on p. 47).

[ZZ20] S. Zhao and K. Zhang. “A distributionally robust stochastic optimization-based model predictive control with distributionally robust chance constraints for cooperative adaptive cruise control under uncertain traffic conditions”. In: *Transportation Research Part B: Methodological* 138 (2020), pp. 144–178 (cit. on p. 85).

APPENDIX A

Convex Wasserstein Robust Chance Constraints

A.1 Preliminary Results

We review properties of α-concave functions, α-concave probability measures, as well as elliptical and star-unimodal distributions.

Lemma 10 (Lemma 4.8 in [SDR09])**.** The mapping $\alpha \mapsto m_\alpha(a, b; \theta)$ is nondecreasing and continuous.

The monotonicity of m_α implies that if f is α-concave, then it is β-concave for all $\beta \leq \alpha$. Under certain conditions, summation preserves α-concavity.

Proposition 22 (Theorem 4.19 in [SDR09])**.** If the function $f\colon \mathbb{R}^n \to \mathbb{R}_+$ is α-concave and the function $g\colon \mathbb{R}^n \to \mathbb{R}_+$ is β-concave, where $\alpha, \beta \geq 1$, then $f(x) + g(x)$ is $\min\{\alpha, \beta\}$-concave.

The next two propositions review the relationship between α-concave probability measures and their densities.

Proposition 23 (Theorem 4.15 in [SDR09])**.** Let Ω be a convex subset of $\mathbb{R}^n$ and s be the dimension of the smallest affine subspace $\mathcal{H}(\Omega)$ containing Ω. The probability measure $\mathbb{P}$ is α-concave with $\alpha \leq 1/s$ if and only if its probability density function (PDF) with respect to the Lebesgue measure on $\mathcal{H}$ is α'-concave with

$$\alpha' := \begin{cases} \alpha/(1 - s\alpha) & \text{if } \alpha \in (-\infty, 1/s), \\ -1/s & \text{if } \alpha = -\infty, \\ +\infty & \text{if } \alpha = 1/s. \end{cases}$$

Proposition 24 (Theorem 2 in [Gup80])**.** Let f_0, f_1 be two non-negative Borel-measurable functions on $\mathbb{R}^n$ with non-empty supports S_0 and S_1, respectively. Assume that f_0 and

f_1 are integrable with respect to the Lebesgue measure on $\mathbb{R}^n$. Let $\theta \in (0,1)$ be a fixed number and f be a non-negative, measurable function on $\mathbb{R}^n$ such that

$$f(x) \geq m_\alpha[f_0(x_0), f_1(x_1); \theta],$$

whenever $x = (1-\theta)x_0 + \theta x_1$ with $x_0 \in S_0, x_1 \in S_1$; $-1/n \leq \alpha \leq +\infty$. Then

$$\int_{(1-\theta)S_0+\theta S_1} f(x)\,\mathrm{d}x \geq m_{\alpha_n^*}\left[\int_{S_0} f_0(x)\,\mathrm{d}x, \int_{S_1} f_1(x)\,\mathrm{d}x; \theta\right],$$

where

$$\alpha_n^* := \begin{cases} \alpha/(1+n\alpha) & \text{if } \alpha > -1/n, \\ 1/n & \text{if } \alpha = +\infty, \\ -\infty & \text{if } \alpha = -1/n. \end{cases}$$

We move on to review preliminary results for elliptical distributions. Following the convention in the literature, we write $X \sim \mathcal{E}_n(\mu, \Sigma, \phi)$ if X is elliptically distributed with parameters μ, Σ, ϕ. Specifically, we say a random vector Y is spherically distributed if $Y \sim \mathcal{E}_n(0, I_n, \phi)$, which implies $\Lambda Y \sim \mathcal{E}_n(0, \Lambda\Lambda^\mathsf{T}, \phi)$ (cf. Proposition 1 in [Fra04]).

Proposition 25 (Theorems 1–3 in [CHS81])**.** $X \sim \mathcal{E}_n(\mu, \Sigma, \phi)$ with $\mathbf{rank}(\Sigma) = k$ if and only if

$$X - \mu \overset{d}{=} R\Lambda U_k,$$

where $\overset{d}{=}$ represents being identical in distribution, U_k is a k-dimensional random vector uniformly distributed on the sphere S^{k-1}, R is a nonnegative random variable independent from U_k, $\mu \in \mathbb{R}^n$, and $\Lambda \in \mathbb{R}^{n\times k}$ with $\mathbf{rank}(\Lambda) = k$ and $\Sigma = \Lambda\Lambda^\mathsf{T}$. Furthermore, if R has a density function $f_R(\cdot)$, then the probability density function $f_X(\cdot)$ of X can be written as

$$f_X(x) = C \cdot t^{-\frac{n-1}{2}} f_R\left(\sqrt{(x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu)}\right), \quad \forall x \in \mathbb{R}^n \setminus \{\mu\},$$

where C is a nonnegative constant depending on Λ and k only.

Remark 7. Suppose that $a \in \mathbb{R}^n$ is a column vector and $X \sim \mathcal{E}_n(\mu, \Lambda\Lambda^\mathsf{T}, \phi)$, then by Proposition 25, we have

$$a^\mathsf{T}(X-\mu) \overset{d}{=} R(a^\mathsf{T}\Lambda)U_k = R\|\Lambda^\mathsf{T}a\|_2 \cdot e_a^\mathsf{T}U_k \overset{d}{=} R\|\Lambda^\mathsf{T}a\|_2 \cdot e_1^\mathsf{T}U_k,$$

where $e_a := \left(\frac{\Lambda^\mathsf{T} a}{\|\Lambda^\mathsf{T} a\|_2}\right)$ is the normalized vector of $\Lambda^\mathsf{T} a$, e_1 denotes the first orthonormal basis of $\mathbb{R}^n$, and the last equality is because U_k is invariant to orthogonal transformations. If we further assume that Σ is positive definite (i.e., $k = n$), then $a^\mathsf{T}(X - \mu)/\|\Lambda^\mathsf{T} a\|_2$ is identically distributed as $R \cdot e_1^\mathsf{T} U_n$, which is a one-dimensional elliptical distribution not dependent on a.

Next, we mention results for star-unimodal distributions.

Proposition 26 (Theorem 2.1 in [DJ88])**.** An n-dimensional random vector X is star-unimodal if and only if X is distributed as $V^{1/n} Z$, where V and Z are independent and V is uniformly distributed on $(0, 1)$.

Proposition 27. Suppose that $X \sim \mathcal{E}_n(0, I_n, \phi)$ is spherically distributed and star-unimodal, then $e_1^\mathsf{T} X$ is unimodal.

We are not aware of this result in the literature, and so we provide a proof in the following.

Proof. By Propositions 25 and 26, random vector X admits two representations:

$$R \cdot I_n \cdot U_n \overset{d}{=} X \overset{d}{=} V^{1/n} \cdot Z,$$

where R is a nonnegative random variable independent from U_n, which is uniformly distributed on the sphere S^{n-1}, V is a random variable uniformly distributed on $(0, 1)$ and is independent from Z, an n-dimensional random vector. Because X is spherically distributed, for any two distinct unit vectors $e_i, e_j \in S^{n-1}, e_i \neq e_j$, we have

$$V^{1/n} e_i^\mathsf{T} Z \overset{d}{=} R \cdot e_i^\mathsf{T} U_n \overset{d}{=} R \cdot e_j^\mathsf{T} U_n \overset{d}{=} V^{1/n} e_j^\mathsf{T} Z.$$

First, we show that Z is spherically distributed. Observe that

$$\begin{aligned}
\mathbb{P}\left[e_j^\mathsf{T} V^{1/n} Z \leq t\right] &= \mathbb{E}_V\left[\mathbb{E}_{Z|V}\left[\mathbb{1}\left\{ V^{1/n} e_j^\mathsf{T} Z \leq t \right\} |V\right]\right] = \int_0^1 \mathbb{P}\left[v^{1/n} \cdot e_j^\mathsf{T} Z \leq t\right] \mathrm{d}v \\
&= \int_{+\infty}^{t} \mathbb{P}\left[e_j^\mathsf{T} Z \leq v'\right] \mathrm{d}\left(\frac{t}{v'}\right)^n = nt^n \int_t^{+\infty} \frac{1}{(v')^{n+1}} \mathbb{P}\left[e_j^\mathsf{T} Z \leq v'\right] \mathrm{d}v',
\end{aligned}$$

where the first equality is by the smoothing property, the second equality is because Z and V are independent, and the third equality is by the variable substitution $v' = t/v^{1/n}$. Because $e_j^\mathsf{T} V^{1/n} Z$ is independent from e_j, we can denote its distribution function as $F_Z(t)$

and have

$$(nt^n)^{-1}F_Z(t) = \int_t^{+\infty} \frac{1}{(v')^{n+1}}\mathbb{P}\left[e_j^\mathsf{T} Z \le v'\right]$$
$$\implies \frac{\mathrm{d}}{\mathrm{d}t}\left[(nt^n)^{-1}F_Z(t)\right] = \frac{\mathrm{d}}{\mathrm{d}t}\int_t^{+\infty} \frac{1}{(v')^{n+1}}\mathbb{P}\left[e_j^\mathsf{T} Z \le v'\right] = -\frac{1}{t^{n+1}}\mathbb{P}\left[e_j^\mathsf{T} Z \le t\right], \quad \forall t \neq 0,$$
$$\implies \mathbb{P}\left[e_j^\mathsf{T} Z \le t\right] = -t^{n+1}\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{1}{nt^n}F_Z(T)\right], \quad \forall t \neq 0,$$

from which we observe that the distribution function of $e_j^\mathsf{T} Z$ does not depend on e_j. In other words, $e_j^\mathsf{T} Z \overset{d}{=} e_i^\mathsf{T} Z$ for all $e_i, e_j \in S^{n-1}, e_i \neq e_j$. Let Z_0 represent a random variable identically distributed as $e_1^\mathsf{T} Z$ and define $\phi_0(t) := \mathbb{E}\left[e^{\mathbf{i}\sqrt{t}Z_0}\right]$. We notice that the characteristic function of Z can be represented as

$$\mathbb{E}\left[e^{\mathbf{i}t^\mathsf{T} Z}\right] = \mathbb{E}\left[e^{\mathbf{i}\|t\|_2 e_t^\mathsf{T} Z}\right] = \mathbb{E}\left[e^{\mathbf{i}\|t\|_2 e_1^\mathsf{T} Z}\right] = \mathbb{E}\left[e^{\mathbf{i}\|t\|_2 Z_0}\right] = \phi_0(t^\mathsf{T} t),$$

implying that Z is spherically distributed (clearly, $\mathbb{E}[Z] = 0$). By Proposition 25, there exists a nonnegative random variable R_Z, independent from U_n, such that $Z \overset{d}{=} R \cdot U_n$.

Second, we show that $e_1^\mathsf{T} X$ is unimodal. Let B_n be a random variable uniformly distributed on the n-dimensional unit ball. Then,

$$e_1^\mathsf{T} X \overset{d}{=} e_1^\mathsf{T}\left(V^{1/n} \cdot R_Z \cdot U_n\right) \overset{d}{=} e_1^\mathsf{T}\left(R_Z \cdot (V^{1/n} U_n)\right) \overset{d}{=} e_1^\mathsf{T}\left(R_Z \cdot B_n\right) \overset{d}{=} R_Z \cdot e_1^\mathsf{T} B_n,$$

where the third equality is because $B_n \overset{d}{=} V^{1/n} U_n$ (see Section $3.1.5$ of [FKN18]). Furthermore, because the density function of $e_1^\mathsf{T} B_n$ is monotone increasing on $(-\infty, 0)$ and monotone decreasing on $(0, +\infty)$ (see Section 3.1.5 of [FKN18]), $e_1^\mathsf{T} B_n$ is unimodal. Therefore, by Proposition 26, there exists a random variable Z_B, independent from V, such that $e_j^\mathsf{T} B_n \overset{d}{=} V Z_B$. It follows that

$$e_1^\mathsf{T} X \overset{d}{=} R_Z \cdot V Z_B \overset{d}{=} V \cdot (R_Z Z_B),$$

i.e., $e_1^\mathsf{T} X$ can be represented as the product between V, a uniform distribution on $(0, 1)$, and $R_Z Z_B$, which is independent of V. Therefore, $e_1^\mathsf{T} X$ is unimodal by Proposition 26. □

Finally, we review the (reverse) Minkowski's inequality.

Proposition 28 (Minkowski's Inequality; see Theorem 9 in Chapter 3 of [Bul13])**.** For $p > 1$

and $a_i, b_i \in \mathbb{R}_+$ for all $i \in [n]$, the following holds:

$$\left(\sum_{i=1}^{n}(a_i+b_i)^p\right)^{1/p} \leq \left(\sum_{i=1}^{n}a_i^p\right)^{1/p} + \left(\sum_{i=1}^{n}b_i^p\right)^{1/p}.$$

If $p < 1$ and $p \neq 0$, then the inequality holds with the inequality sign reversed.

An implication of the Minkowski's inequality is as follows.

Lemma 11. If the function $f\colon \mathbb{R}^n \to \mathbb{R}_+$ is an α-concave function with $\alpha \in \overline{\mathbb{R}}$ and $c \in \mathbb{R}_+$ is a constant, then $g(x) := f(x) - c$ is α-concave on $D := \{\, x \in \mathbb{R}^n\colon f(x) > c \,\}$.

Proof. When $\alpha \geq 1$, the result follows from Proposition 22. When $\alpha = 0$, the result was proved in [BBV04] (see Exercise 3.48). When $\alpha = -\infty$, shifting the function along the vertical direction does not affect the convexity of its super level sets. Hence, it suffices to prove the result when $\alpha < 1$ and $\alpha \neq 0$.

We notice that D is convex as it is the super-level set of the quasi-concave function f. Now, for any $x_1, x_2 \in D$ and $\theta \in (0,1)$, the following holds for $x_\theta := \theta x_1 + (1-\theta)x_2$:

$$f(x_\theta) \geq \Big(\theta \cdot (f(x_1))^\alpha + (1-\theta)\cdot(f(x_2))^\alpha\Big)^{1/\alpha}. \tag{A.1}$$

By Minkowski's Inequality with p set to be α, we have

$$\begin{aligned}
&\left(\Big[\theta^{1/\alpha}\cdot f(x_1)\Big]^\alpha + \Big[(1-\theta)^{1/\alpha}\cdot f(x_2)\Big]^\alpha\right)^{1/\alpha} \\
\geq\ &\left(\Big[\theta^{1/\alpha}\cdot (f(x_1)-c)\Big]^\alpha + \Big[(1-\theta)^{1/\alpha}\cdot (f(x_2)-c)\Big]^\alpha\right)^{1/\alpha} \\
&+ \left(\Big[\theta^{1/\alpha}\cdot c\Big]^\alpha + \Big[(1-\theta)^{1/\alpha}\cdot c\Big]^\alpha\right)^{1/\alpha},
\end{aligned}$$

from which we obtain

$$\Big(\theta\cdot(f(x_1)-c)^\alpha + (1-\theta)\cdot(f(x_2)-c)^\alpha\Big)^{1/\alpha} + c \leq \Big(\theta\cdot(f(x_1))^\alpha + (1-\theta)\cdot(f(x_2))^\alpha\Big)^{1/\alpha}. \tag{A.2}$$

Combining (A.1) and (A.2) concludes the proof:

$$\begin{aligned}
f(x_\theta) - c &\geq \Big(\theta\cdot(f(x_1))^\alpha + (1-\theta)\cdot(f(x_2))^\alpha\Big)^{1/\alpha} - c \\
&\geq \Big(\theta\cdot(f(x_1)-c)^\alpha + (1-\theta)\cdot(f(x_2)-c)^\alpha\Big)^{1/\alpha}.
\end{aligned}$$

□

A.2 Proof of Lemma 1

Proof. By definition of CVaR, we have

$$\mathsf{CVaR}_{1-\epsilon}\left(X^{-}\right) = \mathsf{VaR}_{1-\epsilon}\left(X^{-}\right) + \frac{1}{\epsilon}\cdot\mathbb{E}\left[X^{-} - \mathsf{VaR}_{1-\epsilon}(X^{-})\right]^{+}.$$

We discuss two cases:

(i) If $0 < \mathsf{VaR}_{1-\epsilon}(X)$, then $\mathsf{VaR}_{1-\epsilon}(X^{-}) = 0$, from which

$$\mathsf{CVaR}_{1-\epsilon}(X^{-}) = 0 + \frac{1}{\epsilon}\cdot\mathbb{E}\left[X^{-} - 0\right]^{+} = 0.$$

(ii) If $0 \geq \mathsf{VaR}_{1-\epsilon}(X)$, then $\mathsf{VaR}_{1-\epsilon}(X^{-}) = \mathsf{VaR}_{1-\epsilon}(X)$. It follows that

$$\begin{aligned}
\mathbb{E}\left[X^{-} - \mathsf{VaR}_{1-\epsilon}(X)\right]^{+} &= \mathbb{E}\left[\left(X^{-} - \mathsf{VaR}_{1-\epsilon}(X)\right)\cdot\mathbb{1}\left\{\, X^{-} \geq \mathsf{VaR}_{1-\epsilon}(X)\,\right\}\right] \\
&= \mathbb{E}\left[\left(X - \mathsf{VaR}_{1-\epsilon}(X) - X^{+}\right)\cdot\mathbb{1}\left\{\, X \geq \mathsf{VaR}_{1-\epsilon}(X)\,\right\}\right] \\
&= \mathbb{E}\left[\left(X - \mathsf{VaR}_{1-\epsilon}(X)\right)\cdot\mathbb{1}\left\{\, X \geq \mathsf{VaR}_{1-\epsilon}(X)\,\right\}\right] \\
&\quad - \mathbb{E}\left[X^{+}\cdot\mathbb{1}\left\{\, X \geq \mathsf{VaR}_{1-\epsilon}(X)\,\right\}\right] \\
&= \mathbb{E}\left[\left(X - \mathsf{VaR}_{1-\epsilon}(X)\right)\cdot\mathbb{1}\left\{\, X \geq \mathsf{VaR}_{1-\epsilon}(X)\,\right\}\right] - \mathbb{E}\left[X^{+}\right] \\
&= \mathbb{E}\left[X - \mathsf{VaR}_{1-\epsilon}(X)\right]^{+} - \mathbb{E}\left[X^{+}\right],
\end{aligned}$$

where the first equality is by definitions of positive part $[\cdot]^{+}$ and $\mathbb{1}\{\,\cdot\,\}$, the second is due to $\mathsf{VaR}_{1-\epsilon}(X) \leq 0$ and the definitions of positive and negative parts, and the fourth is because $X < \mathsf{VaR}_{1-\epsilon}(X)$ implies $X^{+} = 0$. We conclude the proof by noticing that

$$\begin{aligned}
&\mathsf{VaR}_{1-\epsilon}\left(X^{-}\right) + \frac{1}{\epsilon}\cdot\mathbb{E}\left[X^{-} - \mathsf{VaR}_{1-\epsilon}(X^{-})\right]^{+} \\
&= \mathsf{VaR}_{1-\epsilon}\left(X\right) + \frac{1}{\epsilon}\cdot\mathbb{E}\left[X - \mathsf{VaR}_{1-\epsilon}(X)\right]^{+} - \frac{1}{\epsilon}\cdot\mathbb{E}\left[X^{+}\right] \\
&= \mathsf{CVaR}_{1-\epsilon}(X) - \frac{1}{\epsilon}\cdot\mathbb{E}\left[X^{+}\right].
\end{aligned}$$

□

A.3 Proof of Lemma 2

Proof. We denote the set of points whose distance to $\mathcal{S}^c(x)$ is exactly y by

$$E := \left\{ \zeta \in \Xi \colon \mathbf{d}\left(\zeta, \mathcal{S}^c(x)\right) = y \right\}.$$

We notice that $\mathbf{d}\left(\zeta, \mathcal{S}^c(x)\right) = \mathbf{d}\left(\zeta, \mathbf{cl}\,\mathcal{S}^c(x)\right)$, where $\mathbf{cl}\,\mathcal{S}^c(x)$ denotes the closure of $\mathcal{S}^c(x)$. Then, by the item (1) of [Erd45], we have $\mathbf{Leb}(E) = 0$, which further implies that $\mathbb{P}(E) = 0$ because $\mathbb{P}$ is absolutely continuous with respect to $\mathbf{Leb}(\cdot)$ (see Theorem 2.2 in [Nor93]).

In addition, the Lebesgue measure of the event $\{\zeta \in \Xi : f(x,\zeta) = y\}$ equals zero because $a_i(x) \neq 0$ for all $i \in [m] \setminus I(x)$. It follows that $f(x,\zeta)$ is atomless because $\mathbb{P}$ is absolutely continuous with respect to $\mathbf{Leb}(\cdot)$. □

A.4 A Generalized Theorem 1 For Quasi-Concave Inequalities

We generalize Theorem 1 to quasi-concave inequalities as follows.

Theorem 13. Suppose that the reference distribution $\mathbb{P}$ of $\mathcal{P}$ is α-concave with $\alpha \geq -1$ and $h\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is quasi-concave. Then, for $\delta > 0$ the set

$$\mathcal{X}_{\mathsf{R}}^p = \left\{ x \in \mathbb{R}^n \colon \inf_{\mathbb{Q}\in\mathcal{P}} \mathbb{Q}\left[h(x,\xi) \geq 0\right] \geq 1 - \epsilon \right\}$$

is convex and closed.

Proof. Recall that for $\zeta \in \mathbb{R}^m$, the distance $\mathbf{d}\left(\zeta, \mathcal{S}^c(x)\right)$ to the unsafe set is

$$\begin{aligned}\mathbf{d}\left(\zeta, \mathcal{S}^c(x)\right) &= \inf_{\xi\in\Xi} \left\{ \|\zeta - \xi\| \colon (x,\xi) \notin \mathcal{H}_{\geq 0} \right\} \\ &= \inf_{\xi\in\Xi} \left\{ \|\zeta - \xi\| \colon (x,\xi) \notin \mathbf{cl}\,\left(\mathcal{H}_{\geq 0}\right) \right\},\end{aligned}$$

where $\mathcal{H}_{\geq 0} := \{ (x,\xi) \colon h(x,\xi) \geq 0 \}$ represents the superlevel set of h at level 0. To see he second equality, we first observe that $\mathcal{H}_{\geq 0} \subseteq \mathbf{cl}\,\left(\mathcal{H}_{\geq 0}\right)$ and so $\inf_{\xi\in\Xi} \left\{ \|\zeta - \xi\| \colon (x,\xi) \notin \mathcal{H}_{\geq 0} \right\} \leq \inf_{\xi\in\Xi} \left\{ \|\zeta - \xi\| \colon (x,\xi) \notin \mathbf{cl}\,\left(\mathcal{H}_{\geq 0}\right) \right\}$. Second, for any $\varepsilon > 0$, there exists a $\xi_0 \in \Xi$ such that $(x,\xi_0) \notin \mathcal{H}_{\geq 0}$ and

$$\inf_{\xi\in\Xi} \left\{ \|\zeta - \xi\| \colon (x,\xi) \notin \mathcal{H}_{\geq 0} \right\} + 2\varepsilon > \|\zeta - \xi_0\| + \varepsilon. \tag{A.3}$$

We discuss the following two cases.

(i) If (x, ξ_0) happens to fall out of **cl** $\left(\mathcal{H}_{\geq 0}\right)$ as well, then

$$\|\zeta - \xi_0\| + \varepsilon > \inf_{\xi \in \Xi} \left\{ \|\zeta - \xi\| \colon (x, \xi) \notin \textbf{cl} \left(\mathcal{H}_{\geq 0}\right) \right\}, \tag{A.4}$$

which implies

$$\inf_{\xi \in \Xi} \left\{ \|\zeta - \xi\| \colon (x, \xi) \notin \mathcal{H}_{\geq 0} \right\} + 2\varepsilon > \inf_{\xi \in \Xi} \left\{ \|\zeta - \xi\| \colon (x, \xi) \notin \textbf{cl} \left(\mathcal{H}_{\geq 0}\right) \right\}$$

for all $\varepsilon > 0$. It follows that $\inf_{\xi \in \Xi} \left\{ \|\zeta - \xi\| \colon (x, \xi) \notin \mathcal{H}_{\geq 0} \right\} \geq \inf_{\xi \in \Xi} \left\{ \|\zeta - \xi\| \colon (x, \xi) \notin \textbf{cl} \left(\mathcal{H}_{\geq 0}\right) \right\}$.

(ii) If $(x, \xi_0) \in \textbf{cl} \left(\mathcal{H}_{\geq 0}\right)$, then $(x, \xi_0) \in \textbf{bd} \left(\mathcal{H}_{\geq 0}\right)$. Hence, any open ball centered around (x, ξ_0) has to intersect with $(\mathbb{R}^n \times \Xi) \setminus \textbf{cl} \left(\mathcal{H}_{\geq 0}\right)$. As a result, there exists an $(x, \xi_0') \notin \textbf{cl} \left(\mathcal{H}_{\geq 0}\right)$ such that $\|\xi_0 - \xi_0'\| \leq \varepsilon/2$ and

$$\begin{aligned} \|\zeta - \xi_0\| + \varepsilon = \|\zeta - \xi_0\| + \varepsilon/2 + \varepsilon/2 \geq \|\zeta - \xi_0\| + \|\xi_0 - \xi_0'\| + \varepsilon/2 \geq \|\zeta - \xi_0'\| + \varepsilon/2 \\ > \inf_{\xi \in \Xi} \left\{ \|\zeta - \xi\| \colon (x, \xi) \notin \textbf{cl} \left(\mathcal{H}_{\geq 0}\right) \right\}. \end{aligned} \tag{A.5}$$

Combining (A.3) and (A.5) yields

$$\inf_{\xi \in \Xi} \left\{ \|\zeta - \xi\| \colon (x, \xi) \notin \mathcal{H}_{\geq 0} \right\} + 2\varepsilon > \inf_{\xi \in \Xi} \left\{ \|\zeta - \xi\| \colon (x, \xi) \notin \textbf{cl} \left(\mathcal{H}_{\geq 0}\right) \right\}$$

for all $\varepsilon > 0$. It follows that $\inf_{\xi \in \Xi} \left\{ \|\zeta - \xi\| \colon (x, \xi) \notin \mathcal{H}_{\geq 0} \right\} \geq \inf_{\xi \in \Xi} \left\{ \|\zeta - \xi\| \colon (x, \xi) \notin \textbf{cl} \left(\mathcal{H}_{\geq 0}\right) \right\}$.

Because **cl** $\left(\mathcal{H}_{\geq 0}\right)$ is a closed convex set, it can be represented as the intersection of hyperplanes:

$$\textbf{cl} \left(\mathcal{H}_{\geq 0}\right) = \left\{ (x, \xi) \in \mathbb{R}^n \times \Xi \colon a_i^{\mathsf{T}} \xi \leq b_i(x),\ \forall i \in \mathcal{I} \right\},$$

where $\mathcal{I}$ is a (possibly infinite) index set, and for any $i \in \mathcal{I}$, $a_i \in \mathbb{R}^m$ is independent of x while $b_i(x)$ is an affine function of x. In other words, we have recast $\mathcal{X}_{\mathsf{R}}^p$ as

$$\mathcal{X}_{\mathsf{R}}^p = \left\{ x \in \mathbb{R}^n \colon \inf_{\mathbb{Q} \in \mathcal{P}} \mathbb{Q} \left[a_i^{\mathsf{T}} \xi \leq b_i(x),\ \forall i \in \mathcal{I} \right] \geq 1 - \epsilon \right\},$$

which coincides with the chance constraint with linear inequalities in Theorem 1. Therefore, the conclusion follows from the proof of Theorem 1. □

A.5 Proof of Lemma 3

Proof. We show that the hypograph of $\mathsf{VaR}_{1-\epsilon}\big(f(x,\zeta)\big)$, i.e.,

$$\mathcal{H} := \Big\{\ (x,\theta)\colon \mathsf{VaR}_{1-\epsilon}\big(f(x,\zeta)\big) \geq \theta\ \Big\}$$

is convex. To this end, we note that

$$\mathsf{VaR}_{1-\epsilon}\big(f(x,\zeta)\big) \geq \theta \iff \mathbb{P}\big\{\ f(x,\zeta) \leq \theta\ \big\} \leq 1-\epsilon \iff \mathbb{P}\big\{\ f(x,\zeta) - \theta \geq 0\ \big\} \geq \epsilon$$

where both equivalences are because $f(x,\zeta)$ is atomless. Since $f(x,\zeta) - \theta$ is jointly concave in (x,ζ,θ) and $\mathbb{P}$ is α-concave, $\mathbb{P}\big\{\ f(x,\zeta) - \theta \geq 0\ \big\}$ is α-concave in (x,θ) on the set

$$\mathcal{H}' := \big\{\ (x,\theta)\colon \exists\ \zeta \text{ such that } f(x,\zeta) - \theta \geq 0\ \big\}$$

by Proposition 3. Now, since $\mathcal{H} \subseteq \mathcal{H}'$, $\mathbb{P}\big\{\ f(x,\zeta) - \theta \geq 0\ \big\}$ is also α-concave on $\mathcal{H}$ and $\mathcal{H}$ is convex because it is a super level set of $\mathbb{P}\big\{\ f(x,\zeta) - \theta \geq 0\ \big\}$. □

A.6 Proof of Lemma 4

Proof. For any $(\widehat{x},\widehat{t}) \in \mathbb{R}^n \times \mathbb{R}_+$, consider a sequence $\{\ (x_k,t_k)\ \}_k$ that converges to $(\widehat{x},\widehat{t})$ as k goes to infinity. Then, for any $\zeta \in \Xi$ such that $f(\widehat{x},\zeta) - \widehat{t} \neq 0$, we have

$$\lim_{k\to\infty} \mathbb{1}\big\{\ f(x_k,\zeta) \geq t_k\ \big\} = \mathbb{1}\Big\{\ f(\widehat{x},\zeta) \geq \widehat{t}\ \Big\}$$

because the function $f(x,\zeta) - t$ is continuous in (x,t). Hence, as a function of ζ, $\mathbb{1}\big\{\ f(x_k,\zeta) \geq t_k\ \big\}$ converges pointwise to $\mathbb{1}\Big\{\ f(\widehat{x},\zeta) \geq \widehat{t}\ \Big\}$ on the complement of

$$\mathcal{U}_0 := \{\ \zeta \in \Xi\colon f(\widehat{x},\zeta) = \widehat{t}\ \}\,.$$

It follows that

$$\begin{aligned}\lim_{k\to\infty}\psi(x_k,t_k)+(1-\epsilon) &= \lim_{k\to\infty}\mathbb{P}\left[f(x_k,\zeta)\geq t_k\right]\\ &= \lim_{k\to\infty}\int_{\Xi\setminus\mathcal{U}_0}\mathbb{1}\left\{\,\zeta\colon f(x_k,\zeta)\geq t_k\,\right\}\mathrm{d}\mathbb{P}(\zeta)\\ &= \int_{\Xi\setminus\mathcal{U}_0}\lim_{k\to\infty}\mathbb{1}\left\{\,\zeta\colon f(x_k,\zeta)\geq t_k\,\right\}\mathrm{d}\mathbb{P}(\zeta)\\ &= \int\mathbb{1}\left\{\,\zeta\colon f(\widehat{x},\zeta)\geq\widehat{t}\,\right\}\mathrm{d}\mathbb{P}(\zeta)=\psi(\widehat{x},\widehat{t})+(1-\epsilon),\end{aligned}$$

where the second and fourth equality are because $\mathbf{Leb}(\mathcal{U}_0(x,t))=0$, and the third equality is by the dominated convergence theorem. The continuity of ϕ can be established in a similar way: let $\{\,(x_k,y_k)\,\}_k$ be a sequence such that converges to $(\widehat{x},\widehat{y})$. Then,

$$\begin{aligned}\lim_{k\to\infty}\phi(x_k,y_k) &= \int_{\mathbb{R}_+}\lim_{k\to\infty}\psi(x_k,t)\cdot\mathbb{1}\left\{\,t\leq y_k\,\right\}\mathrm{d}t=\int_{\mathbb{R}_+\setminus\{\,\widehat{y}\,\}}\psi(\widehat{x},t)\cdot\lim_{k\to\infty}\mathbb{1}\left\{\,t\leq y_k\,\right\}\mathrm{d}t\\ &= \int_{\mathbb{R}_+\setminus\{\,\widehat{y}\,\}}\psi(\widehat{x},t)\cdot\mathbb{1}\left\{\,t\leq\widehat{y}\,\right\}\mathrm{d}t=\phi(\widehat{x},\widehat{y}),\end{aligned}$$

where the first equality is by the dominated convergence theorem, and the second equality is because ψ is continuous and $\mathbb{1}\{\,t\leq y_k\,\}$ has a limit as $k\to\infty$ when $t\neq\widehat{y}$. This completes the proof. □

A.7 Proof of Example 5

Proof. Suppose that there exists a worst-case Gaussian distribution $\nu\sim\mathcal{N}(\mu,\sigma)$ for some μ and σ. Then, ν satisfies the following two conditions simultaneously:

(A) $\nu\in\mathcal{P}$, that is,

$$d_W(\nu,\mathbb{P}_0)=\int_{\mathbb{R}}\left|F_{\mathbb{P}_0}(y)-F_\nu(y)\right|\mathrm{d}y\leq\delta, \tag{A.6}$$

where $\mathbb{P}_0$ denotes the 1-dimensional standard Gaussian distribution, and $F_{\mathbb{P}_0}$ and F_ν represent the cumulative distribution functions of $\mathbb{P}_0$ and ν, respectively. The above expression of $d_W(\nu,\mathbb{P}_0)$ is because both $\mathbb{P}_0$ and ν are 1-dimensional distributions.

(B) ν attains the probability bound, i.e., $\nu[\xi\leq x]=\inf_{\mathbb{P}\in\mathcal{P}}\mathbb{P}[\xi\leq x]$. As a result, on the one hand, $\nu[\xi\leq x]\geq 1-\epsilon$ implies $x\geq\mathsf{VaR}_{1-\epsilon}(Y_\nu)=\mathsf{VaR}_{1-\epsilon}(\sigma Y+\mu)=\sigma\mathsf{VaR}_{1-\epsilon}(Y)+\mu$, where Y and Y_ν represent random variables following distributions

$\mathbb{P}_0$ and ν, respectively. On the other hand, by Corollary 3 (Ex) is equivalent to $x \geq c_p := (\overline{g})_\epsilon^{-1}(\delta)$. It follows that

$$c_p = \sigma \mathsf{VaR}_{1-\epsilon}(Y) + \mu. \tag{A.7}$$

In what follows, we show that if ν satisfies condition (B), then it necessarily violates condition (A), establishing the claim. To this end, we simplify $d_W(\nu, \mathbb{P}_0)$:

$$\begin{aligned}
d_W(\nu, \mathbb{P}_0) &= \int_{\mathbb{R}} \left| F_{\mathbb{P}_0}(y) - F_\nu(y) \right| \mathrm{d}y \\
&= \int_0^1 \left| F_{\mathbb{P}_0}(\mathsf{VaR}_t(Y_\nu)) - t \right| \mathrm{d}(\mathsf{VaR}_t(Y_\nu)) = \int_0^1 \left| \int_{-\infty}^{\sigma \mathsf{VaR}_t(Y)+\mu} \mathrm{d}\mathbb{P}_0 - \int_{-\infty}^{\mathsf{VaR}_t(Y)} \mathrm{d}\mathbb{P}_0 \right| \mathrm{d}(\mathsf{VaR}_t(Y_\nu)) \\
&= \int_0^1 \left| \int_{\mathsf{VaR}_t(Y)}^{\sigma \mathsf{VaR}_t(Y)+\mu} \mathrm{d}\mathbb{P}_0 \right| \mathrm{d}(\mathsf{VaR}_t(Y_\nu)) = \sigma \int_{\mathbb{R}} \left| \int_y^{\sigma y+\mu} \mathrm{d}\mathbb{P}_0 \right| \mathrm{d}y \\
&= \sigma \int_{\mathbb{R}} \int_{\mathbb{R}} \mathbb{1}\left\{ y \wedge (\sigma y + \mu) \leq t \leq y \vee (\sigma y + \mu) \right\} \mathrm{d}\mathbb{P}_0(t)\, \mathrm{d}y \\
&= \sigma \int_{\mathbb{R}} \int_{\mathbb{R}} \mathbb{1}\left\{ \sigma y \leq \sigma t \text{ or } (\sigma y + \mu) \leq t \right\} \cdot \mathbb{1}\left\{ \sigma t \leq \sigma y \text{ or } t \leq (\sigma y + \mu) \right\} \mathrm{d}\mathbb{P}_0(t)\, \mathrm{d}y \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} \mathbb{1}\left\{ \sigma y \leq \sigma t \vee (t - \mu) \right\} \cdot \mathbb{1}\left\{ \sigma t \wedge (t - \mu) \leq \sigma y \right\} \mathrm{d}(\sigma y)\, \mathrm{d}\mathbb{P}_0(t) \\
&= \int_{\mathbb{R}} \left| \mu - (1 - \sigma)t \right| \mathrm{d}\mathbb{P}_0(t),
\end{aligned}$$

where the second and the fifth equalities are due to the change of variable $y = \mathsf{VaR}_t(Y_\nu)$. Now, for any $t \in (\mathsf{VaR}_{1-\epsilon}(Y), c_p)$, we have

$$\mu - (1 - \sigma)t = (c_p - t) + \sigma(t - \mathsf{VaR}_{(1-\epsilon)}(Y)) > c_p - t > 0,$$

where the equality uses condition (B), particularly equality (A.7). Then,

$$d_W(\nu, \mathbb{P}_0) = \int_{\mathbb{R}} \left|\mu - (1-\sigma)t\right| \mathrm{d}\mathbb{P}_0(t) \geq \int_{\mathsf{VaR}_{1-\epsilon}(Y)}^{c_p} (\mu - (1-\sigma)t)\, \mathrm{d}\mathbb{P}_0(t)$$
$$> \int_{\mathsf{VaR}_{1-\epsilon}(Y)}^{c_p} (c_p - t)\, \mathrm{d}\mathbb{P}_0(t) = \delta,$$

where the last equality uses the definition of c_p. It follows that ν necessarily violates inequality (A.6) and so condition (A). □

A.8 Proof of A Fact Connecting $\mathcal{X}_{\mathsf{T}}^p$ and $\mathcal{C}_\delta$

Fact. For any $x \neq 0$, $(x, \ell, u) \in \mathcal{X}_{\mathsf{T}}^p$ if and only if $\left(\frac{\ell}{\|x\|_*}, \frac{u}{\|x\|_*}\right) \in \mathcal{C}_\delta$.

Proof. We define a set

$$\mathcal{X}_{\mathsf{T}_0}^p := \left\{ (\ell, u) \in \mathbb{R}^2 \colon \inf_{\mathbb{Q} \in \mathcal{P}_0} \mathbb{Q}\left[\ell \leq \xi \leq u\right] \geq 1 - \epsilon \right\},$$

where $\mathcal{P}_0$ is centered around $\mathbb{P}_0 \overset{d}{=} R \cdot e_1^{\mathsf{T}} U_n$ and has radius δ. By Proposition 2, $\mathcal{X}_{\mathsf{T}_0}^p$ can be recast as

$$\mathbb{P}_0\left[\ell \leq Y \leq u\right] \geq 1 - \epsilon, \tag{A.8}$$

$$\int_0^{\mathsf{VaR}_\epsilon(f_0(\ell,u,Y))} \Big(\mathbb{P}_0\left[f_0(\ell, u, Y) \geq t\right] - (1-\epsilon)\Big)\, \mathrm{d}t \geq \delta, \tag{A.9}$$

where Y has distribution $\mathbb{P}_0$, and $f_0(\ell, u, Y) := \min\{\, Y - \ell, u - Y \,\}$. We simplify constraint (A.9):

$$\int_0^{\mathsf{VaR}_\epsilon(f_0(\ell,u,Y))} \Big(\mathbb{P}_0\left[f_0(\ell, u, Y) \geq t\right] - (1-\epsilon)\Big)\, \mathrm{d}t = \int_0^{+\infty} \Big[\mathbb{P}_0\left[f_0(\ell, u, Y) \geq t\right] - (1-\epsilon)\Big]^+ \mathrm{d}t$$
$$= \int_0^{+\infty} \left[\Phi(u - t) + \Phi(\ell + t) - (1-\epsilon)\right]^+ \mathrm{d}t,$$

where the first equality is because the integrand is decreasing in t and the second equality is by definition of Φ. Because constraint (A.9) implies that there exists a $t \geq 0$ such that $\Phi(u-t) - \Phi(\ell+t) > (1-\epsilon)$, or equivalently, $\mathbb{P}_0\left[\ell + t \leq Y \leq u - t\right] \geq 1 - \epsilon$, we conclude that $\mathcal{X}^p_{\mathsf{T}_0} = \{\, (\ell, u) \in \mathbb{R}^2 \colon \delta \leq g_\epsilon(\ell, u) \,\} = \mathcal{C}_\delta$. Hence, it remains to show that, for any $x \neq 0$, $(x, \ell, u) \in \mathcal{X}^p_{\mathsf{T}}$ if and only if $\left(\frac{\ell}{\|x\|_*}, \frac{u}{\|x\|_*}\right) \in \mathcal{X}^p_{\mathsf{T}_0}$.

To this end, by Proposition 2, $(x, \ell, u) \in \mathcal{X}^p_{\mathsf{T}}$ if and only if it satisfies

$$\mathbb{P}\left[\ell \leq x^{\mathsf{T}}\zeta \leq u\right] \geq 1 - \epsilon, \tag{A.10}$$

$$\int_0^{\mathsf{VaR}_\epsilon(f(\ell,u,\zeta))} \left(\mathbb{P}\left[f(\ell, u, x^{\mathsf{T}}\zeta) \geq t\right] - (1-\epsilon)\right) \mathrm{d}t \geq \delta, \tag{A.11}$$

where $f(\ell, u, x^{\mathsf{T}}\zeta)$ represents the minimum distance to the unsafe set and can be rewritten using f_0 and Y:

$$\begin{aligned} f(\ell, u, \zeta) &:= \frac{\min\{\, x^{\mathsf{T}}\zeta - \ell, u - x^{\mathsf{T}}\zeta \,\}}{\|x\|_*} = \min\left\{ \frac{x^{\mathsf{T}}\zeta}{\|x\|_*} - \frac{\ell}{\|x\|_*}, \frac{u}{\|x\|_*} - \frac{x^{\mathsf{T}}\zeta}{\|x\|_*} \right\} \\ &= f_0\left(\frac{\ell}{\|x\|_*}, \frac{u}{\|x\|_*}, \frac{x^{\mathsf{T}}\zeta}{\|x\|_*}\right) \overset{d}{=} f_0\left(\frac{\ell}{\|x\|_*}, \frac{u}{\|x\|_*}, Y\right). \end{aligned} \tag{A.12}$$

Likewise, we have

$$\mathbb{P}\left[\ell \leq x^{\mathsf{T}}\zeta \leq u\right] = \mathbb{P}\left[\frac{\ell}{\|x\|_*} \leq \frac{x^{\mathsf{T}}\zeta}{\|x\|_*} \leq \frac{u}{\|x\|_*}\right] = \mathbb{P}_0\left[\frac{\ell}{\|x\|_*} \leq Y \leq \frac{u}{\|x\|_*}\right] \geq 1 - \epsilon. \tag{A.13}$$

Now, take $(x, \ell, u) \in \mathcal{X}^p_{\mathsf{T}}$ with $x \neq 0$, then by definition it satisfies (A.10), (A.11), and together with equations (A.12) and (A.13) we have that $(\frac{\ell}{\|x\|_*}, \frac{u}{\|x\|_*}) \in \mathcal{X}^p_{\mathsf{T}_0}$. Similarly, if (x, ℓ, u) satisfies $(\frac{\ell}{\|x\|_*}, \frac{u}{\|x\|_*}) \in \mathcal{X}^p_{\mathsf{T}_0}$, then (A.12), (A.13), (A.8), and (A.9) imply that $(x, \ell, u) \in \mathcal{X}^p_{\mathsf{T}}$. $\square$

A.9 Proof of Corollary 2

Proof. By Proposition 2, $(x,u) \in \mathcal{X}_{\mathsf{TS}}^p$ if and only if

$$\mathbb{P}\left[\left|\frac{x^\mathsf{T}\zeta}{\|x\|_*}\right| \leq \frac{u}{\|x\|_*}\right] \geq 1-\epsilon, \tag{A.14}$$

$$\int_0^{+\infty}\left(\mathbb{P}\left[\left|\frac{x^\mathsf{T}\zeta}{\|x\|_*}\right| \leq \frac{u}{\|x\|_*} - t\right] - (1-\epsilon)\right)^+ \mathrm{d}t \geq \delta. \tag{A.15}$$

Observe that if (x,u) satisfies (A.15), then there exists a $t>0$ such that

$$\mathbb{P}\left[\left|\frac{x^\mathsf{T}\zeta}{\|x\|_*}\right| \leq \frac{u}{\|x\|_*}\right] \geq \mathbb{P}\left[\left|\frac{x^\mathsf{T}\zeta}{\|x\|_*}\right| \leq \frac{u}{\|x\|_*} - t\right] \geq 1-\epsilon.$$

Therefore, (A.14) is redundant and we recast (A.15) as

$$\begin{aligned}\mathcal{X}_{\mathsf{TS}}^p &= \left\{\, (x,u) \in \mathbb{R}^n \times \mathbb{R}_+ \colon g_\epsilon^s\left(\frac{u}{\|x\|_*}\right) \geq \delta \,\right\} \\ &= \left\{\, (x,u) \in \mathbb{R}^n \times \mathbb{R}_+ \colon u \geq \|x\|_* \cdot \inf_r \left\{\, r \geq 0 \colon g_\epsilon^s(r) \geq \delta \,\right\} \,\right\},\end{aligned}$$

where the second equality is because $g_\epsilon^s(\cdot)$ is increasing. □

A.10 Proof of Proposition 4

Proof. We first show the α_1^*-concavity of $\phi(x,y)$ using a similar argument as in the proof of Theorem 1. Recall that $\psi(x,t) = \mathbb{P}[f(x,\zeta) \geq t] - (1-\epsilon)$ and $\phi(x,y) = \int_0^y \psi(x,t)\,\mathrm{d}t$. Pick any $(x_0,y_0),(x_1,y_1) \in \mathbf{dom}\,\phi$, then their midpoint $(x_{1/2},y_{1/2}) := \frac{1}{2}(x_0,y_0) + \frac{1}{2}(x_1,y_1)$ lies in $\mathbf{dom}\,\phi$ because $\mathbf{dom}\,\phi$ is convex by Lemma 3. Define $S_i = [0,y_i]$ and pick any $t_i \in S_i$ for $i=0,1$. Since $\psi(x,t)$ is α-concave by Lemma 11, it holds that

$$\psi(x_{1/2},t_{1/2}) \geq m_\alpha\left[\psi(x_0,t_0),\psi(x_0,t_0);\frac{1}{2}\right].$$

It follows from Proposition 24 that

$$\int_{\frac{1}{2}S_0+\frac{1}{2}S_1} \psi(x_{1/2},t)\,\mathrm{d}t \geq m_{\alpha_1^*}\left[\int_{S_0}\psi(x_0,t)\,\mathrm{d}t, \int_{S_1}\psi(x_1,t)\,\mathrm{d}t; \frac{1}{2}\right],$$

or equivalently, $\phi(x_{1/2}, y_{1/2}) \geq m_{\alpha_1^*}[\phi(x_0, y_0), \phi(x_1, y_1); 1/2]$. This shows the midpoint α_1^*-concavity of $\phi(x, y)$, which together with its continuity (see Lemma 4) shows the α_1^*-concavity.

Second, the closedness of **dom** ϕ follows from the continuity of ψ by Lemma 4.

Third, we show that constraints (2.9b)–(2.9c) are equivalent to (2.9d). To this end, we pick any x that satisfies (2.9b)–(2.9c). Then, by letting $y := \mathsf{VaR}_\epsilon\big(f(x, \zeta)\big) \geq 0$, we obtain $\delta \leq \phi(x, y)$, which implies constraint (2.9d). On the contrary, pick any x that satisfies (2.9d). Then, by definition there exists a $y \geq 0$ such that $\delta \leq \phi(x, y)$. Since $\delta > 0$ and $\phi(x, y) = \int_0^y \big(\mathbb{P}\big[f(x, \zeta) \geq t\big] - (1 - \epsilon)\big)\, \mathrm{d}t$, there exists a $t \in [0, y]$ such that $\mathbb{P}\big[f(x, \zeta) \geq t\big] \geq (1 - \epsilon)$, which implies that $\mathbb{P}\big[f(x, \zeta) \geq 0\big] \geq (1 - \epsilon)$, i.e., constraint (2.9c). Finally, we notice that $\phi(x, y) \leq \phi\big(x, \mathsf{VaR}_\epsilon\big(f(x, \zeta)\big)\big)$ and hence $\delta \leq \phi\big(x, \mathsf{VaR}_\epsilon\big(f(x, \zeta)\big)\big)$, i.e., constraint (2.9b). This completes the proof. □

A.11 Proofs of Preparatory Lemmas 12, 13, and 14

Lemma 12. Let $\{\, (x_k, y_k) \,\}_k$ represent a sequence of iterates produced by Algorithm 1. Then, all iterates are feasible, i.e., $(x_k, y_k) \in S$ for all k. In addition, it holds that

$$\lim_{k \to \infty} \phi(x_k, y_k) = \lim_{k \to \infty} \phi(x_{k+1}, y_k).$$

Proof. First, recall that $S \equiv$ **dom** $\phi \cap \big\{\, (x, y) \in X \times \mathbb{R}_+ : c^\mathsf{T} x \leq u \,\big\}$ is compact. Since $\phi(x, y)$ is continuous by Lemma 4, it is bounded on S. In addition, we notice that by construction the ϕ-values of the iterates produced by Algorithm 1 are non-decreasing, i.e.,

$$0 < \phi(x_1, y_1) \leq \phi(x_2, y_1) \leq \phi(x_2, y_2) \leq \cdots \leq \phi(x_k, y_k) \leq \phi(x_{k+1}, y_k) \leq \cdots \tag{A.16}$$

Hence, this non-decreasing, bounded sequence converges to a finite value. It follows that the two subsequences $\{\, \phi(x_k, y_k) \,\}_k$ and $\{\, \phi(x_{k+1}, y_k) \,\}_k$ converge to the same limit.

Second, we recall that $(x_1, y_1) \in S$ by construction. For all $k \geq 2$, $\phi(x_{k+1}, y_k) > 0$ by (A.16), which implies that there exists a $t \in [0, y_k]$ such that $\mathbb{P}\big[f(x_{k+1}, \zeta) \geq t\big] > 1 - \epsilon$. Then, $\mathbb{P}\big[f(x_{k+1}, \zeta) \geq 0\big] > 1 - \epsilon$, or equivalently, $\mathsf{VaR}_\epsilon\big(f(x_{k+1}, \zeta)\big) > 0$. It follows that $y_{k+1} \equiv \mathsf{VaR}_\epsilon\big(f(x_{k+1}, \zeta)\big) \geq 0$ and so $(x_{k+1}, y_{k+1}) \in S$. This completes the proof. □

Lemma 13. Let (x^*, y^*) represent a limit point of the sequence $\{\, (x_k, y_k) \,\}_k$. Then, it holds that

$$\phi(x^* + d_x, y^*) \leq \phi(x^*, y^*) \quad \text{and} \quad \phi(x^*, y^* + d_y) \leq \phi(x^*, y^*)$$

for all $d_x \in \mathbb{R}^n, d_y \in \mathbb{R}$ such that $(x^* + d_x, y^*) \in S$ and $(x^*, y^* + d_y) \in S$. In addition, if $(x^* + d_x, y^* + d_y) \in S$, then the directional derivative of $\phi(x, y)$ along (d_x, d_y) satisfies

$$\phi'(x^*, y^*; (d_x, d_y)) := \lim_{s \to 0^+} \frac{1}{s}\Big[\phi(x^* + sd_x, y^* + sd_y) - \phi(x^*, y^*)\Big] \leq 0.$$

Proof. We split the proof into three parts: the perturbation along $(0, d_y)$, the perturbation along $(d_x, 0)$, and the directional derivative $\phi'(x^*, y^*; (d_x, d_y))$. For notation brevity, we assume, by passing to a subsequence if needed, that $\{(x_k, y_k)\}_k$ converges to (x^*, y^*).
(Perturbation along $(0, d_y)$) By definition of (x^*, y^*), it holds that

$$\begin{aligned}\Big|y^* - \mathsf{VaR}_\epsilon\big(f(x^*, \zeta)\big)\Big| &= \Big|\lim_{k\to\infty} y_k - \mathsf{VaR}_\epsilon\big(f(\lim_{k\to\infty} x_k, \zeta)\big)\Big| \\ &= \Big|\lim_{k\to\infty} \Big(y_k - \mathsf{VaR}_\epsilon\big(f(x_k, \zeta)\big)\Big)\Big| \\ &= \lim_{k\to\infty} |\varepsilon_k| = 0,\end{aligned}$$

where the second and third equalities are due to the continuity of $\mathsf{VaR}_\epsilon(f(x, \zeta))$ (see Lemma 3) and $|\cdot|$, respectively. Therefore, $\phi(x^*, y^* + d_y) \leq \phi(x^*, y^*)$ because $y^* = \mathsf{VaR}_\epsilon(f(x^*, \zeta))$ is a maximizer of $\phi(x^*, y)$ for fixed x^*.
(Perturbation along $(d_x, 0)$) First, suppose that $(x^* + d_x, y^*)$ lies in the interior of S, denoted by int(S). Then, since $\{(x_k, y_k)\}_k$ converges to (x^*, y^*), there exist neighborhoods $N \subseteq S$ and $N^d \subseteq S$ of (x^*, y^*) and $(x^* + d_x, y^*)$, respectively, such that $(x_k, y_k) \in N$ and $(x_k + d_x, y_k) \in N^d$ for sufficiently large k. Then, by construction it holds that

$$\phi(x_k + d_x, y_k) \leq \max_x \phi(x, y_k) \leq \phi(x_{k+1}, y_k) + \varepsilon_k.$$

Driving k to infinity yields

$$\phi(x^* + d_x, y^*) \leq \phi(x^*, y^*)$$

by continuity of ϕ and Lemma 12.

Second, suppose that $(x^* + d_x, y^*)$ lies on the boundary of S. Then, for all positive integers M, $(x^* + (1 - 1/M)d_x, y^*) \in$ int(S) by convexity of S. It follows that $\phi(x^* + (1 - 1/M)d_x, y^*) \leq \phi(x^*, y^*)$. Driving M to infinity yields $\phi(x^* + d_x, y^*) \leq \phi(x^*, y^*)$ by continuity of ϕ.
(Directional derivative) Since $\phi(x, y)$ is log-concave and $\phi(x^*, y^*) > 0$, ϕ is directionally differentiable at (x^*, y^*) by Lemma 2.4 in [Nor93]. Hence, $\phi'(x^*, y^*; (d_x, d_y))$ is well-defined. To compute $\phi'(x^*, y^*; (d_x, d_y))$, we define $\varphi(x, t) := \mathbb{P}\big[f(x, \zeta) \geq t\big]$ and recast the finite

difference

$$\begin{aligned}
&\phi(x^* + sd_x, y^* + sd_y) - \phi(x^*, y^*) \\
&= \phi(x^* + sd_x, y^* + sd_y) - \phi(x^* + sd_x, y^*) + \phi(x^* + sd_x, y^*) - \phi(x^*, y^*) \\
&= \int_{y^*}^{y^*+sd_y} \Big(\varphi(x^* + sd_x, t) - (1-\epsilon)\Big)\,\mathrm{d}t + \Big(\phi(x^* + sd_x, y^*) - \phi(x^*, y^*)\Big). \qquad \text{(A.17)}
\end{aligned}$$

For the second term in (A.17), we have

$$\lim_{s\to 0^+} \frac{1}{s}\Big[\phi(x^* + sd_x, y^*) - \phi(x^*, y^*)\Big] = \phi'(x^*, y^*; (d_x, 0)) \le 0$$

because $\phi(x^* + sd_x, y^*) \le \phi(x^*, y^*)$ for all sufficiently small $s > 0$. In what follows, we address the first term in (A.17). To that end, we notice that $\varphi(x, t)$ is log-concave on

$$\textbf{dom}\,\varphi := \big\{\ (x,t) \in \mathbb{R} \times \mathbb{R}_+ \colon \exists\ \zeta \text{ such that } f(x,\zeta) - t \ge 0\ \big\},$$

and $(x^*, y^*) \in \text{int}(\textbf{dom}\,\varphi)$ because

$$\mathbb{P}\left[f(x^*, \zeta) - y^* > 0\right] = \mathbb{P}\left[f(x^*, \zeta) - y^* \ge 0\right] \ge 1 - \epsilon,$$

which implies that there exists a $\widehat{\zeta} \in \Xi$ such that $f(x^*, \widehat{\zeta}) - y^* > 0$. By continuity of f, we also have $f(x', \widehat{\zeta}) - y' \ge 0$ for all (x', y') sufficiently close to (x^*, y^*). Since $\varphi(x^*, y^*)$ is strictly positive and $\ln \varphi(x, t)$ is concave on $\textbf{dom}\,\varphi$, $\ln \varphi(x, t)$ is locally Lipschitz at (x^*, y^*), i.e., there exist $M > 0$ and $r > 0$ such that

$$\big|\ln \varphi(x,t) - \ln \varphi(x^*, y^*)\big| \le M\|(x - x^*, t - y^*)\|_2 \qquad \forall (x,t) \in \mathcal{B}\big((x^*, y^*), r\big),$$

where $\mathcal{B}((x^*, y^*), r)$ denotes a Euclidean ball centered around (x^*, y^*) with radius r. For all $s > 0$ sufficiently small such that $s \cdot \|(d_x, d_y)\|_2 \le r/2$ and all scalar t such that $|t - y^*| < s\,|d_y|$, we have

$$\begin{aligned}
\big|\ln \varphi(x^* + sd_x, t) - \ln \varphi(x^*, t)\big| &\le \big|\ln \varphi(x^* + sd_x, t) - \ln \varphi(x^*, y^*)\big| + \big|\ln \varphi(x^*, y^*) - \ln \varphi(x^*, t)\big| \\
&\le M\|(sd_x, t - y^*)\|_2 + M\|(0, t - y^*)\|_2 \\
&\le M\|(sd_x, sd_y)\|_2 + M\|(0, sd_y)\|_2 \\
&\le 2sM\|(d_x, d_y)\|_2,
\end{aligned}$$

where the first inequality is because of the triangle inequality, the second inequality is

because $\ln \varphi_0$ is locally Lipschitz around (x^*, y^*), and the third inequality is because

$$\|(sd_x, t - y^*)\|_2^2 = \|sd_x\|_2^2 + |t - y^*|^2 < \|sd_x\|_2^2 + \left|sd_y\right|^2 = \|(sd_x, sd_y)\|_2^2.$$

We bound the first term in (A.17) by discussing the following two cases. First, if $d_y > 0$, then it holds that

$$\begin{aligned}
&\int_{y^*}^{y^*+sd_y} \Big(\varphi(x^* + sd_x, t) - (1-\epsilon)\Big)\,\mathrm{d}t = \int_{y^*}^{y^*+sd_y} \Big(\exp\big[\ln\varphi(x^* + sd_x, t)\big] - (1-\epsilon)\Big)\,\mathrm{d}t \\
\le\ &\int_{y^*}^{y^*+sd_y} \Big(\exp\Big[\ln\varphi(x^*, t) + 2sM\|(d_x, d_y)\|_2\Big] - (1-\epsilon)\Big)\,\mathrm{d}t \\
=\ &\exp\big[2sM\|(d_x, d_y)\|_2\big] \int_{y^*}^{y^*+sd_y} \Big(\varphi(x^*, t) - (1-\epsilon)\exp\big[-2sM\|(d_x, d_y)\|_2\big]\Big)\,\mathrm{d}t \\
=\ &\exp\big[2sM\|(d_x, d_y)\|_2\big]\cdot \\
&\left(\int_{y^*}^{y^*+sd_y} \big[\varphi(x^*, t) - (1-\epsilon)\big]\,\mathrm{d}t + (1-\epsilon)\big(1 - \exp\big[-2sM\|(d_x, d_y)\|_2\big]\big)sd_y\right).
\end{aligned}$$

It follows that

$$\begin{aligned}
&\lim_{s\to 0^+} \frac{1}{s}\int_{y^*}^{y^*+sd_y} \big(\varphi(x^* + sd_x, t) - (1-\epsilon)\big)\,\mathrm{d}t \\
\le\ &\left(\lim_{s\to 0^+} \exp\big[2sM\|(d_x, d_y)\|_2\big]\right)\cdot \\
&\left(\phi'(x^*, y^*; (0, d_y)) + \lim_{s\to 0^+} \frac{1}{s}(1-\epsilon)\big(1 - \exp\big[-2sM\|(d_x, d_y)\|_2\big]\big)sd_y\right) \\
=\ &\phi'(x^*, y^*; (0, d_y)),
\end{aligned}$$

where the inequality is because

$$\lim_{s\to 0^+} \frac{1}{s}\int_{y^*}^{y^*+sd_y} \big[\varphi(x^*, t) - (1-\epsilon)\big]\,\mathrm{d}t = \lim_{s\to 0^+} \frac{1}{s}\big[\phi(x^*, y^* + sd_y) - \phi(x^*, y^*)\big] = \phi'(x^*, y^*; (0, d_y)).$$

Second, if $d_y < 0$, then it holds that

$$\begin{aligned}
&\int_{y^*}^{y^*+sd_y} \big(\varphi(x^*+sd_x,t)-(1-\epsilon)\big)\,\mathrm{d}t = \int_{y^*+sd_y}^{y^*} \big(-\exp\big[\ln\varphi(x^*+sd_x,t)\big]+(1-\epsilon)\big)\,\mathrm{d}t \\
\le& \int_{y^*+sd_y}^{y^*} \Big(-\exp\Big[\ln\varphi(x^*,t)-2sM\|(d_x,d_y)\|_2\Big]+(1-\epsilon)\Big)\,\mathrm{d}t \\
=& \exp\big[-2sM\|(d_x,d_y)\|_2\big]\int_{y^*+sd_y}^{y^*}\Big(-\varphi(x^*,t)+(1-\epsilon)\exp\big[2sM\|(d_x,d_y)\|_2\big]\Big)\,\mathrm{d}t \\
=& \exp\big[-2sM\|(d_x,d_y)\|_2\big]\int_{y^*}^{y^*+sd_y}\Big(\varphi(x^*,t)-(1-\epsilon)\exp\big[2sM\|(d_x,d_y)\|_2\big]\Big)\,\mathrm{d}t \\
=& \exp\big[-2sM\|(d_x,d_y)\|_2\big]\cdot \\
&\left(\int_{y^*}^{y^*+sd_y}\big[\varphi(x^*,t)-(1-\epsilon)\big]\,\mathrm{d}t+(1-\epsilon)\Big(1-\exp\big[2sM\|(d_x,d_y)\|_2\big]\Big)sd_y\right),
\end{aligned}$$

where the inequality is because $\ln\varphi(x^*+sd_x,t)\ge\ln\varphi(x^*,t)-2sM\|(d_x,d_y)\|_2$ and that the function $-\exp(\cdot)$ is monotonically decreasing. It follows that

$$\begin{aligned}
&\lim_{s\to0^+}\frac{1}{s}\int_{y^*}^{y^*+sd_y}\big(\varphi(x^*+sd_x,t)-(1-\epsilon)\big)\,\mathrm{d}t \\
\le& \left(\lim_{s\to0^+}\exp\big[-2sM\|(d_x,d_y)\|_2\big]\right)\cdot \\
&\left(\phi'(x^*,y^*;(0,d_y))+\lim_{s\to0^+}\frac{1}{s}(1-\epsilon)\big(1-\exp\big[2sM\|(d_x,d_y)\|_2\big]\big)sd_y\right) \\
=& \phi'(x^*,y^*;(0,d_y)).
\end{aligned}$$

Finally, applying the above analysis on both terms in (A.17) yields

$$\begin{aligned}
\phi'(x^*,y^*;(d_x,d_y)) &= \lim_{s\to0^+}\frac{1}{s}\Big[\phi(x^*+sd_x,y^*+sd_y)-\phi(x^*,y^*)\Big] \\
&\le \phi'\big(x^*,y^*;(0,d_y)\big)+\phi'\big(x^*,y^*;(d_x,0)\big)\le 0,
\end{aligned}$$

which completes the proof. □

Lemma 14. For all $(x,y)\in$ **dom** ϕ with $\phi(x,y)>0$, the directional derivative $\phi'(x,y;\Delta)$ at (x,y) along direction Δ is continuous and positively homogeneous in Δ.

Proof. For notation brevity, we denote $z = (x, y)$. Then, it holds that

$$\begin{aligned}\phi'(z;\Delta) &= \lim_{s\to 0^+} \frac{1}{s}\big[\phi(z+s\Delta) - \phi(z)\big] \\ &= \lim_{s\to 0^+} \left\{ \frac{\exp(\ln\phi(z+s\Delta)) - \exp(\ln\phi(z))}{\ln\phi(z+s\Delta) - \ln\phi(z)} \cdot \frac{\ln\phi(z+s\Delta) - \ln\phi(z)}{s} \right\} \\ &= \phi(z) \lim_{s\to 0^+} \frac{\ln\phi(z+s\Delta) - \ln\phi(z)}{s} = \phi(z)\cdot(\ln\phi)'(z;\Delta),\end{aligned}$$

where the third equality follows from the L'Hôpital's rule. Since $(\ln\phi)'(z;\Delta)$ is convex and positively homogeneous in Δ by Proposition 17.2 in [BC+11], so is $\phi'(z;\Delta)$. The continuity of $\phi'(z;\Delta)$ follows from its convexity, which completes the proof. □

A.12 Proof of Theorem 4

We first present preparatory propositions 29, 30, and 31. Then, we put them together to prove Theorem 4.

Definitions 9. For $\widehat{\mathcal{C}}_N \subseteq \mathcal{C}_\delta$, define

$$\underline{\delta} := \sup_{(\ell,u)\in \mathbf{bd}\,(\widehat{\mathcal{C}}_N)} g_\epsilon(\ell, u),$$

and we say $\widehat{\mathcal{C}}_N$ is *supported* by $\mathcal{C}_{\underline{\delta}}$. In addition, with respect to the new origin (ℓ_0, u_0), define $\underline{\delta}^+ := g_\epsilon(\ell_0, u_0) > \underline{\delta}$ and

$$g_\epsilon^s \colon \mathbb{R}^2_+ \ni (\Delta\ell, \Delta u) \mapsto g_\epsilon(\ell_0 + \Delta\ell, u_0 - \Delta u) \in \mathbb{R}_+, \tag{A.18}$$

$$\overline{g}_\epsilon^s \colon \mathbb{R}^2 \ni (\Delta\ell, \Delta u) \mapsto g_\epsilon^s(|\Delta\ell|, |\Delta u|) \in \mathbb{R}_+, \tag{A.19}$$

where g_ϵ^s is the g_ϵ function restricted to $[\ell_0, +\infty) \times (-\infty, u_0]$, and $\overline{g}_\epsilon^s$ extends g_ϵ^s by reflecting it over the two axes.

An immediate implication of the above definition is that $\mathcal{C}_{\underline{\delta}} \subseteq \widehat{\mathcal{C}}_N$.

Remark 8. Under Assumptions 1 and 2, we see that g_ϵ^s inherits the log-concavity of g_ϵ, and so is $\overline{g}_\epsilon^s$. This is because $\overline{g}_\epsilon^s$ is continuous; and for any $(\Delta\ell_1, \Delta u_1), (\Delta\ell_2, \Delta u_2) \in \mathbb{R}^2$ and

their midpoint $(\Delta\ell_{1/2}, \Delta u_{1/2})$, we have

$$\begin{aligned}
\overline{g}^s_\epsilon(\Delta\ell_{1/2}, \Delta u_{1/2}) &= g^s_\epsilon(\left|\Delta\ell_{1/2}\right|, \left|\Delta u_{1/2}\right|) \geq g^s_\epsilon(m_1(\left|\Delta\ell_1\right|, \left|\Delta\ell_2\right|; 1/2), \left|\Delta u_{1/2}\right|) \\
&\geq g^s_\epsilon(m_1(\left|\Delta\ell_1\right|, \left|\Delta\ell_2\right|; 1/2), m_1(\left|\Delta u_1\right|, \left|\Delta u_2\right|; 1/2)) \\
&= g^s_\epsilon(m_1((\left|\Delta\ell_1\right|, \left|\Delta u_1\right|), (\left|\Delta\ell_2\right|, \left|\Delta u_2\right|); 1/2)) \\
&\geq m_0(g^s_\epsilon(\left|\Delta\ell_1\right|, \left|\Delta u_1\right|), g^s_\epsilon(\left|\Delta\ell_2\right|, \left|\Delta u_2\right|); 1/2) \\
&= m_0(\overline{g}^s_\epsilon(\Delta\ell_1, \Delta u_1), \overline{g}^s_\epsilon(\Delta\ell_2, \Delta u_2); 1/2),
\end{aligned}$$

where m_0, m_1 are defined in Definition 1, and the first two inequalities are due to the definition of g_ϵ and the convexity of $|\cdot|$.

Remark 9. $\overline{g}^s_\epsilon$ is radially and strictly decreasing on $\overline{g}^s_\epsilon > 0$. To see this, it suffices to examine g^s_ϵ. Pick any $t_0 > 1$ and any nonzero $(\Delta\ell_0, \Delta u_0) \in \mathbb{R}^2_+$ at which $g^s_\epsilon(\Delta\ell_0, \Delta u_0) > 0$, we compare $g^s_\epsilon(\Delta\ell_0, \Delta u_0)$ and $g^s_\epsilon(t_0\Delta\ell_0, t_0\Delta u_0)$ by discussing the following two cases:

1. $\Delta\ell_0 \neq 0$ and $\Delta u_0 \neq 0$: Define

   $$\underline{t} := \min\left\{ (t_0 - 1)\min\left\{ \Delta\ell_0, \Delta u_0 \right\}, \sup_{t>0}\left\{ \Phi(u_0 - \Delta u_0 - t) - \Phi(\ell_0 + \Delta\ell_0 + t) > (1-\epsilon) \right\} \right\},$$

   which is strictly positive. Then,

$$\begin{aligned}
& g_\epsilon(\ell_0 + \Delta\ell_0, u_0 - \Delta u_0) \\
=& \int_0^{+\infty} \left[\Phi(u_0 - \Delta u_0 - t) - \Phi(\ell_0 + \Delta\ell_0 + t) - (1-\epsilon)\right]^+ \mathrm{d}t \\
=& \int_0^{\underline{t}} \left[\Phi(u_0 - \Delta u_0 - t) - \Phi(\ell_0 + \Delta\ell_0 + t) - (1-\epsilon)\right]^+ \mathrm{d}t \\
&+ \int_0^{+\infty} \left[\Phi(u_0 - (\Delta u_0 + \underline{t}) - t) - \Phi(\ell_0 + (\Delta\ell_0 + \underline{t}) + t) - (1-\epsilon)\right]^+ \mathrm{d}t \\
\geq& \int_0^{\underline{t}} \left[\Phi(u_0 - \Delta u_0 - t) - \Phi(\ell_0 + \Delta\ell_0 + t) - (1-\epsilon)\right]^+ \mathrm{d}t \\
&+ \int_0^{+\infty} \left[\Phi(u_0 - t_0\Delta u_0 - t) - \Phi(\ell_0 + t_0\Delta\ell_0 + t) - (1-\epsilon)\right]^+ \mathrm{d}t \\
>& \; g_\epsilon(\ell_0 + t_0\Delta\ell_0, u_0 - t_0\Delta u_0),
\end{aligned}$$

where the second equality is due to variable substitution, the first inequality is by the definition of $\underline{t}$, and the last inequality is because $\underline{t} > 0$ and the integrand of the first term in the summation is strictly positive on $(0, t)$.

2. One of $\Delta\ell_0$ and Δu_0 is 0: Since g_ϵ is symmetric, we can assume $\Delta\ell_0 = 0$ and $\Delta u_0 \neq 0$ without loss of generality. Then, by the continuity of g_ϵ and the argument in the previous case, we have

$$\begin{aligned} g_\epsilon(\ell_0, u_0 - \Delta u_0) = \lim_{n\to\infty} g_\epsilon(\ell_0 + \frac{1}{n}, u_0 - \Delta u_0) > \lim_{n\to\infty} g_\epsilon(\ell_0 + \frac{t_0}{n}, u_0 - t_0\Delta u_0) \\ = g_\epsilon(\ell_0, u_0 - t_0\Delta u_0). \end{aligned}$$

The next proposition relates $\overline{g}_\epsilon$ with **bd** $(\mathcal{C}_\delta)$.

Proposition 29. Suppose that $\epsilon \in (0, 1/2)$ and $\delta > 0$. Then, for a sequence of points $\{\, (\ell_n, u_n), n \in \mathbb{N} \,\} \subseteq$ **bd** $(\mathcal{C}_\delta)$, if $\ell_n \searrow -\infty$ as $n \to \infty$, then $u_n \to u^*$ as $n \to \infty$, where u^* is the solution of the equation $\overline{g}_\epsilon(u) = \delta$.

Proof. Since $\ell_n \searrow -\infty$ and $(\ell_n, u_n) \in$ **bd** $(\mathcal{C}_\delta)$, u_n is decreasing in n. Consider the sequence of functions $\{\, g_\epsilon^n, n \in \mathbb{N} \,\}$, where

$$g_\epsilon^n(u) := \int_0^{+\infty} \big(\Phi(u - t) - \Phi(\ell_n + t) - (1 - \epsilon)\big)^+ \,\mathrm{d}t.$$

Evidently, the sequence $\{\, g_\epsilon^n \,\}_{n=1}^\infty$ is increasing, bounded from above by $\overline{g}_\epsilon$, and continuous for all n by the dominated convergence theorem. Take a $\underline{u} > 0$ such that $\overline{g}_\epsilon(\underline{u}) < \delta$ and define a restricted domain $\mathbf{dom}_g := [\underline{u}, u_1]$ for all (g_ϵ^n)'s and $\overline{g}_\epsilon$. Since $g_\epsilon^n(\underline{u}) \leq \overline{g}_\epsilon(\underline{u}) < \delta$ for all n and $g_\epsilon^n(u_1) \geq g_\epsilon^1(u_1) = \delta$, the solution to the equations $\{\, u \colon g_\epsilon^n(u) = \delta \,\} \subseteq \mathbf{dom}_g$ by the intermediate value theorem. First, we show that $g_\epsilon^n \to \overline{g}_\epsilon$ uniformly as $n \to \infty$ on $\mathbf{dom}_g$. Notice that

$$\begin{aligned} \big|g_\epsilon^n(u) - \overline{g}_\epsilon(u)\big| &\leq \int_0^{+\infty} \Big|\big(\Phi(u - t) - \Phi(\ell_n + t) - (1 - \epsilon)\big)^+ - \big(\Phi(u - t) - (1 - \epsilon)\big)^+\Big| \,\mathrm{d}t \\ &\leq \int_0^{+\infty} \Phi(\ell_n + t) \cdot \mathbb{1}\left\{\, \Phi(u - t) \geq (1 - \epsilon) \,\right\} \mathrm{d}t \\ &= \int_0^{+\infty} \Phi(\ell_n + t) \cdot \mathbb{1}\left\{\, t \leq u_1 - \Phi^{-1}(1 - \epsilon) \,\right\} \mathrm{d}t. \end{aligned}$$

For any $u \in \mathbf{dom}_g$, the dominated convergence theorem implies that

$$\lim_{n\to\infty} \left| g_\epsilon^n(u) - \overline{g}_\epsilon(u) \right| \leq \int_0^{+\infty} \lim_{n\to\infty} \Phi(\ell_n + t) \cdot \mathbb{1}\left\{ t \leq u_1 - \Phi^{-1}(1-\epsilon) \right\} \mathrm{d}t = 0.$$

Due to the strict monotonicity of $\overline{g}_\epsilon$ in u, its inverse function $(\overline{g}_\epsilon)^{-1}$ is well-defined. Furthermore, it is continuous because $\mathbf{dom}_g$ is compact. Second, we bound the distance between u_n and u_*. For any $\varepsilon > 0$, there exists an $N_\varepsilon \in \mathbb{N}$ such that

$$n > N_\varepsilon \implies \sup_{u\in\mathbf{dom}_g} \left| g_\epsilon^n(u) - \overline{g}_\epsilon(u) \right| < \varepsilon.$$

Let u_n^* be the solution of $g_\epsilon^n(u) = \delta$, then for all $n > N_\varepsilon$,

$$u_* \leq u_n^* \leq (\overline{g}_\epsilon)^{-1}(\delta + \varepsilon),$$

where the first inequality is because g_ϵ^n is monotone and $g_\epsilon^n(u_n^*) = \delta = \overline{g}_\epsilon(u^*) \geq g_\epsilon^n(u^*)$, and the second inequality is because $(\overline{g}_\epsilon)^{-1}$ is monotone and $\overline{g}_\epsilon(u_n^*) \leq g_\epsilon^n(u_n^*) + \varepsilon$. We complete the proof by noting that

$$\inf_{\varepsilon>0} \sup_{n\geq N_\varepsilon} |u_n^* - u^*| \leq \inf_{\varepsilon>0} \left((\overline{g}_\epsilon)^{-1}(\delta+\varepsilon) - u^* \right) = 0,$$

where the last equality is because $(\overline{g}_\epsilon)^{-1}$ is continuous. □

The next proposition characterizes the level sets of $\overline{g}_\epsilon^s$ through polar coordinates.

Proposition 30. Suppose that $\epsilon \in (0, 1/2)$, $\delta > 0$, $\mathbb{P}_0$ is unimodal with CDF Φ and density function Φ', and $\overline{g}_\epsilon^s$ is defined in (A.19). Let $\mathcal{K}_\delta$ be its δ-superlevel set, then it holds that

$$\mathbf{bd}\left(\mathcal{K}_\delta\right) = \left\{ (\rho\cos\theta, \rho\sin\theta) \in \mathbb{R}^2 \colon \rho \|(\cos\theta, \sin\theta)\|_{\mathcal{C}_\delta} = 1 \right\}, \tag{A.20}$$

where $\|x\|_{\mathcal{K}_\delta} = \inf\left\{ r > 0 \colon x \in r \cdot \mathcal{K}_\delta \right\}$ is the gauge induced by $\mathcal{K}_\delta$. Furthermore, for $0 < \delta_1 < \delta_2 < \underline{\delta}^+$, it holds that

$$\mathcal{K}_{\delta_2} \subseteq \mathcal{K}_{\delta_1} \subseteq \left(1 + \frac{\sqrt{2}\ln\left(\delta_2/\delta_1\right)}{\underline{D} \cdot \underline{\rho}} \right) \mathcal{K}_{\delta_2}, \tag{A.21}$$

where $\underline{D} = \dfrac{\Phi'(u_0)}{(\Phi(u_0) - \Phi(\ell_0) - (1-\epsilon))}$ and $\underline{\rho} = \|(1,0)\|_{\mathcal{C}_{\delta_2}}^{-1}$.

Proof. Remark 9 shows that $\overline{g}^s_\epsilon$ is strictly decreasing along each radial direction. Then,

$$\begin{aligned}\textbf{bd}\left(\mathcal{K}_\delta\right) &= \left\{ (\rho\cos\theta, \rho\sin\theta) \in \mathbb{R}^2 \colon \rho = \sup_{u>0} \left\{ u \colon (u\cos\theta, u\sin\theta) \in \mathcal{K}_\delta \right\} \right\} \\ &= \left\{ (\rho\cos\theta, \rho\sin\theta) \in \mathbb{R}^2 \colon \rho = \sup_{u'>0} \left\{ 1/u' \colon (1/u'\cos\theta, 1/u'\sin\theta) \in \mathcal{K}_\delta \right\} \right\} \\ &= \left\{ (\rho\cos\theta, \rho\sin\theta) \in \mathbb{R}^2 \colon \rho = \left(\|(\cos\theta, \sin\theta)\|_{\mathcal{K}_\delta} \right)^{-1} \right\},\end{aligned}$$

where the second equality is obtained by the change of variable $u' \leftarrow 1/u$.

The first inclusion in (A.21) follows from $\delta_1 < \delta_2$. To prove the second inclusion, we pick an arbitrary $\theta \in [0, \pi/4)$ and focus on the restriction of $\ln \overline{g}^s_\epsilon(\cdot,\cdot)$'s hypograph to direction θ:

$$\mathcal{H}_\theta := \left\{ (\rho, \delta_{\ln}) \in \mathbb{R}_+ \times \mathbb{R} \colon \ln \overline{g}^s_\epsilon(\rho\cos\theta, \rho\sin\theta) \geq \delta_{\ln} \right\}.$$

Because $\overline{g}^s_\epsilon(\cdot,\cdot)$ is log-concave, $\ln \overline{g}^s_\epsilon(\cdot,\cdot)$ is a concave function, and $\mathcal{H}_\theta$ is a convex set which can be approximated from above by hyperplanes. In particular,

$$(\rho_i, \ln(\delta_i)) \in \textbf{bd}\left(\mathcal{H}_\theta\right), \text{ where } \rho_i := \left(\|(\cos\theta, \sin\theta)\|_{\mathcal{K}_{\delta_i}} \right)^{-1}, \forall i \in [2],$$

because clearly $(\rho_i, \ln(\delta_i))$ and for any $\varepsilon > 0$, $(\rho_i + \varepsilon/2, \ln(\delta_i)) \notin \mathcal{H}_\theta$ for $i \in [2]$. Therefore, the supporting hyperplane at $(\rho_2, \ln(\delta_2))$ is an upper bound of $\ln \overline{g}^s_\epsilon(\rho\cos\theta, \rho\sin\theta)$:

$$\ln \overline{g}^s_\epsilon(\rho\cos\theta, \rho\sin\theta) \leq \frac{\mathrm{d}}{\mathrm{d}\rho}(\ln \overline{g}^s_\epsilon(\rho\cos\theta, \rho\sin\theta))\Big|_{\rho=\rho_2} (\rho - \rho_2) + \ln(\delta_2) =: \widehat{g}^s_{\epsilon,\theta,\delta_2}(\rho),$$

and the superlevel sets of $\widehat{g}^s_{\epsilon,\theta,\delta_2}$ are supersets of those of $\ln \overline{g}^s_\epsilon(\rho\cos\theta, \rho\sin\theta)$. In particular,

$$[0, \rho_1] \subseteq [0, \overline{\rho}_1], \text{ where } \overline{\rho}_1 \in \mathbb{R}_+ \text{ is such that } \widehat{g}^s_{\epsilon,\theta,\delta_2}(\overline{\rho}_1) = \ln(\delta_1).$$

Solving for $\overline{\rho}_1$, we obtain that for any $\theta \in [0, 2\pi)$,

$$\overline{\rho}_1 = \left(1 + \frac{\ln(\delta_2/\delta_1)}{-\frac{\mathrm{d}}{\mathrm{d}\rho}(\ln \overline{g}^s_\epsilon(\rho\cos\theta, \rho\sin\theta))\Big|_{\rho=\rho_2} \cdot \rho_2} \right) \rho_2.$$

Next, we seek an upper bound of $\overline{\rho}_1$ that is independent from θ. To this end, we analyze

the derivative in the denominator:

$$\frac{\mathrm{d}}{\mathrm{d}\rho}\ln \overline{g}^s_\epsilon(\rho\cos\theta,\rho\sin\theta) = \frac{1}{\overline{g}^s_\epsilon(\rho\cos\theta,\rho\sin\theta)}\cdot\frac{\mathrm{d}}{\mathrm{d}\rho}\overline{g}^s_\epsilon(\rho\cos\theta,\rho\sin\theta)$$

where

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\rho}\overline{g}^s_\epsilon(\rho\cos\theta,\rho\sin\theta) &= \frac{\mathrm{d}}{\mathrm{d}\rho}\int_0^{+\infty}\left[\Phi(u_0-\rho\cos\theta-t)-\Phi(\ell_0+\rho\sin\theta+t)-(1-\epsilon)\right]^+\mathrm{d}t\\
&= \int_0^{\bar{t}_{\rho,\theta}}\frac{\mathrm{d}}{\mathrm{d}\rho}\left(\Phi(u_0-\rho\cos\theta-t)-\Phi(\ell_0+\rho\sin\theta+t)\right)\mathrm{d}t\\
&= \int_0^{\bar{t}_{\rho,\theta}}\left(\Phi'(u_0-\rho\cos\theta-t)(-\cos\theta)-\Phi'(\ell_0+\rho\sin\theta+t)\sin\theta\right)\mathrm{d}t\\
&\le \int_0^{\bar{t}_{\rho,\theta}}\left(-\Phi'(u_0)(\cos\theta+\sin\theta)\right)\mathrm{d}t = \int_0^{\bar{t}_{\rho,\theta}}-\Phi'(u_0)\sqrt{2}\sin(\theta+\frac{\pi}{4})\,\mathrm{d}t\\
&\le -\Phi'(u_0)\bar{t}_{\rho,\theta},
\end{aligned}$$

where the second equality is by Leibniz integration rule, and $\bar{t}_{\rho,\theta}$ is defined through

$$\bar{t}_{\rho,\theta} := \max\left\{\, t\ge 0\colon \Phi(u_0-\rho\cos\theta-t)-\Phi(\ell_0+\rho\sin\theta+t)\ge 1-\epsilon\,\right\}.$$

We notice that $\bar{t}_{\rho,\theta}$ satisfies

$$\ell_0+\rho\cos\theta+\bar{t}_{\rho,\theta}<0<u_0-\rho\cos\theta-\bar{t}_{\rho,\theta}.$$

The first inequality is because $\theta\in[0,\pi/4)$ and Φ' is symmetric, increasing on $(-\infty,0]$, and decreasing on $[0,\infty)$. The second inequality is because $\sin(\theta+\frac{\pi}{4})$ achieves its minimum at 0 on $[0,\frac{\pi}{4})$. In order to remove the dependence of $\bar{t}_{\rho,\theta}$ on θ, we seek a lower bound for $\bar{t}_{\rho,\theta}$. To this end, by the mean value theorem there exists a $t_0\in[0,\bar{t}_{\rho,\theta}]$ such that

$$\begin{aligned}
\delta &= \int_0^{\bar{t}_{\rho,\theta}}\left[\Phi(u_0-\rho\cos\theta-t)-\Phi(\ell_0+\rho\sin\theta+t)-(1-\epsilon)\right]\mathrm{d}t\\
&= \bar{t}_{\rho,\theta}\left[\Phi(u_0-\rho\cos\theta-t_0)-\Phi(\ell_0+\rho\sin\theta+t_0)-(1-\epsilon)\right],
\end{aligned}$$

and by

$$\Phi(u_0) - \Phi(\ell_0) \geq \Phi(u_0 - \rho\cos\theta - t_0) - \Phi(\ell_0 + \rho\sin\theta + t_0),$$

we obtain a lower bound for $\bar{t}_{\rho,\theta}$,

$$\underline{t}_\delta := \frac{\delta}{(\Phi(u_0) - \Phi(\ell_0) - (1-\epsilon))} \leq \bar{t}_{\rho,\theta},$$

from which

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}\rho}\left(\ln \overline{g}^s_\epsilon(\rho\cos\theta, \rho\sin\theta)\right)\Big|_{\rho=\rho_2} &= \frac{-1}{\overline{g}^s_\epsilon(\rho_2\cos\theta, \rho_2\sin\theta)} \cdot \frac{\mathrm{d}}{\mathrm{d}\rho}\overline{g}^s_\epsilon(\rho\cos\theta, \rho\sin\theta) \\ &\geq \frac{\Phi'(u_0)\bar{t}_{\rho,\theta}}{\overline{g}^s_\epsilon(\rho_2\cos\theta, \rho_2\sin\theta)} \geq \frac{\Phi'(u_0)\underline{t}_{\delta_2}}{\overline{g}^s_\epsilon(\rho_2\cos\theta, \rho_2\sin\theta)} \\ &= \frac{\Phi'(u_0)}{(\Phi(u_0) - \Phi(\ell_0) - (1-\epsilon))} = \underline{D}. \end{aligned}$$

Thus, for any $(\rho\cos\theta, \rho\sin\theta) \in \mathcal{K}_{\delta_1}$, we have

$$\begin{aligned} 0 \leq \rho \leq \left(\|(\cos\theta, \sin\theta)\|_{\mathcal{K}_{\delta_1}}\right)^{-1} = \rho_1 \leq \overline{\rho}_1 &= \left(1 + \frac{\ln(\delta_2/\delta_1)}{-\frac{\mathrm{d}}{\mathrm{d}\rho}(\ln \overline{g}^s_\rho(\rho\cos\theta, \rho\sin\theta))\Big|_{\rho=\rho_2} \cdot \rho_2}\right)\rho_2 \\ &\leq \left(1 + \frac{\ln(\delta_2/\delta_1)}{\underline{D}\cdot\rho_2}\right)\rho_2 \leq \left(1 + \frac{\sqrt{2}\ln(\delta_2/\delta_1)}{\underline{D}\cdot\underline{\rho}}\right)\rho_2, \end{aligned}$$

where the second inequality is by definition of $\underline{D}$. We justify the last inequality as follows. Define $k := \dfrac{\underline{\rho}}{\rho_2(\cos\theta + \sin\theta)} \geq 0$. Since $\underline{\rho}(1,0) \in \textbf{bd}\,(\mathcal{K}_{\delta_2})$, we have

$$\begin{gathered} \delta_2 = \overline{g}^s_\epsilon((\underline{\rho}, 0)) = g_\epsilon(\ell_0 + \underline{\rho}, u_0 - 0) = g_\epsilon(-u_0, -\ell_0 - \underline{\rho}) = g_\epsilon(\ell_0, u_0 - \underline{\rho}) = \overline{g}^s_\epsilon(0, \underline{\rho}) \\ \implies \underline{\rho}(0,1) \in \textbf{bd}\,(\mathcal{K}_{\delta_2}) \\ k\cdot\rho_2(\cos\theta, \sin\theta) = \frac{\underline{\rho}(\cos\theta, \sin\theta)}{(\cos\theta + \sin\theta)} = \frac{\cos\theta}{\cos\theta + \sin\theta}\cdot\underline{\rho}(1,0) + \frac{\sin\theta}{\cos\theta + \sin\theta}\cdot\underline{\rho}(0,1) \implies k \leq 1, \end{gathered}$$

where the second implication is because $k\cdot\rho_2(\cos\theta, \sin\theta)$ is the convex combination of $\underline{\rho}(1,0)$ and $\underline{\rho}(0,1)$ and is parallel to $\rho_2(\cos\theta, \sin\theta)$. Therefore, we have

$$\rho_2 \geq k\cdot\rho_2 = \|\frac{\underline{\rho}(\cos\theta, \sin\theta)}{\sqrt{2}\sin(\theta + \frac{\pi}{4})}\|_2 \geq \frac{\underline{\rho}}{\sqrt{2}}.$$

Since the upper bound above is independent from θ, it holds that

$$\frac{\rho}{\left(1+\frac{\ln \delta_2/\delta_1}{\underline{D}\cdot\underline{\rho}}\right)} \le \rho_2 = \sup_{u>0}\{\, u\colon (u\cos\theta, u\sin\theta)\in\mathcal{K}_{\delta_2}\,\}$$
$$\iff \frac{1}{\left(1+\frac{\ln(\delta_2/\delta_1)}{\underline{D}\cdot\underline{\rho}}\right)}(\rho\cos\theta,\rho\sin\theta)\in\mathcal{K}_{\delta_2},$$

and (A.21) follows. □

Proposition 31. Suppose that $\epsilon\in(0,1/2)$ and $\mathbb{P}_0$ is unimodal with CDF Φ and density function Φ'. For any δ_1,δ_2 such that $0<\delta_1<\delta_2<\underline{\delta}^+$, denote by $g_\epsilon|_{\ell_0}(u)$ the restriction of the function g_ϵ to the vertical line $\ell=\ell_0$, and by $g_\epsilon^{-1}|_{\ell_0}(\cdot)$ its inverse (whose existence is guaranteed by Remark 9):

$$g_\epsilon|_{\ell_0}(u) := g_\epsilon(\ell_0,u),\quad g_\epsilon^{-1}|_{\ell_0}(\delta) := \inf\{\, u>0\colon g_\epsilon(\ell_0,u)\ge\delta\,\}.$$

In addition, define γ_{13},γ_4 as

$$\gamma_{13} := \frac{u_0-(\overline{g}_\epsilon)^{-1}(\delta_1)}{u_0-g_\epsilon^{-1}|_{\ell_0}(\delta_2)},\quad \gamma_4 := 1+\frac{\sqrt{2}\ln(\delta_2/\delta_1)}{\underline{D}\cdot\underline{\rho}},$$

where $\underline{D}$ and $\underline{\rho}$ are defined in Proposition 30. Then, it holds that

(1) $\mathcal{C}_{\delta_1}\cap\mathcal{O}_2=\mathcal{C}_{\delta_2}\cap\mathcal{O}_2$.

(2) $\mathcal{C}_{\delta_1}\cap\mathcal{O}_i\subseteq\gamma_{13}\cdot(\mathcal{C}_{\delta_2}\cap\mathcal{O}_i)$ for any $i\in\{\,1,3\,\}$.

(3) $\mathcal{C}_{\delta_1}\cap\mathcal{O}_4\subseteq\gamma_4\cdot\mathcal{C}_{\delta_2}\cap\mathcal{O}_4$.

Proof. Because $g_\epsilon(\ell_0,u_0)=\underline{\delta}^+>\delta_2>\delta_1>0$, we have $(\ell_0,u_0)\in\mathcal{C}_{\delta_2}$, from which $\mathcal{O}_2\cap\mathcal{C}_{\delta_2}=\mathcal{O}_2=\mathcal{O}_2\cap\mathcal{C}_{\delta_1}$, i.e., item (1) holds.

Because $\mathcal{C}_\delta$ is symmetric with respect to $u+\ell=0$, it is sufficient to prove the case $i=3$ in item (2). To this end, we notice that $(\ell_0,g_\epsilon^{-1}|_{\ell_0}(\delta_2))\in\mathcal{C}_{\delta_2}$, and so $(-\infty,\ell_0]\times[g_\epsilon^{-1}|_{\ell_0}(\delta_2),u_0]\subseteq\mathcal{C}_{\delta_2}\cap\mathcal{O}_3$ because $(-1,0)$ and $(0,1)$ are extreme rays of $\mathcal{C}_\delta$ for all $\delta>0$. In addition,

$$\begin{aligned}\inf_{\ell,u}\{\, u\colon(\ell,u)\in\mathcal{C}_{\delta_1}\cap\mathcal{O}_3\,\} &= \inf_{\ell,u}\{\, u\colon g_\epsilon(\ell,u)\ge\delta_1,(\ell,u)\in(-\infty,\ell_0]\times(0,u_0]\,\}\\ &= \inf_{\ell}\left\{\inf_{0<u\le u_0}\{\, u\colon g_\epsilon(\ell,u)\ge\delta_1\,\}:\ell\in(-\infty,\ell_0]\right\}\\ &= \lim_{n\to\infty}\inf_{0<u\le u_0}\{\, u\colon g_\epsilon(\ell_0-n,u)\ge\delta_1\,\}\\ &= \inf_u\{\, u\colon\overline{g}_\epsilon(u)\ge\delta_1\,\} = (\overline{g}_\epsilon)^{-1}(\delta_1).\end{aligned}$$

where the third equality is because the sequence of functions $\{ g_\epsilon(\ell_0 - n, \cdot) \}_{n=1}^{+\infty}$ is increasing in n, i.e.

$$\cdots \leq g_\epsilon(\ell_0 - n, \cdot) \leq g_\epsilon(\ell_0 - (n-1), \cdot) \leq \cdots \leq g_\epsilon(\ell_0 - 1, \cdot) \leq g_\epsilon(\ell_0, \cdot)$$

and it follows that $\left\{ \inf_{0<u\leq u_0} \{ u \colon g_\epsilon(\ell_0 - n, u) \geq \delta \} \right\}_{n=1}^{+\infty}$ is a decreasing sequence. The fourth equality is because of Proposition 29. Now, item (2) follows from

$$\begin{aligned}
\gamma_{13} \cdot (\mathcal{C}_{\delta_2} \cap \mathcal{O}_3) &\supseteq \gamma_{13} \cdot (-\infty, \ell_0] \times [g_\epsilon^{-1}|_{\ell_0}(\delta_2), u_0] \\
&= (-\infty, \ell_0] \times [u_0 - \gamma_{13} \cdot (u_0 - g_\epsilon^{-1}|_{\ell_0}(\delta_2)), u_0] \\
&= (-\infty, \ell_0] \times [(\overline{g}_\epsilon)^{-1}(\delta_1), u_0] \supseteq \mathcal{C}_{\delta_1} \cap \mathcal{O}_3.
\end{aligned}$$

Finally, since $\mathcal{C}_{\delta_i} \cap \mathcal{O}_4 = \mathcal{K}_{\delta_i} \cap \mathcal{O}_4$ for $i \in \{ 1, 2 \}$, item (3) follows from Proposition 30. □

Now we are ready to prove Theorem 4.

Proof of Theorem 4. Similar to the proof of Proposition 31, we partition $\widehat{\mathcal{C}}_N$ and $\mathcal{C}_\delta$ into four subsets to obtain:

(1) $\mathcal{C}_\delta \cap \mathcal{O}_2 = \widehat{\mathcal{C}}_N \cap \mathcal{O}_2$.

(2) $\mathcal{C}_\delta \cap \mathcal{O}_i \subseteq \frac{u_0 - (\overline{g}_\epsilon)^{-1}(\delta)}{u_0 - (\overline{g}_\epsilon)^{-1}(\delta+\tau)} \cdot (\widehat{\mathcal{C}}_N \cap \mathcal{O}_i), \forall i \in \{ 1, 3 \}$.

(3) $\mathcal{C}_\delta \cap \mathcal{O}_4 \subseteq \left(1 + \frac{\sqrt{2} \ln((\delta+\tau)/\delta)}{\underline{D} \cdot \underline{\rho}} \right) \cdot (\widehat{\mathcal{C}}_N \cap \mathcal{O}_4)$.

Then, the claim follows by taking the maximum between the two approximation error coefficients for $i = 3, 4$. The asymptotic exactness holds by observing that both coefficients reduce to 1 as τ tends to zero. □

A.13 Proof of Theorem 5

Proof. By Theorem 1 in [GK16], (**O-CC**) is equivalent to inequality

$$\sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{Q}[A(x)\xi \leq b(x)] \equiv \min_{\lambda \geq 0} \left\{ \lambda\delta - \mathbb{E}_{\mathbb{P}} \left[\inf_{\xi \in \Xi} \left\{ \lambda \|\zeta - \xi\| - \mathbb{1} \{ \xi \in \mathcal{S}(x) \} \right\} \right] \right\} \geq 1 - \epsilon.$$

Noting that for any fixed $x \in \mathbb{R}^n$ and $\zeta \in \Xi$

$$\inf_{\xi \in \Xi} \Big\{ \lambda \|\zeta - \xi\| - \mathbb{1}\left\{ \xi \in \mathcal{S}(x) \right\} \Big\} = \begin{cases} -1 & \text{if } \zeta \in \mathcal{S}(x) \\ \min\Big\{ \lambda \cdot \mathbf{d}\left(\zeta, \mathcal{S}(x)\right) - 1, 0 \Big\} & \text{if } \zeta \notin \mathcal{S}(x) \end{cases}$$
$$= \min\Big\{ \lambda \cdot \mathbf{d}\left(\zeta, \mathcal{S}(x)\right) - 1, 0 \Big\},$$

we recast $\mathcal{X}^o$ as

$$\lambda\delta + \mathbb{E}_{\mathbb{P}}\Big[\max\Big\{ 1 - \lambda \cdot \mathbf{d}\left(\zeta, \mathcal{S}(x)\right), 0 \Big\} \Big] \geq 1 - \epsilon \qquad \forall \lambda \geq 0.$$

We notice that the above inequality automatically holds when $\lambda = 0$ because, in this case, the LHS equals one. Hence, we can drop this case and assume that $\lambda > 0$. Then, we divide both sides by λ and denote $\gamma = 1/\lambda$ to obtain

$$\delta + \mathbb{E}_{\mathbb{P}}\left[\Big(\gamma - \mathbf{d}\left(\zeta, \mathcal{S}(x)\right), 0 \Big)^+ \right] \geq (1 - \epsilon)\gamma \qquad \forall \gamma \geq 0.$$

We notice that the above inequality holds for all $\gamma < 0$ because, in that case, the LHS is positive and the RHS is negative. Hence, we expand the domain of γ to be the whole real line and finish the proof as follows:

$$\begin{aligned} & (-\gamma) + \frac{1}{1-\epsilon}\mathbb{E}_{\mathbb{P}}\left[\Big(-\mathbf{d}\left(\zeta, \mathcal{S}(x)\right) - (-\gamma), 0 \Big)^+ \right] \geq -\frac{\delta}{1-\epsilon} \qquad \forall \gamma \in \mathbb{R} \\ \iff & \inf_{-\gamma \in \mathbb{R}} \left\{ (-\gamma) + \frac{1}{1-\epsilon}\mathbb{E}_{\mathbb{P}}\left[\Big(-\mathbf{d}\left(\zeta, \mathcal{S}(x)\right) - (-\gamma), 0 \Big)^+ \right] \right\} \geq \frac{-\delta}{1-\epsilon} \\ \iff & \mathsf{CVaR}_{\epsilon}\Big(-\mathbf{d}\left(\zeta, \mathcal{S}(x)\right) \Big) + \frac{\delta}{1-\epsilon} \geq 0. \end{aligned}$$

□

A.14 Proof of Lemma 5

Proof. Since $\mathbf{d}\left(\zeta, \mathcal{S}(x)\right)$ is defined through a convex program, in which the Slater's condition holds, we take the dual to obtain

$$\mathbf{d}\left(\zeta, \mathcal{S}(x)\right) = \max_{\lambda \leq 0}\Big\{ \lambda^{\mathsf{T}}[b(x) - A\zeta] \colon \|A^{\mathsf{T}}\lambda\|_* \leq 1 \Big\}.$$

This completes the proof.

□

A.15 A Generalized Theorem 6 For Quasi-Concave Inequalities

Theorem 14. Suppose that the reference distribution $\mathbb{P}$ of $\mathcal{P}$ is α-concave with $0 \leq \alpha \leq 1/m$. Then, the set

$$\mathcal{X}_{\mathsf{R}}^{o} := \left\{ x \in \mathbb{R}^n \colon \sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{Q}\Big[h(x, \xi) \geq 0 \Big] \geq 1 - \epsilon \right\}$$

is convex and closed for $\delta > 0$, where $h \colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is quasi-concave.

Proof. In this proof, we show that the distance $\mathbf{d}\left(\zeta, \mathcal{S}(x)\right)$ from $\zeta \in \mathbb{R}^m$ to the safe set $\mathcal{S}(x)$ is jointly convex in (ζ, x) on $\Xi \times \mathbb{R}^n$. Then, the conclusion follows from the proof of Theorem 6.

To show the convexity of $\mathbf{d}\left(\zeta, \mathcal{S}(x)\right)$, we recall that $h \colon \mathbb{R}^n \times \Xi \to \mathbb{R}$ is quasi-concave. Then, the superlevel set $\mathcal{H}_{\geq 0} := \left\{ (x, \xi) \colon h(x, \xi) \geq 0 \right\}$ is convex. In addition,

$$\begin{aligned} \mathbf{d}\left(\zeta, \mathcal{S}(x)\right) &= \inf_{\xi \in \Xi} \left\{ \|\xi - \zeta\| \colon (x, \xi) \in \mathcal{H}_{\geq 0} \right\} \\ &= \min_{\xi \in \Xi} \Big\{ \|\xi - \zeta\| \colon (x, \xi) \in \mathbf{cl}\left(\mathcal{H}_{\geq 0}\right) \Big\}, \end{aligned}$$

where the second equality is because $\|\cdot\|$ is continuous. Take $(x_1, \zeta_1), (x_2, \zeta_2) \in \mathbb{R}^n \times \Xi$, then there exist two minimizers $\xi_1, \xi_2 \in \Xi$ such that they are the closest points in $\mathbf{cl}\left(\mathcal{H}_{\geq 0}\right)$ to (x_1, ζ_1) and (x_2, ζ_2), respectively. It follows that, for $\lambda \in (0, 1)$ and $(x_\lambda, \zeta_\lambda) := \lambda(x_1, \zeta_1) + (1 - \lambda)(x_2, \zeta_2)$,

$$\begin{aligned} \mathbf{d}\left(\zeta_\lambda, \mathcal{S}(x_\lambda)\right) &= \min_{\xi \in \Xi} \Big\{ \|\xi - \zeta_\lambda\| \colon (x_\lambda, \xi) \in \mathbf{cl}\left(\mathcal{H}_{\geq 0}\right) \Big\} \\ &\leq \|\lambda\xi_1 + (1 - \lambda)\xi_2 - \zeta_\lambda\| = \|\lambda(\xi_1 - \zeta_1) + (1 - \lambda)(\xi_2 - \zeta_2)\| \\ &\leq \lambda \mathbf{d}\left(\zeta_1, \mathcal{S}(x_1)\right) + (1 - \lambda)\mathbf{d}\left(\zeta_2, \mathcal{S}(x_2)\right), \end{aligned}$$

where the first inequality is because $(x_\lambda, \lambda\xi_1 + (1 - \lambda)\xi_2) \in \mathbf{cl}\left(\mathcal{H}_{\geq 0}\right)$. □

APPENDIX B

Wasserstein Robust Chance-Constrained Set Covering

B.1 Proof of Proposition 8

Proof. Consider a graph $\mathcal{G} := (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$, on which the classic NP-hard vertex cover problem has the following binary linear formulation:

$$\begin{aligned} \min \ & \sum_{u \in \mathcal{V}} x_u, \\ \text{s.t.} \ & x^{\mathsf{T}} \xi_{u,v} \geq 1, \forall (u,v) \in \mathcal{E}, \\ & x_u \in \{\, 0, 1 \,\}, \forall u \in \mathcal{V}, \end{aligned} \tag{VC}$$

where binary variables x_u indicate whether node $u \in \mathcal{V}$ is part of the vertex cover and $\xi_{u,v}$ is a binary vector with two nonzero entries: $\xi_{u,v} = e_u + e_v$ and $x^{\mathsf{T}} \xi_{u,v} = x_u + x_v$ for all $x \in \{\, 0, 1 \,\}^{|\mathcal{V}|}$. In particular, a vertex cover x can cover every edge twice if and only if all nodes are in the cover, i.e., $x = \mathbf{1}$. We provide a polynomial reduction from (VC) to the following instance of (3.8) to finish the proof:

$$\min_{x \in \{\, 0,1 \,\}^{|\mathcal{V}'|}} \ L(x) := \frac{1}{|\mathcal{V}|} \sum_{u \in \mathcal{V}} x_u + \frac{1}{|\mathcal{V}'|} x_{w'} + x_w - \min_{(u,v) \in \mathcal{E}'} \left(x^{\mathsf{T}} \xi_{u,v} \right)^{1/p}, \tag{B.1}$$

where we add two new nodes w and w' and augment the graph $\mathcal{G}$ to obtain $\mathcal{G}' := (\mathcal{V}', \mathcal{E}')$, $\mathcal{V}' := \mathcal{V} \cup \{\, w, w' \,\}$, and $\mathcal{E}' := \mathcal{E} \cup \{\, (w, w') \,\}$. Since the optimal value of (VC) is bounded above by $|\mathcal{V}|$, we are only interested in whether there is a vertex cover of size less than or equal to $K \in \mathbb{Z}_+, 1 \leq K \leq |\mathcal{V}| - 1$. On the one hand, suppose that there exists a vertex cover x with size less than or equal to K, then together with $x_{w'} := 1$ and $x_w := 0$, $x' := (x, x_{w'}, x_w)$ form a feasible solution to (B.1) with objective value less than or equal to

$(K+1)/|\mathcal{V}'|-1$ because

$$L(x') = \frac{1}{|\mathcal{V}'|}\sum_{u\in\mathcal{V}} x_u + \frac{1}{|\mathcal{V}'|}x_{w'} + x_w - \min_{(u,v)\in\mathcal{E}'}\left(x^{\mathsf{T}}\xi_{u,v}\right)^{1/p} \le \frac{1}{|\mathcal{V}'|}(K+1) - 1.$$

On the other hand, suppose that there is a $y := (y_0, y_{w'}, y_w) \in \{0,1\}^{|\mathcal{V}'|}$ such that $L(y) \le (K+1)/|\mathcal{V}'| - 1$. We discuss the following three cases, with regard to the coverage number $C(y) := \min_{(u,v)\in\mathcal{E}'}\left(y^{\mathsf{T}}\xi_{u,v}\right)^{1/p}$, to show that there exists a vertex cover with size at most K.

1. $C(y) = 0$ is impossible because

$$\begin{aligned} C(y) = \min_{(u,v)\in\mathcal{E}'}\left(y^{\mathsf{T}}\xi_{u,v}\right)^{1/p} &\ge \frac{1}{|\mathcal{V}'|}\sum_{u\in\mathcal{V}} y_u + \frac{1}{|\mathcal{V}'|}y_{w'} + y_w - \frac{1}{|\mathcal{V}'|}(K+1) + 1 \\ &\ge -\frac{1}{|\mathcal{V}'|}(K+1) + 1 \ge 1 - \frac{|\mathcal{V}|+1}{|\mathcal{V}|+2} > 0. \end{aligned}$$

2. $C(y) = 1$: suppose that $y_w = 1$ and $y_{w'}$ equals zero or one. Then an alternative solution $y' := (y_0, y'_{w'}, y'_w)$ with $y'_{w'} = 1$ and $y'_w = 0$ to formulation (B.1) satisfies:

$$\begin{aligned} &C(y') = \min\left\{ \min_{(u,v)\in\mathcal{E}}\left(y_0^{\mathsf{T}}\xi_{u,v}\right)^{1/p}, (y'_{w'} + y'_w)^{1/p} \right\} = 1 \quad \text{and} \\ &L(y') \le L(y) \le \frac{1}{|\mathcal{V}'|}(K+1) - 1. \end{aligned}$$

   It follows that y_0 is a vertex cover of $\mathcal{G}$, and we may assume that $y_{w'} = 1, y_w = 0$ in formulation (B.1) without loss of optimality. Hence,

$$L(y) = \frac{1}{|\mathcal{V}'|}\sum_{u\in\mathcal{V}}(y_0)_u + \frac{1}{|\mathcal{V}'|}y_{w'} + y_w - C(y) = \frac{1}{|\mathcal{V}'|}\left(\sum_{u\in\mathcal{V}}(y_0)_u + 1\right) - 1 \le \frac{1}{|\mathcal{V}'|}(K+1) - 1,$$

   which implies that y_0 is a vertex cover of size at most K.

3. $C(y) = 2^{1/p}$ is also impossible. Indeed, since $C(y) = 2^{1/p}$, every edge of $\mathcal{G}'$ is covered twice, implying that $y_{w'} = w_w = (y_0)_u = 1$ for all $u \in \mathcal{V}$. Then,

$$L(y) = \frac{1}{|\mathcal{V}'|}(|\mathcal{V}|+1) + 1 - 2^{1/p} \le \frac{1}{|\mathcal{V}'|}(K+1) - 1 \le \frac{1}{|\mathcal{V}'|}|\mathcal{V}| - 1.$$

   Simplifying the two ends of the above inequalities gives us $2 + |\mathcal{V}'|^{-1} \le 2^{1/p}$, which is impossible for $p \ge 1$. Therefore, $C(y)$ cannot be $2^{1/p}$.

To sum up, if there is a feasible solution y to formulation (B.1) with $L(y) \leq (K+1)/\left|\mathcal{V}'\right| - 1$, then there is a vertex cover of $\mathcal{G}$ with size at most K. □

B.2 Proof of Lemma 9

Proof. Fix $\widehat{x} \in \{0,1\}^n$. We first prove the equality. If $\widehat{x}^\mathsf{T}\xi - \beta \leq 0$, then

$$f\left(\left(\widehat{x}^\mathsf{T}\xi - \beta\right)^+\right) = 0 = \varphi_1\left(\left(\widehat{x}^\mathsf{T}\xi - \beta\right)^+\right) = h_{(1,0)}(\widehat{x}^\mathsf{T}\xi) \geq \min_{(\ell,\rho)\in L_1} h_{(\ell,\rho)}(\widehat{x}^\mathsf{T}\xi),$$

where the first equality is because $f(0) = 0$, the second equality is because $\varphi_1(z) = f(1)z$, and the third equality is because $\varphi_1((\widehat{x}^\mathsf{T}\xi - \beta)^+) = (\varphi_1(\widehat{x}^\mathsf{T}\xi - \beta))^+$. If $\widehat{\ell} := \widehat{x}^\mathsf{T}\xi - \beta \geq 1$, then

$$f\left(\left(\widehat{x}^\mathsf{T}\xi - \beta\right)^+\right) = f(\widehat{\ell}) = a_{\widehat{\ell}} \cdot \widehat{\ell} + b_{\widehat{\ell}} = \left(\varphi_{\widehat{\ell}}(\widehat{x}^\mathsf{T}\xi - \beta)\right)^+ \geq \min_{(\ell,\rho)\in L_1} h_{(\ell,\rho)}(\widehat{x}^\mathsf{T}\xi),$$

where the last equality is because $a_{\widehat{\ell}}, b_{\widehat{\ell}} \geq 0$. Next we show that $f((x^\mathsf{T}\xi - \beta)^+) \leq \min_{(\ell,\rho)\in L_1} h_{(\ell,\rho)}(x^\mathsf{T}\xi)$. Suppose that $(\widehat{\ell}, \widehat{\rho})$ is a minimizer for $\min_{(\ell,\rho)\in L_1} h_{(\ell,\rho)}(\widehat{x}^\mathsf{T}\xi)$, then

$$\begin{aligned}
\min_{(\ell,\rho)\in L_1} h_{(\ell,\rho)}(\widehat{x}^\mathsf{T}\xi) = h_{(\widehat{\ell},\widehat{\rho})}(\widehat{x}^\mathsf{T}\xi) &= \left(\varphi_{\widehat{\ell}}(\widehat{x}^\mathsf{T}\xi - \beta)\right)^+ \\
&= \varphi_{\widehat{\ell}}\left(\left(\widehat{x}^\mathsf{T}\xi - \beta\right)^+\right) \mathbb{1}\left\{\widehat{x}^\mathsf{T}\xi - \beta \geq 0\right\} + \left(\varphi_{\widehat{\ell}}(\widehat{x}^\mathsf{T}\xi - \beta)\right)^+ \mathbb{1}\left\{\widehat{x}^\mathsf{T}\xi - \beta < 0\right\} \\
&\geq f\left(\left(\widehat{x}^\mathsf{T}\xi - \beta\right)^+\right),
\end{aligned}$$

where the last equality is because $a_{\widehat{\ell}}, b_{\widehat{\ell}} \geq 0$ and the inequality is due to the piecewise linear representation of f and the fact that $f(0) = 0$. Therefore, $f\left(\left(x^\mathsf{T}\xi - \beta\right)^+\right) = \min_{(\ell,\rho)\in L_1} h_{(\ell,\rho)}(x^\mathsf{T}\xi)$.

It remains the show $\min_{(\ell,\rho)\in L_1} h_{(\ell,\rho)}(x^\mathsf{T}\xi) \leq h_{(\widehat{\ell},\widehat{\rho})}(x^\mathsf{T}\xi)$ for all $(\widehat{\ell}, \widehat{\rho}) \in L_2$. We note that

$$a_{\widehat{\ell}+1}(\widehat{\rho} + \widehat{\ell}) \leq a_{\widehat{\ell}+1} \cdot \widehat{\ell} + b_{\widehat{\ell}+1} = f(\widehat{\ell}) = a_{\widehat{\ell}} \cdot \widehat{\ell} + b_{\widehat{\ell}} \leq a_{\widehat{\ell}}(\widehat{\rho} + \widehat{\ell}),$$

implying

$$a_{\widehat{\ell}+1} \leq \frac{f(\widehat{\ell})}{(\widehat{\ell} + \widehat{\rho})} \leq a_{\widehat{\ell}}.$$

Therefore, we have

$$
\begin{aligned}
h_{(\widehat{\ell},\widehat{\rho})}(x^{\mathsf{T}}\xi) &= \left(\frac{f(\widehat{\ell})}{(\widehat{\ell}+\widehat{\rho})}\left(x^{\mathsf{T}}\xi-\beta+\widehat{\rho}\right)\right)^{+} = \left(f(\widehat{\ell})+\frac{f(\widehat{\ell})}{(\widehat{\ell}+\widehat{\rho})}\left(x^{\mathsf{T}}\xi-\beta-\widehat{\ell}\right)\right)^{+} \\
&= \left(f(\widehat{\ell})+\frac{f(\widehat{\ell})}{(\widehat{\ell}+\widehat{\rho})}\left(x^{\mathsf{T}}\xi-\beta-\widehat{\ell}\right)\right)^{+} \mathbb{1}\left\{x^{\mathsf{T}}\xi-\beta\leq\widehat{\ell}\right\}+ \\
&\quad \left(f(\widehat{\ell})+\frac{f(\widehat{\ell})}{(\widehat{\ell}+\widehat{\rho})}\left(x^{\mathsf{T}}\xi-\beta-\widehat{\ell}\right)\right)^{+} \mathbb{1}\left\{x^{\mathsf{T}}\xi-\beta\geq\widehat{\ell}+1\right\} \\
&\geq \left(f(\widehat{\ell})+a_{\widehat{\ell}}\left(x^{\mathsf{T}}\xi-\beta-\widehat{\ell}\right)\right)^{+} \mathbb{1}\left\{x^{\mathsf{T}}\xi-\beta\leq\widehat{\ell}\right\}+ \\
&\quad \left(f(\widehat{\ell})+a_{\widehat{\ell}+1}\left(x^{\mathsf{T}}\xi-\beta-\widehat{\ell}\right)\right)^{+} \mathbb{1}\left\{x^{\mathsf{T}}\xi-\beta\geq\widehat{\ell}+1\right\} \\
&= \left(\varphi_{\widehat{\ell}}\left(x^{\mathsf{T}}\xi-\beta\right)\right)^{+} \mathbb{1}\left\{x^{\mathsf{T}}\xi-\beta\leq\widehat{\ell}\right\}+\left(\varphi_{\widehat{\ell}+1}\left(x^{\mathsf{T}}\xi-\beta\right)\right)^{+} \mathbb{1}\left\{x^{\mathsf{T}}\xi-\beta\geq\widehat{\ell}+1\right\} \\
&\geq \min_{(\ell,\rho)\in L_1} h_{(\ell,\rho)}\left(x^{\mathsf{T}}\xi\right).
\end{aligned}
$$

□

B.3 Proof of Theorem 9

Proof of Theorem 9. The proof is established on Propositions 34, 35, and 36. We present their statements and relegate the proofs to Section B.3.1. We denote the intersection on the RHS of the claim as $\mathcal{X}'$. By Lemma 9, any $(\widehat{\theta},\widehat{x})\in\mathcal{X}$ satisfies

$$
\widehat{\theta}\leq f\left(\left(\widehat{x}^{\mathsf{T}}\xi-\beta\right)^{+}\right)=\min_{(\ell,\rho)\in L_1\cup L_2} h_{(\ell,\rho)}(\widehat{x}^{\mathsf{T}}\xi).
$$

Then, $(\widehat{\theta},\widehat{x})\in\mathcal{X}_{(\ell,\rho)}$ for all $(\ell,\rho)\in L_1\cup L_2$, implying $(\widehat{\theta},\widehat{x})\in\mathcal{X}'$ and $\mathbf{conv}(\mathcal{X})\subseteq\mathcal{X}'$. It remains to prove $\mathcal{X}'\subseteq\mathbf{conv}(\mathcal{X})$. Equivalently, we prove that $(\widehat{\theta},\widehat{x})\notin\mathbf{conv}(\mathcal{X})$ implies $(\widehat{\theta},\widehat{x})\notin\mathbf{conv}(\mathcal{X}_{(\ell,\rho)})$ for some $(\ell,\rho)\in L_1\cup L_2$.

First, by the hyperplane separation theorem, there exists nonzero $(\pi_0,\pi)\in\mathbb{R}^{1+n}$ satisfying

$$
\pi_0\widehat{\theta}+\pi^{\mathsf{T}}\widehat{x}>\pi_0\theta+\pi^{\mathsf{T}}x,\quad\forall(\theta,x)\in\mathcal{X}.
$$

If $\widehat{x}\notin[0,1]^n$ then $(\widehat{\theta},\widehat{x})\notin\mathbf{conv}(\mathcal{X}_{(\ell,\rho)})$ because $\mathbf{conv}(\mathcal{X}_{(\ell,\rho)})\subseteq\mathbb{R}\times[0,1]^n$. Hence, we can

assume $\widehat{x} \in [0,1]^n$ w.l.o.g. Since θ is unbounded below in $\mathcal{X}$, π_0 should be nonnegative. If $\pi_0 = 0$ then $\pi^\mathsf{T}\widehat{x} > \pi^\mathsf{T}x$ for all $x \in \{0,1\}^n$, a contradiction, because this nonzero π separates $\widehat{x}$ from $[0,1]^n$. Hence, $\pi_0 > 0$. We can further assume $\pi_0 = 1$ w.l.o.g. and arrive at

$$\exists \pi \in \mathbb{R}^n : \quad \widehat{\theta} + \pi^\mathsf{T}\widehat{x} > \theta + \pi^\mathsf{T}x, \quad \forall (\theta, x) \in \mathcal{X}.$$

In other words, $\widehat{\theta}$ exceeds the optimal value of the following formulation:

$$\min_{\pi \in \mathbb{R}^n} \left\{ \max \left\{ \theta + \pi^\mathsf{T}x : (\theta, x) \in \mathcal{X} \right\} - \pi^\mathsf{T}\widehat{x} \right\}. \tag{B.2}$$

The rest of the proof solves problem (B.2). The main idea can be summarized as follows:

$$\begin{aligned}
&\min_{\pi \in \mathbb{R}^n} \left\{ \max_{(\theta,x) \in \mathcal{X}} \left\{ \theta + \pi^\mathsf{T}x \right\} - \pi^\mathsf{T}\widehat{x} \right\} \\
&= \min_{\pi \in \Pi^*} \left\{ \max_{(\theta,x) \in \mathcal{X}} \left\{ \theta + \pi^\mathsf{T}x \right\} - \pi^\mathsf{T}\widehat{x} \right\} && \text{(by Propositions 34, 35)} \\
&= \min_{\pi \in \Pi^*} \left\{ \min_{(\ell,\rho) \in L_1 \cup L_2} \left[\max_{(\theta,x) \in \mathcal{X}_{(\ell,\rho)}} \left\{ \theta + \pi^\mathsf{T}x \right\} - \pi^\mathsf{T}\widehat{x} \right] \right\}, && \text{(by Proposition 36)}
\end{aligned}$$

where Π^* is a subset of $\mathbb{R}^n$ given by the optimality conditions derived in Propositions 34 and 35. Proposition 36 relates $\mathcal{X}$ with $\{ \mathcal{X}_{(\ell,\rho)} \}_{(\ell,\rho) \in L_1 \cup L_2}$. Hence, if $\widehat{\theta}$ exceeds the optimal value of (B.2), there exist $\pi' \in \Pi^*$ and $(\ell', \rho') \in L_1 \cup L_2$ that separates $(\widehat{\theta}, \widehat{x})$ from $\mathcal{X}_{(\ell',\rho')}$.

Let $\{ \sigma_i \}_{i=1}^{|\mathcal{Z}|}$ be a permutation of $\mathcal{Z} \equiv \{i \in [n] : \xi_i = 1\}$ such that $\widehat{x}_{\sigma_1} \geq \widehat{x}_{\sigma_2} \geq \cdots \geq \widehat{x}_{\sigma_{|\mathcal{Z}|}}$. For the outer minimization problem in formulation (B.2), we derive the following optimality conditions.

Proposition 32. There exists an optimal solution π to formulation (B.2) satisfying the following:

1. $\pi_{\sigma_1} \geq \pi_{\sigma_2} \geq \cdots \geq \pi_{\sigma_{|\mathcal{Z}|}}$;
2. $\pi \leq 0$, $\pi_{\sigma_{\beta+1}} \geq -a_1$, and $\pi_i = 0$ for all $i \notin \mathcal{Z}$. □

Hence, it suffices to consider those $\pi \in \mathbb{R}^n_-$ with the same ordering as $\widehat{x}$ on $\mathcal{Z}$. It follows

that

$$\begin{aligned}\max_{(\theta,x)\in\mathcal{X}}\left\{\theta+\pi^{\mathsf{T}}x\right\} &= \max_{T\in\{0\}\cup[|\mathcal{Z}|]}\max_{x^{\mathsf{T}}\xi=T}\left\{f\left((x^{\mathsf{T}}\xi-\beta)^{+}\right)+\sum_{i\in\mathcal{Z}}\pi_i x_i\right\}\\ &= \max_{T\in\{0\}\cup[|\mathcal{Z}|]}\left\{f\left((T-\beta)^{+}\right)+\sum_{i=1}^{T}\pi_{\sigma_i}\right\} \qquad \text{(B.3)}\\ &= \max_{T\in\{0\}\cup[|\mathcal{Z}|]}\left\{\min_{(\ell,\rho)\in L_1\cup L_2}h_{(\ell,\rho)}(T)+\sum_{i=1}^{T}\pi_{\sigma_i}\right\},\end{aligned}$$

where the first equality uses the definition of $\mathcal{X}$ and the optimality condition $\pi_i = 0$ for all $i \notin \mathcal{Z}$, the second equality uses the ordering of π on $\mathcal{Z}$, and the last equality follows from Lemma 9.

Second, we show that there exists a set $\Pi^* \subseteq \mathbb{R}^n$ of candidate optimal solutions such that a $\pi \in \Pi^*$ solves problem (B.2). In addition, we show that the following minimax-type claim holds for all $\pi \in \Pi^*$:

$$\max_{T\in\{0\}\cup[|\mathcal{Z}|]}\left\{\min_{(\ell,\rho)\in L_1\cup L_2}h_{(\ell,\rho)}(T)+\sum_{i=1}^{T}\pi_{\sigma_i}\right\}=\min_{(\ell,\rho)\in L_1\cup L_2}\max_{T\in\{0\}\cup[|\mathcal{Z}|]}\left\{h_{(\ell,\rho)}(T)+\sum_{i=1}^{T}\pi_{\sigma_i}\right\},$$

To this end, we consider the one-dimensional maximization problem (B.3) and observe that its objective function is nonincreasing on $\{0\}\cup[\beta]$ since $f\left((T-\beta)^{+}\right)$ is always 0. In addition, if $T \in \{\beta+1,\ldots,|\mathcal{Z}|\}$, it has decreasing increments $a_{i-\beta}+\pi_{\sigma_i}$ for any $\beta+1 \le i \le |\mathcal{Z}|$ because f is concave. Since $a_1+\pi_{\sigma_{\beta+1}} \ge 0$ by Proposition 34, we define t to be the last occurrence of nonnegative increment, i.e.,

$$t := \begin{cases}\max\left\{\beta+1\le i\le|\mathcal{Z}|-1 : a_{i-\beta}+\pi_{\sigma_i}\ge 0, a_{i-\beta+1}+\pi_{\sigma_{i+1}}\le 0\right\} & \text{if } a_{|\mathcal{Z}|-\beta}+\pi_{\sigma_{|\mathcal{Z}|}}\le 0\\ |\mathcal{Z}| & \text{if } a_{|\mathcal{Z}|-\beta}+\pi_{\sigma_{|\mathcal{Z}|}}>0.\end{cases}$$

Then, with any fixed π satisfying the optimality conditions in Proposition 34, a maximizer T^* of problem (B.3) is either 0 or t, yielding the optimal value

$$\left(\sum_{i=1}^{\beta}\pi_{\sigma_i}+\sum_{i=\beta+1}^{t}(a_{i-\beta}+\pi_{\sigma_i})\right)^{+}.$$

This allows us to represent problem (B.2) as follows by partitioning the feasible region of π:

$$\text{(B.2)} = \min_{\substack{t\in\mathbb{Z},\\ \beta+1\le t\le|\mathcal{Z}|}} \min_{\pi\in\Pi_t} \left\{ \left(\sum_{i=1}^{\beta} \pi_{\sigma_i} + \sum_{i=\beta+1}^{t} (a_{i-\beta} + \pi_{\sigma_i}) \right)^+ - \sum_{i=1}^{|\mathcal{Z}|} \pi_{\sigma_i} \widehat{x}_{\sigma_i} \right\},$$

$$\text{where} \quad \Pi_t := \left\{ \begin{array}{c} \text{the last occurrence of } a_{i-\beta} + \pi_{\sigma_i} \ge 0 \text{ is at } t \\ \pi \in \mathbb{R}^n : 0 \ge \pi_{\sigma_1} \ge \pi_{\sigma_2} \ge \cdots \ge \pi_{\sigma_{|\mathcal{Z}|}} \\ \pi_i = 0, \quad \forall i \notin \mathcal{Z} \end{array} \right\}, \quad \forall \beta+1 \le t \le |\mathcal{Z}|.$$

For each t, we further identify the structure of an optimal $\pi^* \in \Pi_t$.

Proposition 33. For all $t \in \{\beta+1, \ldots, |\mathcal{Z}|\}$, there exists an optimal solution π^* to the inner minimization problem

$$\min_{\pi\in\Pi_t} \left\{ \left(\sum_{i=1}^{\beta} \pi_{\sigma_i} + \sum_{i=\beta+1}^{t} (a_{i-\beta} + \pi_{\sigma_i}) \right)^+ - \sum_{i=1}^{|\mathcal{Z}|} \pi_{\sigma_i} \widehat{x}_{\sigma_i} \right\}$$

that has the following structure:

1. If $b_{t-\beta} - \beta a_{t-\beta} \ge 0$, then $\pi^*_{\sigma_1} = \pi^*_{\sigma_2} = \cdots = \pi^*_{\sigma_{|\mathcal{Z}|}} = -a_{t-\beta}$.
2. If $b_{t-\beta} - \beta a_{t-\beta} < 0$, then π^* takes one of the following two forms:

   (A) Let $\widetilde{t} := \beta - \frac{b_{t-\beta}}{a_{t-\beta}}$. Then,

   $$0 = \pi^*_{\sigma_1} = \cdots = \pi^*_{\sigma_{\lfloor\widetilde{t}\rfloor}} \ge \pi^*_{\sigma_{\lceil\widetilde{t}\rceil}} = -a_{t-\beta}(\lceil\widetilde{t}\rceil - \widetilde{t}) > \pi^*_{\sigma_{\lceil\widetilde{t}\rceil+1}} = \cdots = \pi^*_{\sigma_{|\mathcal{Z}|}} = -a_{t-\beta}.$$

   (B) If there exists an integer $t' \in \left(\beta - \frac{b_{t-\beta+1}}{a_{t-\beta+1}}, \beta - \frac{b_{t-\beta}}{a_{t-\beta}}\right) \cap \mathbb{Z}_+$, then

   $$0 = \pi^*_{\sigma_1} = \cdots = \pi^*_{\sigma_{t'}} > \pi^*_{\sigma_{t'+1}} = \cdots = \pi^*_{\sigma_{|\mathcal{Z}|}} = -\frac{f(t-\beta)}{(t-t')}.$$ □

Accordingly, for all $t \in \{\beta+1, \ldots, |\mathcal{Z}|\}$, if $b_{t-\beta} - \beta a_{t-\beta} \ge 0$, then we define

$$\Pi^*_t := \left\{ \pi \in \mathbb{R}^n : \pi_{\sigma_1} = \pi_{\sigma_2} = \cdots = \pi_{\sigma_{|\mathcal{Z}|}} = -a_{t-\beta},\ \pi_i = 0, \forall i \notin \mathcal{Z} \right\};$$

and if $b_{t-\beta} - \beta a_{t-\beta} < 0$, then we define

$$\Pi_t^* := \left\{ \pi : \begin{array}{l} \pi_{\sigma_1} = \pi_{\sigma_2} = \cdots = \pi_{\sigma_{\lfloor \widetilde{t} \rfloor}} = 0, \\ \pi_{\sigma_{\lceil \widetilde{t} \rceil}} = -a_{t-\beta}(\lceil \widetilde{t} \rceil - \widetilde{t}), \\ \pi_{\sigma_{\lceil \widetilde{t} \rceil + 1}} = \cdots = \pi_{\sigma_{|\mathcal{Z}|}} = -a_{t-\beta} \\ \pi_i = 0, \quad \forall i \notin \mathcal{Z} \end{array} \right\} \cup \left\{ \pi : \begin{array}{l} \pi_{\sigma_1} = \cdots = \pi_{\sigma_{t'}} = 0, \\ \pi_{\sigma_{t'+1}} = \cdots \pi_{\sigma_{|\mathcal{Z}|}} = -\dfrac{f(t-\beta)}{(t-t')}, \\ \pi_i = 0, \quad \forall i \notin \mathcal{Z} \\ t' \in \left(\beta - \dfrac{b_{t-\beta+1}}{a_{t-\beta+1}}, \beta - \dfrac{b_{t-\beta}}{a_{t-\beta}} \right) \cap \mathbb{Z}_+ \end{array} \right\}.$$

Note that if the interval $(\beta - b_{t-\beta+1}/a_{t-\beta+1}, \beta - b_{t-\beta}/a_{t-\beta})$ does not contain an integer, then the second component of the above union equals $\varnothing$. Let $\Pi^* := \bigcup_t \Pi_t^*$, then Proposition 35 shows that Π_t can be replaced by Π_t^* without loss of optimality. In other words,

$$\begin{aligned} \text{(B.2)} &= \min_{\substack{t \in \mathbb{Z}, \\ \beta+1 \le t \le |\mathcal{Z}|}} \min_{\pi \in \Pi_t^*} \left\{ \left(\sum_{i=1}^{\beta} \pi_{\sigma_i} + \sum_{i=\beta+1}^{t} (a_{i-\beta} + \pi_{\sigma_i}) \right)^+ - \sum_{i=1}^{|\mathcal{Z}|} \pi_{\sigma_i} \widehat{x}_{\sigma_i} \right\} \\ &= \min_{\pi \in \Pi^*} \left\{ \max_{T \in \{0\} \cup [|\mathcal{Z}|]} \left\{ f((T-\beta)^+) + \sum_{i=1}^{T} \pi_{\sigma_i} \right\} - \sum_{i=1}^{|\mathcal{Z}|} \pi_{\sigma_i} \widehat{x}_{\sigma_i} \right\}. \end{aligned}$$

Furthermore, Proposition 36 shows that, for all $\pi \in \Pi^*$,

$$\max_{T \in \{0\} \cup [|\mathcal{Z}|]} \left\{ f\left((T-\beta)^+\right) + \sum_{i=1}^{T} \pi_{\sigma_i} \right\} = \min_{(\ell,\rho) \in L_1 \cup L_2} \max_{T \in \{0\} \cup [|\mathcal{Z}|]} \left\{ h_{(\ell,\rho)}(T) + \sum_{i=1}^{T} \pi_{\sigma_i} \right\}.$$

It follows that, for all $\pi \in \Pi^*$,

$$\begin{aligned} & \max_{T \in \{0\} \cup [|\mathcal{Z}|]} \left\{ f\left((T-\beta)^+\right) + \sum_{i=1}^{T} \pi_{\sigma_i} \right\} - \sum_{i=1}^{|\mathcal{Z}|} \pi_{\sigma_i} \widehat{x}_{\sigma_i} \\ = & \min_{(\ell,\rho) \in L_1 \cup L_2} \max_{x \in \{0,1\}^n} \left\{ h_{(\ell,\rho)}(x^\mathsf{T} \xi) + \pi^\mathsf{T} x \right\} - \pi^\mathsf{T} \widehat{x} \\ = & \min_{(\ell,\rho) \in L_1 \cup L_2} \max_{(\theta,x) \in \mathcal{X}_{(\ell,\rho)}} \left\{ \theta + \pi^\mathsf{T} x \right\} - \pi^\mathsf{T} \widehat{x}, \end{aligned}$$

where the first equality is because π satisfies the optimality conditions in Proposition 34.

Thus, the optimal value of problem (B.2) satisfies

$$\widehat{\theta} > \min_{\pi \in \Pi^*} \min_{(\ell,\rho) \in L_1 \cup L_2} \left\{ \max_{(\theta,x) \in \mathcal{X}_{(\ell,\rho)}} \left\{ \theta + \pi^\mathsf{T} x \right\} - \pi^\mathsf{T} \widehat{x} \right\}.$$

That is, there exist $(\ell, \rho) \in L_1 \cup L_2$ and $\pi' \in \Pi^*$ such that

$$\widehat{\theta} + \widehat{x}^\mathsf{T} \pi' > \theta + x^\mathsf{T} \pi', \quad \forall (\theta, x) \in \mathcal{X}_{(\ell,\rho)}.$$

Therefore, $(1, \pi')$ separates $(\widehat{\theta}, \widehat{x})$ from the set $\mathcal{X}_{(\ell,\rho)}$, implying $(\widehat{\theta}, \widehat{x}) \notin \textbf{conv}\,(\mathcal{X}_{(\ell,\rho)})$. This finishes the proof. □

B.3.1 Proofs of Propositions 34, 35, and 36

Proposition 34. There exists an optimal solution π to formulation (B.2) satisfying the following:

1. $\pi_{\sigma_1} \geq \pi_{\sigma_2} \geq \cdots \geq \pi_{\sigma_{|\mathcal{Z}|}}$;
2. $\pi \leq 0$, $\pi_{\sigma_{\beta+1}} \geq -a_1$, and $\pi_i = 0$ for all $i \notin \mathcal{Z}$.

Proof. First, for any $\pi \in \mathbb{R}^n$ let $\{\, \sigma'_i : i \in \mathcal{Z} \,\}$, possibly different from $\{\, \sigma_i : i \in \mathcal{Z} \,\}$, be a permutation of $\mathcal{Z} \equiv \{i \in [n] : \xi_i = 1\}$ such that $\pi_{\sigma'_1} \geq \cdots \geq \pi_{\sigma'_{|\mathcal{Z}|}}$. Then, we recast formulation (B.2) as

$$\begin{aligned}
&\min_{\pi \in \mathbb{R}^n} \left\{ \max \left\{ \theta + \pi^\mathsf{T} x : (\theta, x) \in \mathcal{X} \right\} - \pi^\mathsf{T} \widehat{x} \right\} \\
= &\min_{\pi \in \mathbb{R}^n} \left\{ \max_{x \in \{\,0,1\,\}^n} \left\{ f\left(\left(x^\mathsf{T} \xi - \beta \right)^+ \right) + \sum_{i \in \mathcal{Z}} \pi_{\sigma'_i} x_{\sigma'_i} + \sum_{i \notin \mathcal{Z}} \pi_i x_i \right\} - \pi^\mathsf{T} \widehat{x} \right\} \\
= &\min_{\pi \in \mathbb{R}^n} \left\{ \max_{T \in \{0\} \cup [|\mathcal{Z}|]} \max_{x^\mathsf{T} \xi = T} \left\{ f\left(\left(x^\mathsf{T} \xi - \beta \right)^+ \right) + \sum_{i \in \mathcal{Z}} \pi_{\sigma'_i} x_{\sigma'_i} + \sum_{i \notin \mathcal{Z}} \pi_i x_i \right\} - \pi^\mathsf{T} \widehat{x} \right\} \\
= &\min_{\pi \in \mathbb{R}^n} \left\{ \max_{T \in \{0\} \cup [|\mathcal{Z}|]} \left\{ f\left((T - \beta)^+ \right) + \sum_{i=1}^{T} \pi_{\sigma'_i} \right\} + \sum_{i \notin \mathcal{Z}} (\pi_i)^+ - \pi^\mathsf{T} \widehat{x} \right\}. \qquad \text{(B.4)}
\end{aligned}$$

We prove the following optimality conditions:

Condition 1. Suppose that $\pi \in \mathbb{R}^n$ is such that $\pi_{\sigma_j} < \pi_{\sigma_k}$ for some $j, k \in [|\mathcal{Z}|]$ and $j < k$.

We show swapping π_{σ_j} and π_{σ_k} yields a lower objective value of (B.4). Specifically,

define π' as

$$\pi'_{\sigma_i} := \begin{cases} \pi_{\sigma_k} & \text{if } i = j, \\ \pi_{\sigma_j} & \text{if } i = k, \quad \forall i \in \mathcal{Z}, \\ \pi_{\sigma_i} & \text{o.w.} \end{cases}$$
$$\pi'_i := \pi_i \quad \forall i \notin \mathcal{Z}.$$

Since π, π' differ only in ordering, the optimal value of the inner maximization problem of (B.4) remains the same after swapping. Then, the difference between their objective values equals

$$\begin{aligned} -\widehat{x}^{\mathsf{T}}\pi - \left(-\widehat{x}^{\mathsf{T}}\pi'\right) &= -(\pi_{\sigma_j}\widehat{x}_{\sigma_j} + \pi_{\sigma_k}\widehat{x}_{\sigma_k}) + (\pi'_{\sigma_j}\widehat{x}_{\sigma_j} + \pi'_{\sigma_k}\widehat{x}_{\sigma_k}) \\ &= (\widehat{x}_{\sigma_j} - \widehat{x}_{\sigma_k})(\pi_{\sigma_k} - \pi_{\sigma_j}) \geq 0. \end{aligned}$$

Therefore, there exists an optimal π that shares the same ordering with $\widehat{x}$ on $\mathcal{Z}$.

Condition 2. To prove that the entries of an optimal π on $[n] \setminus \mathcal{Z}$ are all zero, we show that replacing any nonzero entry π_j, $j \notin \mathcal{Z}$, with 0 lowers the objective value. Specifically, define π' to be the π with π_j replaced by zero. This yields a difference in objective value

$$\left(\left(\pi_j\right)^+ - \pi_j\widehat{x}_j\right) - 0 = \pi_j(1 - \widehat{x}_j)\mathbb{1}\left\{\, \pi_j > 0 \,\right\} - \pi_j\widehat{x}_j\mathbb{1}\left\{\, \pi_j \leq 0 \,\right\} \geq 0.$$

To prove $\pi \leq 0$ without loss of optimality, it suffices to show that π is nonpositive on $\mathcal{Z}$. Suppose that π has strictly positive entries on $\mathcal{Z}$ and let π_{σ_j} be the last such entry, i.e., either $\pi_{\sigma_j} > 0, \pi_{\sigma_{j+1}} \leq 0$ or $\pi_{\sigma_{|\mathcal{Z}|}} > 0$ with $j = |\mathcal{Z}|$. We show that replacing π_{σ_j} with 0 lowers the objective value. Specifically, let π' be the π with π_{σ_j} replaced by 0. Then, π and π' share the same ordering. In addition, both π_{σ_i} and π'_{σ_i} are nonnegative for all $i \in [j]$. As a result, since f is increasing, a maximizer T^* of the inner maximization problem of (B.4) must be at least j with respect to both π and π'. It follows that π and π' yield the same maximizer T^*. Therefore, the difference in objective value equals

$$\pi_{\sigma_j}(1 - \widehat{x}_{\sigma_j}) - 0 \geq 0,$$

implying that the objective value evaluated at π' is smaller. To prove $\pi_{\sigma_{\beta+1}} \geq -a_1$, we note that if $\pi_{\sigma_{\beta+1}} + a_1 < 0$ then $\pi_{\sigma_{\beta+i}} + a_i < 0$ for all $i \in [|\mathcal{Z}| - \beta]$, because $\pi_{\sigma_{\beta+i}}$ and a_i are decreasing in i. Since $\pi \leq 0$, an optimal T^* to the inner maximization problem

of (B.4) is 0. It follows that increasing all π_{σ_i}, $1 \leq i \leq \beta+1$ to $-a_1$ does not increase the objective value of (B.4). Hence, we have $\pi_{\sigma_{\beta+1}} + a_1 \geq 0$ without loss of optimality.

□

Proposition 35. For all $t \in \{\beta+1, \ldots, |\mathcal{Z}|\}$, there exists an optimal solution π^* to the inner minimization problem

$$\min_{\pi \in \Pi_t} \left\{ \left(\sum_{i=1}^{\beta} \pi_{\sigma_i} + \sum_{i=\beta+1}^{t} (a_{i-\beta} + \pi_{\sigma_i}) \right)^+ - \sum_{i=1}^{|\mathcal{Z}|} \pi_{\sigma_i} \widehat{x}_{\sigma_i} \right\} \tag{B.5}$$

that has the following structure:

1. If $b_{t-\beta} - \beta a_{t-\beta} \geq 0$, then $\pi^*_{\sigma_1} = \pi^*_{\sigma_2} = \cdots = \pi^*_{\sigma_{|\mathcal{Z}|}} = -a_{t-\beta}$.
2. If $b_{t-\beta} - \beta a_{t-\beta} < 0$, then π^* takes one of the following two forms:

   (A) Let $\widetilde{t} := \beta - \frac{b_{t-\beta}}{a_{t-\beta}}$. Then,

   $$0 = \pi^*_{\sigma_1} = \cdots = \pi^*_{\sigma_{\lfloor \widetilde{t} \rfloor}} \geq \pi^*_{\sigma_{\lceil \widetilde{t} \rceil}} = -a_{t-\beta}(\lceil \widetilde{t} \rceil - \widetilde{t}) > \pi^*_{\sigma_{\lceil \widetilde{t} \rceil + 1}} = \cdots = \pi^*_{\sigma_{|\mathcal{Z}|}} = -a_{t-\beta}.$$

   (B) If there exists an integer $t' \in \left(\beta - \frac{b_{t-\beta+1}}{a_{t-\beta+1}}, \beta - \frac{b_{t-\beta}}{a_{t-\beta}} \right) \cap \mathbb{Z}_+$, then

   $$0 = \pi^*_{\sigma_1} = \cdots = \pi^*_{\sigma_{t'}} > \pi^*_{\sigma_{t'+1}} = \cdots = \pi^*_{\sigma_{|\mathcal{Z}|}} = -\frac{f(t-\beta)}{(t-t')}.$$

Proof. Fix $t \in \{\beta+1, \ldots, |\mathcal{Z}|\}$ and we discuss the following two cases.

First, if $b_{t-\beta} - \beta a_{t-\beta} \geq 0$ then

$$\begin{aligned} \sum_{i=1}^{\beta} \pi_{\sigma_i} + \sum_{i=\beta+1}^{t} (a_{i-\beta} + \pi_{\sigma_i}) &\geq t\pi_{\sigma_t} + \sum_{i=\beta+1}^{t} a_{i-\beta} \\ &\geq -ta_{t-\beta} + a_{t-\beta}(t-\beta) + b_{t-\beta} \\ &= b_{t-\beta} - \beta a_{t-\beta} \geq 0, \end{aligned}$$

where the first inequality is because $\pi_{\sigma_i} \geq \pi_{\sigma_t}$ for all $i \in [t]$ and the second inequality is because $\pi_{\sigma_t} \geq -a_{t-\beta}$ by the definition of t and $\sum_{i=\beta+1}^{t} a_{i-\beta} = \sum_{i=\beta+1}^{t} (f(i-\beta) - f(i-\beta-1)) = f(t-\beta) = a_{t-\beta}(t-\beta) + b_{t-\beta}$. Then, we recast (B.5) as

$$\text{(B.5)} = \min_{\pi \in \Pi_t} \left\{ \sum_{i=\beta+1}^{t} a_{i-\beta} + \sum_{i=1}^{t} \pi_{\sigma_i}(1 - \widehat{x}_{\sigma_i}) - \sum_{i=t+1}^{|\mathcal{Z}|} \pi_{\sigma_i} \widehat{x}_{\sigma_i} \right\}.$$

Since $\widehat{x} \in [0,1]^n$, decreasing $\{\pi_{\sigma_i} : i \in [t]\}$ and increasing $\{\pi_{\sigma_i} : t+1 \le i \le |\mathcal{Z}|\}$ improve the objective value. It follows that an optimal solution π^* takes the following form:

$$\pi^*_{\sigma_1} = \cdots = \pi^*_{\sigma_t} = -a_{t-\beta} = \pi^*_{\sigma_{t+1}} = \cdots = \pi^*_{\sigma_{|\mathcal{Z}|}}.$$

Second, suppose that $b_{t-\beta} - \beta a_{t-\beta} < 0$. For $\pi \in \Pi_t$ we define

$$S(\pi) := \sum_{i=1}^{\beta} \pi_{\sigma_i} + \sum_{i=\beta+1}^{t} (a_{i-\beta} + \pi_{\sigma_i}).$$

We show $S(\pi) = 0$ at optimality of (B.5). To see that, we discuss the following two cases.

1. If $S(\pi) > 0$, then there exists an $i \in \mathbb{Z}, \beta+1 \le i \le |\mathcal{Z}|$ such that $a_{i-\beta} + \pi_{\sigma_i} > 0$. Let $\underline{t}$ be its last occurrence:

   $$\underline{t} := \max\left\{\, \beta+1 \le i \le |\mathcal{Z}| : a_{i-\beta} + \pi_{\sigma_i} > 0 \,\right\}.$$

   Note that $\underline{t} \le t$ by definition and $a_{i-\beta} + \pi_{\sigma_i} = 0$ for all $\underline{t} < i \le t$. Then,

   $$\begin{aligned}\max\left\{\, S(\pi), 0 \,\right\} - \sum_{i=1}^{|\mathcal{Z}|} \pi_{\sigma_i}\widehat{x}_{\sigma_i} &= \sum_{i=1}^{\beta} \pi_{\sigma_i} + \sum_{i=\beta+1}^{\underline{t}} (a_{i-\beta} + \pi_{\sigma_i}) - \sum_{i=1}^{|\mathcal{Z}|} \pi_{\sigma_i}\widehat{x}_{\sigma_i} \\ &= \sum_{i=\beta+1}^{\underline{t}} a_{i-\beta} + \sum_{i=1}^{\underline{t}-1} \pi_{\sigma_i}(1 - \widehat{x}_{\sigma_i}) + \pi_{\sigma_{\underline{t}}}(1 - \widehat{x}_{\sigma_{\underline{t}}}) - \sum_{i=\underline{t}+1}^{|\mathcal{Z}|} \pi_{\sigma_i}\widehat{x}_{\sigma_i}.\end{aligned}$$

   Therefore, as long as $S(\pi) > 0$, decreasing $\pi_{\sigma_{\underline{t}}}$ lowers the objective value of (B.5).

2. If $S(\pi) < 0$, then there exists an $i \in [\beta]$ such that $\pi_{\sigma_i} < 0$. Let $\bar{t}$ be its first occurrence:

   $$\bar{t} := \min\left\{\, i \in [\beta] : \pi_{\sigma_i} < 0 \,\right\}.$$

   Then, by noting that $\pi_{\sigma_i} = 0$ for all $i \in [\bar{t}-1]$, we have

   $$\max\left\{\, S(\pi), 0 \,\right\} - \sum_{i=1}^{|\mathcal{Z}|} \pi_{\sigma_i}\widehat{x}_{\sigma_i} = -\pi_{\sigma_{\bar{t}}}\widehat{x}_{\sigma_{\bar{t}}} - \sum_{i=\bar{t}+1}^{|\mathcal{Z}|} \pi_{\sigma_i}\widehat{x}_{\sigma_i}.$$

   Therefore, as long as $S(\pi) < 0$, increasing $\pi_{\sigma_{\bar{t}}}$ lowers the objective value of (B.5).

In view of the optimality condition $S(\pi) = 0$, we reformulate (B.5) as the following linear

program:

$$
\begin{aligned}
\text{(B.5)} = \min_{\pi} \; & -\sum_{i=1}^{|\mathcal{Z}|} \pi_{\sigma_i} \widehat{x}_{\sigma_i}, \\
\text{s.t. } & 0 \geq \pi_{\sigma_1}, \\
& \pi_{\sigma_i} \geq \pi_{\sigma_{i+1}}, \; \forall i \in [|\mathcal{Z}| - 1], \\
& \pi_{\sigma_t} + a_{t-\beta} \geq 0, \\
& \pi_{\sigma_{t+1}} + a_{t-\beta+1} \leq 0, && \text{(B.6)} \\
& \sum_{i=1}^{t} \pi_{\sigma_i} + \sum_{i=\beta+1}^{t} a_{i-\beta} = 0, && \text{(B.7)}
\end{aligned}
$$

where constraint (B.7) states $S(\pi) = 0$ and the remaining constraints come from the definition of Π_t. Note that constraint (B.6) does not exist in the case $t = |\mathcal{Z}|$. Although the remaining proof focuses on $t < |\mathcal{Z}|$, it can be directly adapted to the case $t = |\mathcal{Z}|$. In what follows, we analyze the basic feasible solutions of the above linear program to identify the structure of an optimal solution π^*. There are $|\mathcal{Z}|$ decision variables, 1 equality constraint, and $|\mathcal{Z}| + 2$ inequality constraints. By definition of t, the constraint $\pi_{\sigma_{t+1}} + a_{t-\beta+1} \leq 0$ must hold strictly. In addition, variables $\pi_{\sigma_{t+2}}, \ldots, \pi_{\sigma_{|\mathcal{Z}|}}$ are only constrained by their ordering and that $\pi_{\sigma_{t+2}} \leq \pi_{\sigma_{t+1}}$, and their coefficients in the objective function are all negative. As a consequence, $\pi_{\sigma_{t+1}} = \pi_{\sigma_{t+2}} = \cdots = \pi_{\sigma_{|\mathcal{Z}|}}$ at optimality. Hence, we only need to consider those basic feasible solutions, where either (A) $\pi_{\sigma_t} + a_{t-\beta} = 0$ and two among the constraints $0 \geq \pi_{\sigma_1} \geq \cdots \geq \pi_{\sigma_{t+1}}$ are satisfied strictly (or weakly in case of degeneracy), or (B) $\pi_{\sigma_t} + a_{t-\beta} > 0$ and one of the constraints $0 \geq \pi_{\sigma_1} \geq \cdots \geq \pi_{\sigma_{t+1}}$ is satisfied strictly (or weakly in case of degeneracy). We discuss cases (A) and (B) to finish the proof.

(A) Since $\pi_{\sigma_t} + a_{t-\beta} = 0$, there exist $0 \leq t_1 < t_2 \leq t$ such that

$$
0 = \pi_{\sigma_1} = \cdots = \pi_{\sigma_{t_1}} \geq \pi_{\sigma_{t_1+1}} = \cdots = \pi_{\sigma_{t_2}} \geq \pi_{\sigma_{t_2+1}} = \cdots = \pi_{\sigma_{|\mathcal{Z}|}} = -a_{t-\beta}.
$$

By constraint (B.7), we have:

$$
(t_2 - t_1)\pi_{\sigma_{t_2}} = -\sum_{i=t_2+1}^{t} (-a_{t-\beta}) - \sum_{i=\beta+1}^{t} a_{i-\beta} = -t_2 a_{t-\beta} + (\beta a_{t-\beta} - b_{t-\beta}),
$$

where the last equality is because $\sum_{i=\beta+1}^{t} a_{i-\beta} = \sum_{i=\beta+1}^{t} (f(i-\beta) - f(i-\beta-1)) = f(t-\beta) = a_{t-\beta}(t-\beta) + b_{t-\beta}$. Let $\widetilde{t} := \beta - \frac{b_{t-\beta}}{a_{t-\beta}}$. Then, a valid choice of $\pi_{\sigma_{t_2}}$ should

satisfy $-a_{t-\beta} \leq \pi_{\sigma_{t_2}} \leq 0$, which, together with the above equality, implies

$$-(t_2 - t_1)a_{t-\beta} \leq (\widetilde{t} - t_2)a_{t-\beta} \leq 0 \quad \Leftrightarrow \quad t_1 \leq \widetilde{t} \leq t_2.$$

We assume that $\widetilde{t}$ is fractional and will return to the case of integer-valued $\widetilde{t}$ in the end. We show $t_1 = \lfloor \widetilde{t} \rfloor$ and $t_2 = \lceil \widetilde{t} \rceil$ at optimality. Indeed, if $t_1 \leq \lfloor \widetilde{t} \rfloor - 1$ then increasing t_1 by 1 improves the objective value. Specifically, let π' be the π with t_1 replaced by $t_1 + 1$. Then, $\pi'_{\sigma_{t_2}} \leq \pi_{\sigma_{t_2}}$ because $\sum_{i=1}^{t} \pi_{\sigma_i}$ equals a constant due to constraint (B.7). In addition,

$$\begin{aligned}
(t_2 - t_1)\pi_{\sigma_{t_2}} &= \sum_{i=1}^{t} \pi_{\sigma_i} - \sum_{i=t_2+1}^{t} \pi_{\sigma_i} && \text{(by construction of } \pi) \\
&= -\sum_{i=\beta+1}^{t} a_{i-\beta} - \sum_{i=t_2+1}^{t} \pi_{\sigma_i} && \text{(by constraint (B.7))} \\
&= \sum_{i=1}^{t} \pi'_{\sigma_i} - \sum_{i=t_2+1}^{t} \pi'_{\sigma_i} && \text{(by construction of } \pi') \\
&= (t_2 - t_1 - 1)\pi'_{\sigma_{t_2}}.
\end{aligned}$$

Hence, the change in objective value equals

$$\begin{aligned}
-\pi_{\sigma_{t_2}} \sum_{i=t_1+1}^{t_2} \widehat{x}_{\sigma_i} + \pi'_{\sigma_{t_2}} \sum_{i=t_1+2}^{t_2} \widehat{x}_{\sigma_i} &= -\pi_{\sigma_{t_2}} \widehat{x}_{\sigma_{t_1+1}} - (\pi_{\sigma_{t_2}} - \pi'_{\sigma_{t_2}}) \sum_{i=t_1+2}^{t_2} \widehat{x}_{\sigma_i} \\
&\geq -\pi_{\sigma_{t_2}} \widehat{x}_{\sigma_{t_1+1}} - (\pi_{\sigma_{t_2}} - \pi'_{\sigma_{t_2}}) \widehat{x}_{\sigma_{t_1+1}} (t_2 - t_1 - 1) \\
&= -\widehat{x}_{\sigma_{t_1+1}} \left[\pi_{\sigma_{t_2}} + (\pi_{\sigma_{t_2}} - \pi'_{\sigma_{t_2}})(t_2 - t_1 - 1) \right] \\
&= 0,
\end{aligned}$$

where the inequality is because $\widehat{x}_{\sigma_i}$ decreases in i. Therefore, $t_1 = \lfloor \widetilde{t} \rfloor$ at optimality. Likewise, if $t_2 \geq \lfloor \widetilde{t} \rfloor + 1$ then decreasing t_2 by 1 improves the objective value. Specifically, we let π'' be the π with t_2 replaced by $t_2 - 1$. Then, $\pi''_{\sigma_{t_2}} \geq \pi_{\sigma_{t_2}}$ because $\sum_{i=1}^{t} \pi_{\sigma_i}$

equals a constant due to constraint (B.7). In addition,

$$\begin{aligned}
(t_2 - t_1)\pi_{\sigma_{t_2}} &= \sum_{i=1}^{t} \pi_{\sigma_i} - \sum_{i=t_2+1}^{t} \pi_{\sigma_i} && \text{(by construction of } \pi\text{)} \\
&= -\sum_{i=\beta+1}^{t} a_{i-\beta} - \sum_{i=t_2+1}^{t} \pi_{\sigma_i} && \text{(by constraint (B.7))} \\
&= \sum_{i=1}^{t} \pi''_{\sigma_i} - \sum_{i=t_2+1}^{t} \pi''_{\sigma_i} && \text{(by construction of } \pi''\text{)} \\
&= (t_2 - t_1 - 1)\pi''_{\sigma_{t_2}} + (-a_{t-\beta}).
\end{aligned}$$

Hence, the change in objective value equals

$$\begin{aligned}
& - \pi_{\sigma_{t_2}} \sum_{i=t_1+1}^{t_2} \widehat{x}_{\sigma_i} + a_{t-\beta} \sum_{i=t_2+1}^{|\mathcal{Z}|} \widehat{x}_{\sigma_i} + \pi''_{\sigma_{t_2}} \sum_{i=t_1+1}^{t_2-1} \widehat{x}_{\sigma_i} - a_{t-\beta} \sum_{i=t_2}^{|\mathcal{Z}|} \widehat{x}_{\sigma_i} \\
= & - (\pi_{\sigma_{t_2}} - \pi''_{\sigma_{t_2}}) \sum_{i=t_1+1}^{t_2-1} \widehat{x}_{\sigma_i} - (\pi_{\sigma_{t_2}} + a_{t-\beta})\widehat{x}_{\sigma_{t_2}} \\
\geq & - \left[(\pi_{\sigma_{t_2}} - \pi''_{\sigma_{t_2}})(t_2 - t_1 - 1) + \pi_{\sigma_{t_2}} + a_{t-\beta} \right] \widehat{x}_{\sigma_{t_2}} = 0.
\end{aligned}$$

Therefore, $t_2 = \lceil \widetilde{t} \rceil$ at optimality. It follows from constraint (B.7) that at optimality

$$\pi^*_{\sigma_{t_2}} = -a_{t-\beta}(\lceil \widetilde{t} \rceil - \widetilde{t}).$$

Finally, suppose that $\widetilde{t}$ is an integer. Then, following a similar analysis as above, we obtain that either (i) $t_1 = \widetilde{t} - 1$, $t_2 = \widetilde{t}$, $\pi^*_{\sigma_{t_2}} = 0$ or (ii) $t_1 = \widetilde{t}$, $t_2 = \widetilde{t} + 1$, $\pi^*_{\sigma_{t_2}} = -a_{t-\beta}$. Both cases coincide with the claim of this proposition.

(B) Since $\pi_{\sigma_t} + a_{t-\beta} > 0$, there exists a $0 \leq t' \leq t$ such that

$$0 = \pi_{\sigma_1} = \cdots = \pi_{\sigma_{t'}} \geq \pi_{\sigma_{t'+1}} = \cdots = \pi_{\sigma_{|\mathcal{Z}|}}.$$

If $t' = t$ then $\sum_{i=1}^{t} \pi_{\sigma_i} = 0$, which violates constraint (B.7). Hence, $t' \leq t - 1$ and a valid choice of π_{σ_t} satisfies

$$-a_{t-\beta+1} > \pi_{\sigma_t} = -\frac{1}{(t - t')} \sum_{i=\beta+1}^{t} a_{i-\beta} > -a_{t-\beta}.$$

Since $\sum_{i=\beta+1}^{t} a_{i-\beta} = f(t - \beta) = (t - \beta)a_{t-\beta} + b_{t-\beta}$ and $\sum_{i=\beta+1}^{t} a_{i-\beta} = f(t - \beta) =$

$(t-\beta)a_{t-\beta+1}+b_{t-\beta+1}$, these inequalities are equivalent to

$$\beta-\frac{b_{t-\beta+1}}{a_{t-\beta+1}}<t'<\beta-\frac{b_{t-\beta}}{a_{t-\beta}}.$$

Accordingly, at optimality $\pi^*_{\sigma_{t'+1}}$ equals $-\dfrac{\sum_{i=\beta+1}^{t}a_{i-\beta}}{(t-t')}\equiv-\dfrac{f(t-\beta)}{(t-t')}$.

□

Proposition 36. For all $\pi\in\Pi^*$, it holds that

$$\max_{T\in\{0\}\cup[|\mathcal{Z}|]}\left\{f\left((T-\beta)^+\right)+\sum_{i=1}^{T}\pi_{\sigma_i}\right\}=\min_{(\ell,\rho)\in L_1\cup L_2}\max_{T\in\{0\}\cup[|\mathcal{Z}|]}\left\{h_{(\ell,\rho)}(T)+\sum_{i=1}^{T}\pi_{\sigma_i}\right\}. \quad \text{(B.8)}$$

Proof. Since $f((T-\beta)^+)=\min_{(\ell,\rho)\in L_1\cup L_2}h_{(\ell,\rho)}(T)$ by Lemma 9, we switch the order of maximization and minimization to obtain

$$\begin{aligned}\max_{T\in\{0\}\cup[|\mathcal{Z}|]}\left\{f\left((T-\beta)^+\right)+\sum_{i=1}^{T}\pi_{\sigma_i}\right\}&=\max_{T\in\{0\}\cup[|\mathcal{Z}|]}\left\{\min_{(\ell,\rho)\in L_1\cup L_2}h_{(\ell,\rho)}(T)+\sum_{i=1}^{T}\pi_{\sigma_i}\right\}\\&\leq\min_{(\ell,\rho)\in L_1\cup L_2}\left\{\max_{T\in\{0\}\cup[|\mathcal{Z}|]}h_{(\ell,\rho)}(T)+\sum_{i=1}^{T}\pi_{\sigma_i}\right\}.\end{aligned}$$

It remains to prove the other direction of (B.8). $\pi\in\Pi^*$ implies that $\pi\in\Pi^*_t$ for a $t\in\mathbb{Z},\beta+1\leq t\leq|\mathcal{Z}|$. We discuss the following cases to finish the proof.

1. If $b_{t-\beta}-\beta a_{t-\beta}\geq 0$, then by definition of Π^*_t

$$\pi_{\sigma_1}=\pi_{\sigma_2}=\cdots=\pi_{\sigma_{|\mathcal{Z}|}}=-a_{t-\beta}.$$

It follows that

$$\begin{aligned}\min_{(\ell,\rho)\in L_1\cup L_2}\max_{T\in\{0\}\cup[|\mathcal{Z}|]}\left\{ h_{(\ell,\rho)}(T)+\sum_{i=1}^{T}\pi_{\sigma_i}\right\} &\le \max_{T\in\{0\}\cup[|\mathcal{Z}|]}\left\{ \left(\varphi_{t-\beta}(T-\beta)\right)^+ +\sum_{i=1}^{T}\pi_{\sigma_i}\right\}\\ &= \max_{T\in\{0\}\cup[|\mathcal{Z}|]}\left\{ a_{t-\beta}(T-\beta)+b_{t-\beta}-Ta_{t-\beta}\right\}\\ &= \sum_{i=1}^{\beta}(-a_{t-\beta})+\sum_{i=\beta+1}^{t}(a_{i-\beta}-a_{t-\beta})\\ &= \sum_{i=1}^{\beta}\pi_{\sigma_i}+\sum_{i=\beta+1}^{t}(a_{i-\beta}+\pi_{\sigma_i})\\ &= \max_{T\in\{0\}\cup[|\mathcal{Z}|]}\left\{ f\left((T-\beta)^+\right)+\sum_{i=1}^{T}\pi_{\sigma_i}\right\},\end{aligned}$$

where the inequality is because we select $(\ell,\rho)\in L_1$ with $\ell=t-\beta$ and $\rho=b_{t-\beta}/a_{t-\beta}$, the first equality is because $b_{t-\beta}-\beta a_{t-\beta}\ge 0$, the second equality is because $b_{t-\beta}-\beta a_{t-\beta}=b_{t-\beta}+(t-\beta)a_{t-\beta}-ta_{t-\beta}=\sum_{i=\beta+1}^{t}a_{i-\beta}-ta_{t-\beta}$, and the last equality is because $\pi\in\Pi_t^*\subseteq\Pi_t$.

2. If $b_{t-\beta}-\beta a_{t-\beta}<0$ then we discuss the following two cases.

   (A) If $0=\pi_{\sigma_1}=\cdots=\pi_{\sigma_{\lfloor\widetilde{t}\rfloor}}\ge\pi_{\sigma_{\lceil\widetilde{t}\rceil}}=-a_{t-\beta}(\lceil\widetilde{t}\rceil-\widetilde{t})>\pi_{\sigma_{\lceil\widetilde{t}\rceil+1}}=\cdots=\pi_{\sigma_{|\mathcal{Z}|}}=-a_{t-\beta}$ for $\widetilde{t}\equiv\beta-\frac{b_{t-\beta}}{a_{t-\beta}}$, then

$$\begin{aligned}&\min_{(\ell,\rho)\in L_1\cup L_2}\max_{T\in\{0\}\cup[|\mathcal{Z}|]}\left\{ h_{(\ell,\rho)}(T)+\sum_{i=1}^{T}\pi_{\sigma_i}\right\}\\ \le\ &\max_{T\in\{0\}\cup[|\mathcal{Z}|]}\left\{ \left(a_{t-\beta}(T-\beta)+b_{t-\beta}\right)^+ +\sum_{i=1}^{T}\pi_{\sigma_i}\right\}\\ =\ &0\\ =\ &\max\left\{ \sum_{i=1}^{\beta}\pi_{\sigma_i}+\sum_{i=\beta+1}^{t}(a_{i-\beta}+\pi_{\sigma_i}),0\right\}\\ =\ &\max_{T\in\{0\}\cup[|\mathcal{Z}|]}\left\{ f\left((T-\beta)^+\right)+\sum_{i=1}^{T}\pi_{\sigma_i}\right\},\end{aligned}$$

   where the inequality is because we select $(\ell,\rho)\in L_1$ with $\ell=t-\beta$ and

$\rho = b_{t-\beta}/a_{t-\beta}$, the first equality is because

$$\begin{aligned}\big(a_{t-\beta}(T-\beta)+b_{t-\beta}\big)^+ &= \big(Ta_{t-\beta}+b_{t-\beta}-\beta a_{t-\beta}\big)^+ \\ &= \Big(T-\widetilde{t}\Big)^+ a_{t-\beta} \;=\; -\sum_{i=1}^{T}\pi_{\sigma_i}\end{aligned}$$

for all $T \in \{0\} \cup [|\mathcal{Z}|]$, the second equality is because $\sum_{i=\beta+1}^{t} a_{i-\beta} = (t-\widetilde{t})a_{t-\beta}$, and the last equality is because $\pi \in \Pi_t^* \subseteq \Pi_t$.

(B) If $0 = \pi_{\sigma_1} = \cdots = \pi_{\sigma_{t'}} > \pi_{\sigma_{t'+1}} = \cdots \pi_{\sigma_{|\mathcal{Z}|}} = -\dfrac{f(t-\beta)}{(t-t')}$ for a $t' \in \Big(\beta - \frac{b_{t-\beta+1}}{a_{t-\beta+1}}, \beta - \frac{b_{t-\beta}}{a_{t-\beta}}\Big) \cap \mathbb{Z}_+$, then

$$\begin{aligned}&\min_{(\ell,\rho)\in L_1\cup L_2}\max_{T\in\{0\}\cup[|\mathcal{Z}|]}\left\{\, h_{(\ell,\rho)}(T)+\sum_{i=1}^{T}\pi_{\sigma_i}\,\right\} \\ \le\; &\max_{T\in\{0\}\cup[|\mathcal{Z}|]}\left\{\, h_{(t-\beta,\beta-t')}(T)+\sum_{i=1}^{T}\pi_{\sigma_i}\,\right\} \\ =\; &\max_{T\in\{0\}\cup[|\mathcal{Z}|]}\left\{\, \left(\frac{f(t-\beta)}{t-t'}(T-t')\right)^+ +\sum_{i=1}^{T}\pi_{\sigma_i}\,\right\} \\ =\; &0 \\ =\; &\max\left\{\, \sum_{i=1}^{\beta}\pi_{\sigma_i}+\sum_{i=\beta+1}^{t}(a_{i-\beta}+\pi_{\sigma_i}), 0\,\right\} \\ =\; &\max_{T\in\{0\}\cup[|\mathcal{Z}|]}\left\{\, f\left((T-\beta)^+\right)+\sum_{i=1}^{T}\pi_{\sigma_i}\,\right\},\end{aligned}$$

where the inequality is because we select $(\ell,\rho) \in L_2$ with $\ell = t-\beta$ and $\rho = \beta - t'$, the second equality is due to the definition of π, the third equality is because $\sum_{i=\beta+1}^{t} a_{i-\beta} = f(t-\beta)$, and the last equality is because $\pi \in \Pi_t^* \subseteq \Pi_t$.

□

B.4 Proof of Theorem 11

Proof. For all $k \in \Gamma$, let $\mathcal{S}_k := \mathcal{S}_{\mathbb{B}} \cap \{(x,\gamma,z) : \gamma = r_k\}$. We simultaneously lift variables $\gamma_2, \ldots, \gamma_{|\Gamma|}$ back into inequality (3.17). To this end, we search for lifting coefficients

$\alpha_2, \ldots, \alpha_{|\Gamma|}$ such that

$$\max_{j\in[J]} \left\{ \frac{z_j}{d^j} \right\} \geq \sum_{j=1}^{J} (\nu_1^j - \nu_1^{j-1})(\tau_1^j - x^\mathsf{T}\bar{\xi}^j) + \sum_{k\in\Gamma} \alpha_k\gamma_k, \quad \forall (x,\gamma,z) \in \mathcal{S}_{\mathbb{B}},$$

where $\alpha_1 := 0$. Since $\sum_{k\in\Gamma} \gamma_k = 1$, fixing $\gamma_k = 1$ automatically forces all other γ_k's to be zero. Hence, the lifted inequality is valid if and only if

$$\max_{j\in[J]} \left\{ \frac{z_j}{d^j} \right\} \geq \sum_{j=1}^{J} (\nu_1^j - \nu_1^{j-1})(\tau_1^j - x^\mathsf{T}\bar{\xi}^j) + \alpha_k\gamma_k, \quad \forall (x,\gamma,z) \in \mathcal{S}_k,\ \forall k \in \Gamma.$$

If $k = 1$, this validity requirement is valid because $\alpha_1 = 0$ and the inequality reduces to (3.17). If $k \in \Gamma \setminus \{1\}$, the validity requirement is valid if and only if

$$\begin{aligned} \alpha_k \leq \min_{x,z}\ & \max_{j\in[J]} \left\{ \frac{z_j}{d^j} \right\} - \sum_{j=1}^{J} (\nu_1^j - \nu_1^{j-1})(\tau_1^j - x^\mathsf{T}\bar{\xi}^j), \\ \text{s.t.}\ & \frac{z_j}{d^j} \geq -\frac{r_k + d_0^j}{d^j} - x^\mathsf{T}\bar{\xi}^j, \quad \forall j \in [J], \\ & x \in \{0,1\}^n, z \in \mathbb{R}^J_-. \end{aligned} \tag{Lifting}$$

For fixed $x \in \{0,1\}^n$, $z_j \leq 0$ and $d^j < 0$ imply

$$\begin{aligned} \frac{z_j}{d^j} \geq \left(-\frac{r_k + d_0^j}{d^j} - x^\mathsf{T}\bar{\xi}^j \right)^+ &= \left(\nu_k^j + \tau_k^j - x^\mathsf{T}\bar{\xi}^j - 1 \right)^+ \\ &= \nu_k^j \left(\tau_k^j - x^\mathsf{T}\bar{\xi}^j \right)^+ + (1 - \nu_k^j) \left(\tau_k^j - x^\mathsf{T}\bar{\xi}^j - 1 \right)^+ \end{aligned}$$

for all $j \in [J]$. Let $\Upsilon_k := \max \left\{ \tau_k^j - x^\mathsf{T}\bar{\xi}^j \colon j \in [J] \right\}$, $\mathcal{J}_k$ be the corresponding set of maximizers, and $i \in \mathcal{J}_k$ be a maximizer with the largest ν_k^j, i.e.,

$$\begin{aligned} \mathcal{J}_k &:= \left\{ j \in [J] \colon \tau_k^j - x^\mathsf{T}\bar{\xi}^j = \Upsilon_k \right\} \\ \text{and} \quad i &\in \operatorname{argmax} \left\{ \nu_k^j \colon j \in \mathcal{J}_k \right\}. \end{aligned}$$

Then, it is easy to verify that

$$\max_{j\in[J]} \left\{ \frac{z_j}{d^j} \right\} = \nu_k^i \left(\tau_k^i - x^\mathsf{T}\bar{\xi}^i \right)^+ + (1 - \nu_k^i) \left(\tau_k^i - x^\mathsf{T}\bar{\xi}^i - 1 \right)^+$$

at optimality of the (Lifting) problem. Specifically, if $\Upsilon_k \leq 0$, then $(\tau_k^j - x^\mathsf{T}\bar{\xi}^j) \leq 0$ for all $j \in [J]$ and so $\max_{j\in[J]} \left\{ z_j/d^j \right\} = 0$. This implies a lower bound for the objective value of (Lifting):

$$\begin{aligned}\max_{j\in[J]} \left\{ \frac{z_j}{d^j} \right\} - \sum_{j=1}^{J} (\nu_1^j - \nu_1^{j-1})(\tau_1^j - x^\mathsf{T}\bar{\xi}^j) &= \sum_{j=1}^{J} (\nu_1^j - \nu_1^{j-1})(x^\mathsf{T}\bar{\xi}^j - \tau_1^j) \\ &\geq \sum_{j=1}^{J} (\nu_1^j - \nu_1^{j-1})(\tau_k^j - \tau_1^j).\end{aligned}$$

If $\Upsilon_k \geq 1$ then $\max_{j\in[J]} \left\{ z_j/d^j \right\} = (\tau_k^i - x^\mathsf{T}\bar{\xi}^i - 1) + \nu_k^i$, and the lower bound for the objective value of (Lifting) becomes

$$\begin{aligned}&\max_{j\in[J]} \left\{ \frac{z_j}{d^j} \right\} - \sum_{j=1}^{J} (\nu_1^j - \nu_1^{j-1})(\tau_1^j - x^\mathsf{T}\bar{\xi}^j) \\ &= (\tau_k^i - x^\mathsf{T}\bar{\xi}^i - 1) + \nu_k^i - \sum_{j=1}^{J} (\nu_1^j - \nu_1^{j-1})(\tau_k^j - x^\mathsf{T}\bar{\xi}^j) - \sum_{j=1}^{J} (\nu_1^j - \nu_1^{j-1})(\tau_1^j - \tau_k^j) \\ &\geq (\tau_k^i - x^\mathsf{T}\bar{\xi}^i - 1) + \nu_k^i - \sum_{j\in\mathcal{J}_k} (\nu_1^j - \nu_1^{j-1})(\tau_k^i - x^\mathsf{T}\bar{\xi}^i) - \sum_{j\notin\mathcal{J}_k} (\nu_1^j - \nu_1^{j-1})(\tau_k^i - x^\mathsf{T}\bar{\xi}^i - 1) \\ &\quad - \sum_{j=1}^{J} (\nu_1^j - \nu_1^{j-1})(\tau_1^j - \tau_k^j) \\ &= (\tau_k^i - x^\mathsf{T}\bar{\xi}^i - 1)(1 - \nu_1^J) + \nu_k^i - \sum_{j\in\mathcal{J}_k} (\nu_1^j - \nu_1^{j-1}) + \sum_{j=1}^{J} (\nu_1^j - \nu_1^{j-1})(\tau_k^j - \tau_1^j) \\ &\geq \nu_k^i - \nu_1^J + \sum_{j=1}^{J} (\nu_1^j - \nu_1^{j-1})(\tau_k^j - \tau_1^j),\end{aligned}$$

where the first inequality is because $\tau_k^j - x^\mathsf{T}\bar{\xi}^j \leq \tau_k^i - x^\mathsf{T}\bar{\xi}^i - 1$ if $j \notin \mathcal{J}_k$ and the second inequality is because $\tau_k^i - x^\mathsf{T}\bar{\xi}^i = \Upsilon_k \geq 1$, $\nu_1^J \leq 1$, and $\sum_{j\in\mathcal{J}_k} (\nu_1^j - \nu_1^{j-1}) \leq \sum_{j=1}^{J} (\nu_1^j - \nu_1^{j-1}) = \nu_1^J$. The two cases of the value of Υ_k together suggests the following valid choice of α_k:

$$\alpha_k = \min_{j\in[J]} \min \left\{ \nu_k^j - \nu_1^J, 0 \right\} + \sum_{j=1}^{J} (\nu_1^j - \nu_1^{j-1})(\tau_k^j - \tau_1^j).$$

But $\alpha_k = 0$ is also a valid choice because, in that case, the lifted inequality reduces to (3.17), which is valid for $\mathcal{S}_\mathbb{B} \supseteq \mathcal{S}_k$. Therefore, we can take the maximum of these two valid choices. The conclusion follows. □

APPENDIX C

Locating Charging Stations for Battery Electric Buses

C.1 Nomenclature of Chapter 4

Sets

$\mathcal{L}$ Set of all directed links connecting sequential stations, indexed by (i,j).

$\mathcal{L}_k$ Set of the directed links forming bus line k.

$\mathcal{N}$ Set of all stations/nodes in the bus network, i.e., the set of candidate locations to build charging facilities, indexed by i and j.

$\mathcal{N}_k$ Set of the bus stations on bus line k.

Parameters

a_{it}^{chgr} Amortized investment cost of installing a type-t charging facility at bus station i.

a^{bat} Amortized battery investment cost per kWh, considering replacement and the discount rate.

c_{kij} Energy consumption of a line-k BEB on the link (i,j).

$\bar{\epsilon}$ Maximal remaining battery power (in percentage of the battery size).

$\underline{\epsilon}$ Minimal battery power (in percentage of the battery size).

K Number of bus lines.

o_k A node representing the base station of bus line k.

p_t Power of a type-t charging facility.

T Number of charging facility types.

τ_{ki} Dwell time of a line-k BEB at station i.

ζ_k Total number of buses of line k.

Variables

e_{ki} Remaining battery power of a line-k BEB when it leaves station i.

e_k^{max} Battery size a BEB on line k, in the unit of kWh.

x_{it} A binary variable indicating whether to build a type-t fast-charging facility at bus station i (If yes, then $x_i = 1$; otherwise $x_i = 0$).