NORTHWESTERN UNIVERSITY

Decomposing Optimization Problems Under Stochastic Disruptions

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Industrial Engineering and Management Sciences

by

Haoxiang Yang

EVANSTON, ILLINOIS

September 2019

Abstract

This thesis consists of three projects, centered around the aim to better model real-world systems under uncertainty, specifically, under stochastic disruptions, using optimization. A stochastic disruption is a type of infrequent event in which the timing and the magnitude are random. We introduce the concept of stochastic disruptions and a stochastic optimization model is proposed for such problems with a finite time horizon.

We further develop the idea of a stochastic disruption for a specific example, a project crashing problem under a single disruption. When a disruption occurs, the duration of an activity, which has not yet started, can change. Both the magnitude of the change of an activity's duration and the timing of the disruption can be random. We formulate a stochastic mixed integer program (SMIP) with mixed integer recourse. This SMIP can be computationally challenging to solve using existing techniques. We propose an adaptive branch-and-cut algorithm to solve the SMIP and evaluate the computational performance of our approach.

Next, we consider an application in electric power systems in which a disruption can occur due to uncertain demand or uncertain availability of renewable energy resources. We propose a robust optimization model for the alternating current optimal power flow (ACOPF) problem, considering a two-stage model in which potential disruptions occur on a 10-15 minute timescale. We use an uncertainty set to model a disruption in the context of robust optimization. Based on a recently developed convex relaxation for the ACOPF problem, we construct a robust convex optimization problem with recourse. We develop an enhanced cutting-plane algorithm to solve this problem, and we establish convergence and other desirable properties. Experimental results indicate that our robust convex relaxation of the ACOPF problem can provide a tight lower bound and an

acceptable solution for the non-convex robust ACOPF problem.

Finally, we consider a syringe exchange program (SEP) in which a client's behavior is stochastic. Using data from one program in Chicago over ten years, we study the behavior of its clients, focusing on the temporal process governing their visits to service locations and their demographics. The frequency of using the SEP services may be affected by stochastic disruptions such as the client relocating or participating in a treatment program. We construct a phase-type distribution to characterize unobservable changes in a client's status, and we use an affine relationship between model parameters and features of an individual client. The phase-type distribution governs inter-arrival times between reoccurring visits of each client to SEP sites and is informed by characteristics of a client including age, gender, ethnicity, drug-use habits and more. The inter-arrival time model is a sub-model in a simulation model that we construct for the larger system, which allows us to provide a personalized prediction regarding the client's time-to-return to a service location so that better intervention decisions can be made.

Acknowledgments

First and foremost, I would like to thank my advisor, Prof. David Morton for all the effort he put in me to make me not only a better researcher but also a better person. Prof. Morton has been so patient and always encourage me to aim for better results. At the same time, he has taught me to keep my head down, be detail-oriented and rigorous about everything in my life. He is truly a great advisor, and no words can truly express my gratitude for what he has done for me.

Besides Prof. Morton, all faculty and staff members have been very supportive to me during the past five years. I would like to especially thank Prof. Andreas Wächter and Johnathan Gaetz for all the technical support for my computational studies. I do not think this dissertation could take shape without their tireless instructions.

I want to thank my collaborators, Dr. Krishnamurthy Dvijotham, Dr. Alexander Gutfraind, and Dr. Kenneth Kuhn for guiding me through the project and offering me genuine advice about how to move forward with research ideas and career. I really appreciate all the support from the undergraduate students who worked with me during my Ph.D., Yue Hu and Thomas Massion. I want to thank my committee members, Prof. Sanjay Mehrotra, Prof. Chaithanya Bandi, and Prof. Daniel Bienstock. Your advice is precious for my research. I also would like to thank everyone who has helped me with my research. Your kind suggestions and discussions have sparked many of my ideas.

In the past five years, I have also received a tremendous amount of support from my fellow Ph.D. students at Northwestern University. I have had the blessing to stay in the best office in the world for four years, and I am going to miss all the fun we had during this journey. Thank you, Collin Erickson, Mark Semelhago, and Andrea Treviño-Gavito!

My parents have been supporting me unconditionally throughout my years studying abroad, both mentally and fiscally. They always encourage me to chase my dream and never doubt for a second that I am going to succeed. I will always love you both and support you like what you have done for me.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Stochastic optimization and robust optimization are two widely used methods to incorporate uncertainty in a decision-making process. In such models, the uncertainty takes the form of an unknown outcome of parameters, which is realized after making the decision. In stochastic programming, we model uncertainty by assuming the outcome is governed by a known distribution, or stochastic process, and in robust optimization we instead assume the outcome belongs to an uncertainty set.

If we have a multi-stage problem, which involves multiple rounds of decision making under uncertainty, neither the stochastic program nor the robust optimization problem is typically easy to solve. Under certain assumptions, a multi-stage stochastic linear program can be solved by stochastic dual dynamic programming, when the number of scenarios in each stage is modest and the number of stages is not too large. However, for some applications such as air traffic control and energy dispatch, each decision period is short, resulting in a large number of stages. There is limited literature that extends robust optimization models to a multi-stage setting without simplifying the recourse actions.

At the same time, significant stochastic disruptions can occur during multi-stage decision problems, and there is less work directly addressing this type of uncertainty. Model parameters

may be assumed to be deterministic, in part, because of the increased accuracy of predictive models during normal operations. For example, the deviation of energy consumption from a point forecast, within a five-minute window is usually negligible, and so we might treat it as a deterministic value unless we are under an extreme situation. This suggests that we may focus on these infrequent extreme situations and pursue a simplified model of uncertainty.

1.2 Stochastic Disruptions

We consider stochastic disruptions, which include the timing of the disruption in the model of uncertainty. A disruption is a type of event under which uncertain parameters change in a significant way, in value, distribution, etc. We view disruptions as events that occur infrequently, but that can have a significant impact, which means that a disruption could be a one-time event, or an event that happens only a limited number of times over the planning horizon. Using the concept of stochastic disruption, we can model not only the magnitude of the uncertainty, but also the timing, which is not typically handled by stochastic programming or robust optimization.

To illustrate with the simplest case, we assume there is at most one disruption. A sequential decision problem with a single stochastic disruption is illustrated in Figure 1.1.

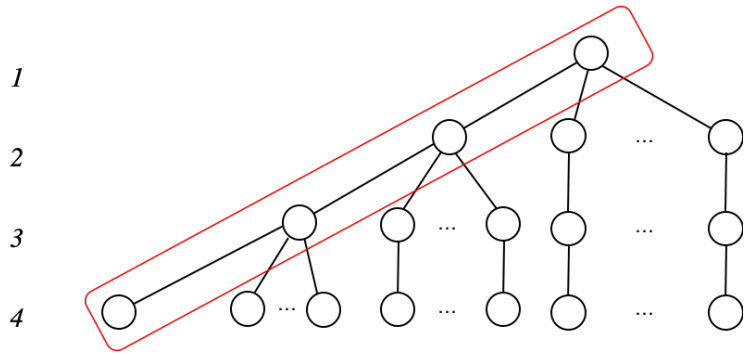


Figure 1.1: An illustration of scenario tree for a sequential decision problem with single disruption

Figure 1.1 shows a sequential decision problem with four time periods. The left diagonal (circled by a red rectangle) represents the nominal scenario in which no disruption occurs. Every branch to the right of the nominal scenario represents a disruption scenario. The depth of branch represents the timing of the disruption and the parallel branches with the same depth represent

different disruption magnitudes. We are interested in finding an optimal policy for the nominal scenario, since we follow this policy until a disruption occurs. Once a disruption occurs, the system illustrated in the figure becomes deterministic and we can re-optimize according to the disruption scenario realized.

1.3 Dissertation Outline

In Chapter 2, we formulate a program evaluation and review technique (PERT) model in which we crash an activity network under stochastic disruptions. Activities within a project can be accelerated with certain costs, and the decision maker has to decide which activities to crash under a limited budget so that the entire project span is minimized in expectation. At some random point along the continuous time horizon, the duration of activities might change due to some external reasons and the crashing strategy must change accordingly. We formulate the model as a two-stage stochastic mixed-integer program (SMIP) in which the timing of the disruption is random. Simple examples give insight into potential solution behavior and justify the use of an SMIP model. Although we show the model to be NP-hard, we propose an effective decomposition algorithm that adaptively partitions the feasible region of continuous first-stage decision variables within a branch-and-cut algorithm. We present computational results to show the value of our model and the effectiveness of our decomposition method compared to solving the extensive formulation with a commercial solver.

In Chapter 3 we discuss solving a robust convex relaxation of an alternating current optimal power flow (ACOPF) problem. In this problem, a disruption can be considered as the time point at which realized net load deviates significantly from a point forecast. Since it is important to prevent an electric power system from failing to satisfy demand, we use a robust optimization model to handle all contingencies in an uncertainty set. We propose a decomposition algorithm, based on generalized Benders' decomposition, to solve a convex relaxation of the robust optimization model. We prove convergence of our algorithm and develop a scenario-appending scheme to improve the computational performance of the cutting-plane algorithm. We show that the lower bound obtained by our model can be tight and the quality of the solution empirically performs well in the non-convex

robust model.

In Chapter 4 we study the stochastic behavior of the clients from a syringe exchange program (SEP) in Chicago. Inspired by the contrast between trends in the use of the SEP and trends in national drug use, we analyze client arrival data between 2005 and 2014, and we formulate and fit a stochastic process for the entire syringe exchange experience of a client. We develop three sub-models of the stochastic process: initiation, reoccurring visits, and termination. We build an optimization model to fit the parameters for the stochastic process and to uncover the relationship between the model parameters and a client's features. We present goodness-of-fit test results to justify our model selection. Using the fitted model, we simulate an active intervention strategy and evaluate its effectiveness.

Finally, in Chapter 5 we summarize the contributions in this dissertation, and we discuss future research directions and potential applications of optimization models that incorporate stochastic disruptions.

Chapter 2

Optimal Crashing of an Activity Network with Disruptions

2.1 Introduction

The management of complex projects through optimization has a rich history in operations research, beginning with the critical path method of Kelly (1961); see Söderlund (2004) for an overview. A project is a collection of activities, between which there are precedence relationships due to logical or technological considerations. A precedence relationship is usually reflected as the start time of one activity following the completion of another. Typically, multiple activities can be processed at the same time, and there is no limit on how many activities can be processed simultaneously, as long as the precedence requirements are satisfied. See Elmaghraby (1977) for a detailed treatment of activity networks. In this setting, "crashing" is an action that consumes a certain amount of one or more resources and shortens the duration of an activity accordingly (Kuhl and Tolentino-Peña 2008). A deterministic optimization model for crashing an activity network was proposed in the 1960s (Fulkerson 1961, Kelly 1961), in which the goal is to complete the project in minimum time by allocating resources under one or more budget constraints.

When the program evaluation and review technique (PERT) was introduced (Malcolm et al. 1959), activity durations were modeled as independent beta random variables, and the project du-

ration approximated by a normal distribution. Extensions that allow for more general assumptions followed (Elmaghraby 1977), and Monte Carlo simulation plays a role in estimating the expected project span, which is difficult to express analytically (Burt and Garman 1971, van Slyke 1963). Heuristics and simulation-based algorithms have been developed to approximately solve the stochastic project crashing problem (Aghaie and Mokhtari 2009, Bowman 1994, Ke 2014, Kim et al. 2007). Another approach to handle uncertainty in activity duration is robust optimization, in which the objective is to minimize the worst case project span over a specified uncertainty set. While affinely adaptive recourse decisions are computationally tractable as linear, or second-order cone, programs, this restriction may lead to suboptimal solutions (Chen et al. 2008, Cohen et al. 2007). However, once recourse decisions can take general form, the robust model is only tractable with rectangular uncertainty sets (Wiesemann et al. 2012). Ahipasaoglu et al. (2016) propose a distributionally robust optimization scheme applied to a PERT network, which reformulates the problem as a semidefinite program or a copositive program, depending on the description of uncertainty. The project crashing optimization problem finds application in project management (Demeulemeester and Herroelen 2006, Jaselskis and Ashley 1991, Tonchia 2018), machine scheduling (Blazewicz et al. 1983, Hall and Sriskandarajah 1996), health services scheduling (Cardoen et al. 2010), chemical processes (Li and Ierapetritou 2008), and digital circuit sizing (Kim et al. 2007).

In this chapter, we propose using the concept of stochastic disruptions to model uncertainty in the duration of activities, which differs from existing approaches in both stochastic programming and robust optimization. A stochastic disruption is an event that may occur at any point in the problem's time horizon and results in a change—typically a significant change—in the system's parameters. A few authors apply this idea in models with discrete time periods, in which the disruption can only occur in a set of specified time periods. Yu and Qi (2004) introduce scenario-based optimization models for airline scheduling. Salmeron et al. (2009) introduce a sealift scheduling problem under a finite number of stochastic disruptions within a stochastic programming structure; this model structure "falls between standard two-stage and multi-stage stochastic programs for a multi-period problem" and reduces the size of the problem (scenario tree) to quadratic, rather than exponential, growth in the number of time periods. Our setting inherits the philosophy of Salmeron

et al. (2009), but enhances the model by allowing the random disruption time to be continuous in the context of an activity network, instead of a prespecified set of fixed time periods.

Given a limited number of disruption scenarios, the problem of optimizing crashing decisions to minimize expected completion time can be formulated as a stochastic mixed-integer program, and we present the model in Section 2.2. If we assume a continuous distribution for the disruption time and magnitude, a sample average approximation (SAA) can be used to create a finite set of scenarios and approximate the original problem by a finite-sized optimization problem. In Section 2.3 we show that the problem is NP-hard even with continuous allocation of crashing effort and just two scenarios. Section 2.4 presents properties of the problem using a serial activity network as a special case. The potentially large scale and the discrete, non-convex nature of the SAA problem's formulation suggest that it may be computationally challenging to solve. In Section 2.5, a method based on Benders' decomposition is developed to solve our problem of optimizing crashing decisions under stochastic disruptions. We show such a decomposition method can solve the integer program in a finite number of iterations. Experiment results are presented in Section 2.6, including the empirical relationship between solution quality and sample size, the comparison between the quality of our solution and solutions of alternative models, and the computational performance of the decomposition method of Section 2.5. We conclude with remarks on potential extensions of our model in Section 2.7.

2.2 Problem Formulation

Nomenclature:

Indices and index sets

I	the set of activities;
J_i	the set of crashing options for activity $i \in I$;
Ω	the index set for disruption scenarios (sample space);
$\mathcal{A}$	set of arcs, which represents precedence relationships;

Parameters

D_{ik} nominal duration between possible start times of activities i and k, $(i,k) \in \mathcal{A}$;

e_{ij} effectiveness of crashing option $j \in J_i$ for activity $i \in I$;

B total budget for crashing options;

b_{ij} cost of crashing option $j \in J_i$ for activity $i \in I$;

H^{ω} disruption time under scenario $\omega \in \Omega$;

d_{ik}^{ω} increase in duration of $(i,k) \in \mathcal{A}$ under $\omega \in \Omega$, if started after the disruption;

p^{ω} the probability of scenario $\omega \in \Omega$;

p^{0} the probability of no disruption;

Decision variables

t_i nominal start time of activity $i \in I$;

x_{ij} crashing of activity $i \in I$ by option $j \in J_i$ in the nominal plan;

t_i^{ω} start time of activity $i \in I$ under scenario $\omega \in \Omega$;

x_{ij}^{ω} crashing of activity $i \in I$ by option $j \in J_i$ under scenario $\omega \in \Omega$;

G_i^{ω} binary indicator whether activity $i \in I$ starts after disruption under $\omega \in \Omega$;

z_{ij}^{ω} binary term to linearize bilinear term, $G_i^{\omega} x_{ij}^{\omega}$, $i \in I, j \in J_i, \omega \in \Omega$.

We first review an optimization model for a deterministic crashing problem; see Fulkerson (1961), Kelly (1961). A set of activities, I, together with precedence relationships, $\mathcal{A} \subseteq I \times I$, form an acyclic activity network $\mathcal{G} = (I, \mathcal{A})$, which represents the project. An arc $(i,k) \in \mathcal{A}$ indicates that activity i has to finish before activity k starts, and its length, D_{ik}, shows that the start time of activity i has to be at least $D_{ik} \geqslant 0$ before the start time of activity k. We create two dummy activities $S, T \in I$ to represent the start and the termination of the entire project. Activity S should precede every activity $i \in I \backslash \{S\}$ and T should succeed every activity $i \in I \backslash \{T\}$, either directly or by implication, and they both have zero duration.

We can apply a finite set of crashing options, $j \in J_i$, to activity $i \in I$. One unit application of each option incurs a cost of b_{ij}, and it decreases the corresponding durations by $D_{ik} e_{ij}$, $\forall (i,k) \in \mathcal{A}$, where $e_{ij} \in [0,1]$ denotes the unit effectiveness of crashing option j. For example, suppose the

duration between the start time of activity 1 and 2 is $D_{12} = 10$, and applying one unit of crashing option 1 to activity 1 decreases the duration by half; i.e., $e_{11} = 0.5$. If we apply 0.4 unit of crashing option 1 to activity 1, $x_{11} = 0.4$, the required separation between activity 1 and 2 becomes $10(1 - 0.4 \cdot 0.5) = 8$. The total cost of crashing cannot exceed a given budget, B. The objective is then to minimize the start time of activity T, and thus, we formulate the deterministic project crashing problem as:

$$\min \quad t_T \tag{2.1a}$$

$$\text{s.t.} \quad t_k - t_i \geqslant D_{ik}\left(1 - \sum_{j \in J_i} e_{ij} x_{ij}\right) \qquad \forall\, (i,k) \in \mathcal{A} \tag{2.1b}$$

$$\sum_{i \in I} \sum_{j \in J_i} b_{ij} x_{ij} \leqslant B \tag{2.1c}$$

$$\sum_{j \in J_i} x_{ij} \leqslant 1 \qquad \forall\, i \in I \tag{2.1d}$$

$$0 \leqslant x_{ij} \leqslant 1 \qquad \forall\, i \in I, j \in J_i \tag{2.1e}$$

$$t_i \geqslant 0 \qquad \forall\, i \in I. \tag{2.1f}$$

In this formulation, t_i represents the start time of activity $i \in I$. We aim to minimize the project span, which is the start time of the terminal activity, t_T. Constraint (2.1b) guarantees the precedence relationship: if activity i precedes activity k, activity k cannot start until time $t_i + D_{ik}(1 - \sum_{j \in J_i} e_{ij} x_{ij})$. Constraint (2.1c) is the budget constraint and constraint (2.1d) ensures that no more than one unit of crashing option can be applied to an activity. Constraint (2.1f) enforces nonnegativity for the start time of all activities.

For a project crashing problem under stochastic disruptions, we assume at most one stochastic disruption can occur at a random time in the project span. While this assumption may be limiting in some settings, it is appropriate when it is unlikely for two or more disruptions to occur during the time horizon, and can apply, e.g., for natural disasters, major market crashes, cyber attacks, and work stoppages. For example, suppose we manage a construction project, and we aim to plan against the potential hazard caused by an earthquake or an employee strike. It may be unlikely

for two major earthquakes or strikes to affect the same project within the relevant time period. We further assume that for each activity $i \in I$, the crashing decision needs to be made prior to the start of that activity, which is reasonable, e.g., when contracts are involved in commitment of resources (Oberlender 1993). We assume a disruption does not affect activities that have already started (including those already finished) at the time of the disruption, but the disruption changes the length of activities that have not yet started according to a known probability distribution. It is usually hard to compute the recourse function directly when random parameters have a continuous distribution, and therefore we use sample average approximation (SAA) (Kim et al. 2015, Shapiro et al. 2009). In this chapter, we assume there is a finite set of scenarios indexed by $\omega \in \Omega$. For each scenario ω, the random realization of parameters, which we denote ξ^ω, consists of the timing of the disruption, H^ω, and the magnitude of the disruption via increases in the duration parameters, $d_{ik}^\omega, \forall (i,k) \in \mathcal{A}$.

Because we assume at most one disruption, we can model the problem as a two-stage stochastic mixed-integer program, in which the timing of the second stage is random. That is, the definition of our stages differs from the usual stochastic programming setting. Here the first stage contains decisions through completion of the project, and we follow this policy if no disruption occurs. And, the second stage characterizes the decisions for each realization of the disruption, which commence at the random time, H^ω. The first stage decision variables are carried out until the disruption if it ever occurs, and after the disruption, the scenario-specific recourse decisions are executed.

The extensive formulation of the two-stage stochastic program is shown as formulation (2.2):

$$z^* = \min \quad p^0 t_T + \sum_{\omega \in \Omega} p^\omega t_T^\omega \tag{2.2a}$$

$$\text{s.t.} \quad t_k - t_i \geqslant D_{ik} \left(1 - \sum_{j \in J_i} e_{ij} x_{ij}\right) \qquad \forall\, (i,k) \in \mathcal{A} \tag{2.2b}$$

$$\sum_{i \in I} \sum_{j \in J_i} b_{ij} x_{ij} \leqslant B \tag{2.2c}$$

$$\sum_{j \in J_i} x_{ij} \leqslant 1 \qquad \forall\, i \in I \tag{2.2d}$$

$$H^\omega + G_i^\omega M \geqslant t_i \qquad \forall\, i \in I, \omega \in \Omega \tag{2.2e}$$

$$H^{\omega} - (1 - G_i^{\omega})M \leqslant t_i \qquad \forall\, i \in I, \omega \in \Omega \tag{2.2f}$$

$$t_i^{\omega} + G_i^{\omega} M_t \geqslant t_i \qquad \forall\, i \in I, \omega \in \Omega \tag{2.2g}$$

$$t_i^{\omega} - G_i^{\omega} M_t \leqslant t_i \qquad \forall\, i \in I, \omega \in \Omega \tag{2.2h}$$

$$x_{ij}^{\omega} + G_i^{\omega} \geqslant x_{ij} \qquad \forall\, i \in I, j \in J_i, \omega \in \Omega \tag{2.2i}$$

$$x_{ij}^{\omega} - G_i^{\omega} \leqslant x_{ij} \qquad \forall\, i \in I, j \in J_i, \omega \in \Omega \tag{2.2j}$$

$$t_k^{\omega} - t_i^{\omega} \geqslant D_{ik} + d_{ik}^{\omega} G_i^{\omega} - \sum_{j \in J_i} D_{ik} e_{ij} x_{ij}^{\omega} - \sum_{j \in J_i} d_{ik}^{\omega} e_{ij} z_{ij}^{\omega} \qquad \forall\, (i,k) \in \mathcal{A}, \omega \in \Omega \tag{2.2k}$$

$$\sum_{i \in I} \sum_{j \in J_i} b_{ij} x_{ij}^{\omega} \leqslant B \qquad \forall\, \omega \in \Omega \tag{2.2l}$$

$$\sum_{j \in J_i} x_{ij}^{\omega} \leqslant 1 \qquad \forall\, i \in I, \omega \in \Omega \tag{2.2m}$$

$$z_{ij}^{\omega} \leqslant G_i^{\omega} \qquad \forall\, i \in I, j \in J_i, \omega \in \Omega \tag{2.2n}$$

$$z_{ij}^{\omega} \leqslant x_{ij}^{\omega} \qquad \forall\, i \in I, j \in J_i, \omega \in \Omega \tag{2.2o}$$

$$z_{ij}^{\omega} \geqslant G_i^{\omega} + x_{ij}^{\omega} - 1 \qquad \forall\, i \in I, j \in J_i, \omega \in \Omega \tag{2.2p}$$

$$t_i \geqslant 0 \qquad \forall\, i \in I \tag{2.2q}$$

$$t_i^{\omega} \geqslant H^{\omega} G_i^{\omega} \qquad \forall\, i \in I, \omega \in \Omega \tag{2.2r}$$

$$0 \leqslant x_{ij} \leqslant 1 \qquad \forall\, i \in I, j \in J_i \tag{2.2s}$$

$$0 \leqslant x_{ij}^{\omega} \leqslant 1 \qquad \forall\, i \in I, j \in J_i, \omega \in \Omega \tag{2.2t}$$

$$0 \leqslant z_{ij}^{\omega} \leqslant 1 \qquad \forall\, i \in I, j \in J_i, \omega \in \Omega \tag{2.2u}$$

$$G_i^{\omega} \in \{0, 1\}. \qquad \forall\, i \in I, \omega \in \Omega. \tag{2.2v}$$

In model (2.2), we minimize the expected project span, weighing the span under each scenario by its probability. We replicate constraints (2.1b)-(2.1d) for the nominal scenario as (2.2b)-(2.2d). In constraints (2.2e)-(2.2f), variable G_i^{ω} takes value 1 if activity i starts after the disruption time; otherwise it takes value 0, and M is a large number to enforce the logic of this relationship. This is important in our problem setting because the duration of each activity depends on its temporal relationship to the disruption time, which is reflected in constraint (2.2k). Also, we must ensure

that decisions made before the disruption time in each scenario match the nominal decisions, and constraints (2.2g)-(2.2j) capture these non-anticipativity conditions. For each scenario, the duration between activity i and k becomes $(D_{ik} + d^{\omega}_{ik}G^{\omega}_i)(1 - \sum_{j\in J_i} e_{ij}x^{\omega}_{ij})$, which expands to the form of constraint (2.2k). If $G^{\omega}_i = 0$, which means activity i starts before the disruption time of scenario ω, this expression is the same as $D_{ik}(1-\sum_{j\in J_i} e_{ij}x_{ij})$ because $x_{ij} = x^{\omega}_{ij}$ is enforced by constraints (2.2i) and (2.2j). If $G^{\omega}_i = 1$, then the duration between activity i and k is changed to $D_{ik} + d^{\omega}_{ik} \geqslant 0$. We allow a negative "increase" in duration d^{ω}_{ik}, but require the overall duration to be nonnegative. The expression $(D_{ik} + d^{\omega}_{ik}G^{\omega}_i)(1 - \sum_{j\in J_i} e_{ij}x^{\omega}_{ij})$ contains a bilinear term $G^{\omega}_i x^{\omega}_{ij}$, which we linearize using binary variable z^{ω}_{ij} and constraints (2.2n)-(2.2p).

2.3 NP-Hardness

We show that the optimal project crashing problem under a stochastic disruption is NP-hard even with a single disruption scenario, which occurs with probability one at time zero. Our proof relies on a transformation from the exactly-one-in-three `3SAT` (`EOIT_3SAT`) problem: Let $U = \{u_1, u_2, \ldots, u_n\}$ be a set of variables. A literal can be either u or $\bar{u} = \neg u$ for $u \in U$. Let $C = \{c_1, c_2, \ldots, c_m\}$ be a set of clauses, each of which is formed by a disjunction of three literals, e.g., $c_i = u_j \vee u_k \vee \bar{u}_\ell$. The `EOIT_3SAT` problem asks whether there is a truth assignment for each $u \in U$ such that each clause in C has exactly one true literal. De et al. (1997) use `EOIT_3SAT` to prove that an activity network problem, in which there are a finite set of alternatives for each activity with different duration and cost, is NP-hard, and we use similar proof constructs.

Starting with an instance of `EOIT_3SAT`, we formulate an activity network using three layers of nodes. The first layer contains $3n$ nodes and represents the truth assignment of each variable in `EOIT_3SAT`. The second layer contains $3m$ nodes and represents the value of `EOIT_3SAT`'s clauses. And, the third layer consists of terminal node T, the end of the project. Each of the first layer's n components corresponds to a variable and contains three nodes, denoted u_{j1}, u_{j2}, and u_{j3}, which are connected as shown in Figure 2.1. The figure also shows how the first layer connects to the third layer. We let $\Omega = \{1\}$, and $H^1 = 0$ with probability $p^1 = 1$. We define the parameter values

associated with the arcs in Figure 2.1 as:

$$D_{u_{j1},u_{j2}} = 1 \qquad d^1_{u_{j1},u_{j2}} = 1 \tag{2.3a}$$

$$D_{u_{j1},u_{j3}} = 2 \qquad d^1_{u_{j1},u_{j3}} = -1 \tag{2.3b}$$

$$D_{u_{j3},T} = 0 \qquad d^1_{u_{j3},T} = 0. \tag{2.3c}$$

No activities in the first layer can be crashed, i.e.,

$$J_{u_{jk}} = \varnothing, \text{ for all } j = 1, 2, \ldots, n, \ k = 1, 2, 3. \tag{2.4}$$

The start node, S, connects to each u_{j1} with zero duration. It is optimal to start each activity u_{j1} at time 0 because the inclusive inequalities in constraints (2.2e)-(2.2f) still allow us to choose $G^1_{u_{j1}} \in \{0, 1\}$ for each variable in U. With this setup, the length of the longer path through the j-th component is always 2, $j = 1, 2, \ldots, n$. Whether the longer path traverses activity u_{j2} (top path in Figure 2.1) or activity u_{j3} (bottom path) depends on the value of $G^1_{u_{j1}}$. If $G^1_{u_{j1}} = 1$ then the top path yields a length of $D_{u_{j1},u_{j2}} + d^1_{u_{j1},u_{j2}} = 1 + 1 = 2$, while the bottom path yields a length of $D_{u_{j1},u_{j3}} + d^1_{u_{j1},u_{j3}} = 2 - 1 = 1$. If $G^1_{u_{j1}} = 0$, then the top path yields a length of $D_{u_{j1},u_{j2}} = 1$, while the bottom path yields a length of $D_{u_{j1},u_{j3}} = 2$. We can consider the value of $G^1_{u_{j1}}$ as the truth assignment of variable u_j. If variable u_j is TRUE, the top path is longer; if it is FALSE the bottom path is longer. The arcs from activities u_{j2} and u_{j3} in Figure 2.1 point to activities in the second layer, which we now construct, again following ideas in De et al. (1997). Consider the clause $c_i = u_j \vee u_k \vee \bar{u}_\ell$, with literals consisting of two original variables, u_j and u_k, and one complement, $\bar{u}_\ell$. We consider an activity, c_{ip}, corresponding to the truth assignment of the variable u_p, for $p \in \{j, k, \ell\}$. For the original variables u_p, $p \in \{j, k\}$, we connect activity u_{p3} to activity c_{ip}, and we connect activity u_{p2} to the other two activities c_{iq}, where $q \in \{j, k, \ell\}, q \neq p$. For the complemented variable $\bar{u}_\ell$, we do the opposite, connecting activity $u_{\ell 2}$ to activity $c_{i\ell}$, and activity $u_{\ell 3}$ to activities c_{iq}, $q \in \{j, k\}$. This illustrates the general rule by which a clause with three literals

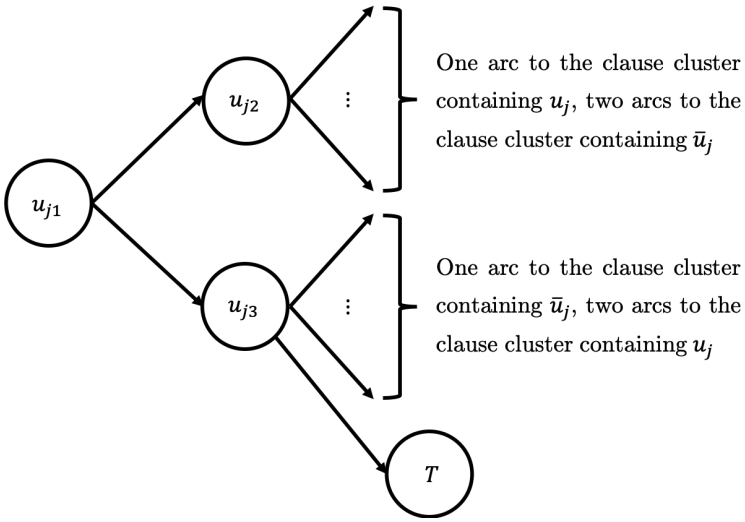


Figure 2.1: A component corresponding to variable u_j, $j = 1, 2, \ldots, n$, in the first layer of the activity network for `EOIT_3SAT`.

(typically a mix of original and complemented variables) yields the network topology:

- *original variables* result in a connection from u_{j3} to c_{ij} via a single arc and a connection from u_{j2} to the other two c_j-activity nodes; and, (2.5a)
- *complemented variables* result in the opposite. (2.5b)

From now on we refer to the activities representing variables as u-activities and those representing clauses as c-activities. We make the following assignments:

$$D_{u_{jp},c_{ij}} = d^1_{u_{jp},c_{ij}} = 0 \qquad \forall i = 1, \ldots, m, j = 1, \ldots, n, p = 2, 3 : u_j \text{ is in clause } c_i \tag{2.6a}$$

$$D_{c_{ij},T} = 1 \text{ and } d^1_{c_{ij},T} = 0 \qquad \forall i = 1, \ldots, m, j = 1, \ldots, n : u_j \text{ is in clause } c_i \tag{2.6b}$$

$$e_{c_{ij},1} = 1 \qquad \forall i = 1, \ldots, m, j = 1, \ldots, n : u_j \text{ is in clause } c_i \tag{2.6c}$$

$$b_{ij} = 1 \qquad \forall i = 1, \ldots, m, j = 1, \ldots, n : u_j \text{ is in clause } c_i \tag{2.6d}$$

$$B = 2m \tag{2.6e}$$

The nominal duration and the disrupted duration for the arcs between u-activities and c-activities

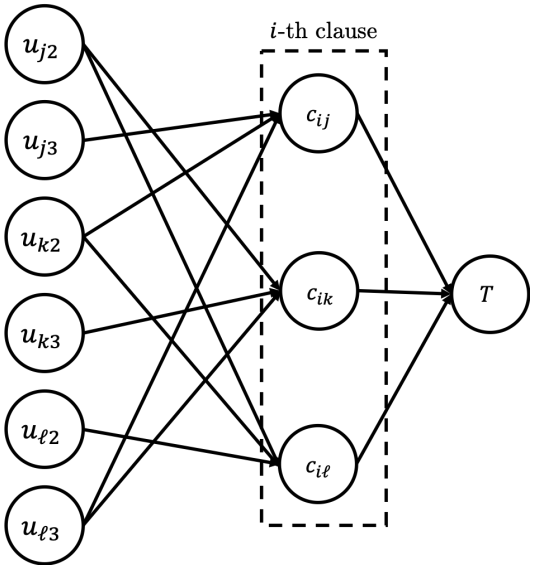


Figure 2.2: The i-th clause, $u_j \vee u_k \vee \bar{u}_\ell$, in the second layer of the constructed activity network for `EOIT_3SAT` with the arcs connecting it with the first and the third layer.

are 0 per equation (2.6a). The nominal and disrupted durations for the arcs between the c-activities and T are specified in equation (2.6b). Unlike the u-activities, each c-activity can be crashed with a single option with unit effectiveness as given in equation (2.6c). We assign the budget in equation (2.6e), where m is the total number of clauses, and assign unit b_{ij} values in equation (2.6d). We illustrate the logic behind this construction using Figure 2.2. For variable u_j, if $G^1_{u_{j1}} = 1$ then the earliest time activity u_{j2} can start is 2, and activity u_{j3} can start at time 1. If $G^1_{u_{j1}} = 0$ then the earliest time activity u_{j3} can start is 2, and activity u_{j2} can start at time 1. The same holds for variable u_k, and the opposite for variable u_ℓ. The truth assignments indicate which path is longer.

Next, we establish two lemmas, which relate start times at certain nodes in the activity network corresponding to an instance of `EOIT_3SAT`.

Lemma 2.3.1. *Consider an instance of* `EOIT_3SAT`*, and the corresponding activity network for this instance. For each clause, c_i, $i = 1, 2, \ldots, m$, there is at most one activity c_{iq^*}, $q^* \in \{j, k, \ell\}$, that has 1 as its earliest start time, and the start time for c_{iq} is 2, for $q \in \{j, k, \ell\}, q \neq q^*$.*

Proof of Lemma 2.3.1. The proof enumerates eight cases, and we begin with $c_i = u_j \vee u_k \vee u_\ell$, which has a component of the activity network illustrated in Figure 2.3. Without loss of generality,

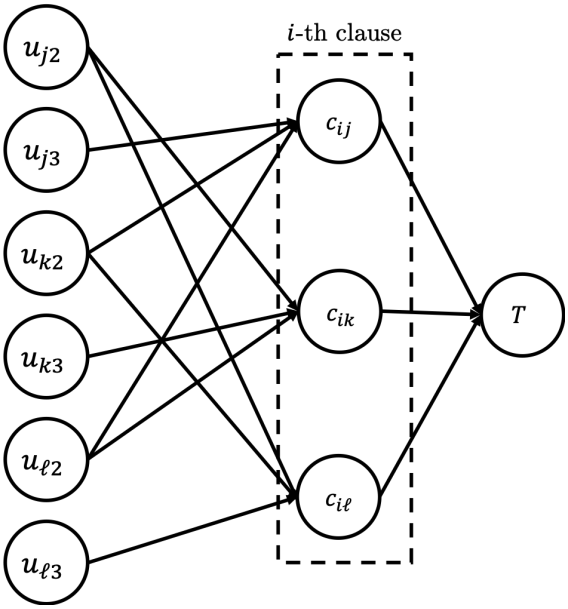


Figure 2.3: The i-th clause, $u_j \vee u_k \vee u_\ell$, in the second layer of the constructed activity network for `EOIT_3SAT` with the arcs connecting it with the first and the third layer

suppose both c_{ij} and c_{ik} can start at time 1. This means that both u_{j2} and u_{j3} need to start at time 1, which is impossible because regardless of $G^1_{u_{j1}}$'s value, at least one activity in $\{u_{j2}, u_{j3}\}$ can start no earlier than time 2. The proof is completed by enumerating the remaining seven cases—with variables u_j, u_k, u_ℓ in all combinations of original or complemented form—in analogous fashion. □

Lemma 2.3.2. *Consider an instance of* `EOIT_3SAT`*, and the corresponding activity network for this instance. For any clause, c_i, activity c_{iq^*}, $q^* \in \{j, k, \ell\}$, has an earliest start time of* 1 *if and only if the corresponding literal, u_{q^*} or $\bar{u}_{q^*}$, is the only literal in the clause to which the truth assignment is TRUE.*

Proof of Lemma 2.3.2. The proof again enumerates eight cases, and we begin with $c_i = u_j \vee u_k \vee u_\ell$; see Figure 2.3. Without loss of generality, we assume $q^* = j$.

($\Longrightarrow$): In turn we suppose u_j is FALSE or u_k is TRUE or u_ℓ is TRUE. First, suppose u_j is FALSE, i.e., $G^1_{u_{j1}} = 0$. Then activity u_{j3} can start no earlier than time 2, which leads to the contradiction that c_{ij} can start as early as time 1; see Figure 2.3. Suppose u_k is TRUE. Since there is an arc from u_{k2} to activity c_{ij}, and since u_{k2} cannot start before time 2 this again contradicts that c_{ij} can

start as early as time 1. The argument for u_ℓ being TRUE is identical. Therefore, if activity c_{ij} can start at time 1 then u_j is the only literal which is TRUE.

($\Longleftarrow$): if only u_j is TRUE, then u_{j3}, u_{k2} and $u_{\ell 2}$ can all start as early as time 1; again, see Figure 2.3. This means that the earliest start time for c_{ij} is 1.

We again complete the proof by enumerating the remaining seven cases. □

As a result of Lemmas 2.3.1 and 2.3.2, we can transform an `EOIT_3SAT` instance to an instance of model (2.2) using the activity network construction process just described. In particular, we know that if there exists a truth assignment to the variables of U that meets the requirement of `EOIT_3SAT`, there are exactly $2m$ c-activities (two per clause) that can start no earlier than time 2. Since we have budget $B = 2m$, we can crash all of those c-activities to achieve a project length of 2. If there is no truth assignment that meets the requirement of `EOIT_3SAT` then model (2.2)'s optimal value is 3. We formalize this in what follows.

Definition 2.3.1. Stochastic Crashing Decision Problem: *Is there a feasible solution,* (t, x, G)*, to model* (2.2) *with objective function value of at most* τ*?*

Theorem 2.3.1. *Consider an instance of* `EOIT_3SAT`*, and the corresponding activity network for this instance. In particular, let* $\Omega = \{1\}$*,* $H^1 = 0$*,* $p^1 = 1$*, and let the network topology and model parameters be given by Figure 2.1, rule* (2.5) *and equations* (2.3)*,* (2.4)*, and* (2.6)*. Let* $\tau = 2$*. The answer to the* Stochastic Crashing Decision Problem *is yes if and only if the given instance of* `EOIT_3SAT` *problem has a solution, i.e., a truth assignment to the variables so that each clause has exactly one true literal.*

Proof of Theorem 2.3.1. The `EOIT_3SAT` problem has n variables and m clauses, and the constructed activity network for the project crashing problem has $3n + 3m + 2$ activities and $4n + 12m$ arcs. Thus the size of the activity network and the time required to construct the network are both polynomial in the size of the original `EOIT_3SAT` instance.

($\Longleftarrow$) Suppose the `EOIT_3SAT` instance has a solution. A feasible solution to the instance of model (2.2) starts every activity as early as possible. Under the `EOIT_3SAT` hypothesis, by Lemmas 2.3.1 and 2.3.2 exactly $2m$ c-activities have earliest start times of 2 and the remaining m

have an earliest start time of 1. Spending the budget, $B = 2m$, to crash all $2m$ c-activities that correspond to the literals with FALSE assignment, yields an objective function value of 2.

($\Longrightarrow$) Suppose the instance of model (2.2) has a solution, $(\hat{t}, \hat{x}, \hat{G})$, with an objective function value of 2. By Lemma 2.3.1 we know that for each clause there is at most one c-activity that can start at time 1, which means there must be at least $2m$ c-activities with a start time of at least 2. Since $B = 2m$, if there are more than $2m$ c-activities that start at time 2 or after, the objective function value of model (2.2) must exceed 2. Hence, there are exactly $2m$ c-activities starting at time 2. Lemmas 2.3.1 and 2.3.2 then imply that exactly one c-activity in each of the m clauses starts at time 1; i.e,. for each clause there is exactly one variable to which the truth assignment is TRUE.

`EOIT_3SAT` is NP-complete (Garey and Johnson 1979). The Stochastic Crashing Decision Problem is in NP because we can check in $O(n+m)$ time whether a given solution is feasible and has an objective function value of at most τ. □

As a result of Theorem 2.3.1 we immediately obtain the following result.

Corollary 2.3.2. *Model* (2.2) *is NP-hard, even under a single disruption scenario, which occurs with probability one at time zero.*

2.4 Illustration of Problem Properties via Examples

We show two examples of serial activity networks to give insight regarding the nature of the project crashing problem under a stochastic disruption, and to draw distinctions relative to its deterministic counterpart. In the deterministic project crashing problem, all activities on the critical path should start as soon as possible. However, with a stochastic disruption, it is sometimes optimal to delay the start of one or more activities. In addition, under a stochastic disruption, it is possible that on a critical path, an activity with a shorter expected duration is crashed with a larger amount of resource, while in the deterministic case, it is always optimal to crash the activity with the longest duration on the critical path, under equal b_{ij} and e_{ij} values. We use two examples to show that the deterministic optimal solution can be significantly inferior in the stochastic setting because of these two properties. Here, we assume that required duration that separates the start

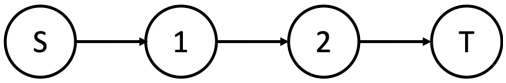


Figure 2.4: Example of a 2-activity serial network project

of activity i and the start of a successor, k, only depends on i; i.e., for each activity i we use D_i to denote the duration of activity i and d_i^ω to denote the change in duration under scenario ω:

$$D_{ik} = D_i \qquad \forall (i,k) \in \mathcal{A}$$
$$d_{ik}^\omega = d_i^\omega \qquad \forall (i,k) \in \mathcal{A}, \omega \in \Omega.$$

Clearly delaying the start of an activity may be beneficial when $d_i^\omega < 0$ for some $i \in I, \omega \in \Omega$ because the expected decrease in duration may exceed the delay required to move the start of activity i after a potential disruption. In the following example, we show value of delay, even if all activities are lengthened by the disruption; i.e., $d_i^\omega > 0, \forall i \in I, \omega \in \Omega$.

Example 2.4.1. *Consider a network with two activities in series, as shown in Figure 2.4 with* $I = \{S, 1, 2, T\}$*, and let parameter* $k > 4$*. Let the nominal durations be* $D_1 = k$ *and* $D_2 = 1$*. We assume only one crashing option for each activity, and so we omit index* j*. We let* $e_1 = e_2 = 1 - \frac{1}{2k}$*, assume* $b_i = 1$ *for all* $i \in I$*, and we let* $B = 1$*. Let* $\Omega = \{1\}$ *so that either we have no disruption with probability* $p^0 = 1 - \frac{1}{k}$*, or we have a disruption that occurs at time* $\varepsilon < \frac{1}{2}$ *with probability* $p^1 = \frac{1}{k}$*. If a disruption occurs, the nominal activity durations are lengthened by* $d_1 = k$ *and* $d_2 = (k-1)^2$*.*

If we start each activity without delay, then $t_1 = 0,\ x_1 = x_1^1$*, and for any* $x_1 \leqslant 1$*,* $t_2 \geqslant \frac{1}{2} \geqslant D_1(1 - e_1 x_1)$*, which means activity* 2 *will start after the disruption. Since* $k > 1$*, the duration of activity* 1*,* $D_1 = k$*, exceeds the expected duration of activity 2,* $D_2 + p^1 d_2 = 1 + \frac{1}{k}(k-1)^2 = k - 1 + \frac{1}{k}$*. As a result, it is optimal to spend the entire budget on activity* 1*:* $x_1 = x_1^1 = 1$ *and* $x_2 = x_2^1 = 0$*, and the expected project duration is:*

$$\begin{aligned} & D_1(1 - e_1 x_1) + p^0 D_2 + p^1 (D_2 + d_2) \\ =& k\left(1 - \left(1 - \frac{1}{2k}\right) \cdot 1\right) + (1 - \frac{1}{k}) \cdot 1 + \frac{1}{k}\left(1 + (k-1)^2\right) \end{aligned}$$

$$= k - \frac{1}{2} + \frac{1}{k}.$$

On the other hand, if we delay the start of activity 1 until $t_1 = \varepsilon$ *then* x_1 *and* x_1^1 *need not be equal. Since for* $k > 2 + \sqrt{2}$, $k = D_1 > D_2 = 1$ *and* $(k-1)^2 + 1 = D_2 + d_2 > D_1 + d_1 = k + k$, *we have* $x_1 = 1$, $x_1^1 = 0$, $x_2^1 = 1$ *in an optimal solution, and the expected duration is:*

$$\begin{aligned}
&\varepsilon + p^0 \left[D_1 \cdot \left(1 - (1 - \frac{1}{2k}) \cdot 1 \right) + D_2 \right] + p^1 \left[(D_1 + d_1) + \left(1 - (1 - \frac{1}{2k}) \cdot 1 \right) (D_2 + d_2) \right] \\
=&\varepsilon + \frac{k-1}{k} \left(k \cdot \frac{1}{2k} + 1 \right) + \frac{1}{k} \left[(k + k) + \frac{1}{2k} \left((k-1)^2 + 1 \right) \right] \\
=&\varepsilon + 4 - \frac{5}{2k} + \frac{1}{k^2}.
\end{aligned}$$

In Example 2.4.1 if require that activity 1 be started without delay then the objective function grows to infinity with k, but the optimal project span by delaying the start of activity 1 by ε has a constant limit of $\varepsilon + 4$. This example shows that the gap between the optimal solution under a no-delay policy and an optimal solution that allows for delay—as we do in model (2.2)—can be arbitrarily large. Because it is possible for an optimal crashing plan to contain a delay for some activities, model (2.2) uses decision variables t_i, $\forall i \in I$, as the start time of each activity, rather than assuming that each activity starts as soon as all of its predecessors are finished.

Example 2.4.2. *We again consider the network with two activities from Figure 2.4. Let* $D_2 > D_1$, $d_1 = 0$, $d_2 > 0$, $e_1 = e_2 = \frac{1}{2}$, *and* $B = 1$. *We again consider a single disruption scenario,* $\Omega = \{1\}$, *so that either we have no disruption,* $p^0 = \frac{1}{2}$, *or we have a disruption that occurs at time* $H^1 = \frac{1}{2}D_1$ *with probability* $p^1 = \frac{1}{2}$. *Here, the optimal solution is to crash the shorter activity, i.e.,* $x_1 = 1$, *which yields an expected project span of* $\frac{1}{2}D_1 + D_2$ *with start times* $t_1 = 0$ *and* $t_2 = \frac{1}{2}D_1$. *In contrast, if* $0 \leqslant x_1 < 1$ *then the expected duration is* $D_1(1 - \frac{1}{2}x_1) + (D_2 + \frac{1}{2}d_2)(1 - \frac{1}{2}x_2)$, *so that the ratio of the objective functions grows arbitrarily large as* d_2 *grows.*

In Example 2.4.2 the intuition behind crashing the shorter activity is that it allows us to initiate activity 2 in time to avoid incurring delay d_2. Both examples in this section suggest that the intuition associated with the deterministic version of the optimal crashing problem does not always apply in the stochastic setting, and provides further motivation for employing a model like

that in formulation (2.2).

2.5 Decomposition Method

Model (2.2) is a two-stage stochastic mixed-integer program, which we can rewrite as follows:

$$
\begin{aligned}
z^* = \min \quad & p^0 t_T + \sum_{\omega \in \Omega} p^\omega f^\omega(t, x) && && \text{(2.7a)}\\
\text{s.t.} \quad & t_k - t_i \geqslant D_{ik}\left(1 - \sum_{j \in J_i} e_{ij} x_{ij}\right) && \forall (i,k) \in \mathcal{A} && \text{(2.7b)}\\
& \sum_{i \in I} \sum_{j \in J_i} b_{ij} x_{ij} \leqslant B && && \text{(2.7c)}\\
& \sum_{j \in J_i} x_{ij} \leqslant 1 && \forall\, i \in I && \text{(2.7d)}\\
& t_i \geqslant 0 && \forall\, i \in I && \text{(2.7e)}\\
& 0 \leqslant x_{ij} \leqslant 1 && \forall\, i \in I, j \in J_i, && \text{(2.7f)}
\end{aligned}
$$

where

$$
\begin{aligned}
(S^\omega) \qquad f^\omega(\hat{t}, \hat{x}) = \min \quad & t_T && && \text{(2.8a)}\\
\text{s.t.} \quad & H^\omega + G_i M \geqslant \hat{t}_i && \forall\, i \in I && \text{(2.8b)}\\
& H^\omega - (1 - G_i) M \leqslant \hat{t}_i && \forall\, i \in I && \text{(2.8c)}\\
& t_i + G_i M_t \geqslant \hat{t}_i && \forall\, i \in I && \text{(2.8d)}\\
& t_i - G_i M_t \leqslant \hat{t}_i && \forall\, i \in I && \text{(2.8e)}\\
& x_{ij} + G_i \geqslant \hat{x}_{ij} && \forall\, i \in I, j \in J_i && \text{(2.8f)}\\
& x_{ij} - G_i \leqslant \hat{x}_{ij} && \forall\, i \in I, j \in J_i && \text{(2.8g)}\\
& t_k - t_i \geqslant D_{ik} + d^\omega_{ik} G_i - \sum_{j \in J_i} D_{ik} e_{ij} x_{ij} - \sum_{j \in J_i} d^\omega_{ik} e_{ij} z_{ij} && \forall\, (i,k) \in \mathcal{A} && \text{(2.8h)}\\
& \sum_{i \in I} \sum_{j \in J_i} b_{ij} x_{ij} \leqslant B && && \text{(2.8i)}
\end{aligned}
$$

$$\sum_{j \in J_i} x_{ij} \leqslant 1 \qquad \forall\, i \in I \tag{2.8j}$$

$$z_{ij} \leqslant G_i \qquad \forall\, i \in I, j \in J_i \tag{2.8k}$$

$$z_{ij} \leqslant x_{ij} \qquad \forall\, i \in I, j \in J_i \tag{2.8l}$$

$$z_{ij} \geqslant G_i + x_{ij} - 1 \qquad \forall\, i \in I, j \in J_i \tag{2.8m}$$

$$t_i \geqslant 0 \qquad \forall i \in I \tag{2.8n}$$

$$t_i \geqslant H^{\omega} G_i \qquad \forall\, i \in I \tag{2.8o}$$

$$0 \leqslant x_{ij} \leqslant 1 \qquad \forall\, i \in I, j \in J_i \tag{2.8p}$$

$$0 \leqslant z_{ij} \leqslant 1 \qquad \forall\, i \in I, j \in J_i \tag{2.8q}$$

$$G_i \in \{0, 1\} \qquad \forall\, i \in I. \tag{2.8r}$$

A number of existing approaches for stochastic mixed-integer programming assume a special structure not satisfied by our model. For example, Gade et al. (2014) solve two-stage stochastic programs with pure binary first stage variables, and general integer second stage variables; they derive a finitely convergent sequential convex approximation and a branch-and-cut framework involving Gomory cuts that are parameterized by the first-stage decision variables. Zou et al. (2016) assume state variables are binary (or general integer via binary expansion) in a multi-stage setting so that the Lagrangian cuts are a tight approximation of the recourse function; see also Philpott et al. (2019). Carøe and Tind (1998) solve a more general case of two-stage models by using integer programming duality, but there is limited computational work investigating their approach. Qi and Sen (2017) allow mixed-integer variables in both the first stage and the recourse problem, and parametric disjunctive cuts convexify recourse problems while Benders' cuts approximate recourse functions (Chen et al. 2012). Although this method suits our problem setting, preliminary computational results found that it was not competitive with the scheme we describe here, which makes use of the special structure of our problem.

A simple approach is to relax the integrality constraints (2.8r) of *subproblem* (S^{ω}), and execute a multi-cut L-shaped decomposition algorithm on this linear programming (LP) relax-

ation. The resulting optimality cuts provide a valid lower approximation of the recourse functions, $f^\omega(t,x), \forall \omega \in \Omega$, but may not be tight. In each iteration of the decomposition, an upper bound can be obtained by solving subproblems (2.8) with the first stage solution. The main challenge is how to iteratively tighten the lower bound while quickly locating a good upper bound. The topics in this section aim to tackle these two issues.

The combinatorial decision, $G_i \in \{0,1\}$, $i \in I$, for each (S^ω) is (almost fully) decided by the first-stage continuous variables, t_i. If $\hat{t}_i > H^\omega$ then $G_i = 1$, if $\hat{t}_i < H^\omega$ then $G_i = 0$, and only if they are equal is there a combinatorial choice. This observation motivates the decomposition algorithm that we develop. We could pull these binary variables to the first stage, but doing so involves $|I||\Omega|$ variables and does not scale well. Instead, we partition an interval, which we denote $[0, T_{\max}]$, containing each t_i, and we adaptively refine that partition. This helps control the number of binary first-stage variables, and has further benefits in terms of tightening lower bounds, as we describe in what follows. We will be specific later regarding the value of $T_{\max}$, but for now we simply assume we have a value such that $t_i \in [0, T_{\max}]$ is a redundant constraint in model (2.2).

We assume that $\Omega = \{1, 2, \ldots, |\Omega|\}$ is such that $\omega < \omega'$ implies $H^\omega \leqslant H^{\omega'}$ with strict inequality if the realizations of H are distinct, and we let $H^0 \equiv 0 \leqslant H^1$ and $H^{|\Omega|+1} \equiv T_{\max} \geqslant H^{|\Omega|}$. We define a partition of $[0, T_{\max}]$ for each $i \in I$ as follows.

Definition 2.5.1. *For each activity $i \in I$, we define a partition of interval $[0, T_{\max}]$ as an ordered set of two-element tuples $\mathcal{P}_i = \{[\underline{H}^q, \bar{H}^q],\ q \in \mathcal{Q}_i\}$ with an index set $\mathcal{Q}_i = \{1, 2, \ldots, |\mathcal{Q}_i|\}$ and the following properties:*

- $\underline{H}^1 = H^0 \equiv 0$
- $\bar{H}^{|\mathcal{Q}_i|} = H^{|\Omega|+1} \equiv T_{\max}$
- $\underline{H}^q < \bar{H}^q \quad \forall q \in \mathcal{Q}_i$
- $\bar{H}^q = \underline{H}^{q+1} \quad \forall q \in \mathcal{Q}_i$.

With the possible exceptions of $\underline{H}^1$ and $\bar{H}^{|\mathcal{Q}_i|}$, each $\underline{H}^q$ and $\bar{H}^q$ corresponds to a disruption time of some scenario, and a simple example is illustrated in Figure 2.5 in which we have five scenarios, $\Omega = \{1, 2, 3, 4, 5\}$. The partition has three intervals as illustrated. The second interval has lower bound H^2 and upper bound H^5.

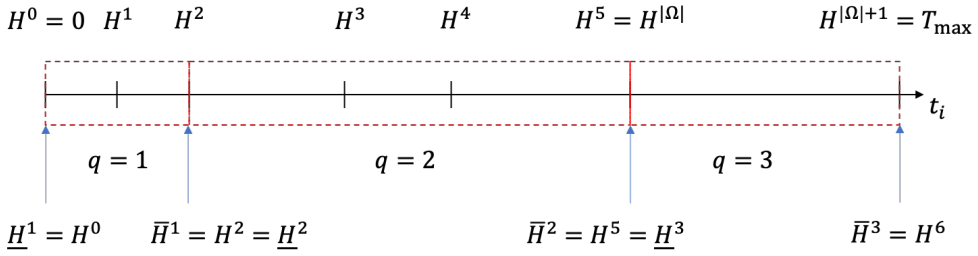


Figure 2.5: An illustration of a partition of interval $[0, T_{\max}]$

For each activity $i \in I$ the first-stage start time t_i lies an interval of $\mathcal{P}_i$, and we introduce a first-stage indicator variable:

$$\sum_{q \in \mathcal{Q}_i} \underline{H}^q y_i^q \leqslant t_i \leqslant \sum_{q \in \mathcal{Q}_i} \bar{H}^q y_i^q \qquad \forall i \in I \tag{2.9a}$$

$$\sum_{q \in \mathcal{Q}_i} y_i^q = 1 \qquad \forall i \in I \tag{2.9b}$$

$$y_i^q \in \{0, 1\}, \qquad \forall i \in I, q \in \mathcal{Q}_i. \tag{2.9c}$$

Constraints (2.9) require that t_i be associated with one of the intervals of the partition. In model (2.2) if $t_i = H^\omega$ then G_i^ω can either be 0 (activity i is said to start before ω's disruption) or 1 (i starts after the disruption). If $t_i = \bar{H}^q = \underline{H}^{q+1} = H^\omega$ for some ω then the y-variables have a similar choice, and our convention is that if the y-variables choose $t_i \in [\underline{H}^q, \bar{H}^q]$ then activity i is said to start before the disruption and if $t_i \in [\underline{H}^{q+1}, \bar{H}^{q+1}]$ then i starts after the disruption.

2.5.1 Tightening Big-M with Partitions

The tightness of (S^ω)'s LP relaxation relies, in part, on the big-M value used in constraints to represent the logical condition of whether activity i starts before or after a disruption. A smaller, but still valid, big-M value yields a tighter relaxation, and can further help prevent numerical issues (e.g., Camm et al. 1990, Klotz and Newman 2013). We can rewrite constraints (2.8b) and (2.8c) as:

$$(\hat{t}_i - H^\omega)/M \leqslant G_i \leqslant (\hat{t}_i - H^\omega)/M + 1. \tag{2.10}$$

Variable G_i can take a wider range of values when M is large. Tightening M hinges on

specifying valid ranges for $\hat{t}_i$. Furthermore, we know that if $\hat{t}_i > H^{\omega'}$ for some ω' then G_i has to take value 1 in all subproblems (S^ω) with $\omega < \omega'$. On the other hand, if $\hat{t}_i < H^{\omega'}$ for some ω' then G_i must be 0 for all subproblems (S^ω) in which $\omega > \omega'$. If we can fix the G_i for all $i \in I$ to either 0 or 1 the resulting optimality cuts will be tight.

Proposition 2.5.1. *Let t_i^0 be the longest S-i path in the activity network $\mathcal{G} = (I, \mathcal{A})$ in which the arc length of $(i,k) \in \mathcal{A}$ is D_{ik} and in which no crashing is allowed. Let t^* denote (part of) an optimal solution to model (2.2). Then there exists a t^* such that $t_i^* \in [0, H^{|\Omega|} + t_i^0],\ \forall i \in I$ provided M and M_t are sufficiently large.*

Proof. Constraint (2.2q) enforces the lower bound of 0.

By hypothesis $t_k^0 - t_i^0 \geqslant D_{ik}, \forall (i,k) \in \mathcal{A}$ since t_i^0 is the longest S-i path of $\mathcal{G}$ in which the arc length of $(i,k) \in \mathcal{A}$ is D_{ik} and in which no crashing is allowed. We prove the upper bound on t_i^* by contradiction.

Suppose there does not exist a t^* such that $t_i^* \in [0, H^{|\Omega|} + t_i^0],\ \forall i \in I$. Then for every t^*, there must be a set $I^* \subseteq I$ such that $t_i^* > H^{|\Omega|} + t_i^0$ for $i \in I^*$. Let the corresponding optimal values of variables $x_{ij}, G_i^\omega, t_i^\omega, x_{ij}^\omega, z_{ij}^\omega$ be denoted $x_{ij}^*, G_i^{\omega,*}, t_i^{\omega,*}, x_{ij}^{\omega,*}, z_{ij}^{\omega,*}$, respectively. We can establish a feasible solution to model (2.2) as follows:

$$\tilde{t}_i = H^{|\Omega|} + t_i^0 \qquad \forall i \in I^* \tag{2.11a}$$

$$\tilde{t}_i = t_i^* \qquad \forall i \in I \backslash I^* \tag{2.11b}$$

$$\tilde{x}_{ij} = x_{ij}^* \qquad \forall i \in I, j \in J_i \tag{2.11c}$$

$$\tilde{G}_i^\omega = G_i^{\omega,*} \qquad \forall i \in I, \omega \in \Omega \tag{2.11d}$$

$$\tilde{t}_i^\omega = t_i^{\omega,*} \qquad \forall i \in I, \omega \in \Omega \tag{2.11e}$$

$$\tilde{x}_{ij}^\omega = x_{ij}^{\omega,*} \qquad \forall i \in I, j \in J_i, \omega \in \Omega \tag{2.11f}$$

$$\tilde{z}_{ij}^\omega = z_{ij}^{\omega,*} \qquad \forall i \in I, j \in J_i, \omega \in \Omega. \tag{2.11g}$$

We see this solution is feasible by examining the constraints of model (2.2):

- For constraint (2.2b), we examine the following four possible cases:

– $i \in I^*, k \in I^*$: since $\tilde{x}_{ij} \geqslant 0$ and $t_k^0 - t_i^0 \geqslant D_{ik}$, we have

$$\tilde{t}_k - \tilde{t}_i = t_k^0 - t_i^0 \geqslant D_{ik} \geqslant D_{ik}\left(1 - \sum_{j \in J_i} e_{ij}\tilde{x}_{ij}\right);$$

– $i \in I^*, k \notin I^*$: we have $\tilde{t}_i < t_i^*$ since $i \in I^*$, and then

$$\tilde{t}_k - \tilde{t}_i > t_k^* - t_i^* \geqslant D_{ik}\left(1 - \sum_{j \in J_i} e_{ij}x_{ij}^*\right) = D_{ik}\left(1 - \sum_{j \in J_i} e_{ij}\tilde{x}_{ij}\right);$$

– $i \notin I^*, k \in I^*$: since $t_i^* \leqslant H^{|\Omega|} + t_i^0$, $\tilde{x}_{ij} \geqslant 0$ and $t_k^0 - t_i^0 \geqslant D_{ik}$, we have

$$\tilde{t}_k - \tilde{t}_i = H^{|\Omega|} + t_k^0 - t_i^* \geqslant H^{|\Omega|} + t_k^0 - \left(H^{|\Omega|} + t_i^0\right) = t_k^0 - t_i^0 \geqslant D_{ik} \geqslant D_{ik}\left(1 - \sum_{j \in J_i} e_{ij}\tilde{x}_{ij}\right);$$

– $i, k \notin I^*$: the constraint is unchanged and feasible;

- For variable G_i^ω:
  – if $i \notin I^*$: we have $\tilde{t}_i = t_i^*$. Therefore, constraints (2.2e)-(2.2h) is unchanged and feasible;
  – if $i \in I^*$: variable $G_i^{*,\omega}$ is forced to take value 1 for all $\omega \in \Omega$. Since $\tilde{t}_i = H^{|\Omega|} + t_i^0 \geqslant H^{|\Omega|}$, for any $\omega \in \Omega$, $\tilde{G}_i^\omega = G_i^{*,\omega} = 1$ remains feasible. Therefore, constraints (2.2e)-(2.2h) hold for M and M_t sufficiently large.
- Since the values for $\tilde{t}_i^\omega, \tilde{x}_{ij}, \tilde{x}_{ij}^\omega, G_i^\omega, z_{ij}^\omega$ all remain the same, constraints (2.2c), (2.2d), (2.2i)-(2.2v) are all satisfied by the solution in (2.11).

□

For simplicity of exposition, we use a uniform upper bound on every t_i for $i \in I$ as $T_{\max} = H^{|\Omega|} + t_T^0$. While Proposition 2.5.1 bounds the start-time variables, t_i, $\forall i \in I$, to the interval $[0, T_{\max}]$, the y-variables of (2.9) allow for tighter bounds. Given a partition for each activity and given a first-stage solution with $\hat{y}_i^q, \forall i \in I, q \in \mathcal{Q}_i$, we can replace constraints (2.8b) and (2.8c) with:

$$H^\omega + G_i\left(\sum_{q \in \mathcal{Q}_i} \bar{H}^q \hat{y}_i^q - H^\omega\right) \geqslant \hat{t}_i \qquad \forall i \in I \tag{2.12a}$$

$$H^\omega - (1 - G_i)\left(H^\omega - \sum_{q \in \mathcal{Q}_i} \underline{H}^q \hat{y}_i^q\right) \leqslant \hat{t}_i \qquad \forall i \in I. \tag{2.12b}$$

In a first-stage solution, $\hat{y}_i^q = 1$ for a specific q satisfying $\underline{H}^q \leqslant \hat{t}_i \leqslant \bar{H}^q$, and verifying the validity of constraints (2.12) in replacing (2.8b)-(2.8c) is straightforward by enumerating the cases $H^\omega \in [\underline{H}^q, \bar{H}^q]$, $H^\omega < \underline{H}^q$, and $H^\omega > \bar{H}^q$. In the degenerate case in which H^ω coincides with $\bar{H}^q$ or $\underline{H}^q$, the corresponding constraint from (2.12) is as tight as possible, i.e., a simple bound on $\hat{t}_i$ involving H^ω. Otherwise constraints (2.12) reduce to the following analog of (2.10):

$$(\hat{t}_i - H^\omega)/(\bar{H}^q - H^\omega) \leqslant G_i \leqslant (\hat{t}_i - H^\omega)/(H^\omega - \underline{H}^q) + 1, \tag{2.13}$$

and we see that smaller values of $\bar{H}^q$ and larger values of $\underline{H}^q$ have the effect of tightening the big-M value in constraints (2.10).

We introduce variables F_i^q to linearize the bilinear terms, and rewrite constraints (2.12) as follows:

$$\sum_{q \in \mathcal{Q}_i} \bar{H}^q F_i^q - H^\omega G_i \geqslant \hat{t}_i - H^\omega \qquad \forall i \in I \tag{2.14a}$$

$$H^\omega G_i - \sum_{q \in \mathcal{Q}_i} \underline{H}^q F_i^q \leqslant \hat{t}_i - \sum_{q \in \mathcal{Q}_i} \underline{H}^q \hat{y}_i^q \qquad \forall i \in I \tag{2.14b}$$

$$F_i^q \leqslant G_i \qquad \forall i \in I, q \in \mathcal{Q}_i \tag{2.14c}$$

$$F_i^q \leqslant \hat{y}_i^q \qquad \forall i \in I, q \in \mathcal{Q}_i \tag{2.14d}$$

$$F_i^q \geqslant G_i + \hat{y}_i^q - 1 \qquad \forall i \in I, q \in \mathcal{Q}_i. \tag{2.14e}$$

We can further tighten the formulation by adding two constraints involving y that cover cases when G_i can be fixed to 0 or 1. Again given a partition of each activity and a first-stage solution $\hat{y}_i^q, \forall i \in I, q \in \mathcal{Q}_i$, we have:

$$\sum_{q \in \mathcal{Q}_i, H^\omega \leqslant \underline{H}^q} \hat{y}_i^q \leqslant G_i \leqslant 1 - \sum_{q \in \mathcal{Q}_i, H^\omega \geqslant \bar{H}^q} \hat{y}_i^q \qquad \forall i \in I. \tag{2.15}$$

For the case in which $H^\omega \in (\underline{H}^q, \bar{H}^q)$ for the q with $\hat{y}_i^q = 1$, constraint (2.15) adds no restriction,

but for the cases in which $H^\omega \leqslant \underline{H}^q$ and $H^\omega \geqslant \bar{H}^q$, G_i is forced to 1 and 0, respectively.

2.5.2 Partition-based Decomposition Method

Our decomposition algorithm to solve model (2.2) iteratively partitions the continuous feasible region of the first-stage t-variables by introducing binary variables that facilitate tighter optimality cuts. With the addition of constraints (2.14) and (2.15), the tightened subproblem is:

$$(S^\omega_{\mathcal{P}}) \qquad f^\omega_{\mathcal{P}}(\hat{t},\hat{x},\hat{y}) = \min \quad t_T \tag{2.16a}$$

$$\text{s.t.} \quad \sum_{q\in\mathcal{Q}_i} \bar{H}^q F_i^q - H^\omega G_i \geqslant \hat{t}_i - H^\omega \qquad \forall i \in I \tag{2.16b}$$

$$H^\omega G_i - \sum_{q\in\mathcal{Q}_i} \underline{H}^q F_i^q \leqslant \hat{t}_i - \sum_{q\in\mathcal{Q}_i} \underline{H}^q \hat{y}_i^q \qquad \forall i \in I \tag{2.16c}$$

$$F_i^q \leqslant G_i \qquad \forall i \in I, q \in \mathcal{Q}_i \tag{2.16d}$$

$$F_i^q \leqslant \hat{y}_i^q \qquad \forall i \in I, q \in \mathcal{Q}_i \tag{2.16e}$$

$$F_i^q \geqslant G_i + \hat{y}_i^q - 1 \qquad \forall i \in I, q \in \mathcal{Q}_i. \tag{2.16f}$$

$$G_i \geqslant \sum_{q\in\mathcal{Q}_i, H^\omega \leqslant \underline{H}^q} \hat{y}_i^q \qquad \forall i \in I \tag{2.16g}$$

$$G_i \leqslant 1 - \sum_{q\in\mathcal{Q}_i, H^\omega \geqslant \bar{H}^q} \hat{y}_i^q \qquad \forall i \in I \tag{2.16h}$$

$$\text{constraints (2.8d)-(2.8q)} \tag{2.16i}$$

$$0 \leqslant F_i^q \leqslant 1 \qquad \forall\, i \in I, q \in \mathcal{Q}_i \tag{2.16j}$$

$$0 \leqslant G_i \leqslant 1. \qquad \forall\, i \in I. \tag{2.16k}$$

We express $(S^\omega_{\mathcal{P}})$ in LP relaxation form, excluding constraints (2.8r). Let ℓ denote the iteration of the decomposition algorithm, and $(\hat{t}^\ell, \hat{x}^\ell, \hat{y}^\ell)$ denote a given first-stage decision. We solve $(S^\omega_{\mathcal{P}})$ for each $\omega \in \Omega$ and construct an optimality cut of the form:

$$\theta^\omega \geqslant v^{\omega,\ell} + \sum_{i\in I} \pi_i^{\omega,\ell}(t_i - \hat{t}_i^\ell) + \sum_{i\in I}\sum_{j\in J_i} \lambda_{ij}^{\omega,\ell}(x_{ij} - \hat{x}_{ij}^\ell) + \sum_{i\in I}\sum_{q\in\mathcal{Q}_i^\ell} \gamma_i^{\omega,\ell,q}\left(y_i^q - \hat{y}_i^{q,\ell}\right). \tag{2.17}$$

Here, θ^ω is a continuous decision variable that forms an outer-linearization of $f^\omega_{\mathcal{P}}(t,x,y)$; parameter $v^{\omega,\ell}$ is the optimal value of $(S^\omega_{\mathcal{P}})$ at $(\hat{t}^\ell, \hat{x}^\ell, \hat{y}^\ell)$; and, coefficients π, λ and γ are appropriate sums of dual variables from the LP relaxation—e.g., $\pi_i^{\omega,\ell}$ involves dual variables from constraints (2.8d)-(2.8e) and (2.16b)-(2.16c). Since we solve a linear relaxation, θ^ω is a lower bound on $f^\omega(\hat{t}^\ell, \hat{x}^\ell, \hat{y}^\ell)$. However, the cut needs to be modified once the partition is updated to maintain validity, and we assume that the update only refines the partition for each $i \in I$:

Definition 2.5.2. *For two partitions $\mathcal{P}_i^1$ and $\mathcal{P}_i^2$, indexed by $\mathcal{Q}_i^1$ and $\mathcal{Q}_i^2$, respectively, we say $\mathcal{P}_i^2$ is a* refinement *of $\mathcal{P}_i^1$ provided:*

$$\forall q^2 \in \mathcal{Q}_i^2, \exists q^1 \in \mathcal{Q}_i^1 \text{ s.t. } \bar{H}^{q^1} \geqslant \bar{H}^{q^2} \text{ and } \underline{H}^{q^1} \leqslant \underline{H}^{q^2}.$$

At the current iteration for each $i \in I$, let the partition $\mathcal{P}_i$ be indexed by $\mathcal{Q}_i$, and assume this partition is formed from earlier partitions by a sequence of refinements satisfying Definition 2.5.2. We can then find a set of intervals in the current partition, $\mathcal{P}_i$, whose union is the q-th interval in partition $\mathcal{P}_i^\ell$ from previous iteration ℓ. We index such a *descendant set* by $\Delta_i(\ell, q)$. Cut (2.17) can then be updated to the following form:

$$\theta^\omega \geqslant v^{\omega,\ell} + \sum_{i\in I} \pi_i^{\omega,\ell}(t_i - \hat{t}_i^\ell) + \sum_{i\in I}\sum_{j\in J_i} \lambda_{ij}^{\omega,\ell}(x_{ij} - \hat{x}_{ij}^\ell) + \sum_{i\in I}\sum_{q\in\mathcal{Q}_i^\ell} \gamma_i^{\omega,\ell,q}\left(\sum_{q'\in\Delta_i(\ell,q)} y_i^{q'} - \hat{y}_i^{q,\ell}\right). \quad (2.18)$$

We show that given a partition, $\mathcal{P} = \times_{i\in I}\mathcal{P}_i$, which is updated by sequential refinement from a previous partition, $\mathcal{P}^\ell = \times_{i\in I}\mathcal{P}_i^\ell$, the optimality cut (2.18) is a valid lower approximation for $f^\omega_{\mathcal{P}}(t,x,y)$.

Proposition 2.5.2. *For each $i \in I$, suppose we have a partition, $\mathcal{P}_i$, indexed by $\mathcal{Q}_i$, which is a refinement of $\mathcal{P}_i^\ell$, indexed by $\mathcal{Q}_i^\ell$. Then at any given feasible (t,x,y) we have*

$$f^\omega_{\mathcal{P}}(t,x,y) \geqslant v^{\omega,\ell} + \sum_{i\in I} \pi_i^{\omega,\ell}(t_i - \hat{t}_i^\ell) + \sum_{i\in I}\sum_{j\in J_i} \lambda_{ij}^{\omega,\ell}(x_{ij} - \hat{x}_{ij}^\ell)\sum_{i\in I}\sum_{q\in\mathcal{Q}_i^\ell} \gamma_i^{\omega,\ell,q}\left(\sum_{q'\in\Delta_i(\ell,q)} y_i^{q'} - \hat{y}_i^{q,\ell}\right). \quad (2.19)$$

Proof. We denote the recourse function corresponding to partition $\mathcal{P}^\ell$ by $f^\omega_{\mathcal{P}^\ell}(t,x,y)$, where y has

the correct dimension according to $\mathcal{P}^\ell$. We first show that

$$f^\omega_{\mathcal{P}}(t,x,y) \geqslant f^\omega_{\mathcal{P}^\ell}(t,x,\tilde{y}), \tag{2.20}$$

where

$$\tilde{y}^q_i = \sum_{q' \in \Delta_i(\ell,q)} y^{q'}_i \qquad \forall i \in I, q \in \mathcal{Q}^\ell_i. \tag{2.21}$$

Suppose for given (t,x,y), we solve $(S^\omega_{\mathcal{P}})$ and obtain an optimal solution $(t^\omega, x^\omega, G^\omega, F^\omega)$. We then form

$$\tilde{F}^{\omega,q}_i = \sum_{q' \in \Delta_i(\ell,q)} F^{\omega,q'}_i \qquad \forall i \in I, q \in \mathcal{Q}^\ell_i,$$

and we obtain a feasible solution $(t^\omega, x^\omega, G^\omega, \tilde{F}^\omega)$ to subproblem $(S^\omega_{\mathcal{P}^\ell})$. Therefore, inequality (2.20) holds. Furthermore, the cut generated under partition $\mathcal{P}^\ell$ is

$$\theta^\omega \geqslant v^{\omega,\ell} + \sum_{i\in I} \pi^{\omega,\ell}_i (t_i - \hat{t}^\ell_i) + \sum_{i\in I}\sum_{j\in J_i} \lambda^{\omega,\ell}_{ij}(x_{ij} - \hat{x}^\ell_{ij}) + \sum_{i\in I}\sum_{q\in\mathcal{Q}^\ell_i} \gamma^{\omega,\ell,q}_i \left(y^q_i - \hat{y}^{q,\ell}_i\right),$$

which means that for any feasible $(t,x,\tilde{y})$, we have

$$f^\omega_{\mathcal{P}^\ell}(t,x,\tilde{y}) \geqslant v^{\omega,\ell} + \sum_{i\in I} \pi^{\omega,\ell}_i (t_i - \hat{t}^\ell_i) + \sum_{i\in I}\sum_{j\in J_i} \lambda^{\omega,\ell}_{ij}(x_{ij} - \hat{x}^\ell_{ij}) + \sum_{i\in I}\sum_{q\in\mathcal{Q}^\ell_i} \gamma^{\omega,\ell,q}_i \left(\tilde{y}^q_i - \hat{y}^{q,\ell}_i\right). \tag{2.22}$$

Using equation (2.21), we replace $\tilde{y}$ by y, and combine inequalities (2.20) and (2.22) to obtain (2.19). □

Proposition 2.5.2 states that by properly modifying the y-variables in cuts generated under earlier partitions, the resulting cuts (2.18) are valid in the sense of providing a lower approximation on the LP relaxation of the recourse function. Therefore, we incorporate the modified cuts (2.18) in the following master problem, given a partition $\mathcal{P}$, which is indexed by $\mathcal{Q}$:

$$(M_{\mathcal{P}}) \quad z^*_{\mathcal{P}} = \min \quad p^0 t_T + \sum_{\omega\in\Omega} p^\omega \theta^\omega \tag{2.23a}$$

$$\text{s.t.} \quad t_k - t_i \geqslant D_{ik}\left(1 - \sum_{j\in J_i} e_{ij}x_{ij}\right) \qquad \forall (i,k) \in \mathcal{A} \tag{2.23b}$$

$$\sum_{i\in I}\sum_{j\in J_i} b_{ij}x_{ij} \leqslant B \tag{2.23c}$$

$$\sum_{j\in J_i} x_{ij} \leqslant 1 \qquad \forall\, i \in I \tag{2.23d}$$

$$\sum_{q\in\mathcal{Q}_i} \underline{H}^q y_i^q \leqslant t_i \leqslant \sum_{q\in\mathcal{Q}_i} \bar{H}^q y_i^q \qquad \forall i \in I \tag{2.23e}$$

$$\sum_{q\in\mathcal{Q}_i} y_i^q = 1 \qquad \forall i \in I \tag{2.23f}$$

$$\theta^\omega \geqslant v^{\omega,\ell} + \sum_{i\in I} \pi_i^{\omega,\ell}(t_i - \hat{t}_i^\ell) + \sum_{i\in I}\sum_{j\in J_i} \lambda_{ij}^{\omega,\ell}(x_{ij} - \hat{x}_{ij}^\ell)$$
$$+ \sum_{i\in I}\sum_{q\in\mathcal{Q}_i^\ell} \gamma_i^{\omega,\ell,q}\left(\sum_{q'\in\Delta_i(\ell,q)} y_i^{q'} - \hat{y}_i^{q,\ell}\right) \qquad \forall \omega\in\Omega, \ell = 1,2,\ldots,L \tag{2.23g}$$

$$y_i^q \in \{0,1\} \qquad \forall i \in I, q \in \mathcal{Q}_i \tag{2.23h}$$

$$t_i \geqslant 0 \qquad \forall\, i \in I \tag{2.23i}$$

$$0 \leqslant x_{ij} \leqslant 1 \qquad \forall\, i \in I, j \in J_i. \tag{2.23j}$$

As we refine the partitions, the generated cuts become tighter and we provide tighter lower bounds on model (2.2)'s optimal value. We refine the partition by selecting the interval of $\mathcal{P}_i$ for each activity $i \in I$, where for some scenarios $\omega \in \Omega$ with $H^\omega \in [\underline{H}^q, \bar{H}^q]$, G_i^ω has a fractional value, and partition the interval as we describe in further detail in Section 2.5.7. The decomposition procedure is given in Algorithm 1.

We prove Algorithm 1 converges in finite number of iterations. Since every partition update is a refinement and we have a finite set of scenario Ω, we can prove the finite convergence of Algorithm 1 as long as with the finest partition we reach the optimum of problem (2.2).

Proposition 2.5.3. *Assume that the realizations of H^ω, $\omega \in \Omega$, are distinct. Assume that for each*

Algorithm 1 Partition-based decomposition algorithm to solve model (2.2)

Initialize cut iteration number $\ell = 0$, lower bound $LB = 0$, upper bound $UB = +\infty$, initial partition $\mathcal{P}^\ell$ with its indexed set $\mathcal{Q}^\ell$, and tolerance parameters $\varepsilon \geqslant \delta \geqslant 0$;
while $\frac{UB-LB}{UB} > \varepsilon$ **do**
  Solve $(M_{\mathcal{P}})$ and obtain solution $\hat{t}^\ell, \hat{x}^\ell, \hat{y}^\ell, \hat{\theta}^\ell$ and optimal value $z^*_{\mathcal{P}}$;
  if $z^*_{\mathcal{P}} > LB$ **then**
    Update $LB = z^*_{\mathcal{P}}$;
  For each $\omega \in \Omega$, solve (S^ω) and obtain $f^\omega(\hat{t}^\ell, \hat{x}^\ell)$ and $\hat{G}^\ell$;
  Calculate $\bar{z} = p^0 \hat{t}^\ell_T + \sum_{\omega\in\Omega} p^\omega f^\omega(\hat{t}^\ell, \hat{x}^\ell)$;
  if $\bar{z} < UB$ **then**
    Update $UB = \bar{z}$ and incumbent solution as $t^* = \hat{t}^\ell, x^* = \hat{x}^\ell$ and $G^* = \hat{G}^\ell$;
  for each $\omega \in \Omega$ **do**
    solve $(S^\omega_{\mathcal{P}})$ given $\hat{t}^\ell, \hat{x}^\ell, \hat{y}^\ell$ and obtain optimal value $v^{\omega,\ell}$ and $\pi^{\omega,\ell}, \lambda^{\omega,\ell}, \gamma^{\omega,\ell}$;
    if $\hat{\theta}^{\omega,\ell} < v^{\omega,\ell} - \delta$ **then** add cut of form (2.17) to $(M_{\mathcal{P}})$;
  if there are cuts added **then**
    Let $\mathcal{P}^{\ell+1} = \mathcal{P}^\ell$ and $\mathcal{Q}^{\ell+1} = \mathcal{Q}^\ell$;
    Let $\ell = \ell + 1$;
  else
    Refine the partition and obtain the new partition $\mathcal{P}^{\ell+1}$ and its indexed sets $\mathcal{Q}^{\ell+1}$;
    Let $\ell = \ell + 1$;
    Update previously generated cuts in $\mathcal{P}^\ell$ to the form of (2.18);
end while
Output UB as the ε-optimal value of model (2.2), and (t^*, x^*, G^*) as the ε-optimal solution.

partition, $\mathcal{P}_i$ indexed by $\mathcal{Q}_i$, and for each H^ω we have $H^\omega = \bar{H}^q$ for some $q \in \mathcal{Q}_i$, i.e., $|\mathcal{Q}_i| = |\Omega| + 1$ so that each partition is as fine as possible. Then

$$z^* = \min_{(t,x,y)\in\mathbb{X}} \; p^0 t_T + \sum_{\omega\in\Omega} p^\omega f^\omega_{\mathcal{P}}(t, x, y), \tag{2.24}$$

where

$$\mathbb{X} = \left\{ (t,x,y) \;\middle|\; \begin{array}{ll} \textit{constraints } (2.23b)\text{-}(2.23f) & \\ y^q_i \in \{0,1\} & \forall i \in I, q \in \mathcal{Q}_i \\ t_i \geqslant 0 & \forall i \in I \\ 0 \leqslant x_{ij} \leqslant 1 & \forall i \in I, j \in J_i \end{array} \right\}$$

and where z^* is the optimal value of model (2.2).

Proof. We can formulate an extensive form for model (2.24), i.e., an analog of model (2.2), using $\mathbb{X}$

and model (2.16), where the decision variables of model (2.16) are now also indexed by ω. Let $\Omega = \{1, 2, \ldots, |\Omega|\}$ so that the realizations of the disruption times are given by $H^1 < H^2 < \cdots < H^{|\Omega|}$. By Definition 2.5.1, under the hypothesis of the proposition, the intervals of each partition are given by the finest partition, $[H^0, H^1], [H^1, H^2], \ldots, [H^{|\Omega|-1}, H^{|\Omega|}], [H^{|\Omega|}, H^{|\Omega|+1}]$, which we can index by $q = 1, 2, \ldots, |\Omega|, |\Omega| + 1$, where $H^0 = 0$ and $H^{|\Omega|+1} = T_{\max}$. Under the assumed partition, constraints (2.16g)-(2.16h) in the extensive form of (2.24) reduce to

$$\sum_{q=\omega+1}^{|\Omega|+1} y_i^q \leqslant G_i^\omega \leqslant 1 - \left(\sum_{q=1}^{\omega} y_i^q \right), \forall i \in I, \omega \in \Omega. \tag{2.25}$$

Under this finest partition, there is a one-to-one mapping between the G- and y-variables under the binary restrictions imposed by models (2.2) and (2.24). In particular,

$$y_i^q = 1 \text{ if and only if } G_i^\omega = 1 \ \forall \omega \geqslant q + 1 \text{ and } G_i^\omega = 0 \ \forall \omega \leqslant q. \tag{2.26}$$

Any feasible solution to model (2.2) necessarily satisfies the condition $G_i^\omega \leqslant G_i^{\omega+1}$ required by (2.26) via constraint (2.2f). And, constraints (2.23f) and (2.23h) ensures the 0-0 or 1-1 nature of the left- and right-hand side of (2.25). By Proposition 2.5.1 we know $t_i \in [0, T_{\max}]$, and hence this restriction imposed by model (2.24) is nonbinding. As a result, model (2.2) and the extensive form of model (2.24) are equivalent and yield the same optimal value. □

Theorem 2.5.1. *Assume that the realizations of H^ω, $\omega \in \Omega$, are distinct, and assume that we obtain a dual extreme-point solution to subproblem $(S_{\mathcal{P}}^\omega)$ in step 11 of Algorithm 1. Then, the algorithm terminates in finite number of iterations to an ε-optimal solution to model* (2.2) *for any $\varepsilon \geqslant 0$.*

Proof. Let z^* denote the optimal value of model (2.2) or equivalently model (2.7). The value of UB in the algorithm is an upper bound on z^* because $(\hat{t}^\ell, \hat{x}^\ell)$ is a feasible solution, and its objective function value in model (2.7) is evaluated in step 7 of the algorithm. The value of LB in the algorithm is a lower bound on z^* because: (i) Proposition 2.5.2 ensures that the cuts (2.23g) are an outer linearizations of $f_{\mathcal{P}}(t, x, y)$; inequality (2.20) in the proof of Proposition 2.5.2 shows that $f_{\mathcal{P}}$

becomes tighter as the partition is refined; and, Proposition 2.5.3 shows that the finest partition yields a model equivalent to model (2.2). Thus, if Algorithm 1 terminates according to step 2 then the current incumbent is an ε-optimal solution.

For a fixed partition, $\mathcal{P}$, Algorithm 1 can only add a finite number of new cuts because each linear program $(S^{\omega}_{\mathcal{P}})$ has a finite number of dual extreme points. After the final iteration in which new cuts are added, partition $\mathcal{P}$ is refined in step 17 of the algorithm. Because there are a finite number of scenarios the finest possible partition, if necessary, will be obtained in a finite number of iterations. From Proposition 2.5.3 we know that with the finest partition the solution to model (2.24) yields an optimal solution, and we will obtain the requisite cuts (2.23g) so that models (2.23) and (2.24) are equivalent in a finite number of iterations. □

2.5.3 Pruning Partitions Using Bound Tightening

Feasibility-based and optimization-based bound tightening schemes have been proved powerful in mixed-integer nonlinear programming to improve computational performance (e.g., Belotti et al. 2012, Coffrin et al. 2015a, Sundar et al. 2018). A feasibility-based bound tightening (FBBT) process is suitable for our problem because the precedence relationships limit the start times of activities. We solve the following linear programs to identify the lower and upper bounds on the first-stage start time of each activity:

$$\min / \max \quad t_i \tag{2.27a}$$

$$\text{s.t.} \quad \text{constraints (2.7b)-(2.7f)} \tag{2.27b}$$

$$\underline{t}_i \leqslant t_i \leqslant \bar{t}_i \qquad \forall i \in I. \tag{2.27c}$$

The bound tightening process starts with a set of initial bounds $\underline{t}_i = 0$ and $\bar{t}_i = T_{\max}$ for each $i \in I$. We solve model (2.27) iteratively and update $\bar{t}_i$ and $\underline{t}_i$ until the bounds converge for every $i \in I$.

We run FBBT at the beginning of our decomposition method to provide the initial partition $\mathcal{P}^0$ to start Algorithm 1. If we can tighten the bounds of some activities, for example, by branch-and-bound or heuristics, we can run FBBT to tighten the bounds for all activities, which leads to a

tighter formulation of the subproblems, ($S^{\omega}_{\mathcal{P}}$), as further detailed in Section 2.5.8.

2.5.4 Obtaining Heuristic Upper Bound

In our computation, we observe that it can take many iterations of Algorithm 1 before we find a good feasible solution. When there is not a tight upper bound to use as a "cutoff value," it takes an integer programming solver longer to solve model ($M_{\mathcal{P}}$). On the other hand, we can quickly solve the extensive formulation (2.2) when the number of scenarios is small; e.g., $|\Omega| = 20$. Solving such a model provides a feasible first-stage solution $(\hat{t}, \hat{x})$, which can be used to generate an upper bound for the problem with the larger, original set of scenarios. Therefore, we can generate N small subsets of scenarios from the original scenario set, i.e., $\Omega_n \subset \Omega,\ n = 1, 2, \ldots, N$, and solve model (2.2) with each of those subsets to generate candidate upper bounds, and select the one with the smallest expected project span under Ω.

In generating Ω_n, we observe that it is beneficial to have diverse scenarios within each subset. As above, we sort the scenarios by disruption time so $H^{\omega} < H^{\omega+1}$. For simplicity, suppose each subset has equal size so that $N \cdot |\Omega_n| = |\Omega|,\ \forall n$. The n-th subset is then:

$$\Omega_n = \{\omega \,|\, \omega = n + (j-1) \cdot |\Omega_n|,\ j = 1, 2, \ldots, N\}.$$

In Section 2.6 we compare the computational time of the decomposition method with and without initial upper bounds, and show the significant performance improvement by including the heuristic upper bound.

2.5.5 Magnanti-Wong Cut Generation

In a Benders' decomposition algorithm, linear programming subproblems can have multiple optimal dual solutions, which means that at a specific incumbent solution, there are multiple valid cuts that could be generated. Magnanti and Wong (1981) provide a method to select Pareto-optimal cuts, which cannot be dominated, and help tighten the LP relaxation of the master program.

We apply a technique that pursues the same goal as Magnanti and Wong in order to tighten cuts at interior points of ($M_{\mathcal{P}}$). We record $(\hat{t}, \hat{x})$ solutions from previous iterations, compute the

corresponding $\hat{y}$ for the current partition, and obtain the average, which we denote (t^0, x^0, y^0). We then solve the subproblems following the procedure of Magnanti and Wong (1981), using this average point as a proxy for the core point, i.e., a point on the relative interior of the LP relaxation of $(M_{\mathcal{P}})$'s feasible region. For each scenario $\omega \in \Omega$, first we solve model (2.16) with the current master solution $(\hat{t}, \hat{x}, \hat{y})$ to obtain the optimal value, $f^{\omega}_{\mathcal{P}}(\hat{t}, \hat{x}, \hat{y})$. Next we need to find dual variables that ensure the dual objective value at $(\hat{t}, \hat{x}, \hat{y})$ is within a small tolerance of $f^{\omega}_{\mathcal{P}}(\hat{t}, \hat{x}, \hat{y})$ (e.g., $\varepsilon_{MW} = 10^{-5}$), while maximizing the dual objective value at (t^0, x^0, y^0). Let the following denote the dual of $(S^{\omega}_{\mathcal{P}})$ in compact form, suppressing dependence on ω:

$$f^{\omega}_{\mathcal{P}}(\hat{t}, \hat{x}, \hat{y}) = \max_{\pi,\lambda,\gamma,\eta} \quad \pi^\top(\hat{t} + b_t) + \lambda^\top(\hat{x} + b_x) + \gamma^\top(\hat{y} + b_y) + \eta^\top b \tag{2.28a}$$

$$\text{s.t.} \quad A^\top_\pi \pi + A^\top_\lambda \lambda + A^\top_\gamma \gamma + A^\top_\eta \eta \leqslant c. \tag{2.28b}$$

We solve the following to obtain Magnanti-Wong cut parameters, v, π, λ, and γ:

$$v = \max_{\pi,\lambda,\gamma,\eta} \quad \pi^\top(t^0 + b_t) + \lambda^\top(x^0 + b_x) + \gamma^\top(y^0 + b_y) + \eta^\top b \tag{2.29a}$$

$$\text{s.t.} \quad \pi^\top(\hat{t} + b_t) + \lambda^\top(\hat{x} + b_x) + \gamma^\top(\hat{y} + b_y) + \eta^\top b \geqslant (1 - \varepsilon_{MW}) f^{\omega}_{\mathcal{P}}(\hat{t}, \hat{x}, \hat{y}) \tag{2.29b}$$

$$A^\top_\pi \pi + A^\top_\lambda \lambda + A^\top_\gamma \gamma + A^\top_\eta \eta \leqslant c. \tag{2.29c}$$

Our average point, (t^0, x^0, y^0), may not be in the relative interior of the master problem's feasible region, e.g., if all solutions contributing to the average have a component of y taking value zero or one. Although this means the resulting cuts may not be Pareto-optimal, they may still improve computational performance, and we investigate this in Section 2.6.

2.5.6 Cut Selection

Algorithm 1 is a multi-cut version of Benders' decomposition procedure (Birge and Louveaux 1988). A multi-cut scheme can converge in fewer iterations, but each iteration is typically more computationally expensive. The latter issue tends to exacerbate as the algorithm proceeds and cuts accumulate, particularly when the master problem is a mixed-integer program.

As we discuss in the next section, when naively keeping all cuts, we see that the time required to solve the master problem can grow quickly as the algorithm proceeds. Therefore, we limit the number of cuts and only keep those that have been tight most recently at the end of each Benders' iteration. Refining a partition can yield many loose cuts. Therefore, keeping only a limited number of cuts eliminates unnecessary constraints, while still providing a valuable lower bound and reducing computational effort.

2.5.7 Refining Partitions

Here, we indicate how a partition is refined in step 17 of Algorithm 1. The most significant contributor to the optimality gap is that variable G can take fractional values in the relaxed problem $(S^{\omega}_{\mathcal{P}})$. For example, if $d^{\omega}_{ik} \gg D_{ik}$ then G_i can take a fractional value far from optimal. This can significantly alter the duration between activity i and k (see constraint (2.8h)) in the LP relaxation, and thus create a large gap between the relaxation and the original mixed-integer subproblem. Let N_p be a parameter that limits the number of new partitions for each activity $i \in I$; we use $N_p = 5$ in our subsequent computation, unless stated otherwise. Suppose subproblem $(S^{\omega}_{\mathcal{P}})$ has as part of its solution $G^{\omega}_i, i \in I$. Then separately for each $i \in I$ our rule selects for partitioning up to N_p scenarios with the largest values of:

$$\rho^{\omega}_i = \begin{cases} \max_{(i,k)\in\mathcal{A}} d^{\omega}_{ik} G^{\omega}_i & \text{if } t_i < H^{\omega} \\ \max_{(i,k)\in\mathcal{A}} d^{\omega}_{ik}(1 - G^{\omega}_i) & \text{if } t_i > H^{\omega} \\ \min\{\max_{(i,k)\in\mathcal{A}} d^{\omega}_{ik} G^{\omega}_i, \max_{(i,k)\in\mathcal{A}} d^{\omega}_{ik}(1 - G^{\omega}_i)\} & \text{if } t_i = H^{\omega}. \end{cases} \tag{2.30}$$

2.5.8 Branch-and-Cut Algorithm

In Algorithm 1 we iteratively refine the partition defining the y-variables, and each time we solve master (2.23) we must solve a mixed-integer linear program, which we do with a commercial solver. Algorithm 1 is not a branch-and-cut (B&C) algorithm in that it does not adaptively generate different cuts at different parts of a branch-and-bound tree. As a potential improvement to Algorithm 1, we propose here a B&C algorithm, which involves a branch-and-bound (B&B) tree

with nodes that we manage. The root node in our B&B tree involves the original partition, and we solve that node using Benders' decomposition, which iteratively adds cuts to master (2.23) until the problem is solved for the fixed partition, as in Algorithm 1. In the B&C algorithm we recursively branch on a continuous variable t_i via $t_i \leqslant H^\omega$ and $t_i \geqslant H^\omega$ for some i-ω pair. Rather than actually branching on the continuous t-variables, this branching is carried out using the partition $(\mathcal{P}_i, \mathcal{Q}_i)$ by fixing the corresponding subset of y_i^q-variables to zero. This helps manage both the number of binary variables and the number of optimality cuts in a master problem. Moreover, in our implementation we solve the nodes in our B&B tree in parallel.

The optimal value of a B&B node provides a lower bound on the optimal value of (2.2). We also continually update a global upper bound each time we obtain a feasible solution in a Benders' decomposition iteration. If the gap between a node's lower bound and the global upper bound is smaller than the tolerance, we mark the node as fathomed. If not, we branch as follows:

$$\text{select } \omega \in \operatorname{argmax}_{\omega\in\Omega} \left[f^\omega(\hat{t}, \hat{x}) - f^\omega_{\mathcal{P}}(\hat{t}, \hat{x}, \hat{y}) \right] \tag{2.31a}$$

$$\text{select } i \in \operatorname{argmax}_{i\in I} \left[\rho_i^\omega \right]. \tag{2.31b}$$

In (2.31a) we select the scenario ω with the largest relaxation gap. Then in (2.31b) we select the activity with the largest ρ_i^ω from equation (2.30). This defines the i-ω pair for branching on $t_i \leqslant H^\omega$ versus $t_i \geqslant H^\omega$ using the y-variables in a form of SOS branching.

After a branch, for each child node we refine the partition on all activities $i \in I$ according to Section 2.5.7. The children inherit the parent node's cuts, updated in a similar fashion as inequality (2.18). To set up notation for Algorithm 2, suppose that for activity $i \in I$ the current B&B node, say node n, has a partition $\mathcal{P}_i^n$ indexed by set $\mathcal{Q}_i^n$, and its parent node m has a partition $\mathcal{P}_i^m$ indexed by set $\mathcal{Q}_i^m$. Then, the cuts inherited from node m are updated for node n as:

$$\begin{aligned}\theta^\omega \geqslant v^{\omega,m,\ell} + \sum_{i\in I} \pi_i^{\omega,m,\ell}(t_i - \hat{t}_i^{m,\ell}) + \sum_{i\in I}\sum_{j\in J_i} \lambda_{ij}^{\omega,m,\ell}(x_{ij} - \hat{x}_{ij}^{m,\ell}) + \\ \sum_{i\in I}\sum_{q\in\mathcal{Q}_i^m} \gamma_i^{\omega,m,q} \left(\sum_{q'\in\Delta_i^n(m,q)} y_i^{q'} - \hat{y}_i^{q,m,\ell} \right) \qquad \forall \ell = 1, 2, \ldots, L.\end{aligned} \tag{2.32}$$

Here, $\Delta_i^n(m,q)$ represents the descendant set of the partition $\mathcal{P}_i^n$ refined from the q-th element in the partition $\mathcal{P}_i^m$. The cuts remain valid by Proposition 2.5.2 because the partitions in the child node refine those of the parent node. The parent node is marked as fathomed after a refinement, and we select the next available node with the smallest lower bound.

Putting all these pieces together, we summarize our partition-based branch-and-cut decomposition method in Algorithm 2. As indicated above, in implementation we execute the algorithm's steps on each available node in parallel.

Algorithm 2 Partition-based branch-and-cut algorithm to solve model (2.2)

Initialize tolerance parameters $\varepsilon \geqslant \delta \geqslant 0$, and a global upper bound UB.
Initialize the B&B tree with node 1, with the following properties: cut iteration number $\ell_1 = 1$, lower bound LB^1, initial partition $\mathcal{P}^1$ with its indexed set $\mathcal{Q}^1$;
while there exists an available node such that $\frac{UB-LB^n}{UB} > \varepsilon$ **do**
    Select available node n with smallest LB^n;
    Append inherited cuts from parent of node n to master $(M_{\mathcal{P}}^n)$ using (2.32);
    repeat
        Solve $(M_{\mathcal{P}}^n)$ and obtain solution $\hat{t}^{\ell^n}, \hat{x}^{\ell^n}, \hat{y}^{\ell^n}, \hat{\theta}^{\ell^n}$ and optimal value $z_{\mathcal{P}}^*$;
        if $z_{\mathcal{P}}^* > LB^n$ **then**
            Update $LB^n = z_{\mathcal{P}}^*$;
        For each $\omega \in \Omega$, solve problem (S^ω) and obtain $f^\omega(\hat{t}^{\ell^n}, \hat{x}^{\ell^n})$ and $\hat{G}^{\ell^n}$;
        Calculate $\bar{z} = p^0 \hat{t}_T^{\ell^n} + \sum_{\omega\in\Omega} p^\omega f^\omega(\hat{t}^{\ell^n}, \hat{x}^{\ell^n})$;
        if $\bar{z} < UB$ **then**
            Update $UB = \bar{z}$ and incumbent solution as $t^* = \hat{t}^{\ell^n}, x^* = \hat{x}^{\ell^n}$ and $G^* = \hat{G}^{\ell^n}$;
        for each $\omega \in \Omega$ **do**
            solve $(S_{\mathcal{P}}^\omega)$ given $\hat{t}^{\ell^n}, \hat{x}^{\ell^n}, \hat{y}^{\ell^n}$ and obtain optimal value v^{ω,ℓ^n} and $\pi^{\omega,\ell^n}, \lambda^{\omega,\ell^n}, \gamma^{\omega,\ell^n}$;
            if $\hat{\theta}^{\omega,\ell} < v^{\omega,\ell} - \delta$ **then** add cut of form (2.17) to $(M_{\mathcal{P}}^n)$;
        Let $\ell^n = \ell^n + 1$;
    until no cut is added;
    if $\frac{UB-LB^n}{UB} > \varepsilon$ **then**
        Branch via (2.31), creating two available children nodes, n_1 and n_2, from node n;
        Refine the partition for nodes n_1 and n_2 to obtain $\mathcal{P}^{n_1}$ and $\mathcal{P}^{n_2}$, respectively;
        Let $LB^{n_1} = LB^{n_2} = LB^n$;
    Mark node n as fathomed;
end while
Output UB as the ε-optimal value of model (2.2), and (t^*, x^*, G^*) as the ε-optimal solution.

2.5.9 Algorithm 2: Numerical Example

We illustrate Algorithm 2 on an example with $\varepsilon = 0$. We consider the network with five activities that accompany the source and terminal, as shown in Figure 2.6.

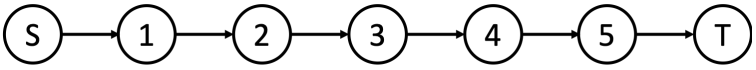


Figure 2.6: A five-activity serial network to illustrate Algorithm 2

Each of the five activities has unit duration, and can be crashed with a single option that decreases the duration by 90% with one unit of resource consumption, i.e., $e_{i1} = 0.9,\ \forall i \in I\backslash\{S,T\}$, and we let $B = 2$. The probability of no disruption is $p^0 = 0.2$, and the disruption can occur at four discrete time points, $H^1 = 1,\ H^2 = 2,\ H^3 = 3,\ H^4 = 4$, with equal probability $p^\omega = 0.2,\ \forall \omega \in \Omega = \{1,2,3,4\}$. The increase in duration under a disruption is $d_i^\omega = 10$ for each activity-scenario pair. By Proposition 2.5.1 we can bound the start time of each activity from above by $T_{\max} = 9$.

Running the FBBT procedure of Section 2.5.3 we obtain $\underline{t}_1 = 0$, $\underline{t}_2 = 0.1$, $\underline{t}_3 = 0.2$, $\underline{t}_4 = 1.2$, $\underline{t}_5 = 2.2$, and $\underline{t}_T = 3.2$. Given these values, and the realizations of H^ω, we initialize node 1 of Algorithm 2 with the following partition, which precludes certain intervals for activities 4, 5, and T:

$$
\begin{aligned}
&\mathcal{P}_1^1 = \{[0,9]\} && \mathcal{Q}_1^1 = \{1\} && \\
&\mathcal{P}_2^1 = \{[0,9]\} && \mathcal{Q}_2^1 = \{1\} && \\
&\mathcal{P}_3^1 = \{[0,9]\} && \mathcal{Q}_3^1 = \{1\} && \\
&\mathcal{P}_4^1 = \{[0,1],[1,9]\} && \mathcal{Q}_4^1 = \{1,2\} && y_4^1 = 0 \\
&\mathcal{P}_5^1 = \{[0,2],[2,9]\} && \mathcal{Q}_5^1 = \{1,2\} && y_5^1 = 0 \\
&\mathcal{P}_T^1 = \{[0,3],[3,9]\} && \mathcal{Q}_T^1 = \{1,2\} && y_T^1 = 0.
\end{aligned}
$$

Executing steps 6-18 of the algorithm, we solve node 1 with Benders' decomposition, converging to an optimal value of $z_{\mathcal{P}}^* = 4.072$, and in the process, we obtain an upper bound of $UB = 6.607$.

The optimal solution is:

$$\hat{t}_1 = 0 \qquad \hat{x}_{11} = 0$$
$$\hat{t}_2 = 1 \qquad \hat{x}_{21} = 0.1124$$
$$\hat{t}_3 = 1.899 \qquad \hat{x}_{31} = 0.1124$$
$$\hat{t}_4 = 2.798 \qquad \hat{x}_{41} = 0.8892$$
$$\hat{t}_5 = 2.997 \qquad \hat{x}_{51} = 0.8860$$
$$\hat{t}_T = 3.2.$$

At this solution, by equation (2.31a) the largest relaxation gap is incurred at $\omega = 1$, and equation (2.31b) corresponds to activity 3, which has a fractional solution of $G_3 = 0.125$ in subproblem $(S^1_{\mathcal{P}})$. We branch on i-ω pair 3-1, creating two children, node 2 and node 3. Node 2 has an additional constraint, $t_3 \geqslant H^1 = 1$ and node 3 has an additional constraint $t_3 \leqslant H^1 = 1$. For each activity i, we refine the partition by selecting the scenario with the largest nonzero ρ_i^ω for $\omega \in \Omega$, i.e., for simplicity we use $N_p = 1$; see Section 2.5.7. If the largest ρ_i^ω is zero for activity i, we do not refine that activity's partition. We again apply the ideas of Section 2.5.3, and solve a series of linear programs (2.27) to tighten the bounds of starting times for nodes 2 and 3, accounting for their respective additional constraints $t_3 \geqslant 1$ and $t_3 \leqslant 1$. This refinement and bound-tightening process yields:

Node 2:

$$\mathcal{P}_1^2 = \{[0,9]\} \qquad \mathcal{Q}_1^2 = \{1\}$$
$$\mathcal{P}_2^2 = \{[0,9]\} \qquad \mathcal{Q}_2^2 = \{1\}$$
$$\mathcal{P}_3^2 = \{[0,1],[1,9]\} \qquad \mathcal{Q}_3^2 = \{1,2\} \qquad y_3^1 = 0$$
$$\mathcal{P}_4^2 = \{[0,1],[1,2],[2,9]\} \qquad \mathcal{Q}_4^2 = \{1,2,3\} \qquad y_4^1 = 0$$
$$\mathcal{P}_5^2 = \{[0,2],[2,3],[3,9]\} \qquad \mathcal{Q}_5^2 = \{1,2,3\} \qquad y_5^1 = 0$$
$$\mathcal{P}_T^2 = \{[0,3],[3,4],[4,9]\} \qquad \mathcal{Q}_T^2 = \{1,2,3\} \qquad y_T^1 = 0$$

Node 3:

$$\mathcal{P}_1^3 = \{[0,1],[1,9]\} \qquad \mathcal{Q}_1^3 = \{1,2\} \qquad y_1^2 = 0$$
$$\mathcal{P}_2^3 = \{[0,1],[1,9]\} \qquad \mathcal{Q}_2^3 = \{1,2\} \qquad y_2^2 = 0$$
$$\mathcal{P}_3^3 = \{[0,1],[1,9]\} \qquad \mathcal{Q}_3^3 = \{1,2\} \qquad y_3^2 = 0$$
$$\mathcal{P}_4^3 = \{[0,1],[1,2],[2,9]\} \qquad \mathcal{Q}_4^3 = \{1,2,3\} \qquad y_4^1 = 0$$
$$\mathcal{P}_5^3 = \{[0,2],[2,3],[3,9]\} \qquad \mathcal{Q}_5^3 = \{1,2,3\} \qquad y_5^1 = 0$$
$$\mathcal{P}_T^3 = \{[0,3],[3,4],[4,9]\} \qquad \mathcal{Q}_T^3 = \{1,2,3\} \qquad y_T^1 = 0.$$

We fathom node 1 and nodes 2 and 3 inherit its cuts and lower bound. Breaking the tie arbitrarily, we select and then solve node 2 with Benders' decomposition and obtain an optimal value of $z_{\mathcal{P}}^* = 6.111$. Node 2 then branches to nodes 4 and 5, which inherit the cuts and the lower bound from node 2. The upper bound value remains $UB = 6.607$.

Node 3 now has the smallest lower bound, 4.072, among unfathomed nodes, and yields an optimal value of $z_{\mathcal{P}}^* = 6$ and optimal solution:

$$\hat{t}_1 = 0 \qquad \hat{x}_{11} = 1 \tag{2.33a}$$
$$\hat{t}_2 = 0.1 \qquad \hat{x}_{21} = \frac{1}{9} \tag{2.33b}$$
$$\hat{t}_3 = 1 \qquad \hat{x}_{31} = 0 \tag{2.33c}$$
$$\hat{t}_4 = 2 \qquad \hat{x}_{41} = 0 \tag{2.33d}$$
$$\hat{t}_5 = 3 \qquad \hat{x}_{51} = \frac{8}{9} \tag{2.33e}$$
$$\hat{t}_T = 3.2. \tag{2.33f}$$

In the process of solving node 3, we also obtain a tighter global upper bound of $UB = 6$ associated with solution (2.33), and hence the optimality gap at node 3 is zero. Nodes 4 and 5 can now be fathomed because their lower bound of 6.111 exceeds UB. The algorithm terminates with the optimal solution in equation (2.33), and the corresponding G-variables indicate that activities 1, 2, and 3 start before disruption scenario 1, activity 4 starts before disruption scenario

2, and activity 5 starts before disruption scenario 3. The optimal objective function value is $6 = 0.6 \cdot [3.2] + 0.2 \cdot [2 + 11 \cdot (1 - 0.9 \cdot 8/9) + 11] + 0.2 \cdot [3 + 11 \cdot (1 - 0.9 \cdot 8/9)]$.

2.6 Experimental Results

In this section, we address the following questions with our computational results:

1. What is the value of model (2.2), which takes account of randomness in both the timing and magnitude of a disruption? In other words, how does the quality of the solution to model (2.2) compare to those of simpler alternatives?
2. How does the solution quality improve as the number of samples grows in a sample average approximation?
3. How do Algorithms 1 and 2 perform versus solving the extensive formulation (2.2) using a state-of-the-art MIP solver? How effective are the computational enhancements of upper-bound generation, Magnanti-Wong cuts, and the cut-selection procedure from Sections 2.5.4-2.5.6?

Section 2.6.1 introduces the PERT networks and probability distributions characterizing the disruptions for our test cases. In Section 2.6.2 we construct deterministic and semi-deterministic alternatives to model (2.2), perform out-of-sample tests, and compare the quality of the resulting solutions to those of model (2.2). We test how the sample size affects solution quality and requisite computational effort in Section 2.6.3. Finally, in Section 2.6.4 we compare the performance of Algorithms 1 and 2 to solving extensive formulation directly.

All tests are run on a server with 30 Intel Xeon cores at 3.1 GHz and 256 GB of RAM. For Algorithm 2, each node is solved by 6 cores and we allow at most 5 nodes to be solved simultaneously so that the maximum number of cores used at any time is again 30. All models are constructed using version 0.18.0 of the JuMP package (Dunning et al. 2017) on the Julia platform. All linear programs and mixed-integer programs are solved by Gurobi 8.01 (Gurobi Optimization, Inc. 2016) with the integer feasibility tolerance and the primal feasibility tolerance both set to 10^{-8}. In addition, we solve all problems using an optimality-gap tolerance of 10^{-2}, including Algorithm 2's $\varepsilon = 10^{-2}$ as well as $\delta = 10^{-4}$.

2.6.1 Test Cases Construction

We construct our test cases based on two activity networks from the literature, along with one we create from "scratch" and one we generate randomly. In particular, we use an activity network from Plambeck et al. (1996) with 11 activities, and one from Elmaghraby (1977) with 19 activities. We also manually create one activity network with 14 activities and randomly generate one network with 35 activities using the tool *RanGen* (Demeulemeester et al. 2003). In the following section, we use "Case X" to denote the test case with X activities. Data for all four test cases are detailed in Appendix A.1.

For each test case, the timing of the disruption is a discrete random variable sampled from a lognormal distribution, which is commonly used to model failure times (e.g., Crow and Shimizu 1987, Mullen 1998). The magnitude of the disruption for each activity follows an exponential distribution (which we again sample) whose parameter varies among the activities; the exponential distribution is widely used to model activity durations such as service times (e.g., Ross et al. 1996).

2.6.2 Value of a Fully Stochastic Model

We compare the quality of five solutions, one from solving model (2.2) and four from solving simpler alternatives, to investigate the value of modeling both the random timing and random magnitude of a disruption. First, we can obtain a solution by solving a deterministic model (2.1), assuming no disruption occurs, i.e., $d^\omega = 0$ (denoted "DET"). Three semi-stochastic alternative models can be solved assuming: both the timing and magnitude of the disruption are deterministic at their expected values (denoted "EXP"); the timing is random but magnitude is fixed at its expected value (denoted "HOnly"); and, the magnitude is random but timing is fixed at its expected value (denoted "dOnly"). Finally, we construct the fully stochastic model (2.2) in which both the timing and magnitude are random (denoted "FULL"). We sample 500 scenarios to solve HOnly, dOnly, and FULL. Twenty batches of samples of size 5,000 are used to estimate an upper bound for each candidate solution. The upper bound point estimate for those five candidate solutions is shown in Figure 2.7, and the 95% confidence interval is shown in Table 2.2.

Figure 2.7 scales the optimal value of each test problem, dividing by that of FULL, and

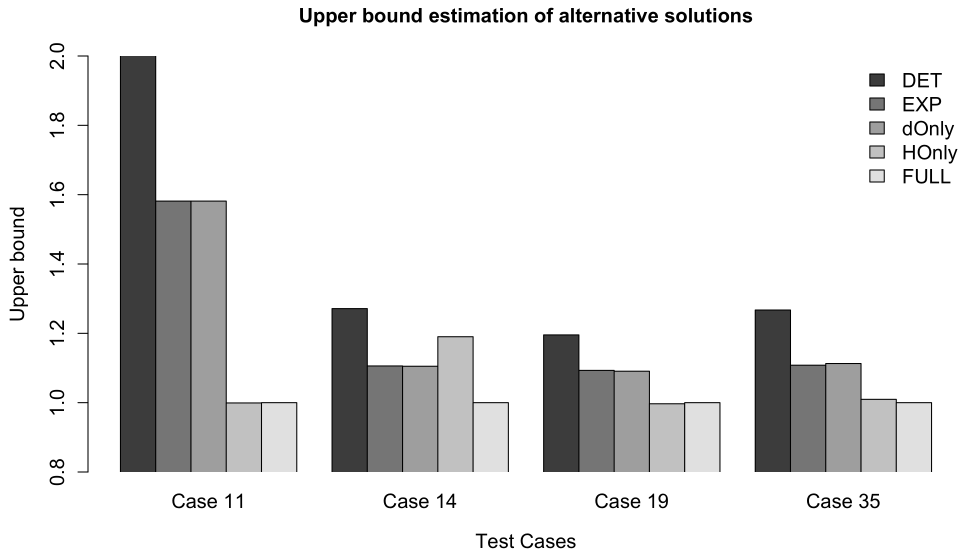


Figure 2.7: Comparison of quality of alternative solutions to the problem (2.2)

shows that the solution quality can be poor without considering the uncertainty in the timing of a disruption. DET's upper bound estimate is at least 20% larger than that of FULL for all test cases. The upper bound estimates of dOnly and EXP are similar and at least 10% larger than that of FULL. For Case 14, the upper bound of HOnly is 20% larger than that of FULL, and is closer to that of FULL for the other three cases. That said, given the sample sizes, the computational effort to solve HOnly and FULL are comparable. These test results indicate that the fully stochastic model can outperform simpler variants.

	DET	EXP	dOnly	HOnly	FULL
Case 11	575.80 ± 16.88	454.33 ± 18.34	454.33 ± 18.34	287.05 ± 14.91	287.29 ± 14.92
Case 14	3309.27 ± 177.18	2878.65 ± 176.70	2876.42 ± 176.55	3098.05 ± 133.06	2602.97 ± 110.12
Case 19	426.09 ± 10.09	389.61 ± 7.88	388.77 ± 8.43	355.29 ± 7.14	356.42 ± 5.86
Case 35	1353.80 ± 21.11	1183.74 ± 18.52	1188.93 ± 19.53	1078.57 ± 17.31	1068.29 ± 17.47

Table 2.2: Compare optimal values from alternatives of the disruption model

2.6.3 Simulation Budget

We examine the quality of solutions obtained by solving SAA problems with different sample sizes. In expectation, the optimal value of an SAA problem provides a lower bound on z^*, the optimal value of the population problem, and in expectation evaluating a sample-mean objective function of an SAA solution using an independent sample provides an upper bound on z^* (Mak et al. 1999). Both estimates approach z^* as the sample size grows large (e.g., Shapiro et al. 2009). Of course, the computational effort to solve the corresponding SAA problem grows with the sample size. Understanding how the solution quality improves with sample size helps us obtain a high-quality solution without excessive computational effort. For each test case, we first solve 20 SAA instances of model (2.2), with sample sizes of 10, 20, 50, 100, 200, and 500. For each solution, we obtain a point estimate of the upper bound using the same 5,000 samples. We also obtain 95% confidence intervals associated with both upper- and lower-bound estimators for each sample size. We present the upper bound results in Table 2.3, the lower bound results in Table 2.4, and a visualization of the estimators in Figure 2.8.

	10	20	50	100	200	500
Case 11	326.05 ± 80.59	293.43 ± 24.20	290.95 ± 28.51	283.06 ± 12.24	280.10 ± 6.59	278.33 ± 2.30
Case 14	3002.42 ± 537.69	2942.09 ± 778.43	2714.87 ± 330.61	2645.71 ± 100.15	2627.94 ± 40.46	2626.90 ± 39.51
Case 19	385.41 ± 42.29	383.67 ± 45.78	358.99 ± 16.76	355.28 ± 10.21	352.57 ± 7.65	350.91 ± 1.83
Case 35	1130.18 ± 117.51	1102.57 ± 45.66	1091.34 ± 33.00	1080.23 ± 21.86	1073.05 ± 10.53	1070.08 ± 5.20

Table 2.3: Upper bound point estimates, and 95% confidence intervals, for SAA solutions with different sample sizes

	10	20	50	100	200	500
Case 11	296.44 ± 255.34	247.24 ± 134.57	266.30 ± 140.43	275.33 ± 73.14	289.32 ± 66.97	274.79 ± 36.59
Case 14	1868.25 ± 2152.85	1991.32 ± 1273.90	2244.39 ± 992.31	2310.86 ± 688.76	2508.60 ± 456.24	2477.13 ± 275.11
Case 19	310.07 ± 123.11	324.00 ± 85.84	331.84 ± 53.49	339.69 ± 36.10	344.84 ± 21.78	345.10 ± 14.26
Case 35	1166.85 ± 586.64	1107.83 ± 282.44	1055.84 ± 202.07	1036.82 ± 120.40	1060.78 ± 82.29	1065.76 ± 54.10

Table 2.4: Lower bound point estimates, and 95% confidence intervals, for SAA solutions with different sample sizes

From the two tables and Figure 2.8, we can see the gap between the upper- and lower-bound estimators shrinks, and both estimators become less variable as the sample size grows. Improvements

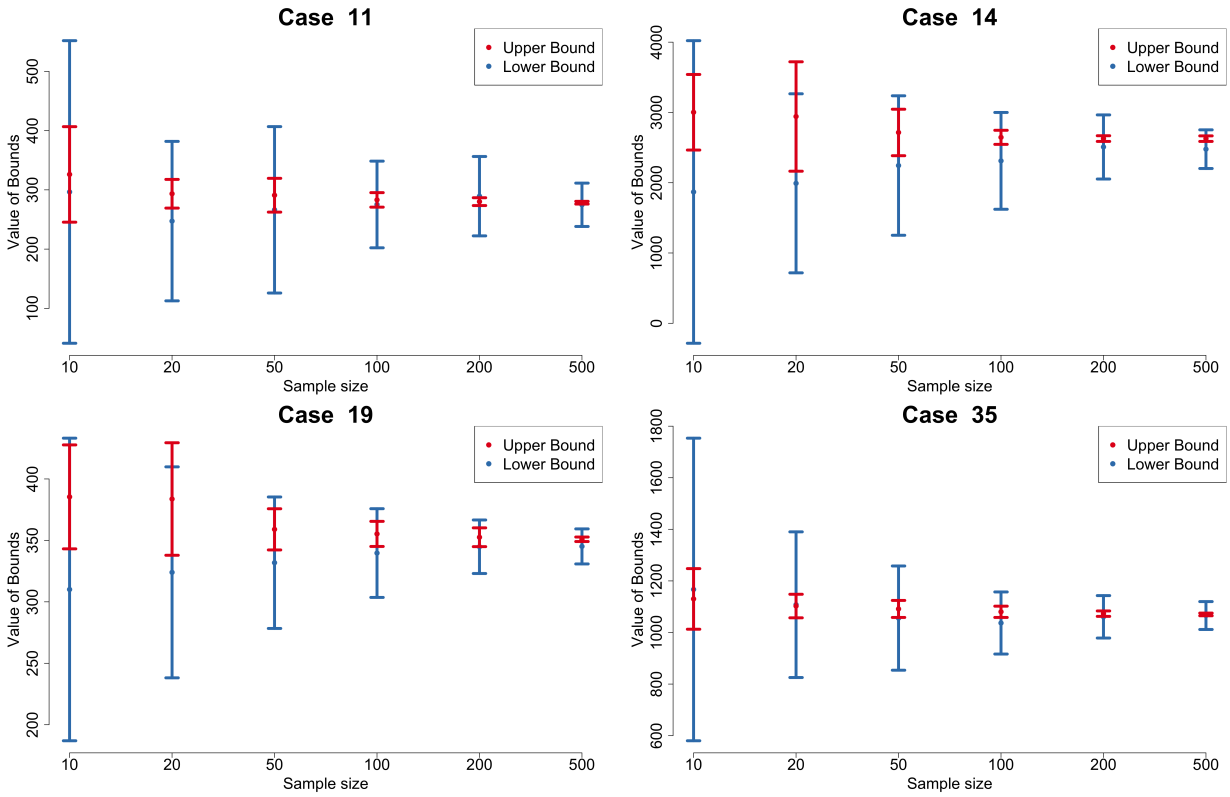


Figure 2.8: Confidence intervals, and point estimates, of the lower upper bounds for different sample sizes

in the upper-bound estimators are small for sample sizes that exceed 100, and while improvements in lower-bound point estimates can appear larger (see Case 14), they are small relative to sampling error for such sample sizes.

2.6.4 Computational Performance

In this section we discuss the computational performance of our decomposition method with its various potential enhancements. As a benchmark, we compare the performance of Algorithms 1 and 2 to direct solution of the extensive formulation (2.2) using a commercial solver.

We first briefly comment on running Algorithm 2, with and without FBBT from Section 2.5.3. Using a sample size of 500, we solve instances of Cases 11, 14, 19, and 35. In so doing, we use all improvements described in Sections 2.5.4-2.5.6. FBBT can significantly improve the value of the LP relaxation at the B&B tree's root node, and this sometimes leads to modest improvements in

the total number of nodes that are explored. That said, the differences in overall run-times, with and without FBBT, are mixed and not particularly large, and hence we do not present these results in detail. We do employ FBBT in the remainder of this section.

Next, we show the quality of the heuristic upper bound described in Section 2.5.4. We solve SAA instances with sample sizes of 100, 200, and 500, and in the notation of Section 2.5.4 we use $N = 10$, 10, and 20 in these three instances, respectively. We compare the heuristic upper bound to the optimal value of the SAA instance. Moreover, for each case/sample-size pair, we replicate the procedure on 20 independent instances. These are the same 20 instances used for sample sizes of 100, 200, and 500 in Section 2.6.3 for which Table 2.4 reports average optimal values. Table 2.5 shows the smallest gap, the average gap, and the largest gap, all as percentages, between the heuristic upper bound and the SAA optimal value across the 20 replications. The table also shows the average time required to compute the heuristic upper bound.

	Sample size	Gap (%)			Average
		Smallest	Average	Largest	time (sec)
Case 11	100	0.01	0.5	2.3	5.6
	200	0	0.15	0.83	17.8
	500	0	0.07	0.37	67.5
Case 14	100	1.34	6.02	16.14	7.7
	200	1.58	3.9	8.91	30.4
	500	0.75	2.35	3.96	104.4
Case 19	100	0.56	2.67	5.28	9.9
	200	0.05	0.99	2.03	37.2
	500	0	0.72	1.59	127.6
Case 35	100	0	0.64	2.65	5.8
	200	0.08	0.46	1.3	25.7
	500	0	0.32	1.25	59.0

Table 2.5: Gap information between the heuristic upper bound and optimal value for twenty random samples

We see that the average upper bound gap exceeds 5% only for Case 14 when the sample size is 100. Among all 240 SAA instances, there are only 15 for which the heuristic upper bound exceeds the optimal value by more than 5%. We also observe that as the sample size grows, the heuristic upper bound's quality improves.

Table 2.6 shows run-time results for finding a heuristic upper bound and generating Magnanti-

Wong cuts while using Algorithm 2 with cut selection. Here we present the running time of a single replication for the sample size of 500, assuming $N = 20$ in the notation of Section 2.5.4. In the table, "UB" means that we only use the heuristic upper bound of Section 2.5.4 and generate regular Benders' cuts, and "MW" means that we generate Magnanti-Wong cuts of Section 2.5.5 without the benefit of the upper bound heuristic. We also show the computation time of using neither and using both techniques.

Running Time (sec.)	Case 11	Case 14	Case 19	Case 35
Neither	70.1	770.0	1118.2	140.1
UB	62.9	534.7	861.0	20.0
MW	141.6	1218.2	952.0	222.3
Both	114.1	795.6	607.5	20.0

Table 2.6: Run-time results for using the heuristic upper bound and Magnanti-Wong cuts in Algorithm 2 with cut selection

Table 2.6 shows that obtaining a heuristic upper bound can reduce computational effort; this occurs because the bound facilitates earlier fathoming of some nodes in the branch-and-bound tree when solving the master MIP. This outweighs the time required to compute the upper bound; see Table 2.5. Magnanti-Wong cuts are effective for Case 19 but not for the other three cases, and the two techniques can have a synergistic effect (Case 19).

Table 2.7 compares the run time of Algorithm 1 (denoted A1), Algorithm 2 (denoted A2), with and without the cut selection procedure (CS). Here we use all improvements described in Section 2.5.3-2.5.5. The table also shows the run-time for solving the extensive formulation (2.2) using Gurobi, again to a relative optimality tolerance of 0.01. Due to the computational effort required to execute some of the less efficient algorithms on large problem instances, we report results for a single replication.

Table 2.7 shows that although directly solving the extensive formulation may be faster when the sample size is small, our decomposition algorithms tend to perform better as the sample size grows; the table's one exception is Algorithm 1 for Case 14. The improved computational performance occurs because, while our decomposition methods must solve the master problem multiple times,

Running Time (sec.)	Sample Size ($\|\Omega\|$)	Methods				
		A1	A1+CS	A2	A2+CS	Extensive
Case 11	100	8.7	7.4	20.0	10.0	5.0
	200	16.7	34.8	20.0	20.0	40.3
	500	273.3	240.9	103.4	114.1	1597.1
	1000	279.9	326.7	73.0	80.1	5019.0
Case 14	100	45.2	31.0	43.7	40.0	13.7
	200	137.5	126.0	150.2	101.0	158.7
	500	1710.7	1578.0	1252.0	795.6	1263.1
	1000	6284.1	7416.6	3135.1	1684.2	9808.3
Case 19	100	201.2	107.4	80.4	60.1	58.4
	200	239.9	194.2	400.6	187.1	279.8
	500	1643.7	1421.0	747.2	607.5	2792.9
	1000	11259.1	8824.9	9785.2	2562.8	33698.5
Case 35	100	33.4	31.7	40.1	30.0	7.0
	200	14.1	13.6	10.0	10.0	35.9
	500	66.1	66.2	20.0	20.0	358.6
	1000	241.0	248.0	40.1	40.1	662.0

Table 2.7: Computational performance with different sample sizes for decomposition methods and directly solving the extensive formulation using Gurobi

the number of binary variables is significantly smaller than that of the extensive formulation.

Algorithm 2 outperforms Algorithm 1 when the sample size exceeds 500. When the sample size is large, Algorithm 1 requires many master iterations to converge, and solving the mixed-integer master program becomes significantly harder. For Algorithm 2, the time saved by processing nodes in parallel outweighs the slightly longer solution time at each node due to using fewer cores. Moreover, because of branching, each master problem at a node in the B&B tree has a small number of binary variables in Algorithm 2, which further accelerates the B&C algorithm. The results also show that selecting cuts can significantly improve the decomposition methods. On the most difficult instances, applying the cut selection scheme yields larger improvements for Algorithm 2 than for Algorithm 1.

2.7 Conclusions

In this chapter, we introduce the concept of a stochastic disruption in the context of a project crashing problem. We consider the case of a single disruption, and formulate the model as a two-

stage stochastic mixed-integer program in which the timing of the stage, i.e., the disruption time, is random. We use examples to illustrate properties of our problem that deviate from its deterministic counterpart, including the fact that it can be optimal to delay the start of an activity or crash a shorter-duration activity, even under proportional reduction. While, conceptually, the underlying problem involves continuous decision variables, we argue that the problem is NP-hard. In our two-stage stochastic mixed-integer program, second-stage binary variables capture the logic of the start time of an activity, relative to the disruption time. The resulting model is computationally challenging, but we propose a decomposition method which exploits the logical temporal relationship just mentioned, and sequentially partitions the feasible region of continuous first-stage decision variables to generate tighter cuts in a Benders' decomposition algorithm. The proposed method can significantly improve computational performance, especially as sample sizes grow large.

The ideas in this chapter can be extended in multiple ways. There may be opportunities to exploit network structure in tailoring the branch-and-cut algorithm that we have developed. The distribution governing the disruption time and, conditional on that time, the magnitude of the disruption may facilitate sampling strategies that reduce variance and improve solution quality. We have considered a model and algorithm that allow for at most one disruption, but handling a small number of disruptions could be attractive.

Chapter 3

Robust Optimization for Electricity Generation

3.1 Introduction

The alternating current optimal power flow (ACOPF) problem has been a topic of interest in the academic literature since the 1960s (Carpentier 1962). The ACOPF problem is used to determine the output for all generators and establish the system's *configuration*, i.e., the voltage and phase angle at each bus and resulting power flows on lines. The goal is usually to minimize the generation cost and keep the system configuration within a stable range; see, for example, Bienstock (2015) for a detailed discussion. While the ACOPF problem can be formulated as a quadratically constrained quadratic program, realistic instances are challenging to solve within time limits commensurate with an operational schedule (usually a few minutes) because of their scale and nonconvexities (Lavaei and Low 2012, Low 2014a, Verma 2010). Linearizing the power flow equations simplifies the nonconvex ACOPF problem to what the literature calls a DCOPF approximation, which is a linear program, and this approximation is frequently applied. However, optimality and feasibility of the solution to the original ACOPF problem cannot be guaranteed because the voltage at each bus is assumed to be fixed and reactive power is ignored; see, e.g., Momoh et al. (1999) and Stott et al. (2009) for reviews of such linear approximations. In recent years,

increasing attention has been paid to convex relaxations of ACOPF problems. Relaxations rooted in semidefinite programming and second-order cone programming have been used to approximate the ACOPF problem (Bai and Wei 2011, Bai et al. 2008, Coffrin et al. 2016, Jabr 2006, Kocuk et al. 2016, Lavaei and Low 2012, Low 2014b), and under some circumstances, these relaxed solutions recover the exact optimal solution of the original nonconvex ACOPF problem.

Electric power systems operate under significant uncertainty due to system load, failure of generation and transmission assets, and uncertain generation from renewable energy sources (RESs) including wind and solar resources. In this context, we seek an economic dispatch decision that is robust to uncertainty in load and RES generation. With stochastic realizations of power from wind farms, Phan and Ghosh (2014) model economic dispatch under ACOPF as a two-stage stochastic program, and use a sample average approximation. Monticelli et al. (1987) introduce a security-constrained variant of an economic dispatch model in which the goal is to obtain a solution that can adapt to failure of a subset of system components explicitly modeled through a set of contingencies. Instead of enforcing feasibility for all modeled contingencies, Lubin et al. (2016) formulate a chance-constrained model that ensures feasibility with high probability. Robust optimization is a natural modeling framework for security-constrained problems in that such models yield solutions that can handle any contingency within a specified uncertainty set. Jabr (2013) and Louca and Bitar (2017) propose an adaptive robust optimization model, in which recourse decisions are represented as an affine function of realizations of uncertainty such as available power from RESs. Attarha et al. (2018) propose a tri-level decomposition algorithm where in the second level a DCOPF relaxation is solved to obtain worst-case scenarios, which are further used to construct a large-scale extensive formulation of the robust ACOPF problem.

Although significant progress has been made both in convex relaxations of nonconvex ACOPF problems and in modeling dispatch under uncertainty, there is much less work that combines these two threads; i.e., most stochastic or robust models for economic dispatch use the linear DCOPF approximation. Liu and Ferris (2015) solve a scenario-based security-constrained ACOPF problem, where for each contingency the ACOPF is relaxed as a semidefinite program (SDP). Lorca and Sun (2018) model a multi-period two-stage robust ACOPF problem using a conic relaxation, which is

similar to our approach, but we focus more on the feasibility guarantee and the properties of our robust solution.

We solve a robust convex approximation, without specifying scenarios but by constructing an uncertainty set, to simultaneously reap the benefit of a tighter relaxation and include uncertainty in our model. We assume an uncontrollable injection represents net load at each bus. Here, net load captures demand and RES generation, which are subject to simple bounds and further constraints that define the uncertainty set. The goal is to find a robust and economical energy generation plan. The robustness here means that for all contingencies modeled by our uncertainty set, we can find a system configuration that satisfies the system's physical and operational constraints. We call such a plan, a robust optimal solution to the ACOPF problem.

Our formulation is unique in that, in addition to using a convex relaxation of the ACOPF problem rather than a DC approximation, we employ a "full recourse" solution rather than relying on simpler approximations like linear decision rules. There are three possible outcomes from solving our model. First, the solution to the convex approximation may be feasible to the robust nonconvex ACOPF problem, which means we exactly recover a robust solution. Second, due to the convex relaxation, the solution we obtain may not be feasible to the robust nonconvex ACOPF problem, but we obtain a lower bound on the optimal cost of the nonconvex counterpart, which yields a bound on the optimality gap when coupled with a heuristically obtained feasible solution. Third, if the convex relaxation is infeasible, we identify infeasibility of the robust nonconvex ACOPF problem.

In Section 3.2, we formulate our convex relaxation of the ACOPF problem. A cutting-plane method is proposed in Section 3.3, and the proof of its convergence is detailed. Experimental results are reported in Section 3.4, and conclusions are drawn in Section 3.5.

3.2 Problem Formulation

In this section we formulate the robust nonconvex ACOPF problem and its convex relaxation. We index the set of buses in the power system by $\mathcal{N}$, and the set of lines by $\mathcal{A}$. The set of controllable generators is denoted by $\mathcal{G}$, and the subset of generators connected to bus i is indexed by $\mathcal{G}_i$. Each

controllable generator $g \in \mathcal{G}$ injects active power s_g^p and reactive power s_g^q at a bus i if $g \in \mathcal{G}_i$. Each bus $i \in \mathcal{N}$ has an uncontrollable injection, which may be negative, consisting of the uncertain net load due to actual demand and RES generation at that bus. At bus $i \in \mathcal{N}$, the uncontrollable active power injected, u_i^p, is bounded within an uncertainty set $[\underline{u}_i^p, \bar{u}_i^p]$, where $\underline{u}_i^p \leqslant \bar{u}_i^p,\ \forall i \in \mathcal{N}$. The uncontrollable reactive power is bounded in a similar way, where $u_i^q \in [\underline{u}_i^q, \bar{u}_i^q],\ \forall i \in \mathcal{N}$.

In addition to simple bounds, we introduce a "budget constraint" in our uncertainty set, which limits the magnitude of deviation from a nominal injection, summed across all buses. Such budget-constrained uncertainty sets have been widely applied in robust optimization, starting with Bertsimas and Sim (2003, 2004). Here we denote the nominal uncontrollable active and reactive power injection as $u^{p,0}$ and $u^{q,0}$, which are both vectors with $|\mathcal{N}|$ components and satisfy $(\underline{u}_i^p, \underline{u}_i^q) \leqslant (u^{p,0}, u^{q,0}) \leqslant (\bar{u}_i^p, \bar{u}_i^q)$, for $i \in \mathcal{N}$. We cluster the set of buses $\mathcal{N}$ into $|\mathcal{M}|$ subgroups, denoted by $\mathcal{N}_m,\ m \in \mathcal{M}$, by solving a facility location problem detailed in Appendix B.1. There has been much research about the geographical correlation of renewable generations and loads in power systems (Xie and Ahmed 2018, Fang et al. 2018, Malvaldi et al. 2017, Klima and Apt 2015, Lohmann et al. 2016, Bernstein et al. 2014), which justifies our choice of clustering. We assume within each cluster the relative magnitude of deviation is the same for both the active power and reactive power at every bus. We define the uncertainty set with the following constraints:

$$0 \leqslant u_i^{p,+} \leqslant \bar{u}_i^p - u_i^{p,0} \qquad 0 \leqslant u_i^{p,-} \leqslant u_i^{p,0} - \underline{u}_i^p \qquad \forall i \in \mathcal{N} \tag{3.1a}$$

$$0 \leqslant u_i^{q,+} \leqslant \bar{u}_i^q - u_i^{q,0} \qquad 0 \leqslant u_i^{q,-} \leqslant u_i^{q,0} - \underline{u}_i^q \qquad \forall i \in \mathcal{N} \tag{3.1b}$$

$$\frac{u_i^{p,+}}{\bar{u}_i^p - u_i^{p,0}} = \frac{u_i^{q,+}}{\bar{u}_i^q - u_i^{q,0}} = u_m^+ \qquad \frac{u_i^{p,-}}{u_i^{p,0} - \underline{u}_i^p} = \frac{u_i^{q,-}}{u_i^{q,0} - \underline{u}_i^q} = u_m^- \qquad \forall m \in \mathcal{M}, i \in \mathcal{N}_m \tag{3.1c}$$

$$\mathcal{U} = \left\{ (u^{p,+}, u^{p,-}, u^{q,+}, u^{q,-}) \in \Re^{4|\mathcal{N}|} \;\middle|\; \begin{array}{l} \text{(3.1a)-(3.1b) and } \exists\, u_m^+, u_m^-, m \in \mathcal{M}, \\ \text{satisfying (3.1c) and } \sum_{m \in \mathcal{M}} \left(u_m^+ + u_m^-\right) \leqslant \Gamma \end{array} \right\}. \tag{3.2}$$

Budget parameter Γ controls the deviation from nominal values, summed across all buses. We can substitute out variables $u_m^+, u_m^-, m \in \mathcal{M}$, and we assume this has been done when referencing $\mathcal{U}$ in

what follows.

In most of the power systems literature, lines are assumed to be undirected, and an orientation indicates the direction of flow. We represent multiple lines by a triple (i, j, n), which uses the orientation to indicate that positive flow is from i to j on the n-th line between these two buses and negative flow is the opposite. Each bus has a voltage, v_i, and a phase angle, θ_i. These configurations, along with the line parameters (complex admittance $y_k = g_k + \sqrt{-1}\, b_k$), the charging susceptance b_k^c, and the shunt admittance of a bus $y_i^{sh} = g_i^{sh} + \sqrt{-1}\, b_i^{sh}$ determine the power flow on line $k = (i, j, n) \in \mathcal{A}$, where P_k and Q_k denote active and reactive power flow, respectively:

$$\begin{aligned} P_k =& g_k \frac{v_i^2}{\tau_{1,k}^2} - g_k \frac{v_i v_j}{\tau_{1,k}\tau_{2,k}} \cos(\theta_i - \sigma_k - \theta_j) - \\ & b_k \frac{v_i v_j}{\tau_{1,k}\tau_{2,k}} \sin(\theta_i - \sigma_k - \theta_j), \qquad \forall k = (i, j, n) \in \mathcal{A} \end{aligned} \tag{3.3a}$$

$$\begin{aligned} Q_k =& -(b_k + \frac{b_k^c}{2}) \frac{v_i^2}{\tau_{1,k}^2} + b_k \frac{v_i v_j}{\tau_{1,k}\tau_{2,k}} \cos(\theta_i - \sigma_k - \theta_j) - \\ & g_k \frac{v_i v_j}{\tau_{1,k}\tau_{2,k}} \sin(\theta_i - \sigma_k - \theta_j), \qquad \forall k = (i, j, n) \in \mathcal{A}. \end{aligned} \tag{3.3b}$$

Here we split the tap ratio for each line $k = (i, j, n) \in \mathcal{A}$ into $\tau_{1,k}$ and $\tau_{2,k}$ to represent the change of voltage at two ends of that line. We have $\tau_{1,k} = \tau$ and $\tau_{2,k} = 1$ if a transformer with tap ratio τ is located at the bus i of line $k = (i, j, n) \in \mathcal{A}$, while we have $\tau_{1,k} = 1$, $\tau_{2,k} = \tau$ if a transformer with tap ratio τ is located at the bus j of line $k = (i, j, n) \in \mathcal{A}$. Similarly, for the transformer phase angle shift, if a transformer with phase angle shift is located at the bus i of line $k = (i, j, n) \in \mathcal{A}$, we set $\sigma_k = \sigma$; otherwise, if a transformer with phase angle shift is located at the bus j of line $k = (i, j, n) \in \mathcal{A}$, we set $\sigma_k = -\sigma$.

At each bus $i \in \mathcal{N}$, we enforce flow conservation of active and reactive power via equations (3.4). The left-hand side of constraint (3.4) is the net active and reactive power flowing out of bus i, and they equal the sum of controllable and uncontrollable injections:

$$\sum_{k=(i,j,n)\in\mathcal{A}} P_k + g_i^{sh}(v_i)^2 = \sum_{g\in\mathcal{G}_i} s_g^p + \left(u_i^{p,0} + u_i^{p,+} - u_i^{p,-}\right),\ \forall i \in \mathcal{N} \tag{3.4a}$$

$$\sum_{k=(i,j,n)\in\mathcal{A}} Q_k - b_i^{sh}(v_i)^2 = \sum_{g\in\mathcal{G}_i} s_g^q + \left(u_i^{q,0} + u_i^{q,+} - u_i^{q,-}\right),\ \forall i \in \mathcal{N}. \tag{3.4b}$$

Constraint (3.5a) bounds the difference in phase angle between adjacent buses, constraint (3.5b) limits the apparent power flowing through each line k, and constraints (3.5c)-(3.5f) provide simple bounds on voltage and phase angle at each bus and active and reactive power at each generator:

$$\underline{\Delta}_k \leqslant \theta_i - \sigma_k - \theta_j \leqslant \bar{\Delta}_k \qquad \forall k = (i,j,n) \in \mathcal{A} \tag{3.5a}$$

$$P_k^2 + Q_k^2 \leqslant W_k^2 \qquad \forall k \in \mathcal{A} \tag{3.5b}$$

$$\underline{v}_i \leqslant v_i \leqslant \bar{v}_i \qquad \forall i \in \mathcal{N} \tag{3.5c}$$

$$\underline{\theta}_i \leqslant \theta_i \leqslant \bar{\theta}_i \qquad \forall i \in \mathcal{N} \tag{3.5d}$$

$$\underline{s}_g^p \leqslant s_g^p \leqslant \bar{s}_g^p \qquad \forall g \in \mathcal{G} \tag{3.5e}$$

$$\underline{s}_g^q \leqslant s_g^q \leqslant \bar{s}_g^q \qquad \forall g \in \mathcal{G}. \tag{3.5f}$$

We denote the cost of controllable injections as $c(s^p, s^q)$, and assume c is convex, where s^p and s^q are $|\mathcal{G}|$-dimensional vectors with respective components s_g^p and s_g^q, $g \in \mathcal{G}$. The first-stage decision variables, s^p and s^q, denote controllable injections that cannot adapt to the realized scenario. We allow small adjustments to these injections via variables $o^{p,+}, o^{p,-}, o^{q,+}, o^{q,-}$, which can be selected once the uncertainty is revealed. These denote near real-time compensation in net generation, which has an upper bound proportional to the generation capacity at each bus. This setting permits greater flexibility than the linearly adaptive control used in previous research (Bienstock et al. 2014, Jabr 2013, Louca and Bitar 2017, Lubin et al. 2016). We seek a robust optimal controllable injection such that for all possible uncontrollable injections in $\mathcal{U}$, there is a feasible system configuration via variables $(v, \theta, o^{p,+}, o^{p,-}, o^{q,+}, o^{q,-}, P, Q)$.

We minimize the set point cost, and consider linear and convex quadratic cost functions:

$$c(s^p, s^q) = \sum_{g\in\mathcal{G}} \left[c_{g,2}^p \left(s_g^p\right)^2 + c_{g,1}^p s_g^p + c_{g,2}^q \left(s_g^q\right)^2 + c_{g,1}^q s_g^q\right],$$

where $c^p_{g,2} \geqslant 0$ and $c^q_{g,2} \geqslant 0$ for all $g \in \mathcal{G}$ and take value zero in the linear case.

Convexity of the cost function is important because, although the ACOPF problem has non-convex constraints, a convex objective function, together with the convex relaxation of the feasible region to be discussed below, yields a convex program. Our robust optimization formulation can be expressed as follows:

$$\min \quad c(s^p, s^q) \tag{3.6a}$$

$$\text{s.t.} \quad \underline{s}^p_g \leqslant s^p_g \leqslant \bar{s}^p_g \qquad \forall g \in \mathcal{G} \tag{3.6b}$$

$$\underline{s}^q_g \leqslant s^q_g \leqslant \bar{s}^q_g \qquad \forall g \in \mathcal{G} \tag{3.6c}$$

$$P^u_k = g_k \frac{(v^u_i)^2}{\tau^2_{1,k}} - g_k \frac{v^u_i v^u_j}{\tau_{1,k}\tau_{2,k}} \cos(\theta^u_i - \sigma_k - \theta^u_j) - b_k \frac{v^u_i v^u_j}{\tau_{1,k}\tau_{2,k}} \sin(\theta^u_i - \sigma_k - \theta^u_j) \qquad \forall k = (i,j,n) \in \mathcal{A},\ u \in \mathcal{U} \tag{3.6d}$$

$$Q^u_k = -(b_k + \frac{b^c_k}{2})\frac{(v^u_i)^2}{\tau^2_{1,k}} + b_k \frac{v^u_i v^u_j}{\tau_{1,k}\tau_{2,k}} \cos(\theta^u_i - \sigma_k - \theta^u_j) - g_k \frac{v^u_i v^u_j}{\tau_{1,k}\tau_{2,k}} \sin(\theta^u_i - \sigma_k - \theta^u_j) \qquad \forall k = (i,j,n) \in \mathcal{A},\ u \in \mathcal{U} \tag{3.6e}$$

$$\underline{\Delta}_k \leqslant \theta^u_i - \sigma_k - \theta^u_j \leqslant \bar{\Delta}_k \qquad \forall k = (i,j,n) \in \mathcal{A},\ u \in \mathcal{U} \tag{3.6f}$$

$$(P^u_k)^2 + (Q^u_k)^2 \leqslant W^2_k \qquad \forall k \in \mathcal{A},\ u \in \mathcal{U} \tag{3.6g}$$

$$\underline{v}_i \leqslant v^u_i \leqslant \bar{v}_i \qquad \forall i \in \mathcal{N},\ u \in \mathcal{U} \tag{3.6h}$$

$$\underline{\theta}_i \leqslant \theta^u_i \leqslant \bar{\theta}_i \qquad \forall i \in \mathcal{N} \tag{3.6i}$$

$$\sum_{k=(i,j,n)\in\mathcal{A}} P^u_k + g^{sh}_i (v^u_i)^2 + o^{p,-,u}_i - o^{p,+,u}_i = \sum_{g\in\mathcal{G}_i} s^p_g + \left(u^{p,0}_i + u^{p,+}_i - u^{p,-}_i\right) \qquad \forall i \in \mathcal{N},\ u \in \mathcal{U} \tag{3.6j}$$

$$\sum_{k=(i,j,n)\in\mathcal{A}} Q^u_k - b^{sh}_i (v^u_i)^2 + o^{q,-,u}_i - o^{q,+,u}_i = \sum_{g\in\mathcal{G}_i} s^q_g + \left(u^{q,0}_i + u^{q,+}_i - u^{q,-}_i\right) \qquad \forall i \in \mathcal{N},\ u \in \mathcal{U} \tag{3.6k}$$

$$o^{p,+,u}_i \leqslant \bar{o}^p_i + \left(h^p_i + \zeta^+_i u^{p,+}_i - \zeta^-_i u^{p,-}_i\right) \qquad \forall i \in \mathcal{N}, u \in \mathcal{U} \tag{3.6l}$$

$$o^{q,+,u}_i \leqslant \bar{o}^q_i + \left(h^q_i + \zeta^+_i u^{q,+}_i - \zeta^-_i u^{q,-}_i\right) \qquad \forall i \in \mathcal{N}, u \in \mathcal{U} \tag{3.6m}$$

$$o_i^{p,-,u} \leqslant \bar{o}_i^p \qquad \forall i \in \mathcal{N}, u \in \mathcal{U} \tag{3.6n}$$

$$o_i^{q,-,u} \leqslant \bar{o}_i^q \qquad \forall i \in \mathcal{N}, u \in \mathcal{U} \tag{3.6o}$$

$$o_i^{p,+,u}, o_i^{p,-,u}, o_i^{q,+,u}, o_i^{q,-,u} \geqslant 0 \qquad \forall i \in \mathcal{N}, u \in \mathcal{U}. \tag{3.6p}$$

Model (3.6) seeks a first stage vector of generation dispatch decisions, (s^p, s^q), that minimizes controllable generation cost. All other decision variables, including power compensations, voltages and phase angles at buses, as well as power flow on lines, adapt to the realization of uncertainty. Constraints (3.6b)-(3.6c) replicate the simple bounds on injections (3.5e)-(3.5f), constraints (3.6d)-(3.6e) replicate the power flow equations (3.3) for each $u \in \mathcal{U}$, and constraints (3.6f)-(3.6i) similarly replicate (3.5a)-(3.5d). Constraints (3.6j) and (3.6k) modify constraints (3.4) by incorporating the deviation variables, whose values are limited by (3.6l)-(3.6o). The maximum adjustment at a bus, due to traditional generators, is denoted by $\bar{o}$. Net load uncertainty includes generation uncertainty, due to renewable sources and demand uncertainty. When an uncertain parameter is larger than its nominal value, this can be because load is low or because RES generation is high. In the latter case, we allow for curtailment of RES generation. Parameters ζ_i^+ and ζ_i^- represent the fraction of total uncertainty due to RES generation, and h_i^p and h_i^q denote nominal renewable generation. The right-hand sides of constraints (3.6l) and (3.6m) capture the option for curtailment, and we discuss this in greater detail in Section 3.4.1.1. It is well known that the power flow equations (3.3), as well as the shunt components in (3.6j) and (3.6k), are nonconvex, and so model (3.6) is an infinite-dimensional nonconvex robust optimization problem with recourse.

There are multiple convex relaxation schemes for ACOPF problems. In the semidefinite programming relaxation of Bai and Wei (2011) and Bai et al. (2008), the vector of voltage variables in model (3.6) is re-expressed as a higher-dimensional matrix, coupled with a rank-one constraint and a positive semidefinite requirement, along with a collection of linear constraints. After dropping the rank-one constraint, the relaxed problem becomes an SDP and can be solved by an interior point method. Experience on realistically sized instances suggests that such SDP formulations are computationally expensive, and so Jabr (2006) proposes a further relaxation of the positive semidefinite constraint, yielding a second-order cone program (SOCP). Although computationally

easier to solve, this SOCP relaxation has the disadvantage of tending to exhibit a larger optimality gap than the SDP relaxation for many test cases. See Low (2014a) for a detailed review of such SDP and SOCP relaxations.

We use the convex relaxation that Coffrin et al. (2016) call the quadratic convex (QC) relaxation. While the QC formulation is also an SOCP, it tightens the relaxation compared to previous SOCP formulations. Coffrin et al. (2016) suggest relaxing equation (3.3) by replacing trigonometric functions by quadratic functions and using a McCormick relaxation to linearize the multi-linear terms. The quadratic terms in (3.6j) and (3.6k), v_i^2, are replaced by $\hat{v}_i$, which is constrained by a linear upper bound and a quadratic lower bound. The formulation of the QC relaxation of model (3.6) is detailed in Appendix B.2. Here, we use generic notation x to represent the system configuration and express the convex relaxation of model (3.6), as formulated in Appendix B.2, more compactly in model (3.7) below.

In what follows, we largely use a vector form to denote the controllable and uncontrollable injections for conciseness. A symbol without a subscript represents a vector, while a subscript-indexed symbol represents a specific component within that vector. Here we denote $u^{p,+}$, $u^{p,-}$, $u^{q,+}$ and $u^{q,-}$ as $|\mathcal{N}|$-dimensional vectors of uncontrollable active and reactive deviation. Similar notation is used for $\bar{u}, \underline{u}, u^0, s, \bar{s}$ and $\underline{s}$ as:

$$\bar{u} = \begin{bmatrix} \bar{u}^p \\ \bar{u}^q \end{bmatrix}, \ \underline{u} = \begin{bmatrix} \underline{u}^p \\ \underline{u}^q \end{bmatrix}, \ u^0 = \begin{bmatrix} u^{p,0} \\ u^{q,0} \end{bmatrix}, \ s = \begin{bmatrix} s^p \\ s^q \end{bmatrix}, \ \bar{s} = \begin{bmatrix} \bar{s}^p \\ \bar{s}^q \end{bmatrix}, \ \underline{s} = \begin{bmatrix} \underline{s}^p \\ \underline{s}^q \end{bmatrix}.$$

In this context, we also represent the active and reactive controllable injections of each bus $i \in \mathcal{N}$ as a linear transformation of the vector of generation s^p and s^q:

$$Ds^p = \left[\sum_{g \in \mathcal{G}_i} s_g^p \right]_{i \in \mathcal{N}} \quad \text{and} \quad Ds^q = \left[\sum_{g \in \mathcal{G}_i} s_g^q \right]_{i \in \mathcal{N}},$$

for an appropriate matrix D. We use ζ^+ and ζ^- to denote $|\mathcal{N}| \times |\mathcal{N}|$ diagonal matrices with entries ζ_i^+ and ζ_i^-, $\forall i \in \mathcal{N}$. This leads to the following compact formulation for the convex relaxation of

model (3.6):

$$\min \quad c(s) \tag{3.7a}$$

$$\text{s.t.} \quad \underline{s} \leqslant s \leqslant \bar{s} \tag{3.7b}$$

$$Ax^u \leqslant b \qquad \forall u \in \mathcal{U} \tag{3.7c}$$

$$\|B_i x^u + a_i\|_2 \leqslant e_i^\top x^u + f_i \qquad \forall i = 1, \ldots, m_c,\ u \in \mathcal{U} \tag{3.7d}$$

$$A^{op} x^u \leqslant \bar{o}^p + h^p + \zeta^+ u^{p,+} - \zeta^- u^{p,-} \qquad \forall u \in \mathcal{U} \tag{3.7e}$$

$$A^{oq} x^u \leqslant \bar{o}^q + h^q + \zeta^+ u^{q,+} - \zeta^- u^{q,-} \qquad \forall u \in \mathcal{U} \tag{3.7f}$$

$$A^p x^u = D s^p + u^{p,0} + u^{p,+} - u^{p,-} \qquad \forall u \in \mathcal{U} \tag{3.7g}$$

$$A^q x^u = D s^q + u^{q,0} + u^{q,+} - u^{q,-} \qquad \forall u \in \mathcal{U}. \tag{3.7h}$$

Constraint (3.7b) replicates the analogous constraints (3.6b) and (3.6c). The linear inequality (3.7c) and the SOCP constraint (3.7d) capture constraint (3.6g), and the relaxation of the nonlinear terms in constraints (3.6d)-(3.6e) and (3.6j)-(3.6k), while the linear inequality (3.7c) also includes (3.6f), (3.6h)-(3.6i), and (3.6n)-(3.6p). Constraints (3.7e) and (3.7f) match their counterparts (3.6l) and (3.6m). Finally, constraints (3.7g) and (3.7h) replicate linearized constraints (3.6j) and (3.6k). Model (3.7) can also represent the robust convex relaxation of the ACOPF problem in which we replace the QC relaxation with alternative convex relaxations discussed in Bai et al. (2008), Jabr (2006) and Kocuk et al. (2016).

Model (3.7) is an infinite-dimensional convex optimization problem, and is an example of robust optimization with recourse. Such models have been discussed in the context of linear programming in Terry (2009) and Thiele et al. (2009). In power systems optimization, similar formulations have been applied to unit commitment problems (Jiang et al. 2012, 2014), electricity markets (Zugno and Conejo 2015), and microgrid operations (Khodaei 2014). There has been limited work on conic programming, or more general convex programming, variants of such models (Terry 2009). In the next section we discuss the reformulation of this problem and the algorithm to solve the reformulated finite-dimensional problem.

3.3 A Cutting-Plane Method

In this section we propose a cutting-plane method to solve the robust convex optimization problem (3.7). To facilitate decomposition of model (3.7), we project onto the set of feasible (s^p, s^q) variables, and we employ an outer approximation to iteratively characterize this set. At each iteration, given a candidate solution, we compute, and add to the master problem, the most-violated inequality. With introduction of auxiliary binary decision variables, we can transform what would otherwise be an infinite number of constraints in (3.7) into a finite formulation and obtain a solution within some acceptable tolerance from the feasible set.

3.3.1 Master Problem and Subproblems

Similar to the generalized Benders' decomposition method of Geoffrion (1972), we can rewrite model (3.7) as:

$$\min \quad c(s) \tag{3.8a}$$

$$\text{s.t.} \quad s \in \mathcal{S} \equiv \cap_{u \in \mathcal{U}} \mathcal{S}^u \cap \{s \mid \underline{s} \leqslant s \leqslant \bar{s}\}, \tag{3.8b}$$

where for each $u \in \mathcal{U}$ we have the induced feasibility set

$$\mathcal{S}^u = \left\{ s \;\middle|\; \exists x \text{ s.t.} \begin{array}{l} Ax \leqslant b \\ \|B_i x + a_i\|_2 \leqslant e_i^\top x + f_i \quad \forall i = 1, \ldots, m_c \\ A^{op} x \leqslant \bar{o}^p + h^p + \zeta^+ u^{p,+} - \zeta^- u^{p,-} \\ A^{oq} x \leqslant \bar{o}^q + h^q + \zeta^+ u^{q,+} - \zeta^- u^{q,-} \\ A^p x = D s^p + u^{p,0} + u^{p,+} - u^{p,-} \\ A^q x = D s^q + u^{q,0} + u^{q,+} - u^{q,-} \end{array} \right\}. \tag{3.9}$$

Here, λ^p and λ^q denote dual variables for constraints (3.11f) and (3.11g), respectively. We introduce auxiliary variables $\ell^{p,+}, \ell^{p,-}, \ell^{q,+}$ and $\ell^{q,-}$ to represent the potential violation of power balance constraints. If optimal value for any of these variables is greater than 0, for the master

solution $\hat{s}$, there is no feasible x such that constraints in (3.9) can be satisfied, and a feasibility cut (3.10c) can be generated. We know from a generalized theorem of the alternative (Geoffrion 1972) that if s violates the inequalities (3.10c) then $s \notin \cap_{u\in\mathcal{U}}\mathcal{S}^u$. (We return to this in Lemma 3.3.2.) Index k corresponds to the k-th inequality, and the scalar cut intercept, z^k, accounts for all objective function terms in the dual of model (3.11) that do not involve $\hat{s}^p$ and $\hat{s}^q$.

This reformulation motivates a cutting-plane algorithm in which we iteratively solve a master problem and a collection of SOCP subproblems. In addition to the simple bounds, constraint (3.8b) requires that s be in the intersection of $\mathcal{S}^u,\ \forall u \in \mathcal{U}$. When we solve the subproblems, we either find a feasible x^u for each $u \in \mathcal{U}$, or we generate linear cuts, each of which is a valid outer approximation for $\mathcal{S}$. The master (M) and the subproblem (S^u) are as follows:

$$
\begin{aligned}
(M)\quad V^* = \min\quad & c(s) && && (3.10a)\\
\text{s.t.}\quad & \underline{s} \leqslant s \leqslant \bar{s} && && (3.10b)\\
& -\lambda^{p,k\top} D s^p - \lambda^{q,k\top} D s^q + z^k \leqslant 0 && \forall k = 1,2,\ldots && (3.10c)
\end{aligned}
$$

$$
\begin{aligned}
(S^u)\quad \min\quad & 1^\top \left(l^{p,+} + l^{p,-} + l^{q,+} + l^{q,-}\right) && && (3.11a)\\
\text{s.t.}\quad & Ax \leqslant b && && (3.11b)\\
& \|B_i x + a_i\|_2 \leqslant e_i^\top x + f_i && \forall i = 1,\ldots,m_c && (3.11c)\\
& A^{op} x \leqslant \bar{o}^p + h^p + \zeta^+ u^{p,+} - \zeta^- u^{p,-} && && (3.11d)\\
& A^{oq} x \leqslant \bar{o}^q + h^q + \zeta^+ u^{q,+} - \zeta^- u^{q,-} && && (3.11e)\\
& A^p x + l^{p,+} - l^{p,-} = D\hat{s}^p + u^{p,0} + u^{p,+} - u^{p,-} && && (3.11f)\\
& A^q x + l^{q,+} - l^{q,-} = D\hat{s}^q + u^{q,0} + u^{q,+} - u^{q,-} && && (3.11g)\\
& l^{p,+}, l^{p,-}, l^{q,+}, l^{q,-} \geqslant 0. && && (3.11h)
\end{aligned}
$$

The decomposition algorithm is not directly implementable because there are infinitely many subproblems, (S^u). So, we instead seek the most violated inequality across all elements of the

uncertainty set, which results in the following max-min problem:

$$\max_{u\in\mathcal{U}} \quad \min \quad 1^\top \left(l^{p,+} + l^{p,-} + l^{q,+} + l^{q,-}\right) \tag{3.12a}$$

$$\text{s.t.} \quad (3.11b) - (3.11h).$$

To reformulate model (3.12) in a computationally tractable manner we first take the dual of the inner minimization. We denote the dual variables for constraints (3.11b) and (3.11d)-(3.11g) by $\lambda, \lambda^{op}, \lambda^{oq}, \lambda^p$, and λ^q. For the second-order cone constraints in (3.11c), we denote the dual variables as $(\mu_i, \nu_i),\ i = 1, \ldots, m_c$. Then taking the dual yields:

$$\max_{u\in\mathcal{U}} \max_{\lambda,\lambda^{op},\lambda^{oq},\lambda^p,\lambda^q,\mu,\nu} \quad -\lambda^\top b - \sum_{i=1}^{m_c}\left(\nu_i f_i + \mu_i^\top a_i\right) -$$

$$\lambda^{op\top}(\bar{o}^p + h^p + \zeta^+ u^{p,+} - \zeta^- u^{p,-}) - \lambda^{oq\top}(\bar{o}^q + h^q + \zeta^+ u^{q,+} - \zeta^- u^{q,-}) -$$

$$\lambda^{p\top}\left(D\hat{s}^p + u^{p,0} + u^{p,+} - u^{p,-}\right) - \lambda^{q\top}\left(D\hat{s}^q + u^{q,0} + u^{q,+} - u^{q,-}\right) \tag{3.13a}$$

$$\text{s.t.} \quad \lambda^\top A + \lambda^{op\top} A^{op} + \lambda^{oq\top} A^{oq} + \lambda^{p\top} A^p + \lambda^{q\top} A^q$$

$$- \sum_{i=1}^{m_c}\left(\mu_i{}^\top B_i + \nu_i e_i^\top\right) = 0^\top \tag{3.13b}$$

$$\|\mu_i\|_2 \leqslant \nu_i \qquad \forall i = 1, \ldots, m_c \tag{3.13c}$$

$$-1 \leqslant \lambda_i^p \leqslant 1 \qquad \forall i \in \mathcal{N} \tag{3.13d}$$

$$-1 \leqslant \lambda_i^q \leqslant 1 \qquad \forall i \in \mathcal{N} \tag{3.13e}$$

$$\lambda, \lambda^{op}, \lambda^{oq} \geqslant 0. \tag{3.13f}$$

The optimal value of the inner maximization problem is a convex function of the $4|\mathcal{N}|$-dimensional vector u. The outer problem maximizes this convex function over the polytope $\mathcal{U}$. We know that an optimal solution can be obtained by restricting attention to the extreme points of $\mathcal{U}$, denoted by $\mathcal{U}^E$ (Enhbat 1996). When $\mathcal{U}$ has an amenable structure this can allow for a finite reformulation. In what follows, we assume that $\mathcal{U}$ is defined as in equation (3.2), and we introduce

$2|\mathcal{M}|$ binary variables to model the extreme points:

$$
\begin{aligned}
(SDI)\ \max \quad & -\lambda^\top b - \sum_{i=1}^{m_c} \left(\nu_i f_i + \mu_i^\top a_i\right) - \sum_{i\in\mathcal{N}} \Big[y^+_{m_i} \left(\bar{u}_i^p - u_i^{p,0}\right) (\lambda_i^p + \zeta_i^+ \lambda_i^{op}) + \\
& y^-_{m_i} \left(\underline{u}_i^p - u_i^{p,0}\right) (\lambda_i^p + \zeta_i^- \lambda_i^{op}) + y^+_{m_i} \left(\bar{u}_i^q - u_i^{q,0}\right) (\lambda_i^q + \zeta_i^+ \lambda_i^{oq}) + \\
& y^-_{m_i} \left(\underline{u}_i^q - u_i^{q,0}\right) (\lambda_i^q + \zeta_i^- \lambda_i^{oq}) \Big] - \lambda^{op\top} (\bar{o}^p + h^p) - \lambda^{oq\top} (\bar{o}^q + h^q) - \\
& \left[\lambda^{p\top} \left(D\hat{s}^p + u^{p,0}\right) + \lambda^{q\top} \left(D\hat{s}^q + u^{q,0}\right) \right] && \text{(3.14a)} \\
\text{s.t.} \quad & \text{(3.13b)} - \text{(3.13e)} \\
& y_m^+ + y_m^- \leqslant 1 \qquad \forall m \in \mathcal{M} && \text{(3.14b)} \\
& \sum_{m\in\mathcal{M}} \left(y_m^+ + y_m^-\right) \leqslant \Gamma && \text{(3.14c)} \\
& y_m^+, y_m^- \in \{0,1\} \qquad \forall m \in \mathcal{M} && \text{(3.14d)} \\
& \lambda, \lambda^{op}, \lambda^{oq} \geqslant 0. && \text{(3.14e)}
\end{aligned}
$$

For $i \in \mathcal{N}_m$, we use y_m^+ to indicate that $u_i^{p,+}$ and $u_i^{q,+}$ take their upper bound and y_m^- to indicate that $u_i^{p,-}$ and $u_i^{q,-}$ take their lower bound. The objective function in (3.14a) includes bilinear terms such as $\lambda_i^p y_{m_i}$, where m_i is used to indicate bus i's cluster. These are linearized in a straightforward way as shown in Appendix B.3.

Constraints (3.14b) enforce that at most one, instead of exactly one, end point of the feasible range is taken, and constraint (3.14c) requires that at most Γ clusters of uncontrollable injections taking their end point value. We include in Appendix B.3 the full formulation of model (3.14), which is derived from the convex quadratic relaxation detailed in Appendix B.2.

Algorithm 3 formalizes our cutting-plane procedure, where at iteration k we solve the master problem, (M), and obtain $(\hat{s}^{p,k}, \hat{s}^{q,k})$. Then, using the uncertainty set defined in equation (3.2), we solve model (3.14), and denote the optimal value by z_{feas}^k and part of the optimal solution by $\lambda^{p,k}, \lambda^{q,k}$. If $z_{feas}^k > 0$, we then generate the most violated cut as:

$$
z_{feas}^k - \lambda^{p,k\top} D(s^p - \hat{s}^{p,k}) - \lambda^{q,k\top} D(s^q - \hat{s}^{q,k}) \leqslant 0. \tag{3.15}
$$

With $z^k = z^k_{feas} + \lambda^{p,k\top} D\hat{s}^{p,k} + \lambda^{q,k\top} D\hat{s}^{q,k}$, inequality (3.15) is of form (3.10c).

Algorithm 3 Cutting-plane algorithm for model (3.7)

Initialize with iteration number $k := 1$ and tolerance $\varepsilon > 0$;
Solve master problem (M) and obtain solution $(\hat{s}^{p,k}, \hat{s}^{q,k})$ and optimal value V^*;
Solve (SDI) with $(\hat{s}^{p,k}, \hat{s}^{q,k})$ and obtain solution $(\lambda^{p,k}, \lambda^{q,k})$ and optimal value z^k_{feas};
while $z^k_{feas} > \varepsilon$ **do**
    Append $z^k_{feas} - \lambda^{p,k\top} D(s^p - \hat{s}^{p,k}) - \lambda^{q,k\top} D(s^q - \hat{s}^{q,k}) \leqslant 0$ to constraints (3.10c) of (M);
    Let $k := k + 1$;
    Solve (M) and obtain solution $(\hat{s}^{p,k}, \hat{s}^{q,k})$;
    if (M) is feasible **then**
        Obtain optimal value V^*;
    else
        Stop and return the status of infeasibility;
    Solve (SDI) with $(\hat{s}^{p,k}, \hat{s}^{q,k})$ and obtain solution $(\lambda^{p,k}, \lambda^{q,k})$ and optimal value z^k_{feas};
end while
Output V^* as a lower bound on the optimal value of model (3.7), and output $(\hat{s}^{p,k}, \hat{s}^{q,k})$ as an ε-feasible solution.

3.3.2 Convergence of the Algorithm

Given $\varepsilon > 0$, we show that in a finite number of iterations Algorithm 3 either finds an ε-feasible solution or terminates with a statement that model (3.7)—and hence model (3.6)—is infeasible. Furthermore, the sequence of solutions generated by our algorithm converges to an optimal solution when the tolerance in the algorithm is $\varepsilon = 0$. We make the notion of an "ε-feasible" solution precise as follows.

Definition 3.3.1. *Let $\varepsilon > 0$. An $s \in \{s \mid \underline{s} \leqslant s \leqslant \bar{s}\}$ is ε-feasible to model (3.7) if for each $u \in \mathcal{U}$ there exists an $\hat{s} \in \mathcal{B}_\varepsilon(s)$ such that*

$$\left\{ x \;\middle|\; \begin{array}{ll} Ax \leqslant b & \\ \|B_i x + a_i\|_2 \leqslant e_i^\top x + f_i \quad \forall i = 1, \ldots, m_c & \\ A^{op}x \leqslant \bar{o}^p + h^p + \zeta^+ u^{p,+} - \zeta^- u^{p,-} & A^{oq}x \leqslant \bar{o}^q + h^q + \zeta^+ u^{q,+} - \zeta^- u^{q,-} \\ A^p x = D\hat{s}^p + u^{p,0} + u^{p,+} - u^{p,-} & A^q x = D\hat{s}^q + u^{q,0} + u^{q,+} - u^{q,-} \end{array} \right\} \neq \varnothing, \quad (3.16)$$

where $\mathcal{B}_\varepsilon(s)$ is an l_1 ball with center s and radius ε.

For an ε-feasible s, the l_1 distance from s to the corresponding $\hat{s}$ is at most ε for each $u \in \mathcal{U}$. The definition does not ensure that there is a uniform $\hat{s}$ that works for all $u \in \mathcal{U}$. To establish convergence properties of Algorithm 3, we make the following assumptions:

Assumption 3.3.1. *Function $c(\cdot)$ is convex and continuous on domain defined by* (3.8b).

Assumption 3.3.2. *Set*

$$\left\{ x \;\middle|\; \begin{array}{l} Ax \leqslant b \\ \|B_i x + a_i\|_2 \leqslant e_i^\top x + f_i \quad \forall i = 1, \ldots, m_c \\ A^{op} x \leqslant \bar{o}^p + h^p + \zeta^+ u^{p,+} - \zeta^- u^{p,-} \\ A^{oq} x \leqslant \bar{o}^q + h^q + \zeta^+ u^{q,+} - \zeta^- u^{q,-} \end{array} \right\}$$

is non-empty, and hence model (3.11) *is feasible, for all $u \in \mathcal{U}$.*

Assumption 3.3.3. *Set $\mathcal{U}$ is defined by* (3.2).

Assumption 3.3.1 is consistent with the power systems literature because the cost for generation is typically modeled via a convex (piecewise) linear or quadratic function. Assumption 3.3.2 should hold with great generality for an actual power system because the set is a relaxation of the system's constraints, which does not include load satisfaction. Assumption 3.3.3 is revisited in Section 3.4.1.1. We now establish convergence properties of the sequence of solutions generated by Algorithm 3.

Lemma 3.3.1. *Let $Z^u(s)$ denote the optimal value of model* (3.11) *for a specific $u \in \mathcal{U}$, where $\hat{s}$ on the right-hand side of constraints* (3.11f) *and* (3.11g) *is replaced by s, and let $Z(s)$ denote the analogous optimal value for model* (3.12)*. If Assumptions 3.3.2 and 3.3.3 hold, then both $Z^u(\cdot)$ and $Z(\cdot)$ are convex on the domain $\Re^{2|\mathcal{G}|}$.*

Proof of Lemma 3.3.1. The function $Z^u(s)$ is the optimal value of model (3.11), which is feasible by Assumption 3.3.2 for any $s \in \Re^{2|\mathcal{G}|}$, and hence has a finite optimal value. Thus $Z^u(s)$ is also the optimal value of the dual of model (3.11). The dual's feasible region is independent of s, and its objective function is an affine function of s. Therefore, $Z^u(\cdot)$ is the maximum of a collection of affine functions in s, and hence convex. Furthermore, $Z(\cdot)$ is the maximum of convex functions

$Z^u(\cdot)$ over the set of $\mathcal{U}$, and so $Z(\cdot)$ is also convex. □

Lemma 3.3.2. *Let* $\mathcal{S}^k = \{s \mid \underline{s} \leqslant s \leqslant \bar{s},\ z^j_{feas} - \lambda^{p,j\top} D(s^p - \hat{s}^{p,j}) - \lambda^{q,j\top} D(s^q - \hat{s}^{q,j}) \leqslant 0,\ \forall j = 1, \ldots, k\}$*, where these cuts are defined in* (3.15)*. If Assumptions 3.3.2 and 3.3.3 hold then* $\mathcal{S} \subseteq \mathcal{S}^k,\ \forall k = 1, 2, \ldots$.

Proof of Lemma 3.3.2. At iteration k of Algorithm 3, the solution to model (3.14) specifies a specific element of $\mathcal{U}$ via binary variables, and we denote this element $u^k \in \mathcal{U}$. By a theorem of the alternative for an SOCP model (see Boyd and Vandenberghe 2004, Section 5.8), any inequality of the form (3.15) satisfies $\mathcal{S}^{u^k} \subseteq \{s \mid z^k_{feas} - \lambda^{p,k\top} D(s^p - \hat{s}^{p,k}) - \lambda^{q,k\top} D(s^q - \hat{s}^{q,k}) \leqslant 0\}$. Since $\mathcal{S} = \cap_{u\in\mathcal{U}} \mathcal{S}^u \cap \{s \mid \underline{s} \leqslant s \leqslant \bar{s}\}$, and each cut is produced for a specific u, we have that $\mathcal{S} \subseteq \mathcal{S}^k$ for all k. □

Theorem 3.3.1. *Let Assumptions 3.3.1-3.3.3 hold, and assume that model* (3.7) *is feasible. Let* $\varepsilon = 0$*, and let* $\{\hat{s}^k\}$ *denote the sequence of iterates produced by Algorithm 3. Every limit point of this sequence solves model* (3.7).

Proof of Theorem 3.3.1. If Algorithm 3 terminates in a finite number of iterations, then it does so with $z^k_{feas} = 0$. In this case, the associated solution solves model (3.7) by Lemma 3.3.2 because the master problem is a relaxation, and the proof is complete. Now assume that the algorithm produces an infinite sequence of iterates, and let $\mathcal{S}$ be defined as in (3.8b). Set $\mathcal{S}$ is compact because it is a closed subset of $\underline{s} \leqslant s \leqslant \bar{s}$. So, $\{\hat{s}^k\}$ has at least one limit point in $\mathcal{S}$, which we denote as $\hat{s}$, and we let $\mathcal{K}$ index a corresponding convergent subsequence; i.e., $\lim_{k\in\mathcal{K}, k\to\infty} \hat{s}^k = \hat{s}$.

Solving model (3.14) yields a $u^k \in \mathcal{U}$ that represents the most violated element of the uncertainty set. Because these solutions correspond to $\mathcal{U}^E$, there are a finite number of possibilities. So, there is at least one $\hat{u} \in \mathcal{U}$ that occurs infinitely many times among the iterations indexed by $\mathcal{K}$, and we let $\mathcal{K}' \subset \mathcal{K}$ denote such a further subsequence. Let $k, k' \in \mathcal{K}'$ with $k' > k$. Then we have

$$\begin{aligned} z^k_{feas} &\leqslant \lambda^{p,k\top} D(\hat{s}^{p,k'} - \hat{s}^{p,k}) + \lambda^{q,k\top} D(\hat{s}^{q,k'} - \hat{s}^{q,k}) \\ &\leqslant \|\lambda^{p,k}\| \|D(\hat{s}^{p,k'} - \hat{s}^{p,k})\| + \|\lambda^{q,k}\| \|D(\hat{s}^{q,k'} - \hat{s}^{q,k})\|. \end{aligned} \tag{3.17}$$

From constraints (3.13d) and (3.13e), we know $\|\lambda^p\|$ and $\|\lambda^q\|$ are bounded. Both $\hat{s}^k$ and $\hat{s}^{k'}$ converge to $\hat{s}$ so

$$\lim_{\substack{k\to\infty \\ k'\to\infty \\ k'>k \\ k',k\in\mathcal{K}'}} \left(\|\lambda^{p,k}\| \|(\hat{s}^{p,k'} - \hat{s}^{p,k})\| + \|\lambda^{q,k}\| \|(\hat{s}^{q,k'} - \hat{s}^{q,k})\| \right) = 0. \tag{3.18}$$

We let $Z^u(s)$ denote the optimal value of model (3.11) for a specific $u \in \mathcal{U}$, which is equivalent to the inner minimization problem of (3.12), and we let $Z(s)$ denote the optimal value of (3.12), where the right-hand side is parametrized by s rather than $\hat{s}$. Thus, we have:

$$\lim_{\substack{k\to\infty \\ k\in\mathcal{K}'}} Z^{\hat{u}}(s^k) = Z^{\hat{u}}(\hat{s}) = Z(\hat{s}) \leqslant 0, \tag{3.19}$$

where the first equality holds by continuity of $Z^{\hat{u}}(\cdot)$ from Lemma 3.3.1, and the second equality holds because $\hat{u}$ corresponds to a most violated point of $\mathcal{U}$. Thus, $\hat{s}$ is feasible to model (3.7). Let z^* denote the optimal value of model (3.7). Then $c(\hat{s}) \geqslant z^*$.

By Lemma 3.3.2, we have $c(\hat{s}^k) \leqslant z^*, \ \forall k \in \mathcal{K}$, and hence by Assumption 3.3.1, we have that

$$\lim_{\substack{k\to\infty \\ k\in\mathcal{K}}} c(\hat{s}^k) = c(\hat{s}) \leqslant z^*. \tag{3.20}$$

Thus, $\hat{s}$ solves model (3.7). □

Finally we show that when Algorithm 3 terminates, it returns an ε-feasible solution to model (3.7) in a finite number of iterations, if model (3.7) is feasible.

Theorem 3.3.2. *Let Assumptions 3.3.1-3.3.3 hold, and assume that model* (3.7) *is feasible. Let* $\varepsilon > 0$. *Algorithm 3 terminates with an* ε*-feasible solution in a finite number of iterations.*

Proof. Proof of Theorem 3.3.2 Model (3.7) is feasible, and hence $\mathcal{S} \neq \varnothing$. By Lemma 3.3.2 $\mathcal{S} \subseteq \mathcal{S}^k$, which is the feasible region of model (3.10) for all $k = 1, 2, \ldots$. Therefore, Algorithm 3 does not terminate with a status of infeasibility because model (3.10) is feasible for all $k = 1, 2, \ldots$.

We first prove by contradiction that the algorithm terminates in a finite number of iterations.

Here ε only determines the stopping critch3erion but does not affect the cuts generated ch3in Algorithm 3. Suppose Algorithm 3 does not terminate after a finite number of iterations. Thus, we have an infinite sequence of solutions $\{\hat{s}^k\}$, and $Z(\hat{s}^k) > \varepsilon,\ \forall k = 1, 2, \ldots$. By the proof of Theorem 3.3.1, every convergent subsequence of $\{\hat{s}^k\}$ indexed by $\mathcal{K}$ with a limit point $\hat{s}$ satisfies $Z(\hat{s}) \leqslant 0$. We have $\lim_{k\in\mathcal{K}, k\to\infty} Z(s^k) = Z(\hat{s}) \leqslant 0$ because $Z(\cdot)$ is convex and hence continuous. However, this contradicts that $Z(s^k) > \varepsilon > 0.\ \forall k = 1, 2, \ldots$. Therefore, the algorithm terminates in a finite number of iterations.

If Algorithm 3 terminates in iteration $k < \infty$, then $z^k_{feas} \leqslant \varepsilon$. By hypothesis, model (3.12) is feasible and has a finite optimal value. Hence, by strong duality, the optimal value of model (3.14) is equal to that of model (3.12) and is at most ε. Let $(\hat{s}^{p,k}, \hat{s}^{q,k})$ denote the input of Algorithm 3 (step 12) to model (3.14), or equivalently, to model (3.12). For each $u \in \mathcal{U}$, let $(x^u, l^{p,+,u}, l^{p,-,u}, l^{q,+,u}, l^{q,-,u})$ denote the optimal solution of the inner minimization problem defined in (3.12). For each $u \in \mathcal{U}$, let $s^{p,u} = \hat{s}^{p,k} - l^{p,+,u} + l^{p,-,u}$ and $s^{q,u} = \hat{s}^{q,k} - l^{q,+,u} + l^{p,-,u}$. From the formulation of model (3.12) we know that $(s^{p,u}, s^{q,u})$ yields $\{x \mid \text{(3.11b)-(3.11g)}\} \neq \varnothing$, and $\|s^u - \hat{s}^k\|_1 = 1^\top(l^{p,+} + l^{p,-} + l^{q,+} + l^{q,-}) = z^k_{feas} \leqslant \varepsilon$; i.e., $\hat{s}$ is an ε-feasible solution. □

3.3.3 Improving Convergence of Algorithm 3

It is well known that cutting-plane algorithms can converge slowly; see, e.g., Nemirovsky and Yudin (1983). This can occur because master problem solutions differ dramatically from one iteration to the next. There are multiple ways to improve such algorithms ranging from trust-region methods to level-set methods to bundle methods. We studied a bundle method by adding a quadratic regularization term to the master's objective function. This approach improved computational performance, but did not facilitate solving our largest test cases. The method is detailed in Appendix B.5.

Therefore we considered a second method in which we identify extreme points $u \in \mathcal{U}^E$ for which $\mathcal{S}^u$ characterizes important parts of the boundary of $\mathcal{S}$. In a Benders' decomposition algorithm for stochastic integer programs, Crainic et al. (2016) include a subset of the scenario subproblems in the master problem in order to reduce generation of feasibility cuts. We employ a similar approach,

but we discover the requisite elements u to be added to the master problem in the cutting-plane process instead of generating them upfront. (Lorca and Sun 2018 employ a similar idea in their Algorithm 2.) In each iteration of Algorithm 3 we record the $\hat{u}$ obtained by solving (SDI), and if a particular $\hat{u}$ is repeatedly generated n_c times then, instead of appending the linear cutting planes (3.15), we add $\hat{u}$ to a set $\widehat{\mathcal{U}}$ and use the modified master program:

$$\min \quad c(s) \tag{3.21a}$$

$$\text{s.t.} \quad \underline{s} \leqslant s \leqslant \bar{s} \tag{3.21b}$$

$$-\lambda^{p,k\top} Ds^p - \lambda^{q,k\top} Ds^q + z^k \leqslant 0 \qquad \forall k = 1, 2, \ldots \tag{3.21c}$$

$$Ax^u \leqslant b \qquad \forall u \in \widehat{\mathcal{U}} \tag{3.21d}$$

$$\|B_i x^u + a_i\|_2 \leqslant e_i^\top x^u + f_i, \qquad \forall i = 1, \ldots, m_c,\ u \in \widehat{\mathcal{U}} \tag{3.21e}$$

$$A^{op} x^u \leqslant \bar{o}^p + h^p + \zeta^+ u^{p,+} - \zeta^- u^{p,-} \qquad \forall u \in \widehat{\mathcal{U}} \tag{3.21f}$$

$$A^{oq} x^u \leqslant \bar{o}^q + h^q + \zeta^+ u^{q,+} - \zeta^- u^{q,-} \qquad \forall u \in \widehat{\mathcal{U}} \tag{3.21g}$$

$$A^p x^u = Ds^p + u^{p,0} + u^{p,+} - u^{p,-} \qquad \forall u \in \widehat{\mathcal{U}} \tag{3.21h}$$

$$A^q x^u = Ds^q + u^{q,0} + u^{q,+} - u^{q,-} \qquad \forall u \in \widehat{\mathcal{U}}. \tag{3.21i}$$

3.4 Experimental Results

3.4.1 Modeling and Implementation Details

In this section, we describe computational results to help understand the nature of our robust convex optimization problem and the performance of Algorithm 3, along with enhancements to that algorithm. The optimal value of model (3.7) is a lower bound on that of the nonconvex model (3.6). It is important to assess the tightness of this lower bound, while also answering the question of whether the robust solution generated by Algorithm 3 is feasible for model (3.6), at least for a selection of points from the uncertainty set. Doing so helps assess the robustness of our solution.

Prior to solving model (3.7), we run a bound tightening process to improve the quality of QC relaxation (detailed in Appendix B.4). Throughout this section we use Algorithm 3 with the

scenario-appending technique of Section 3.3.3. We use test cases from NESTA, the NICTA Energy System Test Case Archive (Coffrin et al. 2014). We select IEEE cases with 5, 9, 14, 118 and 300 buses and the Polish system winter peak cases with 2383 and 2746 buses. We refer to these by the number of buses (e.g., Case 5). All tests are run on a server with 20 Intel Xeon cores at 3.1 GHz and 256 GB of RAM. All models are constructed using version 0.18.0 of the JuMP package (Dunning et al. 2017) on the Julia platform. The mixed integer second-order cone programs (MISOCPs) and SOCPs are solved by Gurobi 7.52 (Gurobi Optimization, Inc. 2016), where we set the option "NumericFocus" to 3 for Case 2383 and Case 2746. All nonconvex optimization problems are solved by Ipopt 3.12.1 (Wächter and Biegler 2006), with the linear solver MA27.

We first introduce some modeling specifics used to build our test instances. Then we detail the tests to characterize our robust convex relaxation of the ACOPF problem and computational performance of our algorithms.

3.4.1.1 Uncertainty Set and Recourse Bounds

We first specify construction of the uncertainty set, $\mathcal{U}$, which includes both generation and demand uncertainty. Then we discuss two specific parameter selection schemes used in our tests. For each bus, $i \in \mathcal{N}$, nominal values of uncertain demand, $(d_i^p, d_i^q) \geqslant 0$, are known from the NESTA datasets. We model uncertain renewable generation at a subset of buses, $\mathcal{N}_G$, where selection of this subset is detailed in Appendix B.1. The nominal active generation from renewables is given by $h_i^p = 0.05|\mathcal{N}_G|^{-1} \sum_{i\in\mathcal{N}} d_i^p$ for $i \in \mathcal{N}_G$ and is zero otherwise. We fix the constant power factor at $\gamma = 98\%$ and calculate $h_i^q = \sqrt{\frac{1}{\gamma^2} - 1}\, h_i^p$. We assume the maximum allowable deviation of both generation and demand is a percentage of their nominal values, and the positive and negative deviation can differ. Therefore, we can parametrize the deviation by a set of percentages $(\alpha^{h,+}, \alpha^{h,-}, \alpha^{d,+}, \alpha^{d,-})$, with $\alpha^{h,+}, \alpha^{h,-}, \alpha^{d,+}, \alpha^{d,-} \in [0, 1]$:

$$\bar{u}_i^p = (1 + \alpha^{h,+})h_i^p - (1 - \alpha^{d,+})d_i^p \qquad \bar{u}_i^q = (1 + \alpha^{h,+})h_i^q - (1 - \alpha^{d,+})d_i^q \tag{3.22a}$$

$$\underline{u}_i^p = (1 - \alpha^{h,-})h_i^p - (1 + \alpha^{d,-})d_i^p \qquad \underline{u}_i^q = (1 - \alpha^{h,-})h_i^q - (1 + \alpha^{d,-})d_i^q. \tag{3.22b}$$

We model asymmetric uncertainty sets because we are concerned with demand spikes because of the right skewness of electricity demand (Maisano et al. 2016, Singh et al. 2010). In particular, at bus i, we set $\alpha^{d,-} = 5\alpha^{d,+}$ to focus on large negative deviation. We consider symmetric generation uncertainty, i.e., $\alpha^{h,+} = \alpha^{h,-}$, which is commonly used for generation uncertainty (Jiang et al. 2012, Attarha et al. 2018).

We assume that at each bus the upper bounds on recourse decisions in constraints (3.6l)-(3.6o) have the same ratio β to their corresponding maximum generation level; i.e.,

$$\bar{o}_i^p = \beta \sum_{g \in \mathcal{G}_i} \bar{s}_g^p \qquad \bar{o}_i^q = \beta \sum_{g \in \mathcal{G}_i} \bar{s}_g^q \qquad \forall i \in \mathcal{N}. \tag{3.23}$$

The curtailment coefficients ζ^+ and ζ^- can be derived under the current setup as follows:

$$\zeta_i^+ = \frac{\alpha^{h,+} h_i^p}{\alpha^{h,+} h_i^p + \alpha^{d,+} d_i^p} \qquad \zeta_i^- = \frac{\alpha^{h,-} h_i^p}{\alpha^{h,-} h_i^p + \alpha^{d,-} d_i^p} \qquad \forall i \in \mathcal{N}. \tag{3.24}$$

Power systems are distinguished by numerous characteristics. The use of α and β sketched above provides a relatively simple way of parameterizing the tests that follow.

3.4.1.2 Measure of Infeasibility

We measure infeasibility of a set point as follows. Given an $(\hat{s}^p, \hat{s}^q)$, it is possible that the nonconvex model (3.6) is infeasible for one or more values of $(u^{p,+}, u^{p,-}, u^{q,+}, u^{q,-}) \in \mathcal{U}$. Therefore, we modify the model to allow for additional flexibility in satisfying constraints (3.6j) and (3.6k) through variables $l_i^{p,+}, l_i^{p,-}, l_i^{q,+}, l_i^{q,-}$. For a given $u \in \mathcal{U}$ we measure the magnitude of infeasibility by:

$$I(\hat{s}, u) = \min \quad \sum_{i \in \mathcal{N}} \frac{\left(l_i^{p,+} + l_i^{p,-} + l_i^{q,+} + l_i^{q,-}\right)}{\sum_{i \in \mathcal{N}} \left(|d_i^p| + |d_i^q|\right)} \tag{3.25a}$$

$$\text{s.t.} \quad \text{constraints (3.6b)-(3.6i)}$$

$$\sum_{k=(i,j,n) \in \mathcal{A}} P_k + g_i^{sh} v_i^2 + o_i^{p,+} - o_i^{p,-} + l_i^{p,+} - l_i^{p,-}$$

$$= \sum_{g \in \mathcal{G}_i} \hat{s}_g^p + (u_i^{p,0} + u_i^{p,+} - u_i^{p,-}) \qquad \forall i \in \mathcal{N} \tag{3.25b}$$

$$\sum_{k=(i,j,n) \in \mathcal{A}} Q_k - b_i^{sh} v_i^2 + o_i^{q,+} - o_i^{q,-} + l_i^{q,+} - l_i^{q,-}$$

$$= \sum_{g \in \mathcal{G}_i} \hat{s}_g^q + (u_i^{q,0} + u_i^{q,+} - u_i^{q,-}) \qquad \forall i \in \mathcal{N} \tag{3.25c}$$

constraints (3.6l)-(3.6o)

$$o_i^{p,+}, o_i^{p,-}, o_i^{q,+}, o_i^{q,-} \geqslant 0 \qquad \forall i \in \mathcal{N} \tag{3.25d}$$

$$l_i^{p,+}, l_i^{p,-}, l_i^{q,+}, l_i^{q,-} \geqslant 0 \qquad \forall i \in \mathcal{N}. \tag{3.25e}$$

The infeasibility measure, $I(s, u)$, given by nonconvex model (3.25) yields a minimum normalized adjustment to the right-hand side of constraints (3.25b) and (3.25c) needed to construct a feasible recourse solution for a given $u \in \mathcal{U}$.

3.4.1.3 Solving Nonconvex Problems

In this section, we discuss two nonconvex problems solved in the chapter. First, as stated in Section 3.4.1.2, we solve model (3.25) to measure the infeasibility of a solution $\hat{s}$ for a given uncertainty scenario u. Since Ipopt can only give a local optimum to this problem, we actually obtain an upper bound of the infeasibility measure. To help measure the infeasibility accurately, we obtain a lower bound by solving the convex relaxation of model (3.25), which is model (3.11) with u specified. We denote this lower bound by $\underline{I}(\hat{s}, u)$.

For the second problem, we solve a relaxation to the nonconvex robust ACOPF problem. We have shown that solving model (3.7) provides a lower bound for the robust ACOPF problem (3.6). However, it is very hard to obtain a valid upper bound with a good quality (Nguyen et al. 2019). In our convex setting, i.e., model (3.7), we can restrict attention to $\mathcal{U}^E$. In the nonconvex setting, the worst-case violation need not be at an extreme point of $\mathcal{U}$. Therefore, we need to check every scenario within uncertainty set $\mathcal{U}$ to guarantee feasibility for model (3.6). In this chapter, we do not aim to obtain a valid upper bound but we try to estimate it with an approximation. To achieve this, we solve a nonconvex optimization model (3.6) with $\mathcal{U}$ substituted by $\mathcal{U}^E$, and obtain the

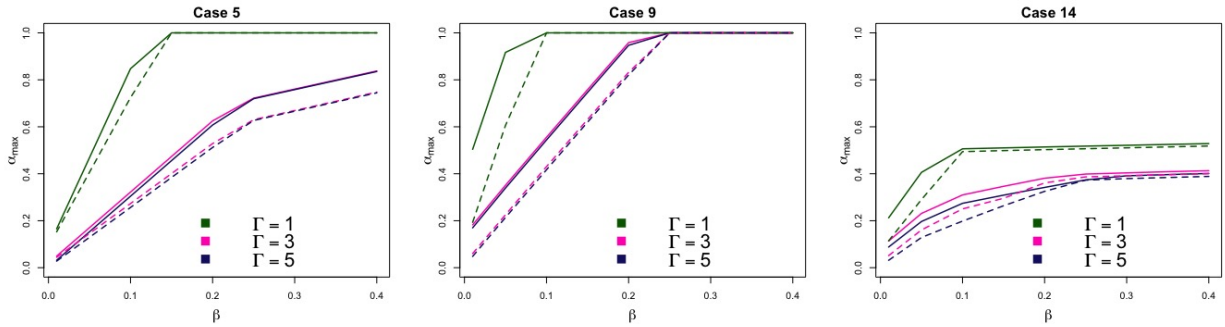


Figure 3.1: The plots specify $\alpha^{d,-}_{\max}$ as a function of β for various values of Γ; see equations (3.22) and (3.23). Here, $\alpha^{d,-}_{\max}$ is the largest value for which model (3.7) (solid line) / model (3.6) (dashed line) is feasible.

optimal solution, s^*. The optimal value obtained by Ipopt serves as a heuristic upper bound. If the solution, s^*, is indeed feasible for all $u \in \mathcal{U}$, Ipopt gives a local optimum, which is an upper bound for model (3.6). Although we cannot guarantee the feasibility of the solution, we compute $I(s^*, u)$ to show the magnitude of infeasibility for simulated $u \in \mathcal{U}$ for $\Gamma = 5$. This measure of infeasibility on random scenarios can reflect the robustness of s^*.

3.4.2 Descriptions of Tests and Results

First we explore the relationship between $\alpha^{d,-}$ and β, as specified in equations (3.22) and (3.23). For fixed values of the other α-parameters, β and Γ, where Γ is used in defining $\mathcal{U}$ in equation (3.2), we determine $\alpha^{d,-}_{\max}$, the largest value of $\alpha^{d,-}$ for which model (3.7) is feasible. Figure 3.1 suggests that, with other parameters fixed, $\alpha^{d,-}_{\max}$ is a concave function of β. To understand the concavity that arises in Figure 3.1 consider model (3.7), where we restrict $u \in \mathcal{U}$ to the extreme points of $\mathcal{U}$ and we replace the objective function by 0. We denote the feasible region of this model's dual by Λ, which is the intersection of multiple polyhedral and second-order cones. Writing this dual compactly we have:

$$\max_{\lambda \in \Lambda} \quad \alpha^{d,-}(a^\top \lambda) + \beta(b^\top \lambda) + c^\top \lambda, \tag{3.26}$$

where terms a and b come from (3.7g) and (3.7h). For the primal to be feasible, we must have

$$\alpha^{d,-}(a^\top \lambda) + \beta(b^\top \lambda) + c^\top \lambda \leqslant 0 \qquad \forall \lambda \in \Lambda. \tag{3.27}$$

This condition implies:

$$\alpha_{\max}^{d,-} = \inf_{\substack{\lambda \in \Lambda \\ a^\top \lambda > 0}} \frac{-\beta(b^\top \lambda) - c^\top \lambda}{a^\top \lambda}, \tag{3.28}$$

which is consistent with the concave functions for model (3.7) in Figure 3.1 (solid lines).

The dashed lines in Figure 3.1 repeat the $\alpha_{\max}^{d,-}$-β relationship for the nonconvex ACOPF model (3.6). Since model (3.7) is a relaxation of model (3.6), we see the former model can accommodate a slightly larger value of $\alpha_{\max}^{d,-}$ for a given β. While the gaps between solid and dashed lines in Figure 3.1 give some insight to the difference between models (3.6) and (3.7), this relationship does not involve the objective function, $c(s)$. Hence a small or large gap in Figure 3.1 does not necessarily correspond to a small or large optimality gap. For example, in results that we report below, Case 5 has the largest optimality gap while the difference between the curves for Case 5 in Figure 3.1 is modest.

All tests use $\alpha^{h,+} = \alpha^{h,-} = 1$, and $\beta = 0.05$; $\alpha^{d,-} = 5\alpha^{d,+}$ is set at the case-specific $\alpha_{\max}^{d,-}$ for $\beta = 0.05$, and the stopping tolerance is $\varepsilon = 10^{-4}$ $(\sum_{i \in \mathcal{N}} |d_i^p| + |d_i^q|)$ for all tests. We first assess three properties of our robust ACOPF problem: the quality of the lower bound generated by model (3.7), the robustness of the solution to model (3.7) in the nonconvex setting, and the performance of this robust solution relative to a deterministic alternative.

For each budget parameter Γ, a robust convex ε-feasible solution, $\hat{s}$, is first obtained by executing Algorithm 3 with the scenario-appending technique from Section 3.3.3. The cost associated with this solution, $c(\hat{s})$, provides a lower bound for model (3.6). We denote the lower bound obtained by solving the robust convex relaxation problem as C_R, and the estimated upper bound obtained by solving model (3.6) with $\mathcal{U}$ substituted by its extreme points as described in Section 3.4.1.3 to a local minimum as C_N. The gap is defined by $g = 100 \times \frac{C_N - C_R}{C_N}$.

We measure the infeasibility of this solution by solving model (3.25) at every extreme point of the uncertainty set, and report the maximum as $I(\hat{s}, \hat{u})$. We also obtain a deterministic nominal solution, $\hat{s}^0$, and optimal value, C_0, by solving the deterministic QC relaxation; i.e., model (3.7) with the singleton $\mathcal{U}$ defined under $\alpha^{d,-} = \alpha^{d,+} = \alpha^{h,-} = \alpha^{h,+} = 0$, and keeping the same recourse adjustment range so that the results are comparable. We again measure the infeasibility of this

Test Case	Γ	C_N	C_R	$g\%$	C_0	$I(\hat{s},\hat{u})$	$I(\hat{s}^0,\hat{u})$	$\underline{I}(\hat{s}^0,\hat{u})$
Case 5	1	17101.9	15005.7	12.26	12651.2	5.59×10^{-2}	9.86×10^{-2}	4.63×10^{-2}
	3	19863.1	17715.2	10.81	12651.2	5.71×10^{-2}	1.43×10^{-1}	1.19×10^{-1}
	5	20181.0	17979.6	10.91	12651.2	5.98×10^{-2}	1.46×10^{-1}	1.34×10^{-1}
Case 9	1	4751.4	4751.4	0.00	4059.1	4.46×10^{-5}	7.23×10^{-2}	7.23×10^{-2}
	3	5917.4	5917.3	0.00	4059.1	4.09×10^{-6}	1.81×10^{-1}	1.81×10^{-1}
	5	7208.9	6035.6	16.27	4059.1	3.92×10^{-3}	1.91×10^{-1}	1.91×10^{-1}
Case 14	1	233.0	232.9	0.02	209.0	0	6.07×10^{-2}	6.06×10^{-2}
	3	252.9	252.9	0.02	209.0	1.70×10^{-2}	1.11×10^{-1}	1.11×10^{-1}
	5	260.3	260.2	0.02	209.0	1.41×10^{-2}	1.34×10^{-1}	1.31×10^{-1}
Case 30	1	187.8	186.7	0.54	164.7	3.28×10^{-3}	5.15×10^{-2}	4.57×10^{-2}
	3	201.6	200.6	0.50	164.7	1.38×10^{-2}	8.01×10^{-2}	7.65×10^{-2}
	5	209.8	208.8	0.47	164.7	1.01×10^{-2}	9.87×10^{-2}	9.65×10^{-2}
Case 118	1	3456.2	3426.5	0.86	3110.6	1.87×10^{-2}	5.21×10^{-2}	4.81×10^{-2}
	3	3808.3	3777.3	0.81	3110.6	2.19×10^{-2}	1.04×10^{-1}	1.02×10^{-1}
	5	4045.4	4008.1	0.92	3110.6	1.59×10^{-2}	1.37×10^{-1}	1.35×10^{-1}
Case 300	1	15743.7	15116.7	3.98	13915.0	2.93×10^{-3}	2.56×10^{-2}	2.51×10^{-2}
	3	17794.6	16832.5	5.41	13915.0	2.66×10^{-3}	5.75×10^{-2}	5.69×10^{-2}
	5	18455.0	17522.3	5.05	13915.0	2.07×10^{-3}	7.96×10^{-2}	7.91×10^{-2}
Case 2383	1	1629795.2	1611397.2	1.13	1562639.8	5.48×10^{-3}	1.27×10^{-2}	8.93×10^{-3}
	3	1714285.1	1696575.3	1.03	1562639.8	5.79×10^{-3}	2.71×10^{-2}	2.59×10^{-2}
	5	1789041.5	1772009.8	0.95	1562639.8	5.48×10^{-3}	4.29×10^{-2}	4.18×10^{-2}
Case 2746	1	1483630.4	1480689.5	0.20	1440355.8	9.94×10^{-4}	1.51×10^{-2}	1.42×10^{-2}
	3	1564579.5	1561091.4	0.22	1440355.8	1.04×10^{-3}	4.20×10^{-2}	4.12×10^{-2}
	5	1623337.0	1619767.2	0.22	1440355.8	1.31×10^{-3}	6.10×10^{-2}	6.02×10^{-2}

Table 3.1: Robustness results of the robust convex relaxation solution and nominal solution

nominal solution at every extreme point of the uncertainty set, and report the maximum as $I(\hat{s}^0,\hat{u})$.

From Table 3.1 we observe that many of the gaps between the lower bound and estimated upper bound are below 1.5%, while the gaps for Case 5, Case 9 ($\Gamma = 5$), and Case 300 ($\Gamma = 3$ and $\Gamma = 5$) exceed 5%. This result suggests that solving model (3.7) can provide a tight lower bound for model (3.6). The large gap of Case 5 is caused by the convex relaxation, given a specific realization of uncontrollable injections, not being tight. In the deterministic setting, Coffrin et al. (2015b) report a gap of about 9.3% for Case 5. We can also find a general trend that the robust optimality gaps shown in Table 3.1 are larger than their deterministic counterpart described in Coffrin et al. (2015b). For a specific uncontrollable injection, the nonconvex feasible region may coincide with that of the QC relaxation near the optimum, but once we take the intersection of feasible regions under the robust setting, this may no longer be true. The degree of this phenomenon depends on

the power system structure and level of uncertainty, which may explain the larger gaps in Cases 9 and 300.

For Case 5, the infeasibility measure for model (3.7)'s solution is about 6% of the total demand. It is under 2.5% of the total demand for other cases, and does not grow with the size of uncertainty set (Γ). For example, the unmet demand is 0.12% of the total demand for $\Gamma = 1$ of Case 14, which equals to 0.31MW, and has limited impact in real world operations. These results contrast with the corresponding infeasibility of the nominal solution Column $I(\hat{s}^0, u)$, where the magnitude can be significantly larger. For all test cases but one (Case 5, $\Gamma = 1$), the lower bound of such infeasibility measure, $\underline{I}(\hat{s}^0, u)$, is closed to the upper bound, $I(\hat{s}^0, u)$, and also significantly larger than the upper bound of the infeasibility measure for the robust solution, $I(\hat{s}, u)$. This confirms the result that the nominal solution is inferior to the robust solution in terms of feasibility under the worst-case scenario.

To construct set $\mathcal{U}$, we correlate the uncontrollable injections at different buses as described in Section 3.2 and Appendix B.1. Of course, injections may not occur in a worst-case manner or in a manner with this type of correlation. To assess the performance of our solution in a stochastic environment, we assume that $u = (u^p, u^q)$ is a uniform random vector in the box specified by the bounds in equation (3.22), and we sample 1000 realizations. Given the solution ($\hat{s}$ obtained from model (3.7) for a specific $\Gamma \in \{1, 3, 5\}$, for each realization, we first solve the nonconvex model (3.25) to compute the upper bound of infeasibility measure, I. Next, we solve the convex relaxation of model (3.25) as stated in Section 3.4.1.3 to compute the lower bound of infeasibility measure, $\underline{I}$. Finally, given the nonconvex heuristic solution for $\Gamma = 5$, s^*, we solve the nonconvex model (3.25) to compute the upper bound of infeasibility measure and evaluate its robustness.

We show the computational results in Table 3.2. Among each batch of 1000 realizations, we denote the mean violation under the nonconvex setting by μ_I, and the expected maximum violation by $I_{\max}$. The corresponding infeasibility measures under the convex relaxation setting are denoted as $\mu_{\underline{I}}$ and $\underline{I}_{\max}$ Due to its probabilistic nature, we replicate this test 20 times to obtain a point estimate for $I_{\max}$ and μ_I as well as 95% confidence intervals. As expected, our robust solution for $\Gamma = 5$ is feasible for all $u \in \mathcal{U}$ for the convex relaxation (3.7) by construction, and infeasibility

Test Case	Γ	$I_{\max}\pm$ CI half width	$\underline{I}_{\max}\pm$ CI half width	$\mu_I\pm$ CI half width	$\mu_{\underline{I}}\pm$ CI half width
Case 5	1	$9.05 \times 10^{-2} \pm 8.71 \times 10^{-3}$	$6.33 \times 10^{-2} \pm 1.32 \times 10^{-2}$	$2.52 \times 10^{-2} \pm 1.34 \times 10^{-3}$	$3.08 \times 10^{-3} \pm 5.94 \times 10^{-4}$
	3	$4.93 \times 10^{-2} \pm 6.89 \times 10^{-3}$	$1.63 \times 10^{-4} \pm 1.43 \times 10^{-3}$	$7.24 \times 10^{-3} \pm 6.03 \times 10^{-4}$	$1.63 \times 10^{-7} \pm 1.43 \times 10^{-6}$
	5	$5.54 \times 10^{-2} \pm 3.14 \times 10^{-3}$	0 ± 0	$1.71 \times 10^{-2} \pm 6.38 \times 10^{-4}$	0 ± 0
Case 9	1	$9.24 \times 10^{-2} \pm 1.64 \times 10^{-2}$	$9.24 \times 10^{-2} \pm 1.64 \times 10^{-2}$	$4.17 \times 10^{-3} \pm 7.58 \times 10^{-4}$	$4.17 \times 10^{-3} \pm 7.58 \times 10^{-4}$
	3	0 ± 0	0 ± 0	0 ± 0	0 ± 0
	5	$6.69 \times 10^{-4} \pm 1.53 \times 10^{-3}$	0 ± 0	$7.93 \times 10^{-7} \pm 2.34 \times 10^{-6}$	0 ± 0
Case 14	1	$4.50 \times 10^{-2} \pm 1.08 \times 10^{-2}$	$4.49 \times 10^{-2} \pm 1.08 \times 10^{-2}$	$1.24 \times 10^{-3} \pm 2.17 \times 10^{-4}$	$1.23 \times 10^{-3} \pm 2.16 \times 10^{-4}$
	3	$1.57 \times 10^{-2} \pm 9.62 \times 10^{-4}$	$6.36 \times 10^{-4} \pm 3.18 \times 10^{-3}$	$1.71 \times 10^{-3} \pm 2.08 \times 10^{-4}$	$7.61 \times 10^{-7} \pm 4.05 \times 10^{-6}$
	5	$1.16 \times 10^{-2} \pm 9.65 \times 10^{-4}$	0 ± 0	$7.95 \times 10^{-4} \pm 1.28 \times 10^{-4}$	0 ± 0
Case 30	1	$2.94 \times 10^{-2} \pm 5.68 \times 10^{-3}$	$2.84 \times 10^{-2} \pm 6.36 \times 10^{-3}$	$1.06 \times 10^{-3} \pm 2.57 \times 10^{-4}$	$6.81 \times 10^{-4} \pm 2.04 \times 10^{-4}$
	3	$1.26 \times 10^{-2} \pm 6.46 \times 10^{-4}$	$6.42 \times 10^{-4} \pm 2.14 \times 10^{-3}$	$1.65 \times 10^{-3} \pm 2.38 \times 10^{-4}$	$9.73 \times 10^{-7} \pm 3.56 \times 10^{-6}$
	5	$8.56 \times 10^{-3} \pm 6.47 \times 10^{-4}$	0 ± 0	$6.24 \times 10^{-4} \pm 1.33 \times 10^{-4}$	0 ± 0
Case 118	1	$6.56 \times 10^{-2} \pm 1.45 \times 10^{-2}$	$5.72 \times 10^{-2} \pm 1.69 \times 10^{-2}$	$1.14 \times 10^{-2} \pm 7.56 \times 10^{-4}$	$3.29 \times 10^{-3} \pm 4.52 \times 10^{-4}$
	3	$2.42 \times 10^{-2} \pm 1.04 \times 10^{-2}$	$6.88 \times 10^{-3} \pm 1.32 \times 10^{-2}$	$9.44 \times 10^{-4} \pm 1.82 \times 10^{-4}$	$7.94 \times 10^{-6} \pm 1.60 \times 10^{-5}$
	5	$3.98 \times 10^{-3} \pm 2.84 \times 10^{-3}$	0 ± 0	$2.32 \times 10^{-5} \pm 1.79 \times 10^{-5}$	0 ± 0
Case 300	1	$3.62 \times 10^{-2} \pm 7.35 \times 10^{-3}$	$3.52 \times 10^{-2} \pm 7.38 \times 10^{-3}$	$2.57 \times 10^{-3} \pm 2.63 \times 10^{-4}$	$2.03 \times 10^{-3} \pm 2.45 \times 10^{-4}$
	3	$3.78 \times 10^{-3} \pm 4.04 \times 10^{-3}$	$3.15 \times 10^{-3} \pm 4.66 \times 10^{-3}$	$1.11 \times 10^{-4} \pm 2.08 \times 10^{-5}$	$4.03 \times 10^{-5} \pm 1.17 \times 10^{-5}$
	5	$1.38 \times 10^{-3} \pm 4.04 \times 10^{-3}$	0 ± 0	$7.99 \times 10^{-5} \pm 1.51 \times 10^{-5}$	0 ± 0
Case 2383	1	$2.41 \times 10^{-2} \pm 5.39 \times 10^{-3}$	$2.29 \times 10^{-2} \pm 5.44 \times 10^{-3}$	$3.31 \times 10^{-3} \pm 1.98 \times 10^{-4}$	$1.23 \times 10^{-3} \pm 2.04 \times 10^{-4}$
	3	$7.52 \times 10^{-3} \pm 4.03 \times 10^{-3}$	$5.71 \times 10^{-3} \pm 5.13 \times 10^{-3}$	$2.05 \times 10^{-3} \pm 6.60 \times 10^{-5}$	$1.38 \times 10^{-5} \pm 1.48 \times 10^{-5}$
	5	$4.89 \times 10^{-3} \pm 3.41 \times 10^{-4}$	0 ± 0	$2.06 \times 10^{-3} \pm 5.76 \times 10^{-5}$	0 ± 0
Case 2746	1	$3.22 \times 10^{-2} \pm 7.91 \times 10^{-3}$	$3.12 \times 10^{-2} \pm 7.90 \times 10^{-3}$	$1.95 \times 10^{-3} \pm 2.35 \times 10^{-4}$	$1.71 \times 10^{-3} \pm 2.21 \times 10^{-4}$
	3	$5.75 \times 10^{-3} \pm 5.06 \times 10^{-3}$	$4.59 \times 10^{-3} \pm 4.82 \times 10^{-3}$	$2.16 \times 10^{-5} \pm 1.14 \times 10^{-5}$	$7.73 \times 10^{-6} \pm 8.70 \times 10^{-6}$
	5	$3.74 \times 10^{-4} \pm 1.35 \times 10^{-4}$	0 ± 0	$4.18 \times 10^{-6} \pm 1.43 \times 10^{-6}$	0 ± 0

Table 3.2: Computational results for solving model (3.7) for a range of values of Γ, and then assessing feasibility, along with 95% confidence intervals, using 20 replications over 1000 uniformly distributed realizations over the hyper-rectangle governing u.

measures decrease monotonically as Γ increases. We observe a similar trend when checking the nonconvex infeasibility measures for all test cases but Case 5 and Case 9. We omit the infeasibility measure upper bound for solution s^* since they are all zeros, which indicates that our heuristic solution, s^*, is feasible for all simulated scenarios within $\mathcal{U}$. The evidence shows that s^* is likely to be feasible and the objective value associated to it, C_N, is likely to be a valid upper bound.

Next, we report the computational performance of Algorithm 3 and its scenario-appending improvement scheme. If we append scenario-specific constraints to the master problem, the master becomes larger and takes longer to solve, but this helps decrease the number of iterations of the cutting-plane algorithm. In Table 3.3 we show computational results for Cases 118 and 300, which best exemplify the effectiveness of the scenario-appending scheme. The table shows that direct application of Algorithm 3 fails to obtain an ε-feasible solution within 300 iterations, but by appending scenarios to the master, we solve Cases 118 and 300 in at most 15 iterations. The results that we report elsewhere in this section all use the improvement of appending scenarios to the

Parameters	No. of iterations		ε-feasibility achieved		T (sec.)	
	Case 118	Case 300	Case 118	Case 300	Case 118	Case 300
Algorithm 3	300	300	No	No	2414	23042
$n_c = 1$	5	6	Yes	Yes	61	550
$n_c = 2$	7	9	Yes	Yes	82	761
$n_c = 3$	8	15	Yes	Yes	89	1172
$n_c = 4$	9	11	Yes	Yes	96	761
$n_c = 5$	10	12	Yes	Yes	104	802

Table 3.3: Computational results of solving model (3.7) with different improvement techniques for Case 118 and Case 300 with $\Gamma = 3$.

master program in which n_c, the number of replications after which a scenario $\hat{u} \in \mathcal{U}$ is appended to the master program, is $n_c = 1$.

The scenario-appending method aims to reduce the number of iterations of Algorithm 3. To decrease the running time of each iteration, we compare the computational performance of two alternatives: solving the MISOCP (3.14) directly, or enumerating all extreme points of $\mathcal{U}$ and solving the corresponding SOCPs individually, which is possible when $|\mathcal{M}|$ is modest. For example, if $\Gamma = 1$, there are only $2|\mathcal{M}|$ extreme points, and solving this moderate number of SOCPs each iteration may reduce computation time, especially when parallelizing the calculations. Furthermore, rather than identifying $u \in \mathcal{U}$ for a most violated constraint, we can generate multiple feasibility cuts in one iteration to again attempt to reduce the number of overall iterations. For our tests, we generate cuts at the 10 most violated scenarios at each iteration, and we solve SOCPs in parallel with 20 threads.

We present the test result in Table 3.4. The number of extreme points of $\mathcal{U}$ is denoted by N and the number of iterations until ε-feasibility is achieved is denoted "iter." and the clock time of the two approaches is denoted by "time." For all test cases, solving the SOCPs in parallel requires significantly less time than solving MISOCPs. When $|\mathcal{M}|$ is small, the number of extreme points of $\mathcal{U}$ is modest. Therefore, solving SOCPs corresponding to every extreme point of $\mathcal{U}$ in parallel is more efficient than solving the MISOCPs.

In our implementation, we note that solving the MISOCP is the computational bottleneck. In spite of our effort to use the default Gurobi setting to generate linear outer approximation and

			MISOCP		SOCP		Nonconvex		Nominal
Test Case	Γ	N	iter.	time (sec.)	iter.	time (sec.)	iter.	time (sec.)	time (sec.)
Case 5	1	10	2	1.4	2	0.4	2	0.2	0.05
	3	80	2	1.1	2	0.7	2	0.4	0.07
	5	32	2	1.7	1	0.2	2	0.3	0.05
Case 9	1	10	2	2.2	2	0.4	2	0.3	0.09
	3	80	2	2.1	2	0.8	2	0.5	0.10
	5	32	2	3.3	1	0.3	2	0.4	0.08
Case 14	1	10	2	4.3	2	0.9	2	0.4	0.2
	3	80	2	5.1	2	1.5	3	1.3	0.2
	5	32	2	7.0	1	0.4	2	0.5	0.1
Case 30	1	10	2	11.0	2	1.4	2	0.6	0.3
	3	80	2	12.9	2	2.7	2	1.6	0.4
	5	32	2	18.6	1	0.8	2	0.9	0.3
Case 118	1	10	5	106.3	3	15.6	2	3.3	1.7
	3	80	5	154.8	2	18.5	3	13.5	1.6
	5	32	2	106.9	1	4.1	2	6.5	1.6
Case 300	1	10	6	501.8	3	44.2	3	17.2	4.3
	3	80	6	1278.9	2	61.6	2	25.7	4.7
	5	32	3	980.0	1	14.8	1	6.0	4.6
Case 2383	1	10	4	7225.5	2	344.8	2	87.1	76.4
	3	80	3	12853.2	2	848.6	3	366.4	76.2
	5	32	2	19407.3	1	257.7	1	56.1	79.4
Case 2746	1	10	–	–	2	557.6	2	250.8	102.5
	3	80	–	–	2	1175.3	2	330.5	121.9
	5	32	–	–	1	250.8	2	272.3	95.6

Table 3.4: Comparison between solving the MISOCP (3.14) and solving a set of SOCPs for the extreme points of $\mathcal{U}$, with $n_c = 1$ and Gurobi parameter NumericalFocus= 3.

utilize the warm start to solve the MISOCPs, we still encounter numerical problems when solving the MISOCP for Case 2746. On the other hand, once the number of extreme points becomes large, finding the most-violated scenario without going through all of them remains challenging and requires further research.

3.5 Conclusions

In this chapter we present a model to relax the nonconvex robust ACOPF problem to a robust convex program with recourse. A cutting-plane algorithm is proposed to solve the convex relaxation, and within each iteration of the cutting-plane algorithm, an MISOCP is solved to generate a cut

separating the incumbent solution from the robust convex feasible region. In summary, we:

- formulated a two-stage robust model that permits full recourse decisions rather than simpler, e.g., linear, decision rules;
- established desirable convergence properties of a cutting-plane algorithm;
- showed that our algorithm can provide a good lower bound for the nonconvex ACOPF problem (3.6);
- found the solution to the robust convex relaxation model (3.7) is robust in the nonconvex setting, provided its deterministic QC relaxation is reasonably tight; and,
- reduced solution time in the cutting-plane algorithm by appending a small number of key scenarios to the master program.

There are many possible ways to extend the result of this research. One important direction is to reduce the computational effort required to solve the "separation problem," which is currently modeled as an MISOCP. Doing so would further facilitate scaling our algorithm to larger problems. Finding a valid upper bound or prove the validity of our heuristic upper bound not only bears theoretical significance, but also has potential to be combined with optimization based bound tightening process (Sundar et al. 2018) to further tighten the formulation and the lower bound obtained in this chapter. Furthermore, our scenario-appending algorithms can be used in other applications of robust optimization with a convex recourse, such as microgrid planning (Khodaei 2014), location transportation (Gabrel et al. 2014), call center staffing (Zhao and Zeng 2012).

Chapter 4

Analyzing Client Behavior in a Syringe Exchange Program

4.1 Introduction

Three major agencies provide syringe exchange programs (SEPs) in the Chicago metropolitan area: Community Outreach Intervention Projects (COIP), Chicago Recovery Alliance (CRA), and Test Positive Aware Network (TPAN). Each agency offers equipment and educational services and conducts research on drug users. With the goal of supporting persons who inject drugs (PWIDs) and helping prevent the spread of infectious diseases, they provide services including street outreach, counseling and training for preventing HIV and hepatitis C, case management for persons living with HIV, assistance in entering treatment for substance use, and HIV medical, mental, and pharmacy care. Their locations include storefronts and mobile vans, which may operate according to a flexible schedule. The benefits of an SEP are twofold. First, multiple studies have shown that SEPs are effective in reducing risk behavior such as sharing syringes Bluthenthal et al. (2007, 2000), Braine et al. (2004), Holtzman et al. (2009), Huo and Ouellet (2007), thereby lowering rates of HIV and hepatitis C transmission. Second, higher utilization of an SEP provides PWIDs with more opportunities to learn about treatment programs, which further reduces drug use Huo et al. (2006), DeSimone (2005).

This chapter focuses on one of the SEPs in Chicago, and we refer to program participants as *clients.* Service locations, including storefronts and mobile vans, accept used syringes and, in exchange, provide clients with new syringes along with other devices that help prevent the spread of disease, such as condoms, cookers, purified water, and bleach. On their first visit to a service location, clients are asked to take a voluntary survey involving demographic information and the nature of their drug use (frequency, types of drugs, etc.), which we detail in Section 4.2. Once a client is established in the system, the SEP keeps a record of frequency of drug use, health condition, and general living condition. During a visit, an SEP employee will have a personal conversation with a client, e.g., about recent life changes, employment status, and family situation. The SEP will further provide the client with information to help with health issues, and seek to introduce the client to drug treatment programs. Evidence has shown that this type of personal interaction can help clients obtain peer support to recover from substance addiction Clarke et al. (2016), Hay et al. (2017), Kidorf and King (2008).

Nationwide, about 681,000 Americans aged 12 years or older reported using heroin in 2013 Abuse and Administration (2014), and the number of reported users grew every year from 2007 to 2013, with new users growing about 70% from 2002 to 2013. The volume of heroin seized by officials, and the number of heroin overdoses, both grew over the same period in Chicago Abuse and Administration (2014). The contrast between the significant growth in the use of heroin and slightly lower use of services at the target SEP (see Section 4.2) motivates our study. In order to promote its services and tailor them to individual clients, the agency needs a better understanding of client behavior in using the SEP.

The arrival process of clients to SEP sites is key to understanding their use of SEP services, and we have data on arrival times and locations over a ten-year period. Our main focus is to develop a *contextual* understanding of the arrival process; i.e., we want to understand inter-arrival times given *features* of an individual client obtained, in part, from the voluntary demographic survey. Our data suggest some clients "establish care" with the SEP, returning consistently, while others use SEP services once and never return. We seek to model and understand the "life cycle" of an individual client from initiation, reoccurring visits, and termination with the program. Through

interviews with SEP staff, we learned that ethnicity, gender, age, geographic location, and drug history affect how a client will use SEP services. For example, African Americans are less active in SEP programs because they tend to prefer not to disclose their drug habits, and because African Americans who use drugs tend to snort rather than inject. A client's interaction with the SEP is affected by life events; e.g., a person who moves farther from a service location may visit less frequently. These are snapshots of a pervasive phenomenon that suggest many factors can influence client behavior when using SEP services.

To address these issues, we propose a model for how a client engages with the SEP using three integrated sub-models for initiation, reoccurring visits, and termination. The latter two sub-models are fully integrated and involve a phase-type distribution, which may be viewed as a continuous-time Markov chain with hidden states. To build a contextual model, we express the Markov chain's parameters as a function of client features using linear and logistic regression. Analysis with our model can help the SEP understand the importance of different features and hence estimate the distribution governing a specific client's next arrival time. This, in turn, can help SEP staff better allocate limited resources to improve the program's effectiveness. Armed with a probability distribution for the timing of a client's next visit, SEP staff can be alerted when a specific client has not used SEP services for an unusual period of time, which may point to risky drug-use behavior. SEP employees can then contact the client (e.g., via a text message), or dispatch a mobile van to specific locations and message nearby clients. Using simulation, we illustrate the potential value of using our model in this manner.

Modeling arrival processes plays a key role in many application domains; see Lakshmi and Iyer (2013) for a review of relevant literature in healthcare. Homogeneous and nonhomogeneous Poisson processes are widely used to model an arrival process Aksin et al. (2007), Fomundam and Herrmann (2007), Govil and Fu (1999), but do not address our primary goal, i.e., to provide insights regarding a client's return time based on individual predictors associated with that client. In principle, we could partition clients into different categories based on their features and fit such arrival processes based on these categories. However, such an approach scales poorly given the number of features we consider.

As we will discuss, our data on inter-arrival times of clients to SEP sites have heavier tails than that of an exponential distribution, and this is true even when we develop models that condition on sub-populations of the clients. This may arise for two related reasons: (i) the "establishing care" nature of some clients discussed above, and (ii) clients effectively transitioning between hidden states, which capture active and passive engagement with the SEP. The three sub-model approach that we propose allows us to capture these effects, and its contextual nature captures heterogeneity. We use a negative binomial model for initiation, which is consistent with an over-dispersed mixture of Poisson distributions that captures heterogeneity. We use a two-state continuous-time Markov chain model with unobservable states to capture reoccurring visits and termination. These two integrated sub-models allow us to reasonably represent issues (i) and (ii) and their conditional transition- and system-exit probabilities allow us to capture heterogeneity.

Hidden Markov chain models have been used to model scenarios in which an agent transitions between a modest number of states, each associated with certain patterns of behavior. Paddock et al. Paddock et al. (2012) construct a Markov chain model to understand trajectories of a marijuana user with the goal of using simulation to assess alternative treatment and prevention policies. Liu et al. Liu et al. (2015) learn a continuous-time hidden Markov chain model for disease progression in glaucoma and Alzheimer's disease. Chehrazi et al. (Chehrazi et al. 2019, Appendix C) discuss using a two-state Markov chain to model the repayment behavior associated with delinquent credit card accounts, in which the two states represent high and low repayment rates, with the goal of directing credit collection efforts. Hidden Markov chains enable modelers to capture plausible, but unobservable, transitions of an agent. Our approach can further enhance such models by allowing the parameters of the Markov chain to depend on agent-specific features using regression. Given requisite data, the types of models just sketched could benefit from our approach, increasing model fidelity and insights from analysis, by linking agent heterogeneity to the Markov chain model. A technical challenge that we address in this chapter—at least for the family of serial Coxian models that we detail—involves parameter estimation when using regression to map the features of an agent to the Markov chain's parameters.

We first provide, in Section 4.2, an overview of the demographic and arrival data collected

by the SEP between 2005 and 2014. In Section 4.3, we provide details of the derivation of the three sub-models: initiation, reoccurring visits, and termination, which together capture individual features associated with clients. We present our computational techniques and results in Section 4.4, with model validation and an example of active intervention that our model can recommend. We conclude the chapter in Section 4.5 by summarizing the model and insights from our analysis.

4.2 Description of the Data

The data from the SEP consist of results from a survey and records of individual syringe exchange transactions. The transaction data were collected from January 2001 to November 2014 with 139,488 entries. The survey data were collected between July 2005 and November 2014 with 6,843 surveys. Each survey entry corresponds to a unique client. When a new client arrives at a service site, the client is assigned a unique study number (henceforth, client ID) and is asked to complete an enrollment survey. This client ID is then used throughout the client's sojourn in the system.

We use the data from July 2005 to November 2014 because the surveys are aligned consistently with the transaction records in this period. After removing incomplete and contaminated records, we combine the transaction and survey data to obtain a merged dataset with 63,960 entries, each with 50 data fields, which we detail in the following section.

4.2.1 Survey Data

The survey contains 31 questions, which yield 33 predictors, covering basic demographic information, ZIP code of residence, and the client's drug use habits, which can be categorized as follows:

- Basic personal information: age, ethnicity, gender;
- Length of time using/injecting drugs;
- Frequency of using/injecting drugs in the past 30 days;
- Frequency of reusing one's own syringes in the past 30 days;
- Frequency of sharing syringes with others in the past 30 days;

- Involvement in group injection in the past 30 days;
- Sources of syringes;
- Types of drugs used;
- Drug treatment program participation in the past six months;
- Reasons and frequency of being in the area with an SEP location in the past 30 days.

Among 5,903 clients in our merged dataset, 4,101 are male, 1,800 are female, and two are transgender. The demographics of these clients are summarized in Tables 4.1 and 4.2.

Ethnicity	Number of clients	Percentage (%)
White	3,045	51.58
African American	1,384	23.45
Puerto Rican	873	14.78
Mexican	364	6.17
Other Latino	72	1.22
Other	165	2.80
Total	5,903	100

Table 4.1: Ethnicity of clients

Age (years)	Number of clients	Percentage (%)
<15	1	0.02
$16-30$	2,602	44.08
$31-45$	2,108	35.71
$46-60$	1,111	18.82
>60	81	1.37
Total	5,903	100

Table 4.2: Age of clients

The average age of first drug use among all clients is 23.5 years, about 67.9% of clients inject drugs daily, and 81.7% injected drugs more than 20 out of 30 days before taking the survey. For all reasonable responses to the survey question regarding injection frequency (defined as at most 10 injections per day), the average number of injections per day is 2.77. About 82.6% of clients reported that they did not use someone else's syringe in the past 30 days, and 77.9% reported

that no one used their syringes over the same time period. Among 986 clients who injected with syringes others had used in the past 30 days, 558 shared with their spouse, 53 shared with a family member, 350 shared with a friend, 24 shared with an acquaintance, and 11 shared with a stranger. About 72.1% of the clients did not share cookers, cotton or water during injection, and 8.6% of the clients stated that they injected drugs in a shooting gallery within the 30 days before they began using the syringe exchange service. About 69.2% of the clients were in the neighborhood of the SEP location where they took the survey more than 20 days in a typical 30-day month, 25.1% of whom lived in the neighborhood with an SEP location, 54.5% of whom had come mainly to buy drugs, 2.0% of whom had come mainly to exchange syringes, and 4.6% of whom had come to visit friends. The voluntary nature of the survey could lead to a non-response bias in the results; see, for example, the illustrated cases and discussion in Locker et al. (1981), Cheung et al. (2017). Because every SEP transaction is labeled with a client ID, we can compute the fraction of clients for whom we have a completed survey, which is 98.6%. Discussions with SEP staff confirmed that although the survey is voluntary, almost every new client takes the survey upon the recommendation of SEP staff, and so the overall effect of non-response bias is likely small.

4.2.2 Transaction Data

The transaction data include details of syringe exchanges that occurred in multiple SEP storefronts and on mobile van routes. During each transaction, SEP employees record the client ID, number of syringes exchanged, number of other preventive devices distributed, size of the group coming with the client, and type(s) of health education material given to the client. In between July 2005 and November 2014, the SEP distributed 3,647,384 syringes, 160,895 male and female condoms, and 63,667 sets of educational material. The mean size of the group coming with the client during a single transaction was 1.89 with a standard deviation of 2.37. The mean number of syringes exchanged in one transaction was 57.03 with a standard deviation of 110.88. Our discussions with the SEP suggested that there are ample staff to process clients' visits, and that the SEP always had enough syringes and rarely ran out of other drug-injection equipment. Thus we see the level of censoring in the demand data as minimal.

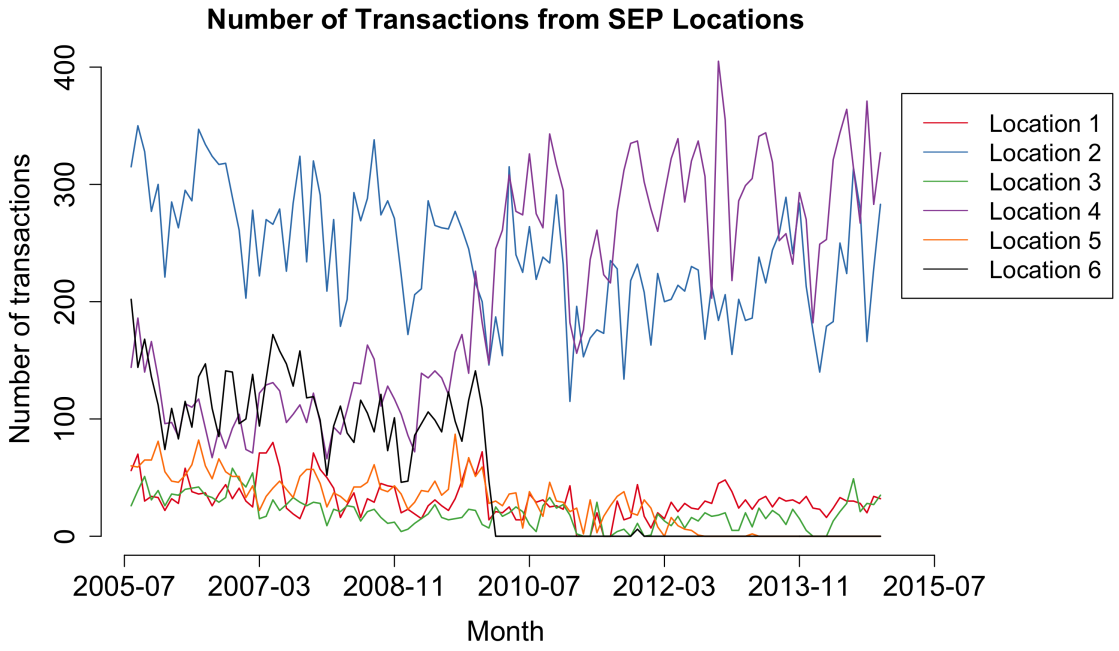


Figure 4.1: The time series of transactions from eight SEP service locations

We present a time series of the number of transactions from July 2005 to November 2014 in Figure 4.1. The figure shows transactions at four storefronts and mobile distribution in two areas, where a mobile van was dispatched to certain locations with a flexible schedule. According to SEP staff, the schedule for the van was communicated to clients during their visits and at some shooting galleries. Clients could also inquire about the schedule through phone calls. The service at Location 6 was terminated in January 2010, and some of its clients started visiting Location 4 afterward, consistent with the increasing trend for Location 4. Figure 4.1 also suggests a declining number of visits for Location 2. Figure 4.2 shows the fluctuation of monthly aggregated transactions. The figure suggests a slight decrease in the number of monthly transactions over the ten-year time span. The number of syringes exchanged surged between 2009 and 2012 but appeared to decrease afterward.

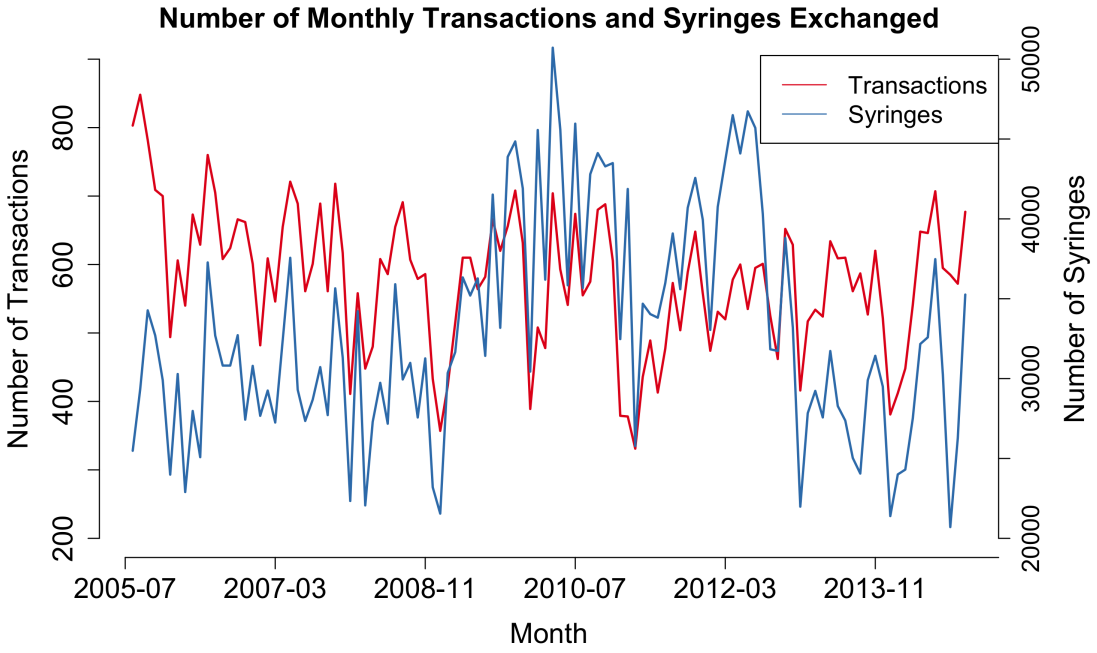


Figure 4.2: The time series of aggregated monthly transactions and syringes exchanged

4.3 Model of Client Arrival Process

We seek to develop a predictive model for a client's arrival process based on a number of predictors such as race, gender, age, and more. To this end, we segment a client's experience in the system into three sub-processes: initiation, reoccurring visits, and termination. Features of the client are used as covariates to estimate parameters of the model of reoccurring visits and termination. We can achieve two objectives with our model. First, we can forecast the next arrival of a specific client, given that client's features and most recent arrival time. Second, we can simulate the system and perform sensitivity analysis on specific model parameters. The former can help the SEP identify irregular behavior and take prompt intervention measures. The latter can guide initiatives to improve system-wide performance.

The overall structure of the model we formulate is that we build sub-models of these three individual sub-processes. While we have analytical models of these sub-components, our overall model, which combines these sub-models, can only be executed as a simulation, as we describe after

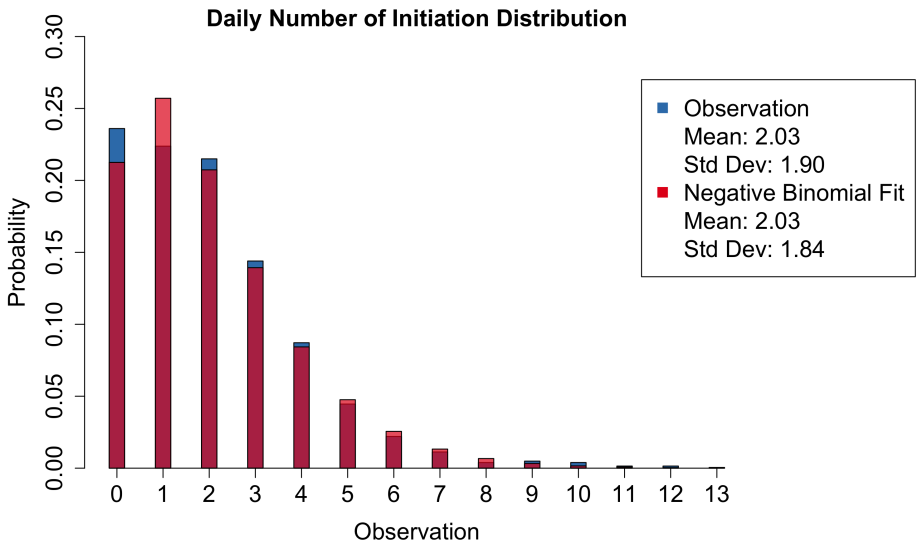


Figure 4.3: Empirical distribution fit of number of daily initiations to a negative binomial distribution

characterizing the sub-models.

4.3.1 Initiation

The first time that a client uses the SEP is called the *initiation* of the client's arrival process. The recorded initiation data are clear, and so we focus on how to simulate new initiations. We examine the distribution of the number of initiations per day, i.e., the arrivals generated by clients who have never previously visited a service location. Since our SEP started recording survey data four years after recording transactions, some returning clients were asked to complete the survey starting in July 2005, even though their true initiation was earlier. In an attempt to avoid inflating some initiation counts, we use data starting from January 2007 in order to fit a distribution to estimate the initiation process. The blue bars in Figure 4.3 show the empirical distribution of initiations per day. We fit a negative binomial distribution with parameter $(3, 0.59725)$, shown as the red bars in Figure 4.3. The negative binomial distribution can be seen as an over-dispersed version of a Poisson distribution and used to model discrete data whose sample variance exceeds the sample mean; see, e.g., Gardner et al. (1995). In Section 4.4.3, we detail goodness-of-fit measures for this and other distributional estimates.

For the purpose of simulation, each day we first generate a number of total new clients using the negative binomial distribution. In addition to simulating new arrivals of clients, we must assign attributes to those clients. In our simulation, we do so by drawing a client at random (with replacement) from the collection of 5,903 clients in our dataset. From the survey data, 33 numerical and categorical characteristics describe the client, and these are summarized in Appendix C.1.

4.3.2 Reoccurring Visits

We track the history of clients who visit the SEP service sites multiple times and plot the distribution of inter-arrival times. The inter-arrival time is defined here as the duration between two consecutive visits made by the same client. Figure 4.4 suggests that the distribution has a heavy tail, i.e., it shrinks to zero more slowly than an exponential.

As we indicate in Section 4.1, using a Poisson process to model inter-arrival times does not help with our main goal, which is to provide contextual, i.e., client-specific predictions. Putting this aside for a moment, we tested the goodness of fit associated with a Poisson process for our aggregate dataset, and for datasets associated with sub-populations of clients based on ethnicity, age, and gender. Moreover, we investigated both homogeneous and non-homogeneous (e.g., piecewise constant arrival rate by week) Poisson processes. For the models we assessed, statistical tests yielded p-values that were vanishingly small, suggesting that such models fail to provide an adequate representation. This is consistent with our observation from Figure 4.4, giving further evidence that modeling inter-arrival times using an exponential distribution may not be appropriate.

We model clients inter-arrival times with a phase-type distribution, because of its goodness of fit and its potential interpretability. The distribution of any nonnegative random variable can be approximated with high accuracy using a phase-type distribution; see, e.g., Asmussen (2003). A phase-type distribution can be expressed as the time required for a continuous-time Markov chain (CTMC) to enter an absorbing state (say, state 0) from a randomly selected transient state, $\{1, 2, \ldots, n\}$; see, e.g., Buchholz et al. (2014). A probability mass function, denoted by $\alpha = (\alpha_i)_{i=1,\ldots,n}$, governs the initial state of the CTMC. The infinitesimal generator is constructed

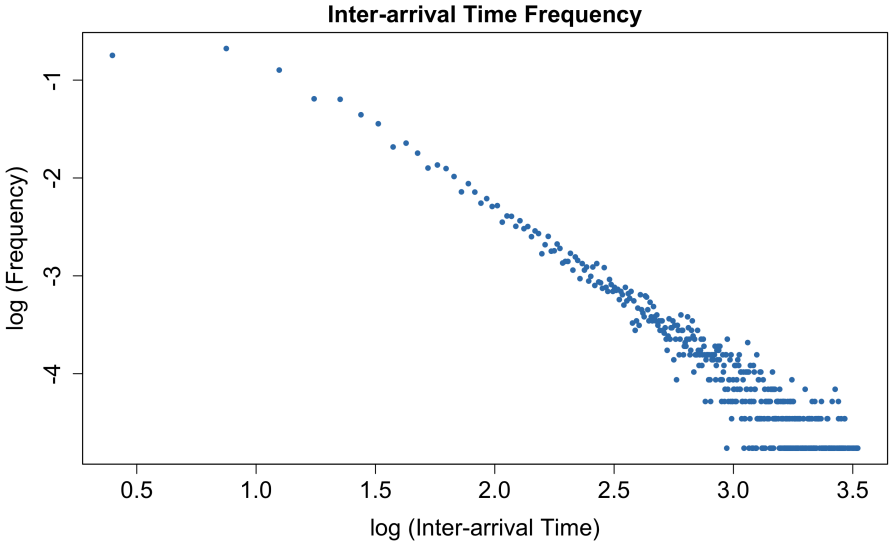


Figure 4.4: Log-log relationship between frequency and inter-arrival time. The logarithms in the figure are base 10, and underlying inter-arrival times are in days

in the following manner:

$$\begin{bmatrix} 0 & 0 \\ a & Q \end{bmatrix}, \tag{4.1}$$

where a is an n-dimensional vector specifying the transition rates from the transient states to the absorbing state, and Q is an $n \times n$ matrix specifying the transition rates among transient states, where $Q(i,j)$ denotes the rate of transitioning from state i to state j, and $Q(i,i) = -[\sum_{j \neq i, j=1,\ldots,n} Q(i,j) + a(i)]$. The first row in the generator of equation (4.1) corresponds to the transition rates from the absorbing state to any other transient state, which are always 0. The probability density function (pdf) of the phase-type distribution can be characterized as $f(t) = \alpha e^{Qt} a$, and the cumulative distribution function (cdf) is given by $F(t) = 1 - \alpha e^{Qt} \mathbf{1}$, where $\mathbf{1}$ is the n-dimensional vector of all 1's.

There are multiple ways to fit a phase-type distribution to data; see, e.g., Nelson and Gerhardt (2010). Here, we formulate a nonlinear optimization model rooted in maximum-likelihood estimation (MLE), coupled with a regression model that uses covariates of the clients. We first describe the MLE approach in the context of the Coxian distribution, a special case of phase-type

distributions.

For a Coxian distribution, the embedded Markov chain has $n+1$ states, as shown in Figure 4.5. The stochastic process starts in state 1 (i.e., $\alpha = (1, 0, \ldots, 0)$), and from each transient state, $i = 1, \ldots, n-1$, we can transition to only the adjacent transient state, $i+1$ or to the absorbing state, 0. The rate at which we depart state i is γ_i, and we transition to the absorbing state with probability q_i and to the adjacent transient state with probability $1-q_i$. As Figure 4.5 also depicts, transient state n can only transition to the absorbing state.

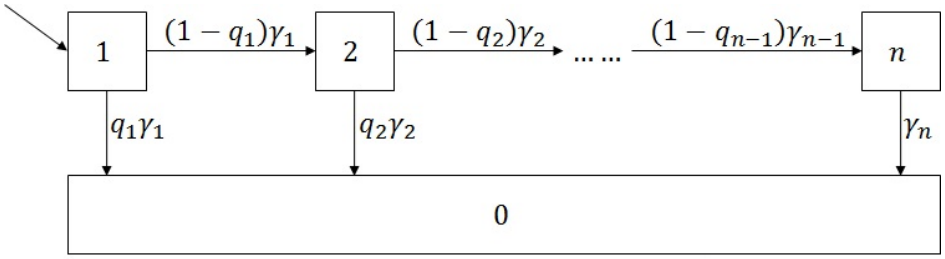


Figure 4.5: CTMC depiction of the Coxian distribution

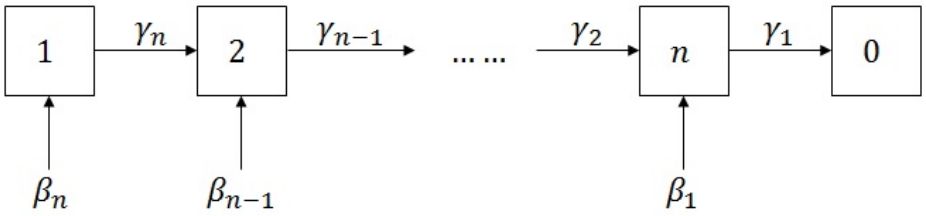


Figure 4.6: An equivalent CTMC to the Coxian distribution's model in Figure 4.5

A Coxian distribution is more parsimonious than a general phase-type distribution. The former model contains $2n-1$ parameters while the latter has up to n^2+n. Bobbio and Cumani (1992) show that the Coxian model in Figure 4.5 is equivalent to another CTMC model that is depicted in Figure 4.6. The latter formulation is helpful in our setting because it offers a linear structure for capturing client-specific features as we describe below. The relationship between the q_i parameters in the first model and the β_i parameters in the second model is given by the following equations:

$$q_0 = \beta_0 = 0$$

$$\beta_i = q_i \prod_{k=0}^{i-1}(1-q_k) \qquad i = 1, 2, \ldots, n$$

$$q_i = \frac{\beta_i}{1 - \sum_{k=0}^{i-1} \beta_k} \qquad i = 1, 2, \ldots, n.$$

A CTMC lends itself to interpretation by associating transitions in the model with assumed phases of a client using syringes after leaving a service location. In addition to references mentioned earlier which use CTMCs with hidden states (Chehrazi et al. 2019, Paddock et al. 2012, Liu et al. 2015), we note that such an approach has also been used to model the length of stay of hospital patients (Faddy et al. 2009, Faddy and McClean 1999, 2005), including work in which serial (Coxian) CTMCs are employed. We use a similar philosophy by inferring transition rates from unobservable states and, moreover, connecting them to features of a client.

We can interpret a Coxian distribution with $n = 2$ in our setting as follows. After visiting an SEP location, the client enters an (unobservable) "active state" with probability β_1, and subsequently returns to an SEP site after an exponentially distributed delay with parameter γ_1. Alternatively, with probability $\beta_2 = 1 - \beta_1$ the client enters a "passive state." Returning to an SEP site is then the sum of two independent exponential random variables with rates γ_1 and γ_2, where we expect $\gamma_1 > \gamma_2$. The passive state could correspond to the client temporarily seeking another source of syringes, for example.

Bobbio and Cumani (1992) present an MLE procedure to fit the parameters for the Coxian distribution. Their method, however, is not directly applicable when we express β and γ as affine functions of predictors associated with clients. A result of Bibinger (2013) allows us to express the pdf and cdf of a sum of independent exponential random variables, and we can use this result to write the pdf and cdf of a Coxian random variable as:

$$f(t) = \sum_{i=1}^{n} \beta_{n+1-i} f_i(t) = \sum_{i=1}^{n} \left[\beta_{n+1-i} \left(\prod_{l=1}^{n+1-i} \gamma_l \right) \sum_{j=1}^{n+1-i} \frac{e^{-\gamma_j t}}{\prod_{k=1, k\neq j}^{n+1-i} (\gamma_k - \gamma_j)} \right] \tag{4.2}$$

$$F(t) = \sum_{i=1}^{n} \beta_{n+1-i} F_i(t) = \sum_{i=1}^{n} \left[\beta_{n+1-i} \left(\prod_{l=1}^{n+1-i} \gamma_l \right) \sum_{j=1}^{n+1-i} \frac{(1 - e^{-\gamma_j t})/\gamma_j}{\prod_{k=1, k\neq j}^{n+1-i} (\gamma_k - \gamma_j)} \right]. \tag{4.3}$$

Here f_i and F_i are the pdf and cdf of the inter-arrival time, conditioned on beginning in state i. We index the collection of inter-arrival times by $\mathcal{S}$, and denote each inter-arrival time by t_s, $s \in \mathcal{S}$, and similarly denote censored inter-arrival times by t_u, $u \in \mathcal{U}$. For every client, the time from the last arrival to the end of the observation horizon can be considered a right-censored inter-arrival time. The likelihood function is the product of the pdf of each inter-arrival time and the complement of the cdf of each censored inter-arrival time. Maximizing the log-likelihood function then leads to the following problem:

$$\max_{\beta,\gamma \geqslant 0} \quad \sum_{s\in\mathcal{S}} \log(f(t_s)) + \sum_{u\in\mathcal{U}} \log\left(1 - F(t_u)\right) \tag{4.4a}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \beta_i = 1 \tag{4.4b}$$

$$f, F \text{ defined as in (4.2) and (4.3), } \forall t_s, s \in \mathcal{S} \text{ and } t_u, u \in \mathcal{U}, \text{ respectively.} \tag{4.4c}$$

The first term in the objective function of model (4.4) corresponds to observed inter-arrival times, and the second term corresponds to right-censored data in which we do not know the inter-arrival time, only that it exceeds, t_u; see, e.g., Papaioannou (2014) for such treatments of right-censored data. We note that a limiting analysis shows that equations (4.2)-(4.3) remain valid even when rates at distinct states are identical (Bibinger 2013). That said, this can cause numerical difficulties and we return to this issue below.

So far, the described fitting procedure assumes all clients behave according to the same model. As discussed above, we seek to incorporate the features of clients when we construct the parameters of the Coxian distribution. Here we use an affine relationship to connect those features to the parameters of the Coxian distribution. We use $\mathcal{V}$ to represent the set of clients, and we use $j = 1, 2, \ldots, m$ to index the characteristics of clients. We use $x_{j,v},\ \forall j = 1, \ldots, m,\ v \in \mathcal{V}$, to denote these predictors. We also use $v(s) \in \mathcal{V}$ to specify the client associated with the s-th inter-arrival time, and we similarly define $v(u)$ for the client associated with the u-th censored inter-arrival time. Our extension of model (4.4) to incorporate client-specific predictors is given by:

$$\max_{\beta,\gamma,b,g,\varepsilon} \quad \sum_{s\in\mathcal{S}} \log(f(t_s)) + \sum_{u\in\mathcal{U}} \log\left(1 - F(t_u)\right) - \eta^{\beta} \|\varepsilon^{\beta}\|_2^2 \tag{4.5a}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \beta_{i,v} = 1 \qquad \forall v \in \mathcal{V} \tag{4.5b}$$

$$\beta_{i,v} = \sum_{j=1}^{m} b_{i,j} x_{j,v} + b_{i,0} + \varepsilon_{i,v}^{\beta} \qquad \forall i = 1, 2, \ldots, n, v \in \mathcal{V} \tag{4.5c}$$

$$\gamma_{i,v} = \sum_{j=1}^{m} g_{i,j} x_{j,v} + g_{i,0} \qquad \forall i = 1, 2, \ldots, n, v \in \mathcal{V} \tag{4.5d}$$

$$f(t_s) = \sum_{i=1}^{n} \left[\beta_{n+1-i,v(s)} \left(\prod_{l=1}^{n+1-i} \gamma_{l,v(s)} \right) \sum_{j=1}^{n+1-i} \frac{e^{-\gamma_{j,v(s)} t_s}}{\prod_{k=1,k\neq j}^{n+1-i} \left(\gamma_{k,v(s)} - \gamma_{j,v(s)} \right)} \right] \qquad \forall s \in \mathcal{S} \tag{4.5e}$$

$$F(t_u) = \sum_{i=1}^{n} \left[\beta_{n+1-i,v(u)} \left(\prod_{l=1}^{n+1-i} \gamma_{l,v(u)} \right) \sum_{j=1}^{n+1-i} \frac{e^{-\gamma_{j,v(u)} t_u} / \gamma_{j,v(u)}}{\prod_{k=1,k\neq j}^{n+1-i} \left(\gamma_{k,v(u)} - \gamma_{j,v(u)} \right)} \right] \qquad \forall u \in \mathcal{U} \tag{4.5f}$$

$$\beta_{i,v} \geqslant 0 \qquad \forall i = 1, 2, \ldots, n, v \in \mathcal{V} \tag{4.5g}$$

$$\gamma_{i,v} \geqslant 0 \qquad \forall i = 1, 2, \ldots, n, v \in \mathcal{V}. \tag{4.5h}$$

The idea behind model (4.5) is that we have predictors associated with each client, and constraints (4.5c) and (4.5d) express the parameters of the Coxian model as an affine function of these predictors. Constraint (4.5b) replicates constraint (4.4b), and constraints (4.5e) and (4.5f) define the pdf and cdf terms that appear in the log-likelihood in the first two terms of the objective function.

Model (4.5) combines elements of regression and maximum likelihood estimation. The first two terms in the objective function maximize log-likelihood in the spirit of model (4.4). Constraints (4.5c)-(4.5d) define the regression model. Parameters $\beta_{i,v}$ and $\gamma_{i,v}$ are unobservable, and so, in principle, we could have no residual term in the regression model. However, constraint (4.5b) requires that the conditional probabilities that we return to states $1, 2, \ldots, n$ sum to 1. So, to maintain feasibility we add a residual, ε_v^{β}, in equation (4.5b) and penalize its two-norm using a positive weight η^{β} in the final term in the objective function (4.5a). While not explicit in model (4.5)'s statement, to help prevent overfitting, we regularize the regression parameters b and g by adding terms $-\eta^b \|b\|_2^2$ and $-\eta^g \|g\|_2^2$ to the objective function (4.5a).

While a limiting analysis shows the validity of equation (4.2)-(4.3) even when some of the components of γ are identical, allowing this when optimizing can cause numerical problems. Moreover, our motivation, sketched above, includes the idea that the sojourn times should be larger in

the passive state than in the active state, and for both reasons we add the following constraint to model (4.5):

$$\gamma_{i,v} - \gamma_{i+1,v} \geqslant \delta, \quad \forall i = 1, 2, \ldots, n-1, v \in \mathcal{V}, \text{ where } \delta > 0. \tag{4.6}$$

4.3.3 Termination

We assume that each client has a possibility to exit the system after each visit to the SEP. Specifically, as soon as the CTMC hits the absorbing state, we assume that with probability p_v, the client $v \in \mathcal{V}$ will stop visiting our SEP.

However, we cannot observe a client leaving the system because we only observe their visits. If a client visits service locations multiple times, then we know that for every visit before the last one, the client is still in the system, and so the likelihood function is conditioned on the client remaining in the system. After the last visit, a client may stay in the system or may leave. We need to incorporate this information in the likelihood function. Conditioned on the client remaining in the system, the likelihood function is $1 - F(t_u)$, where $t_u,\ u \in \mathcal{U}$, is the time between the client's last visit and the end of the observation horizon. As a result, the log-likelihood function can be revised as:

$$\sum_{s \in \mathcal{S}} \left(\log(f(t_s)) + \log(1 - p_{v(s)}) \right) + \sum_{u \in \mathcal{U}} \log \left((1 - F(t_u)) (1 - p_{v(u)}) + p_{v(u)} \right). \tag{4.7}$$

We again model parameter p_v via a functional relationship with client v's covariates. Instead of an affine relationship, we use a logistic function as follows:

$$p_v = \left(1 + e^{-(\rho_0 + \sum_{j=1}^{m} \rho_j x_{j,v})} \right)^{-1}, \tag{4.8}$$

where we will optimize the fit via parameters ρ_j and ρ_0. Given that $p_v \in (0, 1)$, the logistic function is a natural choice; we do note that we also tested a linear relationship but obtained poorer results.

We can fit the termination parameter p by combining the results of (4.5), (4.7), and (4.8). Since the likelihood function is conditioned on whether the client has exited the system, we need to solve a nonlinear optimization problem as follows, which fits parameters for both reoccurring visits

and termination, integrating our latter two sub-models:

$$\max_{\beta,\gamma,b,g,\rho,p,\varepsilon} \quad \sum_{s\in\mathcal{S}} \left(\log(f(t_s)) + \log(1-p_{v(s)})\right) + \sum_{u\in\mathcal{U}} \log\left((1-F(t_u))\,(1-p_{v(u)}) + p_{v(u)}\right) - \eta^{\beta}\|\varepsilon^{\beta}\|_2^2 \tag{4.9a}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \beta_{i,v} = 1 \qquad \forall v \in \mathcal{V} \tag{4.9b}$$

$$\beta_{i,v} = \sum_{j=1}^{m} b_{i,j}x_{j,v} + b_{i,0} + \varepsilon_{i,v}^{\beta} \qquad \forall i = 1,2,\ldots,n, v \in \mathcal{V} \tag{4.9c}$$

$$\gamma_{i,v} = \sum_{j=1}^{m} g_{i,j}x_{j,v} + g_{i,0} \qquad \forall i = 1,2,\ldots,n, v \in \mathcal{V} \tag{4.9d}$$

$$p_v = \left(1 + e^{-(\rho_0 + \sum_{j=1}^{m}\rho_j x_{j,v})}\right)^{-1} \qquad \forall v \in \mathcal{V} \tag{4.9e}$$

$$f(t_s) = \sum_{i=1}^{n}\left[\beta_{n+1-i,v(s)}\left(\prod_{l=1}^{n+1-i}\gamma_{l,v(s)}\right)\sum_{j=1}^{n+1-i}\frac{e^{-\gamma_{j,v(s)}t_s}}{\prod_{k=1,k\neq j}^{n+1-i}\left(\gamma_{k,v(s)} - \gamma_{j,v(s)}\right)}\right] \qquad \forall s \in \mathcal{S} \tag{4.9f}$$

$$F(t_u) = \sum_{i=1}^{n}\left[\beta_{n+1-i,v(u)}\left(\prod_{l=1}^{n+1-i}\gamma_{l,v(u)}\right)\sum_{j=1}^{n+1-i}\frac{e^{-\gamma_{j,v(u)}t_u}/\gamma_{j,v(u)}}{\prod_{k=1,k\neq j}^{n+1-i}\left(\gamma_{k,v(u)} - \gamma_{j,v(u)}\right)}\right] \qquad \forall u \in \mathcal{U} \tag{4.9g}$$

$$\beta_{i,v} \geqslant 0 \qquad \forall i = 1,2,\ldots,n, v \in \mathcal{V} \tag{4.9h}$$

$$\gamma_{i,v} \geqslant 0 \qquad \forall i = 1,2,\ldots,n, v \in \mathcal{V}. \tag{4.9i}$$

Solving problem (4.9) maximizes the log-likelihood function for the combination of reoccurring visits and termination. Similar to fitting the reoccurring visits, we also add a regularization term $-\eta^{\rho}\|\rho\|_2^2$ in the objective function to prevent overfitting. Constraint (4.9e) models the logistic relationship between the parameter p and the covariates x. Given the fit value of ρ^* and the features of client v, we can calculate the termination probability of that client, p_v.

4.4 Experimental Results

In this section, we first discuss how we solve model (4.9) and its simpler variants, along with preliminary results in which we do *not* use the covariates of the clients. Then we present the results of the fit model, provide insights as to how different features predict client behavior, test elements of model validity, and show an example of how our personalized arrival model can guide active intervention.

4.4.1 Computational Issues and Preliminary Results

We use a Coxian model with $n = 2$ transient states. We interpret one phase as the client being in an active state with the SEP, i.e., with more frequent visits to exchange syringes, and we interpret the other phase as a passive state with less frequent visits.

Model (4.9) and its variants are computationally challenging nonconvex optimization problems. We use Ipopt 3.12.1 (Wächter and Biegler 2006), with linear solver MA27, to solve instances of these optimization problems. Due to nonconvexity, we only obtain locally optimal solutions. In addition, because numerical issues can arise, we briefly sketch ways in which we "help" the solver.

We scale all continuous data, i.e., client predictor data, so that it is normalized. As discussed at the end of Section 4.3.2, we enforce $\gamma_{1,v} \geqslant \gamma_{2,v} + \delta$, and we use $\delta = 0.005$ in our computation. For numerical reasons, we also bound the γ-parameters away from zero, by enforcing $\gamma_2 \geqslant \underline{\gamma} \equiv 0.0005$.

We start by solving a simplified variant of model (4.9) in which we remove the predictors and directly optimize β, γ, and p. We do this for two reasons. First, it provides insight regarding typical values of these parameters, and second, as we discuss in further detail below, it helps provide a good initial solution for model (4.9). In particular we solve:

$$\max_{\beta,\gamma,p} \quad \sum_{s\in\mathcal{S}} \left(\log(f(t_s)) + \log(1-p)\right) + \sum_{u\in\mathcal{U}} \log\left((1-F(t_u))(1-p)+p\right) \tag{4.10a}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \beta_i = 1 \tag{4.10b}$$

$$f(t_s) = \sum_{i=1}^{n} \left[\beta_{n+1-i} \left(\prod_{l=1}^{n+1-i} \gamma_l \right) \sum_{j=1}^{n+1-i} \frac{e^{-\gamma_j t_s}}{\prod_{k=1,k\neq j}^{n+1-i} (\gamma_k - \gamma_j)} \right] \qquad \forall s \in \mathcal{S} \tag{4.10c}$$

$$F(t_u) = \sum_{i=1}^{n} \left[\beta_{n+1-i} \left(\prod_{l=1}^{n+1-i} \gamma_l \right) \sum_{j=1}^{n+1-i} \frac{e^{-\gamma_j t_u}/\gamma_j}{\prod_{k=1,k\neq j}^{n+1-i} (\gamma_k - \gamma_j)} \right] \qquad \forall u \in \mathcal{U} \tag{4.10d}$$

$$\gamma_i - \gamma_{i+1} \geqslant \delta \qquad \forall i = 1, 2, \ldots, n-1 \tag{4.10e}$$

$$\beta_i \geqslant 0 \qquad \forall i = 1, 2, \ldots, n \tag{4.10f}$$

$$\gamma_n \geqslant \underline{\gamma} \tag{4.10g}$$

$$0 \leqslant p \leqslant 1. \tag{4.10h}$$

Solving model (4.10) leads to parameters for a "featureless" client, as follows:

$$\hat{\beta}_1 = 0.8194,\ \hat{\beta}_2 = 0.1806$$

$$\hat{\gamma}_1 = 0.0520,\ \hat{\gamma}_2 = 0.0030$$

$$\hat{p} = 0.0981.$$

This result suggests that after each visit, the featureless client has a 9.8% chance of exiting the SEP system. Conditional on the client visiting an SEP site again, the client returns via the active state (frequent visits) with a probability of about 0.82. The mean time from this state is $1/\gamma_1 \approx 19$ days. With probability about 0.18, the client returns via the passive state, and the expected time to visit the SEP is then $1/\gamma_2 + 1/\gamma_1 \approx 350$ days.

Rather than optimizing over the intercept terms, b_0, g_0 and ρ_0 in model (4.9), we fixed these terms as:

$$b_{1,0} = \hat{\beta}_1 \quad b_{2,0} = \hat{\beta}_2$$

$$g_{1,0} = \hat{\gamma}_1 \quad g_{2,0} = \hat{\gamma}_2$$

$$\rho_0 = \ln\left(\frac{\hat{p}}{1-\hat{p}}\right).$$

Fixing the intercept terms in this way allows us to interpret parameters $b_{i,j}$, $g_{i,j}$, and ρ_j for $j = 1, 2, \ldots, m$ as deviations from the featureless client. Moreover, fixing these parameters helps improve the numerical performance of Ipopt when solving the nonconvex problem, in part by effectively providing a good initial solution.

After adding the regularization terms for b, g and ρ described in Section 4.3.2 and 4.3.3, and fixing the value of b_0, g_0 and ρ_0, we solve the following nonlinear program:

$$\max_{\beta,\gamma,b,g,\rho,p,\varepsilon} \quad \sum_{s\in\mathcal{S}} \left(\log(f(t_s)) + \log(1-p_{v(s)})\right) + \sum_{u\in\mathcal{U}} \log\left((1-F(t_u))\,(1-p_{v(u)}) + p_{v(u)}\right)$$
$$- \eta^\beta \|\varepsilon^\beta\|_2^2 - \eta^b \|b\|_2^2 - \eta^g \|g\|_2^2 - \eta^\rho \|\rho\|_2^2 \tag{4.11a}$$
$$\text{s.t.} \quad \sum_{i=1}^{n} \beta_{i,v} = 1 \qquad \forall v \in \mathcal{V} \tag{4.11b}$$

$$\beta_{i,v} = \sum_{j=1}^{m} b_{i,j} x_{j,v} + b_{i,0} + \varepsilon_{i,v}^{\beta} \qquad \forall i = 1, 2, \ldots, n, v \in \mathcal{V} \tag{4.11c}$$

$$\gamma_{i,v} = \sum_{j=1}^{m} g_{i,j} x_{j,v} + g_{i,0} \qquad \forall i = 1, 2, \ldots, n, v \in \mathcal{V} \tag{4.11d}$$

$$p_v = \left(1 + e^{-(\rho_0 + \sum_{j=1}^{m} \rho_j x_{j,v})}\right)^{-1} \qquad \forall v \in \mathcal{V} \tag{4.11e}$$

$$f(t_s) = \sum_{i=1}^{n} \left[\beta_{n+1-i,v(s)} \left(\prod_{l=1}^{n+1-i} \gamma_{l,v(s)} \right) \sum_{j=1}^{n+1-i} \frac{e^{-\gamma_{j,v(s)} t}}{\prod_{k=1,k\neq j}^{n+1-i} \left(\gamma_{k,v(s)} - \gamma_{j,v(s)}\right)} \right] \qquad \forall s \in \mathcal{S} \tag{4.11f}$$

$$F(t_u) = \sum_{i=1}^{n} \left[\beta_{n+1-i,v(u)} \left(\prod_{l=1}^{n+1-i} \gamma_{l,v(u)} \right) \sum_{j=1}^{n+1-i} \frac{e^{-\gamma_{j,v(u)} t_u} / \gamma_{j,v(u)}}{\prod_{k=1,k\neq j}^{n+1-i} \left(\gamma_{k,v(u)} - \gamma_{j,v(u)}\right)} \right] \qquad \forall u \in \mathcal{U} \tag{4.11g}$$

$$\gamma_{i,v} - \gamma_{i+1,v} \geqslant \delta \qquad \forall i = 1, 2, \ldots, n-1, v \in \mathcal{V} \tag{4.11h}$$

$$\beta_{i,v} \geqslant 0 \qquad \forall i = 1, 2, \ldots, n, v \in \mathcal{V} \tag{4.11i}$$

$$\gamma_{n,v} \geqslant \underline{\gamma} \qquad \forall v \in \mathcal{V} \tag{4.11j}$$

$$b_{i,0} = \hat{\beta}_i, g_{i,0} = \hat{\gamma}_i, \rho_0 = \ln\left(\frac{\hat{p}}{1-\hat{p}}\right) \qquad \forall i = 1, 2, \ldots, n. \tag{4.11k}$$

With modest tuning effort, we select the following weights on the regularization terms:

$$\eta^{\beta} = 100,\ \eta^{b} = 100,\ \eta^{g} = 1000,\ \eta^{\rho} = 10.$$

4.4.2 Results and Analysis

The results from fitting the parameters using the method in Section 4.4.1 are displayed in Table 4.3. The estimators are represented by ρ, b, and g. A positive $\rho_j, j = 1, 2, \ldots, m$, indicates that having feature j increases the probability that the client will exit the system. Given that the client stays in the system, a positive value of parameter $b_{1,j}$ increases the probability that the client returns to the active state. We do not report $b_{2,j}$ because its coefficient differs from $b_{1,j}$'s by a sign. Positive coefficients $g_{1,j}$ and $g_{2,j}$ lead to increased frequencies, i.e., shorter mean times, associated with the active and passive states, respectively. The coefficients in Table 4.3 are given in either regular font or gray font. The former category is significant, and the latter is not, where "significant" is defined as having at least 90% of bootstrapped replications having the same sign, as detailed in Appendix C.2.

Column Δ in Table 4.3 shows the amount by which the conditional expected inter-arrival time

	Factor (j)	ρ_j	$b_{1,j}$	$g_{1,j}$	$g_{2,j}$	Δ	T
	Gallery	0.5448	-0.0387	0.0027	0.0003	5.9	-268.2
	From Other Locations	0.2047	-0.0156	0.0023	0.0004	-3.2	-162.2
	From Other SEP	-0.3156	0.0248	-0.0063	0.0000	-6.2	186.3
	From Friends	-0.1194	-0.0045	-0.0011	-0.0007	20.9	330.1
	From Strangers	-0.2697	0.0142	0.0020	0.0010	-19.3	-26.8
	Speedball	-0.0293	-0.0704	0.0146	-0.0008	48.1	524.9
	Heroin	-0.0437	-0.0258	0.0061	-0.0008	31.4	365.6
	In Treatment	-0.3254	0.0243	-0.0126	0.0002	-4.7	215.7
	Been in Treatment	-0.1137	-0.0178	0.0063	-0.0001	5.9	154.8
	Female	0.0154	0.0217	-0.0076	0.0008	-14.8	-159.8
	White	0.0000	0.0218	-0.0005	0.0002	-9.9	-101.0
	African American	0.3623	-0.0073	-0.0026	0.0001	1.6	-209.6
	Puerto Rican	-0.7594	0.0084	0.0101	0.0001	-7.4	673.3
	Mexican	-0.3383	0.0014	0.0014	0.0010	-15.6	76.0
	Other	-0.1994	0.0088	-0.0054	0.0000	-0.7	151.9
	Age of First Drug Use	1.1219	0.0117	0.0007	0.0003	-8.8	-525.7
	Drug Use Span	1.8506	-0.0042	-0.0030	-0.0001	4.4	-602.3
	FUD	0.0332	0.0073	-0.0011	-0.0001	0.4	-19.6
	FROS	0.0380	-0.0051	0.0000	0.0000	1.7	-10.5
	FBSA	-0.1968	-0.0033	0.0058	0.0002	-5.2	95.4

Note: Here, ρ_j, $b_{1,j}$, and $g_{1,j}/g_{2,j}$ are factor-specific regression coefficients for the probability of exiting the system, probability of returning to the active state, and mean transition times in the CTMC, respectively. Based on the factor in each row, parameter Δ denotes the amount by which the conditional expected inter-arrival time changes, and T similarly denotes changes in the system sojourn time, both in days.

Table 4.3: Fitted parameters of the Coxian process

changes (in days) if a client has that row's feature but is otherwise a featureless client. Column T similarly shows the magnitude by which the expected sojourn time in the system changes due to a single feature. For context, the mean sojourn time of a featureless client is 806.5 days. The acronyms FUD, FROS, and FBSA in the table respectively stand for frequency of using drugs, frequency of reusing own syringes, and frequency of being the area of an SEP location.

Table 4.3 provides information on how a client's attributes affect the probability the client leaves the system and the frequency with which the client makes use of SEP services, even though the factors are from different categories, e.g., type of drugs the client uses versus where the client obtains syringes versus ethnicity. Some observations from the table include:

1. If the client attends a shooting gallery, it is more likely for the client to exit the system or

become passive. The former factor dominates in that the overall expected time in the system decreases relative to a featureless client (T).

2. Clients who can obtain syringes from other locations, such as pharmacies, are more likely to exit the system, perhaps because they are not reliant on SEP services. On the other hand, if the client obtains syringes from other sources (other SEPs, friends, and strangers), which may not be as reliable, it is more likely for the client to remain in the system.
3. The b_1-coefficient associated with speedball (a type of drug mixing cocaine with heroin or morphine) is strongly negative, meaning the client is less likely to stay in the active state, leading to an increase in expected inter-arrival time.
4. A client in a treatment program is more likely to stay in the SEP system, as indicated by a negative value of ρ and a positive value of T.
5. A female client is more likely to visit frequently, but remain in the system for a shorter period of time.
6. The probability of exiting the system differs significantly according to ethnicity: African-American clients are more likely to exit the system while the opposite is true for Puerto Rican and Mexican clients.
7. Not surprisingly, clients who are more frequently near an SEP site (i.e., have larger values of FBSA) are less likely to leave the system, are more likely to visit a site frequently, and overall have longer sojourn times in the system.

Such observations from the model we fit may allow our SEP to tailor promotion of their services to specific target populations. For example, the starkly different behavior of African-American clients may warrant special attention from the SEP in early encounters, relative to PWID of other ethnicities. Bean (1993) states that African-American drug users are less likely to inject drugs than White drug users. However, among all clients, it is not clear why African Americans are less likely to seek the syringe exchange services offered by our SEP. Investigating whether African Americans are more likely to quickly abandon injection drug use, versus continue use but not seek SEP services, would likely be needed to guide such strategies. Our results show that it would be beneficial to increase the frequency of clients being in the service location area, and one possible solution is to

increase the frequency of the mobile van service in certain locations to improve accessibility.

4.4.3 Model Validation via Simulation

The model of Section 4.3 quantifies client behavior. In this section, we perform statistical tests and compare results of the simulation outputs with observed data in order to assess model validity.

We begin with further details on the implementation of the simulation model. We simulate the arrival of clients to the SEP over 2,310 days, equivalent to the number of days that our SEP was open between July 2005 and November 2014. We also simulate an initial 5,000 days as a warm-up period, since our SEP was established more than 15 years before 2005. We assume SEP staff are always available to serve a visiting client, there are no shortages of syringes or other resources that would alter client behavior, and the location and operating schedule of storefronts and mobile vans are fixed. In other words, we assume the nature of client visits is governed solely by the features of the clients, in a manner consistent with historical data, and is not affected by service or resource availability. This matches our understanding of the actual system, as we discuss in Section 4.2. In our simulation of 7,310 days in total, for each day we first simulate the number of new clients according to the negative binomial distribution. We assign features to these clients by drawing a client at random, with replacement, from the list of 5,903 clients described in Section 4.2. Given the features of a client, x, and the parameters, ρ, b, and g, obtained from model (4.11), we can calculate the parameters of the CTMC, $\beta_i, \gamma_i,\ i = 1, 2$, and p, by:

$$\beta_i = \sum_{j=1}^{m} b_{i,j} x_j \qquad i = 1, 2 \tag{4.12a}$$

$$\gamma_i = \sum_{j=1}^{m} g_{i,j} x_j \qquad i = 1, 2 \tag{4.12b}$$

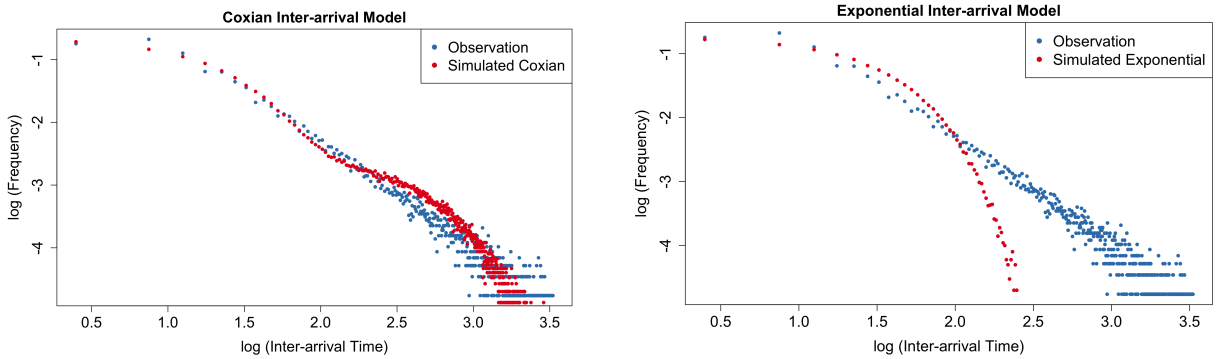
$$p = \left(1 + e^{-(\rho_0 + \sum_{j=1}^{m} \rho_j x_j)}\right)^{-1}. \tag{4.12c}$$

With the CTMC built for every client, upon the arrival of a specific client, we simulate whether this client exits the system using p from equation (4.12c). And, if the client does not exit the system, we simulate the time of the next arrival using the β and γ values from equations (4.12a)-(4.12b). After the warm-up period and simulating the 2,310 days of interest, we obtain summary statistics

as outputs based on inter-arrival times, sojourn times, and number of visits of each client.

We perform statistical tests to assess the quality of our sub-models for initiation, reoccurring visits, and termination. For initiation, we selected the negative binomial distribution for reasons that we discuss in Section 4.3.1. Using a Pearson's chi-squared goodness-of-fit test we obtain a p-value of 0.250, suggesting that we should not reject the null hypothesis that the data are consistent with the fit distribution. For comparison, we also fit other commonly used distributions (geometric, binomial, uniform, Poisson, and hyper-geometric), and we performed the same goodness-of-fit tests. None of those distributions had a p-value that exceeded a 0.05 level of significance.

For reoccurring visits, we compare the distribution of inter-arrival times obtained from the simulation model with an exponential distribution with a mean of 67.50 days, which is the mean of observed inter-arrival times. Figure 4.7 illustrates such a comparison. Results from the simulated Coxian process are shown on the left, and those from the exponential distribution are on the right, both in red. In addition, we plot actual observations in blue. The figure suggests that our Coxian-based simulation model provides a better fit to the observed data.



(a) Coxian Inter-arrival Model

(b) Exponential Inter-arrival Model

Figure 4.7: Part (a) of the figure shows the log-log relationship between frequency and the Coxian inter-arrival time model. Part (b) shows the analogous relationship for the exponential inter-arrival time model. In both subplots the (simulated) model values are shown with red dots and observed data are shown with blue dots. The logarithms are base 10 and the underlying inter-arrival times in days.

Since we do not directly observe whether a client has left the system, we test the right-censored sojourn time of clients in the system. We run a Pearson's chi-squared procedure to test the null

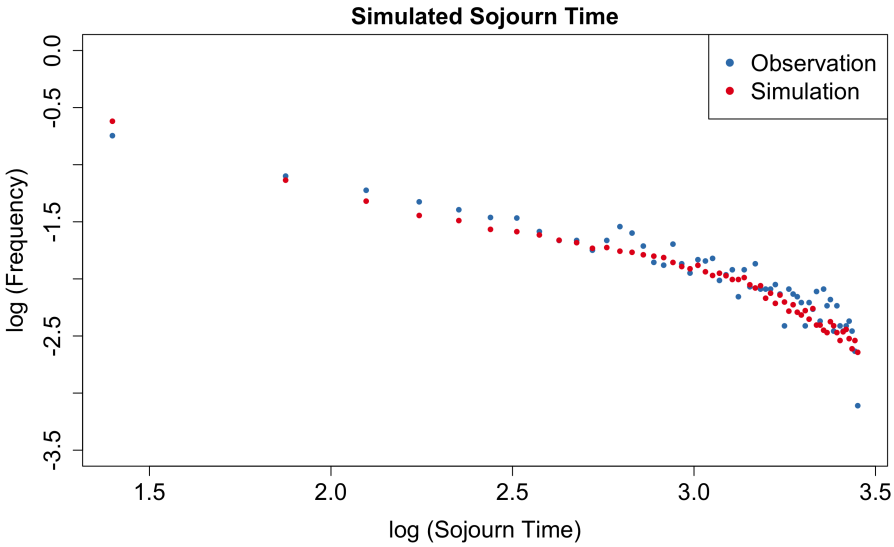


Figure 4.8: The log-log (base 10) relationship between frequency and simulated sojourn time is shown with red dots, and the same relationship between frequency and observed sojourn time is shown with blue dots.

hypothesis that the simulated distribution is consistent with observed data. The p-value of the test is 0.6262, which suggests that we should not reject the null hypothesis. Figure 4.8 suggests that the simulated sojourn times from the simulation appear consistent with the observation data.

In addition to assessing the validity of our three sub-models, Table 4.4 suggests that our simulation model can accurately capture demographic information such as ethnicity. The second and the third columns represent new clients by ethnicity. These columns are, of course, very close because we simply draw from the observed set of clients. The values help give a sense of variability due to sampling. The two right-most columns are the percentage of visits to SEP sites based on ethnicity. The consistency of these simulated values with observed percentages hinges on the Coxian model accurately capturing ethnicity-based inter-arrival times and exits from the system. For example, the simulation model captures well the lower values for African Americans, and larger values for Puerto Ricans, relative to their rates of initiation in the system.

Ethnicity Group	Initiation		SEP Locations Visits	
	Observed	Simulated	Observed	Simulated
White	51.58%	51.80%	47.53%	48.05%
African American	23.45%	23.66%	16.36%	15.99%
Puerto Rican	14.78%	14.21%	24.16%	23.62%
Mexican	6.17%	6.23%	7.64%	8.00%
Other Latino	1.22%	1.17%	1.14%	1.17%
Other	2.80%	2.93%	3.17%	3.16%

Table 4.4: Comparison of the observed percentage of number of clients for each ethnicity group, the simulated percentage of number of clients for each ethnicity group, the observed percentage of number of arrivals for each ethnicity group, and the simulated percentage of number of clients in each ethnicity group

4.4.4 Guiding Active Intervention

Our simulation model can facilitate analysis to provide SEP staff with insights regarding: (i) specific clients who are likely to enter a passive state or exit the system, and (ii) dispatch policies for the mobile van. We discuss both of these in turn.

4.4.4.1 Simple Client-Specific Intervention

While not immediate from Table 4.3, our discrete-event simulation model can be used to estimate that a 40-year old Puerto Rican male, with a history of using drugs for 21 years, who injects heroin 10 times per day, wants treatment, uses syringes after others once every 30 days, and frequents the area of an SEP site, has a 95% chance of having entered the passive state if he does not visit a service location within 56 days, assuming he has not already terminated contact with the SEP system. As a result, SEP staff could send a text message to such a client as a reminder if he has not returned within two months.

4.4.4.2 Intervention with Mobile Van Dispatch

The value of actively reaching out to clients based on insights from our simulation model may be further enhanced by *mobile* exchange of syringes. Here, we simulate mobile van dispatch, with a personalized notification push, to show its potential to improve current SEP operations. The corresponding simulation model has the following constructs and assumptions.

Clients. We simulate the initiation and reoccurring visits of clients in the same way described in Section 4.4.3 with the following exception: We assume that when a client exits the system, there is a probability, denoted by p_r, that the client is eligible to return to the system with active intervention. We say that a client is *at risk* if they reuse syringes, either their own or the syringes of another, and we note that 22.4% of our 5,903 clients are at risk.

Mobile van. We assume the SEP has one van, and each weekday the van is dispatched to one of five ZIP codes. We further assume that any client that the SEP contacts within a five-mile radius of that ZIP code is eligible to be served by the van. We selected the five ZIP codes by solving a facility-location model, which maximizes coverage of at-risk clients. The van visits each of these five ZIP codes in turn, Monday-Friday, each week over the simulation horizon.

Risky behavior. We assume only at-risk clients engage in risky behavior. And, we assume an at-risk client does *not* engage in risky behavior if the client is in the CTMC's active state, but otherwise the client does exhibit risky behavior.

Intervention. The SEP cannot observe a client's state or behavior, and hence intervention decisions are made knowing the time since the client's last visit and the client's predictors. In particular, γ_1 is the rate associated with the exponential distribution governing the return time, *if* the client is in the active state. We assume the SEP contacts a client if: (i) the time since the last visit exceeds the 0.9-level quantile for the active state's exponential return-time distribution, and (ii) the client's ZIP code is within five miles of the ZIP code the van is visiting that day.

Re-engaging a client. An intervention can be successful if the client is in the active state, passive state, or has exited but is eligible to return (with probability p_r). Among these clients, we let p_s denote the probability that a contacted client will visit the van, and hence re-engage with the SEP. If a client ignores the notification, we assume the client stays in the same state: active, passive, or exited. We further assume the SEP stops contacting a client after *three* notification attempts have been ignored.

Using the simulation model, we compare active intervention with current SEP operations, and we estimate the relative effectiveness by examining: (i) the number of additional arrivals to the system, and (ii) the number of times an SEP intervention re-engages a client who would otherwise

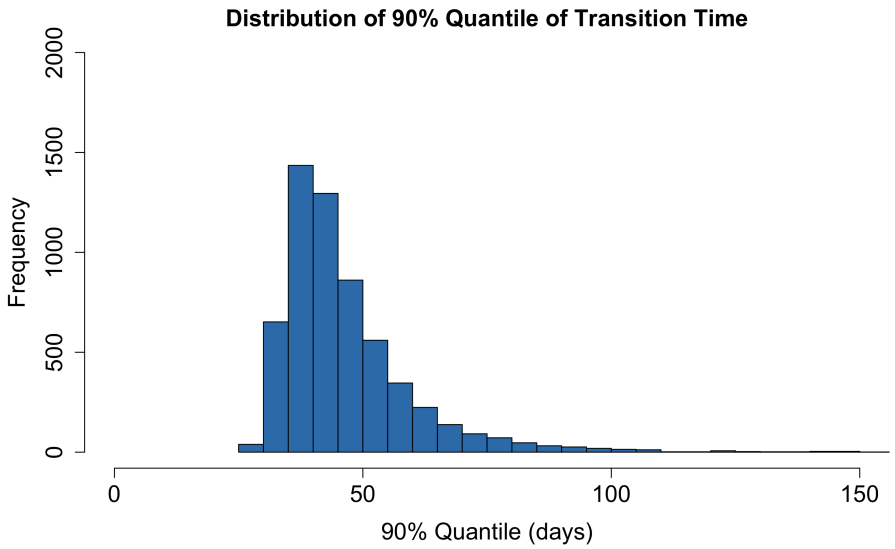


Figure 4.9: Histogram of the 90% client-specific quantile values for the return-time distributions from the CTMC's active state for the 5,903 surveyed clients.

be engaging in risky behavior. We estimate p_r as follows. The 0.975-level quantile of the observed inter-arrival time is 552 days, and we use this as a proxy for whether the client has exited the system. We denote the number of inter-arrival times exceeding 552 days by N_r, and the number of clients who have not visited any service location after 552 days by N_e. We estimate p_r via:

$$p_r = \frac{N_r}{N_r + N_e}.$$

With our data, $N_r = 1354$ and $N_e = 4281$, and so $p_r \approx 0.24$. As we indicate above, we use the 0.9-level quantile for client-specific return-time distributions from the active state to make intervention decisions, and Figure 4.9 shows a histogram of these quantiles among the surveyed clients.

We run the simulation model under two different system designs: (i) with active intervention via van dispatch and client notification, and (ii) without active intervention, approximating current SEP operations. The results are shown in Table 4.5 as we parametrically vary the success probability, p_s. After the warm-up period of 5,000 days, the average number of daily arrivals in the last 2,310 days has a mean of 24.6 and a 95% confidence interval halfwidth of 0.9. For each

value of p_s in the table, we run the simulation 20 times and present 95% confidence intervals. The "Total" column under "No. of Interventions" shows the number of times over 2,310 days that a notified client visited the mobile van, and the "Risky" column indicates the subset of those successful interventions involving at-risk clients in the passive or exited state.

p_s	Average Daily Arrivals	No. of Interventions	
		Risky	Total
0	24.6 ± 0.9	0	0
0.01	27.1 ± 0.9	79.1 ± 16.4	398.3 ± 37.3
0.03	28.2 ± 1.0	254.5 ± 38.4	1251.6 ± 86.2
0.05	29.7 ± 0.9	439.1 ± 54.0	2171.9 ± 117.7
0.07	31.0 ± 1.1	641.5 ± 66.4	3173.0 ± 162.4
0.09	32.4 ± 1.1	852.3 ± 95.9	4235.3 ± 217.8

Table 4.5: Simulation results of nominal SEP operations ($p_s = 0$) and SEP with active intervention via van dispatch and client notification. Here, p_s denotes the success probability associated with client notification, and the $\pm$ values represent 95% confidence interval half widths.

Under current SEP operations ($p_s = 0$), we see about 56,800 client arrivals over the 2,310-day horizon. Table 4.5 shows that a success probability of $p_s = 0.05$ leads to about one intervention per day over that time horizon, i.e., less than 4% of the nominal total number of arrivals would be the immediate result of an intervention. However, because a successful intervention leads to a re-engaged client returning to the active state, we see a significant 20% growth of about five arrivals per day relative to the $p_s = 0$ case. And, this corresponds to about one client per week who stopped risky behavior because of the intervention. Having only one client per day visiting the van may seem low, but a regularly scheduled van would also attract other clients not included in our model. More importantly, a 20% growth in average daily arrivals would be a welcome improvement countering some of the trends we point to in Section 4.2.2. We see these results as suggesting that there is value in the SEP investigating an active intervention scheme similar to our example. Moreover, our simulation model makes it possible to exploit the contextual nature of our model of inter-arrival times to determine a personalized threshold for push notification.

4.5 Conclusions

In this chapter, we examine the survey and transaction data of one major syringe exchange service provider in the Chicago metropolitan area. We find there is a discrepancy between a slightly decreasing trend in the number of client transactions with our SEP and an increasing number of heroin users in Chicago and the United States. We also discover significant differences in the behavior of clients in terms of how they engage with the SEP based on demographic attributes and further personal characteristics. Based on our observations, we focus on producing personalized predictions for clients that can aid the SEP in improving the system such as intervention initiatives for clients with certain attributes.

Standard stochastic models, such as Poisson processes, fail to accurately capture the observed inter-arrival process. Therefore, we formulate a CTMC-based simulation model to represent a client's path through the system. Our model consists of three sub-models: initiation, reoccurring visits, and termination, with their parameters learned from linear and logistic regression models integrated into the procedure by which we estimate the model's parameters. With the aid of this model, SEP staff and researchers can analyze the system parameters to draw useful conclusions for groups with different traits, so that proper actions can be taken towards a specific target group, or even the individual PWID. The quantitative model, combined with the personal interaction with each client can inform SEP staff of timely intervention opportunities. Such personalized recommendations may be particularly useful when the SEP faces challenges in tracking a large number of clients. Our simulation model can also help SEP staff evaluate the effectiveness of candidate initiatives.

In the future, our method can be enhanced by finding an algorithm to fit the parameters for higher fidelity Markov chain models since our optimization model for parameter estimation depends on the closed-form representation of the Coxian distribution. Other functional forms for the predictive models can also be integrated into the fitting procedure. Optimization over the location and route of the van can be investigated using our simulation platform. Further sensitivity analysis can also be performed to evaluate other initiatives beyond dispatching the mobile van.

Chapter 5

Conclusions

Stochastic disruptions can adversely affect many systems in a severe manner, but modeling them via mathematical optimization has not been systematically studied in the past. This dissertation focused on the case of a single disruption, and has developed a class of stochastic optimization models—a stochastic mixed integer program in Chapter 2 and a robust optimization model with convex recourse in Chapter 3—for stochastic disruptions. For the corresponding problems, the dissertation presented enhanced decomposition algorithms based on cutting-plane methods, and achieved superior computational performance compared to other state-of-the-art algorithms. The research presented in this dissertation can find application in various fields, such as project management, energy systems, disaster relief coordination, and public health policy.

5.1 Research Contributions

Our specific contributions include the following:

- We have established a modeling framework and specific modeling concepts for sequential decision problems under stochastic disruptions.
- We have developed a two-stage stochastic mixed integer program, in which the timing of the stage is random, to model the project crashing problem under a single stochastic disruption. We have showed that solutions to this model outperform alternatives, but is also NP-hard.
- We have proposed decomposition algorithms that sequentially tighten the linear programming

relaxation in a cutting-plane process by refining partitions on continuous first-stage variables. Convergence of our algorithm has been shown, and we have further shown the superior computational performance of our algorithm compared to solving the extensive formulation.

- We have developed a two-stage robust optimization model with convex recourse for the ACOPF problem under uncertainty, with a consideration of uncertainty from demand and renewable energy generation.
- For the robust ACOPF model, we have presented a cutting-plane algorithm and proved that it converges to an ε-feasible solution in a finite number of iterations. We have further provided a scenario-generation algorithm that significantly improves the computational performance compared to a naive implementation of the cutting-plane algorithm.
- We have examined the transaction and survey data of a syringe exchange program (SEP) in Chicago between 2005 and 2014, and we have modeled the arrival process of a client with three sub-models. Those sub-models integrate stochastic processes with regression and provide insights on how the demographic features and drug-use behavior of clients can affect their arrival frequency. We have tested the statistical validity of our model and presented an example of how our model can assist active intervention initiatives for SEPs.

5.2 Future Work

There are several potential extensions of our optimization models under stochastic disruptions. Higher-fidelity models and effective algorithms are required for future applications in real-world problems. The development of our models and algorithms can also help improve optimization problems with mixed integer decision variables in a more general setting. Specifically, we discuss the potential development in the following two major areas:

- **Multi-disruption models and algorithms:**

  The modeling framework presented in Section 1.2 shows the scenario tree of a sequential decision problem under a single stochastic disruption. It is possible to extend the model to incorporate more than one disruption. Suppose there are at most k disruptions during a fixed

time horizon. The stochastic programming model becomes a $(k+1)$-stage problem and, under appropriate convexity assumptions, stochastic dual dynamic programming method could be used to solve such a model. However, there are more opportunities to share cuts between nodes of the scenario tree in the same time period and the number of value functions that we need to approximate is significantly smaller than the full multi-stage stochastic programming model. An effective decomposition algorithm and further numerical tests will be valuable for more pervasive use of our stochastic disruption model.

It is interesting to compare the computational performance, solution, and optimal value obtained from such a multi-disruption model and those from a full multi-stage stochastic programming model. Similar to the work in Section 2.6, evaluating the policy generated by alternative models against established probabilistic models of certain types of disruptions is important to further justify the value of modeling disruptions.

The multi-disruption model can be applied to the operation of an electricity distribution system under stochastic contingencies. A contingency, such as an outage of a line or a transformer, can occur at a random time, and take a random period of time to fix. Assuming that the number of such contingencies is limited, we can combine the results from Chapter 2 and Chapter 3 to formulate a multi-stage convex program.

- **Identification of important disruption scenarios:**
  As stated in Chapter 3, a scenario-appending algorithm can be effective when solving a robust optimization problem with convex recourse. In this dissertation, we either solve the recourse problem at every extreme point of the uncertainty set or solve an MISOCP, and select the most violated scenario to append to the master problem. There would be value in work to identify such "most violated" scenarios efficiently without solving complex optimization problems. Identifying such scenarios without heavy computational costs could also benefit a solution method for the multi-disruption model.

Bibliography

Substance Abuse and Mental Health Services Administration. Results from the 2013 national survey on drug use and health: Summary of national findings. *NSDUH Series H-48, HHS Publication No. (SMA) 14-4863.*, 2014.

A. Aghaie and H. Mokhtari. Ant colony optimization algorithm for stochastic project crashing problem in PERT networks using MC simulation. *The International Journal of Advanced Manufacturing Technology*, 45(11):1051–1067, 2009.

S. D. Ahipasaoglu, K. Natarajan, and D. Shi. Distributionally robust project crashing with partial or no correlation information. *Optimization-Online*, 2016. URL `http://www.optimization-online.org/DB_FILE/2016/11/5715.pdf`.

Z. Aksin, M. Armony, and V. Mehrotra. The modern call center: a multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6):665–688, 2007.

S. Asmussen. Applied probability and queues. *Stochastic Modelling and Applied Probability*, 51(1):355–426, 2003.

A. Attarha, N. Amjady, and A. J. Conejo. Adaptive robust AC optimal power flow considering load and wind power uncertainties. *International Journal of Electrical Power & Energy Systems*, 96:132–142, 2018.

X. Bai and H. Wei. A semidefinite programming method with graph partitioning technique for optimal power flow problems. *International Journal of Electrical Power & Energy Systems*, 33(7):1309–1314, 2011.

X. Bai, H. Wei, K. Fujisawa, and Y. Wang. Semidefinite programming for optimal power flow problems. *International Journal of Electrical Power & Energy Systems*, 30(6):383–392, 2008.

P. Bean. *Cocaine and Crack: Supply and Use.* Palgrave Macmillan, London, 1993.

P. Belotti, S. Cafieri, J. Lee, and L. Liberti. On feasibility based bounds tightening. *Optimization-Online*, 2012. URL `http://www.optimization-online.org/DB_FILE/2012/01/3325.pdf`.

A. Bernstein, D. Bienstock, D. Hay, M. Uzunoglu, and G. Zussman. Power grid vulnerability to geographically correlated failuresanalysis and control implications. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, pages 2634–2642. IEEE, 2014.

D. Bertsimas and M. Sim. Robust discrete optimization and network flows. *Mathematical Programming*, 98 (1):49–71, 2003.

D. Bertsimas and M. Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004.

M. Bibinger. Notes on the sum and maximum of independent exponentially distributed random variables with different scale parameters. arXiv:1307.3945, 2013. URL `https://arxiv.org/pdf/1307.3945.pdf`.

D. Bienstock. *Electrical Transmission System Cascades and Vulnerability: An Operations Research Viewpoint*. SIAM, 2015.

D. Bienstock, M. Chertkov, and S. Harnett. Chance-constrained optimal power flow: Risk-aware network control under uncertainty. *SIAM Review*, 56(3):461–495, 2014.

J. R. Birge and F. V. Louveaux. A multicut algorithm for two-stage stochastic linear programs. *European Journal of Operational Research*, 34:384–392, 1988.

J. Blazewicz, J. K. Lenstra, and A. H. G. R. Kan. Scheduling subject to resource constraints: classification and complexity. *Discrete Applied Mathematics*, 5(1):11–24, 1983.

R. N. Bluthenthal, A. H. Kral, L. Gee, E. A. Erringer, and B. R. Edlin. The effect of syringe exchange use on high-risk injection drug users: a cohort study. *AIDS*, 14(5):605–611, 2000.

R. N. Bluthenthal, R. Anderson, N. M. Flynn, and A. H. Kral. Higher syringe coverage is associated with lower odds of HIV risk and does not increase unsafe syringe disposal among syringe exchange program clients. *Drug and Alcohol Dependence*, 89(2):214–222, 2007.

A. Bobbio and A. Cumani. ML estimation of the parameters of a PH distribution in triangular canonical form. *Computer Performance Evaluation*, 22:33–46, 1992.

R. A. Bowman. Stochastic gradient-based time-cost tradeoffs in PERT networks using simulation. *Annals of Operations Research*, 53(1):533–551, 1994.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

N. Braine, D. C. Des Jarlais, S. Ahmad, D. Purchase, and C. Turner. Long-term effects of syringe exchange on risk behavior and HIV prevention. *AIDS Education and Prevention*, 16(3):264–275, 2004.

P. Buchholz, J. Kriege, and I. Felko. *Input Modeling with Phase-Type Distributions and Markov Models: Theory and Applications*, chapter 2. Springer Briefs in Mathematics, 2014.

J. M. Burt and M. B. Garman. Conditional Monte Carlo: A simulation technique for stochastic network analysis. *Management Science*, 18(3):207–217, 1971.

J. D. Camm, A. S. Raturi, and S. Tsubakitani. Cutting big M down to size. *Interfaces*, 20(5):61–66, 1990.

B. Cardoen, E. Demeulemeester, and J. Beliën. Operating room planning and scheduling: A literature review. *European Journal of Operational Research*, 201(3):921–932, 2010.

C. Carøe and J. Tind. L-shaped decomposition of two-stage stochastic programs with integer recourse. *Mathematical Programming*, 83(1-3):451–464, 1998.

J. Carpentier. Contribution á l'étude du dispatching économique. *Bulletin de la Société Française des Electriciens*, 3:431–447, 1962.

N. Chehrazi, P. W. Glynn, and T. A. Weber. Dynamic credit-collections optimization. *Management Science*, 2019. URL `https://doi.org/10.1287/mnsc.2018.3070`.

B. Chen, S. Küçükyavuz, and S. Sen. A computational study of the cutting plane tree algorithm for general mixed-integer linear programs. *Operations Research Letters*, 40(1):15–19, 2012.

X. Chen, M. Sim, P. Sun, and J. Zhang. A linear decision-based approximation approach to stochastic programming. *Operations Research*, 56(2):344–357, 2008.

K. L. Cheung, P. M. Klooster, C. Smit, H. de Vries, and M. E. Pieterse. The impact of non-response bias due to sampling in public health studies: A comparison of voluntary versus mandatory recruitment in a Dutch national survey on adolescent health. *BMC Public Health*, 17(1):276–286, 2017.

K. Clarke, D. Harris, J. A. Zweifler, M. Lasher, R. B. Mortimer, and S. Hughes. The significance of harm reduction as a social and health care intervention for injecting drug users: An exploratory study of a needle exchange program in Fresno, California. *Social Work Public Health*, 31(5):398–407, 2016.

C. Coffrin, D. Gordon, and P. Scott. NESTA, the NICTA energy system test case archive. *arXiv preprint arXiv:1411.0359*, 2014. URL `https://arxiv.org/pdf/1411.0359.pdf`.

C. Coffrin, H. L. Hijazi, and P. Van Hentenryck. Strengthening convex relaxations with bound tightening for power network optimization. In *International Conference on Principles and Practice of Constraint Programming*, pages 39–57. Springer, 2015a.

C. Coffrin, H. L. Hijazi, and P. Van Hentenryck. Strengthening the SDP relaxation of AC power flows with convex envelopes, bound tightening, and lifted nonlinear cuts. *arXiv preprint arXiv:1512.04644*, 2015b.

C. Coffrin, H. L. Hijazi, and P. Van Hentenryck. The QC relaxation: A theoretical and computational study on optimal power flow. *IEEE Transactions on Power Systems*, 31(4):3008–3018, 2016.

I. Cohen, B. Golany, and A. Shtub. The stochastic time–cost tradeoff problem: a robust optimization approach. *Networks*, 49(2):175–188, 2007.

T. G. Crainic, M. Hewitt, F. Maggioni, and W. Rei. Partial benders decomposition strategies for two-stage stochastic integer programs. Technical report, CIRRELT, 2016.

E. L. Crow and K. Shimizu. *Lognormal Distributions: Theory and Applications*. CRC Press, 1987.

P. De, E. J. Dunne, J. B. Ghosh, and C. E. Wells. Complexity of the discrete time-cost tradeoff problem for project networks. *Operations Research*, 45(2):302–306, 1997.

E. Demeulemeester, M. Vanhoucke, and W. Herroelen. RanGen: A random network generator for activity-on-the-node networks. *Journal of scheduling*, 6(1):17–38, 2003.

E. L. Demeulemeester and W. S. Herroelen. *Project scheduling: a research handbook*, volume 49. Springer Science & Business Media, 2006.

J. DeSimone. Needle exchange programs and drug injection behavior. *Journal of Policy Analysis and Management*, 24(3):559–577, 2005.

I. Dunning, J. Huchette, and M. Lubin. JuMP: A modeling language for mathematical optimization. *SIAM Review*, 59(2):295–320, 2017. doi: 10.1137/15M1020575.

S. E. Elmaghraby. *Activity networks: project planning and control by network models*. Wiley, 1977.

R. Enhbat. An algorithm for maximizing a convex function over a simple set. *Journal of Global Optimization*, 8(4):379–391, 1996.

M. Faddy and S. McClean. Analysing data on lengths of stay of hospital patients using phase-type distributions. *Applied Stochastic Models in Business and Industry*, 15:311–317, 1999.

M. Faddy and S. McClean. Markov chain modelling for geriatric patient care. *Methods of Information in Medicine*, 44:369–373, 2005.

M. Faddy, N. Graves, and A. Pettitt. Modeling length of stay in hospital and other right skewed data: comparison of phase-type, gamma and log-normal distributions. *Value in Health*, 12(2):309–314, 2009.

X. Fang, B. Hodge, E. Du, N. Zhang, and F. Li. Modelling wind power spatial-temporal correlation in multi-interval optimal power flow: A sparse correlation matrix approach. *Applied energy*, 230:531–539, 2018.

S. Fomundam and J. W. Herrmann. A survey of queuing theory applications in healthcare. Technical report, Institute for Systems Research, University of Maryland, 2007. URL `https://drum.lib.umd.edu/bitstream/handle/1903/7222/tr_2007-24.pdf`.

D. R. Fulkerson. A network flow computation for project cost curves. *Management Science*, 7(2):167–178, 1961.

V. Gabrel, M. Lacroix, C. Murat, and N. Remli. Robust location transportation problems under uncertain demands. *Discrete Applied Mathematics*, 164:100–111, 2014.

D. Gade, S. Küçükyavuz, and S. Sen. Decomposition algorithms with parametric Gomory cuts for two-stage stochastic integer programs. *Mathematical Programming*, 144(1-2):39–64, 2014.

W. Gardner, E. P. Mulvey, and E. C. Shaw. Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin*, 118(3):392–404, 1995.

M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., 1979.

A. M. Geoffrion. Generalized Benders decomposition. *Journal of Optimization Theory and Applications*, 10 (4):237–260, 1972.

M. K. Govil and M. C. Fu. Queueing theory in manufacturing: a survey. *Journal of Manufacturing Systems*, 18(3):214–240, 1999.

Gurobi Optimization, Inc. *Gurobi Optimizer Reference Manual*, 2016. URL `http://www.gurobi.com`.

N. G. Hall and C. Sriskandarajah. A survey of machine scheduling problems with blocking and no-wait in process. *Operations Research*, 44(3):510–525, 1996.

B. Hay, C. Henderson, J. Maltby, and J. J. Canales. Influence of peer-based needle exchange programs on mental health status in people who inject drugs: A nationwide New Zealand study. *Frontiers in Psychiatry*, 7:211, 2017.

D. Holtzman, V. Barry, L. J. Ouellet, D. C. Des Jarlais, D. Vlahov, E. T. Golub, S. M. Hudson, and R. S. Garfein. The influence of needle exchange programs on injection risk behaviors and infection with hepatitis C virus among young injection drug users in select cities in the United States, 1994-2004. *Preventive Medicine*, 49:68–73, 2009.

D. Huo and L. J. Ouellet. Needle exchange and injection-related risk behaviors in Chicago: a longitudinal study. *Journal of Acquired Immune Deficiency Syndromes*, 45(1):108–114, 2007.

D. Huo, S. L. Bailey, and L. J. Ouellet. Cessation of injection drug use and change in injection frequency: the Chicago needle exchange evaluation study. *Addiction*, 101(11):1606–1613, 2006.

R. A. Jabr. Radial distribution load flow using conic programming. *IEEE Transactions on power systems*, 21(3):1458–1459, 2006.

R. A. Jabr. Adjustable robust OPF with renewable energy sources. *IEEE Transactions on Power Systems*, 28(4):4742–4751, 2013.

E. J. Jaselskis and D. B. Ashley. Optimal allocation of project management resources for achieving success. *Journal of Construction Engineering and Management*, 117(2):321–340, 1991.

R. Jiang, J. Wang, and Y. Guan. Robust unit commitment with wind power and pumped storage hydro. *IEEE Transactions on Power Systems*, 27(2):800–810, 2012.

R. Jiang, M. Zhang, G. Li, and Y. Guan. Two-stage network constrained robust unit commitment problem. *European Journal of Operational Research*, 234(3):751–762, 2014.

H. Ke. A genetic algorithm-based optimizing approach for project time-cost trade-off with uncertain measure. *Journal of Uncertainty Analysis and Applications*, 2(1):8, 2014.

J. E. Kelly. Critical-path planning and scheduling: Mathematical basis. *Operations Research*, 9(3):296–320, 1961.

A. Khodaei. Resiliency-oriented microgrid optimal scheduling. *IEEE Transactions on Smart Grid*, 5(4): 1584–1591, 2014.

M. Kidorf and V. L. King. Expanding the public health benefits of syringe exchange programs. *The Canadian Journal of Psychiatry*, 53(8):487–495, 2008.

S. Kim, S. P. Boyd, S. Yun, D. D. Patil, and M. A. Horowitz. A heuristic for optimizing stochastic activity networks with applications to statistical digital circuit sizing. *Optimization and Engineering*, 8(4): 397–430, 2007.

S. Kim, R. Pasupathy, and S. G. Henderson. A guide to sample average approximation. In *Handbook of Simulation Optimization*, pages 207–243. Springer, 2015.

K. Klima and J. Apt. Geographic smoothing of solar PV: results from Gujarat. *Environmental Research Letters*, 10(10):104001, 2015.

E. Klotz and A. M. Newman. Practical guidelines for solving difficult mixed integer linear programs. *Surveys in Operations Research and Management Science*, 18(1-2):18–32, 2013.

B. Kocuk, S. S. Dey, and X. A. Sun. Strong SOCP relaxations for the optimal power flow problem. *Operations Research*, 64(6):1177–1196, 2016.

M. E. Kuhl and R. A. Tolentino-Peña. A dynamic crashing method for project management using simulation-based optimization. In *Proceedings of the 40th Conference on Winter Simulation*, pages 2370–2376. Winter Simulation Conference, 2008.

C. Lakshmi and S. A. Iyer. Application of queueing theory in health care: a literature review. *Operations Research for Health Care*, 2(1):25–39, 2013.

J. Lavaei and S. H. Low. Zero duality gap in optimal power flow problem. *IEEE Transactions on Power Systems*, 27(1):92–107, 2012.

Z. Li and M. Ierapetritou. Process scheduling under uncertainty: Review and challenges. *Computers & Chemical Engineering*, 32(4-5):715–727, 2008.

Y. Liu and M. C. Ferris. Security constrained economic dispatch using semidefinite programming. In *2015 IEEE Power and Energy Society General Meeting*. Institute of Electrical and Electronics Engineers (IEEE), 2015.

Y. Liu, S. Li, F. Li, L. Song, and J. M. Rehg. Efficient learning of continuous-time hidden Markov models for disease progression. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 3600–3608. Curran Associates, Inc., 2015.

D. Locker, R. Wiggins, Y. Sittampalam, and D. L. Patrick. Estimating the prevalence of disability in the community: the influence of sample design and response bias. *Journal of Epidemiology & Community Health*, 35(3):208–212, 1981.

G. M. Lohmann, A. H. Monahan, and D. Heinemann. Local short-term variability in solar irradiance. *Atmospheric Chemistry and Physics*, 16(10):6365–6379, 2016.

Á. Lorca and X. A. Sun. The adaptive robust multi-period alternating current optimal power flow problem. *IEEE Transactions on Power Systems*, 33(2):1993–2003, 2018.

R. Louca and E. Bitar. Robust AC optimal power flow. *arXiv preprint arXiv:1706.090199*, 2017.

S. H. Low. Convex relaxation of optimal power flow - part I: Formulations and equivalence. *IEEE Transactions on Control of Network Systems*, 1(1):15–27, 2014a.

S. H. Low. Convex relaxation of optimal power flow - part II: Exactness. *IEEE Transactions on Control of Network Systems*, 1(2):177–189, 2014b.

M. Lubin, Y. Dvorkin, and S. Backhaus. A robust approach to chance constrained optimal power flow with renewable generation. *IEEE Transactions on Power Systems*, 31(5):3840–3849, 2016.

T. L. Magnanti and R. T. Wong. Accelerating Benders decomposition: Algorithmic enhancement and model selection criteria. *Operations research*, 29(3):464–484, 1981.

J. Maisano, A. Radchik, and T. Ling. A lognormal model for demand forecasting in the national electricity market. *The ANZIAM Journal*, 57(3):369–383, 2016.

W. K. Mak, D. P. Morton, and R. K. Wood. Monte Carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters*, 24(1):47–56, 1999.

D. G. Malcolm, J. H. Roseboom, C. E. Clark, and W. Fazar. Application of a technique for research and development program evaluation. *Operations research*, 7(5):646–669, 1959.

A. Malvaldi, S. Weiss, D. Infield, J. Browell, P. Leahy, and A. M. Foley. A spatial and temporal correlation analysis of aggregate wind power in an ideally interconnected Europe. *Wind Energy*, 20(8):1315–1329, 2017.

J. A. Momoh, M. E. El-Hawary, and R. Adapa. A review of selected optimal power flow literature to 1993. Part II. Newton, linear programming and interior point methods. *IEEE Transactions on Power Systems*, 14(1):105–111, 1999.

A. Monticelli, M. V. F. Pereira, and S. Granville. Security-constrained optimal power flow with post-contingency corrective rescheduling. *IEEE Transactions on Power Systems*, 2(1):175–180, 1987.

R. E. Mullen. The lognormal distribution of software failure rates: origin and evidence. In *Proceedings Ninth International Symposium on Software Reliability Engineering (Cat. No.98TB100257)*, pages 124–133, Nov 1998.

B. L. Nelson and I. Gerhardt. On capturing dependence in point processes: matching moments and other techniques. Technical report, Northwestern University, 2010. URL `http://users.iems.northwestern.edu/~nelsonb/Publications/GerhardtNelsonSurvey.pdf`.

A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. A Wiley-Interscience publication. Wiley, 1983.

H. D. Nguyen, K. Dvijotham, and K. Turitsyn. Constructing convex inner approximations of steady-state security regions. *IEEE Transactions on Power Systems*, 34(1):257–267, 2019.

G. D. Oberlender. *Project management for engineering and construction*, volume 2. McGraw-Hill New York, 1993.

S. M. Paddock, B. Kilmer, J. P. Caulkins, M. J. Booth, and R. L. Pacula. An epidemiological model for examining marijuana use over the life course. *Epidemiology Research International*, volume 2012, article ID 520894, 2012.

T. Papaioannou. Censoring. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, and J. L. Teugels, editors, *Wiley StatsRef: Statistics Reference Online*. American Cancer Society, 2014.

D. Phan and S. Ghosh. Two-stage stochastic optimization for optimal power flow under renewable generation uncertainty. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 24(1):1–22, 2014.

A. B. Philpott, F. Wahid, and J. F. Bonnans. MIDAS: A mixed integer dynamic approximation scheme. *Mathematical Programming*, Feb 2019. URL `https://doi.org/10.1007/s10107-019-01368-1`.

E. L. Plambeck, B. Fu, S. M. Robinson, and R. Suri. Sample-path optimization of convex stochastic performance functions. *Mathematical Programming*, 75(2):137–176, 1996.

Y. Qi and S. Sen. The ancestral Benders cutting plane algorithm with multi-term disjunctions for mixed-integer recourse decisions in stochastic programming. *Mathematical Programming*, 161(1-2):193–235, 2017.

S. M. Ross, J. J. Kelly, R. J. Sullivan, W. J. Perry, D. Mercer, R. M. Davis, T. D. Washburn, E. V. Sager, J. B. Boyce, and V. L. Bristow. *Stochastic processes*, volume 2. Wiley New York, 1996.

J. Salmeron, R. K. Wood, and D. P. Morton. A stochastic program for optimizing military sealift subject to attack. *Military Operations Research*, 14(2):19–39, 2009.

A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

R. Singh, B. C. Pal, and R. A. Jabr. Statistical representation of distribution system loads using gaussian mixture model. *IEEE Transactions on Power Systems*, 25(1):29–37, 2010.

J. Söderlund. Building theories of project management: past research, questions for the future. *International Journal of Project Management*, 22(3):183–191, 2004.

B. Stott, J. Jardim, and O. Alsaç. DC power flow revisited. *IEEE Transactions on Power Systems*, 24(3): 1290–1300, 2009.

K. Sundar, H. Nagarajan, S. Misra, M. Lu, C. Coffrin, and R Bent. Optimization-based bound tightening using a strengthened QC-relaxation of the optimal power flow problem. *arXiv preprint arXiv:1809.04565*, 2018. URL `https://arxiv.org/pdf/1809.04565.pdf`.

T. L. Terry. *Robust linear optimization with recourse: Solution methods and other properties.* PhD thesis, University of Michigan, 2009.

A. Thiele, T. Terry, and M. Epelman. Robust linear optimization with recourse. *Optimization-Online*, 2009. URL `http://www.optimization-online.org/DB_FILE/2009/03/2263.pdf`.

S. Tonchia. *Industrial project management.* Springer, 2018.

R. M. van Slyke. Letter to the editor—Monte Carlo methods and the PERT problem. *Operations Research*, 11(5):839–860, 1963.

A. Verma. *Power grid security analysis: An optimization approach.* PhD thesis, Columbia University, 2010.

A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical programming*, 106(1):25–57, 2006.

W. Wiesemann, D. Kuhn, and B. Rustem. Robust resource allocations in temporal networks. *Mathematical Programming*, 135(1):437–471, 2012.

W. Xie and S. Ahmed. Distributionally robust chance constrained optimal power flow with renewables: A conic reformulation. *IEEE Transactions on Power Systems*, 33(2):1860–1867, 2018.

G. Yu and X. Qi. *Disruption Management: Framework, Models and Applications.* World Scientific, 2004.

L. Zhao and B. Zeng. An exact algorithm for two-stage robust optimization with mixed integer recourse problems. Technical report, University of Florida, 2012.

J. Zou, S. Ahmed, and X. A. Sun. Stochastic dual dynamic integer programming. *Optimization-Online*, 2016. URL `http://www.optimization-online.org/DB_FILE/2016/05/5436.pdf`.

M. Zugno and A. J. Conejo. A robust optimization approach to energy and reserve dispatch in electricity markets. *European Journal of Operational Research*, 247(2):659–671, 2015.

Appendix A

Appendices for Chapter 2

A.1 Test Cases Data

We present the data of four test cases here. For all test cases we assume there is only one possible crashing option for each activity. The option consumes 1 unit of resource and has the effectiveness parameter of $e_{i1} = 0.5$ for all $i \in I$. The nominal scenario probability is $p^0 = 0.1$ and $p^\omega = \frac{1-p^0}{|\Omega|}$. The timing of the disruption follows a lognormal distribution with parameters μ and σ where the mode is $e^{\mu-\sigma^2}$. We also assume that the duration only depends on the predecessor, i.e., $D_{ik} = D_i$ and $d^\omega_{ik} = d^\omega_i$. All d_i follow an exponential distribution with a mean of μ_i. The value of D_i and μ_i are shown in the following tables.

- Case 11: $B = 3, \mu = \ln 6, \sigma = 0.5$

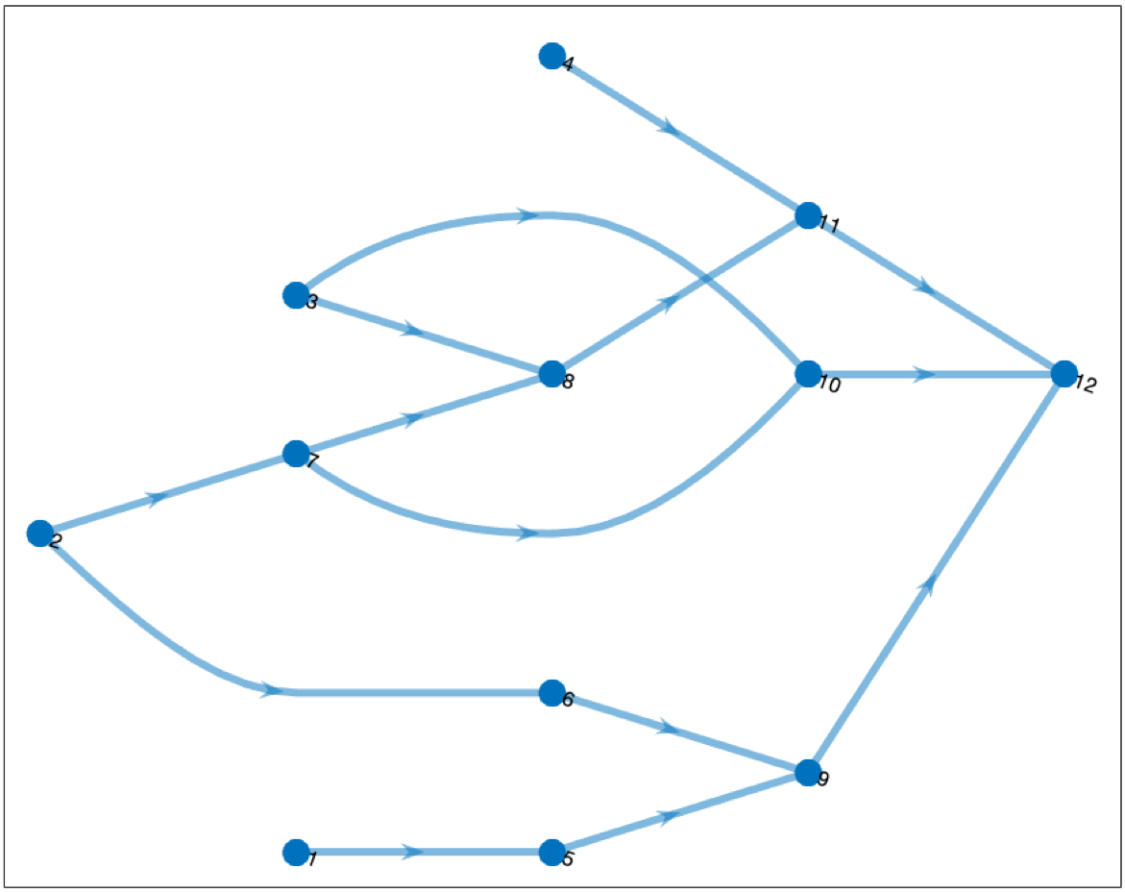


Figure A.1: Activity network of Case 11

Activity	D_i	λ_i
1	10	10^{-5}
2	2	4
3	10	2
4	12	30
5	3	1500
6	10	1

Activity	D_i	λ_i
7	7.3	1
8	4.9	50
9	11.1	40
10	3.5	40
11	9.9	5

Table A.1: Activity duration D_i and the mean of disruption magnitude λ_i for Case 11

- Case 14: $B = 4, \mu = \ln 35, \sigma = 0.5$

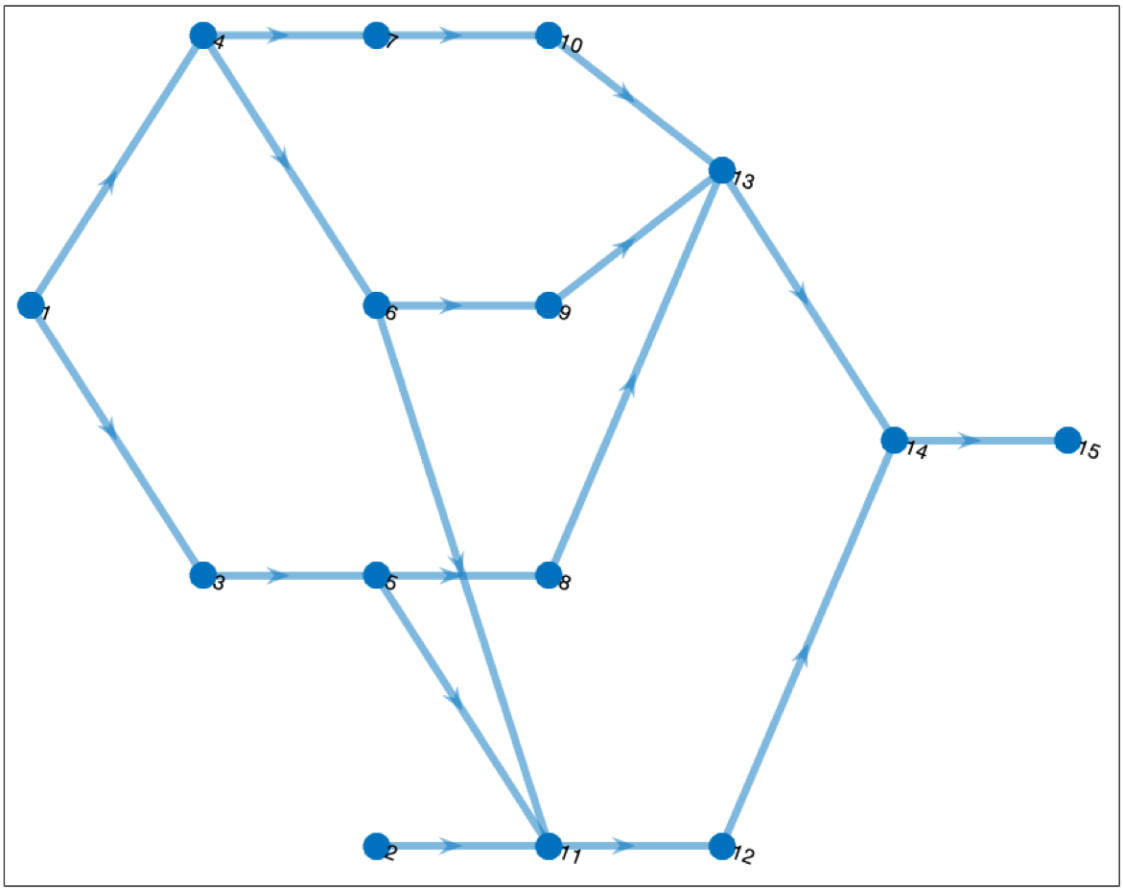


Figure A.2: Activity network of Case 14

Activity	D_i	λ_i
1	5	10^{-5}
2	30	5
3	25	40000
4	20	40000
5	15	1500
6	24	20000
7	30	20000

Activity	D_i	λ_i
8	49	4000
9	40	4000
10	30	3000
11	45	4000
12	25	5
13	21	5
14	5	5

Table A.2: Activity duration D_i and the mean of disruption magnitude λ_i for Case 14

- Case 19: $B = 4, \mu = \ln 8, \sigma = 0.5$

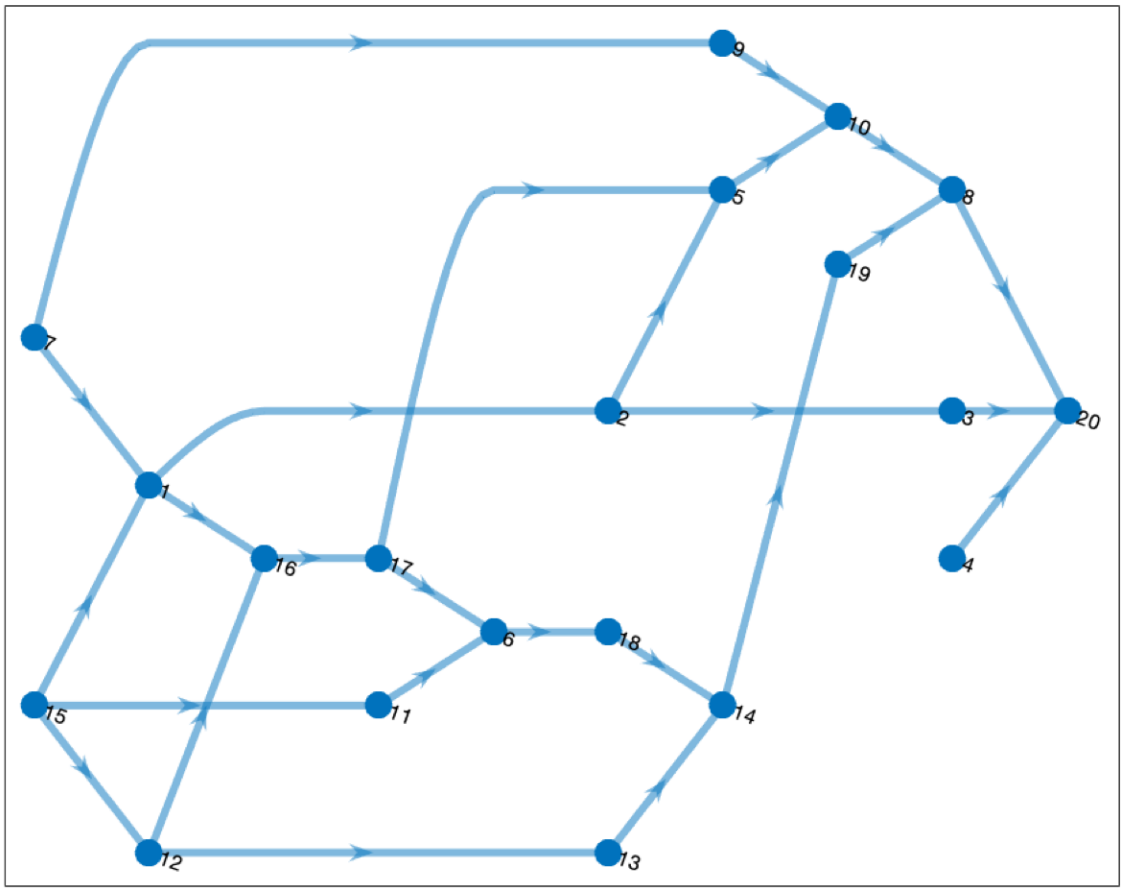


Figure A.3: Activity network of Case 19

Activity	D_i	λ_i
1	5	10
2	10	100
3	18	300
4	35	300
5	6	50
6	6	50
7	2.5	400
8	8	20
9	10	1000
10	1	50

Activity	D_i	λ_i
11	6	400
12	4.5	300
13	12	1000
14	1.5	10
15	2.1	200
16	8.1	2
17	6.1	100
18	0.001	40
19	0.001	300

Table A.3: Activity duration D_i and the mean of disruption magnitude λ_i for Case 19

- Case 35: $B = 8, \mu = \ln 4, \sigma = 0.3$

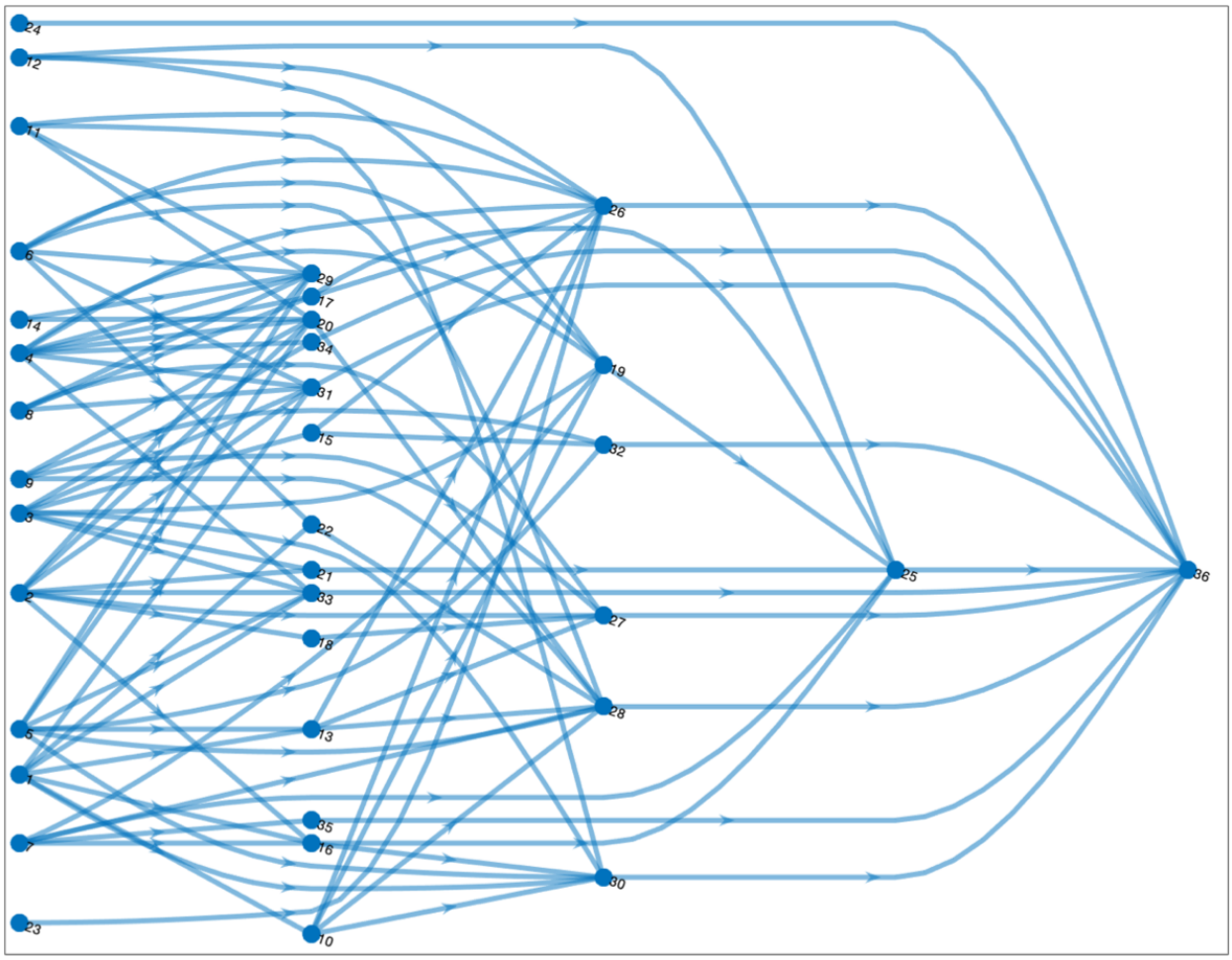


Figure A.4: Activity network of Case 35

Activity	D_i	λ_i	Activity	D_i	λ_i
1	9	10	19	8	10
2	7	40	20	8	500
3	3	30	21	1	500
4	4	100	22	5	500
5	6	50	23	2	10
6	3	10	24	7	10
7	10	10	25	1	300
8	4	20	26	4	400
9	3	10	27	3	200
10	6	1000	28	4	1000
11	9	10	29	10	300
12	8	500	30	7	500
13	5	200	31	2	200
14	2	10	32	9	100
15	5	400	33	7	100
16	2	10	34	1	100
17	10	10	35	7	200
18	4	2000			

Table A.4: Activity duration D_i and the mean of disruption magnitude λ_i for Case 35

Appendix B

Appendices for Chapter 3

B.1 Grouping Buses

To construct the uncertainty set $\mathcal{U}$, we solve a facility location problem for each test case to cluster the buses. That is, while we solve the test instances with the full resolution of network topology, as indicated in equation (3.1c), uncertain injections at buses occur in concert within a cluster. We assume the distance between two directly connected buses is 1, and more generally, the distance between two buses is the length of the shortest path (counted in hops) between them. We select a total of $|\mathcal{M}| = 5$ buses to be the "facilities" and assign each bus to a facility. All buses that are assigned to a facility are considered a cluster, i.e., elements of $\mathcal{N}_m$.

The detailed formulation is expressed as follows:

Indices and index sets

$i \in \mathcal{N}$	set of buses;
$J_i \subseteq \mathcal{N}$	set of buses eligible to be associated with bus i, $i \in \mathcal{N}$;

Parameters

d_{ij}	distance between bus i and j, $i, j \in \mathcal{N}$;
N	number of facilities ($|\mathcal{M}|$);

Decision variables

x_{ij}	indicator of whether bus i is assigned to bus j, $i, j \in \mathcal{N}$;
y_i	indicator of whether bus i is selected as a facility, $i \in \mathcal{N}$.

Facility Location Problem:

$$\min \quad \sum_{i\in\mathcal{N}}\sum_{j\in J_i} d_{ij}x_{ij} \tag{B.1a}$$

$$\text{s.t.} \quad \sum_{j\in J_i} x_{ij} = 1 \qquad \forall i \in \mathcal{N} \tag{B.1b}$$

$$\sum_{i\in\mathcal{N}} y_i = N \tag{B.1c}$$

$$x_{ij} \in \{0,1\} \qquad \forall i \in \mathcal{N}, j \in J_i \tag{B.1d}$$

$$y_i \in \{0,1\} \qquad \forall i \in \mathcal{N}. \tag{B.1e}$$

To facilitate computational tractability, we control the size of J_i via a distance threshold so that we can only assign one bus to another if their distance is within the threshold. The clusters formed by model (B.1) define the sets $\mathcal{M}$ and $\mathcal{N}_m$ used in equation (3.1c) to construct $\mathcal{U}$ as described at the beginning of Section 3.2.

In addition to clustering buses, we distinguish uncertainty in load and in renewable generation as described in Section 3.4.1.1. The latter occurs only at a subset of buses, denote $\mathcal{N}_G$, and we now describe construction of this set. Once clusters are formed, we select the two buses in each cluster that have the largest capacity, defined by the sum of the capacities of the incident lines. If the cluster is a singleton, then only that bus is selected, and the process for other clusters remains the same. The buses selected in this way form set $\mathcal{N}_G$.

B.2 QC Relaxation

We use the quadratic convex (QC) relaxation from Coffrin et al. (2016). Nonconvex functions in the ACOPF problem, such as quadratic, cosine and sine functions are transformed into a collection of second-order cone constraints and linear constraints. A McCormick relaxation is applied to linearize multi-linear terms. We also assume the difference in phase angles at adjacent buses i and j satisfies $-\frac{\pi}{6} \leqslant \underline{\Delta}_k \leqslant \theta_i - \sigma_k - \theta_j \leqslant \bar{\Delta}_k \leqslant \frac{\pi}{6}$. The detailed formulation of the QC relaxation is expressed as follows:

Indices and index sets

$i \in \mathcal{N}$	set of buses;
$k = (i,j,n) \in \mathcal{A}$	set of lines;

$g \in \mathcal{G}$	set of generators;
$g \in \mathcal{G}_i$	set of generators that are connected to bus $i \in \mathcal{N}$;

Parameters

u_i^p	uncontrollable active power injection at bus $i \in \mathcal{N}$;
u_i^q	uncontrollable reactive power injection at bus $i \in \mathcal{N}$;
$\underline{s}_g^p, \bar{s}_g^p$	lower and upper bound of active power generation by generator g, $g \in \mathcal{G}$;
$\underline{s}_g^q, \bar{s}_g^q$	lower and upper bound of reactive power generation by generator g, $g \in \mathcal{G}$;
$\underline{v}_i, \bar{v}_i$	lower and upper bound of voltage magnitude at bus i, $i \in \mathcal{N}$;
$\underline{\theta}_i, \bar{\theta}_i$	lower and upper bound of phase angle at bus $i \in \mathcal{N}$;
$\underline{\Delta}_k, \bar{\Delta}_k$	lower and upper bound of phase angle difference of adjacent buses on line $k \in \mathcal{A}$;
$\underline{cs}_k, \bar{cs}_k$	lower and upper bound of cosine of phase angle difference of adjacent buses on line $k \in \mathcal{A}$;
$\underline{ss}_k, \bar{ss}_k$	lower and upper bound of sine of phase angle difference of adjacent buses on line $k \in \mathcal{A}$;
g_k	conductance of line k, $k \in \mathcal{A}$;
b_k	susceptance of line k, $k \in \mathcal{A}$;
b_k^c	charging susceptance of line k, $k \in \mathcal{A}$;
g_i^{sh}	shunt conductance of bus i, $i \in \mathcal{N}$;
b_i^{sh}	shunt susceptance of bus i, $i \in \mathcal{N}$;
W_k	maximum apparent power flow on line k, $k \in \mathcal{A}$;
$\tau_{1,k}$	tap ratio of transformer at bus i on line k, $k \in \mathcal{A}$;
$\tau_{2,k}$	tap ratio of transformer at bus j on line k, $k \in \mathcal{A}$;
σ_k	phase angle shift of transformer on line k, $k \in \mathcal{A}$;
θ_k^u	upper bound of absolute value of phase angle difference of line k, $k \in \mathcal{A}$, $\theta_k^u = \max(\lvert\bar{\Delta}_k\rvert, \lvert\underline{\Delta}_k\rvert)$;
$\underline{v}_i^\delta$	sum of lower and upper bound of voltage magnitude at bus i, $i \in \mathcal{N}$, $v_i^\delta = \bar{v}_i + \underline{v}_i$;
θ_k^ϕ	mid-point of range of difference in phase angles on line k, $k \in \mathcal{A}$, $\theta_k^\phi = \frac{(\bar{\Delta}_k + \underline{\Delta}_k)}{2}$;
θ_k^δ	range of difference in phase angles on line k, $k \in \mathcal{A}$, $\theta_k^\delta = \frac{(\bar{\Delta}_k - \underline{\Delta}_k)}{2}$;

Decision variables

s_g^p	active power generation at generator g, $g \in \mathcal{G}$;
s_g^q	reactive power generation at generator g, $g \in \mathcal{G}$;
v_i	voltage magnitude at bus i, $i \in \mathcal{N}$;

θ_i	phase angle at bus i, $i \in \mathcal{N}$;
P_k	active power flow on line k, $k \in \mathcal{A}$;
Q_k	reactive power flow on line k, $k \in \mathcal{A}$;
$\widehat{cs}_k$	approximation term of $\cos(\theta_i - \sigma_k - \theta_j)$, $k = (i,j,n) \in \mathcal{A}$;
$\widehat{ss}_k$	approximation term of $\sin(\theta_i - \sigma_k - \theta_j)$, $k = (i,j,n) \in \mathcal{A}$;
$\hat{v}_i$	approximation term of v_i^2, $i \in \mathcal{N}$;
$\widehat{vv}_k$	approximation term of $\frac{v_i v_j}{\tau_{1,k}\tau_{2,k}}$, $k = (i,j,n) \in \mathcal{A}$;
$\widehat{wc}_k$	approximation term of $\frac{v_i v_j}{\tau_{1,k}\tau_{2,k}}\cos(\theta_i - \sigma_k - \theta_j)$, $k = (i,j,n) \in \mathcal{A}$;
$\widehat{ws}_k$	approximation term of $\frac{v_i v_j}{\tau_{1,k}\tau_{2,k}}\sin(\theta_i - \sigma_k - \theta_j)$, $k = (i,j,n) \in \mathcal{A}$.

Formulation:

$$\min \quad c(s^p, s^q) \tag{B.2a}$$

$$\text{s.t.} \quad \underline{v}_i \leqslant v_i \leqslant \bar{v}_i \qquad \forall i \in \mathcal{N} \tag{B.2b}$$

$$\underline{\Delta}_k \leqslant \theta_i - \sigma_k - \theta_j \leqslant \bar{\Delta}_k \qquad \forall k = (i,j,n) \in \mathcal{A} \tag{B.2c}$$

$$\underline{\theta}_i \leqslant \theta_i \leqslant \bar{\theta}_i \qquad \forall i \in \mathcal{N} \tag{B.2d}$$

$$\underline{cs}_k \leqslant \widehat{cs}_k \leqslant \bar{cs}_k \qquad \forall k = (i,j,n) \in \mathcal{A} \tag{B.2e}$$

$$\underline{ss}_k \leqslant \widehat{ss}_k \leqslant \bar{ss}_k \qquad \forall k = (i,j,n) \in \mathcal{A} \tag{B.2f}$$

$$\underline{s}_g^p \leqslant s_g^p \leqslant \bar{s}_g^p \qquad \forall g \in \mathcal{G} \tag{B.2g}$$

$$\underline{s}_g^q \leqslant s_g^q \leqslant \bar{s}_g^q \qquad \forall g \in \mathcal{G} \tag{B.2h}$$

$$P_k = g_k \frac{\hat{v}_i}{(\tau_{1,k})^2} - g_k \frac{\widehat{wc}_k}{\tau_{1,k}\tau_{2,k}} - b_k \frac{\widehat{ws}_k}{\tau_{1,k}\tau_{2,k}} \qquad \forall k = (i,j,n) \in \mathcal{A} \tag{B.2i}$$

$$Q_k = -(b_k + \frac{b_k^c}{2})\frac{\hat{v}_i}{(\tau_{1,k})^2} + b_k \frac{\widehat{wc}_k}{\tau_{1,k}\tau_{2,k}} - g_k \frac{\widehat{ws}_k}{\tau_{1,k}\tau_{2,k}} \qquad \forall k = (i,j,n) \in \mathcal{A} \tag{B.2j}$$

$$P_k^2 + Q_k^2 \leqslant W_k^2 \qquad \forall k \in \mathcal{A} \tag{B.2k}$$

$$\sum_{k=(i,j,n)\in\mathcal{A}} P_k + g_i^{sh}\hat{v}_i = \sum_{g\in\mathcal{G}_i} s_g^p + u_i^p \qquad \forall i \in \mathcal{N} \tag{B.2l}$$

$$\sum_{k=(i,j,n)\in\mathcal{A}} Q_k - b_i^{sh}\hat{v}_i = \sum_{g\in\mathcal{G}_i} s_g^q + u_i^q \qquad \forall i \in \mathcal{N} \tag{B.2m}$$

$$\widehat{vv}_k^2 \leqslant \frac{\hat{v}_i}{\tau_{1,k}^2}\frac{\hat{v}_j}{\tau_{2,k}^2} \qquad \forall k = (i,j,n) \in \mathcal{A} \tag{B.2n}$$

$$\widehat{cs}_k + \frac{1-\cos(\theta_k^u)}{(\theta_k^u)^2}(\theta_i - \sigma_k - \theta_j)^2 \leqslant 1 \qquad \forall k = (i,j,n) \in \mathcal{A} \tag{B.2o}$$

$$\widehat{cs}_k \geqslant \frac{\cos(\bar{\Delta}_k) - \cos(\underline{\Delta}_k)}{\bar{\Delta}_k - \underline{\Delta}_k}(\theta_i - \sigma_k - \theta_j - \underline{\Delta}_k) + \cos(\underline{\Delta}_k) \qquad \forall k = (i,j,n) \in \mathcal{A} \tag{B.2p}$$

$$\widehat{ss}_k - \cos\left(\frac{\theta_k^u}{2}\right)(\theta_i - \sigma_k - \theta_j) \leqslant \sin\left(\frac{\theta_k^u}{2}\right) - \frac{\theta_k^u}{2}\cos\left(\frac{\theta_k^u}{2}\right) \qquad \forall k = (i,j,n) \in \mathcal{A} \tag{B.2q}$$

$$-\hat{ss}_k + \cos\left(\frac{\theta_k^u}{2}\right)(\theta_i - \sigma_k - \theta_j) \leqslant \sin\left(\frac{\theta_k^u}{2}\right) - \frac{\theta_k^u}{2}\cos\left(\frac{\theta_k^u}{2}\right) \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.2r)}$$

$$v_i^2 - \hat{v}_i \leqslant 0 \qquad \forall i \in \mathcal{N} \quad \text{(B.2s)}$$

$$\hat{v}_i - (\bar{v}_i + \underline{v}_i)v_i \leqslant -\bar{v}_i\underline{v}_i \qquad \forall i \in \mathcal{N} \quad \text{(B.2t)}$$

$$\widehat{vv}_k \geqslant \frac{\underline{v}_i}{\tau_{1,k}}\frac{v_j}{\tau_{2,k}} + \frac{\underline{v}_j}{\tau_{2,k}}\frac{v_i}{\tau_{1,k}} - \frac{\underline{v}_i}{\tau_{1,k}}\frac{\underline{v}_j}{\tau_{2,k}} \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.2u)}$$

$$\widehat{vv}_k \geqslant \frac{\bar{v}_i}{\tau_{1,k}}\frac{v_j}{\tau_{2,k}} + \frac{\bar{v}_j}{\tau_{2,k}}\frac{v_i}{\tau_{1,k}} - \frac{\bar{v}_i}{\tau_{1,k}}\frac{\bar{v}_j}{\tau_{2,k}} \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.2v)}$$

$$\widehat{vv}_k \leqslant \frac{\underline{v}_i}{\tau_{1,k}}\frac{v_j}{\tau_{2,k}} + \frac{\bar{v}_j}{\tau_{2,k}}\frac{v_i}{\tau_{1,k}} - \frac{\underline{v}_i}{\tau_{1,k}}\frac{\bar{v}_j}{\tau_{2,k}} \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.2w)}$$

$$\widehat{vv}_k \leqslant \frac{\bar{v}_i}{\tau_{1,k}}\frac{v_j}{\tau_{2,k}} + \frac{\underline{v}_j}{\tau_{2,k}}\frac{v_i}{\tau_{1,k}} - \frac{\bar{v}_i}{\tau_{1,k}}\frac{\underline{v}_j}{\tau_{2,k}} \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.2x)}$$

$$\widehat{wc}_k \geqslant \frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\hat{cs}_k + \underline{cs}_k\widehat{vv}_k - \frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\underline{cs}_k \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.2y)}$$

$$\widehat{wc}_k \geqslant \frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\hat{cs}_k + \bar{cs}_k\widehat{vv}_k - \frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\bar{cs}_k \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.2z)}$$

$$\widehat{wc}_k \leqslant \frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\hat{cs}_k + \bar{cs}_k\widehat{vv}_k - \frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\bar{cs}_k \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.2aa)}$$

$$\widehat{wc}_k \leqslant \frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\hat{cs}_k + \underline{cs}_k\widehat{vv}_k - \frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\underline{cs}_k \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.2ab)}$$

$$\widehat{ws}_k \geqslant \frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\hat{ss}_k + \underline{ss}_k\widehat{vv}_k - \frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\underline{ss}_k \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.2ac)}$$

$$\widehat{ws}_k \geqslant \frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\hat{ss}_k + \bar{ss}_k\widehat{vv}_k - \frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\bar{ss}_k \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.2ad)}$$

$$\widehat{ws}_k \leqslant \frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\hat{ss}_k + \bar{ss}_k\widehat{vv}_k - \frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\bar{ss}_k \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.2ae)}$$

$$\widehat{ws}_k \leqslant \frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\hat{ss}_k + \underline{ss}_k\widehat{vv}_k - \frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\underline{ss}_k \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.2af)}$$

$$\widehat{ws}_k - \tan(\bar{\Delta}_k)\widehat{wc}_k \leqslant 0 \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.2ag)}$$

$$\widehat{ws}_k - \tan(\underline{\Delta}_k)\widehat{wc}_k \geqslant 0 \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.2ah)}$$

$$\frac{v_i^\delta v_j^\delta}{\tau_{1,k}\tau_{2,k}}(\widehat{wc}_k\cos(\theta_k^\phi) + \widehat{ws}_k\sin(\theta_k^\phi)) - \frac{\bar{v}_j}{\tau_{2,k}}\cos(\theta_k^\delta)\frac{v_j^\delta}{\tau_{2,k}}\frac{\hat{v}_i}{\tau_{1,k}^2}$$
$$-\frac{\bar{v}_i}{\tau_{1,k}}\cos(\theta_k^\delta)\frac{v_i^\delta}{\tau_{1,k}}\frac{\hat{v}_j}{\tau_{2,k}^2} \geqslant \frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\cos(\theta_k^\delta)(\frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}} - \frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}}) \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.2ai)}$$

$$\frac{v_i^\delta v_j^\delta}{\tau_{1,k}\tau_{2,k}}(\widehat{wc}_k\cos(\theta_k^\phi) + \widehat{ws}_k\sin(\theta_k^\phi)) - \frac{\underline{v}_j}{\tau_{2,k}}\cos(\theta_k^\delta)\frac{v_j^\delta}{\tau_{2,k}}\frac{\hat{v}_i}{\tau_{1,k}^2}$$
$$-\frac{\underline{v}_i}{\tau_{1,k}}\cos(\theta_k^\delta)\frac{v_i^\delta}{\tau_{1,k}}\frac{\hat{v}_j}{\tau_{2,k}^2} \geqslant -\frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\cos(\theta_k^\delta)(\frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}} - \frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}}) \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.2aj)}$$

$$\hat{cs}_k = \hat{cs}_{\tilde{k}} \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.2ak)}$$

$$\widehat{ss}_k = -\widehat{ss}_{\tilde{k}} \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.2al)}$$

$$\widehat{wc}_k = \widehat{wc}_{\tilde{k}} \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.2am)}$$

$$\widehat{ws}_k = -\widehat{ws}_{\tilde{k}} \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.2an)}$$

$$\widehat{vv}_k = \widehat{vv}_{\tilde{k}} \qquad \forall k = (i,j,n) \in \mathcal{A}. \quad \text{(B.2ao)}$$

Constraints (B.2b)-(B.2f) are simple bounds for voltage magnitude, difference in phase angles, phase angles and approximation terms for $\cos(\theta_i - \sigma_k - \theta_j)$ and $\sin(\theta_i - \sigma_k - \theta_j)$, respectively. Using the bound tightening techniques in Coffrin et al. (2015a), we can derive the upper bounds and lower bounds of $\cos(\theta_i - \theta_j)$ and $\sin(\theta_i - \theta_j)$ as:

$$\begin{aligned}
&\bar{cs}_k = \cos(\bar{\Delta}_k), \quad \underline{cs}_k = \cos(\underline{\Delta}_k) && \text{if } \bar{\Delta}_k \leqslant 0 \\
&\bar{cs}_k = \cos(\underline{\Delta}_k), \quad \underline{cs}_k = \cos(\bar{\Delta}_k) && \text{if } \underline{\Delta}_k \geqslant 0 \\
&\bar{cs}_k = 1, \quad \underline{cs}_k = \min\{\cos(\underline{\Delta}_k), \cos(\bar{\Delta}_k)\} && \text{if } \bar{\Delta}_k \geqslant 0 \text{ and } \underline{\Delta}_k \leqslant 0,
\end{aligned}$$

and

$$\begin{aligned}
\bar{ss}_k &= \sin(\bar{\Delta}_k) \\
\underline{ss}_k &= \sin(\underline{\Delta}_k).
\end{aligned}$$

Constraints (B.2i) and (B.2j) are linearized versions of the power flow constraints in (3.3), and constraint (B.2k) replicates (3.5b), while active and reactive power flow balance constraints (B.2l) and (B.2m) replicate (3.4). Since we are modeling the deterministic ACOPF problem here, we do not have the recourse freedom variables $o^{p,+}, o^{p,-}, o^{q,+}, o^{q,-}$ in this formulation. Constraint (B.2n) represents the correct quantitative relationship between $\widehat{vv}_k$ and $\hat{v}_i\hat{v}_j$, where $\widehat{vv}$ can be considered as a relaxation of the bilinear term v_iv_j. Constraint (B.2o)-(B.2t) are linear and convex quadratic bounding approximations of cosine, sine and quadratic functions. For multi-linear terms, a McCormick scheme is applied to relax v_iv_j, $v_iv_j\cos(\theta_i - \theta_j)$ and $v_iv_j\sin(\theta_i - \theta_j)$ in constraints (B.2u)-(B.2af).

The model includes valid inequalities described in Coffrin et al. (2015b) to further tighten this convex relaxation. One set of valid constraints uses the trigonometric relationship $\tan(\theta) = \frac{\sin(\theta)}{\cos(\theta)}$ to build valid inequalities on $\widehat{wc}$ and $\widehat{ws}$ as in (B.2ag) and (B.2ah), using the fact that $\widehat{wc} \geqslant 0$. Another

set of valid constraints, constraints (B.2ai)-(B.2aj), is called a lifted nonlinear cut. See Coffrin et al. (2015b) for a detailed derivation of the lifted nonlinear cut. Finally, we use constraints (B.2ak)-(B.2an) to make sure variables describing the forward flow are consistent with those describing the backward flow between the same pair of buses, where $\tilde{k} = (j, i, n)$ represents the backward flow of the flow $k = (i, j, n)$.

B.3 Detailed Formulation of Model (3.14)

In this section we first expand the formulation of model (3.14), and then we explain the corresponding relationship between the dual variables and the primal constraints in formulation (3.14) and Appendix B.2.

$$
\begin{aligned}
\max \quad & -\sum_{i\in\mathcal{N}}\Bigg[\sum_{g\in\mathcal{G}_i}\left(\hat{s}_g^p\lambda_i^p+\hat{s}_g^q\lambda_i^q\right)+(\bar{u}_i^p-u_i^{p,0})(r_i^{p,+}+\zeta_i^+r^{op,+})+\\
& (\underline{u}_i^{p,-}-u_i^{p,0})(r_i^{p,-}+\zeta_i^-r^{op,-})+(\bar{u}_i^q-u_i^{q,0})(r_i^{q,+}+\zeta_i^+r^{oq,+})+\\
& (\underline{u}_i^{q,-}-u_i^{q,0})(r_i^{q,-}+\zeta_i^+r^{oq,-})+u_i^{p,0}\lambda_i^p+u_i^{q,0}\lambda_i^q-\frac{1}{4}\mu_{4,i2}+\frac{1}{4}\nu_{4,i}+\\
& \bar{v}_i\lambda_i^{vu}+\underline{v}_i\lambda_i^{vl}+\bar{o}_i^p\lambda_i^{op,-}+\bar{o}_i^q\lambda_i^{oq,-}+\bar{v}_i\underline{v}_i\lambda_i^v+\lambda_i^{op}(\bar{o}_i^p+h_i^p)+\lambda_i^{oq}(\bar{o}_i^q+h_i^q)\Bigg]-\\
& \sum_{k\in\mathcal{A}}W_k\nu_{1,k}-\sum_{k=(i,j,n)\in\mathcal{A}}\Bigg[-\frac{3}{4}\mu_{3,k2}+\frac{5}{4}\nu_{3,k}-\sqrt{\frac{1-\cos\theta_k^u}{(\theta_k^u)^2}}\mu_{3,k1}+\\
& \bar{cs}_k\lambda_k^{cs1}+\underline{cs}_k\lambda_k^{cs2}+\bar{ss}_k\lambda_k^{ss1}+\underline{ss}_k\lambda_k^{ss2}+(\bar{\Delta}_k+\sigma_k)\lambda_k^{\theta1}+(\underline{\Delta}_k+\sigma_k)\lambda_k^{\theta2}+\\
& \left(\cos\underline{\Delta}_k-\frac{\cos\bar{\Delta}_k-\cos\underline{\Delta}_k}{\bar{\Delta}_k-\underline{\Delta}_k}(\underline{\Delta}_k+\sigma_k)\right)\lambda_k^{cs3}+\\
& \left(\sin\frac{\theta_k^u}{2}-\frac{\theta_k^u}{2}\cos\frac{\theta_k^u}{2}\right)(\lambda_k^{ss3}-\lambda_k^{ss4})-\\
& \frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{vv1}-\frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{vv2}-\frac{\underline{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{vv3}-\frac{\bar{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{vv4}+\\
& \frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\cos\theta_k^\delta(\frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}}-\frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}})\lambda_k^{lnc1}-\frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\cos\theta_k^\delta(\frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}}-\frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}})\lambda_k^{lnc2}-\\
& \underline{cs}_k\frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{wc1}-\bar{cs}_k\frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{wc2}-\bar{cs}_k\frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{wc3}-\underline{cs}_k\frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{wc4}-\\
& \underline{ss}_k\frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{ws1}-\bar{ss}_k\frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{ws2}-\bar{ss}_k\frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{ws3}-\underline{ss}_k\frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{ws4}\Bigg] \qquad \text{(B.3a)}\\
\text{s.t.} \quad & \lambda_k^{pt}-\mu_{1,k1}+\lambda_i^p=0 \qquad \forall k=(i,j,n)\in\mathcal{A} \quad \text{(B.3b)}\\
& \lambda_k^{qt}-\mu_{1,k2}+\lambda_i^q=0 \qquad \forall k=(i,j,n)\in\mathcal{A} \quad \text{(B.3c)}\\
& \|\mu_{1,k}\|\leqslant\nu_{1,k} \qquad \forall k\in\mathcal{A} \quad \text{(B.3d)}
\end{aligned}
$$

$$\|\mu_{2,k}\| \leqslant \nu_{2,k} \qquad \forall k \in \mathcal{A} \quad \text{(B.3e)}$$

$$\|\mu_{3,k}\| \leqslant \nu_{3,k} \qquad \forall k \in \mathcal{A} \quad \text{(B.3f)}$$

$$\|\mu_{4,i}\| \leqslant \nu_{4,i} \qquad \forall i \in \mathcal{N} \quad \text{(B.3g)}$$

$$-\sum_{k=(i,j,n)\in\mathcal{A}} \left(\frac{\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{vv1} + \frac{\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{vv2} + \frac{\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{vv3} + \frac{\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{vv4} \right) -$$

$$\sum_{k=(j,i,n)\in\mathcal{A}} \left(\frac{\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{vv1} + \frac{\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{vv2} + \frac{\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{vv3} + \frac{\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{vv4} \right) -$$

$$\mu_{4,i1} - (\bar{v}_i + \underline{v}_i)\lambda_i^v + \lambda_i^{vu} + \lambda_i^{vl} = 0 \qquad \forall i \in \mathcal{N} \quad \text{(B.3h)}$$

$$\sum_{k=(i,j,n)\in\mathcal{A}} \left[-g_k \frac{\lambda_k^{pt}}{\tau_{1,k}^2} + (b_k + \frac{b_k^c}{2})\frac{\lambda_k^{qt}}{\tau_{2,k}^2} \right] - \mu_{4,i2} - \nu_{4,i} + \lambda_i^v + \lambda_i^p g_i^{sh} - \lambda_i^q b_i^{sh} -$$

$$\sum_{k=(i,j,n)\in\mathcal{A}} \left[\frac{\mu_{2,k2}}{\tau_{1,k}^2\sqrt{2}} + \frac{\nu_{2,k}}{\tau_{1,k}^2\sqrt{2}} + \cos\theta_k^\delta \frac{v_j^\delta}{\tau_{2,k}} \left(\frac{\bar{v}_j}{\tau_{2,k}}\frac{\lambda_k^{lnc1}}{\tau_{1,k}^2} + \frac{\underline{v}_j}{\tau_{2,k}}\frac{\lambda_k^{lnc2}}{\tau_{1,k}^2} \right) \right] -$$

$$\sum_{k=(j,i,n)\in\mathcal{A}} \left[\frac{\mu_{2,k3}}{\tau_{2,k}^2\sqrt{2}} + \frac{\nu_{2,k}}{\tau_{2,k}^2\sqrt{2}} + \cos\theta_k^\delta \frac{v_j^\delta}{\tau_{1,k}} \left(\frac{\bar{v}_j}{\tau_{1,k}}\frac{\lambda_k^{lnc1}}{\tau_{2,k}^2} + \frac{\underline{v}_j}{\tau_{1,k}}\frac{\lambda_k^{lnc2}}{\tau_{2,k}^2} \right) \right] = 0 \qquad \forall i \in \mathcal{N} \quad \text{(B.3i)}$$

$$\lambda_k^{vv1} + \lambda_k^{vv2} + \lambda_k^{vv3} + \lambda_k^{vv4} + \lambda_k^{vve} - \lambda_{\tilde{k}}^{vve} - \mu_{2,k1} - \underline{cs}_k\lambda_k^{wc1} - \bar{cs}_k\lambda_k^{wc2} -$$

$$\bar{cs}_k\lambda_k^{wc3} - \underline{cs}_k\lambda_k^{wc4} - \underline{ss}_k\lambda_k^{ws1} - s\bar{s}_k\lambda_k^{ws2} - s\bar{s}_k\lambda_k^{ws3} - \underline{ss}_k\lambda_k^{ws4} = 0 \qquad \forall k \in \mathcal{A} \quad \text{(B.3j)}$$

$$g_k\lambda_k^{pt} - b_k\lambda_k^{qt} - \tan\bar{\Delta}_k\lambda_k^{\tan 1} - \tan\underline{\Delta}_k\lambda_k^{\tan 2} + \lambda_k^{wc1} + \lambda_k^{wc2} + \lambda_k^{wc3} + \lambda_k^{wc4} +$$

$$\lambda_k^{wce} - \lambda_{\tilde{k}}^{wce} + \frac{v_i^\delta v_j^\delta}{\tau_{1,k}\tau_{2,k}}\cos\theta_k^\phi(\lambda_k^{lnc1} + \lambda_k^{lnc2}) = 0 \qquad \forall k \in \mathcal{A} \quad \text{(B.3k)}$$

$$b_k\lambda_k^{pt} + g_k\lambda_k^{qt} + \lambda_k^{ws1} + \lambda_k^{ws2} + \lambda_k^{ws3} + \lambda_k^{ws4} + \lambda_k^{wse} + \lambda_{\tilde{k}}^{wse} +$$

$$\lambda_k^{\tan 1} + \lambda_k^{\tan 2} + \frac{v_i^\delta v_j^\delta}{\tau_{1,k}\tau_{2,k}}\sin\theta_k^\phi(\lambda_k^{lnc1} + \lambda_k^{lnc2}) = 0 \qquad \forall k \in \mathcal{A} \quad \text{(B.3l)}$$

$$-\frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{wc1} - \frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{wc2} - \frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{wc3} - \frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{wc4} +$$

$$\lambda_k^{cs1} + \lambda_k^{cs2} + \lambda_k^{cs3} - \mu_{3,k2} + \nu_{3,k} + \lambda_k^{cse} - \lambda_{\tilde{k}}^{cse} = 0 \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.3m)}$$

$$-\frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{ws1} - \frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{ws2} - \frac{\underline{v}_i\underline{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{ws3} - \frac{\bar{v}_i\bar{v}_j}{\tau_{1,k}\tau_{2,k}}\lambda_k^{ws4} +$$

$$\lambda_k^{ss1} + \lambda_k^{ss2} + \lambda_k^{ss3} + \lambda_k^{ss4} + \lambda_k^{sse} + \lambda_{\tilde{k}}^{sse} = 0 \qquad \forall k = (i,j,n) \in \mathcal{A} \quad \text{(B.3n)}$$

$$\sum_{k=(i,j,n)\in\mathcal{A}} \left(\lambda_k^{\theta 1} + \lambda_k^{\theta 2} - \frac{\cos\bar{\Delta}_k - \cos\underline{\Delta}_k}{\bar{\Delta}_k - \underline{\Delta}_k}\lambda_k^{cs3} - \cos\frac{\theta_k^u}{2}(\lambda_k^{ss3} + \lambda_k^{ss4}) - \right.$$

$$\left. \sqrt{\frac{1-\cos\theta_k^u}{\theta_k^u}}\mu_{3,k1} \right) + \sum_{k=(j,i,n)\in\mathcal{A}} \left(-\lambda_k^{\theta 1} - \lambda_k^{\theta 2} + \frac{\cos\bar{\Delta}_k - \cos\underline{\Delta}_k}{\bar{\Delta}_k - \underline{\Delta}_k}\lambda_k^{cs3} + \right.$$

$$\cos\frac{\theta_k^u}{2}(\lambda_k^{ss3}+\lambda_k^{ss4})+\sqrt{\frac{1-\cos\theta_k^u}{\theta_k^u}}\mu_{3,k1}\Bigg)=0 \qquad \forall i\in\mathcal{N} \qquad \text{(B.3o)}$$

$$\lambda_i^{op,-}\geqslant\lambda_i^p \qquad \forall i\in\mathcal{N} \qquad \text{(B.3p)}$$

$$\lambda_i^{op}\geqslant-\lambda_i^p \qquad \forall i\in\mathcal{N} \qquad \text{(B.3q)}$$

$$\lambda_i^{oq,-}\geqslant\lambda_i^q \qquad \forall i\in\mathcal{N} \qquad \text{(B.3r)}$$

$$\lambda_i^{oq}\geqslant-\lambda_i^q \qquad \forall i\in\mathcal{N} \qquad \text{(B.3s)}$$

$$-y_m^+\leqslant r_i^{p,+}\leqslant y_m^+ \qquad \forall m\in\mathcal{M}, i\in\mathcal{N}_m \qquad \text{(B.3t)}$$

$$-y_m^-\leqslant r_i^{p,-}\leqslant y_m^- \qquad \forall m\in\mathcal{M}, i\in\mathcal{N}_m \qquad \text{(B.3u)}$$

$$-y_m^+\leqslant r_i^{q,+}\leqslant y_m^+ \qquad \forall m\in\mathcal{M}, i\in\mathcal{N}_m \qquad \text{(B.3v)}$$

$$-y_m^-\leqslant r_i^{q,-}\leqslant y_m^- \qquad \forall m\in\mathcal{M}, i\in\mathcal{N}_m \qquad \text{(B.3w)}$$

$$\lambda_i^p-1+y_m^+\leqslant r_i^{p,+}\leqslant\lambda_i^p+1-y_m^+ \qquad \forall m\in\mathcal{M}, i\in\mathcal{N}_m \qquad \text{(B.3x)}$$

$$\lambda_i^p-1+y_m^-\leqslant r_i^{p,-}\leqslant\lambda_i^p+1-y_m^- \qquad \forall m\in\mathcal{M}, i\in\mathcal{N}_m \qquad \text{(B.3y)}$$

$$\lambda_i^q-1+y_m^+\leqslant r_i^{q,+}\leqslant\lambda_i^q+1-y_m^+ \qquad \forall m\in\mathcal{M}, i\in\mathcal{N}_m \qquad \text{(B.3z)}$$

$$\lambda_i^q-1+y_m^-\leqslant r_i^{q,-}\leqslant\lambda_i^q+1-y_m^- \qquad \forall m\in\mathcal{M}, i\in\mathcal{N}_m \qquad \text{(B.3aa)}$$

$$-y_m^+\leqslant r_i^{op,+}\leqslant y_m^+ \qquad \forall m\in\mathcal{M}, i\in\mathcal{N}_m \qquad \text{(B.3ab)}$$

$$-y_m^-\leqslant r_i^{op,-}\leqslant y_m^- \qquad \forall m\in\mathcal{M}, i\in\mathcal{N}_m \qquad \text{(B.3ac)}$$

$$-y_m^+\leqslant r_i^{oq,+}\leqslant y_m^+ \qquad \forall m\in\mathcal{M}, i\in\mathcal{N}_m \qquad \text{(B.3ad)}$$

$$-y_m^-\leqslant r_i^{oq,-}\leqslant y_m^- \qquad \forall m\in\mathcal{M}, i\in\mathcal{N}_m \qquad \text{(B.3ae)}$$

$$\lambda_i^{op}-1+y_m^+\leqslant r_i^{op,+}\leqslant\lambda_i^{op}+1-y_m^+ \qquad \forall m\in\mathcal{M}, i\in\mathcal{N}_m \qquad \text{(B.3af)}$$

$$\lambda_i^{op}-1+y_m^-\leqslant r_i^{op,-}\leqslant\lambda_i^{op}+1-y_m^- \qquad \forall m\in\mathcal{M}, i\in\mathcal{N}_m \qquad \text{(B.3ag)}$$

$$\lambda_i^{oq}-1+y_m^+\leqslant r_i^{oq,+}\leqslant\lambda_i^{oq}+1-y_m^+ \qquad \forall m\in\mathcal{M}, i\in\mathcal{N}_m \qquad \text{(B.3ah)}$$

$$\lambda_i^{oq}-1+y_m^-\leqslant r_i^{oq,-}\leqslant\lambda_i^{oq}+1-y_m^- \qquad \forall m\in\mathcal{M}, i\in\mathcal{N}_m \qquad \text{(B.3ai)}$$

$$y_m^++y_m^-\leqslant 1 \qquad \forall m\in\mathcal{M} \qquad \text{(B.3aj)}$$

$$\sum_{m\in\mathcal{M}}\left(y_m^++y_m^-\right)\leqslant\Gamma \qquad \text{(B.3ak)}$$

$$-1\leqslant\lambda^p\leqslant 1 \qquad \text{(B.3al)}$$

$$-1\leqslant\lambda^q\leqslant 1 \qquad \text{(B.3am)}$$

$$\lambda^{cs1},\lambda^{ss1},\lambda^{ss3},\lambda^{v},\lambda^{vu},\lambda^{\theta 1},\lambda^{vv3},\lambda^{vv4},\lambda^{wc3},\lambda^{wc4},\lambda^{ws3},\lambda^{ws4}\geqslant 0 \qquad \text{(B.3an)}$$

$$\lambda^{op},\lambda^{oq},\lambda^{op,-},\lambda^{oq,-},\lambda^{\tan 1},\nu_1,\nu_2,\nu_3,\nu_4\geqslant 0 \qquad \text{(B.3ao)}$$

$$\lambda^{vl},\lambda^{cs2},\lambda^{cs3},\lambda^{ss2},\lambda^{ss4},\lambda^{\theta 2},\lambda^{vv1},\lambda^{vv2},\lambda^{wc1},\lambda^{wc2},\lambda^{ws1},\lambda^{ws2},\lambda^{\tan 2},\lambda^{lnc1},\lambda^{lnc2}\leqslant 0 \qquad \text{(B.3ap)}$$

$$y^+,y^-\in\{0,1\}^{|\mathcal{M}|}. \qquad \text{(B.3aq)}$$

The formulation above is an exact form of problem (3.14). In this formulation, λ^{pt} and λ^{qt} are the dual variables of line transmission constraints (B.2i) and (B.2j), while λ^{p} and λ^{q} are the dual variables of the equivalent constraints of flow balance constraints (B.2l) and (B.2m) in problem (3.14). For approximation of function cos and sin, we use $\lambda^{cs1}, \lambda^{cs2}, \lambda^{ss1}$ and λ^{ss2} to represent the dual variables of the upper bound and the lower bound of $\widehat{cs}$ and $\widehat{ss}$ respectively, and $\lambda^{cs3}, \lambda^{ss3}$ and λ^{ss4} for constraints (B.2p)-(B.2r). Dual variables λ^{v} correspond to constraint (B.2t). We denote the dual variables for the recourse freedom bounds as $\lambda^{op,-}$, $\lambda^{oq,-}$, λ^{op} and λ^{oq}.

We use λ^{vv1}-λ^{vv4}, λ^{wc1}-λ^{wc4}, λ^{ws1}-λ^{ws4} as the dual variables of McCormick relaxation constraints (B.2u)-(B.2af). The dual variables $\lambda^{\tan 1}$ and $\lambda^{\tan 2}$ correspond to the tangent tightening constraints (B.2ag) and (B.2ah), and λ^{lnc1} and λ^{lnc2} correspond to the lifted nonlinear cuts (B.2ai) and (B.2aj). For variable equality enforcement constraints (B.2ak)-(B.2ao), we use $\lambda^{cse}, \lambda^{sse}, \lambda^{wce}, \lambda^{wse}$ and λ^{vve} as their dual variables. The remaining SOCP constraints (B.2k), (B.2n), (B.2o) are (B.2s) have their dual variables as (μ_1, ν_1), (μ_2, ν_2), (μ_3, ν_3) and (μ_4, ν_4), respectively. Notice that those SOCP constraints can be rewritten in a standard form for duality derivation:

$$
\begin{aligned}
&\|(P_k, Q_k)\|_2 \leqslant W_k && \forall k \in \mathcal{A} \\
&\left\| \left(\widehat{vv}_k, \frac{\hat{v}_i}{\tau_{1,k}^2 \sqrt{2}}, \frac{\hat{v}_j}{\tau_{2,k}^2 \sqrt{2}} \right) \right\|_2 \leqslant \frac{\hat{v}_i/\tau_{1,k}^2 + \hat{v}_j/\tau_{2,k}^2}{\sqrt{2}} && \forall k = (i,j,n) \in \mathcal{A} \\
&\left\| \left(\sqrt{\frac{1-\cos\theta_k^u}{\theta_k^{u\,2}}} (\theta_i - \sigma_k - \theta_j), \widehat{cs}_k - \frac{3}{4} \right) \right\|_2 \leqslant \frac{5}{4} - \widehat{cs}_k && \forall k = (i,j,n) \in \mathcal{A} \\
&\left\| \left(v_i, \hat{v}_i - \frac{1}{4} \right) \right\|_2 \leqslant \hat{v}_i + \frac{1}{4} && \forall i \in \mathcal{N},
\end{aligned}
$$

which means that $\mu_{1,k}$, $\mu_{2,k}$, $\mu_{3,k}$ and $\mu_{4,i}$ are vectors with cardinality $2, 3, 2$ and 2, respectively, while $\nu_{1,k}$, $\nu_{2,k}$, $\nu_{3,k}$ and $\nu_{4,i}$ are scalars. The dual SOCP constraints are formed as (B.3d)-(B.3g).

With the dual variables established, we build constraint (B.3b) for the primal variable P, (B.3c) for Q, (B.3h) for v, (B.3i) for $\hat{v}$, (B.3j) for $\widehat{vv}$, (B.3k) for $\widehat{wc}$, (B.3l) for $\widehat{ws}$, (B.3m) for $\widehat{cs}$, (B.3n) for $\widehat{ss}$, and (B.3o) for θ. Constraints (B.3p)-(B.3s) characterize the dual constraints for primal variables $o^{p,+}, o^{p,-}, o^{q,+}$ and $o^{q,-}$. Constraints (B.3t)-(B.3ai) are the direct replicate of the

linearization constraints of bilinear terms in problem (3.14). Constraints (B.3aj) guarantees that for each bus the uncontrollable injection can be either at the nominal value or at one of the bounds. Constraint (B.3ak) is the budget constraint.

B.4 Bound Tightening Process

Coffrin et al. (2015a,b) introduce a bound tightening process for the QC relaxation involving the bounds that appear in constraints (3.6f) and (3.6h). A new variable, θ_k^d, is created to represent the phase angle difference between two buses of the line $k = (i, j, n) \in \mathcal{A}$, and appended to x. The constraints defining θ_k^d, $\theta_k^d = \theta_i - \theta_j$, are included in the general form linear constraints $Ax \leqslant b$. The process iteratively updates the bounds $v_i = \underline{v}_i$ or $\bar{v}_i$ at bus $i \in \mathcal{N}$ and the phase angle difference $\theta_k^d = \underline{\Delta}_k$ or $\bar{\Delta}_k$ of line $k = (i, j, n) \in \mathcal{A}$, by a set of QC relaxation problems with the objective function substituted by $v_i,\ \forall i \in \mathcal{N}$ or $\theta_k^d,\ \forall k = (i, j, n) \in \mathcal{A}$:

$$\min_{s,x,u} \text{ or } \max_{s,x,u} \quad x_{loc} \tag{B.4a}$$

$$\text{s.t.} \quad \underline{s} \leqslant s \leqslant \bar{s} \tag{B.4b}$$

$$Ax \leqslant b \tag{B.4c}$$

$$\|B_i x + a_i\|_2 \leqslant e_i^\top x + f_i \qquad \forall i = 1, \ldots, m_c \tag{B.4d}$$

$$A^p x = D s^p + u^p \tag{B.4e}$$

$$A^q x = D s^q + u^q \tag{B.4f}$$

$$A^{op} x \leqslant \bar{o}^p + (1 + \alpha^{h,+}) h^p \tag{B.4g}$$

$$A^{oq} x \leqslant \bar{o}^q + (1 + \alpha^{h,+}) h^q \tag{B.4h}$$

$$\underline{u}^p \leqslant u^p \leqslant \bar{u}^p \tag{B.4i}$$

$$\underline{u}^q \leqslant u^q \leqslant \bar{u}^q. \tag{B.4j}$$

In this formulation we treat uncertain uncontrollable injections as decision variables so that the resulting upper and lower bounds are valid for all (u^p, u^q) where $\underline{u}^p \leqslant u^p \leqslant \bar{u}^p$ and $\underline{u}^q \leqslant u^q \leqslant \bar{u}^q$, hence for all $u \in \mathcal{U}$. In the objective function the subscript loc references the position of v_i or

Test Case	Runtime (sec.)
Case 5	1.8
Case 9	3.6
Case 14	8.0
Case 30	29.2
Case 118	809.3
Case 300	5759.0
Case 2383	61001.3
Case 2746	175095.7

Table B.3: Runtime results of the bound tightening process.

θ_k^d in the decision vector x. The bounds are iteratively updated using the optimal solutions from problem (B.4) for each v_i and θ_k^d. The process terminates when changes in the bounds are negligible.

We perform bound tightening as a preprocessing step prior to running optimization. We focus on the bounds on v_i and θ_k^d because the tightness of our linear-quadratic relaxation for the sine and cosine functions and that of the McCormick relaxation for the multi-linear terms depends on the bounds of v_i and θ_k^d. As illustrated in Coffrin et al. (2015b), tightening these bounds tightens the QC relaxation and allows for a tighter lower bound on the nonconvex ACOPF problem. We employ bound tightening in all results reported in Section 3.4. We show the runtimes of the bound tightening process for each test case but do not give detailed improvements from this process beyond indicating here that the optimal values of instances of model (3.7) grow by 1-10% by tightening these simple bounds.

B.5 Regularized Cutting-plane Algorithm

In this section we describe the regularized cutting-plane algorithm mentioned in Section 3.3.3. Given a current incumbent solution, $\hat{s}$, we modify the master problem from model (3.10) by adding a quadratic regularization term, as indicated in model (B.5). In general, the regularization term prevents large changes in incumbent solutions between iterations, which can stabilize the algorithm and encourage faster converge.

$$(M^R) \quad \min \quad c(s^p, s^q) + \frac{\rho}{2}\|(s^p, s^q) - (\hat{s}^p, \hat{s}^q)\|_2^2 \tag{B.5a}$$

$$\text{s.t.} \quad \underline{s} \leqslant s \leqslant \bar{s} \tag{B.5b}$$

$$-\lambda^{p,k\top} D s_i^p - \lambda^{q,k\top} D s^q + z^k \leqslant 0 \qquad \forall k = 1, 2, \ldots \tag{B.5c}$$

However, additional steps need to be taken to obtain a valid lower bound. When an ε-feasible solution is reached, since the regularization term is appended to the master problem as shown in (B.5), $c(\hat{s}^p, \hat{s}^q)$ may not be a lower bound on the optimal value of model (3.7). However, we can solve the original master problem (3.10) with all the feasibility cuts (but without the regularization term) to obtain a valid lower bound, V^*. Although a valid lower bound is obtained, the solution $(\tilde{s}^p, \tilde{s}^q)$ of this non-regularized master problem may not be equal to $(\hat{s}^p, \hat{s}^q)$, and it may not be an ε-feasible solution. The algorithm needs to proceed until we obtain a ε-feasible solution from solving the regularized master problem and the difference between V^* and $c(\hat{s}^p, \hat{s}^q)$ is negligible (less than some tolerance η) so that we can approximate the lower bound value with the cost of this ε-feasible solution. The modified algorithm is presented as Algorithm 4.

Algorithm 4 Regularized cutting-plane algorithm for model (3.7)

Let (M^R) denote regularized master (B.5) and (M) denote non-regularized master (3.10); initialize iteration number $k := 1$, tolerances $\varepsilon, \eta > 0$, and regularization weight, $\rho > 0$;
Solve (M^R) and obtain solution $(\hat{s}^{p,k}, \hat{s}^{q,k})$ and optimal value V^*;
Solve (SDI) with $(\hat{s}^p, \hat{s}^q) = (\hat{s}^{p,k}, \hat{s}^{q,k})$ and obtain solution $(\lambda^{p,k}, \lambda^{q,k})$ and optimal value z^k_{feas};
while $z^k_{feas} > \varepsilon$ or $\frac{UB - V^*}{V^*} > \eta$ **do**
    Append $z^k_{feas} - \lambda^{p,k\top} D(s^p - \hat{s}^{p,k}) - \lambda^{q,k\top} D(s^q - \hat{s}^{q,k}) \leqslant 0$ to constraints (B.5c) of (M^R), (3.10c) of (M);
    Let $k := k + 1$;
    Solve (M^R) and obtain solution $(\hat{s}^{p,k}, \hat{s}^{q,k})$;
    if (M^R) is feasible **then**
        Solve (SDI) with $(\hat{s}^p, \hat{s}^q) = (\hat{s}^{p,k}, \hat{s}^{q,k})$ and obtain solution $(\lambda^{p,k}, \lambda^{q,k})$ and optimal value z^k_{feas};
        if $z^k_{feas} \leqslant \varepsilon$ **then**
            Obtain optimal value $UB = c(\hat{s}^{p,k}, \hat{s}^{q,k})$;
            Solve (M) and obtain solution $(\tilde{s}^p, \tilde{s}^q)$ and optimal value V^*;
            Solve (SDI) with $(\hat{s}^p, \hat{s}^q) = (\tilde{s}^p, \tilde{s}^q)$ and obtain solution $(\lambda^{p,k}, \lambda^{q,k})$ and optimal value z^k_{feas};
    else
        Stop and return the status of infeasibility;
end while
Output V^* as lower bound on optimal value of model (3.7), and output $(\hat{s}^{p,k}, \hat{s}^{q,k})$ as an ε-feasible solution.

Table B.4 compares the computational performance of Algorithms 3 and 4 on Cases 118

and 300 with $\rho = 0.1, 1, 10$ and $\eta = 10^{-4}$. It takes more than 300 iterations for Algorithm 3 to reach an ε-feasible solution, with both the violation and the lower bound improving slowly. We can see that adding a regularization term may decrease the number of iterations to convergence, but the average time for each iteration increases as ρ increases. To understand this effect, we plot

Parameters	No. of iterations		ε-feasibility achieved		Time (sec.)	
	Case 118	Case 300	Case 118	Case 300	Case 118	Case 300
$\rho = 0$	300	300	No	No	2414	23042
$\rho = 0.1$	300	300	No	No	2640	30480
$\rho = 1$	178	300	Yes	No	2114	33950
$\rho = 10$	300	226	No	Yes	4292	27757

Table B.4: Computational results for solving instances of model (3.7) for Cases 118 and 300 with Algorithms 3 and 4 with $\Gamma = 3$ and with a limit of 300 iterations.

the violation (in base-10 log scale) and lower bound as a function of the iteration for Case 118 in Figure B.1. The red dots represent the value of UB corresponding to the ε-feasible solutions from running Algorithm 4. Without regularization, the master solution in the next iteration tends to move far from the incumbent solution. The corresponding cuts provide a global characterization of the feasible region, but it takes a long time to generate enough cuts to obtain an ε-feasible solution. On the other hand, the regularized algorithm tends to generate cuts within a local area, as the new probing solution is close to the incumbent and moves quickly towards the feasible region. It takes longer to solve (SDI) at a solution closer to the feasible region, which leads to a longer average time per iteration for Algorithm 4.

Every time an ε-feasible solution is obtained, the non-regularized master problem (M) is solved to generate a lower bound. If there is still a large enough gap between the cost of that solution and the lower bound, the algorithm moves to the incumbent solution of (M), which may lead to a large feasibility violation. This explains the large spikes in the plots of $\rho = 1$ and $\rho = 10$ in Figure B.1. The process between two spikes can be considered as exploitation of a local area. When $\rho = 10$, there are many spikes which indicates that the algorithm reaches an ε-feasible solutions frequently, but in this case the cuts generated only characterize the feasible region locally, which eventually requires many rounds of exploitation before convergence. Even with an appropriately chosen ρ, the computational performance of Algorithm 4 is inferior to the scenario-appending technique presented

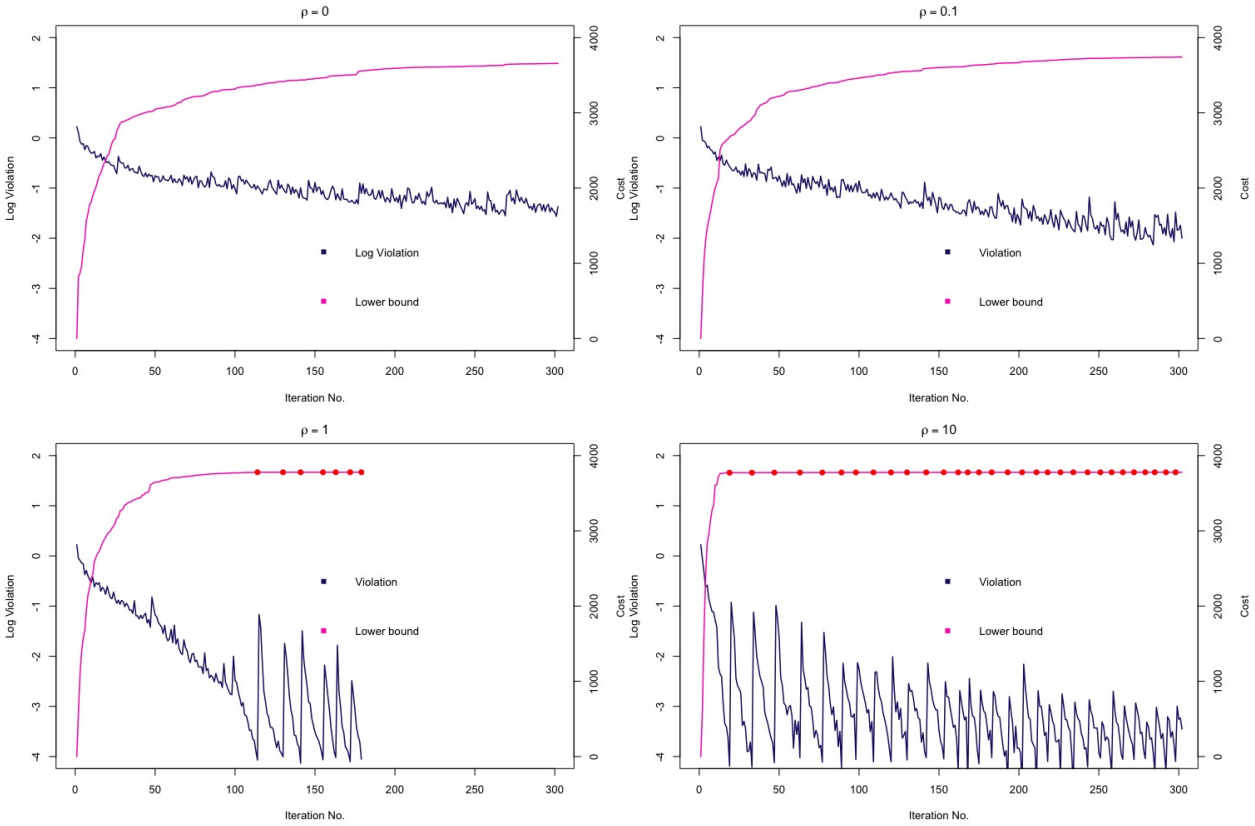


Figure B.1: Computational performance of Algorithms 3 and 4 for Case 118 with $\rho = 0, 0.1$, 1 and 10.

in Section 3.3.3.

Appendix C

Appendices for Chapter 4

C.1 Covariates for the PWID Population

When a client initiation occurs in the simulation, we draw a client at random, with replacement, from our dataset of 5,903 unique clients. In this appendix, we detail those attributes and indicate the fraction of the population with each attribute.

Gender:

Male	Female	Transgender
0.6947	0.3049	0.0004

Ethnicity:

White	African American	Puerto Rican	Mexican	Other Latino	Other
0.5158	0.2345	0.1478	0.0617	0.0122	0.0280

Snort before injection:

Yes	No
0.3446	0.6554

Participation in shooting galleries:

Yes	No
0.0864	0.9136

Participation in treatment programs:

	Currently in	Been in	Tried to get into	Interested in
Yes	0.1011	0.1870	0.0923	0.4738
No	0.8989	0.8130	0.9077	0.5362

Drugs used in the past 30 days:

	Speedball	Heroin	Cocaine	Ritalin Heroin	Other
Yes	0.0486	0.9582	0.0581	0.0005	0.0185
No	0.9514	0.0418	0.9419	0.9995	0.9815

Source of syringes:

	Family	Friends	Acquaintance	Strangers	Other SEP	Other
Yes	0.0586	0.2529	0.0530	0.0163	0.1360	0.6131
No	0.9414	0.7471	0.9470	0.9837	0.8640	0.3869

Reuse own syringes:

Yes	No
0.1579	0.8521

Use syringes behind others:

Yes	No
0.1972	0.8028

For each of the continuous factors, we present descriptive statistics and histograms of their distributions. We use μ to denote the mean of the factor and σ to denote the standard deviation.

- Age: $\mu = 34.79,\ \sigma = 11.22$;
- Age of clients at their first injection: $\mu = 23.44,\ \sigma = 7.86$;

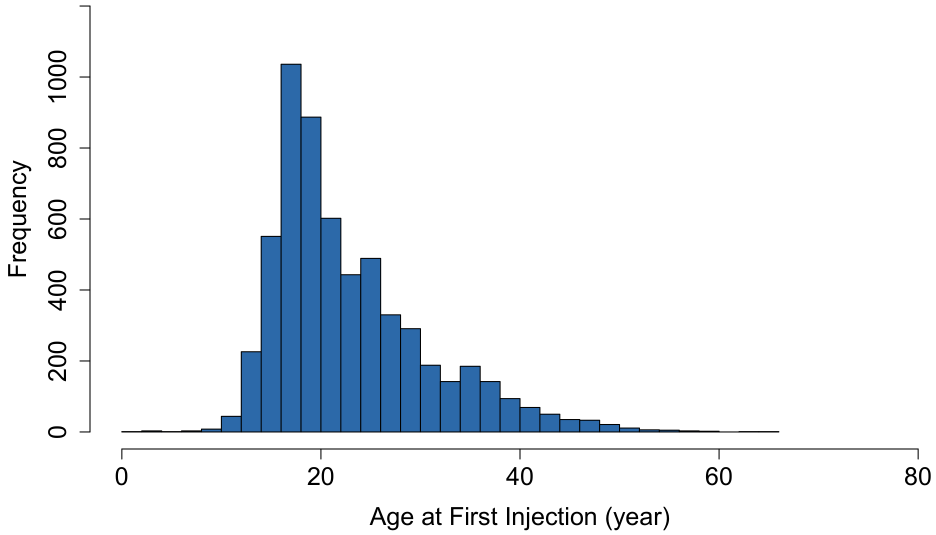


Figure C.1: Distribution of the age of clients at their first injection

- Length of drug injection history: $\mu = 11.36,\ \sigma = 11.36$;

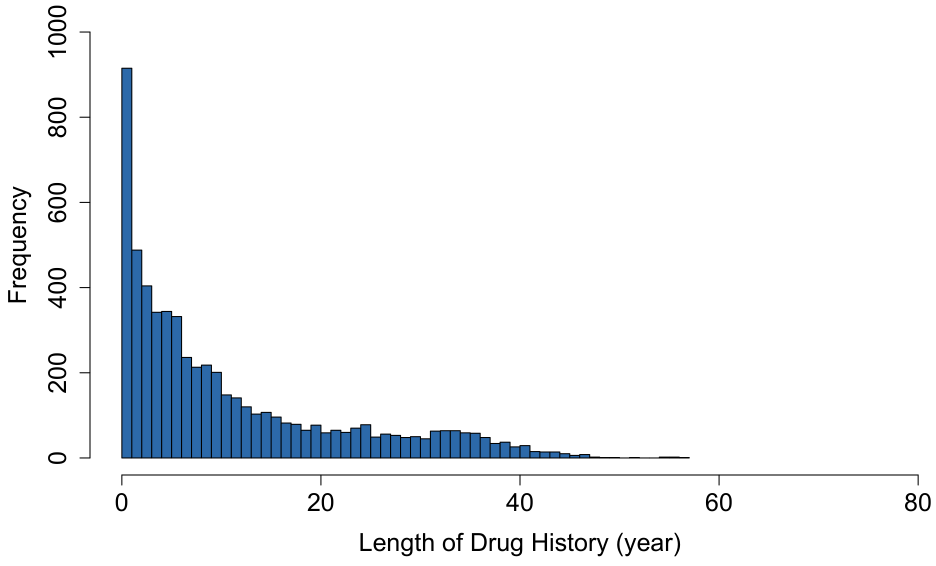


Figure C.2: Distribution of the length of drug injection history

- Number of daily drug injections: $\mu = 2.77,\ \sigma = 1.87$;

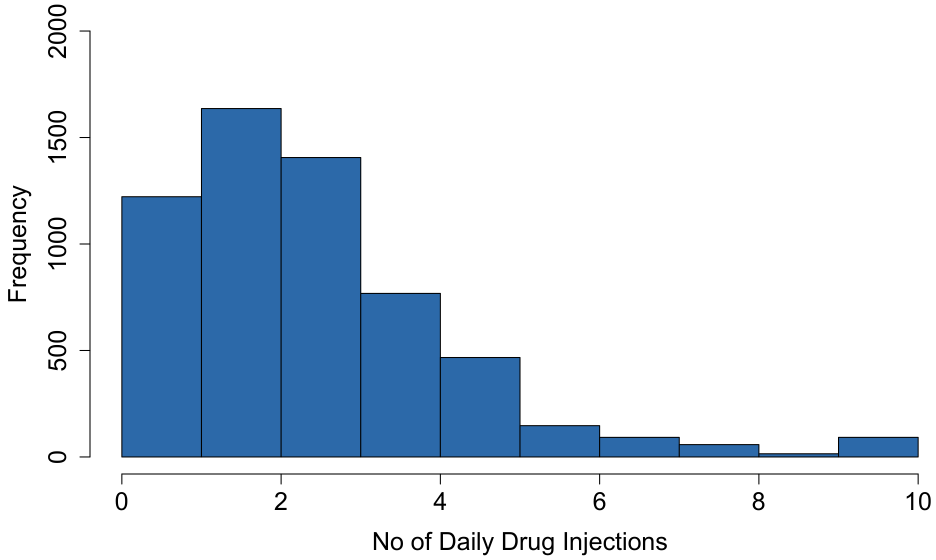


Figure C.3: Distribution of the daily drug injections

- Number of times reusing own syringes in 30 days: $\mu = 1.61,\ \sigma = 6.15$

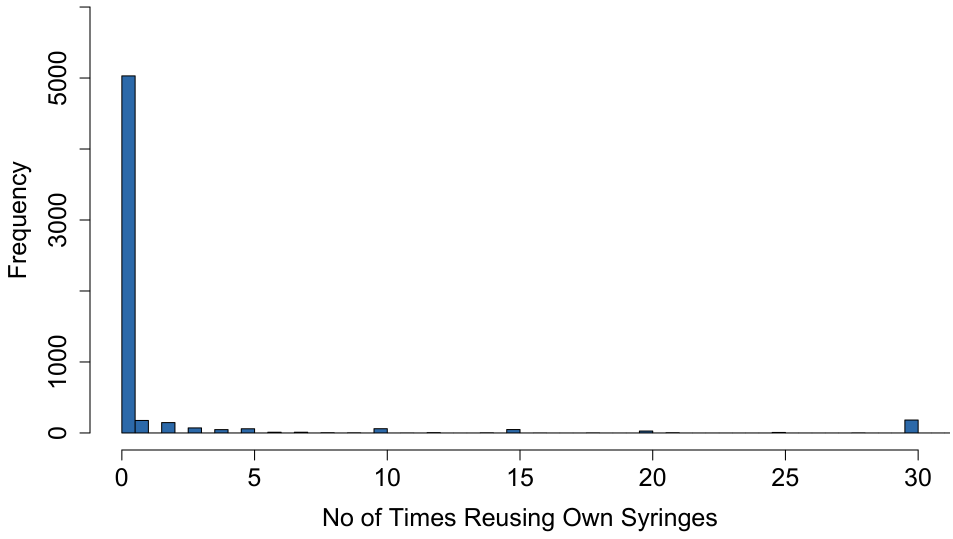


Figure C.4: Distribution of the number of times reusing own syringes in 30 days

- Number of times using others' used syringes in a 30 days: $\mu = 0.34,\ \sigma = 1.29$

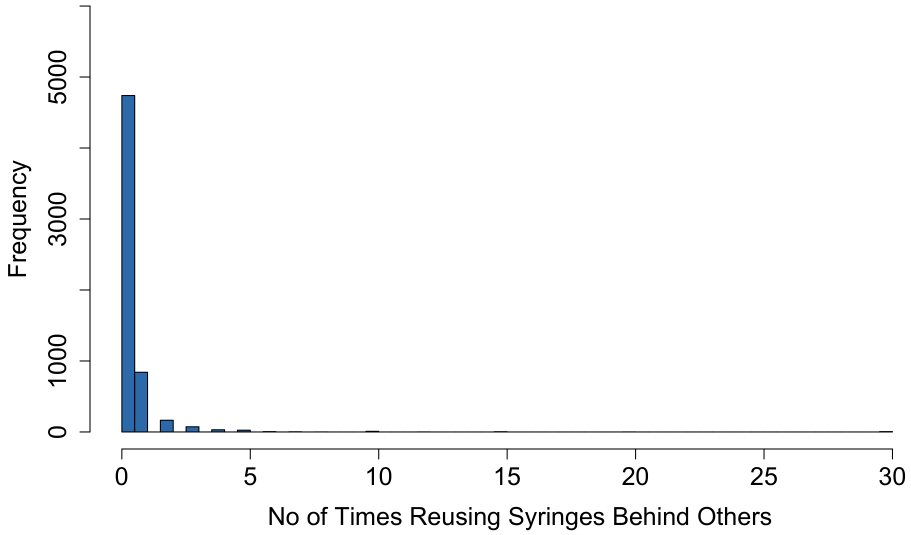


Figure C.5: Distribution of the number of times using others' used syringes in a 30 days

- Number of times visiting the area of service locations in 30 days: $\mu = 23.88,\ \sigma = 9.75$

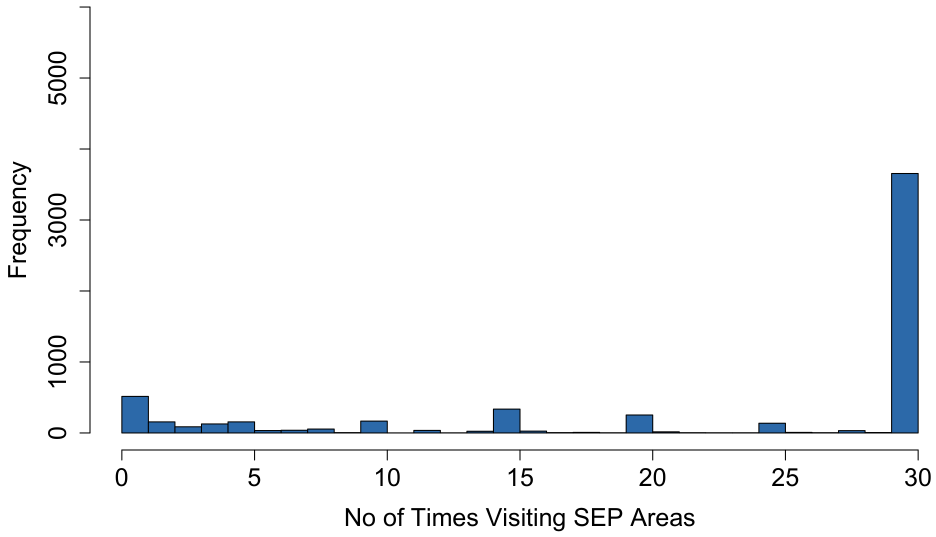


Figure C.6: Distribution of the number of times visiting the area of service locations in 30 days

C.2 Statistical Significance Results of Fitted Parameters

We aim to test the significance of each fitted coefficient, by using a bootstrap resampling scheme. We take 100 bootstrap replicates with the same size of the data. For each replicate, we solve the nonlinear optimization model (4.11) and obtain a set of coefficients. Then we calculate Δ and T, as the inter-arrival time change and the expected sojourn time change as described in Section 4.4.2, for each replicate based on its fitted coefficients. We count how many of those 100 replicates of $b_{i,j}$, $g_{i,j}$, ρ_j, Δ_j and T_j are positive. If a large portion of them, like 90%, is positive or negative, then we can conclude that the parameter is significantly nonzero.

Each column in Table C.1 shows the number of samples, among 100 samples, of which the fitted coefficient is positive. In addition to acronyms defined in the main text, FUSBO stands for frequency of using syringes behind (i.e., after) others.

	Factor j	ρ_j	$b_{1,j}$	$g_{1,j}$	$g_{2,j}$	Δ	T
	Snort	36	16	89	54	67	69
	Gallery	100	7	57	77	88	1
	From Other Locations	97	13	81	96	20	0
	From Other SEP	1	96	7	59	13	98
	From Family	68	55	45	27	73	40
	From Friends	5	52	26	0	100	100
	From Acquaintance	41	63	53	45	41	48
	From Strangers	9	57	62	93	0	49
	Speedball	19	1	91	4	100	99
	Heroin	0	15	67	10	96	100
	Cocaine	42	54	52	56	34	48
	Ritalin Heroin	29	32	3	13	65	64
	Other Drug	62	53	29	37	75	58
	In Treatment	0	94	0	64	20	99
	Been in Treatment	9	10	90	32	88	98
	Attempted Treatment	41	30	83	7	95	84
	Want Treatment	73	21	4	22	100	74
	Female	45	77	17	93	8	7
	Male	100	75	25	85	20	20
	Transsexual	99	99	99	99	99	15
	White	100	96	46	80	2	2
	African American	100	19	30	64	70	2
	Puerto Rican	0	60	99	68	12	100
	Mexican	0	42	64	99	1	82
	Other Latino	27	63	0	44	77	82
	Other	10	82	12	43	50	91
	Age	44	57	7	89	16	55
	Age of First Drug Use	52	100	68	100	0	41
	Drug Use Span	60	6	2	23	100	43
	FUD	65	98	15	5	55	31
	FROS	91	12	54	61	78	19
	FUSBO	34	71	35	63	26	50
	FBSA	0	51	100	99	0	100

Table C.1: Statistics of bootstrap samples for estimating coefficients ρ, b, and g, as well as changes to the mean time to return to an SEP site, Δ, and the expected sojourn time in the system, T