SERVICE SYSTEM DESIGN PROBLEMS UNDER DEMAND UNCERTAINTY

by

Nazanin Madani

Submitted in partial fulfillment of the requirements
for the degree of Master of Applied Science

at

Dalhousie University
Halifax, Nova Scotia
August 2020

© Copyright by Nazanin Madani, 2020

Table of Contents

List of Tables

List of Figures

Abstract

The service system design problem aims to select the location and capacity of service facilities and customers' assignments to minimize the setup, access, and waiting time costs. This thesis addresses the case when there is uncertainty about the demand for service, considering two service systems that can be modelled as independent networks of M/M/1 and G/M/1 queues. Robust optimization, with both Budgeted and Ellipsoidal uncertainty sets, is used when the demand rate is unknown. However, the arrival pattern can still be reasonably approximated as a Poisson process or follow a General distribution, respectively. We use distributionally robust optimization with a Wasserstein ambiguity set to address the case when the demand distribution is estimated from a limited sample. For both models, we can reformulate both the robust and distributionally-robust problems as mixed-integer second-order conic programs. For the M/M/1 model, these problems can be solved directly on commercial solvers, even though the nominal problem has a non-convex cost function. Extensive numerical experiments on benchmark test instances are conducted to compare the different approaches used to handle uncertainty and investigate the effect of problem size and parameters. On the other hand, for the G/M/1 model, we use a Lagrangian-Relaxation approach to solve the problems and conduct numerical experiments based on small instances to confirm the validity of the proposed reformulations.

Acknowledgements

I wish to thank all the people whose assistance was a milestone in the completion of this thesis, especially my supervisor Dr. Ahmed Saif for his guidance and support. Furthermore, I would like to thank Dr. Claver Diallo and Dr. Alexander Engau as my committee members for their interests and feedback. Thank you to the IndusTrial Engineering Department at Dalhousie University for providing an excellent environment to work on my research.

I can not thank my husband and my family enough for their help, incentive, and patience throughout these two years. Everything would have been much harder for me without their support.

Chapter 1

Introduction

This thesis focuses on the service system design problem (SSDP), also known in the literature as the facility location problems with immobile (fixed) servers, stochastic demands, and congestion [17]. Besides the setup and transportation/access costs considered in classical facility location problems, the implicit cost of customers' waiting time for service is an integral part, and a distinctive feature, of service system design problems. Generally speaking, the service system design problem aims to locate service centers (SCs), determine their capacities, and assign customers to those centers to minimize the total cost, including the costs of installing and accessing the SCs and the queuing delay costs. This problem arises in different planning contexts, such as locating emergency medical centers [60], grocery stores, government offices, refuse collection and disposal centers, and designing private communication networks.

In an era when the service sector represents approximately 65% of the global GDP [2], optimizing the design and operation of service systems has become a task of paramount importance. Apart from their economic role, service systems impact our everyday experiences. A tangible example of this can be patient wait times in Nova Scotia (NS), Canada. According to a new study conducted by the Fraser Institute [1, 8], an independent Canadian public policy research and educational organization, patients in Nova Scotia, in 2019, waited 33.3 weeks from referral by a family doctor to treatment. This study shows a 190% increase in wait times from 1993 to 2019. The authors estimate the economic cost of this wait time to be $134 million dollars for a median wait time of 17.1 weeks after seeing a specialist. Additionally, this study shows that NS has higher wait time costs than some other provinces, including some with larger populations (*e.g.,* Saskatchewan), and NS has the largest wait-list size for health services in Canada amounting to 5.8% of its total population.

Therefore, designers of service systems strive to balance economic considerations, such as cost and server utilization, with service quality considerations, such as availability and waiting times, to maximize customers' satisfaction. In particular, avoiding excessive waiting times has been considered a primary objective in the literature [5, 6, 32]. Another way to establish this balance is to minimize the costs of providing the service and add a constraint that puts a minimum threshold on the service quality [46, 49]. This thesis seeks to design a service system using the first approach, *i.e.,* minimizing the total cost, including the service quality cost. Given that customers' arrival in most cases is random and uncontrollable, installing a sufficient service capacity in SCs and cleverly allocating customers to these centers is crucial for avoiding congested systems and thus dissatisfied customers. There are two approaches suggested in the literature to incorporate the service capacity into the model as a decision variable. In some references, a finite set of different capacity levels is considered [3, 32, 54], and in others it is assumed to be a continuous real-valued decision variable [23, 33, 56]. In this thesis, the second approach is chosen.

Moving to the customers' allocation to the service facilities, two scenarios can be considered: user-choice or direct-choice allocations. In a user-choice allocation, customers will choose a facility in a *utility-maximizing* fashion [18]. In contrast, in a direct-choice allocation, the goal is to allocate the customers to the facilities in a way that the total cost of the system is minimized, which includes the setup, access, and queueing delay costs [32, 54, 33]. Our models in this thesis belong to the second category.

The problem is further aggravated by the facts that service systems are usually designed under considerable uncertainty about the future demand for service, and waiting times in queuing systems are quite sensitive to the arrival rate (*i.e.,* demand) of customers, especially when the servers' utilization factor is close to unity. The combined effect of these factors makes the practice of ignoring the uncertainty in arrival rates and designing service systems based solely on the *expected/most likely* arrival rates, often leading to undesired outcomes. For instance, in a single-server facility with Markovian arrival and service patterns (*i.e.,* modelled as an M/M/1 queuing system) and an estimated utilization factor of 90%, if the *real* arrival rate turns out to have been underestimated by 5%, the average waiting time in the system will be almost double of what has been originally estimated. On the other hand, the facility/queuing system will become unstable if the arrival rate turns out to have been underestimated by 10% or more. Therefore, with the inevitable presence of uncertainty about future demand, one is motivated to utilize a robust approach to design the service system. To address this demand uncertainty, we focus on M/M/1 and G/M/1 systems. We assume that the service provision at each SC can be modelled as a Poisson process with a finite rate. By using this assumption, as a service provider, we can control our service rate and have enough information to be sure about it. First, we consider the M/M/1 system as it is simple to use, and researchers have extensively studied this system in the literature. Besides, it provides a good approximation when the utilization factor is high. However, it is unlikely to be sure about the arrival pattern in real-life cases, whether it is Markovian or not. Thus, that is why we also study the G/M/1 system.

In this thesis, we address the issue of demand uncertainty in service systems and try to mitigate its impacts on the economic and service quality metrics by proposing two robust frameworks for service systems design. In both frameworks, we assume that the system designer has access to a finite number of independent future demand scenarios, which could be based on historical data or experts' opinions. In the *robust optimization* framework, we use these scenarios to construct uncertainty sets of specific structures, and then optimize the total cost of the service system while considering the worst-case demand *realization* in the uncertainty set. Two uncertainty set structures are considered: ellipsoidal and budgeted uncertainty sets. In the *distributionally-robust optimization* framework, we assume that the system designer aims to minimize the worst-case expected value of the service system's total cost, where the expectation is taken with respect to the worst-case probability distribution among a *distributional ambiguity set* that is constructed based on the future demand scenarios. In particular, the distributional ambiguity set includes all probability distributions that are within a specific *distance* from the empirical distribution constructed from the sample data points, where the distance between probability distributions is measured using a *1-Wasserstein metric*. We compare the solutions and objective values of the two frameworks against a nominal problem that assumes that the demand is certain. In all cases, the total cost includes the setup cost of servers, the access cost of customers to service facilities, and the implicit cost of customers' waiting time in the system. The mathematical models corresponding to all considered cases were reformulated as mixed-integer second-order conic programs. For the M/M/1 case, these reformulated programs can be directly handled using commercial solvers, whereas, for the G/M/1 model, Lagrangian relaxation and decomposition techniques are used to solve these programs.

The remainder of this thesis is organized as follows. The next chapter provides brief reviews of the SSDP and the frameworks for decision making under uncertainty utilized in this work. Chapter 3 describes the service system design problems which can be modelled as a network of independent M/M/1 queues. It also presents the formulations for the nominal, robust, and distributionally robust optimization problems. Chapter 4 has the same scheme as Chapter 3, but it studies the service system design problems that can be modelled as a network of independent G/M/1 queues. Chapter 5 presents the experiments performed on all the models and the results. Chapter 6 concludes this thesis and proposes ideas for future research. Throughout this thesis, we use upright characters for vectors and italicized characters for scalars.

Chapter 2

Literature Review

This thesis focuses on the SSDP. This Chapter provides a brief review of this problem and the frameworks for decision making under uncertainty used in this work.

2.1 Service System Design Problem

Among the first studies that addressed the SSDP is the work of Amiri [5], which considers a basic setting in which the arrival of customers' demands can be modelled as a Poisson process, whereas the service times in each SC are independently and identically distributed according to an exponential distribution. Hence, this problem can be modelled as a network of independent M/M/1 queues, where the decision variables to be determined are the number, locations, and capacities of SCs. The author applied the proposed model, which is presented as an integer programming problem, to design a telecommunication network. Still, it can also be used by planners to design other types of service systems. In the model proposed in [5], the number, locations, and capacities of service systems are decision variables that need to be determined, and the waiting time (queueing) cost is incorporated in the total cost that should be minimized. The contribution of [5] is to present a realistic model for the SSDP and develop an effective solution procedure for the problem. Later, Amiri [6] emphasized the importance of considering a back-up service in a reliable SSDP, which means customers are assigned to a primary and secondary or back-up facility, and this assumption is added to the formulation of the basic model proposed in [5]. Also, Amiri [7] considered the same basic model in [5] under the time-varying demand conditions as the demand requirements of the customers could vary during different busy-hours. Amiri's basic model [5] was later extended by Wang *et al.* [55]. In contrast to the centrally located customers in [5], they assumed that customers are free to choose and will logically choose the closest open SC. This *closest assignment* assumption was enforced by adding an explicit constraint. They also included restrictions on the maximum expected waiting time at any open facility and the number of facilities to be opened. Later, Wang *et al.* [56] proposed several models for locating the facilities subject to congestion. Contrary to their model presented in [55], Wang *et al.* [56] considered this problem from both the service provider and the customers' perspective together. Thus, the key point in [56] is balancing the service costs against service quality, which can be measured through travel and service time delays.

Assumptions of the M/M/1 queuing model might become quite restrictive for real-life situations. For example, while customers' arrival to a SC is usually random (*i.e.,* Markovian), the service time is often quite controllable and thus can have a general probability distribution. To address this case, Vidyarthi and Jayaswal [54] modelled the SSDP as an M/G/1 queueing network and proposed an exact (ϵ-optimal) algorithm to solve it. Furthermore, SCs typically have multiple parallel servers. Therefore, treating each SC as an M/M/1 queue, in this case, is just an approximation that becomes better as the utilization factor approaches one. Castillo *et al.* [23] studied the SSDP considering two capacity choice scenarios: the situation where each open facility has one server whose service rate can be any positive number, and the situation where the number of parallel servers at each open facility can be any positive integer but the service rate per server is fixed. Besides, the second scenario uses approximations for the expected number of customers and the optimal number of servers for two reasons. First, as the exact performance expressions can only be defined for the number of servers with integer values, they solve the problem's continuous relaxation. Second, there are no exact results that allow them to express the optimal number of servers at each facility in a closed-form. Moreover, these approximations lead to the expressions that they obtained for single-server facilities. Hence, they express the optimal service rate and the optimal number of servers in closed-forms in the first and second scenarios. As a result, they were able to eliminate both the service rates and the number of servers from their models and tractably formulated them as mixed-integer nonlinear programs. Besides, they showed that the problems for both scenarios are structurally identical, which implies that the facilities with multiple servers can be modelled and compared with single-server facilities. Similarly, Syam [53] developed and solved a comprehensive nonlinear location-allocation model for SSDP that incorporates several relevant costs and considerations, including, access, service, and waiting costs, and queueing considerations such as multiple servers, multiple order priority levels, multiple service sites, and service distance limits.

All of the previous models considered the demands to be inelastic. Still, Aboolian *et al.* [3] studied the problem of maximizing the overall profit of a system while considering the elasticity of demand. Their models belong to the class of location models with immobile servers with equilibrium constraints. They modelled the problem as a network of M/M/1 and M/M/s queues separately. This work can be applied for finding exact optimal solutions for large-scale instances when they separated capacity assignment from the customer assignment and location subproblems. Berman and Kaplan [16] were the first to explicitly model demand losses resulting from the elasticity while considering the travel distance and congestion for single-facility systems. Besides, they assumed that a finite set of facility locations are given, and they did not impose any service level constraints in their model. Berman *et al.* [15] also considered distance-sensitive demand models, but in contrast to Aboolian *et al.* [3] , no equilibrium-type constraints were imposed in their work. Compared to Berman and Kaplan [16], Berman *et al.* [15] assumed that the facilities could be located at any point in the network. Moreover, they included a service level constraint in their model and considered opening more than one facility. Marianov *et al.* [44] and Marianov *et al.* [45] also focused on these types of problems without any explicit equilibrium constraints. Later, Zhang *et al.* [60], studied a multilocation model with elastic demand and congestion, in which they modelled each facility as an M/M/1 queue. They used the total time (travel, waiting, and service) as a proxy for accessibility, and assumed that customers at the same demand zones would choose the same facility with the minimum total time. However, this assumption prevented them from identifying an equilibrium allocation of customers to facilities. Later, Zhang *et al.* [59] extended the work of Zhang *et al.* [60] by incorporating the possibility that customers from the same demand zones can patronize different facilities, which usually guarantees the existence of an equilibrium allocation and result in a completely different modelling approach. In contrast to Aboolian's work, the objective of these two papers is maximizing accessibility.

In designing a service system network, the location of SCs has a significant impact on the congestion at each of them, and affects the quality of service. The locations of facilities should be determined in a way such that they would be accessible from demand zones within a reasonable time. Besides, customers' waiting time should also be as short as possible, *i.e.,* SCs should have sufficient capacities. As a result, ensuring both convenience and enough capacity should be considered while designing a service system network. To address both considerations, Marianov and Serra [46] presented several probabilistic, maximal covering, location-allocation (MCL) models for congested systems. In their first model, they considered an M/M/1 queueing system that addresses the issue of the location of a given number of facilities so that the maximum number of customers is served within a standard distance. Then, in their second model, they formulated several maximal coverage models, using one or more servers per SC. Using probabilistic constraints, these models restrict either the response time or the queue length to be smaller than a predetermined value. The contribution of these models is that the value of service quality can be explicitly observed in the optimization model. Thus, when designing a system, these models would allow the designer to trade off investment and operating cost versus service quality. Marianov *et al.* [47] used the same design scheme but for the hierarchical location-allocation models in which facilities at different levels provide different types of services. In [46], either the number or the capacity of the facilities (or both) are assumed to be fixed. The demand arrivals are supposed to be Poisson, and the service time follows an exponential distribution. In contrast to [46], Baron *et al.* [9] worked on models with the general spatial distribution of the demand arrivals and service processes, without fixing either the number or the capacity of the facilities, or their potential locations in advance. They also assumed that the demand arrivals are distributed over a certain space, such as a line, or a network. Like Wang *et al.* [55], Baron *et al.* [9] imposed the *closest assignment* assumption in their models and included a service-level constraint that limits the waiting times at facilities. The contribution of their work is defining a location vector, which ensures identical customer demand at all facilities. However, Castillo *et al.* [23] argued that this assumption is not appropriate for models with immobile servers, and it is essential to consider what information customers have about waiting time. Although customer choice processes are not incorporated explicitly in their models, their results show that customers choose a facility that is not close but has less waiting time.

All the models above treated the capacity cost as a linear function of service rate or the number of servers. However, in reality, capacity costs are often affected by economies-of-scale. Recently, Elhedhli *et al.* [33] considered the service system design problem with a general continuous capacity case and accounted for economies-of-scale in its cost through an increasing concave function. This problem is formulated as a non-linear mixed-integer program with linear constraints and an objective function with both convex and concave terms. Furthermore, two solution approaches were proposed. In the first, the problem was reformulated as a mixed-integer nonlinear program that could be approximated using piecewise linearization; whereas in the second, they used Lagrangian relaxation to decompose the problem and reformulate the subproblems as mixed-integer second-order cone programs.

In most cases listed above, the demand arrival process is usually assumed to be Poisson, and the service process is typically assumed to be exponential. Similarly, in this thesis, we first consider a SSDP that can be modelled as a network of independent

M/M/1 queues. On the other hand, by assuming a general distribution for demand arrivals, as opposed to Poisson distribution, we propose a more realistic and general case in which a system can be modelled as a network of independent G/M/1 queues. Generally speaking, from the Queuing Theory's perspective, predicting the mean waiting time or queue length in a steady-state condition could be very challenging when considering models with G/G/1 queues. Therefore, approximations are used to deal with this challenge. One of the widely used approximations for waiting time developed by Kingman [42], which can also be applied for G/M/1 and M/G/1 queues. Doshi [28] also studied the G/G/1 queues with vacations or setup times and developed a decomposition for the waiting time distribution. Later, Aden *et al.* [4] studied the G/M/1 queues with setup times and retrieved the waiting time decomposition result of Doshi [28]. Besides, they established a decomposition for the attained waiting time. The martingale techniques, transform techniques, and sample-path arguments are the methods they used in their work. Moreover, they made use of the duality between attained and virtual waiting time process. Chu and Ke [25] developed an estimation of mean response time for a G/M/1 queue using the Empirical Laplace function approach. They obtained an estimate of the response time by applying a data-based computation procedure.

2.2 Optimization Under Uncertainty

2.2.1 Robust Optimization (RO)

In all previous models, it has been assumed that the SC designer knows the demand arrival and the service rates with certainty. However, it is often the case that these parameters are uncertain and known only to lie in an *uncertainty set*. In such cases, it might be desirable to protect against this uncertainty in demand arrival or service time by employing a Robust Optimization (RO) approach that requires the constraints to hold for all realizations of the uncertain parameters within the uncertainty set, and minimizes the cost function corresponding to the worst-case among these realizations. So, this so-called Robust Counterpart (RC) can also be tractably reformulated as an optimization problem that depends on both the uncertainty set and objective function/constraints with uncertain parameters. The RO approach is particularly appealing when the probability distribution of the uncertain parameter is unknown; thus, it can be a good alternative for stochastic programming (SP).

Soyster [51] was the first to apply RO on a linear optimization model to generate a feasible solution for all the parameters that lie within a box, *i.e.,* hyper-cubic set. Although the box uncertainty sets are easy to handle and often lead to a deterministic problem, they are too conservative. Later, Ben-Tal and Nemirovski [12, 13, 14], and El-Ghaoui *et al.* [29, 30] made a significant improvement in addressing this over-conservatism by proposing an Ellipsoidal uncertainty set. Though this approach leads to a nonlinear model in the form of a conic quadratic problem and could be more expensive computationally, it is still convex. Moreover, Bertimas and Sim [20] introduced a new class of uncertainty set, referred to as the *budgeted uncertainty set* that preserves the linearity of the problem and allows the degree of conservatism to be fully controlled by selecting the uncertainty budget.

According to Charnes and Cooper [24], when a constraint is affected by an uncertain parameter, the goal is to satisfy that constraint with a certain probability, *e.g.,* $1-\epsilon$, where $\epsilon \geq 0$. Thus, a smaller ϵ applies more protection and assures that the constraint is satisfied for more realizations. It has been shown that chance constraints can be conservatively approximated using robust optimization, which enables uncertainty sets to be calibrated such that they provide a probabilistic guarantee of feasibility [20]. More precisely, for a given sample of observations of the uncertain parameter, we can define an uncertainty set that is large enough to contain $(1-\epsilon)\times 100\%$ of these realizations, and the solution obtained will be feasible in at least $(1-\epsilon)\times 100\%$ of cases, which makes this approximation conservative. In this thesis, we use two types of uncertainty set, *Budgeted* and *Ball*, and calibrate them using the chance constraint approximation.

2.2.2 Distributionally-Robust Optimization (DRO)

The RO approach assumes an *oblivious* decision-maker, *i.e.,* one that does not know the probability distribution of the uncertain parameter. Although RO solutions protect from extreme unfavourable scenarios, they are considered too conservative and often lead to poor expected performances. On the other extreme, if the decision-maker has access to sample data that enables reasonably accurate estimation of the uncertain parameters' *true* probability distribution, implementing a *risk-neutral* approach like SP might be a more favourable alternative. In reality, however, decision-makers often have small-size samples of reliable historical data or future predictions they can utilize. In such a case, implementing a classical SP might lead to substantial disappointments when implementing the solution obtained with *out-of-sample* data, an over-fitting phenomenon referred to as the *optimizers' curse* [50]. One way to overcome this issue is by considering a family of distributions (referred to as the *ambiguity set*) that contains the *true* probability distribution with a high probability, instead of a single distribution, when making a decision. With a *risk-averse* decision-maker who desires to protect itself against the worst-case distribution within the ambiguity set, this is called Distributionally Robust Optimization (DRO). In other words, DRO bridges SP and RO and serves as a unifying framework for them. More specifically, when the ambiguity set includes only a single (nominal/empirical) distribution, it reduces to SP. When it includes all the distributions supported on the uncertainty set, it reduces to RO.

Different classes of ambiguity sets have been considered in DRO, including moment-bases ambiguity sets that contain all distributions that satisfy certain moment constraints [27, 36, 58]. On the other hand, statistical-distance-based ambiguity sets are defined as balls in the space of probability distributions by using a probability distance function such as the *Prohorov metric* [34], the *Kullback-Leibler divergence*

[40], or the *Wasserstein metric*, also known as the *Kantorovich metric*, [48]. These statistical-distance-based ambiguity sets include all the distributions that are close enough to a nominal or most likely distribution with for the prescribed probability metric. In this case, the radius of the ambiguity set can be tuned, which means that the level of the conservatism of the optimization problem can be restrained. Hence, the ambiguity set is a vital ingredient of any distributionally robust optimization model.

In this thesis, Wasserstein ambiguity set is used as it has powerful properties that are demonstrated by Mohajerani Esfahani *et al.* [35] as follow: 1. *Finite Sample Gaurantee*: For a carefully chosen size of the ambiguity set, the optimal value of the DRO problem offers a confidence bound on the out-of-sample performance of the optimal solution of the DRO problem. 2. *Asymptotic Consistency*: As the number of realizations goes to infinity, the optimal value and the data-driven optimal solution converge to the optimal value and the optimal solution of the stochastic programming, respectively. 3. *Tractability*: For many objective functions and feasible sets, the DRO is computationally tractable. These properties were originally identified by Bertimas *et al.* [19] as desirable properties of data-driven solutions for stochastic programs. Moreover, the Wasserstein ambiguity set makes it possible to control the model's conservativeness and contains all the continuous and discrete distributions that are sufficiently close to a discrete empirical distribution (the center of the ambiguity set).

2.3 Research Gap and Contributions

All of the prior SSDP models assume that the demand arrival rates are known with certainty, which is not the case in a realistic situation. Among the few recent references that consider uncertainty in SSDP is the work of Juan Ma *et al.* [43], which studies a capacity planning problem for a service provider that process transactions arrival from its client, where the arrival rate is uncertain. In their work, they assume that the transaction arrival rate is uniformly distributed over a predefined interval, and propose a chance-constrained model as a standard M/G/1 queue. Considering the uniformly distributed assumption makes it possible for them to solve their model analytically.

In this thesis, we also focus on designing and configuring service systems when the customer arrival rate is uncertain. In contrast to [43], both RO and DRO approaches are considered to deal with the uncertainty. In the RO framework, two uncertainty sets structures are considered: Ball and Budgeted uncertainty sets. In the DRO framework, Wasserstein ambiguity set is used. Besides, we propose new mixed-integer second-order conic (MISOC) mathematical reformulations for all the considered cases. We model the SSDP as a network of independent M/M/1 and G/M/1 queues. For the M/M/1 case, all the proposed formulations can be solved directly using commercial solvers, whereas for the G/M/1 case, Lagrangian relaxation and decomposition techniques are used to solve these models as the structure of this problem is difficult to handle directly using commercial solvers.

Chapter 3

Service System Design Problems Modelled as a Network of M/M/1 Queues

3.1 Problem Description

Let $I := \{DZ_i\}_{i=1}^{m}$ be a set of demand zones and $J := \{SC_j\}_{j=1}^{n}$ be a set of potential SC locations. Each demand zone needs to be assigned to a single SC to satisfy its demand. We assume that a demand zone arrival to its assigned SC follows a Poisson process (*i.e.,* inter-arrival times of individual customers are exponential *i.i.d.* random variables) having rate $\xi_i,\ i \in I$. Initially, we will assume that these rates are known with certainty, whereas the uncertain case will be addressed later. Likewise, we assume that service provision at SC_j can be reasonably modelled as a Poisson process with a finite service rate (*i.e.,* capacity) μ_j, which is not known *a priori* but can be determined by the system designer. The single assignment assumption might not be optimal as the *Hakimi property* ([37, 38]) does not hold for the SSDP, but practical considerations might require it.

With that, each SC can be modelled as an M/M/1 queueing system, and the service system becomes a network of independent M/M/1 queues. The M/M/1 model is a single service facility with one server, infinite buffers to store demands for service, and the first-come-first-served queue discipline. The setup cost of SC is proportional to their service rate, *i.e.,* each service rate unit at SC_j costs f_j. There is a demand zone access cost c_{ij} per unit demand if DZ_i is assigned to SC_j. Furthermore, to discourage excessive waiting, customers' waiting time in the system is penalized at a constant rate of t per unit time. The objective is to determine the locations of SCs to open, their service capacities to install, and the assignment of customers to SC to minimize the total expected cost, which includes setup, access and waiting time costs. To formulate the problem, we use the following decision variables:

$$y_{ij} = \begin{cases} 1 & \text{if } DZ_i \text{ is assigned to } SC_j \\ 0 & \text{otherwise} \end{cases}$$

$$\mu_j = \text{service rate/capacity of } SC_j.$$

Therefore, the SSDP can be formulated as

$$[NP] : \min_{\mathrm{y},\mu} \quad \sum_{j\in J} f_j\mu_j + \sum_{j\in J}\sum_{i\in I} c_{ij}\xi_i y_{ij} + t\sum_{j\in J} \frac{\sum_{i\in I}\xi_i y_{ij}}{\mu_j - \sum_{i\in I}\xi_i y_{ij}} \tag{3.1a}$$

$$\text{s.t.} \quad \sum_{j\in J} y_{ij} = 1 \qquad \forall i \in I \tag{3.1b}$$

$$\sum_{i\in I}\xi_i y_{ij} \le \mu_j \qquad \forall j \in J \tag{3.1c}$$

$$y_{ij} \in \{0,1\}, \quad \mu_j \ge 0 \qquad \forall i \in I, \forall j \in J. \tag{3.1d}$$

The three terms in the objective function (3.1a) represent the capacity-dependent setup cost, the customers' access cost from demand zone i to SC_j, and their gross waiting time cost, respectively. constraints (3.1b) ensure that every demand zone is assigned to exactly one SC. Constraints (3.1c) guarantee that each demand zone is assigned to an open SC only and that the total demand arrival rate to the SC does not exceed its service capacity.

The proposed formulation results in a nonlinear mixed-integer program with linear constraints. When $\xi_i,\ i \in I$ is known with certainty, we refer to (3.1) as the *nominal problem (NP)*. At first glance, one might suspect that (3.1) is a convex optimization problem. However, careful examination reveals that it is not, and hence cannot be solved using classical convex optimization techniques. The following lemma states this observation. But for ease of exposition, let us first define the variable $s_j = \sum_{i\in I}\xi_i y_{ij},\ j \in J$.

Lemma 3.1. *In the domain* $\mathrm{s} \in [0, \mu]$*, the function* $f(\mathrm{s}, \mu) = \dfrac{\mathrm{s}}{\mu - \mathrm{s}}$ *is element-wise convex in* s *and* μ*, but not jointly convex in both.*

Proof. We know that a function $f(x)$ is convex if and only if

$$\frac{\partial^2 f}{\partial x^2} \geq 0, \tag{3.2}$$

and a function $f(x_1, x_2)$ is convex if and only if all the following conditions are met for all possible values of x_1 and x_2

$$\frac{\partial^2 f}{\partial x_1^2} \geq 0 \tag{3.3a}$$

$$\frac{\partial^2 f}{\partial x_2^2} \geq 0 \tag{3.3b}$$

$$\frac{\partial^2 f}{\partial x_1^2} . \frac{\partial^2 f}{\partial x_2^2} - \left[\frac{\partial^2 f}{\partial x_1 \partial x_2}\right]^2 \geq 0. \tag{3.3c}$$

Using the definitions (3.2) and (3.3), and knowing that $\mathrm{s} \leq \mu$, we will get

$$\frac{\partial^2 f}{\partial \mathrm{s}^2} = \frac{2\mu(\mu - \mathrm{s})}{(\mu - \mathrm{s})^4} \geq 0 \tag{3.4}$$

$$\frac{\partial^2 f}{\partial \mu^2} = \frac{2\mathrm{s}(\mu - \mathrm{s})}{(\mu - \mathrm{s})^4} \geq 0 \tag{3.5}$$

$$\frac{\partial^2 f}{\partial \mathrm{s}^2} . \frac{\partial^2 f}{\partial \mu^2} - \left[\frac{\partial^2 f}{\partial \mathrm{s} \partial \mu}\right]^2 = \frac{4\mathrm{s}\mu}{(\mu - \mathrm{s})^6} - \frac{(\mu + \mathrm{s})^2}{(\mathrm{s} - \mu)^6} - \frac{(\mu + \mathrm{s})^2}{(\mu - \mathrm{s})^6} = \frac{-2(\mu^2 + \mathrm{s}^2)}{(\mu - \mathrm{s})^6} \leq 0. \tag{3.6}$$

Now, from (3.4) and (3.5), we conclude that f is element-wise convex in s and μ but (3.6) does not satisfy (3.3c) which means that f is nonconvex. □

Before tackling the uncertain case, we begin by reformulating the nominal problem into a structure that is easier to handle.

3.1.1 Reformulation Into a Mixed-Integer Second-Order Conic Programming Problem

For a given feasible $\bar{y}$, let $\bar{s}_j = \sum_{i \in I} \xi_i \bar{y}_{ij}$. Thus, the nominal problem reduces to

$$\min_{\mu} \Big[\sum_{j \in J} f_j \mu_j + t \sum_{j \in J} \frac{\bar{s}_j}{\mu_j - \bar{s}_j} \Big] + \sum_{j \in J} \sum_{i \in I} c_{ij} \xi_i \bar{y}_{ij} \tag{3.7a}$$

$$\text{s.t. } \bar{s}_j \leq \mu_j \qquad \forall j \in J. \tag{3.7b}$$

which decomposes by j to n subproblems. Each subproblem can be stated as $V_j(\bar{s}_j) = \min_{\mu_j \geq \bar{s}_j} f_j \mu_j + \dfrac{t\bar{s}_j}{\mu_j - \bar{s}_j}$, which is a single-variable convex minimization problem. By setting its first derivative equal to zero, we get

$$f_j + \frac{-t\bar{s}_j}{(\mu_j - \bar{s}_j)^2} = 0 \Rightarrow f_j = \frac{t\bar{s}_j}{(\mu_j - \bar{s}_j)^2}$$

$$\mu_j - \bar{s}_j = \sqrt{\frac{t\bar{s}_j}{f_j}} \Rightarrow \mu_j^* = \bar{s}_j + \sqrt{\frac{t\bar{s}_j}{f_j}} \geq \bar{s}_j,$$

which renders constraints set (3.7b) redundant. By substituting μ_j^* back in the subproblem we have

$$\begin{aligned} V_j^*(\bar{s}_j) &= f_j \left[\bar{s}_j + \sqrt{\frac{t\bar{s}_j}{f_j}} \right] + \frac{t\bar{s}_j}{\sqrt{\frac{t\bar{s}_j}{f_j}}} \\ &= f_j \bar{s}_j + \sqrt{tf_j}\sqrt{\bar{s}_j} + \sqrt{tf_j}\sqrt{\bar{s}_j} \\ &= f_j \bar{s}_j + 2\sqrt{tf_j}\sqrt{\bar{s}_j}. \end{aligned} \tag{3.8}$$

Thus, by using result (3.8), the SSDP can be reformulated as

$$\min_{y} \sum_{j \in J} \sum_{i \in I} c_{ij} \xi_i y_{ij} + \sum_{j \in J} \sum_{i \in I} f_j \xi_i y_{ij} + 2 \sum_{j \in J} \sqrt{tf_j} \sqrt{\sum_{i \in I} \xi_i y_{ij}} \tag{3.9a}$$

$$\text{s.t. } \sum_{j \in J} y_{ij} = 1 \qquad \forall i \in I \tag{3.9b}$$

$$y_{ij} \in \{0, 1\} \qquad \forall i \in I, \forall j \in J. \tag{3.9c}$$

Next, let us define $z_j \geq \sqrt{\sum_{i \in I} \xi_i y_{ij}}$, and replace $\sqrt{\sum_{i \in I} \xi_i y_{ij}}$ in the objective function with z_j, and add this constraints set to the mathematical model. Furthermore, since

$y_{ij} \in \{0,1\}$, we can replace it with y_{ij}^2. These transformations enable us to rewrite (3.1) as

$$\min_{\mathrm{y,z}} \sum_{j\in J}\sum_{i\in I} c_{ij}\xi_i y_{ij} + \sum_{j\in J}\sum_{i\in I} f_j \xi_i y_{ij} + 2\sum_{j\in J}\sqrt{tf_j}z_j \tag{3.10a}$$

$$\text{s.t.} \quad \sum_{j\in J} y_{ij} = 1 \qquad \forall i \in I \tag{3.10b}$$

$$z_j \geq \sqrt{\sum_{i\in I}\xi_i y_{ij}^2} \qquad \forall j \in J \tag{3.10c}$$

$$y_{ij} \in \{0,1\},\ z_j \geq 0 \qquad \forall i \in I, \forall j \in J. \tag{3.10d}$$

The objective function is linear in both y and z, whereas constraint (3.10c) is a second-order cone (SOC) constraint. This mathematical model can be solved directly on commercial solvers like Cplex or Gurobi.

3.2 The Robust Optimization (RO) Problem

It is often the case that customers' demand is not known with certainty, but can rather be represented as a parameter that lies within an *uncertainty set*. In such case, it might be desirable to protect against this uncertainty in demand by employing a robust optimization approach. In general, in this approach, a *risk-averse* decision-maker who aims to avoid large losses might opt to minimize the *worst-case-scenario* loss, which means that, for a given x, if $h(\mathrm{x}) : \sup_{\xi\in\Xi} g(\mathrm{x}, \xi)$, where Ξ is the uncertainty set, we aim to find $\min_{\mathrm{x}\in X} h(\mathrm{x})$, or equivalently $\min_{\mathrm{x}\in X}\sup_{\xi\in\Xi} g(\mathrm{x},\xi)$, which is called the *Robust Counterpart* (*i.e.,* select $\mathrm{x} \in X$ such that when the most adverse scenario $\xi \in \Xi$ is realized, the loss is minimized). Using this perspective, the robust counterpart of (3.9) can be stated as follows:

$$\min_{\mathrm{y}} \sup_{\xi \in \Xi} \Big[\sum_{j \in J} \sum_{i \in I} (c_{ij} + f_j) \xi_i y_{ij} + 2 \sum_{j \in J} \sqrt{t f_j} \sqrt{\sum_{i \in I} \xi_i y_{ij}^2} \Big] \tag{3.11a}$$

$$\text{s.t.} \quad \sum_{j \in J} y_{ij} = 1 \qquad \forall i \in I \tag{3.11b}$$

$$y_{ij} \in \{0, 1\} \qquad \forall i \in I, \forall j \in J. \tag{3.11c}$$

In this section, we consider two classes of uncertainty sets, Budgeted and Ball uncertainty set, and show how the robustified problem can be tractably formulated.

3.2.1 Budgeted Uncertainty Set

First, we consider the budgeted uncertainty set introduced by Bertsimas and Sim [20], defined as $\Xi_{Bu} := \{\xi \in \mathbb{R}_+^m \mid \xi_i = \xi_i^{nom} + \widehat{\xi}_i w_i, \quad \sum_{i=1}^m |w_i| \leq \Gamma, \quad |w_i| \leq 1\}$. For this set, w_i is the primary uncertain parameter, $\Gamma \in [0, m]$ is the uncertainty budget, and $\widehat{\xi}$ is the maximum absolute deviation from the nominal value. To tractably reformulate (3.11a), we utilize the scheme based on Fenchel duality proposed by Ben-Tal, Hertog, and Vial [11]. This is possible since the objective function is concave in y_{ij} which is the optimization variable. We first state the Theorem we will use in this reformulation.

Theorem 3.1 ([11],Theorem 2)**.** *The vector* $\mathrm{y} \in Y$ *satisfies the robust constraint* $g(\mathrm{y}, \xi) \leq 0, \forall \xi \in \Xi$ *if and only if* y *and* $\mathrm{v} \in \mathbb{R}^m$ *satisfy the single inequality*

$$(FRC) \quad \delta^*(\mathrm{v} \mid \Xi) - g_*(\mathrm{y}, \mathrm{v}) \leq 0,$$

where δ^* *is the support function of set* Ξ*, defined as*

$$\delta^*(\mathrm{v} \mid \Xi) := \sup_{\xi \in \Xi} \xi^\intercal \mathrm{v} \tag{3.12}$$

and, $g_*(.,.)$ *is the partial concave conjugate with respect to the first variable and defined as*

$$g_*(\mathrm{y}, \mathrm{v}) := \inf_{\xi \in \Xi_g} \mathrm{v}^\intercal \xi - g(\mathrm{y}, \xi), \tag{3.13}$$

and $g(.,.)$ is a mapping defined over the convex domain $Y_g \times \Xi_g$ with $Y_g \subseteq \mathbb{R}^n$ and $\Xi_g \subseteq \mathbb{R}^m$.

This theorem represents a general Fenchel Robust Counterpart (FRC) formulation for a general robust constraint g which indicates that the computations involving g_* are completely independent from those involving Ξ. Based on this theorem, Ben-Tal *et al.* [11] illustrate how to compute $\delta^*(\mathrm{v} \mid \Xi)$ and $g_*(\mathrm{y}, \mathrm{v})$ in (FRC) for several choices of Ξ and g, respectively. One of this results states that the robust counterpart of $\sum_{k=1}^{K} f'_k(\mathrm{y}, \boldsymbol{\xi})$ can be shown as follow:

$$\begin{cases} \delta^*(\mathrm{v} \mid \Xi) - \sum_{k=1}^{K} (f'_k)_*(\mathrm{p_k}, \mathrm{y}) \leq 0 \\ \sum_{k=1}^{K} \mathrm{p_k} = \mathrm{v}. \end{cases} \tag{3.14}$$

Corollary 3.1. *When the uncertainty set is Ξ_{Bu}, objective function (3.11a) can be tractably reformulated as*

$$\begin{aligned}
\min_{\mathrm{y},\boldsymbol{\delta},\mathrm{p_1},\mathrm{p_2},\mathrm{v},\theta,\boldsymbol{\phi}} \quad & \sum_{i \in I} \xi_i^{nom} v_i + \Gamma\theta + \sum_{i \in I} \phi_i + t \sum_{j \in J} f_j \delta_j \\
s.t. \quad & \theta + \phi_i \geq \widehat{\xi}_i v_i && \forall i \in I \\
& p_{1i} \geq \sum_{j \in J} (c_{ij} + f_j) y_{ij} && \forall i \in I \\
& \delta_j p_{2i} \geq y_{ij}^2 && \forall i \in I, \forall j \in J \\
& p_{1i} + p_{2i} = v_i && \forall i \in I \\
& y_{ij} \in \{0,1\},\ \delta_j, p_{1i}, p_{2i}, v_i, \theta, \phi_i \geq 0 && \forall i \in I, \forall j \in J.
\end{aligned}$$

Proof. First, considering result (3.14), we need to find the support function, which becomes

$$
\begin{aligned}
\delta^*(\mathrm{v} \mid \Xi_{Bu}) &= \sup_{\xi \in \Xi_{Bu}} \xi^\intercal \mathrm{v} \\
&= \sum_{i \in I} \xi_i^{nom} v_i + \sup_{w_i} \sum_{i \in I} \widehat{\xi}_i w_i v_i \\
\text{s.t.} \quad & \sum_{i \in I} w_i \leq \Gamma && (\theta) \\
& 0 \leq w_i \leq 1 && (\phi_i).
\end{aligned}
$$

and can be written in the dual form as

$$
\begin{aligned}
& \sum_{i \in I} \xi_i^{nom} v_i + \inf_{\theta, \phi} \Big[\Gamma\theta + \sum_{i \in I} \phi_i \Big] \\
\text{s.t.} \quad & \theta + \phi_i \geq \widehat{\xi}_i v_i && \forall i \in I \\
& \theta, \phi_i \geq 0 && \forall i \in I.
\end{aligned} \tag{3.15}
$$

Next, to calculate $\sum\limits_{k=1}^{K} (f'_k)_*(\mathrm{p_k}, \mathrm{y})$, we define f'_1 and f'_2 as follows

$$
\begin{cases}
f'_1 := \sum\limits_{j \in J} \sum\limits_{i \in I} (c_{ij} + f_j) \xi_i y_{ij} \\
\\
f'_2 := 2 \sum\limits_{j \in J} \sqrt{t f_j} \sqrt{\sum\limits_{i \in I} \xi_i y_{ij}^2}.
\end{cases} \tag{3.16}
$$

So, the conjugate functions for f'_1 becomes

$$
\begin{aligned}
(f'_1)_*(\mathrm{p}_1, \mathrm{y}) &= \inf_{\xi \geq 0} \boldsymbol{p_1}^\intercal \xi - f_1(\boldsymbol{y}, \xi) \\
&= \inf_{\xi \geq 0} \sum_{i \in I} p_{1i} \xi_i - \sum_{j \in J} \sum_{i \in I} (c_{ij} + f_j) \xi_i y_{ij} \\
&= \inf_{\xi \geq 0} \sum_{i \in I} \xi_i \Big[p_{1i} - \sum_{j \in J} (c_{ij} + f_j) y_{ij} \Big].
\end{aligned} \tag{3.17}
$$

This minimization over $\xi_i \geq 0$ returns 0 if $p_{1i} \geq \sum\limits_{j \in J} (c_{ij} + f_j) y_{ij}, \forall i \in I$ and $-\infty$ otherwise. So, $(f'_1)_* = 0$, and the constraint $p_{1i} \geq \sum\limits_{j \in J} (c_{ij} + f_j) y_{ij}, \forall i \in I$ is added to the reformulated problem. Next, by decomposing f'_2 by j, the conjugate function of

f'_{2j} can be written as

$$\begin{aligned}
(f'_{2j})_*(\mathrm{p}_2, \mathrm{y}) &= \inf_{\xi \geq 0} \mathrm{p}_2{}^\intercal \xi - f_{2j}(\mathrm{y}, \xi) \\
&= \inf_{\xi \geq 0} \sum_{i \in I} p_{2i}\xi_i - 2\sqrt{tf_j}\sqrt{\sum_{i \in I} \xi_i y_{ij}^2} \\
&= \inf_{\xi \geq 0, \psi_j} \sum_{i \in I} p_{2i}\xi_i - 2\sqrt{tf_j}\sqrt{\psi_j} \\
\text{s.t.} \quad \psi_j &\leq \sum_{i \in I} \xi_i y_{ij}^2 \qquad\qquad (\eta_j).
\end{aligned}$$

In this case, we assemble the dual problem based on Lagrangian duality. For any $\eta_j \geq 0$, the Lagrangian function

$$L(\mathrm{y}, \mathrm{p}_2, \eta_j) = \inf_{\xi \geq 0, \psi_j} \Big[\sum_{i \in I} p_{2i}\xi_i - 2\sqrt{tf_j}\sqrt{\psi_j} + \eta_j\Big(\psi_j - \sum_{i \in I} \xi_i y_{ij}^2\Big)\Big]$$

provides a lower bound for $(f'_{2j})_*(\mathrm{p}_2, \mathrm{y})$. Since $(f'_{2j})_*$ is convex, and satisfies the weak Slater's condition, the strong duality holds; thus, the bound is tight at optimality [21], and we have

$$\begin{aligned}
(f'_{2j})_*(\mathrm{p}_2, \mathrm{y}) &= \max_{\eta_j \geq 0} L(\mathrm{y}, \mathrm{p}_2, \eta_j) \\
&= \max_{\eta_j \geq 0} \Big[\inf_{\xi \geq 0} \Big(\sum_{i \in I} p_{2i}\xi_i - \eta_j \sum_{i \in I} \xi_i y_{ij}^2 \Big) + \inf_{\psi_j} \Big(\eta_j\psi_j - 2\sqrt{tf_j}\sqrt{\psi_j} \Big) \Big].
\end{aligned}$$

The first inner minimization over $\xi_i \geq 0$ can be written as $\inf_{\xi \geq 0} \sum_{i \in I} \xi_i(p_{2i} - \eta_j y_{ij}^2)$, which equals 0 if $p_{2i} \geq \eta_j y_{ij}^2$, $\forall i \in I$ and $-\infty$ otherwise. The second minimization over ψ_j is a convex function; therefore, it can be solved by setting its first derivative equal to zero:

$$\eta_j - \frac{2\sqrt{tf_j}}{2\sqrt{\psi_j}} = 0 \Rightarrow \eta_j = \frac{\sqrt{tf_j}}{\sqrt{\psi_j}} \Rightarrow \psi_j^* = \frac{tf_j}{\eta_j^2}.$$

by substituting this value in the second minimization problem, it reduces to $-\dfrac{tf_j}{\eta_j}$. With that, the partial concave conjugate function for every j becomes

$$\begin{aligned}
(f'_{2j})_*(\mathrm{p}_2, \mathrm{y}) &= \max_{\gamma} -\frac{tf_j}{\eta_j} \\
\text{s.t.} \quad p_{2i} &\geq \eta_j y_{ij}^2 \qquad\qquad \forall i \in I.
\end{aligned}$$

which, by defining $\delta_j = \dfrac{1}{\eta_j}$, can be written as

$$
\begin{aligned}
(f'_{2j})_*(\mathrm{p}_2, \mathrm{y}) = \max_{\gamma} & -(tf_j)\delta_j \\
\text{s.t.} \quad & p_{2i}\delta_j \geq y_{ij}^2 && \forall i \in I.
\end{aligned}
$$

So, $(f'_2)_*$ becomes

$$
\begin{aligned}
(f'_2)_*(\mathrm{p}_2, \mathrm{y}) = -\min_{\delta} & \; t\sum_{j\in J} f_j\delta_j \\
\text{s.t.} \quad & \delta_j p_{2i} \geq y_{ij}^2 && \forall i \in I, \forall j \in J.
\end{aligned} \tag{3.18}
$$

Now, $(f'_2)_*$ is a linear function with rotated second-order cone constraints. Substituting (3.15),(3.18) and the result from (3.17) in (3.14) completes the proof. $\square$

Therefore, the robust counterpart of the nominal problem becomes

$$
\min_{\mathrm{y},\delta,\mathrm{p}_1,\mathrm{p}_2,\mathrm{v},\theta,\phi} \sum_{i\in I} \xi_i^{nom} v_i + \Gamma\theta + \sum_{i\in I}\phi_i + t\sum_{j\in J} f_j\delta_j \tag{3.19a}
$$

$$
\text{s.t.} \quad \sum_{j\in J} y_{ij} = 1 \qquad \forall i \in I \tag{3.19b}
$$

$$
\theta + \phi_i \geq \widehat{\xi}_i v_i \qquad \forall i \in I \tag{3.19c}
$$

$$
p_{1i} \geq \sum_{j\in J}(c_{ij} + f_j)y_{ij} \qquad \forall i \in I \tag{3.19d}
$$

$$
\delta_j p_{2i} \geq y_{ij}^2 \qquad \forall i \in I, \forall j \in J \tag{3.19e}
$$

$$
p_{1i} + p_{2i} = v_i \qquad \forall i \in I \tag{3.19f}
$$

$$
y_{ij} \in \{0,1\},\; \delta_j, p_{1i}, p_{2i}, v_i, \theta, \phi_i \geq 0 \quad \forall i \in I, \forall j \in J. \tag{3.19g}
$$

The objective function is linear, whereas constraint (3.19e) is a second-order cone constraint. This mathematical model can be solved directly on commercial solvers.

3.2.2 Ball Uncertainty Set

Next, consider the case when Ξ is a Ball uncertainty set of the form $\Xi_{Ba} := \{\xi \in \mathbb{R}^m_+ \mid \xi = \xi^{nom} + \widehat{\xi}, \quad \|\widehat{\xi}\|_2 \leq r\}$. This is a special case (with $\Sigma = 1/r$) of the Ellipsoidal uncertainty set introduced by Ben-Tal and Nemirovski [13] that takes the form $\Xi_E := \{\xi \in \mathbb{R}^m_+ \mid \xi^\intercal \Sigma \xi \leq 1, \ \Sigma \succ 0\}$. Note that the Fenchel duality scheme we used with the budgeted uncertainty set effectively decomposes the dependence of the reformulation between the uncertainty set and the constraint function. Therefore, in order to tractably reformulate the objective function (3.11a) with any other uncertainty set, we need only to replace the support function $\delta^*(\mathrm{v} \mid \Xi)$, whereas the concave conjugate function remains unchanged. With the ball uncertainty set, the support function becomes

$$
\begin{aligned}
\delta^*(\mathrm{v} \mid \Xi_{Ba}) &= \sup_{\xi \in \Xi_{Ba}} \{\xi^\intercal \mathrm{v} \mid \xi = \xi^{nom} + \widehat{\xi}\} \\
&= \sup_{\widehat{\xi}} \widehat{\xi}^\intercal \mathrm{v} + (\xi^{nom})^\intercal \mathrm{v} \\
&\text{s.t.} \quad \|\widehat{\xi}\|_2 \leq r.
\end{aligned}
$$

where $\sup_{\|\widehat{\xi}\|_2 \leq r} \widehat{\xi}^\intercal \mathrm{v}$ is the definition of the dual norm of the Euclidean norm, and evaluates to $r\|\mathrm{v}\|_2$. More generally, the dual of the l_p-norm is the l_q-norm, where q satisfies $\frac{1}{p} + \frac{1}{q} = 1$. Thus, the objective function (3.11a) can be replaced with

$$
\begin{aligned}
\min_{\mathrm{y},\delta,\mathrm{p}_1,\mathrm{p}_2,\mathrm{v}} \quad & \sum_{i \in I} \xi_i^{nom} v_i + r\|\mathrm{v}\|_2 + t \sum_{j \in J} f_j \delta_j \\
\text{s.t.} \quad & p_{1i} \geq \sum_{j \in J} (c_{ij} + f_j) y_{ij} && \forall i \in I \\
& \delta_j p_{2i} \geq y_{ij}^2 && \forall i \in I, \forall j \in J \\
& p_{1i} + p_{2i} = v_i && \forall i \in I \\
& y_{ij} \in \{0,1\}, \ \delta_j, p_{1i}, p_{2i}, v_i \geq 0 && \forall i \in I, \forall j \in J.
\end{aligned}
$$

Next, let us define $u \geq \sqrt{\sum_{i \in I} v_i^2}$, and replace $\sqrt{\sum_{i \in I} v_i^2}$ in the objective function with u and add this constraint to the mathematical model. Therefore, the robust problem can be reformulated as

$$\min_{\mathrm{y},\delta,\mathrm{p}_1,\mathrm{p}_2,\mathrm{v},u} \sum_{i \in I} \xi_i^{nom} v_i + ru + t \sum_{j \in J} f_j \delta_j \tag{3.20a}$$

$$\text{s.t.} \quad \sum_{j \in J} y_{ij} = 1 \qquad \forall i \in I \tag{3.20b}$$

$$p_{1i} \geq \sum_{j \in J} (c_{ij} + f_j) y_{ij} \qquad \forall i \in I \tag{3.20c}$$

$$\delta_j p_{2i} \geq y_{ij}^2 \qquad \forall i \in I, \forall j \in J \tag{3.20d}$$

$$p_{1i} + p_{2i} = v_i \qquad \forall i \in I \tag{3.20e}$$

$$\sum_{i \in I} v_i^2 \leq u^2 \tag{3.20f}$$

$$y_{ij} \in \{0,1\}, \ \delta_j, p_{1i}, p_{2i}, v_i, u \geq 0 \quad \forall i \in I, \forall j \in J. \tag{3.20g}$$

Again, this is a mixed-integer programming problem with the second-order cone constraints (3.20d) and (3.20f) that can be solved using commercial solvers.

3.3 The Distributionally-Robust Optimization (DRO) Problem

In the RO framework, the realizations of the demand (uncertain parameter) are not representative of the demand distribution, and we assume that we do not have any information about this distribution. Instead, we use these realizations merely to construct the uncertainty set. However, if we are confident that these data points can be representative of the population distribution, DRO is a good alternative for approximating the true demand distribution.

Formally, the DRO problem is stated as $\min_{\mathrm{x} \in \mathcal{X}} \sup_{F_{\xi} \in \mathcal{D}} \mathbb{E}_{F_{\xi}}[g(\mathrm{x}, \xi)]$, where the uncertain parameter ξ follows a probability distribution F_{ξ} that belongs to a distributional ambiguity set (DAS) $\mathcal{D}$, *i.e.,* we minimize the *worst-case expected loss*, where the expectation is taken with respect to the probability distributions in the DAS. In this section, we will use the Wasserstein-metric-based ambiguity set introduced in [35], which can be described as follows: Given a finite set $\widehat{\Xi} := \left\{\widehat{\xi}^1, \ldots, \widehat{\xi}^N\right\}$ of sample points, each representing a historical or predicted realization of the uncertain parameters, an empirical distribution $\widehat{F}_\xi$ can be constructed such that each discrete point in the sample set has an equal probability of $\frac{1}{N}$, *i.e.,* $\widehat{F}_\xi := \frac{1}{N}\sum_{n=1}^{N} \delta_{\widehat{\xi}^n}$, where $\delta_\xi : \Sigma \mapsto \{0,1\}$, $\delta_{\widehat{\xi}^n}(\mathcal{A}) = \begin{cases} 1 & \text{if } \widehat{\xi}^n \in \mathcal{A} \\ 0 & \text{otherwise} \end{cases}$ is a *Dirac measure* concentrating unit mass at $\widehat{\xi}^n$, and Σ is a Borel σ-algebra on Ξ. The *Wasserstein ambiguity set* $\mathcal{D}_\epsilon(\widehat{F}_\xi, \Xi)$ will be constructed as a ball around the empirical distribution and includes all probability distributions supported on $\Xi \subset \mathbb{R}^m$ that are within a distance $\epsilon \geq 0$ of the reference/empirical distribution $\widehat{F}_\xi$, where the distance is measured using the *Wasserstein metric*, which is also referred to as the Kantorovich-Rubinstein metric [41]. Formally, the Wasserstein ambiguity set can be stated as

$$\mathcal{D}_\epsilon(\widehat{F}_\xi, \Xi) := \left\{F_\xi \in \mathcal{M}(\Xi) \,:\, d_{\mathrm{W}}\left(\widehat{F}_\xi, F_\xi\right) \leq \epsilon, \mathbb{P}(\xi \in \Xi) = 1\right\},$$

where the Wasserstein metric $d_{\mathrm{W}} : \mathcal{M}(\Xi) \times \mathcal{M}(\Xi) \to \mathbb{R}$ is defined as

$$d_{\mathrm{W}}(F_1, F_2) := \inf \left\{ \int_{\Xi^2} \|\xi_1 - \xi_2\| \Pi(\mathrm{d}\xi_1, \mathrm{d}\xi_2) \;\middle|\; \begin{array}{l} \Pi \text{ is a joint distribution of } \xi_1 \text{ and } \xi_2 \\ \text{with marginals } F_1 \text{ and } F_2 \text{ respectively} \end{array} \right\},$$

where $\|\cdot\|$ represents an arbitrary norm on $\mathbb{R}^m$ and the probability space $\mathcal{M}(\Xi)$ contains all probability distributions supported on Ξ. The decision variable Π can be viewed as a *transportation plan* for moving a mass distribution described by F_1 to another one described by F_2. Thus, the Wasserstein distance between F_1 and F_2 represents the cost of an optimal mass transportation plan, where the norm $\|\cdot\|$ encodes the transportation costs.

Starting with the reformulated problem (3.9), the single-stage distributionally-robust SSDP can be stated

$$\min_{\mathrm{y}} \sup_{F_{\xi}\in\mathcal{D}_{\epsilon}(\widehat{F}_{\xi})} \mathbb{E}_{F_{\xi}}\Big[\sum_{j\in J}\sum_{i\in I}(c_{ij}+f_j)\xi_i y_{ij} + 2\sum_{j\in J}\sqrt{tf_j}\sqrt{\sum_{i\in I}\xi_i y_{ij}}\Big] \tag{3.21a}$$

$$\text{s.t.} \quad \sum_{j\in J} y_{ij} = 1 \qquad \forall i\in I \tag{3.21b}$$

$$y_{ij}\in\{0,1\} \qquad \forall i\in I, \forall j\in J. \tag{3.21c}$$

Assumption 3.1. *In the Wasserstein distributional ambiguity set $\mathcal{D}_{\varepsilon}(\widehat{F}_{\xi},\Xi)$, (i) the support set is a bounded polyhedron defined as $\Xi := \{\xi\in\mathbb{R}^n \mid \mathrm{C}\xi\le \mathrm{d}\}$, for some $\mathrm{C}\in\mathbb{R}^{|L|\times n}$ and $\mathrm{d}\in\mathbb{R}^{|L|}$; and (ii) the norm used in the Wasserstein metric definition is an l_1-norm.*

Now, moving to the distributionally-robust objective function (3.21a), we utilize Theorem 4.2 in [35], which applies since the inner function inside the brackets is concave in ξ and Ξ is a convex and closed set (Assumption 4.1, [35]).

Lemma 3.2. *Objective function (3.21a) is equivalent to:*

$$\begin{aligned}
\min_{\mathrm{y},\lambda,\delta,\mathrm{r},\mathrm{s},\alpha} \quad & \lambda\epsilon + \frac{1}{N}\sum_{n=1}^{N} s_n \\
s.t. \quad & t\sum_{j\in J} f_j\delta_{jn} + \sum_{l\in L} d_l\alpha_{ln} + \sum_{i\in I} r_{ni}\widehat{\xi}_i^n \le s_n && \forall n\in N \\
& y_{ij}^2 \le \delta_{jn}\Big[r_{ni} - (c_{ij}+f_j)y_{ij} + \sum_{l\in L}\alpha_{ln}C_{ln}\Big] && \forall i\in I,\ j\in J, \forall n\in N \\
& r_{in}\le\lambda && \forall i\in I,\forall n\in N \\
& -r_{in}\le\lambda && \forall i\in I,\forall n\in N \\
& y_{ij}\in\{0,1\},\ \lambda,\delta_{jn},r_{ni},s_n,\alpha_{ln}\ge 0 && \forall i\in I,\forall j\in J,\forall n\in N,\forall l\in L.
\end{aligned}$$

Proof. According to ([35], Theorem 4.2) the DRO problem

$$\sup_{F_{\xi}\in\mathcal{D}_{\epsilon}(\widehat{F}_{\xi},\Xi)} \mathbb{E}_{F_{\xi}}\left[g(\mathrm{y},\xi)\right]$$

is equivalent to

$$\begin{aligned} \inf_{\lambda,\mathrm{s},\mathrm{r},\nu} \quad & \lambda\epsilon + \frac{1}{N}\sum_{n=1}^{N} s_n \\ \text{s.t.} \quad & [-g]^*(r_n - \nu_n, y) + \sigma_\Xi(\nu_n) - r_n^\intercal \widehat{\xi}^n \le s_n \quad && \forall n \in N \\ & \|r_n\|_* \le \lambda && \forall n \in N. \end{aligned} \tag{3.23}$$

where $[-g]^*(r_n - \nu_n)$ denotes the conjugate of $-g$ evaluated at $r_n - \nu_n$ and σ_Ξ represents the support function of Ξ. Besides, in the proof of the same theorem, it has been shown that the optimal value of (3.23) coincides with the optimal value of the following RO problem

$$\inf_{\lambda,\mathrm{s},\mathrm{r}} \quad \lambda\epsilon + \frac{1}{N}\sum_{n=1}^{N} s_n \tag{3.24a}$$

$$\text{s.t.} \sup_{\boldsymbol{\xi}\in\Xi} \left(g(\mathrm{y},\boldsymbol{\xi}) - r_n^\intercal \boldsymbol{\xi}\right) + r_n^\intercal \widehat{\boldsymbol{\xi}}^n \le s_n \quad \forall n \in N \tag{3.24b}$$

$$\|r_n\|_* \le \lambda \qquad \forall n \in N. \tag{3.24c}$$

To prove Lemma (3.2), we are going to use result (3.24) that uses the robust constraint (3.24b), which can be written as follow

$$\sup_{\boldsymbol{\xi}\in\Xi} \left(g(\mathrm{y},\boldsymbol{\xi}) - r_n^\intercal \boldsymbol{\xi}\right) = -\inf_{\boldsymbol{\xi}\in\Xi} \left(r_n^\intercal \xi - g(\mathrm{y},\boldsymbol{\xi})\right), \tag{3.25}$$

where

$$g = \sum_{j\in J}\sum_{i\in I}(c_{ij} + f_j)\xi_i y_{ij} + 2\sum_{j\in J}\sqrt{tf_j}\sqrt{\sum_{i\in I}\xi_i y_{ij}}.$$

which can be decomposed by j. Now, for each j, (3.25) can be represented as

$$\begin{aligned} -\inf_{\xi\in\Xi} \quad & \left[\sum_{i\in I} r_{ni}\xi_i - \sum_{i\in I}(c_{ij} + f_j)\xi_i y_{ij} - 2\sqrt{tf_j}\sqrt{\sum_{i\in I}\xi_i y_{ij}}\right] \\ \text{s.t.} & \sum_{i\in I} C_{li}\xi_i \le d_l \quad \forall l \in L. \end{aligned} \tag{3.26}$$

By defining a new variable $\zeta_j = \sum\limits_{i \in I} \xi_i y_{ij}^2$ (3.26) becomes

$$
\begin{aligned}
-\inf_{\xi,\zeta_j} \quad & \sum_{i \in I} r_{ni}\xi_i - \sum_{i \in I}(c_{ij} + f_j)\xi_i y_{ij} - 2\sqrt{tf_j}\sqrt{\zeta_j} && (3.27) \\
\text{s.t.} \quad & \sum_{i \in I} C_{li}\xi_i \leq d_l && (\alpha_{ln}) \\
& \zeta_j \leq \sum_{i \in I} \xi_i y_{ij}^2 && (\gamma_{jn}).
\end{aligned}
$$

In this case, we have to assemble the duality problem based on Lagrangian duality. For any $\gamma_{jn} \geq 0$, the Lagrangian function

$$
\begin{aligned}
L(\mathrm{y}, \alpha, \gamma_{jn}) = -\inf_{\xi,\zeta_j} \Big[& \sum_{i \in I} r_{ni}\xi_i - \sum_{i \in I}(c_{ij} + f_j)\xi_i y_{ij} - 2\sqrt{tf_j}\sqrt{\zeta_j} \\
& + \sum_{l \in L} \alpha_{ln}\Big(\sum_{i \in I} C_{li}\xi_i - d_l\Big) + \gamma_{jn}\Big(\zeta_j - \sum_{i \in I}\xi_i y_{ij}^2\Big)\Big].
\end{aligned}
$$

provides an upper bound for (3.27). Since it satisfies the strong duality conditions, the bound is tight at optimality and we have

$$
\begin{aligned}
\min_{\alpha,\gamma_{jn}} L(\mathrm{y}, \alpha, \gamma_{jn}) = \min_{\alpha,\gamma_{jn}} \Big[& -\inf_{\xi}\Big(\sum_{i \in I} \xi_i\Big[r_{ni} - (c_{ij} + f_j)y_{ij} + \sum_{l \in L}\alpha_{ln}C_{li} - \gamma_{jn}y_{ij}^2\Big]\Big) \\
& -\inf_{\zeta_j}\Big(-2\sqrt{tf_j}\sqrt{\zeta_j} + \gamma_{jn}\zeta_j\Big)\Big] + \sum_{l \in L}\alpha_{ln}d_l. \qquad (3.28)
\end{aligned}
$$

The first inner minimization over $\xi_i \geq 0$ would be equal to 0 if $r_{ni} - (c_{ij} + f_j)y_{ij} + \sum\limits_{l} \alpha_{ln}C_{li} \geq \gamma_{jn}y_{ij}^2, \forall i \in I$ and $-\infty$ otherwise. The second minimization over ζ_j is a convex function; therefore, it can be solved by setting its first derivative equal to zero

$$
\gamma_{jn} - \frac{2\sqrt{tf_j}}{2\sqrt{\zeta_j}} = 0 \Rightarrow \zeta_j^* = \frac{tf_j}{\gamma_{jn}^2}.
$$

by substituting this value in the second minimization problem, it reduces to $\dfrac{tf_j}{\gamma_{jn}}$. With that, (3.28) becomes

$$
\begin{aligned}
\min_{\alpha,\gamma_{jn}} \quad & \frac{tf_j}{\gamma_{jn}} + \sum_{l \in L}\alpha_{ln}d_l \\
\text{s.t.} \quad & \gamma_{jn}y_{ij}^2 \leq r_{ni} - (c_{ij} + f_j)y_{ij} + \sum_{l \in L}\alpha_{ln}C_{li} && \forall i \in I, \forall n \in N.
\end{aligned}
$$

which by defining $\delta_{jn} = \dfrac{1}{\gamma_{jn}}$, the problem can be written as

$$
\begin{aligned}
\min_{\alpha,\delta_{jn}} \quad & tf_j\delta_{jn} + \sum_{l\in L} \alpha_{ln} d_l \\
\text{s.t.} \quad & y_{ij}^2 \le \delta_{jn} r_{ni} - \delta_{jn}(c_{ij}+f_j)y_{ij} + \delta_{jn}\sum_{l\in L}\alpha_{ln}C_{li} \qquad \forall i\in I, \forall n \in N.
\end{aligned}
$$

So, (3.25) becomes

$$
\begin{aligned}
\min_{\alpha,\delta_{jn}} \quad & t\sum_{j\in J} f_j\delta_{jn} + \sum_{l\in L} \alpha_{ln} d_l && (3.29)\\
\text{s.t.} \quad & y_{ij}^2 \le \delta_{jn} r_{ni} - \delta_{jn}(c_{ij}+f_j)y_{ij} + \delta_{jn}\sum_{l\in L}\alpha_{ln}C_{li} \qquad \forall i\in I, \forall j \in J, \forall n \in N.
\end{aligned}
$$

Finally, the norm constraint simply reduces to $|r_{in}| \le \lambda$, $\forall i \in I, \forall n \in N$. Combining the aforementioned derivations leads to the desired result. □

With that, the distributionally-robust service system design problem can be tractably formulated as a mixed-integer second-order conic program

$$
\begin{aligned}
\min_{\mathrm{y},\lambda,\delta,\mathrm{r},\mathrm{s},\alpha} \quad & \lambda\epsilon + \frac{1}{N}\sum_{n=1}^{N} s_n && && (3.30a)\\
\text{s.t.} \quad & \sum_{j\in J} y_{ij} = 1 && \forall i \in I && (3.30b)\\
& t\sum_{j\in J} f_j\delta_{jn} + \sum_{l\in L} d_l\alpha_{ln} + \sum_{i\in I} r_{ni}\widehat{\xi}_i^n \le s_n && \forall n \in N && (3.30c)\\
& y_{ij}^2 \le \delta_{jn}\Big[r_{ni} - (c_{ij}+f_j)y_{ij} + \sum_{l\in L}\alpha_{ln}C_{ln}\Big] && \forall i \in I,\ j\in J, \forall n \in N && (3.30d)\\
& r_{in} \le \lambda && \forall i \in I, \forall n \in N && (3.30e)\\
& -r_{in} \le \lambda && \forall i \in I, \forall n \in N && (3.30f)\\
& y_{ij} \in \{0,1\},\ \lambda, \delta_{jn}, r_{ni}, s_n, \alpha_{ln} \ge 0 && \forall i \in I, \forall j \in J, \forall n \in N, \forall l \in L. && (3.30g)
\end{aligned}
$$

In constraint (3.30d), the term inside the brackets can be replaced by a single variable.

Chapter 4

Service System Design Problems Modelled as a Network of G/M/1 Queues

In this chapter, we are going to focus on a SSDP that can be modelled as a network of G/M/1 queues. Since we assume that we do not know the demand distribution for sure, and we use RO, and DRO to deal with it, it is also unlikely that we can be sure about the arrival pattern, whether it is Markovian or not.

4.1 Problem Description

Let $I := \{DZ_i\}_{i=1}^{m}$ be a set of demand zones and $J := \{SC_j\}_{j=1}^{n}$ be a set of potential SC locations. Each demand zone needs to be assigned to a single SC to satisfy its demand. We assume that a demand zone arrival to its assigned SC follows a General distribution (*i.e.,* inter-arrival times of individual customers, T_i, are generally distributed *i.i.d.* random variables with variance $\sigma_i^2, i \in I$) with the mean arrival rate $\xi_i = 1/T_i, i \in I$. Initially, we will assume that these rates are known with certainty, whereas the uncertain case will be addressed later. Likewise, we assume that service provision at SC_j can be reasonably modelled as a Poisson process with a finite rate (*i.e.,* capacity) μ_j, which is a decision variable.

With that, each SC can be modelled as an G/M/1 queueing system and the service system becomes a network of G/M/1 queues. The G/M/1 model is a single service facility with one server, which has infinite buffers to store demands for service, and the first-come first-served queue discipline. The setup cost of SC is proportional to their service rate, *i.e.,* each service rate unit at SC_j costs f_j. There is a demand zone access cost c_{ij} per unit demand if DZ_i is assigned to SC_j. Furthermore, to discourage excessive waiting, customers' waiting time in the system is penalized at a constant rate of t per unit time. The objective is to determine the locations of SCs to open, their service capacities to install, and the assignment of customers to SC to minimize the total expected cost, including setup, access and waiting time costs. To formulate the problem, we use the following decision variables:

$$y_{ij} = \begin{cases} 1 & \text{if } DZ_i \text{ is assigned to } SC_j \\ 0 & \text{otherwise} \end{cases}$$

$$\mu_j = \text{service rate/capacity of } SC_j.$$

The simplest queueing model is the M/M/1 model presented in Chapter 3, in which the expected waiting time could be calculated exactly. However, for more realistic queueing models, finding exact solutions becomes more challenging to achieve. There are many relatively accurate but complicated approximations, such as the ones proposed by Buzacott and Shanthikumar [22], and Connors *et al.* [26] for G/G/s queueing models. The most widely used approximation was developed by Kingman [42] for the G/G/1 queueing models. Later, Hopp and Spearman [52] developed an estimate on the expected waiting time, based on Kingman's G/M/1 and Whitt's G/G/s [57] approximations as follow:

$$E[W_q] = \left(\frac{C_a^2 + C_s^2}{2}\right)\left(\frac{\rho^{\sqrt{2(s+1)}-1}}{1-\rho}\right)\left(\frac{1}{\mu}\right), \tag{4.1}$$

where C_a and C_s are the coefficients of variation of inter-arrival times and service times, respectively, and ρ is the utilization factor. Now, considering (4.1), and using *Little's Formula*, the expected waiting time (including service time) of customers at SC_j in a service system with G/M/1 queues can be written as

$$E[w_j] = \left(\frac{C_a^2 + 1}{2}\right)\frac{\rho_j}{\mu_j(1-\rho_j)} + \frac{1}{\mu_j}. \tag{4.2}$$

Here, we assume that $C_a = \dfrac{\sigma}{T}$ is constant across the system, given a homogeneous and infinite calling population. By substituting $\rho_j = \dfrac{\Lambda_j}{\mu_j}$, where $\Lambda_j = \sum_{i \in I} \xi_i y_{ij}$, in (4.2) we get

$$E[w_j] = \left(\frac{C_a^2+1}{2}\right)\frac{\Lambda_j}{\mu_j(\mu_j-\Lambda_j)} + \frac{1}{\mu_j}. \tag{4.3}$$

If we state the waiting time in terms of y_{ij} and μ_j, we would have

$$E[w_j(\mathrm{y},\mu)] = \left(\frac{C_a^2+1}{2}\right)\frac{\sum\limits_{i\in I}\xi_i y_{ij}}{\mu_j(\mu_j-\sum\limits_{i\in I}\xi_i y_{ij})} + \frac{1}{\mu_j}.$$

Therefore, the service system design problem can be formulated as

$$\min_{\mathrm{y},\mu} \sum_{j\in J} f_j\mu_j + \sum_{j\in J}\sum_{i\in I} c_{ij}\xi_i y_{ij} + t\sum_{j\in J}\sum_{i\in I}\xi_i y_{ij}E[w_j(\mathrm{y},\mu)] \tag{4.4a}$$

$$\text{s.t.} \quad \sum_{j\in J} y_{ij} = 1 \qquad \forall i \in I \tag{4.4b}$$

$$\sum_{i\in I}\xi_i y_{ij} \le \mu_j \qquad \forall j \in J \tag{4.4c}$$

$$y_{ij} \in \{0,1\},\ \mu_j \ge 0 \qquad \forall i \in I, \forall j \in J. \tag{4.4d}$$

In this formulation, the three terms in the objective function (4.4a) represent the capacity-dependent setup cost, access cost, and the waiting cost, respectively. Constraints (4.4b) ensure that every demand zone i is going to be assigned to exactly one SC. Constraints set in (4.4c) guarantee that each demand zone is assigned to an open SC only, and the total demand arrival rate to the SC does not exceed its service capacity. The proposed formulation results in a nonlinear mixed-integer program with linear constraints. When $\xi_i,\ i \in I$ is known with certainty, we refer to (4.4) as the Nominal Problem.

By defining $R = \dfrac{C_a^2+1}{2}$ for each facility, the last term in (4.4a) can be written as

$$t\sum_{j\in J}\Lambda_j\Big[\frac{R\Lambda_j}{\mu_j(\mu_j-\Lambda_j)} + \frac{1}{\mu_j}\Big] = t\sum_{j\in J}\Big[\frac{R\Lambda_j^2}{\mu_j(\mu_j-\Lambda_j)} + \frac{\Lambda_j}{\mu_j}\Big] = t\sum_{j\in J}\Big[\frac{R\rho_j^2}{1-\rho_j} + \rho_j\Big].$$

As a result, the nominal problem becomes

$$\min_{y,\mu,\rho} \sum_{j\in J} f_j\mu_j + \sum_{j\in J}\sum_{i\in I} c_{ij}\xi_i y_{ij} + t\sum_{j\in J}\left[\frac{R\rho_j^2}{1-\rho_j} + \rho_j\right] \tag{4.5a}$$

$$\text{s.t.} \quad \sum_{j\in J} y_{ij} = 1 \qquad \forall i \in I \tag{4.5b}$$

$$\sum_{i\in I} \xi_i y_{ij} \leq \mu_j \qquad \forall j \in J \tag{4.5c}$$

$$\sum_{i\in I} \xi_i y_{ij} = \rho_j\mu_j \qquad \forall j \in J \tag{4.5d}$$

$$y_{ij} \in \{0,1\},\ \mu_j \geq 0,\ \ 0 \leq \rho_j < 1 \qquad \forall i \in I, \forall j \in J. \tag{4.5e}$$

Note that in problem (4.5), if $R = 1$, and we simplify the term $\left[\frac{\rho_j^2}{1-\rho_j} + \rho_j\right]$ in the objective function (4.5a), (4.5) reduces to the nominal problem (3.1) in the M/M/1 model.

4.1.1 A Piecewise Linear Approximation

In this section, we apply an approximate solution approach to solve the nominal problem (4.5) based on a piecewise linearization of the nonlinear function $g(\rho) = \frac{\rho^2}{1-\rho}$. First, we show how the breakpoints of linear segments are defined so that the approximation error does not exceed a predefined threshold ϵ, *i.e.,* the piecewise-linear function $\hat{g}$ should satisfy $0 \leq g(\rho) - \widehat{g}(\rho) \leq \epsilon$ for every possible ρ, an approach that was first applied in Elhedli's work [31]. After identifying the breaking points, we provide the approximated model's formulation using special ordered sets of type 2 (SOS2) introduced by Beale and Forrest [10].

Let us assume that $\widehat{g}$, the piecewise-linear function, has $n+1$ breakpoints located at $p_0, p_1, ..., p_n$, and its line segments are tangent to the original function g at n points $q_1, q_2, ..., q_n$ where $p_{k-1} < q_k < p_k$. Assuming, without loss of generality, that $p_0 = 0$, and given that $\widehat{g}$ is linear in the interval $[p_{k-1}, p_k]$, it is possible to find both q_k and p_k when p_{k-1} is known. As a result, all the breakpoints and points of tangency can be identified recursively. Each line segment of $\widehat{g}$ can be divided into two smaller parts at the tangency point, *i.e.*, the line segment in the interval $[p_{k-1}, p_k]$ can be divided into two parts in intervals $[p_{k-1}, q_k]$ and $[q_k, p_k]$, respectively. To find q_k, we consider the first part and solve

$$g(q_k) = \widehat{g}(p_{k-1}) + g'(q_k)(q_k - p_{k-1}), \tag{4.6}$$

which is the equation of the segment in the interval $[p_{k-1}, q_k]$. Then, using the calculated q_k and considering the line segment in the interval $[q_k, p_k]$, we can find p_k. First, consider the segment in the interval $[q_k, p_k]$

$$\widehat{g}(p_k) = g(q_k) + g'(q_k)(p_k - q_k), \tag{4.7}$$

and as we want to ensure that the difference between g and $\widehat{g}$ does not exceed ϵ, we should substitute (4.7) in the following equation

$$g(p_k) = \widehat{g}(p_k) + \epsilon. \tag{4.8}$$

Thus, (4.8) becomes

$$g(p_k) = g(q_k) + g'(q_k)(p_k - q_k) + \epsilon. \tag{4.9}$$

Now we can find p_k by solving (4.9). For the next recursion, we can similarly use p_k to find q_{k+1} and p_{k+1}, and so on. This algorithm terminates when p_k reaches or exceeds a pre-defined upper limit of p. Using the results (4.6) and (4.9) for $g(\rho)$, the approximation formulas are

$$\begin{aligned} \frac{q_k^2}{1-q_k} &= \widehat{g}(p_{k-1}) + \left[\frac{1}{(1-q_k)^2} - 1\right](q_k - p_{k-1}) \\ \frac{p_k^2}{1-p_k} &= \frac{q_k^2}{1-q_k} + \left[\frac{1}{(1-q_k)^2} - 1\right](p_k - q_k) + \epsilon, \end{aligned}$$

and as $\rho < 1$, the stopping criterion is selected to be $p \geq 0.99$. If $\rho = 1$, $g(\rho)$ goes to infinity, and the system becomes unstable. Hence, $p = 1$ can not be included in the set of breaking points as it violates constraint (4.8).

Moreover, since $\rho < 1$, constraint (4.5c) becomes redundant in the presence of constraint (4.5d), and can be eliminated. Besides, since $y_{ij} \in \{0, 1\}$, it can be replaced with y_{ij}^2. Therefore, the piecewise approximation of (4.5) can be reformulated as

$$\min_{\mathrm{y},\boldsymbol{\mu},\boldsymbol{\rho},\boldsymbol{\theta},\boldsymbol{\lambda}} \sum_{j \in J} f_j \mu_j + \sum_{j \in J} \sum_{i \in I} c_{ij} \xi_i y_{ij} + tR \sum_{j \in J} \theta_j + t \sum_{j \in J} \rho_j \tag{4.10a}$$

$$\text{s.t.} \quad \sum_{j \in J} y_{ij} = 1 \qquad \forall i \in I \tag{4.10b}$$

$$\rho_j = \sum_{k \in K} \lambda_{jk} p_k \qquad \forall j \in J \tag{4.10c}$$

$$\theta_j = \sum_{k \in K} \lambda_{jk} \widehat{g}(p_k) \qquad \forall j \in J \tag{4.10d}$$

$$\sum_{k \in K} \lambda_{jk} = 1 \qquad \forall j \in J \tag{4.10e}$$

$$\sum_{i \in I} \xi_i y_{ij}^2 \leq \rho_j \mu_j \qquad \forall j \in J \tag{4.10f}$$

$$\lambda_{jk} \geq 0, \text{SOS2} \qquad \forall j \in J, \forall k \in K \tag{4.10g}$$

$$y_{ij} \in \{0, 1\}, \ \mu_j, \theta_j \geq 0, \ \ 0 \leq \rho_j < 1 \qquad \forall i \in I, \forall j \in J. \tag{4.10h}$$

In this formulation, the objective function (4.10a) is linear in $\mathrm{y}, \boldsymbol{\mu}, \boldsymbol{\rho}, \boldsymbol{\theta}$, and λ, whereas constraint (4.10f) is a second-order cone constraint that is converted from equality to inequality. Since problem (4.10), tries to minimize the objective function over $\boldsymbol{\mu}$ and $\boldsymbol{\rho}$, which have positive coefficients in the objective function, it forces $\boldsymbol{\mu}$ and $\boldsymbol{\rho}$ to take the minimum values possible; thus, the equality holds at optimality. Constraints (4.10b) ensures that every demand zone is assigned to only one SC. Besides, constraints (4.10c)-(4.10e) are SOS2 constraints using $|K|$ breakpoints and the variable $\boldsymbol{\lambda}$ is the ordered set of non-negative variables λ_{jk}, of which at most two consecutive ones can be non-zero.

4.2 The Robust Optimization (RO) Problem

In this section, we introduce the robust counterpart of (4.10), which can be stated as follows:

$$\min_{\mathrm{y},\mu,\rho,\theta,\lambda} \quad \sum_{j\in J} f_j\mu_j + tR\sum_{j\in J}\theta_j + t\sum_{j\in J}\rho_j + \sup_{\xi\in\Xi}\sum_{j\in J}\sum_{i\in I} c_{ij}\xi_i y_{ij} \tag{4.11a}$$

$$\text{s.t.} \quad \sum_{j\in J} y_{ij} = 1 \qquad \forall i \in I \tag{4.11b}$$

$$\rho_j = \sum_{k\in K}\lambda_{jk}p_k \qquad \forall j \in J \tag{4.11c}$$

$$\theta_j = \sum_{k\in K}\lambda_{jk}\widehat{g}(p_k) \qquad \forall j \in J \tag{4.11d}$$

$$\sum_{k\in K}\lambda_{jk} = 1 \qquad \forall j \in J \tag{4.11e}$$

$$\sup_{\xi\in\Xi}\sum_{i\in I}\xi_i y_{ij}^2 \leq \rho_j\mu_j \qquad \forall j \in J \tag{4.11f}$$

$$\lambda_{jk} \geq 0, \mathrm{SOS2} \qquad \forall j \in J, \forall k \in K \tag{4.11g}$$

$$y_{ij} \in \{0,1\},\ \mu_j,\theta_j \geq 0,\ \ 0 \leq \rho_j < 1 \qquad \forall i \in I, \forall j \in J. \tag{4.11h}$$

Solving this problem provides a conservative approximation of the robust counterpart of the nominal problem (4.10) as the objective function (4.11a) and constraints (4.11f) are going to be robustified individually. Similar to the M/M/1 case, we consider two classes of uncertainty sets: Budgeted and Ball uncertainty set, and show how the robustified problem can be tractably formulated as a mixed-integer second-ordered cone programming problem.

4.2.1 Budgeted Uncertainty Set

To tractably reformulate (4.11), we need to reformulate the objective function (4.11a) and constraint (4.11f), which are both linear in ξ.

Note that the term $\sup_{\xi\in\Xi} \sum_{j\in J}\sum_{i\in I} c_{ij}\xi_i y_{ij}$ can be written as

$$\begin{aligned}
&\sum_{j\in J}\sum_{i\in I} c_{ij}\xi_i^{nom} y_{ij} + \sup_{w_i} \sum_{j\in J}\sum_{i\in I} c_{ij}\widehat{\xi}_i w_i y_{ij} \\
\text{s.t.} &\sum_i w_i \le \Gamma && (\theta') \\
&0 \le w_i \le 1 && (\phi_i'),
\end{aligned}$$

and its robust counterpart can be obtained directly through LP duality as

$$\sum_{j\in J}\sum_{i\in I} c_{ij}\xi_i^{nom} y_{ij} + \inf_{\theta',\phi'} (\Gamma\theta' + \sum_i \phi_i') \tag{4.12a}$$

$$\text{s.t.} \quad \theta' + \phi_i' \ge \sum_j c_{ij}\widehat{\xi}_i y_{ij} \qquad \forall i \in I \tag{4.12b}$$

$$\theta', \phi_i' \ge 0 \qquad \forall i \in I. \tag{4.12c}$$

Moreover, by using the same approach, the left hand side of constraint (4.11f) becomes

$$\begin{aligned}
&\sum_{i\in I} \xi_i^{nom} y_{ij}^2 + \sup_{w_i} \sum_{i\in I} \widehat{\xi}_i w_i y_{ij}^2 \\
\text{s.t.} &\sum_i w_i \le \Gamma && (\gamma_j) \\
&0 \le w_i \le 1 && (\eta_{ij}).
\end{aligned}$$

and its robust counterpart can be obtained directly through LP duality as

$$\sum_{i\in I} \xi_i^{nom} y_{ij}^2 + \inf_{\gamma,\eta} \left(\Gamma\gamma_j + \sum_{i\in I} \eta_{ij}\right) \tag{4.13a}$$

$$\text{s.t.} \quad \gamma_j + \eta_{ij} \ge \widehat{\xi}_i y_{ij} \qquad \forall i \in I, \forall j \in J \tag{4.13b}$$

$$\gamma_j, \eta_{ij} \ge 0 \qquad \forall i \in I, \forall j \in J. \tag{4.13c}$$

By substituting results (4.12) and (4.13) into the objective function (4.11a) and constraint set (4.11f) respectively, (4.11) can be tractably reformulated as

$$
\begin{aligned}
\min_{y,\mu,\rho,\theta,\lambda,\theta',\phi',\gamma,\eta} \quad & \sum_{j\in J} f_j\mu_j + \sum_{j\in J}\sum_{i\in I} c_{ij}\xi_i^{nom} y_{ij} + \Gamma\theta' + \sum_{i\in I}\phi_i' && && (4.14a)\\
& + tR\sum_{j\in J}\theta_j + t\sum_{j\in J}\rho_j \\
\text{s.t.} \quad & \sum_{j\in J} y_{ij} = 1 && \forall i\in I && (4.14b)\\
& \theta' + \phi_i' \geq \sum_{j\in J} c_{ij}\xi_i y_{ij} && \forall i\in I && (4.14c)\\
& \rho_j = \sum_{k\in K}\lambda_{jk}p_k && \forall j\in J && (4.14d)\\
& \theta_j = \sum_{k\in K}\lambda_{jk}\widehat{g}(p_k) && \forall j\in J && (4.14e)\\
& \sum_{k\in K}\lambda_{jk} = 1 && \forall j\in J && (4.14f)\\
& \sum_{i\in I}\xi_i^{nom}y_{ij}^2 + \Gamma\gamma_j + \sum_{i\in I}\eta_{ij} \leq \rho_j\mu_j && \forall j\in J && (4.14g)\\
& \gamma_j + \eta_{ij} \geq \widehat{\xi}_i y_{ij} && \forall i\in I, \forall j\in J && (4.14h)\\
& \lambda_{jk}\geq 0, \mathrm{SOS2} && \forall j\in J, \forall k\in K && (4.14i)\\
& y_{ij}\in\{0,1\},\ \mu_j,\theta_j,\theta',\phi_i',\gamma_j,\eta_{ij}\geq 0,\ 0\leq\rho_j<1 && \forall i\in I,\forall j\in J. && (4.14j)
\end{aligned}
$$

The objective function is linear, whereas constraint (4.14g) is a second-order cone constraint.

4.2.2 Ball Uncertainty Set

To tractably reformulate (4.11) using the Ball uncertainty set, we need to robustify the objective function (4.11a) and constraint (4.11f) which are both linear in ξ.

First, let us consider the objective function, in which $\sup_{\boldsymbol{\xi}\in\Xi} \sum\limits_{j\in J}\sum\limits_{i\in I} c_{ij}\xi_i y_{ij}$ can be written as

$$\sup_{\widehat{\boldsymbol{\xi}}} \sum_{j\in J}\sum_{i\in I} c_{ij}\widehat{\xi}_i y_{ij} + \sum_{j\in J}\sum_{i\in I} c_{ij}\xi_i^{nom} y_{ij} \tag{4.15}$$

$$\text{s.t. } \|\widehat{\boldsymbol{\xi}}\|_2 \leq r. \tag{4.16}$$

where $\sup_{\|\widehat{\boldsymbol{\xi}}\|_2\leq r} \sum\limits_{j\in J}\sum\limits_{i\in I} c_{ij}\widehat{\xi}_i y_{ij}$ evaluates to $r\|\mathrm{c}^\intercal \mathrm{y}\|_2$.

Next, let us define $u \geq \sqrt{\sum\limits_{j\in J}\sum\limits_{i\in I} c_{ij}^2 y_{ij}^2}$, replace $\sqrt{\sum\limits_{j\in J}\sum\limits_{i\in I} c_{ij}^2 y_{ij}^2}$ with u, and add this constraint to the mathematical model. Thus, the objective function becomes

$$\min_{\mathrm{y},\boldsymbol{\mu},\boldsymbol{\rho},\boldsymbol{\theta},u} \sum_{j\in J} f_j\mu_j + \sum_{j\in J}\sum_{i\in I} c_{ij}\xi_i^{nom} y_{ij} + ru + tR\sum_{j\in J}\theta_j + t\sum_{j\in J}\rho_j \tag{4.17}$$

$$\text{s.t.} \quad \sum_{j\in J}\sum_{i\in I} c_{ij}^2 y_{ij}^2 \leq u^2. \tag{4.18}$$

Moving to the left hand side of constraint (4.11f), it can be written as

$$\sup_{\widehat{\boldsymbol{\xi}}} \sum_{i\in I} \widehat{\xi}_i y_{ij} + \sum_{i\in I} \xi_i^{nom} y_{ij}^2 \tag{4.19}$$

$$\text{s.t. } \|\widehat{\boldsymbol{\xi}}\|_2 \leq r. \tag{4.20}$$

where $\sup_{\|\widehat{\boldsymbol{\xi}}\|_2\leq r} \sum\limits_{i\in I} \widehat{\xi}_i y_{ij}$ evaluates to $r\|\mathrm{y}\|_2$. By defining $u'_j \geq \sqrt{\sum\limits_{i\in I} y_{ij}^2}$, replacing $\sqrt{\sum\limits_{i\in I} y_{ij}^2}$ with u'_j in constraint (4.11f), and adding it back to the model, the robust counterpart of problem (4.11) becomes

$$
\begin{aligned}
\min_{\mathbf{y},\boldsymbol{\mu},\boldsymbol{\rho},\boldsymbol{\theta},\boldsymbol{\lambda},u,\mathbf{u}'} \quad & \sum_{j \in J} f_j \mu_j + \sum_{j \in J} \sum_{i \in I} c_{ij} \xi_i^{nom} y_{ij} + ru + tR \sum_{j \in J} \theta_j && && (4.21a)\\
& + t \sum_{j \in J} \rho_j \\
\text{s.t.} \quad & \sum_{j \in J} y_{ij} = 1 && \forall i \in I && (4.21b)\\
& \sum_{j \in J} \sum_{i \in I} c_{ij}^2 y_{ij}^2 \leq u^2 && && (4.21c)\\
& \rho_j = \sum_{k \in K} \lambda_{jk} p_k && \forall j \in J && (4.21d)\\
& \theta_j = \sum_{k \in K} \lambda_{jk} \widehat{g}(p_k) && \forall j \in J && (4.21e)\\
& \sum_{k \in K} \lambda_{jk} = 1 && \forall j \in J && (4.21f)\\
& \sum_{i \in I} \xi_i^{nom} y_{ij}^2 + ru'_j \leq \rho_j \mu_j && \forall j \in J && (4.21g)\\
& {u'_j}^2 \geq \sum_{i \in I} y_{ij}^2 && \forall j \in J && (4.21h)\\
& \lambda_{jk} \geq 0, \mathrm{SOS2} && \forall j \in J, \forall k \in K && (4.21i)\\
& y_{ij} \in \{0,1\},\ \mu_j, \theta_j, u, u'_j \geq 0,\ \ 0 \leq \rho_j < 1 && \forall i \in I, \forall j \in J. && (4.21j)
\end{aligned}
$$

The objective function is linear, whereas constraint (4.21c), (4.21g), and (4.21h) are second-order cone constraints.

4.3 The Distributionally-Robust Optimiztion (DRO) Problem

Starting with the reformulated problem (4.10), the single-stage distributionally-robust SSDP can be stated as

$$
\begin{aligned}
\min_{\mathrm{y},\mu,\rho,\theta,\lambda} \quad & \sum_{j\in J} f_j\mu_j + \sup_{F_\xi\in\mathcal{D}_\epsilon(\widehat{F}_\xi)} \mathbb{E}_{F_\xi}\Big[\sum_{j\in J}\sum_{i\in I} c_{ij}\xi_i y_{ij}\Big] + tR\sum_{j\in J}\theta_j && (4.22a)\\
& + t\sum_{j\in J}\rho_j \\
\text{s.t.} \quad & \sum_{j\in J} y_{ij} = 1 && \forall i\in I \quad (4.22b)\\
& \rho_j = \sum_{k\in K}\lambda_{jk}p_k && \forall j\in J \quad (4.22c)\\
& \theta_j = \sum_{k\in K}\lambda_{jk}\hat{g}(p_k) && \forall j\in J \quad (4.22d)\\
& \sum_{k\in K}\lambda_{jk} = 1 && \forall j\in J \quad (4.22e)\\
& \sup_{F_\xi\in\mathcal{D}_\epsilon(\widehat{F}_\xi)} \mathbb{E}_{F_\xi}\Big[\sum_{i\in I}\xi_i y_{ij}^2\Big] \le \rho_j\mu_j && \forall j\in J \quad (4.22f)\\
& \lambda_{jk}\ge 0, \mathrm{SOS2} && \forall j\in J, \forall k\in K \quad (4.22g)\\
& y_{ij}\in\{0,1\},\ \mu_j,\theta_j\ge 0,\ 0\le\rho_j<1 && \forall i\in I, \forall j\in J. \quad (4.22h)
\end{aligned}
$$

Under Assumption (3.1), and focusing on the distributionally-robust objective function (4.22a), we utilize the following Corollary from [35]:

Corollary 4.1 ([35],Corollary 5.1)**.** *Suppose that the uncertainty set is a polytope, that is,* $\Xi = \{\xi\in\mathbb{R}^m : \mathrm{C}\xi\le \mathrm{d}\}$ *where* C *is a matrix and* d *a vector of appropriate dimensions, and consider the affine function* $a(\xi) := a^\intercal\xi + b$*. The worst-case* *expectation* $sup_{F_{\xi}\in\mathcal{D}_\epsilon(\widehat{F}_{\xi})}\mathbb{E}_{F_{\xi}}\big[a(\boldsymbol{\xi})\big]$ *evaluates to*

$$\begin{aligned}
&\inf_{\tau,s_n,\pi_n} \tau\epsilon + \frac{1}{N}\sum_{n\in N} s_n \\
&s.t. \quad b + a^{\intercal}\widehat{\boldsymbol{\xi}}^n + \pi_n(\mathrm{d} - \mathrm{C}^{\intercal}\widehat{\boldsymbol{\xi}}^n) \le s_n \quad \forall n \in N \\
&\qquad \|\mathrm{C}^{\intercal}\pi_n - a\|_* \le \tau \quad \forall n \in N \\
&\qquad \pi_n \ge 0 \quad \forall n \in N.
\end{aligned}$$

Thus, according to this Corollary, objective function (4.22a) can be tractably reformulated as

$$\begin{aligned}
\min_{\mathrm{y},\boldsymbol{\mu},\boldsymbol{\rho},\boldsymbol{\theta},\tau,\mathrm{s},\pi\ge 0} & \sum_{j\in J} f_j\mu_j + \tau\epsilon + \frac{1}{N}\sum_{n\in N} s_n + tR\sum_{j\in J}\theta_j + t\sum_{j\in J}\rho_j && (4.23)\\
\text{s.t.} \quad & \sum_{j\in J}\sum_{i\in I} c_{ij}\widehat{\xi}_i^n y_{ij} + \sum_{l\in L}\left(d_l - \sum_{i\in I} C_{li}\widehat{\xi}_i^n\right)\pi_{ln} \le s_n \quad \forall n \in N \\
& \left|\sum_{l\in L} C_{li}\pi_{ln} - \sum_{j\in J} c_{ij}y_{ij}\right| \le \tau \quad \forall i \in I, \forall n \in N.
\end{aligned}$$

The norm constraint in the Corollary reduces to a constraint on the absolute value since l_∞-norm is the dual norm in this case. Now, moving to the distributionally-robust constraint (4.22f) and using the Corollary again, $\sup\limits_{F_{\xi}\in\mathcal{D}_\epsilon(\widehat{F}_{\xi})} \mathbb{E}_{F_{\xi}}\Big[\sum\limits_{i\in I}\xi_i y_{ij}^2\Big]$ can be tractably reformulated as

$$\inf_{\tau',s',\pi'\ge 0} \tau'_j\epsilon + \frac{1}{N}\sum_{n\in N} s'_{nj} \tag{4.24a}$$

$$\text{s.t.} \quad \sum_{i\in I}\widehat{\xi}_i^n y_{ij} + \sum_{l\in L}\left(d_l - \sum_{i\in I} C_{li}\widehat{\xi}_i^n\right)\pi'_{lnj} \le s'_{nj} \quad \forall n \in N, \forall j \in J \tag{4.24b}$$

$$\left|\sum_{l\in L} C_{li}\pi'_{lnj} - y_{ij}\right| \le \tau'_j \quad \forall i \in I, \forall j \in J, \forall n \in N. \tag{4.24c}$$

and by defining $a'_{nj} = s'_{nj} - \sum\limits_{i\in I}\widehat{\xi}_i^n y_{ij}$, we can rewrite (4.24a) and (4.24b) as follow

$$\inf_{\tau',s',\pi'\ge 0} \tau'_j\epsilon + \frac{1}{N}\sum_{n}\big(a'_{nj} + \sum_{i\in I}\widehat{\xi}_i^n y_{ij}^2\big) \tag{4.25a}$$

$$\text{s.t.} \quad \sum_{l\in L}\left(d_l - \sum_{i\in I} C_{li}\widehat{\xi}_i^n\right)\pi'_{lnj} \le a'_{nj} \quad \forall n \in N, \forall j \in J. \tag{4.25b}$$

Using results (4.23), and (4.25), problem (4.22) can be tractably reformulated as

$$\min \quad \sum_{j\in J} f_j\mu_j + \tau\epsilon + \frac{1}{N}\sum_{n\in N} s_n + tR\sum_{j\in J}\theta_j + t\sum_{j\in J}\rho_j \tag{4.26a}$$

$$\text{s.t.} \quad \sum_{j\in J}\sum_{i\in I} c_{ij}\widehat{\xi}_i^n y_{ij} + \sum_{l\in L}\left(d_l - \sum_{i\in I} C_{li}\widehat{\xi}_i^n\right)\pi_{ln} \le s_n \qquad \forall n\in N \tag{4.26b}$$

$$\sum_{l\in L} C_{li}\pi_{ln} - \sum_{j\in J} c_{ij}y_{ij} \le \tau \qquad \forall i\in I, \forall n\in N \tag{4.26c}$$

$$\sum_{j\in J} c_{ij}y_{ij} - \sum_{l\in L} C_{li}\pi_{ln} \le \tau \qquad \forall i\in I, \forall n\in N \tag{4.26d}$$

$$\sum_{j\in J} y_{ij} = 1 \qquad \forall i\in I \tag{4.26e}$$

$$\rho_j = \sum_{k\in K}\lambda_{jk}p_k \qquad \forall j\in J \tag{4.26f}$$

$$\theta_j = \sum_{k\in K}\lambda_{jk}\hat{g}(p_k) \qquad \forall j\in J \tag{4.26g}$$

$$\sum_{k\in K}\lambda_{jk} = 1 \qquad \forall j\in J \tag{4.26h}$$

$$a'_{nj} = s'_{nj} - \sum_{i\in I}\widehat{\xi}_i^n y_{ij} \qquad \forall n\in N, \forall j\in J \tag{4.26i}$$

$$\frac{1}{N}\sum_{n\in N}\sum_{i\in I}\widehat{\xi}_i^n y_{ij}^2 + \tau'_j\epsilon + \frac{1}{N}\sum_{n\in N} a'_{nj} \le \rho_j\mu_j \qquad \forall j\in J \tag{4.26j}$$

$$\sum_{l\in L}\left(d_l - \sum_{i\in I} C_{li}\widehat{\xi}_i^n\right)\pi'_{lnj} \le a'_{nj} \qquad \forall n\in N, \forall j\in J \tag{4.26k}$$

$$\sum_{l\in L} C_{li}\pi'_{lnj} - y_{ij} \le \tau'_j \qquad \forall i\in I, \forall j\in J, \forall n\in N \tag{4.26l}$$

$$y_{ij} - \sum_{l\in L} C_{li}\pi'_{lnj} \le \tau'_j \qquad \forall i\in I, \forall j\in J, \forall n\in N \tag{4.26m}$$

$$\lambda_{jk} \ge 0, \mathrm{SOS2} \qquad \forall j\in J, \forall k\in K \tag{4.26n}$$

$$y_{ij}\in\{0,1\},\ \mu_j,\theta_j,\tau,s_n,\pi_{ln},\tau'_j,\pi'_{lnj}\ge 0,\ 0\le\rho_j<1 \quad \forall i\in I, \forall j\in J, \forall n\in N, \forall l\in L. \tag{4.26o}$$

Again, this is a mixed-integer second-order conic programming problem.

4.4 Solution Method

In this section, we propose a Lagrangian Relaxation (LR) approach to solve the deterministic and robust optimization models for the G/M/1 problem. This method enables us to decompose problems (4.10), (4.14), and (4.21) into smaller problems, which make it easier to solve. Moreover, as the solution obtained from the LR is, in general, not feasible, we use Dantzing-Wolfe decomposition to get a feasible solution for these problems.

4.4.1 Deterministic Problem

Consider problem (4.10). Then, by relaxing the constraints set (4.10b), using multiplier $\boldsymbol{\delta} \in \mathbb{R}^m_+$, we get the Lagrangian subproblem

$$
\begin{aligned}
[LSP]: \min_{\mathbf{y},\boldsymbol{\mu},\boldsymbol{\rho},\boldsymbol{\theta},\boldsymbol{\lambda}} \quad & \sum_{j\in J} f_j\mu_j + \sum_{j\in J}\sum_{i\in I} c_{ij}\xi_i y_{ij} + tR\sum_{j\in J}\theta_j + t\sum_{j\in J}\rho_j & & \text{(4.27a)}\\
& + \Big[\sum_{i\in I}\delta_i(1-\sum_{j\in J} y_{ij})\Big] & & \\
\text{s.t.} \quad & \rho_j = \sum_{k\in K}\lambda_{jk}p_k & \forall j\in J & \quad \text{(4.27b)}\\
& \theta_j = \sum_{k\in K}\lambda_{jk}\hat{g}(p_k) & \forall j\in J & \quad \text{(4.27c)}\\
& \sum_{k\in K}\lambda_{jk} = 1 & \forall j\in J & \quad \text{(4.27d)}\\
& \sum_{i\in I}\xi_i y_{ij}^2 \le \rho_j\mu_j & \forall j\in J & \quad \text{(4.27e)}\\
& \lambda_{jk}\ge 0, \mathrm{SOS2} & \forall j\in J, \forall k\in K & \quad \text{(4.27f)}\\
& y_{ij}\in\{0,1\},\ \mu_j,\theta_j\ge 0,\ 0\le\rho_j<1 & \forall i\in I, \forall j\in J. & \quad \text{(4.27g)}
\end{aligned}
$$

By simplifying the objective function, (4.27) becomes

$$[LSP]: \min_{\mathbf{y},\boldsymbol{\mu},\boldsymbol{\rho},\boldsymbol{\theta},\boldsymbol{\lambda}} \quad \sum_{j\in J} f_j\mu_j + \sum_{j\in J}\sum_{i\in I}(c_{ij}\xi_i - \delta_i)y_{ij} + tR\sum_{j\in J}\theta_j + t\sum_{j\in J}\rho_j + \sum_{i\in I}\delta_i \tag{4.28a}$$

$$\text{s.t.} \quad (4.27b) - (4.27g). \tag{4.28b}$$

Moreover, (4.28) can be decomposed by j to n subproblems. Hence, the decomposed subproblem for every potential SC location $j \in J$ is

$$[LSPj]: \beta_j = \min_{\mathbf{y},\boldsymbol{\mu},\boldsymbol{\rho},\boldsymbol{\theta},\boldsymbol{\lambda}} \quad f_j\mu_j + \sum_{i\in I}(c_{ij}\xi_i - \delta_i)y_{ij} + (tR)\theta_j + t\rho_j \tag{4.29a}$$

$$\text{s.t.} \quad \rho_j = \sum_{k\in K}\lambda_{jk}p_k \tag{4.29b}$$

$$\theta_j = \sum_{k\in K}\lambda_{jk}\hat{g}(p_k) \tag{4.29c}$$

$$\sum_{k\in K}\lambda_{jk} = 1 \tag{4.29d}$$

$$\sum_{i\in I}\xi_i y_{ij}^2 \leq \rho_j\mu_j \tag{4.29e}$$

$$\lambda_{jk} \geq 0, \text{SOS2} \qquad \forall k \in K \tag{4.29f}$$

$$y_{ij} \in \{0,1\},\ \mu_j, \theta_j \geq 0,\ 0 \leq \rho_j < 1 \qquad \forall i \in I. \tag{4.29g}$$

Solving (4.28) provides a lower bound for given $\delta_i, \forall i \in I$. Thus, to find the best (the highest) bound, we solve the Lagrangian Dual Problem (LDP)

$$\max_{\beta_j,\delta_i} \sum_{j\in J}\beta_j + \sum_{i\in I}\delta_i,$$

where β_j is the optimal value of LSP$_j$.

LDP can be reformulated as a linear program. To do that, let $H_j = \{h_j\}$ be the index set of feasible solutions of (4.29). Thus, β_j can be written as an optimization over the set H_j, *i.e.*,

$$\beta_j = \min_{h_j\in H_j} f_j\mu_j^{h_j} + \sum_{i\in I}(c_{ij}\xi_i - \delta_i)y_{ij}^{h_j} + (tR)\theta_j^{h_j} + t\rho_j^{h_j}.$$

With that, the Lagrangian Dual Problem can be formulated as

$$[DMP] : \max_{\beta,\delta} \sum_{j\in J} \beta_j + \sum_{i\in I} \delta_i \tag{4.30a}$$

$$\text{s.t.} \quad \beta_j + \sum_{i\in I} y_{ij}^{h_j}\delta_i \le f_j\mu_j^{h_j} + \sum_{i\in I} c_{ij}\xi_i y_{ij}^{h_j} + (tR)\theta_j^{h_j} + t\rho_j^{h_j} \qquad (w_{jh_j}) \qquad \forall j \in J, \forall h_j \in H_j. \tag{4.30b}$$

which is referred to as the dual master Problem.

In general, the solution obtained from the Lagrangian Relaxation is not feasible to problem (4.10) as it violates constraint (4.10b), and the optimality gap is strictly positive. Thus, to get a feasible solution, one way is to apply the Dantzing-Wolfe decomposition approach, in which we consider an integer version of the dual problem of $[DMP]$. The $[DMP]$ is an LP; hence, its dual problem (with the integrality constraint) is

$$[MP] : \min_{\mathrm{w}} \quad \sum_{j\in J}\sum_{h_j\in H_J} \Big[f_j\mu_j^{h_j} + \sum_{i\in I} c_{ij}\xi_i y_{ij}^{h_j} + (tR)\theta_j^{h_j} + t\rho_j^{h_j} \Big] w_{jh_j} \tag{4.31a}$$

$$\text{s.t.} \quad \sum_{j\in J}\sum_{h_j\in H_J} y_{ij}^{h_j} w_{jh_j} = 1 \qquad (\delta_i) \tag{4.31b}$$

$$\sum_{h_j\in H_J} w_{jh_j} = 1 \qquad (\beta_j) \tag{4.31c}$$

$$w_{jh_j} \in \{0,1\} \qquad \forall j \in J, \forall h_j \in H_j, \tag{4.31d}$$

which is called the (Dantzing-Wolfe) master problem. Note that to obtain a feasible solution for the original problem, we must force the integrality of w_{jh_j}. The description of the algorithm, known as Kelly's Cutting Plane algorithm, will be provided at the end of this section. Furthermore, the pseudocode of this algorithm, for the G/M/1 nominal problem, is shown in Algorithm 1.

4.4.2 RO Problem (Budgeted Uncertainty Set)

Recall problem (4.14). Then, by relaxing constraints sets (4.14b), and (4.14c), using multipliers $\boldsymbol{\upsilon}$ and $\boldsymbol{\chi} \in \mathbb{R}_+^m$, respectively, the Lagrangian subproblem becomes

$$
\begin{aligned}
[LSP]:\min \sum_{j\in J} f_j\mu_j &+ \sum_{j\in J}\sum_{i\in I} c_{ij}\xi_i^{nom} y_{ij} + \Gamma\theta' + \sum_{i\in I}\phi_i' + tR\sum_{j\in J}\theta_j && (4.32a)\\
&+ t\sum_{j\in J}\rho_j + \Big[\sum_{i\in I}\upsilon_i\Big(1-\sum_{j\in J} y_{ij}\Big)\Big] \\
&+ \Big[\sum_{j\in J}\sum_{i\in I} c_{ij}\widehat{\xi}_i y_{ij}\chi_i - \theta'\sum_{i\in I}\chi_i - \sum_{i\in I}\phi_i'\chi_i\Big] \\
\text{s.t. } \rho_j &= \sum_{k\in K}\lambda_{jk} p_k \qquad \forall j\in J && (4.32b)\\
\theta_j &= \sum_{k\in K}\lambda_{jk}\hat{g}(p_k) \qquad \forall j\in J && (4.32c)\\
&\sum_{k\in K}\lambda_{jk} = 1 \qquad \forall j\in J && (4.32d)\\
&\sum_{i\in I}\xi_i^{nom} y_{ij}^2 + \Gamma\gamma_j + \sum_{i\in I}\eta_{ij} \le \rho_j\mu_j \qquad \forall j\in J && (4.32e)\\
&\gamma_j + \eta_{ij} \ge \widehat{\xi}_i y_{ij} \qquad \forall i\in I, \forall j\in J && (4.32f)\\
&\lambda_{jk}\ge 0, \text{SOS2} \qquad \forall j\in J, \forall k\in K && (4.32g)\\
&y_{ij}\in\{0,1\},\ \mu_j,\theta_j,\theta',\phi_i',\gamma_j,\eta_{ij}\ge 0,\ 0\le\rho_j<1 \qquad \forall i\in I, \forall j\in J. && (4.32h)
\end{aligned}
$$

By simplifying the objective function, (4.32) becomes

$$
\begin{aligned}
[LSP]: \min_{\mathbf{y},\boldsymbol{\mu},\boldsymbol{\rho},\boldsymbol{\theta},\boldsymbol{\lambda},\theta',\boldsymbol{\phi}',\boldsymbol{\gamma},\boldsymbol{\eta}} &\sum_{j\in J} f_j\mu_j + \sum_{j\in J}\sum_{i\in I}\Big[c_{ij}\xi_i^{nom} + c_{ij}\widehat{\xi}_i\chi_i - \upsilon_i\Big]y_{ij} - \theta'\Big[\sum_{i\in I}\chi_i - \Gamma\Big] \\
&- \sum_{i\in I}\phi_i'(\chi_i - 1) + tR\sum_{j\in J}\theta_j + t\sum_{j\in J}\rho_j + \sum_{i\in I}\upsilon_i && (4.33a)\\
\text{s.t.} \qquad &(4.32b) - (4.32h). && (4.33b)
\end{aligned}
$$

This problem is feasible only when $\sum_{i\in I}\chi_i \le \Gamma$, and $\chi_i \le 1$, and they force both θ', and ϕ_i' to take value of zero, respectively. Moreover, (4.33) can be decomposed by j to

n subproblems. Hence, the decomposed subproblem for every potential SC location $j \in J$ is

$$[LSP_j] : \beta_j = \min_{\mathrm{y},\mu,\rho,\theta,\lambda,\gamma,\eta} f_j\mu_j + \sum_{i \in I} \Big[c_{ij}\xi_i^{nom} + c_{ij}\widehat{\xi}_i\chi_i - \upsilon_i\Big] y_{ij} + tR\theta_j + t\rho_j \tag{4.34a}$$

$$\text{s.t.} \qquad \rho_j = \sum_{k \in K} \lambda_{jk} p_k \tag{4.34b}$$

$$\theta_j = \sum_{k \in K} \lambda_{jk} \widehat{g}(p_k) \tag{4.34c}$$

$$\sum_{k \in K} \lambda_{jk} = 1 \tag{4.34d}$$

$$\sum_{i \in I} \xi_i^{nom} y_{ij}^2 + \Gamma\gamma_j + \sum_{i \in I} \eta_{ij} \leq \rho_j \mu_j \tag{4.34e}$$

$$\gamma_j + \eta_{ij} \geq \widehat{\xi}_i y_{ij} \qquad \forall i \in I \tag{4.34f}$$

$$\lambda_{jk} \geq 0, \mathrm{SOS2} \qquad \forall k \in K \tag{4.34g}$$

$$y_{ij} \in \{0,1\},\ \mu_j, \theta_j, \gamma_j, \eta_{ij} \geq 0, 0 \leq \rho_j < 1 \qquad \forall i \in I. \tag{4.34h}$$

Solving (4.33) provides a lower bound for given values of χ_i, and $\upsilon_i, \forall i \in I$. Thus, to find the best (highest) bound, we solve the Lagrangian Dual Problem (LDP)

$$\max_{\beta_j, \upsilon_i} \sum_{j \in J} \beta_j + \sum_{i \in I} \upsilon_i,$$

where β_j is the optimal value of LSP$_j$.

LDP can be reformulated as a linear program. To do that, let $H_j = \{h_j\}$ be the index set of feasible solutions of (4.34). Thus, β_j can be written as an optimization over the set H_j, *i.e.*,

$$\beta_j = \min_{h_j \in H_j} f_j \mu_j^{h_j} + \sum_{i \in I} \Big[c_{ij}\xi_i^{nom} + c_{ij}\widehat{\xi}_i\chi_i - \upsilon_i\Big] y_{ij}^{h_j} + (tR)\theta_j^{h_j} + t\rho_j^{h_j}.$$

With that, the Lagrangian Dual Problem can be formulated as

$$[DMP]: \max_{\chi \geq 0, \beta, \upsilon} \sum_{j \in J} \beta_j + \sum_{i \in I} \upsilon_i \tag{4.35a}$$

$$\text{s.t.} \quad \beta_j + \sum_{i \in I} y_{ij}^{h_j} \upsilon_i - \sum_{i \in I} c_{ij} \widehat{\xi}_i y_{ij}^{h_j} \chi_i \leq f_j \mu_j^{h_j} \tag{4.35b}$$

$$+ \sum_{i \in I} c_{ij} \xi_i^{nom} y_{ij}^{h_j} + (tR)\theta_j^{h_j} + t\rho_j^{h_j} \qquad \forall j \in J, \forall h_j \in H_j \quad (\alpha_{jh_j})$$

$$\sum_{i \in I} \chi_i \leq \Gamma \qquad (\omega) \tag{4.35c}$$

$$0 \leq \chi_i \leq 1 \qquad \forall i \in I \quad (\delta_i). \tag{4.35d}$$

which is referred to as the dual master Problem.

In general, the solution obtained from the Lagrangian Relaxation is not feasible to problem (4.14) as it violates constraints (4.14b) and (4.14c), and the optimality gap is strictly positive. Thus, to get a feasible solution, one way is to apply the Dantzing-Wolfe decomposition approach, in which we need to solve an integer version of the dual problem of $[DMP]$. The $[DMP]$ is an LP; hence, its dual problem (with the integrality constraint) is

$$[MP]: \min_{\alpha, \omega, \delta} \quad \sum_{j \in J} \sum_{h_j \in H_J} \Big[f_j \mu_j^{h_j} + \sum_{i \in I} c_{ij} \xi_i^{nom} y_{ij}^{h_j} + (tR)\theta_j^{h_j} + t\rho_j^{h_j} \Big] \alpha_{jh_j}$$

$$+ \omega \Gamma + \sum_{i \in I} \delta_i \tag{4.36a}$$

$$\text{s.t.} \quad - \sum_{j \in J} \sum_{h_j \in H_J} \alpha_{jh_j} (c_{ij} \widehat{\xi}_i y_{ij}^{h_j}) + \omega + \delta_i \geq 0 \qquad (\chi_i) \tag{4.36b}$$

$$\sum_{j \in J} \sum_{h_j \in H_J} y_{ij}^{h_j} \alpha_{jh_j} = 1 \qquad (\upsilon_i) \tag{4.36c}$$

$$\sum_{h_j \in H_J} \alpha_{jh_j} = 1 \qquad (\beta_j) \tag{4.36d}$$

$$\alpha_{jh_j} \in \{0, 1\}, \omega, \delta_i \geq 0 \qquad \forall i \in I, \forall h_j \in H_j. \tag{4.36e}$$

which is called the (Dantzing-Wolfe) master problem. Note that to obtain a feasible solution for the original problem, we must force the integrality of α_{jh_j}.

4.4.3 RO Problem (Ball Uncertainty Set)

Recall the robust counterpart of problem (4.21). Let us rewrite the objective function as follow

$$\min f_j\mu_j + \sum_{j\in J}\sum_{i\in I} c_{ij}\xi_i^{nom} y_{ij} + r\|\mathrm{c}^\intercal y\|_2 + tR\sum_{j\in J}\theta_j + t\sum_{j\in J}\rho_j.$$

Now, by using the Cauchy-Schwarz inequality we have

$$\|\mathrm{c}^\intercal \mathrm{y}\|_1 = \sum_{j\in J}\sum_{i\in I} \mid c_{ij}y_{ij} \mid .1 \leq \Big(\sum_{j\in J}\sum_{i\in I} \mid c_{ij}y_{ij} \mid^2\Big)^{1/2}\Big(\sum_{j\in J}\sum_{i\in I} 1^2\Big)^{1/2} = \sqrt{m\times n}\|\mathrm{c}^\intercal \mathrm{y}\|_2$$

$$\Rightarrow \frac{1}{\sqrt{m\times n}}\|\mathrm{c}^\intercal \mathrm{y}\|_1 = \frac{1}{\sqrt{m\times n}}\sum_{j\in J}\sum_{i\in I} c_{ij}y_{ij} \leq \|\mathrm{c}^\intercal \mathrm{y}\|_2.$$

Thus, we can rewrite the approximated problem for (4.21) as follow:

$$\begin{aligned}
\min_{\mathrm{y},\mu,\rho,\theta,\lambda,\mathrm{u}'} \quad & \sum_{j\in J} f_j\mu_j + \sum_{j\in J}\sum_{i\in I} c_{ij}\xi_i^{nom} y_{ij} + \frac{r}{\sqrt{m\times n}}\sum_{j\in J}\sum_{i\in I} c_{ij}y_{ij} && (4.37a)\\
& + tR\sum_{j\in J}\theta_j + t\sum_{j\in J}\rho_j && \\
\text{s.t.} \quad & \sum_{j\in J} y_{ij} = 1 \qquad \forall i\in I && (4.37b)\\
& (4.21d) - (4.21j). && (4.37c)
\end{aligned}$$

Now, by relaxing constraint (4.37b), using the multiplier $\boldsymbol{\alpha} \in \mathbb{R}^m_+$, the Lagrangian subproblem becomes

$$
\begin{aligned}
[LSP]: \min_{\mathbf{y},\boldsymbol{\mu},\boldsymbol{\rho},\boldsymbol{\theta},\boldsymbol{\lambda},\mathbf{u}'} & \sum_{j \in J} f_j \mu_j + \sum_{j \in J} \sum_{i \in I} c_{ij} \xi_i^{nom} y_{ij} + \frac{r}{\sqrt{m \times n}} \sum_{j \in J} \sum_{i \in I} c_{ij} y_{ij} && \text{(4.38a)} \\
& + tR \sum_{j \in J} \theta_j + t \sum_{j \in J} \rho_j + \Big[\sum_{i \in I} \alpha_i \Big(1 - \sum_{j \in J} y_{ij} \Big) \Big] \\
\text{s.t.} \quad & \rho_j = \sum_{k \in K} \lambda_{jk} p_k && \forall j \in J \quad \text{(4.38b)} \\
& \theta_j = \sum_{k \in K} \lambda_{jk} \widehat{g}(p_k) && \forall j \in J \quad \text{(4.38c)} \\
& \sum_{k \in K} \lambda_{jk} = 1 && \forall j \in J \quad \text{(4.38d)} \\
& \sum_{i \in I} \xi_i^{nom} y_{ij}^2 + r u'_j \leq \rho_j \mu_j && \forall j \in J \quad \text{(4.38e)} \\
& {u'_j}^2 \geq \sum_{i \in I} y_{ij}^2 && \forall j \in J \quad \text{(4.38f)} \\
& \lambda_{jk} \geq 0, \text{SOS2} && \forall j \in J, \forall k \in K \quad \text{(4.38g)} \\
& y_{ij} \in \{0,1\},\ \mu_j, \theta_j, u'_j \geq 0,\ 0 \leq \rho_j < 1 && \forall i \in I, \forall j \in J. \quad \text{(4.38h)}
\end{aligned}
$$

By simplifying the objective function, (4.38) can be written as

$$
\begin{aligned}
[LSP]: \min_{\mathbf{y},\boldsymbol{\mu},\boldsymbol{\rho},\boldsymbol{\theta},\boldsymbol{\lambda},\mathbf{u}'} & \sum_{j \in J} f_j \mu_j + \sum_{j \in J} \sum_{i \in I} \Big[c_{ij} \big(\xi_i^{nom} + \frac{r}{\sqrt{m \times n}} \big) - \alpha_i \Big] y_{ij} && \text{(4.39a)} \\
& + tR \sum_{j \in J} \theta_j + t \sum_{j \in J} \rho_j + \sum_{i \in I} \alpha_i \\
\text{s.t.} \quad & \text{(4.38b)} - \text{(4.38h)}. && \text{(4.39b)}
\end{aligned}
$$

Besides, (4.38) can be decomposed by j to n subproblems. Hence, the decomposed subproblem for every potential SC location $j \in J$ is

$$
\begin{aligned}
[LSPj] : \beta_j = \min_{\mathrm{y},\mu,\rho,\theta,\lambda,\mathrm{u}'} \quad & f_j\mu_j + \sum_{i\in I}\Big[c_{ij}(\xi_i^{nom} + \frac{r}{\sqrt{m\times n}}) - \alpha_i\Big]y_{ij} && (4.40a)\\
& + (tR)\theta_j + t\rho_j \\
\text{s.t.} \quad & \rho_j = \sum_{k\in K}\lambda_{jk}p_k && (4.40b)\\
& \theta_j = \sum_{k\in K}\lambda_{jk}\widehat{g}(p_k) && (4.40c)\\
& \sum_{k\in K}\lambda_{jk} = 1 && (4.40d)\\
& \sum_{i\in I}\xi_i^{nom}y_{ij}^2 + ru_j' \le \rho_j\mu_j && (4.40e)\\
& {u_j'}^2 \ge \sum_{i\in I}y_{ij}^2 && (4.40f)\\
& \lambda_{jk} \ge 0, \mathrm{SOS2} \qquad \forall k \in K && (4.40g)\\
& y_{ij} \in \{0,1\},\ \mu_j,\theta_j,u_j' \ge 0,\ 0 \le \rho_j < 1 \qquad \forall i \in I. && (4.40h)
\end{aligned}
$$

Solving (4.39) provides a lower bound for given $\alpha_i, \forall i \in I$. Thus, to find the best (the highest) bound, we solve the Lagrangian Dual Problem (LDP)

$$\max_{\beta_j,\alpha_i}\sum_{j\in J}\beta_j + \sum_{i\in I}\alpha_i,$$

where β_j is the optimal value of LSP$_j$.

LDP can be reformulated as a linear program. To do that, let $H_j = \{h_j\}$ be the index set of feasible solutions of (4.40). Thus, β_j can be written as an optimization over the set H_j, *i.e.*,

$$\beta_j = \min_{h_j\in H_j} f_j\mu_j^{h_j} + \sum_{i\in I}\Big[c_{ij}(\xi_i^{nom} + \frac{r}{\sqrt{I\times J}}) - \alpha_i\Big]y_{ij}^{h_j} + (tR)\theta_j^{h_j} + t\rho_j^{h_j}.$$

With that, the Lagrangian Dual Problem can be formulated as

$$[DMP]:\max_{\beta,\alpha} \sum_{j\in J}\beta_j + \sum_{i\in I}\alpha_i \tag{4.41a}$$

$$\text{s.t.}\quad \beta_j + \sum_{i\in I} y_{ij}^{h_j}\alpha_i \le f_j\mu_j^{h_j} + \sum_{i\in I}\Big[c_{ij}(\xi_i^{nom} + \frac{r}{\sqrt{I\times J}}) - \alpha_i\Big]y_{ij}^{h_j} \tag{4.41b}$$

$$+ (tR)\theta_j^{h_j} + t\rho_j^{h_j} \qquad (w_{jh_j}) \qquad \forall j\in J, \forall h_j\in H_j,$$

which is referred to as the dual master problem.

In general, the solution obtained from the Lagrangian Relaxation is not feasible to problem (4.21) as it violates the constraint (4.21b), and the optimality gap is strictly positive. Thus, to get a feasible solution, one way is to apply the Dantzing-Wolfe decomposition approach, in which we solve an integer version of the dual problem of $[DMP]$. Since the $[DMP]$ is an LP; hence, its dual problem (with the integerality constraint) is

$$[MP]:\min_{w} \sum_{j\in J}\sum_{h_j\in H_J}\Big[f_j\mu_j^{h_j} + \sum_{i\in I}\Big(c_{ij}\xi_i^{nom} + \frac{rc_{ij}}{\sqrt{I\times J}} - \alpha_i\Big)y_{ij}^{h_j} \tag{4.42a}$$

$$+ (tR)\theta_j^{h_j} + t\rho_j^{h_j}\Big]w_{jh_j}$$

$$\text{s.t.}\quad \sum_{j\in J}\sum_{h_j\in H_J} y_{ij}^{h_j} w_{jh_j} = 1 \qquad (\alpha_i) \tag{4.42b}$$

$$\sum_{h_j\in H_J} w_{jh_j} = 1 \qquad (\beta_j) \tag{4.42c}$$

$$w_{jh_j}\in\{0,1\} \qquad \forall j\in J, \forall h_j\in H_j, \tag{4.42d}$$

which is called the (Dantzing-Wolfe) master problem. Note that to obtain a feasible solution for the original problem, we must enforce the integrality of w_{jh_j}.

In all the aforementioned models presented in this section, we start with initial multipliers for the $[LSPj]$ and solve these problems to get a lower bound and a set of solutions. Then upon solving the $[DMP]$, we obtain new multipliers for the $[LSPj]$

and an upper bound. In each iteration, the new multipliers are updated and used in the subproblems to get new solutions and a new lower bound. Besides, all the solutions from the subproblems are used to generate new cuts that are added to the $[DMP]$. We iterate between theses two problems until the lower bound and the upper bound converge to the Lagrangian bound. Algorithm 1 shows a pseudocode of the solution method.

Algorithm 1: Kelly's Cutting Plane Algorithm For The G/M/1 NP

Initialization: $\delta \geq 0$, $H_j \leftarrow \emptyset$, $UB \leftarrow \infty$, $LB \leftarrow -\infty$;

while $UB - LB > \epsilon$ **do**

$\forall j \in J$, solve $[LSPj](\delta^k)$ to obtain $X^k = (\mathrm{y}^k, \mu^k, \rho^k, \theta^k)$, and LB^k.

Update the lower bound as $LB = \max(LB, LB^k)$;

Generate a new cut from X^k as $\beta \leq \mathrm{f}\mu^k + (\mathrm{c}^\intercal \xi - \delta)\mathrm{y}^k + (tR)\theta^k + t\rho^k$ and append it to $[DMP]$; *i.e.,* $H_j \leftarrow H_j \cup \{k\}$;

Solve $[DMP]$ to update UB and obtain new multipliers δ^{k+1} for the next iteration.

end

Declare LB as the Lagrangian bound.

As we mentioned earlier, we need to solve the master problem to find a set of feasible solutions for the original problem as some of the constraints are violated. For each $j \in J$, decision variables w_{jh_j}, in the Deterministic and RO-Ball master problems, and α_{jh_j}, in RO-Budgeted Mater problem, corresponding to a set of feasible solutions obtained from subproblem j in all iterations. Besides, for each j, only one w_{jh_j}, and α_{jh_j} would be equal to one, and the rest becomes zero. In problem (4.31), let $w^*_{jh_j}$ be the optimal solution of the binary master problem, then we can retrieve a feasible solution for the original problem as follows:

$$\mu_j = \sum_{h_j \in H_j} \mu_j^{h_j} w^*_{jh_j}$$
$$y_{ij} = \sum_{h_j \in H_j} y_{ij}^{h_j} w^*_{jh_j}.$$

The same applies to other cases. Moreover, the relative optimality gap is computed as the difference between the optimal value obtained from the Dantzing-Wolfe decomposition and the Lagrangian bound, divided by the Lagrangian bound. Constraints (4.31b), (4.36c), and (4.42b) guarantee that every demand zone i is assigned to one facility only. Constraint sets (4.31c), (4.36d), and (4.42c) ensure that only one assignment is selected for each facility.

Chapter 5

Numerical Results

5.1 Test Problems

We test on benchmark instances introduced in Holmberg *et al.* [39], which were originally developed for the capacitated facility location problems with single sourcing. They consist of four sets of test problems, randomly generated with different sizes and properties. To evaluate the performance of proposed models, we use two test problems with different sizes from the first set, one problem from the second set, and one from the last set of Holmberg test problems. The reason for this selection is that the test problems from these three sets are meant to test the effect of changing the setup cost (f), the capacity, and the different sizes. Instances of the same size, in each set, have the same demands and access costs. Since the models in this thesis consider the capacity as a decision variable, we only pick one instance of each size $m \times n = 50 \times 10$, 50×20, 150×30, and 200×30 to show the effect of the problem size on the computational performance with a different setup and waiting time costs. However, the third set's test problems are quite different as they have the same setup cost f and capacity for the problems with the same J, but differ in demands and access costs. The setup costs are assumed to be \$10 and \$20 per customer per unit time, the access costs are considered for per unit of demand, and the waiting cost t is assumed to be \$100 and \$500 per unit time. Moreover, for the RO models, two sizes of the uncertainty sets, which contain 70% and 90% of the sample data, are considered for the test problems. Data samples of size $N = 10$ were drawn uniformly and random from $U(0, 2\xi^{nom})$, where ξ^{nom} is the nominal (deterministic) demand.

For the DRO problem, the values of 100 and 500 are used for ϵ, which indicates our allowance for moving the masses between probability distributions, *i.e.,* as ϵ gets bigger, we could move more masses between the distributions. The realizations used in RO problems are also used here as the historical or predicted realizations of the uncertainty parameter (the demand). We used a box support set defined as $0 \leq \boldsymbol{\xi} \leq 2\boldsymbol{\xi}^{nom}$. The cut-off time and optimality gap are set to $10,000$ seconds and 0.1% for all the instances. All these models are coded in MATLAB and solved using Gurobi 9.0.1.

5.2 Results For The M/M/1 Problem

Tables 5.1 and 5.2 depict the computational results of direct solution with Gurobi for the M/M/1 Deterministic Problem with different values of t and f. The tables report the total cost($), the computation time in seconds (CPU), the number of open facilities (OF), and the optimality gap (%). Besides, the contribution of each cost component (setup cost (SC), access cost (AC), and waiting time cost (WTC)) in the objective function is reported with the maximum and minimum utilization of the open facilities (U-Max, and U-Min).

Table 5.1: Computational Performance: Deterministic Problem, t = 100

	f	m	n	TC	CPU (s)	OF	SC (%)	AC (%)	WTC (%)	U-Min (%)	U-Max (%)	Gap (%)
p1a	10	50	10	26171	2.00	5	55.63	24.15	20.22	80.97	87.43	0.0000
p1b	20	50	10	42857	13.29	4	67.95	16.54	15.52	85.75	91.20	0.0573
p2a	10	50	20	42688	56.42	7	66.88	12.35	20.77	83.39	88.45	0.0959
p2b	20	50	20	74286	64.80	4	76.87	10.55	12.59	89.82	93.62	0.0613
p3a	10	150	30	46865	10004.00	5	63.37	20.71	15.91	80.75	91.15	0.6647
p3b	20	150	30	79529	10005.00	4	74.69	13.16	12.15	90.86	93.29	0.2566
p4a	10	200	30	92561	10007.00	14	65.64	14.85	19.51	75.70	88.82	0.4833
p4b	20	200	30	160230	10002.00	11	75.84	9.92	14.24	87.94	92.80	1.0226

Table 5.2: Computational Performance: Deterministic Problem, t = 200

	f	m	n	TC	CPU (s)	OF	SC (%)	AC (%)	WTC (%)	U-Min (%)	U-Max (%)	Gap (%)
p1a	10	50	10	28297	18.88	4	51.45	25.05	23.50	75.05	83.82	0.0166
p1b	20	50	10	45611	11.48	4	63.84	15.75	20.41	78.85	88.10	0.0979
p2a	10	50	20	45736	90.52	4	62.42	17.13	20.45	81.52	88.00	0.0000
p2b	20	50	20	77475	503.06	2	73.70	13.97	12.33	91.95	92.57	0.0978
p3a	10	150	30	49838	10008.81	4	59.59	20.97	19.44	84.06	87.05	0.5482
p3b	20	150	30	83533	10010.00	4	71.11	12.55	16.34	87.29	90.75	0.4017
p4a	10	200	30	99453	10078.00	11	61.09	16.00	22.91	78.48	86.87	1.6409
p4b	20	200	30	169470	10008.00	10	71.71	10.32	17.98	82.84	90.50	2.3525

Considering these two tables and Figures 5.1 and 5.2, and putting the test problems in four different groups, the following observations can be made:

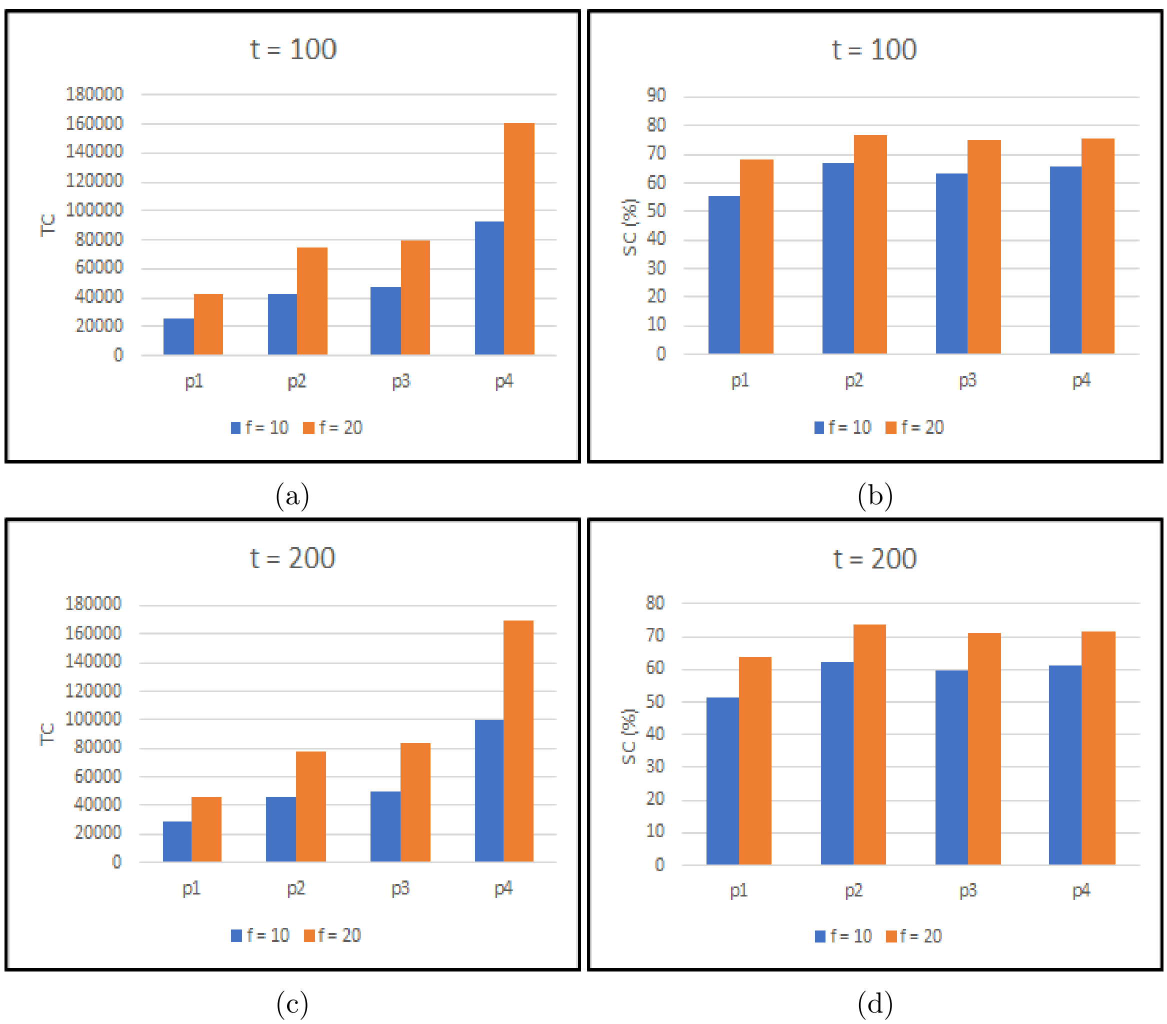


Figure 5.1: TC and SC For M/M/1 Deterministic Problem Using Different Setup Costs

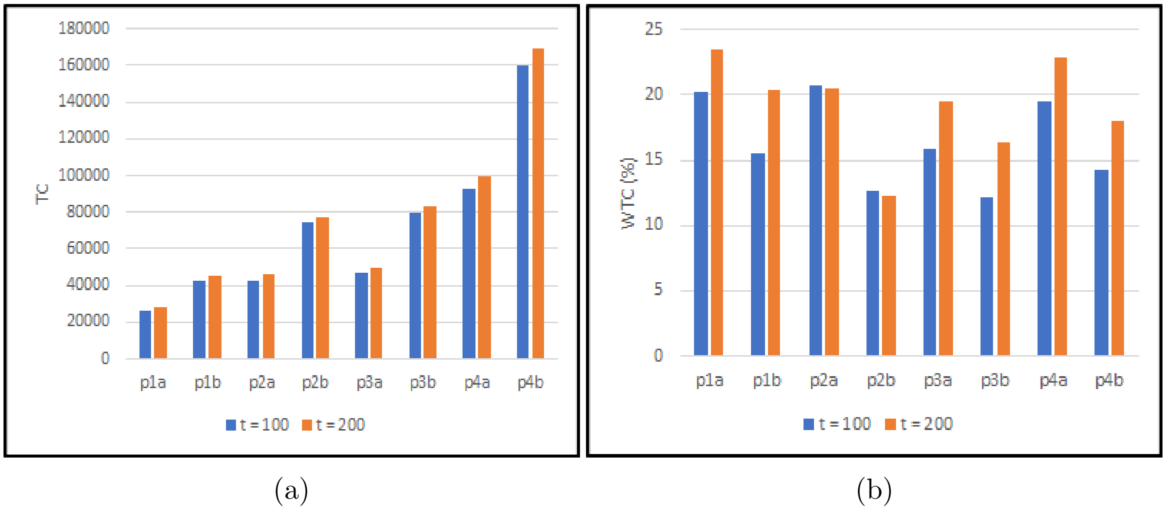


Figure 5.2: TC and WTC For M/M/1 Deterministic Problem Using Different t Values

- In each group, an increase in f increases SC and TC, which is expected. Increasing the setup cost means that opening a facility would be more expensive, and as it is proportional to the capacity, the system tries to counter this increase by attaining higher utilization of existing servers and decreasing the total capacity. Besides, attaining higher utilization leads to having smaller WTC as a percentage of TC. Here is an example from Table 5.2 that shows the decrease in the total capacity when f increases: in p1a and p1b, the system opens facilities $1, 2, 5$, and 7 for both problems, with the following capacities: p1a: 640.64, 298.97, 607.70, and 241.17, and p1b: 610.28, 279.06, 623.10, and 176.28 for facilities $1, 2, 5$, and 7, respectively.

- Increasing t leads to an increase in TC. An increase in t means a larger waiting penalty for congestion, which leads to a system with more uniform utilization among the open facilities or a system with a larger total capacity. In both cases, the maximum utilization decreases, which means fewer customers are waiting in the system. However, in most cases, WTC increases as t increases, and the reason is that a decrease in the number of customers in the system is too small compared with the increase in t, leading to an increase in WTC. As an example, for p3b with different t, the system opens the same facilities, but the total capacity for $t = 100$ is 3211.60, and when $t = 200$, the total capacity becomes 3311.27. However, in p2a, when $t = 100$, the system opens facilities $1, 10, 11, 14, 16, 18, 20$ with the total capacity of 3298.269, whereas with $t = 200$ it opens facilities $2, 7, 8, 14$ with the total capacity 3322.589.

- As the number of demand zones and the potential facility locations increases, the computational time increases and the problems become harder to solve to optimality.

Tables 5.3 and 5.4 summarize the computational results for the M/M/1 RO problem with the Budgeted uncertainty set, and Tables 5.5 and 5.6 illustrate the results for M/M/1 RO problem with Ball uncertainty set. These are the results of a direct solution with Gurobi with different values of f and t. Because of the random nature of the realizations, we report the computational results for three randomly generated realizations of data samples. According to the tables, an increase in f and t increases TC for both types of uncertainty sets. Besides, other observations can be made as follow:

- For each Trial, in both RO problems, an increase in the uncertainty budget increases TC as we become more conservative. In cases where the number of facilities remains unchanged, the total capacity increases as the uncertainty budgets increase to accommodate higher demands. Moreover, when the number of facilities increases or decreases, the total capacity also increases or decreases.
- Figure 5.3 compares the costs of using the RO problem with both uncertainty sets (we consider Trial 3 as an example) when using the same uncertainty budget. As shown in the figures, for the same problem, there is not much difference between the costs obtained from using the budgeted or ball uncertainty sets as we use the same realizations to calibrate the uncertainty sets.

Tables 5.7 and 5.8 summarize the computational results of direct solution with Gurobi for the M/M/1 DRO Problem, with different values of f, t, and ϵ. For this model, we

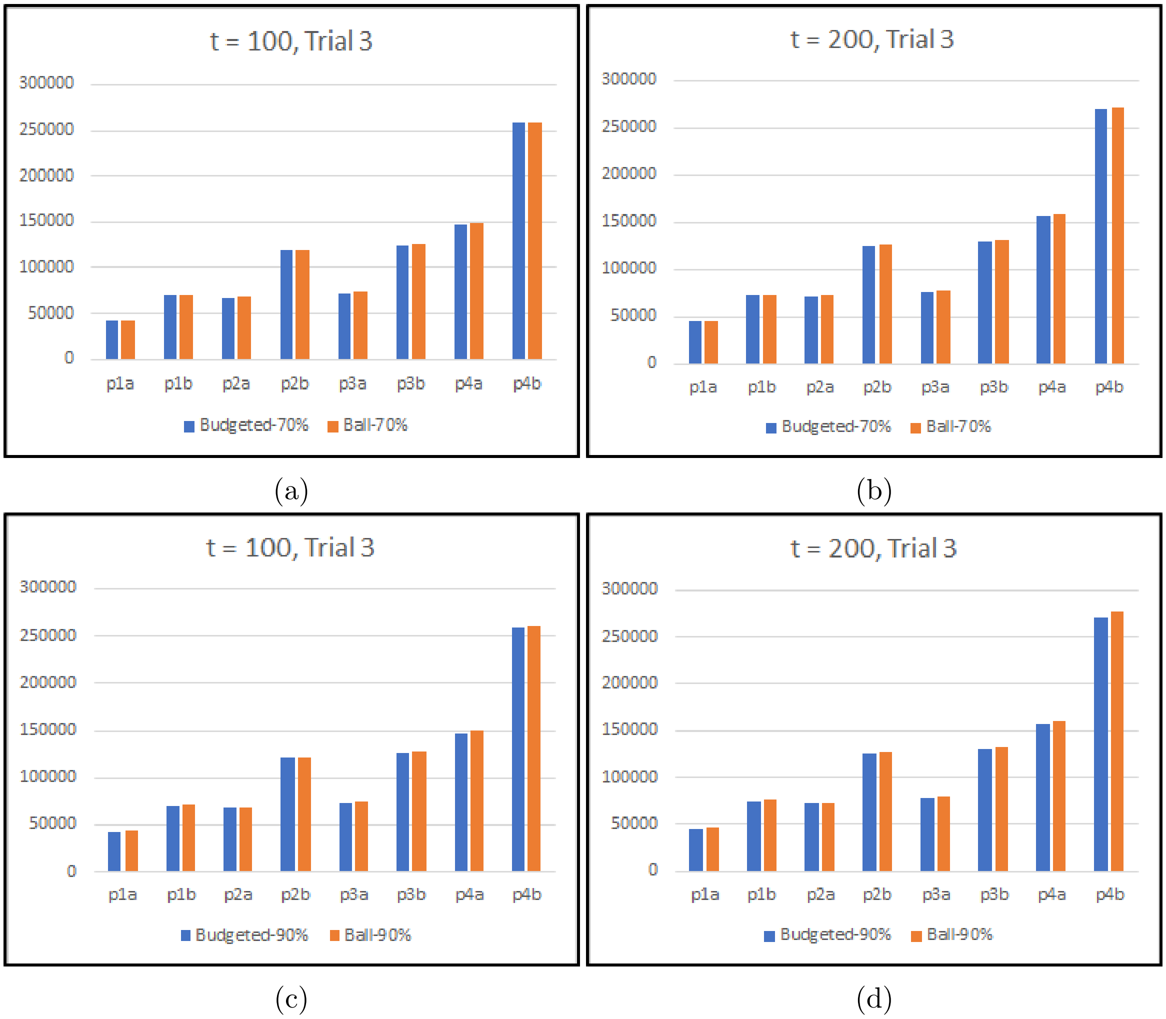


Figure 5.3: Costs of Using RO-Budgeted and RO-Ball with The Same Uncertainty Budget

could only solve the smallest instance of the selected test problems with optimality gap of less than 7% within the cut-off time. Hence, we tried smaller instances based on the Holmberg instances of different sizes $m \times n = 15 \times 5$, 25×5, and 25×10. Moreover, because of the random nature of the realizations, we report the results for three randomly generated realizations of data samples. According to the results, similar observations to the RO models could be made regarding the increase in f and t. Besides, as ϵ increases, TC also increases, which is expected. The reason for that is when we increase ϵ, we are allowing more probability mass to be transported between scenarios, including high-cost ones, which costs us more. Now, the question is how we should properly choose the ϵ. Mohajerani Esfahani and Kuhn [35] provide a formula for the out-of-sample probabilistic performance guarantees as a function of ϵ. They observe that by starting from $\epsilon = 0$, the out-of-sample performance is high, and as ϵ increases, the out-of-sample improves up to a certain point (critical Wasserstein radius) and then increases again. Thus, we should try to choose ϵ to optimize the out-of-sample performance, and ensure a certain expected out-of-sample performance guarantee; *i.e.,* the out-of-sample expected cost should not be higher than the objective value of the DRO problem. Besides, it seems here that in all three Trials for p1a and p1b, when $\epsilon = 500$, the entire probability mass was transported to the worst-case scenario, so we are solving the RO problem for these instances.

Table 5.3: Computational Performance: RO-Budgeted for Three Trials, t = 100

				$\Gamma_{70\%}$				$\Gamma_{90\%}$			
	f	m	n	TC	CPU (s)	OF	Gap(%)	TC	CPU (s)	OF	Gap(%)
p1a1	10	50	10	40392	28.62	6	0.0000	41221	37.18	6	0.0270
p1a2	10	50	10	40795	25.90	7	0.0001	41738	31.62	7	0.0359
p1a3	10	50	10	41600	33.60	6	0.0000	41959	33.84	7	0.0000
p1b1	20	50	10	67203	41.02	4	0.0681	68641	66.79	5	0.0006
p1b2	20	50	10	68020	43.87	5	0.0669	69649	35.77	5	0.0000
p1b3	20	50	10	69356	73.55	5	0.0004	69939	95.93	5	0.0767
p2a1	10	50	20	65628	8418.50	8	0.0987	66030	10011.00	8	0.2880
p2a2	10	50	20	65636	8646.00	7	0.0967	66117	8198.90	7	0.0939
p2a3	10	50	20	67275	10028.59	8	0.9339	67865	10017.00	8	0.7098
p2b1	20	50	20	116310	10008.00	4	1.1935	117040	10015.00	4	1.4605
p2b2	20	50	20	116690	10012.00	4	1.2120	117570	10013.00	4	1.3755
p2b3	20	50	20	119240	10009.00	4	1.7098	120460	10012.00	5	2.7844
p3a1	10	150	30	72591	10013.00	7	1.5312	73044	10013.00	7	1.5976
p3a2	10	150	30	72816	10010.00	7	1.4744	73066	10014.00	6	1.7001
p3a3	10	150	30	72423	10011.00	7	1.6583	73241	10010.00	6	1.5789
p3b1	20	150	30	124690	10010.00	5	1.2752	125480	10007.00	5	0.8265
p3b2	20	150	30	124900	10016.00	5	1.3684	125340	10009.00	5	1.3690
p3b3	20	150	30	124350	10011.00	5	1.2145	125800	10008.00	5	1.3414
p4a1	10	200	30	144990	10019.00	16	1.2216	146550	10015.00	15	2.1173
p4a2	10	200	30	143710	10011.00	15	1.1939	144410	10013.00	15	1.1869
p4a3	10	200	30	146920	10010.00	16	2.1956	147240	10016.00	16	1.2808
p4b1	20	200	30	255680	10008.00	13	1.3528	258510	10011.00	12	2.9755
p4b2	20	200	30	253400	10008.00	12	2.1296	254710	10017.00	12	2.9486
p4b3	20	200	30	258920	10014.00	11	2.3929	259390	10020.00	12	1.3912

Table 5.4: Computational Performance: RO-Budgeted for Three Trials, t = 200

				$\Gamma_{70\%}$				$\Gamma_{90\%}$			
	f	m	n	TC	CPU (s)	OF	Gap(%)	TC	CPU (s)	OF	Gap(%)
p1a1	10	50	10	43349	232.92	5	0.0545	44198	94.77	5	0.0010
p1a2	10	50	10	43768	79.96	5	0.0000	44748	96.72	5	0.0000
p1a3	10	50	10	44576	253.63	5	0.0489	44943	52.59	5	0.0000
p1b1	20	50	10	70740	980.86	4	0.0819	72211	224.05	4	0.0005
p1b2	20	50	10	71629	969.37	4	0.0962	73290	219.90	4	0.0758
p1b3	20	50	10	73117	1140.10	4	0.0959	73726	145.91	4	0.0021
p2a1	10	50	20	69967	10004.00	4	1.8503	70394	10017.00	4	2.1436
p2a2	10	50	20	70290	10008.00	4	2.7963	70790	10011.00	6	3.0172
p2a3	10	50	20	71737	10011.00	4	2.7351	72456	10011.00	5	2.6672
p2b1	20	50	20	121270	10005.00	4	2.8179	122010	10021.00	4	2.5912
p2b2	20	50	20	121590	10003.00	4	4.5537	122510	10004.00	4	4.0223
p2b3	20	50	20	124250	10007.00	4	4.4186	125300	10003.00	3	4.3572
p3a1	10	150	30	76696	10011.00	5	2.2043	77166	10011.00	5	2.2482
p3a2	10	150	30	76893	10015.00	5	1.9900	77159	10012.00	5	2.3063
p3a3	10	150	30	76530	10014.00	5	2.1420	77401	10010.00	5	2.2750
p3b1	20	150	30	129990	10008.00	4	3.0686	130820	10015.00	4	3.1985
p3b2	20	150	30	130191	10009.69	4	1.8567	130640	10012.00	4	3.0387
p3b3	20	150	30	129785	10026.48	4	2.1542	131280	10005.00	4	2.1478
p4a1	10	200	30	154600	10011.00	12	2.2881	156520	10011.00	13	2.4511
p4a2	10	200	30	153530	10017.00	12	3.5097	154120	10014.00	12	2.3192
p4a3	10	200	30	156560	10010.00	12	2.3585	156970	10010.00	12	2.4769
p4b1	20	200	30	266960	10015.00	8	3.5754	270000	10008.00	10	2.2324
p4b2	20	200	30	264560	10010.00	7	2.4643	265990	10009.00	9	2.2249
p4b3	20	200	30	270290	10013.00	9	2.1086	270820	10007.00	8	2.1790

Table 5.5: Computational Performance: RO-Ball for Three Trials, t = 100

				$r_{70\%}$				$r_{90\%}$			
	f	m	n	TC	CPU (s)	OF	Gap(%)	TC	CPU (s)	OF	Gap(%)
p1a1	10	50	10	41432	78.91	6	0.0033	41536	196.89	6	0.0916
p1a2	10	50	10	41930	43.44	7	0.0000	42772	69.41	8	0.0145
p1a3	10	50	10	42305	91.91	7	0.0151	43599	84.79	7	0.0000
p1b1	20	50	10	67925	2588.00	6	0.0977	68097	169.30	6	0.0204
p1b2	20	50	10	68740	5632.90	6	0.0998	70113	2558.70	6	0.0996
p1b3	20	50	10	69356	561.86	6	0.0999	71481	485.31	6	0.0621
p2a1	10	50	20	65824	10009.00	8	0.8448	66573	10010.00	8	2.2932
p2a2	10	50	20	66885	10016.35	8	0.9867	67561	10009.00	8	1.0268
p2a3	10	50	20	67424	10008.00	8	1.1160	68013	10011.00	8	1.2101
p2b1	20	50	20	116880	10012.00	4	3.1535	117740	10010.00	7	2.7803
p2b2	20	50	20	119070	10006.00	6	3.0962	120080	10007.00	5	2.7695
p2b3	20	50	20	119510	10019.00	6	2.4738	120810	10006.00	5	2.9089
p3a1	10	150	30	73426	10008.00	7	1.5370	73912	10009.00	6	1.8950
p3a2	10	150	30	73135	10009.00	7	1.7055	73671	10008.00	7	1.7382
p3a3	10	150	30	73258	10010.32	6	1.7187	74346	10017.00	7	1.7648
p3b1	20	150	30	125690	10007.00	6	1.7361	126580	10004.00	6	2.0347
p3b2	20	150	30	125290	10010.00	6	1.9193	126280	10010.00	5	1.9271
p3b3	20	150	30	125730	10020.30	6	2.3996	127330	10005.00	6	1.7641
p4a1	10	200	30	147950	10012.00	21	2.1819	149970	10011.00	21	1.4799
p4a2	10	200	30	146750	10011.00	20	2.4547	148810	10008.00	22	1.7673
p4a3	10	200	30	148610	10012.00	22	2.7922	150720	10010.00	22	3.3924
p4b1	20	200	30	258360	10008.00	15	1.5991	263320	10022.00	18	2.1781
p4b2	20	200	30	255500	10011.18	14	1.6256	264060	10011.00	20	4.2212
p4b3	20	200	30	258310	10010.00	16	1.8922	260670	10034.00	14	1.9283

Table 5.6: Computational Performance: RO-Ball for Three Trials, t = 200

				$r_{70\%}$				$r_{90\%}$			
	f	m	n	TC	CPU (s)	OF	Gap(%)	TC	CPU (s)	OF	Gap(%)
p1a1	10	50	10	44478	4320.50	6	0.0838	44585	330.26	6	0.0842
p1a2	10	50	10	44998	1476.10	6	0.0791	45873	407.34	6	0.0000
p1a3	10	50	10	45391	1756.10	6	0.0938	46745	111.79	7	0.0993
p1b1	20	50	10	71924	10006.00	4	1.4279	72142	10004.00	5	1.1347
p1b2	20	50	10	72789	10005.00	4	1.0153	74213	10005.00	4	0.9249
p1b3	20	50	10	73439	895.45	5	0.0779	75632	10012.00	5	0.7374
p2a1	10	50	20	70803	10005.00	7	3.8798	71493	10010.00	6	5.6245
p2a2	10	50	20	71991	10003.00	7	4.3702	72592	10007.00	6	3.6476
p2a3	10	50	20	72425	10005.00	7	3.6793	73046	10006.00	5	3.8198
p2b1	20	50	20	122504	10007.85	5	4.8673	123330	10009.00	4	5.1318
p2b2	20	50	20	124739	10015.83	6	5.4745	127020	10018.00	6	6.1330
p2b3	20	50	20	126120	10011.00	6	5.7015	127570	10005.00	5	6.2035
p3a1	10	150	30	77984	10007.00	6	2.9448	78333	10018.00	6	2.8361
p3a2	10	150	30	77459	10008.00	6	2.5006	78137	10011.00	6	2.7758
p3a3	10	150	30	77762	10014.63	6	2.8483	79434	10007.00	6	4.3585
p3b1	20	150	30	130860	10014.00	4	1.9835	131510	10005.00	4	0.0214
p3b2	20	150	30	130820	10006.00	5	1.8477	132530	10006.00	5	3.1304
p3b3	20	150	30	131330	10006.00	5	2.0248	132420	10007.00	4	1.5064
p4a1	10	200	30	160400	10018.00	16	5.0855	161550	10010.00	19	3.3296
p4a2	10	200	30	157730	10021.00	18	4.5382	161220	10009.00	17	5.4007
p4a3	10	200	30	158550	10016.00	16	2.8482	160040	10010.00	13	3.0893
p4b1	20	200	30	273710	10009.00	12	3.4997	278320	10007.00	13	3.8224
p4b2	20	200	30	270860	10015.00	14	3.8021	275250	10009.00	14	3.8447
p4b3	20	200	30	272040	10014.00	12	3.1091	278380	10016.00	13	4.5543

Table 5.7: Computational Performance: DRO for Three Trials, t = 100

				$\epsilon = 100$				$\epsilon = 500$			
	f	m	n	TC	CPU (s)	OF	Gap(%)	TC	CPU (s)	OF	Gap(%)
p1a1	10	15	5	11343	1211.71	5	0.0989	16177	596.78	5	0.0805
p1a2	10	15	5	10853	324.88	5	0.0851	16177	347.30	5	0.0772
p1a3	10	15	5	11153	799.30	5	0.0228	16177	573.62	5	0.0772
p1b1	20	15	5	17329	953.83	5	0.0246	25485	588.35	5	0.0863
p1b2	20	15	5	16619	1730.80	5	0.0632	25485	806.92	5	0.0859
p1b3	20	15	5	17019	4281.60	5	0.0813	25485	1022.50	5	0.0792
p2a1	10	25	5	16393	632.74	5	0.0946	23674	708.37	5	0.0775
p2a2	10	25	5	16569	746.21	5	0.0932	23899	772.79	5	0.0750
p2a3	10	25	5	16912	647.28	5	0.0694	24014	560.36	5	0.0836
p2b1	20	25	5	25230	880.09	5	0.0927	36979	1365.30	5	0.0960
p2b2	20	25	5	25751	1400.10	5	0.0805	37416	1145.80	5	0.0806
p2b3	20	25	5	26153	2802.60	5	0.0785	37582	1043.00	5	0.0911
p3a1	10	25	10	15310	7101.90	10	0.0016	22018	5773.20	10	0.0809
p3a2	10	25	10	15547	3702.80	10	0.0822	22291	8107.70	10	0.0914
p3a3	10	25	10	15774	7331.50	10	0.0832	22373	1760.30	10	0.0774
p3b1	20	25	10	24389	10001.00	10	1.6309	35879	10001.00	10	1.8664
p3b2	20	25	10	24994	10001.00	10	2.4197	36269	10004.00	10	1.3297
p3b3	20	25	10	24994	10001.00	10	2.0358	36406	10001.00	10	1.4408
p4a1	10	50	10	27846	10007.00	10	1.3027	35154	10009.00	10	1.3868
p4a2	10	50	10	27703	10002.00	10	2.7408	34843	10003.00	10	2.5811
p4a3	10	50	10	29209	10002.00	10	2.2486	36471	10002.00	10	2.6596
p4b1	20	50	10	45333	10006.00	10	1.4481	57113	10004.00	10	1.2302
p4b2	20	50	10	45172	10002.00	10	3.7029	56815	10002.00	10	2.8510
p4b3	20	50	10	48175	10002.00	10	3.4178	59803	10002.00	10	2.5798

Table 5.8: Computational Performance: DRO for Three Trials, t = 200

				$\epsilon = 100$				$\epsilon = 500$			
	f	m	n	TC	CPU (s)	OF	Gap(%)	TC	CPU (s)	OF	Gap(%)
p1a1	10	15	5	12449	1806.06	5	0.0342	17625	525.76	5	0.0116
p1a2	10	15	5	11955	1704.00	5	0.0685	17625	1360.30	5	0.0201
p1a3	10	15	5	12272	1305.30	5	0.0966	17625	705.65	5	0.0089
p1b1	20	15	5	18706	3035.19	5	0.0692	27474	1658.30	5	0.0973
p1b2	20	15	5	18016	387.27	5	0.0973	27474	431.99	5	0.0792
p1b3	20	15	5	18388	1597.50	5	0.0713	27474	1405.10	5	0.0814
p2a1	10	25	5	17657	1863.80	5	0.0850	25406	884.18	5	0.0782
p2a2	10	25	5	17993	2108.70	5	0.0288	25658	1057.00	5	0.0849
p2a3	10	25	5	18343	2237.30	5	0.0012	25773	1193.70	5	0.0898
p2b1	20	25	5	26945	1649.10	5	0.0005	39231	1908.20	5	0.0005
p2b2	20	25	5	27531	4252.90	5	0.0055	39686	1038.90	5	0.0919
p2b3	20	25	5	27915	2236.20	5	0.0831	39884	1744.80	5	0.0869
p3a1	10	25	10	16825	10002.40	10	1.9736	24266	10004.00	10	2.6749
p3a2	10	25	10	17226	10009.00	10	1.4586	24511	10001.00	10	0.5461
p3a3	10	25	10	17494	10001.00	10	1.9603	24597	10002.00	10	1.5517
p3b1	20	25	10	26524	10001.00	10	3.3008	38764	10001.00	10	3.6196
p3b2	20	25	10	27795	10001.00	10	6.1498	39377	10002.00	10	3.8310
p3b3	20	25	10	27783	10002.00	10	5.3197	39467	10003.00	10	3.0318
p4a1	10	50	10	30786	10004.16	10	4.9518	38215	10002.00	10	3.9563
p4a2	10	50	10	29965	10002.00	10	3.7541	37881	10002.00	10	4.7544
p4a3	10	50	10	31880	10002.00	10	6.4004	39762	10002.00	10	4.3208
p4b1	20	50	10	49261	10002.00	10	6.2528	61935	10002.00	10	5.6710
p4b2	20	50	10	48502	10002.00	10	5.5064	60417	10002.00	10	4.0772
p4b3	20	50	10	51432	10002.00	10	5.3749	63846	10002.00	10	4.9091

5.2.1 Deterministic VS. Uncertain Demands

In this section, we will compare the Deterministic model with the models when uncertainty is considered. For this purpose, we choose Trial 2 as an example. Figure 5.4 shows the objective function values (costs) for each problem, considering the Deterministic model and the RO models using both uncertainty sets. One may ask the reason for proposing the RO models as the costs in these problems are almost double the costs of the Deterministic problem. There are a couple of observations that can be made from the Deterministic model:

- Sensitivity to the Capacities and the Demands: Considering the expected number of customers in the system $\sum_{j\in J} \frac{\sum_{i\in I} \xi_i y_{ij}}{\mu_j - \sum_{i\in I} \xi_i y_{ij}}$, we realize that if $\mu_j = \sum_{i\in I} \xi_i y_{ij}$, this number can go to infinity; therefore, the whole system can become unstable. Hence, this problem can be further complicated and become sensitive to the capacities installed and the demand experienced. As a result, it is crucial to include uncertainty when designing a service system.

- Out-of-Sample Data: Now, let us move to the feasibility constraint $\sum_{i\in I} \xi_i y_{ij} \leq \mu_j$. Here, to have a better insight, we use test problem p1b from the Deterministic problem when $t = 100$. Considering the optimal solution (μ^*, y^*) for this problem, we generated 30 realizations from $U(0, 2\boldsymbol{\xi}^{nom})$ within the uncertainty set and the feasibility constraint was tested using these new realizations (demands). We found out that 80% of the realizations was not feasible for the Deterministic problem as the feasibility constraint was violated. Although the other 20% of the realizations were feasible, we know that the cost with the same optimal solution would be higher as they are not the initial demands that we solve the problem to optimality with.

- The Number of Open Facilities and Customers' Assignment: Comparing the results for the Deterministic Problem and RO problems, we can see that the number of open facilities can vary when considering uncertain demands. Besides, with a different number of open facilities, the assignment of customers would be different. However, there are some cases that the number of open facilities remains unchanged. In this case, the system's total with uncertain demands would be larger, and the assignment of customers could be different.

As we mentioned earlier, the RO approaches are meant to protect the model from the worst-case scenario (demand); therefore, they are considered too conservative and have a poor performance. Moreover, they are based on the assumption of having no knowledge about the probability distribution of the uncertain parameter. Thus, as an

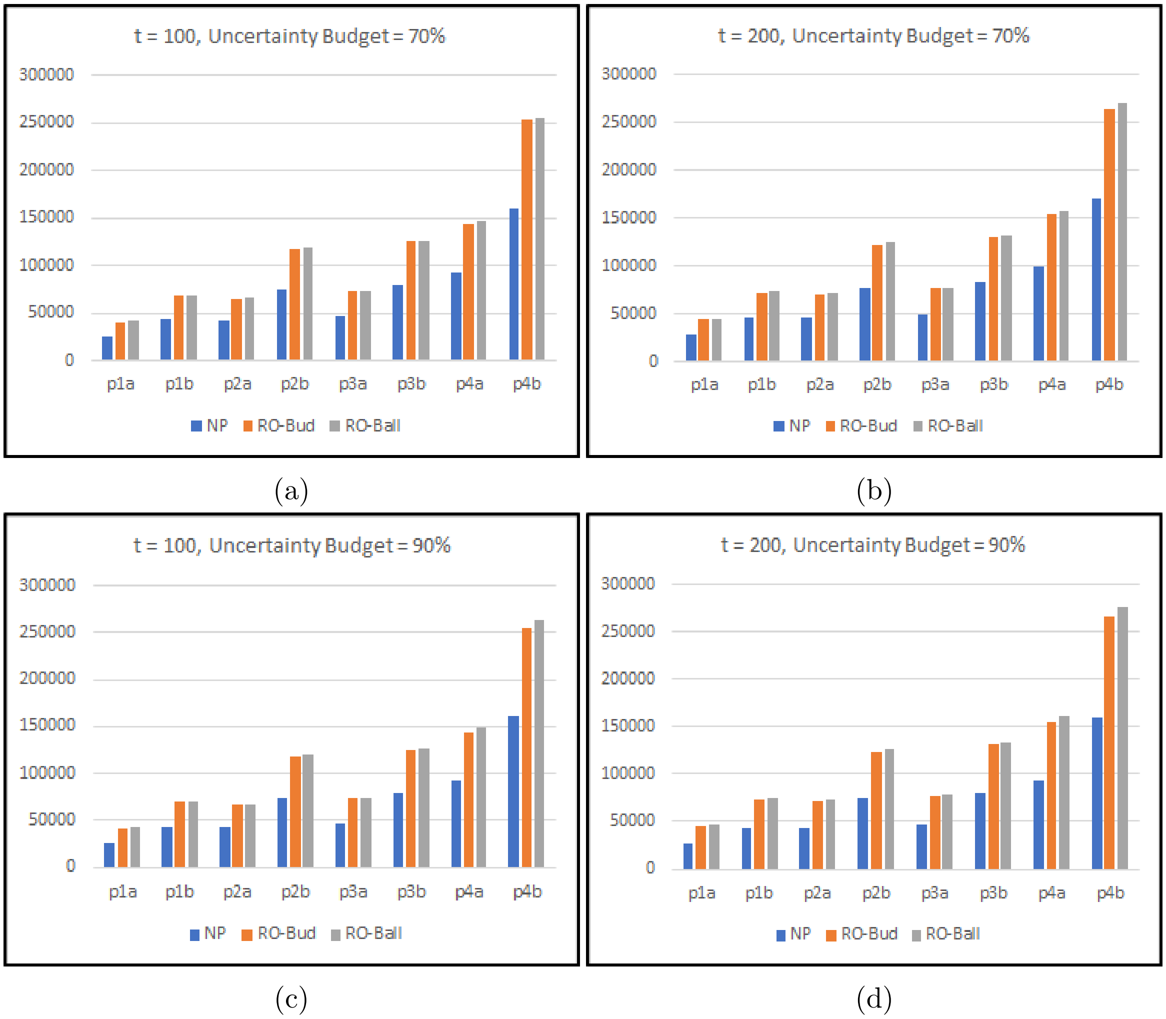


Figure 5.4: Costs of Considering The Deterministic Model VS. RO Models (Trial 2)

alternative, we proposed the DRO model. We choose two small instances, with different values of f and t, from the DRO test problems to compare the results from the Deterministic, RO, and DRO models. Figure 5.5 compares the costs obtained using these three approaches for Trial 1, in which we tried four values for ϵ to demonstrate the results better. As we use a Box as a support set in the DRO problem, we report the RO problem's results when using a Box as an uncertainty set. The observations that can be made from the results and Figure 5.5 are as follows:

- The costs attained from the DRO problems are in between the ones achieved from the Deterministic and RO problems, which is expected. DRO is still

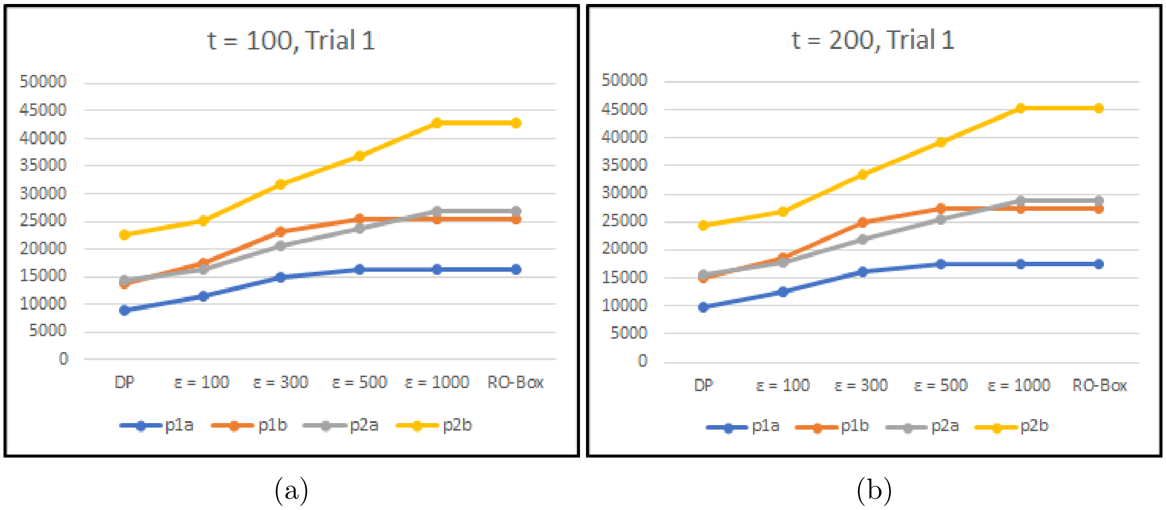


Figure 5.5: Costs Obtained From Three Models

conservative, but less conservative than the RO approach.

- As ϵ increases, the costs also increase, and at some point, when ϵ is big enough for the problems, the objective function values remain unchanged. This is when the unchanged costs would be equal to the costs obtained from the RO problem.

5.3 Results For The G/M/1 Problem

In this section, we summarize the results for the Deterministic, RO-Ball, and DRO problems. For these models, we could not solve big instances, so we generated smaller instances based on the Holmberg test problems with sizes of $m \times n = 6 \times 3, 10 \times 5, 15 \times 5,$ and 15×10 to evaluate the performance of the proposed models.

Tables 5.9 and 5.10 summarize the computational results for the G/M/1 Deterministic problem using the Lagrangian-Relaxation approach, with different values of $t, f,$ and C_a. The tables report the total cost, the number of iterations (Iter.), the computation time in seconds (CPU), the number of open facilities (OF), and the optimality gap (%). Besides, the contribution of each term (setup cost (SC), access cost (AC), and the waiting time cost (WTC)) in the objective function is reported with the maximum and minimum utilization of the open facilities (U-Max, and U-Min). Considering these two tables, and putting the test problems in four different groups, the following observations can be made:

Table 5.9: Computational Performance: Deterministic Problem, t = 100

	f	m	n	TC	Iter.	CPU (s)	OF	SC (%)	AC (%)	WTC (%)	U-Min (%)	U-Max (%)	Gap
$C_a = 0.5$													
p1a	10	6	3	4139	14	86.40	2	53.75	34.00	12.25	73.42	81.55	0.00%
p1b	20	6	3	6281	13	66.64	2	66.43	22.40	11.17	79.59	86.36	0.00%
p2a	10	10	5	5883	23	237.60	3	60.47	25.89	13.65	75.88	82.04	0.00%
p2b	20	10	5	9283	31	568.45	2	69.77	20.69	9.54	83.30	89.06	0.00%
p3a	10	15	5	13108	52	10451.00	3	68.82	21.34	9.84	84.62	87.82	0.20%
p3b	20	15	5	22204	37	10060.00	3	77.77	14.37	7.86	88.84	89.17	1.80%
p4a	10	15	10	13008	31	4053.80	5	71.44	16.16	12.40	68.67	87.27	0.00%
p4b	20	15	10	22461	30	10636.00	5	78.97	10.88	10.15	79.95	90.32	3.80%
$C_a = 2$													
p1a	10	6	3	4826	13	56.16	2	54.75	29.67	15.58	59.39	69.88	0.50%
p1b	20	6	3	7259	21	209.50	1	61.10	27.29	11.61	79.37	79.37	1.90%
p2a	10	10	5	6858	27	504.82	2	57.58	28.01	14.41	65.15	74.50	0.00%
p2b	20	10	5	10627	37	992.35	2	68.35	18.08	13.58	71.87	80.51	0.00%
p3a	10	15	5	15331	48	6253.60	4	66.06	20.55	13.38	8.60	77.46	2.10%
p3b	20	15	5	24259	55	13725.00	1	72.48	19.87	7.65	88.83	88.83	3.20%
p4a	10	15	10	14988	39	8818.10	7	66.57	20.43	13.00	13.13	50.55	0.00%
p4b	20	15	10	24259	33	13561.00	1	72.48	19.87	7.65	88.83	88.83	19.50%

- In each group, an increase in f increases SC and TC, which is expected. Increasing the setup cost means that opening a facility would be more expensive, and as it is proportional to the capacity, the system tries to counter this increase by attaining larger utilization and decreasing the total capacity. Besides, reaching larger utilization leads to having smaller WTC.

- Increasing t leads to an increase in TC. An increase in t means a larger waiting penalty for congestion, which leads to having a system with more uniform utilization among the open facilities or a system with a larger total capacity. In both cases, the maximum utilization decreases, which means fewer customers are waiting in the system. However, in most cases, WTC increases as t increases, and the reason is that an decrease in the number of customers in the system is too small compared with an increase in t.

Table 5.10: Computational Performance: Deterministic Problem, t = 200

	f	m	n	TC	Iter.	CPU (s)	OF	SC (%)	AC (%)	WTC (%)	U-Min (%)	U-Max (%)	Gap
$C_a = 0.5$													
p1a	10	6	3	4566	11	43.47	2	53.01	30.82	16.18	65.32	76.05	0.00%
p1b	20	6	3	6896	18	83.81	2	64.52	20.77	14.71	73.42	81.55	2.10%
p2a	10	10	5	6501	30	448.46	2	56.36	29.55	14.09	70.90	79.95	0.00%
p2b	20	10	5	10018	34	858.70	2	68.25	19.18	12.57	77.47	84.99	0.00%
p3a	10	15	5	14190	42	3546.40	3	67.02	19.71	13.27	79.37	83.82	0.00%
p3b	20	15	5	24237	42	10314.00	3	73.63	16.52	9.85	80.14	88.49	7.60%
p4a	10	15	10	14295	38	8427.50	7	66.24	21.05	12.71	6.48	82.61	1.10%
p4b	20	15	10	23304	37	14784.00	1	72.99	20.68	6.32	91.83	91.83	4.60%
$C_a = 2$													
p1a	10	6	3	5438	20	58.96	2	54.73	26.33	18.94	51.96	62.48	2.90%
p1b	20	6	3	7799	20	75.91	1	62.01	23.64	14.35	72.79	72.79	0.00%
p2a	10	10	5	7655	32	601.34	2	57.37	25.10	17.54	57.26	67.68	0.00%
p2b	20	10	5	11676	54	1836.10	1	62.57	24.69	12.73	77.47	77.47	0.00%
p3a	10	15	5	17187	70	10021.00	3	63.62	20.77	15.62	59.96	75.53	6.60%
p3b	20	15	5	25775	61	14071.00	1	71.30	18.70	9.99	84.99	84.99	5.20%
p4a	10	15	10	17168	49	10388.00	3	63.44	20.99	15.58	64.47	68.47	5.30%
p4b	20	15	10	29595	56	14179.00	5	70.66	13.63	15.72	58.88	73.14	26.90%

- An increase in C_a increases TC but may increase or decrease WTC. Generally, as C_a increases, we have higher variability in the system, thereby resulting in increasing the WTC. On the other hand, as an increase in C_a can be interpreted as having a more congested system, the system tries to overcome this increase by having more uniform utilization among the open facilities or installing a larger total capacity for the system, leading to a decrease in WTC.

- As the number of demand zones and the potential facility locations increase, the computational time increases, and the problem becomes harder to solve to optimality.

- We use a piecewise linear approximation to solve the G/M/1 Nominal problem. If we generate enough breaking points for this approximation, and for $C_a = 1$, the objective values would be equal or very close to the objective values obtained from the M/M/1 problem. Following is the results for some of the problems, with $t = 100$ that solved to optimality:

	M/M/1	G/M/1
p1a	4323	4322
p1b	6551	6551
p2a	6176	6176
p2b	9652	9652
p3a	13648	13648
p4a	13605	13597

Table 5.11: Computational Performance: RO-Ball, t = 100

$C_a = 0.5$

				r = 70%					r = 90%				
	f	m	n	TC	Iter.	CPU (s)	OF	Gap	TC	Iter.	CPU (s)	OF	Gap
p1a	10	6	3	4818	13	37.24	2	0.00%	4970	15	125.68	2	0.00%
p1b	20	6	3	6967	13	32.71	2	0.00%	7119	14	88.83	2	0.00%
p2a	10	10	5	6445	32	606.83	3	0.00%	6494	24	351.55	3	0.00%
p2b	20	10	5	10032	24	236.72	4	1.60%	10078	24	813.48	4	1.50%
p3a	10	15	5	13883	57	3862.90	4	0.00%	13947	54	8161.30	4	0.00%
p3b	20	15	5	22703	43	10281.00	3	1.00%	22774	41	10495.00	5	0.30%
p4a	10	15	10	13463	39	4805.80	5	0.00%	13480	32	4435.60	5	0.00%
p4b	20	15	10	22602	32	10199.00	6	0.60%	22984	30	10076.00	5	2.60%

$C_a = 2$

				r = 70%					r = 90%				
	f	m	n	TC	Iter.	CPU (s)	OF	Gap	TC	Iter.	CPU (s)	OF	Gap
p1a	10	6	3	5519	15	37.59	2	0.60%	5642	11	37.80	2	0.00%
p1b	20	6	3	7920	20	64.21	1	0.00%	8170	19	144.95	2	0.90%
p2a	10	10	5	7598	28	261.18	4	1.30%	7550	23	176.02	3	0.00%
p2b	20	10	5	11331	32	538.62	2	0.00%	11599	29	482.59	4	1.80%
p3a	10	15	5	15857	68	10141.00	3	0.00%	15928	68	10022.00	3	0.10%
p3b	20	15	5	26084	52	10543.00	3	3.60%	26396	48	10411.00	4	4.70%
p4a	10	15	10	15767	47	8193.00	6	0.90%	15790	39	10310.00	6	0.90%
p4b	20	15	10	27428	36	10881.00	5	17.80%	27023	36	12657.00	5	10.70%

Tables 5.11 and 5.12 summarize the computational results for the G/M/1 RO problem with the Ball uncertainty set using the Lagrangian-Relaxation approach, with different values of $t, f,$ and C_a. An increase in t, f, C_a, and the uncertainty budget leads to an increase in TC. In this thesis, we could only evaluate the performance of the RO model with the Ball uncertainty set, but, the model for RO problem with the Budgeted uncertainty set was derived but not numerically tested.

Table 5.12: Computational Performance: RO-Ball, t = 200

$C_a = 0.5$

				r = 70%					r = 90%				
	f	m	n	TC	Iter.	CPU (s)	OF	Gap	TC	Iter.	CPU (s)	OF	Gap
p1a	10	6	3	5251	14	55.68	2	0.00%	5405	14	70.22	2	0.00%
p1b	20	6	3	7588	17	82.04	2	0.50%	7744	14	98.10	2	0.40%
p2a	10	10	5	7269	26	317.06	3	2.10%	7276	23	234.52	4	1.50%
p2b	20	10	5	10886	25	400.38	4	1.50%	10784	30	533.28	2	0.00%
p3a	10	15	5	14975	58	4638.00	3	0.00%	15047	71	10116.00	3	0.10%
p3b	20	15	5	24514	46	10791.00	3	1.50%	24776	41	10058.00	4	2.80%
p4a	10	15	10	14880	43	7565.60	6	0.90%	14903	44	6622.20	6	0.90%
p4b	20	15	10	26382	35	10382.00	6	11.80%	26714	32	11058.00	6	16.30%

$C_a = 2$

				r = 70%					r = 90%				
	f	m	n	TC	Iter.	CPU (s)	OF	Gap	TC	Iter.	CPU (s)	OF	Gap
p1a	10	6	3	6130	17	65.76	2	0.80%	6253	16	42.52	2	0.00%
p1b	20	6	3	8913	19	88.25	2	3.70%	8778	27	132.53	1	0.00%
p2a	10	10	5	8359	31	363.96	2	0.00%	8420	34	349.32	2	0.00%
p2b	20	10	5	12500	38	1496.10	2	0.00%	12561	33	519.96	2	0.00%
p3a	10	15	5	17373	72	10270.00	2	0.10%	17462	57	5834.30	2	0.00%
p3b	20	15	5	27924	63	10238.00	3	6.90%	29110	53	10349.00	3	10.00%
p4a	10	15	10	17160	50	10412.00	2	0.00%	18112	43	10311.00	4	6.00%
p4b	20	15	10	30631	40	11067.00	5	23.30%	30965	41	11464.00	6	28.70%

Tables 5.13 and 5.14 summarize the computational results of direct solution with Gurobi for the G/M/1 DRO problem, with different values of f, t, ϵ, and C_a. According to the results, similar observations as to the two previous models could be made regarding an increase in f, t, and C_a. Besides, as ϵ increases, TC also increases, which is expected.

As for the DRO problem, we could solve a small number of instances to optimality or with a small gap. In this thesis, we use the support set as a Box, as everything is linear and easier to deal with; however, trying other support sets may improve the reformulation or even the model's performance. Moreover, without the optimality proven for the test problems, it is hard to make accurate comparisons as we did for the M/M/1 DRO problem.

Table 5.13: Computational Performance: DRO, t = 100

$C_a = 0.5$

				$\epsilon = 100$				$\epsilon = 500$			
	f	m	n	TC	CPU (s)	OF	Gap(%)	TC	CPU (s)	OF	Gap(%)
p1a	10	6	3	6911	178.82	1	0.0810	7714	10026.00	2	4.8577
p1b	20	6	3	10099	871.91	1	0.0855	11753	10002.00	2	3.9023
p2a	10	10	5	9015	10010.00	3	15.7623	10839	10035.00	3	13.8333
p2b	20	10	5	13379	10016.00	1	7.3772	17800	10005.00	4	15.9758
p3a	10	15	5	16482	10018.00	1	5.0145	24309	10005.00	5	12.8207
p3b	20	15	5	26167	10026.00	1	2.8376	37448	10009.00	1	1.1625
p4a	10	15	10	15631	10026.00	2	7.7138	24758	10004.00	6	20.7164
p4b	20	15	10	24239	36.73	1	0.0819	35445	10020.00	1	2.7818

$C_a = 2$

				$\epsilon = 100$				$\epsilon = 500$			
	f	m	n	TC	CPU (s)	OF	Gap(%)	TC	CPU (s)	OF	Gap(%)
p1a	10	6	3	7605	10010.00	1	3.2271	8704	10007.00	2	18.1216
p1b	20	6	3	11133	10025.00	1	0.6781	13289	10177.00	2	13.3278
p2a	10	10	5	10084	10006.00	2	22.8208	12454	10058.00	3	24.1376
p2b	20	10	5	14575	10016.00	1	12.7750	19547	10026.00	2	21.5405
p3a	10	15	5	17824	10041.00	1	12.3047	25278	10010.00	1	15.0671
p3b	20	15	5	28126	10012.00	1	9.7179	39831	10018.00	1	5.3354
p4a	10	15	10	17448	10047.00	2	15.0336	24688	10006.00	1	18.6072
p4b	20	15	10	26101	10033.00	1	5.4342	37750	10024.00	1	9.2348

Table 5.14: Computational Performance: DRO, t = 200

$C_a = 0.5$

				$\epsilon = 100$				$\epsilon = 500$			
	f	m	n	TC	CPU (s)	OF	Gap(%)	TC	CPU (s)	OF	Gap(%)
p1a	10	6	3	7289	113.71	1	0.0873	8318	10545.00	2	9.3262
p1b	20	6	3	10624	743.11	1	0	12568	10668.00	2	5.9834
p2a	10	10	5	9810	10031.00	3	21.0569	12131	10005.00	3	22.8334
p2b	20	10	5	14300	10061.00	2	13.0218	18632	10012.00	3	19.7392
p3a	10	15	5	17878	10057.00	2	12.1609	24419	10022.00	1	11.7942
p3b	20	15	5	27083	10004.00	1	5.6251	38541	10035.00	1	3.2342
p4a	10	15	10	15996	10036.00	1	8.5172	26379	10004.00	6	24.7127
p4b	20	15	10	25112	10062.00	1	1.6630	36505	10019.00	1	6.7209

$C_a = 2$

				$\epsilon = 100$				$\epsilon = 500$			
	f	m	n	TC	CPU (s)	OF	Gap(%)	TC	CPU (s)	OF	Gap(%)
p1a	10	6	3	8206	10005.00	1	5.3993	9679	10005.00	2	25.6414
p1b	20	6	3	12012	10002.00	1	7.1373	14599	10017.00	2	19.5934
p2a	10	10	5	11022	10005.00	2	27.9566	13968	10005.00	3	31.0607
p2b	20	10	5	15587	10019.00	1	17.6210	22445	10010.00	4	32.0077
p3a	10	15	5	18956	10023.00	1	15.9588	27410	10007.00	2	21.4801
p3b	20	15	5	29766	10042.00	1	10.5955	41821	10004.00	1	12.757
p4a	10	15	10	17722	10022.00	1	15.8480	28482	10006.00	2	28.8512
p4b	20	15	10	27664	10012.00	1	10.3544	39675	10012.00	1	12.5253

Table 5.15 shows the Lagrangian Relaxation performance for the G/M/1 Deterministic problem. Although we could get solutions directly from Gurobi for some small instances in a short time, for medium and large instances, it exhibits a poor performance within the cut-off time. In contrast, the Lagrangian Relaxation approach led to better solutions and good bounds within the same cut-off time or even in a shorter time.

Table 5.15: Lagrangian Relaxation Performance For the Deterministic Problem

	NP		NP-LR			NP		NP-LR		
	CPU(s)	Gap(%)	CPU(s)	Iter.	Gap(%)	CPU(s)	Gap(%)	CPU(s)	Iter.	Gap(%)
	$C_a = 0.5, t = 100$					$C_a = 0.5, t = 200$				
p1a	5.50	0.0005	86.40	14	0.0000	14.13	0.0000	43.47	11	0.0000
p1b	8.21	0.0019	66.64	13	0.0000	6.08	0.0050	83.81	18	2.1000
p2a	122.68	0.0000	237.60	23	0.0000	107.54	0.0000	448.46	30	0.0000
p2b	268.26	0.0000	568.45	31	0.0000	246.93	0.0000	858.70	34	0.0000
p3a	10012.00	18.2859	10451.00	52	0.2000	10014.00	18.7790	3546.40	42	0.0000
p3b	10010.00	30.7822	10060.00	37	1.8000	10010.00	32.8750	10314.00	42	7.6000
p4a	10007.00	31.8389	4053.80	31	0.0000	10009.00	31.9003	8427.50	38	1.1000
p4b	10008.00	43.5678	10636.00	30	3.8000	10009.00	48.9858	14784.00	37	4.6000
	$C_a = 2, t = 100$					$C_a = 2, t = 200$				
p1a	17.57	0.0000	56.16	13	0.5000	15.08	0.0001	58.96	20	2.9000
p1b	6.86	0.0000	209.50	21	1.9000	20.04	0.0000	75.91	20	0.0000
p2a	137.65	0.0000	504.82	27	0.0000	162.54	0.0000	601.34	32	0.0000
p2b	367.06	0.0000	992.35	37	0.0000	351.61	0.0000	1836.10	54	0.0000
p3a	10010.00	20.0170	6253.60	48	2.1000	10014.00	21.6630	10021.00	70	6.6000
p3b	10010.00	29.9094	13725.00	565	3.2000	10010.00	30.0803	14071.00	61	5.2000
p4a	10012.00	33.7112	8818.10	39	0.0000	10016.00	34.3758	10388.00	49	5.3000
p4b	10009.00	44.5217	13561.00	33	19.5000	10015.00	47.3098	14179.00	56	26.9000

5.3.1 Deterministic VS. Uncertain Demands

This section discusses the importance of considering uncertainty when designing a service system, as shown in section 5.2.1. Figure 5.6 shows the objective function values (costs) for each problem, using the deterministic model and the RO model using the Ball uncertainty set. Similar observations could be made as in section 5.2.1:

- Sensitivity to the Capacities and the Demands: Considering the expected number of customers in the system $\sum_{j\in J} \frac{R\Lambda_j^2}{\mu_j(\mu_j - \Lambda_j)} + \frac{\Lambda_j}{\mu_j}$, where $\Lambda_j = \sum_{i\in I} \xi_i y_{ij}$, we realize that if $\mu_j = \Lambda_j$, this number can go to infinity; therefore, the whole system can become unstable. Hence, this problem can be further complicated and become sensitive to the capacities installed and the demand experienced. As a result, it is crucial to include uncertainty when designing a service system.

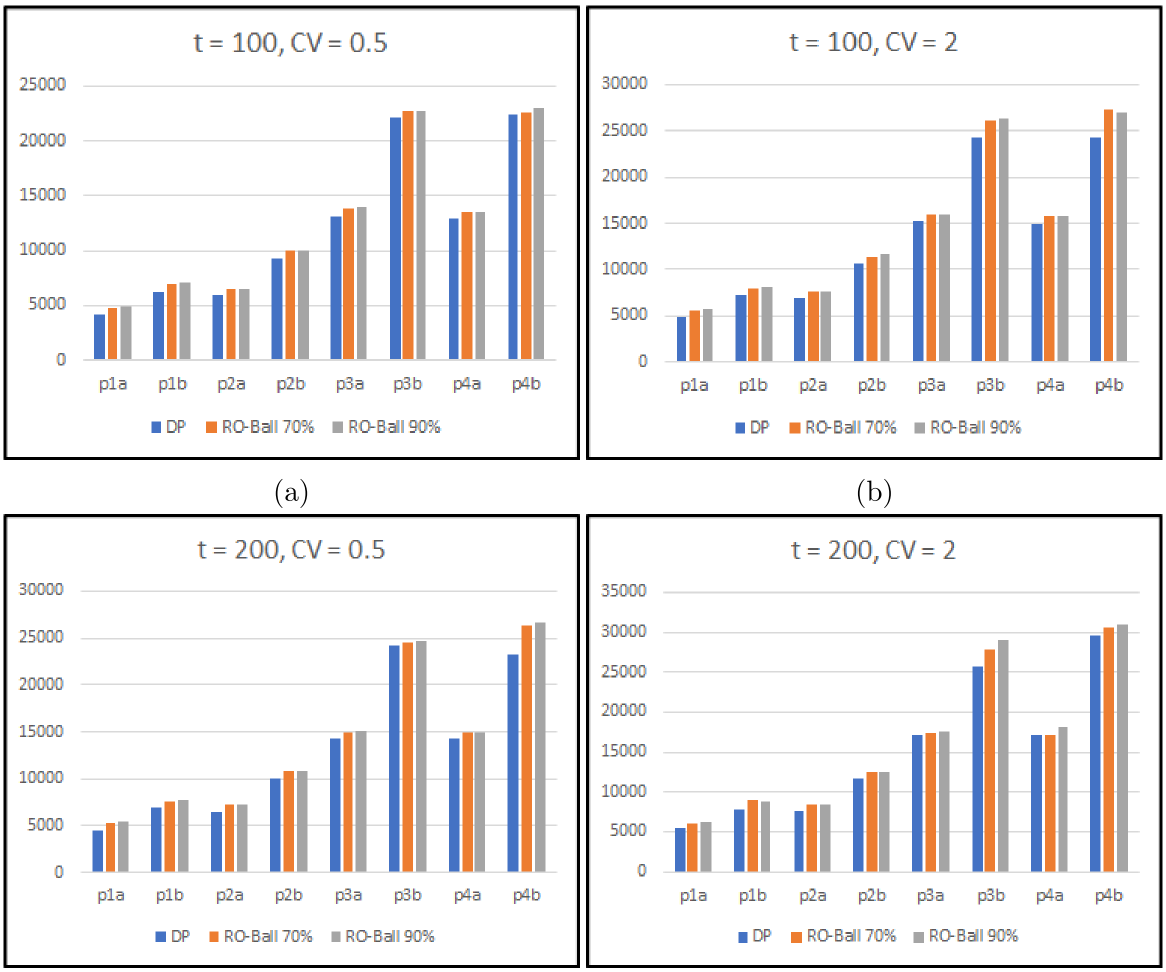
- Out-of-Sample Data: Now, let us move to the feasibility constraint $\sum_{i\in I} \xi_i y_{ij} \leq \mu_j$. Here, to have a better insight, we use test problem p4b from the Deterministic problem when $t = 200$ and $C_a = 2$. Considering the optimal solution $(\boldsymbol{\mu}^*, \mathrm{y}^*)$ for this problem, we generated 30 realizations from $U(0, 2\boldsymbol{\xi}^{nom})$ within the uncertainty set (the Box uncertainty set), and the feasibility constraint was tested using these new realizations (demands). We found out that 67% of them were not feasible for the Deterministic problem as the feasibility constraint violated. Although the other 33% of the realizations were feasible, we know that the cost with the same optimal solution would be higher as they are not the initial demands that we solve the problem to optimality with.

- The Number of Open Facilities and Customers' Assignment: Here, the same observations can be made about the M/M/1 model. Let us consider p4b, with $t = 100, r = 70\%$ and $C_a = 0.5$ as an example. The number of open facilities in the Deterministic problem is 6, with the total capacity of 1736.85, and the number of open facilities in RO-Ball is also 6 (the same facilities), but with the total capacity of 1778.80.

Again, as we use a Box as a support set in the DRO problem, we should consider the results from the RO problem when using a Box as an uncertainty set when it comes to comparison. However, since we could not solve most of the DRO problems to optimality, this comparison would be meaningless, but it is suspected that the following results will be realized:

- The costs obtained from the DRO problem should be between the ones achieved from the Deterministic and RO problems.

- As ϵ increases, the costs should also increase, and at some point when ϵ is big enough, the objective values should remain unchanged. At this moment, the unchanged costs should be equal to the costs attained from the RO problem.



(a) (b) (c) (d)

Figure 5.6: Costs of Considering The Deterministic VS. RO Model

Chapter 6

Conclusion

This thesis focused on investigating and developing novel approaches for service system design that account for the uncertainty in demand for service. In Chapter 3, under the assumption that the demands arrival can be modelled as a Poisson process, and the service process follows the exponential distribution, we modelled this problem as a network of independent M/M/1 queues. Moreover, in Chapter 4, we only changed the assumption for the demand arrival rate and modelled the problem as a network of independent G/M/1 queues. Modern methods in robust and distributionally-robust optimization were used to address some variations of both problems, applying different uncertainty (ambiguity) schemes.

For the M/M/1 network, we proposed MISOC models for the Nominal, RO, and DRO problems, which could be solved using the commercial solvers, such as Cplex or Gurobi. Testing for the Deterministic and RO problems reveals that these models can reach good results for small/medium problems in a reasonable time, but not for big size problems. However, the DRO model only shows a good performance for small problems. Moreover, due to very high sensitivity of the problem to the demand patterns, we explained the importance of designing a system that can be immune against the uncertainty in demands, although it would be more expensive.

For the G/M/1 network, we started with the MISOC reformulation of the Nominal problem, combined with a piecewise linear approximation based on the SOS2 constraints. Using this model, we proposed MISOC reformulations for the RO and DRO problems. We then proposed a Lagrangian Relaxation approach to deal with larger problems for Nominal and RO problems, in which the subproblems are also MISOC programs. We could only solve small problems for this part, and testing for the Nominal and RO problems reveals their good performance on the tested instances. However, the DRO model only showed a good performance for a very limited number of problems. We also explained that it is crucial to consider uncertainty when designing a service system, and introduced the DRO approach as an alternative to RO approaches.

Future research directions may include extending the proposed approaches for situations when each service facility has more than one server (M/M/s or G/M/s) or has a general service time (M/G/1 and M/G/s). Moreover, the capacities could be allowed to be selected from a finite number of discrete levels instead of being continuous decision variables. In this thesis, we used a Box as a support set for the DRO problem; hence, another extension could be using DRO without any support set or infinite support. Besides, the models' performance can be improved by trying other alternatives, such as using valid cuts, or using other approximation schemes. Also, meta-heuristic algorithms are another approach to solve large-scale problems.

Bibliography

[1] The Private Cost of Public Queues for Medically Necessary Care, 2020. https://www.fraserinstitute.org/studies/private-cost-of-public-queues-for-medically-necessary-care-2020.

[2] The World Bank. https://data.worldbank.org/indicator/NV.SRV.TOTL.ZS.

[3] Robert Aboolian, Oded Berman, and Dmitry Krass. Profit maximizing distributed service system design with congestion and elastic demand. *Transportation Science*, 46(2):247–261, 2012.

[4] Ivo Adan, Onno Boxma, and David Perry. The G/M/1 queue revisited. *Mathematical Methods of Operations Research*, 62(3):437–452, 2005.

[5] Ali Amiri. Solution procedures for the service system design problem. *Computers & Operations Research*, 24(1):49–60, 1997.

[6] Ali Amiri. The design of service systems with queueing time cost, workload capacities and backup service. *European Journal of Operational Research*, 104(1):201 – 217, 1998.

[7] Ali Amiri. The multi-hour service system design problem. *European Journal of Operational Research*, 128(3):625 – 638, 2001.

[8] Barua Bacchus and Moir Mackenzie. Waiting your turn: Wait times for health care in canada, 2019 report. Fraser Institute, 2019.

[9] Opher Baron, Oded Berman, and Dmitry Krass. Facility location with stochastic demand and constraints on waiting time. *Manufacturing & Service Operations Management*, 10(3):484–505, 2008.

[10] EML Beale and John JH Forrest. Global optimization using special ordered sets. *Mathematical Programming*, 10(1):52–69, 1976.

[11] Aharon Ben-Tal, Dick Den Hertog, and Jean-Philippe Vial. Deriving robust counterparts of nonlinear uncertain inequalities. *Mathematical Programming*, 149(1-2):265–299, 2015.

[12] Aharon Ben-Tal and Arkadi Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, 23(4):769–805, 1998.

[13] Aharon Ben-Tal and Arkadi Nemirovski. Robust solutions of uncertain linear programs. *Operations Research Letters*, 25(1):1–13, 1999.

[14] Aharon Ben-Tal and Arkadi Nemirovski. Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical Programming*, 88(3):411–424, 2000.

[15] Oded Berman and Zvi Drezner. Location of congested capacitated facilities with distance-sensitive demand. *IIE Transactions*, 38(3):213–221, 2006.

[16] Oded Berman and Edward H Kaplan. Facility location and capacity planning with delay-dependent demand. *International Journal of Production Research*, 25:1773–80, 1987.

[17] Oded Berman and Dmitry Krass. 11 facility location problems with stochastic demands and congestion. *Facility Location: Applications and Theory*, 329, 2001.

[18] Oded Berman and Dmitry Krass. Stochastic location models with congestion. In *Location Science*, pages 477–535. Springer, 2019.

[19] Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Robust sample average approximation. *Mathematical Programming*, 171(1-2):217–282, 2018.

[20] Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004.

[21] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[22] John A Buzacott and J George Shanthikumar. *Stochastic models of manufacturing systems*, volume 4. Prentice Hall Englewood Cliffs, NJ, 1993.

[23] Ignacio Castillo, Armann Ingolfsson, and Thaddeus Sim. Social optimal location of facilities with fixed servers, stochastic demand, and congestion. *Production and Operations Management*, 18(6):721–736, 2009.

[24] Abraham Charnes and William W Cooper. Chance-constrained programming. *Management Science*, 6(1):73–79, 1959.

[25] Yunn-Kuang Chu and Jau-Chuan Ke. Interval estimation of mean response time for a G/M/1 queueing system: empirical Laplace function approach. *Mathematical Methods in the Applied Sciences*, 30(6):707–715, 2007.

[26] Daniel P Connors, Gerald E Feigin, and David D Yao. A queueing network model for semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, 9(3):412–427, 1996.

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

[28] Bharat T Doshi. A note on stochastic decomposition in a GI/G/1 queue with vacations or set-up times. *Journal of Applied Probability*, 22(2):419–428, 1985.

[29] Laurent El Ghaoui and Hervé Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 18(4):1035–1064, 1997.

[30] Laurent El Ghaoui, Francois Oustry, and Hervé Lebret. Robust solutions to uncertain semidefinite programs. *SIAM Journal on Optimization*, 9(1):33–52, 1998.

[31] Samir Elhedhli. Exact solution of a class of nonlinear knapsack problems. *Operations Research Letters*, 33(6):615–624, 2005.

[32] Samir Elhedhli. Service system design with immobile servers, stochastic demand, and congestion. *Manufacturing & Service Operations Management*, 8(1):92–97, 2006.

[33] Samir Elhedhli, Yan Wang, and Ahmed Saif. Service system design with immobile servers, stochastic demand and concave-cost capacity selection. *Computers & Operations Research*, 94:65–75, 2018.

[34] Emre Erdoğan and Garud Iyengar. Ambiguous chance constrained problems and robust optimization. *Mathematical Programming*, 107(1-2):37–61, 2006.

[35] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

[36] Joel Goh and Melvyn Sim. Distributionally robust optimization and its tractable approximations. *Operations Research*, 58(4-part-1):902–917, 2010.

[37] Seifollah Louis Hakimi. Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations Research*, 12(3):450–459, 1964.

[38] Seifollah Louis Hakimi. Optimum distribution of switching centers in a communication network and some related graph theoretic problems. *Operations Research*, 13(3):462–475, 1965.

[39] Kaj Holmberg, Mikael Rönnqvist, and Di Yuan. An exact algorithm for the capacitated facility location problems with single sourcing. *European Journal of Operational Research*, 113(3):544–559, 1999.

[40] Zhaolin Hu and L Jeff Hong. Kullback-Leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 2013.

[41] Leonid Vasilevich Kantorovich and Gennady S Rubinstein. On a space of completely additive functions. *Vestnik Leningrad. Univ*, 13(7):52–59, 1958.

[42] JFC Kingman. The single server queue in heavy traffic. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 57, pages 902–904. Cambridge University Press, 1961.

[43] Juan Ma, Ying Tat Leung, and Manjunath Kamath. Service system design under information uncertainty: Insights from an M/G/1 model. *Service Science*, 11(1):40–56, 2019.

[44] Vladimir Marianov, Miguel Rios, and Francisco Javier Barros. Allocating servers to facilities, when demand is elastic to travel and waiting times. *RAIRO-Operations Research*, 39(3):143–162, 2005.

[45] Vladimir Marianov, Miguel Ríos, and Manuel José Icaza. Facility location for market capture when users rank facilities by shorter travel and waiting times. *European Journal of Operational Research*, 191(1):32–44, 2008.

[46] Vladimir Marianov and Daniel Serra. Probabilistic, maximal covering location—allocation models forcongested systems. *Journal of Regional Science*, 38(3):401–424, 1998.

[47] Vladimir Marianov and Daniel Serra. Hierarchical location–allocation models for congested systems. *European Journal of Operational Research*, 135(1):195–208, 2001.

[48] Georg Pflug and David Wozabal. Ambiguity in portfolio selection. *Quantitative Finance*, 7(4):435–442, 2007.

[49] Francisco Silva and Daniel Serra. Locating emergency services with different priorities: the priority queuing covering location problem. *Journal of the Operational Research Society*, 59(9):1229–1238, 2008.

[50] James E Smith and Robert L Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006.

[51] Allen L Soyster. Convex programming with set-inclusive constraints and applications to inexact linear programming. *Operations Research*, 21(5):1154–1157, 1973.

[52] Mark L Spearman and Wallace J Hopp. Factory physics: Foundations of manufacturing management. *Irwin, Chicago, IL*, 439, 1996.

[53] Siddhartha S Syam. A multiple server location–allocation model for service system design. *Computers & Operations Research*, 35(7):2248–2265, 2008.

[54] Navneet Vidyarthi and Sachin Jayaswal. Efficient solution of a class of location–allocation problems with stochastic demand and congestion. *Computers & Operations Research*, 48:20–30, 2014.

[55] Qian Wang, Rajan Batta, and Christopher M Rump. Algorithms for a facility location problem with stochastic customer demand and immobile servers. *Annals of Operations Research*, 111(1-4):17–34, 2002.

[56] Qian Wang, Rajan Batta, and Christopher M Rump. Facility location models for immobile servers with stochastic demand. *Naval Research Logistics (NRL)*, 51(1):137–152, 2004.

[57] Ward Whitt. The queueing network analyzer. *The Bell System Technical Journal*, 62(9):2779–2815, 1983.

[58] Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

[59] Yue Zhang, Oded Berman, Patrice Marcotte, and Vedat Verter. A bilevel model for preventive healthcare facility network design with congestion. *IIE Transactions*, 42(12):865–880, 2010.

[60] Yue Zhang, Oded Berman, and Vedat Verter. Incorporating congestion in preventive healthcare facility network design. *European Journal of Operational Research*, 198(3):922–935, 2009.