Statistical Analysis of Baggage Handling Quality in the Aviation Industry based on RFID Tags

Shahbazi, Shima

DOI (link to publication from Publisher):
10.5278/vbn.phd.engsci.00037

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

STATISTICAL ANALYSIS OF BAGGAGE HANDLING QUALITY IN THE AVIATION INDUSTRY BASED ON RFID TAGS

BY
SHIMA SHAHBAZI

DISSERTATION SUBMITTED 2015

AALBORG UNIVERSITY
DENMARK

Statistical Analysis of Baggage Handling Quality in the Aviation Industry based on RFID Tags

Ph.D. Dissertation

Shima Shahbazi

Dissertation submitted 2015

AALBORG UNIVERSITY
DENMARK

Department of Mathematical Sciences

Thesis submitted:	September 2015
PhD supervisor:	Assoc. Prof. Kasper Klitgaard Berthelsen Aalborg University
Assistant PhD supervisor:	Assoc. Prof. Esben Høg Aalborg University
PhD committee:	Professor Jesper Møller (chairman) Aalborg University
	Professor Jens Lysgaard Aarhus University
	Associate Professor Bo Friis Nielsen Technical University of Denmark
PhD Series:	Faculty of Engineering and Science, Aalborg University

ISSN (online): 2246-1248
ISBN (online): 978-87-7112-359-3

Published by:
Aalborg University Press
Skjernvej 4A, 2nd floor
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2015

Thesis Details

Thesis Title:	Statistical Analysis of Baggage Handling Quality in the Aviation Industry based on RFID Tags
Ph.D. Student:	Shima Shahbazi
Principal Supervisor:	Assoc. Prof. Kasper Klitgaard Berthelsen, Aalborg University
Secondary Supervisor:	Assoc. Prof. Esben Høg, Aalborg University

The main body of this thesis consists of the following papers.

[A] Shima Shahbazi, Esben Høg, Kasper K. Berthelsen, "Optimizing RFID Tagging in the Aviation Industry," *SIAM Journal on Optimization* (under revision).

[B] Shima Shahbazi, Kasper K. Berthelsen, "Review of Statistical Models for Analyzing RFID Data in the Aviation Industry," *Journal of Applied Statistics* (under revision).

This thesis has been submitted for assessment in partial fulfillment of the PhD degree. The thesis is based on the submitted scientific papers which are listed above. Parts of the papers are used directly or indirectly in the extended summary of the thesis. As part of the assessment, co-author statements have been made available to the assessment committee and are also available at the Faculty. The thesis is not in its present form acceptable for open publication but only in limited and closed circulation as copyright may not be ensured.

Preface

This thesis is the result of my Ph.D. studies under supervision of Assoc. Prof. Kasper Klitgaard Berthelsen and Assoc. Prof. Esben Høg at the department of Mathematical Sciences, Aalborg University, Denmark. The topic of this thesis is studying baggage handling in the aviation industry. The study is based on data from some airports in the Scandinavian area provided by Lyngsoe Systems.

My Ph.D. research is a statistical part of a project called BagTrack funded by the Danish National Advance Technology Foundation. The main statistical achievements are developing strategies for bag tagging by using a stochastic optimization problem and analyzing the handling quality by using a statistical model.

With regard to my past background in statistics, I needed to make myself more educated with a number of basic concepts, definitions and methods in optimization theory. These concepts, definitions and methods are presented in the first chapter of the thesis and provide a path to fully understand the stochastic optimization programming.

The thesis is a collection of two papers with an introductory chapter describing the relevant background for reading the papers. Accordingly, chapter 1 is divided into two main parts each relevant to one of these papers, and chapter 2 contains the papers with the following titles:

- Optimizing RFID Tagging in the Aviation Industry
- Review of Statistical Models for Analyzing RFID Data in the Aviation Industry

The papers have been submitted to SIAM Journal on Optimization and Journal of Applied Statistics respectively.

I wish to express my sincere thanks to my supervisor, Assoc. Prof. Kasper Klitgaard Berthelsen, for sharing his expertise and being generous with his time all through my study. I am also greatly grateful to my co-supervisor, Assoc. Prof. Esben Høg, for his useful comments and valuable guidance and encouragement extended to me. I consider myself very fortunate for having a chance to work with a group of nice colleagues including senior researchers,

post-docs, PhD students and secretaries in the department of Mathematical Sciences and I appreciate the positive work environment they made for me.

Shima
Aalborg University, September 2, 2015

Summary

Every year, baggage mishandling costs a considerable amount of money for the aviation industry. This problem has led some airports and airlines to use more updated technologies in their handling system to improve baggage handling and possibly cut down on mishandled bags. Regarding this issue, some airports and airlines in the Scandinavian area have started using Radio Frequency Identification (RFID) technology to tag and track the bags.

The following thesis mainly deals with two research questions arising before and after using RFID technology. The first one is determining RFID tagging rates in the airports under the study based on a limited amount of budget, and the second one is determining a probability model which describes the quality of baggage handling based on the data obtained from RFID technology.

Since the airports under study have a limited budget to tag the bags, and tagging with RFID is more expensive than the other common types of tagging, we consider tagging a (random) subset of the bags by RFID. Accordingly, we define RFID tagging rates at each of these airports in a way to attain the best desired outcome. This leads to an optimization problem.

The defined optimization problem involves parameters which vary from one time period to another time period. Therefore, we need to study different approaches to deal with the randomness corresponding to the problem-specific parameters. The final proposed problem is a stochastic optimization problem including both a stochastic objective function and a chance constraint. These two together are uncommon as well as challenging.

One method for solving the defined stochastic optimization problem is the method of sample average approximation. This method leads to an interval that contains the optimal value of the problem. Further, we suggest a method based on an extension of the majorization definition. This method also leads to an interval for the optimal value of the problem. The results of the two methods are presented and compared with each other in paper A.

The second challenge that we deal with in this thesis is the analysis of the quality of baggage handling based on the data obtained from RFID technology. Regarding this issue, the mishandling problem in one specific airport is

studied. The study is limited to the transfer bags which are left-behind in this airport. A transfer bag has a statue of either left-behind or not left-behind. Accordingly, a binary random variable is defined which represents the transfer bags' statuses in the airport. This leads us to study statistical models for analyzing binary random variables.

Inspired by the data, we introduce a logit-nonlinear relationship between the probability of being left-behind and the corresponding connection time of the transfer bags. The logistic-nonlinear regression model is fitted to the observed data through a Bayesian approach. We suggest an informal visual test to check the validity of this model. The test is based on the linearized form of Ripley's K-function. This function let us see the clustering pattern of the bags' statuses.

The high clustering pattern suggests that there is a source of dependency, other than connection times, between the bags' statuses. That means the assumptions of the logistic-nonlinear regression model are violated. It is well-known that the correlation between binary data and invalidity of the assumptions of the logistic regression model cause overdispersion phenomenon. Therefore, we need to study statistical models appropriate for modeling overdispersed data. Finally, a beta-binomial logistic-nonlinear regression model is fitted to the observed data through a Bayesian approach. The results are presented in paper B.

Sammendrag

Hvert år påføres luftfartsindustrien en betragtelig udgift som følge af fejlhåndtering af bagage. Dette problem har fået nogle lufthavne og flyselskaber til at indføre ny teknologi til håndtering af bagage for at forbedre procedurerne og nedbringe antallet af fejlhåndteringer. I denne forbindelse har nogle af de skandinaviske lufthavne og flyselskaber introduceret *Radio Frequency Identification* (RFID) – teknologien til mærkning og sporing af bagage.

I nærværende afhandling arbejdes der primært med to spørgsmål af videnskabelig interesse, som opstår før og efter brugen af RFID. Det første spørgsmål omhandler fastsættelse af de andele af bagage der skal mærkes med RFID under hensyntagen til et begrænset budget. Det andet omhandler opstilling af en sandsynlighedsteoretisk model, som kan beskrive kvaliteten af bagagehåndteringen, baseret på de data, som RFID teknologien giver.

Da de involverede lufthavne råder over et begrænset budget til mærkning af bagage – og da RFID teknologiens bagagemærkning er dyrere end andre anvendte typer mærkning, så vil vi med hensyn til det første spørgsmål ovenfor betragte en (tilfældig) delmængde af bagagen, som mærkes med RFID. Vi angiver en RFID mærknings-rate for hver af de involverede lufthavne for at opnå det bedst mulige resultat. Dette leder efterfølgende til et optimeringsproblem.

Optimeringsproblemet involverer parametre som varierer over tid. Det opstillede problem bliver et stokastisk optimeringsproblem, som indeholder både en stokastisk objektfunktion og stokastiske restriktioner.

Vi beskriver to metoder til at løse det stokastiske optimeringsproblem: Den første er baseret på den såkaldte *method of sample average approximation*, og den anden er baseret på en udvidelse af definitionen af såkaldt *majorisering*. Begge metoder fører til beregning af et interval, som indeholder det stokastiske optimeringsproblems optimale værdi. Resultaterne af de to metoder præsenteres og sammenlignes i artikel A.

Med henblik på besvarelse af det andet spørgsmål ovenfor, som vedrører kvaliteten af bagagehåndteringen, fokuseres der på resultaterne i én specifik lufthavn. Studiet er afgrænset til at omhandle transfer-bagage, som fejlagtigt ikke kommer med på flyet. Transfer-bagage har status som enten efterladt

(*left-behind*) eller ikke-efterladt *not left-behind*. I overensstemmelse hermed defineres en binær stokastisk variabel, som repræsenterer transfer-bagagens status.

Inspireret af data fra lufthavnen introducerer vi en logistisk-ikke-lineær model for relationen mellem sandsynligheden for at en bagage-enhed er *left-behind* og transfer tid for bagage.

Modellen tilpasses data via en Bayesiansk analyse. Vi foreslår en uformel visuel test som modelkontrol. Testen er baseret på en lineariseret udgave af Ripley's K-funktion. Testen antyder at transfer tid ikke er tilstrækkelig til at forklare transfer-bagages status, således at antagelserne om den logistisk-ikke-lineære model er brudte. Modellen forbedres så den kan tage højde for overspredning, hvorved en beta-binomial logistisk-ikke-lineær model tilpasses data, igen via en Bayesiansk analyse. Resultaterne præsenteres i artikel B.

Contents

Part I

Introduction and Overview of methods

Introduction

The topic of this thesis is studying baggage handling in the aviation industry. Today, the problem of mishandling is an important issue in this industry. This problem has led some airports and airlines to use more updated technologies in their handling system. Regarding this issue, some airports and airlines in the Scandinavian area have started using Radio Frequency Identification (RFID) technology to track the bags. Particularly, in this thesis, we study two things. One is determining RFID tagging rates in the airports under the study based on a limited amount of budget. This is done by solving a proper optimization problem. The other is determining a probability model which describes the quality of baggage handling in one of these airports. This is done by using a beta-binomial logistic-nonlinear regression model.

This introductory chapter is divided in two main parts. In the first part, we study mathematical optimization problems in general. Specifically, we study modeling a stochastic optimization problem and the methods to solve them. These methods have been used in paper A for finding RFID tagging rates. In the second part, we study binary correlated random variables and the methods to deal with them. A beta-binomial logistic-nonlinear regression model is used in paper B to describe these types of data.

1 Mathematical optimization

A *mathematical optimization* problem (also known as a *mathematical programming* problem) is the problem of maximizing or minimizing a real function regarding a set of data. To solve an optimization problem, there are some calculus based (analytic) methods which lead to closed form formulas for solving a problem e.g., the method of Lagrange multipliers and some numerical methods which use an iterative algorithm to attain the optimal value, e.g., the interior point methods. Some well-known categories of optimization problems are constrained or unconstrained, linear or nonlinear, convex or non-convex, and finally deterministic or stochastic problems. The main topic of this chapter is reviewing the stochastic optimization problems. These

types of optimization problems have been developed in recent years rapidly and can be solved by some advanced methods.

Stochastic optimization is a class of optimization problems involving some uncertain parameters[1] while it is assumed that the parameters' statistical distributions are known or at least can be estimated. The theory behind this is a combination of optimization theory and statistical theory. Due to the existence of random parameters in the stochastic problems, these types of problems are very practicable, and they arise in many applications, see, e.g., Wallace and Ziemba (2005).

This chapter is divided into three parts. First, some basic concepts and methods in optimization problems are studied. Then, some types of stochastic problems are introduced. Finally, methods to solve the introduced stochastic problems are explained.

1.1 Introduction to optimization

Optimization is an old branch of mathematics. Mathematical optimization is the science of determining the best (maximum or minimum) solution to a mathematical formulation. One example is the problem of finding a solution to the following formulation:

$$\begin{aligned} &\min_{\mathbf{x}} \ F(\mathbf{x}) \\ &\text{subject to:} \\ &\begin{cases} g_j(\mathbf{x}) \leq 0, j = 1, 2, \ldots, m, \\ h_j(\mathbf{x}) = 0, j = 1, 2, \ldots, p, \end{cases} \end{aligned} \tag{1}$$

where $F : \mathbb{R}^n \to \mathbb{R}$ and $g_j(\mathbf{x}), j = 1, 2, \ldots, m$ and $h_j(\mathbf{x}), j = 1, 2, \ldots, p$ are all scalar functions of the vector $\mathbf{x}$. The above problem is a minimization problem, and it can simply be reformulated as a maximization one.

The components $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ are called the *decision variables*, and the function $F(\mathbf{x})$ is called the *objective function*. The functions $g_j(\mathbf{x})$ and $h_j(\mathbf{x})$ denote the inequality and equality constraints respectively. Any $\mathbf{x}$ which satisfy all of the constraints is a *feasible point*, and the set of all these points is called the *feasible set*.

If some variables in problem (1) are restricted to be integer, it is called a *mixed-integer optimization* problem.

There are a wide range of algorithms for solving an optimization problem. For example, the simplex algorithm, interior point methods and subgradient methods which are used to solve linear and nonlinear convex programming problems and the branch and bound algorithm designed to solve

[1]The word "parameter" is not used in its common meaning in the context of probability and statistics but as its common meaning in the optimization theory.

mixed-integer programming problems. These methods are briefly discussed in sections 1.1.2, 1.1.3 and 1.1.4. Earlier, the definition of convexity is given in section 1.1.1.

1.1.1 Convexity (Concavity)

The convexity (concavity) of the objective function and the convexity (non-convexity) of the feasible set play an important role in a minimization (maximization) problem. A convex (concave) function has at most one minimum (maximum) point. This property simplify the optimization problem. On the other hand, the powerful tool of convex analysis can be used for an optimization problem with a convex feasible set. In this section, the definitions of convex (concave) functions and convex (non-convex) sets are given. Also, a useful theorem to investigate the convexity of a function is presented. We use these definitions in the next sections specifically in section 1.2.2, where we discuss the convexity of a chance constraint problem.

The definitions of convex set and function are well known in the literature, and the definition of an α-convex function was first introduced by Avriel (1972).

Definition 1. *A set $\mathcal{C} \subset \mathbb{R}^n$ is said to be convex if for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{C}$ and $0 \leq \lambda \leq 1$, we have $\lambda \mathbf{x}_1 + (1-\lambda)\,\mathbf{x}_2 \in \mathcal{C}$. Otherwise, it is called a non-convex set.*

Definition 2. *Fix $-\infty \leq \alpha \leq \infty$, a function, $f(\mathbf{x})$, defined on a convex set $\mathcal{C} \subset \mathbb{R}^n$ is said to be α-convex if for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{C}$ and $0 \leq \lambda \leq 1$, the following holds:*

$$f(\lambda \mathbf{x}_1 + (1-\lambda)\,\mathbf{x}_2) \leq m_\alpha(f(\mathbf{x}_1), f(\mathbf{x}_2), \lambda),$$

where,

$$m_\alpha(f(\mathbf{x}_1), f(\mathbf{x}_2), \lambda) = \begin{cases} f(\mathbf{x}_1)^\lambda f(\mathbf{x}_2)^{(1-\lambda)} & \text{if } \alpha = 0, \\ \min\{f(\mathbf{x}_1), f(\mathbf{x}_2)\} & \text{if } \alpha = \infty, \\ \max\{f(\mathbf{x}_1), f(\mathbf{x}_2)\} & \text{if } \alpha = -\infty, \\ \left(\lambda f(\mathbf{x}_1)^\alpha + (1-\lambda) f(\mathbf{x}_2)^\alpha\right)^{1/\alpha} & \text{otherwise.} \end{cases}$$

In the case of $\alpha = 0$, the function is also assumed to be non-negative.

For $\alpha = 1, 0, -\infty$, this is simply called a convex, a log-convex and a quasi-convex function respectively.

In a similar way, an α-concave function is defined, see, e.g., Shapiro et al. (2009, p. 94).

Definition 3. *For $-\infty \leq \alpha \leq \infty, \alpha \neq 0$, a function is called α-concave if the negative of the function is α-convex. When $\alpha = 1$ and $\alpha = -\infty$, this is simply called a concave and a quasi-concave function respectively.*

In the case of $\alpha = 0$, a non-negative function is 0-concave (also known as log-concave) if $\log(f(.))$ is a concave function.

Every non-negative concave function is log-concave. In general, it can be shown that for a non-negative function, α-concavity entails β-concavity if $\beta \leq \alpha$, see, e.g., Shapiro et al. (2009, p. 96). Therefore, a non-negative concave function is also quasi-concave, but the reverse is not necessarily true.

The next theorem helps us to recognize convex functions by their closure properties. The proof can be found in Lange (2004, Chap. 5).

Theorem 1. *Convex functions satisfy the following:*

(a) The sum of non-negative convex functions is convex.

(b) If $f(\mathbf{x})$ and $g(\mathbf{x})$ are convex functions, then $\max\{f(\mathbf{x}), g(\mathbf{x})\}$ is also convex.

(c) An affine function $f(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$ for fixed values of A and $\mathbf{b}$ is convex.

(d) If $f(\mathbf{x})$ is convex, then the composition of $f(\mathbf{x})$ and an affine function $g(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$ for fixed values of A and $\mathbf{b}$ is convex.

1.1.2 Linear programming (LP)

A standard linear optimization problem in matrix form is written as

$$\min_{\mathbf{x}} \left\{ \mathbf{c}^T\mathbf{x} \,|\, A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq 0 \right\},$$

where A is an $m \times n$ matrix and $\mathbf{x}, \mathbf{c}$ are $n \times 1$ vectors and $\mathbf{b}$ is an $m \times 1$ vector. Any linear inequality constraint can simply be reformulated as an equality by using *slack* variables or *surplus* variables. So, there is no loss of generality in only considering equality constraints.

Since $A\mathbf{x} = \mathbf{b}$ is a system of linear equations, it has a solution if and only if the number of linearly independent columns in the matrix A is not less than the number of independent columns in the matrix $A|\mathbf{b}$. Assume that there is at least one feasible solution and $n \geq m$. If we set $(n - m)$ variables to zero and solve m equations with m independent variables, the solution is called a *basis* solution. The variables that are chosen to be non-zero are called *basis variables*, and the rest are called *non-basis*. Two basis feasible solutions are *adjacent*, if they have $(n - 1)$ common basis variables. It can be shown that with a finite number of linear inequality constraints, there can only be a finite number of basis feasible solutions, see, e.g., Bertsimas and Tsitsiklis (1997, p. 52). In geometric terms, basis solutions correspond to *extreme points* of the *polyhedron* $P = \{\mathbf{x} \,|\, A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq 0\}$.

It can be shown that the optimal value in a LP problem of $\min_{\mathbf{x}} \mathbf{c}^T\mathbf{x}$ over a polyhedron P, is either an extreme point or infinity. Based on this fact, Dantzig's *simplex* algorithm is introduced as follows:

1 Start with a basis solution.

2 Move to an unobserved adjacent basis feasible solution.

3 If all the basis feasible solutions are observed, then determine the optimal value. Otherwise, go to the second step.

There are many computer solvers able to solve a simple linear optimization problem. For more details, see, e.g., Dantzig and Thapa (1997).

1.1.3 Nonlinear programming and convex analysis

When the objective function and/or constraints in problem (1) are nonlinear, the problem is called a *nonlinear programming* problem. Furthermore, if the feasible set and the objective function in this problem are convex, the nonlinear programming problem would be a *convex problem*. A convex function can be differentiable or non-differentiable. In each case, there are some well-known methods to solve the problem. In this section, two methods for solving such problems are explained. One is the interior point method and the other is the projected sub-gradient method. The first one is used when the objective function and the constraints are differentiable, and the second one can be used for non-differentiable functions.

An example of a convex nonlinear optimization problem with differentiable functions is second-order cone programming (or convex quadratically constrained linear programming) problem. Problems (A.5), (A.16) and (A.17) in paper A (pages 62, 79 and 62) are examples of second-order cone programming problems. Thus, after explaining the interior point methods in general, the method is explained in particular for the second-order cone programming problems.

Interior point methods Assume that the objective function and the constraints in problem (1) are convex and twice continuously differentiable. Then, it is possible to define the necessary and sufficient conditions that an optimal solution $\mathbf{x}^*$ and some multipliers, $\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*$, called *Lagrangian multipliers*, must satisfy. These conditions require a regularity condition called the *Slater condition*. The Slater condition is that there exist one point, say $\hat{\mathbf{x}}$, such that $g_j(\hat{\mathbf{x}}) < 0, j = 1, 2, \ldots, m$, and the constraints $h_j(\mathbf{x}) = 0, j = 1, 2, \ldots, p$, are linear constraints.

According to the above point, we assume that the constraints $h_j(\mathbf{x}) = 0$ for $j = 1, 2, \ldots, p$, in problem (1), are affine and reformulate them as $A\mathbf{x} = \mathbf{b}$,

where A is a $p \times n$ matrix and $\mathbf{b}$ is a $p \times 1$ vector. This assumption does not introduce any restriction to the problem since a nonlinear equality can be reformulated as an inequality by using surplus variables.

Given the Slater condition, the $\mathbf{x}^*$ is an optimal solution to the convex problem (1) with linear equality constraints, if and only if the Karush-Kuhn-Tucker (KKT) conditions, introduced by Kuhn and Tucker (1951), hold as follows:

$$\nabla F(\mathbf{x}^*) + \sum_{j=1}^{m} \lambda_j^* \nabla g_j(\mathbf{x}^*) + A^T \nu_j^* = 0, \tag{2}$$

$$g_j(\mathbf{x}^*) \leq 0, j = 1, 2, \ldots, m, \tag{3}$$

$$A\mathbf{x}^* = \mathbf{b}, \tag{4}$$

$$\lambda_i^* g_j(\mathbf{x}^*) = 0, j = 1, 2, \ldots, m, \tag{5}$$

$$\lambda_i^* \geq 0. \tag{6}$$

In the above, ∇ is the gradient operator.

To solve the above Karush-Kuhn-Tucker conditions, we can use interior point methods as described below.

The interior point methods are kind of hierarchical methods to solve an optimization problem. In this method, first, a linear equality and nonlinear inequality constrained problem is reduced to a linear equality constrained problem with twice differentiable objective function and then, it is efficiently solved by using the well-known Newton's methods (Boyd and Vandenberghe, 2004, p. 561). Therefore, problem (1) should be formulated as an equality constrained problem.

Let us reformulate problem (1) to the equivalent problem

$$\min_{\mathbf{x}} \left\{ F(\mathbf{x}) + \sum_{j=1}^{m} I_-\left(g_j(\mathbf{x})\right) \mid A\mathbf{x} = \mathbf{b} \right\},$$

where $I_-(.)$ is defined as follows:

$$I_-(u) = \begin{cases} 0 & \text{if } u \leq 0, \\ \infty & \text{if } u > 0. \end{cases}$$

Now, problem (1) has been converted to an equality constrained problem, but the objective function is not twice differentiable. So, we approximate this

problem with the following differentiable problem:

$$\begin{aligned} &\min_{\mathbf{x}} \; F(\mathbf{x}) - \sum_{j=1}^{m} \frac{1}{t} \log\left(-g_j(\mathbf{x})\right) \\ &\text{subject to: } A\mathbf{x} = \mathbf{b}. \end{aligned} \tag{7}$$

The objective function here is convex and differentiable, and Newton's method can be used to solve it. The KKT conditions for problem (7) are as follows:

$$\begin{aligned} &\nabla F(\mathbf{x}^*) - \sum_{j=1}^{m} \frac{1}{t g_j(\mathbf{x}^*)} \nabla g_j(\mathbf{x}^*) + A^T \boldsymbol{\nu}^* = 0, \\ &A\mathbf{x}^* = \mathbf{b}, \\ &g_j(\mathbf{x}^*) < 0, j = 1, 2, \ldots, m. \end{aligned}$$

Let $\lambda_j^* = -1/t g_j(\mathbf{x}^*)$, then the above KKT conditions are the same as the KKT conditions in equations (2)-(6) apart from the condition $\lambda_j^* g_j(\mathbf{x}^*) = 0$ which has been replaced with $\lambda_j^* g_j(\mathbf{x}^*) = 1/t$. Thus, for large t, the obtained optimal solution to problem (7), $\mathbf{x}^*$, almost satisfy the KKT conditions of problem (1).

We now show that the optimal value of problem (7) is a lower bound for the optimal value of problem (1). To do so, in the Lagrangian function $L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = F(\mathbf{x}) + \sum_{j=1}^{m} \lambda_j g_j(\mathbf{x}) + \boldsymbol{\nu}^T (A\mathbf{x} - \mathbf{b})$, set $\mathbf{x} = \mathbf{x}^*$, $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$ and $\boldsymbol{\nu} = \boldsymbol{\nu}^*$, then we have

$$\begin{aligned} L(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) &= F(\mathbf{x}^*) + \sum_{j=1}^{m} \lambda_j^* g_j(\mathbf{x}^*) + \boldsymbol{\nu}^{*T} (A\mathbf{x}^* - \mathbf{b}) \\ &= F(\mathbf{x}^*) - \frac{m}{t} \end{aligned}$$

It means the optimal value of problem (7) is no more than m/t less than the optimal value of problem (1). This confirms that $\mathbf{x}^*$ converges to an optimal solution as $t \to \infty$. Based on this fact, the *barrier* algorithm is introduced as follows:

1 Choose the starting points $\mathbf{x}_0$, $t_0 > 0$ and $\mu > 1$.

2 Solve problem (7) with Newton's method, start at $\mathbf{x}_0$.

3 Update $\mathbf{x}_0 = \mathbf{x}^*$.

4 If m/t is small enough, terminate the algorithm. Otherwise, update $t_0 = t_0 \mu$ and go to the second step.

There are several studies on the convergence rate of interior point methods and how to choose good primary values $\mathbf{x}_0$, t_0 and μ, see, e.g., Boyd and

Vandenberghe (2004, Chap. 11). For a review on recent developments of interior point methods, see, e.g., Potra and Wright (2000) and Wright (2004).

Example: As mentioned previously, problems (A.5) and (A.16) and (A.17) in paper A (pages 62, 79 and 62) are examples of second-order cone programming problems. A nonlinear convex optimization problem in the form

$$\begin{aligned} &\min \mathbf{u}^T\mathbf{x} \\ &\text{subject to:} \\ &\|A_j\mathbf{x}+\mathbf{b}_j\|_2 \leq \mathbf{c}_j\mathbf{x}+d_j, j=1,2,\ldots,m. \end{aligned} \tag{8}$$

is a second-order cone programming problem (SOCP). The Euclidean norm, $\|A_j\mathbf{x}+\mathbf{b}_j\|_2$, for $j=1,2,\ldots,m$, are not differentiable at $\{\mathbf{x}\,|\,A_j\mathbf{x}=\mathbf{b}_j\}$. Therefore, it seems that the interior point methods are not appropriate for solving a SOCP problem. However, we can square the two sides of the constraint functions in the SOCP problem and reformulate it to the following problem:

$$\begin{aligned} &\min \mathbf{u}^T\mathbf{x} \\ &\text{subject to:} \\ &\begin{cases} \|A_j\mathbf{x}+\mathbf{b}_j\|_2^2/(\mathbf{c}_j\mathbf{x}+d_j) \leq \mathbf{c}_j\mathbf{x}+d_j, j=1,2,\ldots,m, \\ \mathbf{c}_j\mathbf{x}+d_j \geq 0, j=1,2,\ldots,m. \end{cases} \end{aligned} \tag{9}$$

Now, the constraint functions of problem (9) are convex and twice differentiable. Therefore, it can be solved with interior point methods.

Note that the two problems (8) and (9) are not exactly equivalent. If for an optimal solution to problem (8), $\mathbf{x}^*$, there exist a j such that $\mathbf{c}_j\mathbf{x}^*+d_j=0$, then we can not obtain the optimal solution from problem (9) since $\mathbf{x}^*$ is not in its domain. In spite of that, it can be shown that the interior point method, applied to problem (9), produces an accurate solution to problem (8), see, e.g., Boyd and Vandenberghe (2004, p. 624).

Projected sub-gradient method The sub-gradient method is an iterative algorithm to find an optimal value of a convex function. The advantage of this method is that it can be applied even to a non-differentiable function. The projected sub-gradient method is an extension of the sub-gradient method to find the optimal value of a convex function when it is restricted to a convex set of constraints (Shor, 1998, Chap. 2.1).

In the following, some useful definitions are given and subsequently, the projected sub-gradient algorithm is explained.

Definition 4. *Suppose that $F:\mathbb{R}^n\rightarrow\mathbb{R}$ is a real valued convex function on a convex set $\mathcal{C}$ in the Euclidean space R^n, a vector $\mathbf{h}(\mathbf{x}_0)\in\mathcal{C}$ is called a sub-gradient*

of F at $\mathbf{x}_0 \in \mathcal{C}$ if for all $\mathbf{x} \in \mathcal{C}$,

$$F(\mathbf{x}) \geq F(\mathbf{x}_0) + \mathbf{h}(\mathbf{x}_0)^T (\mathbf{x} - \mathbf{x}_0).$$

The set of all sub-gradients at $\mathbf{x}_0$ is called the *sub-differential set* at $\mathbf{x}_0$ and is denoted by $\partial F(\mathbf{x}_0)$. Clearly, if the function F is differentiable, the sub-differential set at $\mathbf{x}_0$ only contains the gradient of F at $\mathbf{x}_0$.

The sub-gradient algorithm for solving problem (1) with convex functions is as follows:

1 Choose a feasible initial value $\mathbf{x}^{(1)}$ and let $F_{best}^{(1)} = F\left(\mathbf{x}^{(1)}\right)$.

2 For $l = 1, 2, \ldots, L$, do the following steps:

2-1 let $\mathbf{x}^{(l+1)} = \mathbf{x}^{(l)} - \beta_l \mathbf{h}\left(\mathbf{x}^{(l)}\right)$, where $\mathbf{x}^{(l)}$ is the lth iterate value, β_l is the lth step size (e.g., it can be assumed to be a square summable but not summable step size like $1/l$) and $\mathbf{h}\left(\mathbf{x}^{(l)}\right)$ is a sub-gradient of $F(\mathbf{x})$ at $\mathbf{x}^{(l)}$.

2-2 Project the point $\mathbf{x}^{(l+1)}$ onto the feasible set. This projection is the solution of the following problem:

$$\begin{aligned} &\min_{\mathbf{x}} \left\| \mathbf{x}^{(l+1)} - \mathbf{x} \right\|_2^2 \\ &\text{subject to:} \\ &\begin{cases} g_j(\mathbf{x}) \leq 0, j = 1, 2, \ldots, m, \\ h_j(\mathbf{x}) = 0, j = 1, 2, \ldots, p, \end{cases} \end{aligned}$$

where $\|.\|_2$ is the Euclidean norm.

2-3 Let $F_{best}^{(l+1)} = \min\left\{ F\left(\mathbf{x}^{(l+1)}\right), F_{best}^{(l)} \right\}$.

It can be shown that if, for all l, $\left\|\mathbf{h}\left(\mathbf{x}^{(l)}\right)\right\|_2$ are bounded, then the obtained $F_{best}^{(l)}$ converges in probability to the optimal value of problem (1) when $L \to \infty$, see, e.g., Boyd and Park (2007).

In section 1.3.1, the definition of a sub-gradient is extended to a noisy sub-gradient, and a similar algorithm is used to solve a stochastic optimization problem with an expected value objective function.

1.1.4 The branch and bound algorithm

The *Branch and Bound* algorithm is one of the most common algorithms used to solve mixed-integer problems. In this method, we first find the optimal solution for the desired optimization problem while ignoring the integer constraints (This is called *relaxation*). If the solution is integer, the problem has

been solved. Otherwise, we divide the problem into sub-problems by decomposing the set of all possible integer alternatives. Now, we find the integer optimal solution in each sub-problem and compare them to find the best. The point in this algorithm is that we do not need to enumerate all possible integer alternatives since we use some rules to prune some sub-problems. In addition, if solving a sub-problem is not straightforward, we can divide it into sub-problems which are easy to solve.

For example, consider a linear optimization problem with some integer decision variables, such as

$$\begin{aligned} &\min_{\mathbf{x}} \ \mathbf{c}^T\mathbf{x} \\ &\text{subject to:} \\ &\begin{cases} A\mathbf{x} = \mathbf{b}, \\ x_j \in \mathbb{Z}^+, j = 1, 2, \ldots, k, \\ x_j \in \mathbb{R}^+, j = k+1, k+2, \ldots, n, \end{cases} \end{aligned} \tag{10}$$

where A is an $m \times n$ matrix and $\mathbf{x}, \mathbf{c}$ are $n \times 1$ vectors and $\mathbf{b}$ is an $m \times 1$ vector. A simple example of the branch and bound algorithm to solve such a problem is as follows:

1 Find the optimal solution to problem (10) while ignoring the integer constraints (relaxed problem) and call it $\mathbf{x}^*$.

2 If the obtained values $x_1^*, x_2^*, \ldots, x_k^*$ are integer, the algorithm is terminated. Otherwise, let $U = \infty$ and $\Omega = \emptyset$, and do the following steps:

3 Choose one of the variables $x_1, x_2, \ldots, x_k$, e.g., x_2, and divide the problem into two sub-problems by adding the constraints $x_2 \leq [x_2^*]$ and $x_2 \geq [x_2^*] + 1$, where $[u]$ is the largest integer not greater than u.

4 Call the two sub-problems $Prob_i$, $Prob.T_i$ and let $\Omega = \Omega \cup \{Prob_i, Prob.T_i\}$.

5 If $\Omega \neq \emptyset$, do the following steps:

6 Choose one of the sub-problems in Ω and subtract it from the set Ω.

7 Solve the chosen sub-problem and find its optimal solution while ignoring the integer constraints (relaxed problem) and call it $\mathbf{x}^*$. If the problem is not solvable, ignore the next step and go to step 5.

8 If the obtained values $x_1^*, x_2^*, \ldots, x_k^*$ are integer, let $U = min\{U, F(\mathbf{x}^*)\}$ and go to step 5. Otherwise, go back to step 3.

Initializing $U = \infty$ in the above algorithm is valid but not efficient. Accordingly, some methods have been introduced to select a better upper bound, U.

Likewise, more advanced strategies to branch the problem to sub-problems can make the algorithm faster, see, e.g., Achterberg et al. (2005).

There is also another algorithm called *branch and cut* which has been designed to solve mixed-integer linear programming problems. This is a combination of the branch and bound algorithm and a so called *cutting plane* strategy to tighten the linear relaxed problem. For example in step 8 of the above algorithm, if the obtained optimal solution is not integer, we add an extra inequality to the relaxed problem in step 7 and solve it again. This inequality is chosen in a way to cut the optimum from the true feasible set. We hope that resolving the problem attain an integer solution (Mitchell, 2002).

Problem (A.13) (page 70) is an example of a mixed-integer programming problem which we solve in paper A. There are plenty of computer solvers which can solve such a problem with branch and bound or branch and cut algorithms. We use the `Rmosek` package in R and obtain the optimal solutions.

1.2 Stochastic programming (modeling)

In many areas of applications of optimization theory, one encounters a problem that involves parameters which are random variables. There are different approaches to deal with the randomness corresponding to the problem-specific parameters, see, e.g., King and Wallace (2012, Chap. 1). When the parameters are within certain bounds, one may use *robust analysis*. That is formulating and solving an optimization problem, where the optimum solution is feasible for all random values (Ben-Tal et al., 2009, p. 26). When the probability distributions of the random parameters are known or at least can be estimated, one may use *stochastic programming*. That is formulating and solving an optimization problem, where the objective function and/or some of its constraints are expressed in terms of probabilistic statements (Birge and Louveaux, 2011, p. 71).

To explain the above further, consider the following problem:

$$\begin{aligned} &\min_{\mathbf{x}} \; F(\mathbf{x}, \boldsymbol{\xi}) \\ &\text{subject to:} \\ &\begin{cases} g_j(\mathbf{x}, \boldsymbol{\xi}) \leq 0, j = 1, 2, \ldots, m, \\ \mathbf{x} \in \mathcal{X}, \end{cases} \end{aligned} \tag{11}$$

where $F : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}$, $g_j : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}$, $\mathcal{X} \subset \mathbb{R}^n$.

When $\boldsymbol{\xi}$ is an s dimensional specific known vector, the problem is just a *deterministic* optimization problem. When $\boldsymbol{\xi}$ is a random vector, there are various approaches to consider such randomness. For example, the following

problem is a robust optimization problem:

$$\begin{aligned}&\min_{\mathbf{x}} \max_{\boldsymbol{\xi}\in\Xi} F(\mathbf{x},\boldsymbol{\xi})\\&\text{subject to:}\\&\begin{cases} g_j(\mathbf{x},\boldsymbol{\xi}) \leq 0, j = 1,2,\ldots,m, \forall \boldsymbol{\xi} \in \Xi,\\ \mathbf{x} \in \mathcal{X}, \end{cases}\end{aligned}$$

where Ξ is a bounded set.

The optimum solution to the above problem is feasible for all possible values of $\boldsymbol{\xi} \in \Xi$. Now, consider the following problem which is an example of stochastic programming problems:

$$\begin{aligned}&\min_{\mathbf{x}} E(F(\mathbf{x},\boldsymbol{\xi}))\\&\text{subject to:}\\&\begin{cases} E\left(g_j(\mathbf{x},\boldsymbol{\xi})\right) \leq 0, j = 1,2,\ldots,m,\\ \mathbf{x} \in \mathcal{X}, \end{cases}\end{aligned}$$

where the expectations are taken with respect to the probability distribution of $\boldsymbol{\xi}$ which is assumed to be known. Roughly speaking, the optimum solution to the above problem is feasible for almost all possible values of $\boldsymbol{\xi} \in \Xi$.

When the sensitivity of the optimum solution regarding the existent randomness is very important, a robust analysis seems appropriate. However, this analysis is very conservative since it contemplates all possible values, including those that are extremely unlikely to happen. When the decisions are made repeatedly over time, it seems appropriate to find a solution which works well on average and use a stochastic optimization problem. However, these types of problems are mostly computationally complex. Therefore, there is a fundamental trade-off between these two approaches (Giuseppe and Fabrizio, 2006, Preface).

Wald (1945) introduced a minimax problem which is an example of a robust optimization analysis. The origin of the stochastic approach dates back to the work of Dantzig (1955). In this thesis, the main focus is on stochastic optimization problems, specifically the ones which include a chance constraint and initialed with Charnes and Cooper (1959).

In the next two sections, some standard models for formulating a stochastic optimization problem are presented. Section 1.2.2 is devoted to one of these standard forms which includes a chance constraint.

1.2.1 Examples of stochastic optimization problems

Consider an optimization problem in the form of problem (11). Assume that $\boldsymbol{\xi}$ is a random vector and $E(\boldsymbol{\xi})$ and $Var(\boldsymbol{\xi})$ are known and well defined. One

simple method for incorporating the randomness of $\boldsymbol{\xi}$ into the optimization problem is to replace all random variables with their corresponding mean values. For example, problem (11) is reformulated as follows:

$$\begin{aligned} &\min_{\mathbf{x}} \ F\left(\mathbf{x}, E\left(\boldsymbol{\xi}\right)\right) \\ &\text{subject to:} \\ &\begin{cases} g_j\left(\mathbf{x}, E\left(\boldsymbol{\xi}\right)\right) \leq 0, j = 1, 2, \ldots, m, \\ \mathbf{x} \in \mathcal{X}. \end{cases} \end{aligned} \tag{12}$$

This is called the *expected value (EV) problem* or *mean value problem*. Problem (A.4) in paper A (page 62) is an example of an EV problem. In practice, we can not trust in the solution to an EV problem unless there is no or little dependency between the optimal value of this problem and the random variables $\boldsymbol{\xi}$ (Birge and Louveaux, 2011, p. 165). One simple modification is to incorporate the dispersion of the random variables into the model, such as

$$\begin{aligned} &\min_{\mathbf{x}} \ F\left(\mathbf{x}, E\left(\boldsymbol{\xi}\right) + h_0\left(\mathbf{x}, Var\left(\boldsymbol{\xi}\right)\right)\right) \\ &\text{subject to:} \\ &\begin{cases} g_j\left(\mathbf{x}, E\left(\boldsymbol{\xi}\right)\right) + h_j\left(\mathbf{x}, Var\left(\boldsymbol{\xi}\right)\right) \leq 0, j = 1, 2, \ldots, m, \\ \mathbf{x} \in \mathcal{X}, \end{cases} \end{aligned} \tag{13}$$

where $h_j : \mathbb{R}^n \times \mathbb{R}^{s \times s} \rightarrow \mathbb{R}, \forall j$. Problem (A.5) in paper A (page 62) is an example of the modification of the EV problem.

Another simple approach is when the decision variables are chosen based on the expectation of the objective function. The formulation is

$$\begin{aligned} &\min_{\mathbf{x}} \ E\left(F\left(\mathbf{x}, \boldsymbol{\xi}\right)\right) \\ &\text{subject to:} \\ &\begin{cases} g_j\left(\mathbf{x}, E\left(\boldsymbol{\xi}\right)\right) \leq 0, j = 1, 2, \ldots, m, \\ \mathbf{x} \in \mathcal{X}, \end{cases} \end{aligned} \tag{14}$$

where $\boldsymbol{\xi}$ has a known distribution function and the expected value functions are well defined.

If the optimization problem is convex, using Jensen's inequality, it is easy to show that $F\left(\mathbf{x}, E\left(\boldsymbol{\xi}\right)\right) \leq E\left(F\left(\mathbf{x}, \boldsymbol{\xi}\right)\right)$. Therefore, the optimal value of problem (14) is always larger than the optimal value of the EV problem.

Problem (14) minimizes the expectation of the objective function and requires the satisfaction of the constraints on average, when $\boldsymbol{\xi}$ is a random vector. This kind of formulating a stochastic problem is also not suitable when some of the constraint functions, $g_j\left(\mathbf{x}, \boldsymbol{\xi}\right), j = 1, 2, \ldots, m$, have high variability (Dentcheva, 2006, p. 50). Finally, another standard way of incorporating the randomness into an optimization problem is formulating a problem with

chance constraints. This is discussed in detail in the next section.

1.2.2 Chance constraint problems

As mentioned above, problem (14) is also not suitable when some of the constraint functions have high variability. In fact, this kind of formulating a stochastic problem is not satisfactory in incorporating the dispersion of the random vectors into the optimization problem. A solution, which was introduced by Charnes and Cooper (1959), is to define constraints with probability functions instead. One can define a stochastic problem, such as

$$\begin{aligned} &\min_{\mathbf{x}} \; E\left(F\left(\mathbf{x}, \boldsymbol{\xi}\right)\right) \\ &\text{subject to:} \\ &\begin{cases} P\left(g_j\left(\mathbf{x}, \boldsymbol{\xi}\right) \leq 0, j = 1, 2, \ldots, m\right) \geq 1 - \alpha_0, \\ \mathbf{x} \in \mathcal{X}, \end{cases} \end{aligned} \tag{15}$$

where $0 < \alpha_0 < 1$ is a fixed value. This means that for a given decision variable, $\mathbf{x}$, we do not reject the statistical hypothesis that the constraints $g_j\left(\mathbf{x}, \boldsymbol{\xi}\right) \leq 0, j = 1, 2, \ldots, m$, are satisfied (Shapiro et al., 2009, p. 87). This is certainly a weaker condition than satisfying the constraints $g_j\left(\mathbf{x}, \boldsymbol{\xi}\right) \leq 0$ for all possible realizations of $\boldsymbol{\xi}$, as we seek in the robust optimization analysis.

Constraints defined with probability functions are called *chance constraints* (or *probabilistic constraints*). The constraint

$$P\left(g_j\left(\mathbf{x}, \boldsymbol{\xi}\right) \leq 0, j = 1, 2, \ldots, m\right) \geq 1 - \alpha_0,$$

is called a *joint chance constraint*. One can also define *individual chance constraints* as below:

$$P\left(g_j\left(\mathbf{x}, \boldsymbol{\xi}\right) \leq 0\right) \geq 1 - \alpha_{0j}, j = 1, 2, \ldots, m,$$

where $0 < \alpha_{0j} < 1, j = 1, 2, \ldots, m$, are m fixed values. The latter is studied in the context of *stochastic ordering/dominance* constraint problems.

A chance constraint which can be written as $P\left(g_i\left(\mathbf{x}\right) \geq \xi_i, i = 1, 2, \ldots, s\right) \geq 1 - \alpha_0$, where $g_i : \mathbb{R}^n \to \mathbb{R}$, is called a *separable* chance constraint. Otherwise, it is called a *non-separable* constraint. A separable chance constraint problem may lead to a simpler problem, since this constraint is equivalent to $\mathcal{F}\left(g_1\left(\mathbf{x}\right), g_2\left(\mathbf{x}\right), \ldots, g_s\left(\mathbf{x}\right)\right) \geq 1 - \alpha_0$, where $\mathcal{F} : \mathbb{R}^s \to \mathbb{R}$ is the joint cumulative distribution function of the vector $\left(\xi_1, \xi_2, \ldots, \xi_s\right)$. Problem (A.6) in paper A (page 63) is an example of non-separable chance constraint problems.

In recent years, chance constraint problems have been applied in different fields for example wind power, see, e.g., Elshahed et al. (2013) and Wang et al. (2012), production planning, see, e.g., Lejeune and Ruszczynski (2007)

and chemical processing, see, e.g., Henrion and Moller (2003).

In general, solving a chance constraint problem numerically is a difficult task since

1 It is usually difficult to compute the exact value of $P\left(g\left(\mathbf{x},\boldsymbol{\xi}\right)\leq 0\right)$ at $\mathbf{x}\in\mathcal{X}$, even for a simple function $g\left(\mathbf{x},\boldsymbol{\xi}\right)$, e.g., linear function. Thus, it is difficult to check the feasibility of a solution and in fact, sometimes the only way to check this is by using Monte Carlo sampling.

2 The feasible set defined by a chance constraint can be non-convex even if $\mathcal{X}$ is a convex set and $g\left(\mathbf{x},\boldsymbol{\xi}\right)$ is a convex function of $\mathbf{x}$ for every possible realization of $\boldsymbol{\xi}$.

Accordingly, in the following section, we study the conditions that may lead to the convexity of the feasible set in the chance constraint problems.

Convexity of the feasible set in a chance constraint optimization problem When a mathematical optimization problem is formulated with a chance constraint, it is very important to clarify whether the problem is convex. If the convexity is satisfied, a numerical solution for the problem is possible.

Regarding this issue, in the past 50 years, some new mathematical concepts and the proof of some basic theorems have been released. For example, Prekopa (1971) introduced and studied the concept of logarithmic concave measures. Borell (1974); Rinott (1976) and Brascamp and Lieb (1976) generalized this definition to the definition of α-concavity of the measures. We review some of these concepts and theorems in the following.

In section 1.1.1, a convex (non-convex) set and an α-convex (α-concave) function were defined. The definition of α-concavity can be extended for a probability measure function in the following way:

Definition 5. *Fix $-\infty\leq\alpha\leq\infty$, a probability measure function, P, defined on the Borel subsets of a convex set $\Omega\subset\mathbb{R}^s$ is said to be α-concave if for all sets $A,B\subset\Omega$ and $0\leq\lambda\leq 1$, the following inequality holds true:*

$$P\left(\lambda A+\left(1-\lambda\right)B\right)\geq m_{\alpha}\left(P\left(A\right),P\left(B\right),\lambda\right),$$

where $\lambda A+\left(1-\lambda\right)B=\left\{\lambda\mathbf{a}+\left(1-\lambda\right)\mathbf{b}\,|\mathbf{a}\in A,\mathbf{b}\in B\right\}$ and

$$m_{\alpha}\left(P\left(A\right),P\left(B\right),\lambda\right)=\begin{cases}P\left(A\right)^{\lambda}P\left(B\right)^{\left(1-\lambda\right)} & \text{if } \alpha=0,\\ \max\left\{P\left(A\right),P\left(B\right)\right\} & \text{if } \alpha=\infty,\\ \min\left\{P\left(A\right),P\left(B\right)\right\} & \text{if } \alpha=-\infty,\\ \left(\lambda P\left(A\right)^{\alpha}+\left(1-\lambda\right)P\left(B\right)^{\alpha}\right)^{1/\alpha} & \text{otherwise.}\end{cases}$$

With this definition of the α-concavity for the probability measures, it is only possible to define an α-concave distribution function on a continuous

set. This is because, the set $\Omega \subset \mathbb{R}^s$ is assumed to be convex. This definition is extended to the discrete distribution functions as follows:

Definition 6. *Fix $-\infty \leq \alpha \leq \infty$, a distribution function, $F(\boldsymbol{\xi})$, defined on the set $\mathcal{A} \subset \mathbb{R}^s$ is said to be α-concave if for all $\mathbf{z}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \in \mathcal{A}$ and $0 \leq \lambda \leq 1$, the following inequality holds true:*

$$F(\mathbf{z}) \geq m_\alpha \left(F(\boldsymbol{\xi}_1), F(\boldsymbol{\xi}_2), \lambda\right),$$

where $\mathbf{z} \geq \lambda \boldsymbol{\xi}_1 + (1-\lambda)\boldsymbol{\xi}_2$[2]. This is an α-concave continuous distribution function if $\mathcal{A} = \mathbb{R}^s$.

Although, many multivariate probability distribution functions are not concave, there are a wide range of distributions which are quasi-concave or even log-concave (Prekopa, 1971). The normal, the Wishart, the beta, the Dirichlet, and the gamma distributions are some examples of continuous log-concave multivariate distribution functions, see, e.g., Prekopa (1971) and Shapiro et al. (2009, p. 102). Likewise, the binomial, the Poisson, the geometric, and the negative binomial distributions are some examples of discrete log-concave distributions, see, e.g., An (1995).

The main results in the convexity theory of optimization problems with chance constraints are the following two theorems. The proofs can be found in Shapiro et al. (2009, pp. 107-108).

Theorem 2. *Let the functions $g_j : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}, j = 1, 2, \ldots, m$, be quasi-convex jointly in both arguments[3] and $\boldsymbol{\xi} \in \mathbb{R}^s$ be a random vector with an α-concave probability distribution, then the function*

$$G(\mathbf{x}) = P\left(g_j(\mathbf{x}, \boldsymbol{\xi}) \leq 0, j = 1, 2, \ldots, m\right)$$

is α-concave on the set $R = \left\{\mathbf{x} \in \mathbb{R}^n \,\middle|\, \exists \boldsymbol{\xi} \in \mathbb{R}^s \text{such that } g_j(\mathbf{x}, \boldsymbol{\xi}) \leq 0, \forall j\right\}$.

As a consequence of the above theorem, we conclude the convexity of the feasible set in a chance constraint optimization problem in the following theorem.

Theorem 3. *Let the functions $g_j : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}, j = 1, 2, \ldots, m$, be quasi-convex jointly in both arguments and $\boldsymbol{\xi} \in \mathbb{R}^s$ be a random vector with an α-concave probability distribution, then the following set is convex and closed:*

$$\mathcal{C} = \left\{\mathbf{x} \in \mathbb{R}^n \,\middle|\, P\left(g_j(\mathbf{x}, \boldsymbol{\xi}) \leq 0, j = 1, 2, \ldots, m\right) \geq 1 - \alpha_0\right\}.$$

[2]The inequality relationship between two vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^s$ is defined as follows: $\mathbf{x} \geq \mathbf{y}$ if $x_i \geq y_i, i = 1, 2, \ldots, s$.

[3]That means for all $(\mathbf{x}_1, \boldsymbol{\xi}_1), (\mathbf{x}_2, \boldsymbol{\xi}_2) \in \Omega \subset \mathbb{R}^n \times \mathbb{R}^s$, we have

$$g_j(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2, \lambda \boldsymbol{\xi}_1 + (1-\lambda)\boldsymbol{\xi}_2) \leq max\left\{g_j(\mathbf{x}_1, \boldsymbol{\xi}_1), g_j(\mathbf{x}_2, \boldsymbol{\xi}_2)\right\}$$

Applying this theorem in a sophisticated chance constraint problem may lead to a convex optimization problem which can be solved numerically. Shapiro et al. (2009, Chap. 4) discussed the continuity and differentiability of an α-concave function and based on that presented some numerical solutions for chance constraint problems. However, it is generally difficult to obtain these numerical solutions and other approaches, such as sample average approximation, are preferred.

Now, look again at problem (A.6) (page 63) which we solve in paper A. Considering the definitions and theorems of section 1.1.1 and the current section, we want to investigate the convexity of the problem. In this problem, the decision variables are denoted by $\mathbf{r} = (r_1, r_2, \ldots, r_k)$ and the parameters, which are random, are denoted by $\mathbf{B} = (B_1, B_2, \ldots, B_k)$ and $\mathbf{D} = (D_{11}, D_{12}, \ldots, D_{ks})$. In addition, $(n_{min})_j, j = 1, 2, \ldots, s$, and K are assumed to be some non-negative constant values.

To investigate the convexity of the objective function in problem (A.6), based on theorem 1, we know that the affine function $(n_{min})_j - \sum_{i=1}^{k} r_i D_{ij}$ for a specific $\mathbf{D}$ is convex. In the same theorem, it is stated that the maximum of two convex functions is convex. Thus, $max\left\{(n_{min})_j - \sum_{i=1}^{k} r_i D_{ij}, 0\right\}$ is convex for each $\mathbf{D}$. Clearly, the sum of some non-negative convex functions is convex and consequently $\sum_{j=1}^{s} max\left\{(n_{min})_j - \sum_{i=1}^{k} r_i D_{ij}, 0\right\}$ is convex for each $\mathbf{D}$. Finally, since the expected value of a convex function is convex (Boyd and Mutapcic, 2006), the objective function of problem (A.6) is convex.

To investigate the convexity of the feasible set, we look into theorem 3. Sufficient conditions for the convexity of the feasible set in problem (A.6) are that the random vector $\mathbf{B}$ has a log-concave probability distribution and $g(\mathbf{r}, \mathbf{B}) = \sum_{i=1}^{k} r_i B_i - K$ is a quasi-convex function jointly in both arguments $\mathbf{r}$ and $\mathbf{B}$. Unfortunately, $g(\mathbf{r}, \mathbf{B})$ is not necessarily quasi-convex in both arguments (Shapiro et al., 2009, p. 109), and we can not prove the convexity of the feasible set.

To sum up, we could not prove the convexity of problem (A.6) to find a numerical solution to this problem. Instead, we use a sample average approximation algorithm to obtain the solution. In the next section, the algorithms to solve stochastic optimization problems are discussed.

1.3 Stochastic programming (algorithms)

In section 1.2, different approaches for incorporating the randomness of the parameters into an optimization problem were discussed. In this part, we explain some methods for solving such problems. In the following, we assume that the set $\mathcal{X}$ is convex. In addition, we assume that the functions $F(\mathbf{x}, \boldsymbol{\xi})$ and $g_j(\mathbf{x}, \boldsymbol{\xi}), j = 1, 2, \ldots, m$, are convex functions of $\mathbf{x}$ for each $\boldsymbol{\xi}$.

Problem (12) is simply equivalent to the deterministic problem (11) when

the random variables are replaced with their corresponding mean values. In fact, in the application, they are mostly replaced with the estimation of their corresponding mean values. In any case, we need to solve a convex problem, and we can use the well-known methods of solving convex optimization problems. As an example, problem (A.4) in paper A (page 62), which is in the form of problem (12), is simply reformulated as a linear optimization problem and can be solved with the simplex method.

In addition, if the functions $h_j, j = 1, 2, \ldots, m$, in problem (13) are convex functions of $\mathbf{x}$ for each $\boldsymbol{\xi}$, then this problem would also be convex. As an example, problem (A.5) in paper A (page 62), which is in the form of problem (13), is a convex problem and can be solved with the interior point methods which were explained in section 1.1.3.

Likewise, it can be shown that with the assumption of convexity for the functions F and $g_j, j = 1, 2, \ldots, m$, problem (14) is a convex problem (Boyd and Mutapcic, 2006). Consider a case where the expected value function $E\left(F\left(\mathbf{x}, \boldsymbol{\xi}\right)\right)$ can not be written in a closed form even though the function $F\left(\mathbf{x}, \boldsymbol{\xi}\right)$ is easily computable in $\mathbf{x}$ for each $\boldsymbol{\xi}$. For example, consider the following problem:

$$\begin{aligned} &\min_{\mathbf{x}} \; E\left(\sum_{j=1}^{s} \max\left\{n_j - \boldsymbol{\xi}_j^T \mathbf{x}, 0\right\}\right) \\ &\text{subject to:} \\ &\begin{cases} \mathbf{a}^T \mathbf{x} \leq \mathbf{b}, \\ \mathbf{x} \in \mathcal{X}, \end{cases} \end{aligned} \tag{16}$$

where $n_j, j = 1, 2, \ldots, s$, $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ and $\mathbf{b} = (b_1, b_2, \ldots, b_n)$ are fixed known values and $\boldsymbol{\xi}_j = \left(\xi_{1j}, \xi_{2j}, \ldots, \xi_{nj}\right), j = 1, 2, \ldots, s$, are random vectors. The objective function of this problem is called the *expected total violation* function (Boyd and Mutapcic, 2006). Note that the objective function of problem (A.6) in paper A (page 63) is in the form of the objective function of problem (16).

There are various methods to calculate the optimal value of such a problem. One way is to use a stochastic projected sub-gradient method, see, e.g., Boyd and Mutapcic (2006). This method is explained in section 1.3.1. We can also approximate the objective function with Monte Carlo simulations. We refer to this as the sample average approximation method, see, e.g., Kleywegt et al. (2001). We explain this method in section 1.3.2.

Another stochastic problem which was discussed in section 1.2 is a chance constraint problem. As mentioned in page 16, except in some very special cases, solving a chance constraint problem numerically is a difficult task. There are two reasons for that. First, for a given $\mathbf{x} \in \mathcal{X}$, the quantity $P\left(g\left(\mathbf{x}, \boldsymbol{\xi}\right) \leq 0\right)$ may be hard to be computed, since it requires a multi-

dimensional integration. Second, the feasible set defined by a chance constraint may be a non-convex set. For example, we explained that problem (A.6) is not necessarily a convex problem. Accordingly, for solving a chance constraint problem, some sampling approximation methods are developed.

Sampling approximation methods for chance constraint problems went into two different directions. One is the scenario approximation method, and the other is the sample average approximation method. In the scenario approximation method, a finite number of observations are drawn from either the exact distribution of the random vector or the estimated distribution. Then, the problem is discretized and solved as a deterministic problem, see, e.g., Dentcheva et al. (2000). In the sample average approximation method, the original distribution of the random vector is replaced with an empirical distribution. This method is also based on the Monte Carlo sampling. This means that a finite number of observations are drawn from the exact/estimated distribution of the random vector and then, the problem is reformulated to be solved. In fact, the two methods are very similar except that in the scenario approximation method, the constraint should be satisfied for all drawn samples while in the sample average approximation method, it is not necessary, see, e.g., Luedtke and Ahmed (2008) and Pagnoncelli et al. (2009).

The scenario approximation method is very conservative, and it is similar to the robust optimization problems when the number of samples is increased. The sample average approximation method is less conservative, and it allows a few samples to violate the constraints. Instead, the sample average approximation method is usually computationally more complex than the scenario approximation method. In this thesis, we use the sample average approximation method.

The sample average approximation method for a chance constraint problem is a variation of the well-known sample average approximation method for stochastic problems with expected value objective functions. As mentioned earlier, the sample average approximation method, for solving stochastic problems with expected value objective functions, is explained in section 1.3.2. In this section, we also explain the sample average approximation method for solving chance constraint problems.

In paper A, we suggest another method to solve a chance constraint problem, namely the majorization method. Section (1.3.3) is devoted to explain this method in more details.

1.3.1 Stochastic projected sub-gradient method

One way to calculate the optimal value of problem (16) is to use a stochastic projected sub-gradient method. This method is almost the same as the sub-gradient method, but it uses a noisy sub-gradient instead of an ordinary sub-

gradient in each iteration. Hence also, the stochastic projected sub-gradient method is an extended version of the projected sub-gradient method. This method is used when the true sub-gradient can not be computed easily e.g., for an expected function (Shor, 1998, Chap. 2.4).

In the present section, the concept of a noisy sub-gradient is explained and it is shown how one can obtain a noisy sub-gradient for an expected function. Furthermore, the convergence analysis of the stochastic sub-gradient method is explained, and based on that an algorithm for obtaining the optimal value of problem (16) is presented as an example.

Definition 7. *Suppose $F : \mathbb{R}^n \to \mathbb{R}$ is a real valued convex function on a convex set $\mathcal{C}$ in the Euclidean space $\mathbb{R}^n$. A random vector $\tilde{\mathbf{h}}(\mathbf{x}_0) \in \mathcal{C}$ is called a noisy (unbiased) sub-gradient of F at $\mathbf{x}_0 \in \mathcal{C}$ if $\mathbf{h}(\mathbf{x}_0) = E\left(\tilde{\mathbf{h}}(\mathbf{x}_0)\right) \in \partial F(\mathbf{x}_0)$.*

Below, it is explained that how one can obtain a noisy sub-gradient for an expected function like $E(F(\mathbf{x}, \boldsymbol{\xi}))$.

Noisy sub-gradient of an expected function value In the following, we compute a noisy sub-gradient for the function $f(\mathbf{x}) = E(F(\mathbf{x}, \boldsymbol{\xi}))$ at $\mathbf{x}$, where $F : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}$ and for each $\boldsymbol{\xi}$, $F(\mathbf{x}, \boldsymbol{\xi})$ is a convex function of $\mathbf{x}$.

Let $\tilde{\mathbf{h}} : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}^n$ be a sub-gradient of $F(\mathbf{x}, \boldsymbol{\xi})$ at $\mathbf{x}_0$ for each $\boldsymbol{\xi}$, i.e., $\tilde{\mathbf{h}}(\mathbf{x}_0, \boldsymbol{\xi}) \in \partial_{\mathbf{x}} F(\mathbf{x}_0, \boldsymbol{\xi})$. We show that $\mathbf{h}(\mathbf{x}_0) = E\left(\tilde{\mathbf{h}}(\mathbf{x}_0, \boldsymbol{\xi})\right)$ is a sub-gradient for the function $f(\mathbf{x}) = E(F(\mathbf{x}, \boldsymbol{\xi}))$ at $\mathbf{x}_0$.

By definition, for any $\mathbf{x}$ and each $\boldsymbol{\xi}$, the following inequality holds true:

$$F(\mathbf{x}, \boldsymbol{\xi}) \geq F(\mathbf{x}_0, \boldsymbol{\xi}) + \tilde{\mathbf{h}}(\mathbf{x}_0, \boldsymbol{\xi})^T (\mathbf{x} - \mathbf{x}_0).$$

Multiplying this by the density of the random vector $\boldsymbol{\xi}$, which is non-negative, and integrating gives

$$\begin{aligned} E(F(\mathbf{x}, \boldsymbol{\xi})) &\geq E(F(\mathbf{x}_0, \boldsymbol{\xi})) + E\left(\tilde{\mathbf{h}}(\mathbf{x}_0, \boldsymbol{\xi})\right)^T (\mathbf{x} - \mathbf{x}_0) \\ &= f(\mathbf{x}_0) + \mathbf{h}(\mathbf{x}_0)^T (\mathbf{x} - \mathbf{x}_0). \end{aligned}$$

Therefore, $\tilde{\mathbf{h}}(\mathbf{x}_0, \boldsymbol{\xi})$ is a noisy sub-gradient of the expected value function at $\mathbf{x}_0$ for each $\boldsymbol{\xi}$.

One can generate a sample from the distribution of the random vector $\boldsymbol{\xi}$ and calculate the noisy sub-gradient of $E(F(\mathbf{x}, \boldsymbol{\xi}))$ at $\mathbf{x}_0$. Another solution is to generate M samples from the distribution of the random vector $\boldsymbol{\xi}$ and calculate the mean of the noisy sub-gradients at $\mathbf{x}_0$. This helps us to use a sub-gradient method without computing the sub-gradients of a complicated expected value function.

Convergence analysis Consider problem (14) with a convex objective function and a convex feasible set denoted by $\mathcal{C}$, i.e, $\mathcal{C} = \left\{\mathbf{x} \in \mathcal{X} \,\middle|\, g_j(\mathbf{x}, E(\boldsymbol{\xi})) \leq 0, \forall j\right\}$.

Assume that the sub-gradient of the objective function can not be easily computed but it is easy to compute the sub-gradient of $F(\mathbf{x}, \boldsymbol{\xi})$ for each $\boldsymbol{\xi}$. We can apply the projected sub-gradient algorithm when the sub-gradients have been replaced with the noisy sub-gradients. This is called a stochastic projected sub-gradient method. In the following, we discuss the convergence of a stochastic projected sub-gradient method.

Assume $\mathbf{x}^*$ is the optimum value of problem (14). Given $\mathbf{x}^{(1)}$, we generate the sequence $\mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \ldots, \mathbf{x}^{(L)}$ in a stochastic projected sub-gradient method by the formula

$$\mathbf{x}^{(l+1)} = \Pi_{\mathcal{C}}\left(\mathbf{x}^{(l)} - \beta_l \tilde{\mathbf{h}}\left(\mathbf{x}^{(l)}, \boldsymbol{\xi}\right)\right), l = 1, 2, \ldots, L-1,$$

where $\Pi_{\mathcal{C}}$ is the orthogonal projection of a point onto the feasible set $\mathcal{C}$ and $\tilde{\mathbf{h}}\left(\mathbf{x}^{(l)}, \boldsymbol{\xi}\right)$ is a noisy sub-gradient of F at $\mathbf{x}^{(l)}$.

Assuming $E\left\|\tilde{\mathbf{h}}(\mathbf{x}, \boldsymbol{\xi})\right\|_2 \leq G$, $\forall \mathbf{x} \in \mathcal{C}$ and $\max_{\mathbf{x}} \left\|\mathbf{x}^{(1)} - \mathbf{x}\right\|_2 \leq B$, where G and B are some known constants, we have

$$\min_{l=1,\ldots,L} E\left(f\left(\mathbf{x}^{(l)}\right)\right) - f(\mathbf{x}^*) \leq \frac{B^2 + G^2 \sum_{l=1}^{L} \beta_l^2}{2\sum_{l=1}^{L} \beta_l}. \tag{17}$$

The proof can be found in Boyd and Mutapcic (2006).

Based on the above inequality, for $\beta_l = 1/l$, the right hand side converges to zero, i.e., $\min_{l=1,\ldots,L} E\left(f\left(\mathbf{x}^{(l)}\right)\right)$ converges to $f(\mathbf{x}^*)$. Subsequently, by using Jensen's inequality and the concavity of the minimum function, we can show that $E\left(\min_{l=1,\ldots,l} f\left(\mathbf{x}^{(l)}\right)\right)$ converges to $f(\mathbf{x}^*)$. Finally, by Chebyshev's inequality, we can show the convergence in probability.

If we fix the number of iterations and use the constant step size policy, i.e., $\beta_l = \beta$, we can find the best constant step size which minimizes the convex function $f(\beta) = \left(B^2 + G^2 L \beta^2\right)/(2L\beta)$ with respect to β. By differentiating $f(\beta)$ and setting it equal to zero, we get

$$\beta = \frac{B}{G\sqrt{L}}.$$

For such a constant step size, the inequality (17) is written as

$$E\left(\min_{l=1,\ldots,l}\left(f\left(\mathbf{x}^{(l)}\right)\right)\right) - f(\mathbf{x}^*) \leq \frac{BG}{2\sqrt{L}}.$$

By Chebyshev inequality, for $\epsilon \geq 0$, we have

$$P\left(\min_{l=1,\ldots,L}\left(f\left(\mathbf{x}^{(l)}\right)\right) - f\left(\mathbf{x}^*\right) \geq \epsilon\right) \leq \epsilon^{-1} E\left(\min_{l=1,\ldots,L}\left(f\left(\mathbf{x}^{(l)}\right)\right) - f\left(\mathbf{x}^*\right)\right) \leq \frac{BG}{2\epsilon\sqrt{L}}.$$

If $P\left(\min_{l=1,\ldots,l}\left(f\left(\mathbf{x}^{(l)}\right)\right) - f\left(\mathbf{x}^*\right) \geq \epsilon\right) = 1-\alpha$, then for $L \geq B^2G^2/\left(4\epsilon^2\left(1-\alpha\right)^2\right)$, the approximate solution, $\min_{l=1,\ldots,L}\left(f\left(\mathbf{x}^{(l)}\right)\right)$ will converge in probability to the solution $f\left(\mathbf{x}^*\right)$.

Now, as an example, we give the stochastic projected sub-gradient algorithm to find an optimum solution to problem (16).

Solving problem (16) with stochastic projected sub-gradient method The objective function and the feasible set in problem (16) are convex. In addition, the objective function is piece-wise linear, and it is not differentiable everywhere. Thus, the sub-gradient method seems appropriate for this example. In the following, we outline an algorithm to solve problem (16) via the stochastic projected sub-gradient method.

1 Choose the initial value $\mathbf{x}^{(1)}$. It is usually the value obtained by the equivalent deterministic problem.

2 Let $f_{best}^{(1)} = E\left(F\left(\mathbf{x}^{(1)}, \boldsymbol{\xi}\right)\right)$, this can be calculated for example by Monte Carlo sampling.

3 Set $L = B^2G^2/\left(4\epsilon^2\left(1-\alpha\right)^2\right)$, where G is a value such that $E\left\|\tilde{\mathbf{h}}\left(\mathbf{x}, \boldsymbol{\xi}\right)\right\|_2 \leq G$, $\forall \mathbf{x} \in \mathcal{C}$ and B is a value that satisfy $\max_{\mathbf{x}}\left\|\mathbf{x}^{(1)} - \mathbf{x}\right\|_2 \leq B$, and ϵ, α are arbitrary small values.

4 For $l = 1, 2, \ldots, L$, do the following steps:

4-1 Generate one or more samples from the distribution of the random vector $\boldsymbol{\xi} = \left(\xi_{11}, \xi_{12}, \ldots, \xi_{ns}\right)$.

4-2 Suppose $\tilde{\mathbf{h}}\left(\mathbf{x}^{(l)}, \boldsymbol{\xi}\right)$ is a noisy sub-gradient of the function

$$E\left(F\left(\mathbf{x}, \boldsymbol{\xi}\right)\right) = E\left(\sum_{j=1}^{s} \max\left\{n_j - \sum_{i=1}^{n} r_i \xi_{ij}, 0\right\}\right),$$

at $\mathbf{x}^{(l)}$. This means that

$$\tilde{\mathbf{h}}\left(\mathbf{x}^{(l)}, \boldsymbol{\xi}\right) = \sum_{j=1}^{s}\sum_{t=1}^{M} \tilde{\mathbf{h}}_j\left(\mathbf{x}^{(l)}, \boldsymbol{\xi}^{(t)}\right)/M,$$

where $\tilde{\mathbf{h}}_j\left(\mathbf{x}^{(l)}, \boldsymbol{\xi}^{(t)}\right)$ for $t = 1, 2, \ldots, M$ and $j = 1, 2, \ldots, s$ equals

$$\tilde{\mathbf{h}}_j\left(\mathbf{x}^{(l)}, \boldsymbol{\xi}^{(t)}\right) = \begin{cases} \mathbf{0} & \text{if } n_j \leq \sum_{i=1}^{n} r_i^{(l)} \xi_{ij}^{(t)}, \\ -\left(\xi_{1j}^{(t)}, \xi_{2j}^{(t)}, \ldots, \xi_{nj}^{(t)}\right) & \text{otherwise,} \end{cases}$$

and $\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(M)}$ are M generated values obtained from the previous step.

Note that, if $\tilde{\mathbf{h}}\left(\mathbf{x}^{(l)}, \boldsymbol{\xi}\right)$ equals zero, we should not stop the algorithm since it may happen because of the randomness of the vector $\boldsymbol{\xi}$.

4-3 Let

$$\mathbf{x}^{(l+1)} = \mathbf{x}^{(l)} - \frac{B}{G\sqrt{L}}\tilde{\mathbf{h}}\left(\mathbf{x}^{(l)}, \boldsymbol{\xi}\right),$$

where L, G and B are defined as before.

4-4 Project the point $\mathbf{x}^{(l+1)}$ onto the set $\mathcal{C} = \left\{\mathbf{x} \in \mathcal{X} \,\middle|\, \mathbf{a}^T\mathbf{x} \leq \mathbf{b}\right\}$. This projection is $\arg\min_{\mathbf{x}\in\mathcal{C}} \left\|\mathbf{x}^{(l+1)} - \mathbf{x}\right\|_2^2$, where $\|.\|_2$ is the Euclidean norm. Since $\mathcal{C}$ is convex and closed, this minimizer exists and it is unique.

4-5 Calculate the value $f\left(\mathbf{x}^{(l+1)}\right) = E\left(F\left(\mathbf{x}^{(l+1)}, \boldsymbol{\xi}\right)\right)$, using the generated samples in step 4-1.

4-6 Let $f_{best}^{(l+1)} = \min\left\{f\left(\mathbf{x}^{(l+1)}\right), f_{best}^{(l)}\right\}$.

As stated previously, the objective function of problem (16) is similar to the objective function of problem (A.6) in paper A (page 63). However, the existence of a chance constraint in problem (A.6) makes it impossible to use the stochastic sub-gradient method for solving this problem.

1.3.2 The sample average approximation method

In the previous section, we explained the stochastic projected sub-gradient method to solve an optimization problem with an expected value objective function. Another method for solving such a problem is the sample average approximation (SAA) method. The SAA method can also be used in chance constraint problems. Therefore, we divide this section into two parts. In the first part, the SAA method for solving an optimization problem with the objective function in the form of an expectation is explained, and in the second part, the SAA method for solving an optimization problem with chance constraints is explained.

An optimization problem with an expected value objective function Consider problem (14) with a convex objective function and a closed convex feasible set $\mathcal{C} = \left\{\mathbf{x} \in \mathcal{X} \,\middle|\, g_j\left(\mathbf{x}, E\left(\boldsymbol{\xi}\right)\right) \leq 0, \forall j\right\}$. Assume that the expected value function $f\left(\mathbf{x}\right) = E\left(F\left(\mathbf{x}, \boldsymbol{\xi}\right)\right)$ can not be written in a closed form and it is also difficult to be computed while the function $F\left(\mathbf{x}, \boldsymbol{\xi}\right)$ is easily computable for given $\mathbf{x}$ and $\boldsymbol{\xi}$. Let us call this problem the *true* problem.

Suppose that we can generate independent samples $\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(N)}$ from the distribution of the random vector $\boldsymbol{\xi}$. Using these samples, the sample average approximation problem associated with the true problem is written as $\min\limits_{\mathbf{x}\in\mathcal{C}} \hat{f}_N\left(\mathbf{x}\right)$, where

$$\hat{f}_N\left(\mathbf{x}\right) = \frac{1}{N}\sum_{t=1}^{N} F\left(\mathbf{x}, \boldsymbol{\xi}^{(t)}\right).$$

Convergence analysis It is clear that for each $\mathbf{x}$, the function $\hat{f}_N\left(\mathbf{x}\right)$ is an unbiased estimator of $f\left(\mathbf{x}\right)$ and based on the law of large numbers (LLN), for every fixed $\mathbf{x}$, $\hat{f}_N\left(\mathbf{x}\right)$ converges to $f\left(\mathbf{x}\right)$, w.p.1 as $N \to \infty$. However, the optimal value of the sample average approximation problem does not necessarily converge to the optimal value of the true problem. To ensure such a convergence, we need a uniform convergence of $\hat{f}_N\left(\mathbf{x}\right)$ to $f\left(\mathbf{x}\right)$ as it is defined below.

Definition 8. *$\hat{f}_N$ uniformly converges to f if, for any $\mathbf{x}$, the following two conditions hold true:*

1 *For any sequence $\mathbf{x}_N$ converging to $\mathbf{x}$ one has $\liminf\limits_{N\to\infty} \hat{f}_N\left(\mathbf{x}_N\right) \geq f\left(\mathbf{x}\right)$.*

2 *There exists a sequence $\mathbf{x}_N$ converging to $\mathbf{x}$ such that $\limsup\limits_{N\to\infty} \hat{f}_N\left(\mathbf{x}_N\right) \leq f\left(\mathbf{x}\right)$.*

The following theorem implies the conditions that satisfy the uniform convergence of $\hat{f}_N\left(\mathbf{x}\right)$ to $f\left(\mathbf{x}\right)$. The proof can be found in Ruszczynski and Shapiro (2003, pp. 363-364).

Theorem 4. *Suppose that (i) the LLN conditions hold (ii) $\mathcal{C}$ is compact and nonempty, (iii) the function $F\left(\mathbf{x}, \boldsymbol{\xi}\right)$ is continuous on $\mathcal{C}$ for every $\boldsymbol{\xi}$, and (iv) $F\left(\mathbf{x}, \boldsymbol{\xi}\right)$, $\mathbf{x} \in \mathcal{C}$, is dominated by an integrable function. Then, the function $f\left(\mathbf{x}\right)$ is finite-valued and continuous on $\mathcal{C}$ and $\hat{f}_N\left(\mathbf{x}\right)$ converges uniformly to $f\left(\mathbf{x}\right)$ as $N \to \infty$.*

Let u^* and S^* denote the optimal value and the set of optimal solutions of the true problem respectively. Likewise, let $\hat{u}$ and $\hat{S}$ denote their counterpart SAA estimators, i.e, $\hat{u}$ and $\hat{S}$ are the optimal value and the set of optimal solutions of the SAA problem. The following theorem discusses the convergence

property of the SAA estimators. The proof can be found in Ruszczynski and Shapiro (2003, p. 362).

Theorem 5. *Suppose that (i) $\mathcal{C}$ is compact, (ii) the sets of optimal solutions to the true problem, S^*, and to the SAA problem, $\hat{S}$, are non-empty and contained in $\mathcal{C}$, (iii) the function $f(\mathbf{x})$ is finite-valued and continuous on $\mathcal{C}$, and (iv) the function $\hat{f}_N(\mathbf{x})$ converges uniformly to $f(\mathbf{x})$. Then, as $N \to \infty$, the optimal value, $\hat{u}$, converges to u^* and the deviation between the sets S^* and $\hat{S}$ converges to zero, where the deviation between two sets S^* and $\hat{S}$ is defined as*

$$\mathbb{D}\left(S^*, \hat{S}\right) = \sup_{\mathbf{x} \in S^*} \left(\inf_{\mathbf{x}' \in \hat{S}} \left\| \mathbf{x} - \mathbf{x}' \right\| \right).$$

The above theorem implies that the SAA estimator, $\hat{u}$, is a consistent estimator for u^* since it converges w.p.1 to u^* as the Monte Carlo sample size N, goes to infinity.

Convergence rate and sample size estimation So far, the SAA problem associated with the true problem was defined, and a theorem was given that clarified the sufficient conditions for the convergence of the SAA estimators. In practice, we need to choose a finite N value rather than an infinite number of samples. So, a value N is estimated such that the SAA estimators provide a given accuracy. For example, the estimate of the sample size can be obtained based on the convergence rate. The convergence rate of the SAA estimators has been studied extensively, see, e.g., Ruszczynski and Shapiro (2003, pp. 371-382). We explain this subject briefly in the following.

First let us define sets of ϵ-optimal solutions. A feasible point $\bar{\mathbf{x}}$ is said to be an ϵ-optimal solution for $\epsilon \geq 0$, if $f(\bar{\mathbf{x}}) \leq u^* + \epsilon$. The set of ϵ-optimal solutions of the SAA problem and the true problem are denoted by $\hat{S}^\epsilon$ and S^ϵ respectively. Clearly, when $\epsilon = 0$ then $\hat{S}^\epsilon = \hat{S}$ and $S^\epsilon = S^*$.

Consider δ and ϵ such that $0 \leq \delta \leq \epsilon$. It can be shown that $P\left(\hat{S}^\delta \subset S^\epsilon\right) \geq 1 - \alpha$, if the sample size, N, holds the following inequality, see, e.g., Kleywegt et al. (2001):

$$N \geq \frac{3\sigma_{max}^2}{(\epsilon - \delta)^2} \log\left(\frac{|\mathcal{C}|}{\alpha}\right),$$

where $\mathcal{C}$ is a finite set and $|\mathcal{C}|$ is the number of elements in the set $\mathcal{C}$ and

$$\sigma_{max}^2 = \max_{\mathbf{x} \in \mathcal{C} - S^\epsilon} Var(F(u(\mathbf{x}), \boldsymbol{\xi}) - F(\mathbf{x}, \boldsymbol{\xi})).$$

This suggests an estimate of the sample size required to find an ϵ-optimal solution with probability at least $1 - \alpha$. However, this estimation has two shortcomings. First, it is not easy to compute σ_{max}^2 and/or $|\mathcal{C}|$. Second, the obtained estimates of N are typically too conservative for a practical use

(Kleywegt et al., 2001). In practice, it has been suggested to use a relatively smaller sample size and investigate the validity of the SAA estimators later.

Validation analysis Suppose $\hat{\mathbf{x}}$ is an optimal solution to the SAA problem with a finite sample size. We define $u^* - f(\hat{\mathbf{x}})$ as the *optimality gap*. In the following, we describe a technique to estimate the optimality gap of a candidate solution, $\hat{\mathbf{x}}$. This technique is based on the construction of an upper and a lower bound for the true optimal value, u^*. The distance between the upper and lower bounds is estimated and used as an estimate of the optimality gap. This method was suggested by Mak et al. (1999).

Clearly for all $\mathbf{x}' \in \mathcal{C}$, $\min_{\mathbf{x}\in\mathcal{C}} \left\{\hat{f}_N(\mathbf{x})\right\} \leq \hat{f}_N(\mathbf{x}')$. By taking the expectation on both sides and minimizing the right hand side over all $\mathbf{x}' \in \mathcal{C}$, we have

$$\begin{aligned} E\left(\min_{\mathbf{x}\in\mathcal{C}} \left\{\hat{f}_N(\mathbf{x})\right\}\right) &\leq \min_{\mathbf{x}'\in\mathcal{C}} \left\{E\left(\hat{f}_N(\mathbf{x}')\right)\right\}, \text{ or} \\ E(\hat{u}) &\leq u^*. \end{aligned}$$

So, $\hat{u}$ is a negative biased estimator for u^*. On the other hand, since u^* is the minimum value of $E(F(\mathbf{x}, \boldsymbol{\xi}))$ for all $\mathbf{x} \in \mathcal{C}$, then $u^* \leq E(F(\hat{\mathbf{x}}, \boldsymbol{\xi}))$. Therefore,

$$E(\hat{u}) \leq u^* \leq E(F(\hat{\mathbf{x}}, \boldsymbol{\xi})). \tag{18}$$

To estimate the lower bound, we generate M independent samples of the random vector $\boldsymbol{\xi}$ each of size N and solve the SAA counterpart problem M times. Assume $\hat{u}_l, l = 1, 2, \ldots, M$, denote the obtained optimal values. Then the quantity

$$\bar{u} = \frac{1}{M}\sum_{l=1}^{M} \hat{u}_l$$

is an unbiased estimator of $E(\hat{u})$. In addition, an estimate of the variance of the above estimator can be computed as

$$S_M^2 = \frac{1}{M-1}\sum_{l=1}^{M} (\hat{u}_l - \bar{u})^2.$$

For the upper bound, we generate a sample of the random vector $\boldsymbol{\xi}$ with size N', say, and estimate the true objective value $f(\hat{\mathbf{x}}) = E(F(\hat{\mathbf{x}}, \boldsymbol{\xi}))$. Since computing f at specific points is usually easy, we can choose N' much larger than N. The quantity

$$\bar{F}_{N'}(\hat{\mathbf{x}}) = \frac{1}{N'}\sum_{t=1}^{N'} F\left(\hat{\mathbf{x}}, \boldsymbol{\xi}^{(t)}\right),$$

is an unbiased estimator of $f(\hat{\mathbf{x}})$. In addition, an estimate of the variance of the above estimator can be computed as

$$S^2_{N'}(\hat{\mathbf{x}}) = \frac{1}{N'-1}\sum_{t=1}^{N'}\left(F\left(\hat{\mathbf{x}}, \boldsymbol{\xi}^{(t)}\right) - \bar{F}_{N'}(\hat{\mathbf{x}})\right)^2.$$

Finally, an estimate of the interval (18) is as follows:

$$\bar{u} - \psi^{-1}(1-\alpha)\frac{S_M}{\sqrt{M}} \leq u^* \leq \bar{F}_{N'}(\hat{\mathbf{x}}) + \phi^{-1}(1-\alpha)\frac{S_{N'}(\hat{\mathbf{x}})}{\sqrt{N'}},$$

where ψ is the cumulative distribution function of the t-distribution with $M-1$ degrees of freedom and ϕ is the standard normal cumulative distribution function.

If the estimated optimality gap is not larger than some pre-specified threshold value, it means that the sample size N has given a sufficiently good optimal solution. Otherwise, we need to solve the SAA problem with a larger sample size.

Algorithm We state the following algorithm to solve an optimization problem with an expected value objective function. As an example, this algorithm can be used to solve problem (16). Before running the algorithm, we choose the values M, N and N' for the number of replications of the algorithm and the sizes of the Monte Carlo sampling. If the value

$$\frac{(UpperBound - LowerBound)}{LowerBound} \times 100\%$$

is not sufficiently small, e.g., less than 5%, we would repeat the algorithm for another value of N.

1 For $l = 1, 2, \ldots, M$, do the following steps:

1-1 **Generate i.i.d. samples:** Generate $\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(N)}$ from the distribution of the random vectors $\boldsymbol{\xi}$.

1-2 **Solve the SAA problem:** Solve the SAA problem associated with the true problem using the generated samples in step 1-1. Let $\hat{\mathbf{x}}_l$ and $\hat{u}_l$ be the optimal solution and the optimal value obtained in iteration l.

1-3 **Calculate $\bar{F}_{N'}(\hat{\mathbf{x}}_l)$:** Generate $\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(N')}$ and calculate the value $\bar{F}_{N'}(\hat{\mathbf{x}}_l)$.

2 Calculate $\bar{u} = \frac{1}{M}\sum_{l=1}^{M}\hat{u}_l$ and $S_M^2 = \frac{1}{M-1}\sum_{l=1}^{M}(\hat{u}_l - \bar{u})^2$. Let

$$LowerBound = \bar{u} - \psi^{-1}(1-\alpha)\frac{S_M}{\sqrt{M}}.$$

3 Find $\min_l \bar{F}_{N'}(\hat{\mathbf{x}}_l)$ and let $\hat{\mathbf{x}}_{Optimal}$ be its relevant optimal solution.

4 Generate $\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(N')}$ and recalculate $\bar{F}_{N'}\left(\hat{\mathbf{x}}_{Optimal}\right)$ and evaluate

$$S_{N'}^2\left(\hat{\mathbf{x}}_{Optimal}\right) = \frac{1}{N'-1}\sum_{t=1}^{N'}\left(F\left(\hat{\mathbf{x}}_{Optimal}, \boldsymbol{\xi}^{(t)}\right) - \bar{F}_{N'}\left(\hat{\mathbf{x}}_{Optimal}\right)\right)^2.$$

Finally, let

$$UpperBound = \bar{F}_{N'}\left(\hat{\mathbf{x}}_{Optimal}\right) + \phi^{-1}(1-\alpha)\frac{S_{N'}\left(\hat{\mathbf{x}}_{Optimal}\right)}{\sqrt{N'}}.$$

5 Compute the gap between the upper and lower bounds.

In the above, we studied a stochastic problem with an expected value objective function while the feasible set is independent of the sample. The feasible set in problem (A.6), which we solve in paper A, is not independent of the sample. In fact, problem (A.6) (page 63) is a chance constraint problem with an expected value objective function. The SAA method for solving a chance constraint problem is explained in the following section.

An optimization problem with chance constraints Consider problem (15) with a convex objective function and a closed set $\mathcal{X}$. Let us call this problem the *true* problem. We restrict our discussion to the case where $m = 1$. In fact, there is no loss of generality in using $m = 1$, since the joint chance constraint in problem (15) can be written as

$$P\left(g_j(\mathbf{x}, \boldsymbol{\xi}) \leq 0, j = 1, 2, \ldots, m\right) = P\left(g(\mathbf{x}, \boldsymbol{\xi}) \leq 0\right) \geq 1 - \alpha_0,$$

where $g(\mathbf{x}, \boldsymbol{\xi}) = \max_j g_j(\mathbf{x}, \boldsymbol{\xi})$. For the sake of simplicity, we first assume that the expected value function $f(\mathbf{x}) = E(F(\mathbf{x}, \boldsymbol{\xi}))$ is given explicitly and only the chance constraint should be approximated. However, in problem (A.6), the objective function can not be simply computed and we need to approximate the objective function and the chance constraint together. In this case, as it is explained in paper A, we combine the two SAA algorithms.

Suppose that we can generate independent samples $\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(N)}$ from the distribution of the random vector $\boldsymbol{\xi}$. Clearly, we can write a proba-

bility function as an expectation function, e.g., $P(x \in A) = E(\mathbb{1}_A(x))$, where $\mathbb{1}_A(.)$ is the indicator function such that $\mathbb{1}_A(x) = 1$ when $x \in A$ and equals zero otherwise. Therefore, the sample average approximation problem associated with the true problem is written as $\min_{\mathbf{x}\in\mathcal{X}} \{f(\mathbf{x}) \,|\hat{p}_N(\mathbf{x}) \geq 1-\alpha\}$, where

$$\hat{p}_N(\mathbf{x}) = \frac{1}{N}\sum_{t=1}^{N} \mathbb{1}_{(-\infty,0)}\left(g\left(\mathbf{x}, \boldsymbol{\xi}^{(t)}\right)\right).$$

In this definition, we let the SAA problem have a risk level $\alpha \in (0,1]$ which may be different from α_0 in problem (15).

Convergence analysis The function $\hat{p}_N(\mathbf{x})$ is the proportion of times that $g\left(\mathbf{x}, \boldsymbol{\xi}^{(t)}\right) \leq 0$, $t = 1, 2, \ldots, N$. $\hat{p}_N(\mathbf{x})$ is clearly an unbiased estimator of $p(\mathbf{x}) = P(g(\mathbf{x}, \boldsymbol{\xi}) \leq 0)$. In addition, based on the LLN, $\hat{p}_N(\mathbf{x})$ converges to $p(\mathbf{x})$, w.p.1 as $N \to \infty$ for every fixed $\mathbf{x}$. To ensure the convergence of the optimal value of the SAA problem to the optimal value of the true problem, we need a uniform convergence of $\hat{p}_N(\mathbf{x})$ to $p(\mathbf{x})$.

The following theorem implies the conditions that satisfy the uniform convergence of $\hat{p}_N(\mathbf{x})$ to $p(\mathbf{x})$. The proof can be found in Shapiro et al. (2009, p. 211).

Theorem 6. *Suppose that (i) the LLN conditions hold (ii) $g(\mathbf{x}, \boldsymbol{\xi})$ is Caratheodory, i.e., measurable for every $\mathbf{x} \in \mathcal{X}$ and continuous for a.e. $\boldsymbol{\xi}$, and (iii) for every $\mathbf{x} \in \mathcal{X}$, it holds that*

$$P\{\boldsymbol{\xi} \in \Xi \,|g(\mathbf{x}, \boldsymbol{\xi}) = 0\} = 0.$$

Then, the function $p(\mathbf{x})$ is continuous on $\mathcal{X}$ and $\hat{p}_N(\mathbf{x})$ converges uniformly to $p(\mathbf{x})$ as $N \to \infty$.

Let u^* and S^* denote the optimal value and the set of optimal solutions of the true problem respectively. Likewise, let $\hat{u}$ and $\hat{S}$ denote their counterpart SAA estimators. The following theorem discusses the convergence property of the SAA estimators, where $\alpha = \alpha_0$. The proof can be found in Shapiro et al. (2009, p. 211).

Theorem 7. *Suppose that (i) the LLN conditions hold (ii) $\mathcal{X}$ is compact, (iii) the function $f(\mathbf{x})$ is continuous, (iv) $g(\mathbf{x}, \boldsymbol{\xi})$ is Caratheodory, and (v) there exist an optimal solution $\bar{\mathbf{x}}$ such that for any $\epsilon > 0$, there is a solution $\mathbf{x} \in \mathcal{X}$ with $\|\mathbf{x} - \bar{\mathbf{x}}\| \leq \epsilon$ and $P(g(\mathbf{x}, \boldsymbol{\xi}) \leq 0) \geq 1 - \alpha_0$. Then, as $N \to \infty$, the optimal value, $\hat{u}$, converges to u^* and the deviation between the sets S^* and $\hat{S}$ converges to zero.*

Theorem 7 clarifies the sufficient conditions for the convergence of the SAA estimators as the sample size, N, goes to infinity. In practice, we need to choose a finite N value rather than an infinite number of samples. Thus, we

use an arbitrary sample size and investigate the validity of the SAA estimators later.

Validation analysis Suppose $\hat{\mathbf{x}}$ is an optimal solution to the SAA problem with a finite sample size. There are two issues about this point that we need to verify. First, we need to define whether it is a feasible solution for the true problem. Second, we need to estimate the optimality gap $u^* - f(\hat{\mathbf{x}})$. In the following, we describe a technique to investigate the feasibility of a candidate solution, $\hat{\mathbf{x}}$, and estimate the optimality gap.

To verify the feasibility of a candidate point, $\hat{\mathbf{x}}$, we generate samples $\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(N')}$ and compute $\hat{p}_{N'}(\hat{\mathbf{x}})$ which is the proportion of times that $g\left(\hat{\mathbf{x}}, \boldsymbol{\xi}^{(t)}\right) \leq 0$, $t = 1, 2, \ldots, N'$. If the sample size, N', is sufficiently large, then $\hat{p}_{N'}(\hat{\mathbf{x}})$ is asymptotically normal with mean $p(\hat{\mathbf{x}})$ and variance $\hat{p}_{N'}(\hat{\mathbf{x}})\left(1 - \hat{p}_{N'}(\hat{\mathbf{x}})\right)/N'$. Therefore, an approximate upper bound, with $1-\beta$ confidence, for the probability $p(\hat{\mathbf{x}})$ is given by

$$\hat{p}_{N'}(\hat{\mathbf{x}}) + \phi^{-1}(1-\beta)\sqrt{\frac{\hat{p}_{N'}(\hat{\mathbf{x}})\left(1 - \hat{p}_{N'}(\hat{\mathbf{x}})\right)}{N'}},$$

where ϕ is the standard normal cumulative distribution function.

The point $\hat{\mathbf{x}}$ is considered as a feasible point for the true problem if the above value is larger than $1 - \alpha_0$.

Now, we find an estimate of the optimality gap by constructing an upper and a lower bound on u^*. To estimate the lower bound, we generate M independent samples of the random vector $\boldsymbol{\xi}$ each of size N and solve the SAA counterpart problem M times. Assume $\hat{u}_l, l = 1, 2, \ldots, M$, denote the obtained optimal values. Rearrange the values $\hat{u}_l$ in a non-decreasing order to obtain the order statistics $\hat{u}_{[l]}$ for $l = 1, 2, \ldots, M$. It can be shown that for any $M \geq \mathcal{M}$, $l \leq \mathcal{M}$ and $\beta \geq \sum_{i=0}^{\mathcal{M}} \binom{\mathcal{M}}{i}(1/2)^{\mathcal{M}}$, we have

$$P\left(\hat{u}_{[l]} \leq u^*\right) \geq 1 - \beta.$$

For example, for $M = 10$ and $l = 2$, we have $P\left(\hat{u}_{[2]} \leq u^*\right) \geq 0.989$. The proof can be found in Shapiro et al. (2009, pp. 218-220). Therefore, $\hat{u}_{[l]}$ is a $100(1-\beta)\,\%$ confidence lower bound for the true optimal value, u^*.

To obtain an upper bound on u^*, we solve the SAA problem with $\alpha \leq \alpha_0$, M times. Let $\hat{u}_l$ and $\hat{\mathbf{x}}_l$ denote the optimal value and one optimal solution of iteration l. An upper bound on u^* is given by the smallest $\hat{u}_l$ which corresponds to a feasible solution.

Algorithm We state the following algorithm to solve a chance constraint problem. Before running the algorithm, we choose the values M, N and N'

for the number of replications of the algorithm and the sizes of the Monte Carlo sampling. If the value

$$\frac{(UpperBound - LowerBound)}{LowerBound} \times 100\%$$

is not sufficiently small, e.g., less than 5%, we would repeat the algorithm for another value of N.

1 For $l = 1, 2, \ldots, M$, do the following steps:

1-1 **Generate i.i.d. samples:** Generate $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(N)}$ from the distribution of the random vectors ξ.

1-2 **Solve SAA (for lower bound):** Solve the SAA problem associated with the true problem, with $\alpha = \alpha_0$, using the generated samples in step 1-1. Let $\hat{\mathbf{x}}_{1_l}$ and $\hat{u}_{1_l}$ be the optimal solution and the optimal value obtained in iteration l.

1-3 **Solve SAA (for upper bound):** Solve the SAA problem associated with the true problem, with $\alpha \leq \alpha_0$, using the generated samples in step 1-1. Let $\hat{\mathbf{x}}_{2_l}$ and $\hat{u}_{2_l}$ be the optimal solution and the optimal value obtained in iteration l.

1-4 **Posteriori check:** Generate $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(N')}$ and estimate $p_N\left(\hat{\mathbf{x}}_{2_l}\right)$ to verify whether $\hat{\mathbf{x}}_{2_l}$ is feasible or non-feasible.

2 Rearrange $f\left(\hat{\mathbf{x}}_{1_1}\right), f\left(\hat{\mathbf{x}}_{1_2}\right), \ldots, f\left(\hat{\mathbf{x}}_{1_M}\right)$ in a non-decreasing order as follows:

$$f\left(\hat{\mathbf{x}}_{1_{[1]}}\right) \leq f\left(\hat{\mathbf{x}}_{1_{[2]}}\right) \leq \ldots \leq f\left(\hat{\mathbf{x}}_{1_{[M]}}\right),$$

where $f\left(\hat{\mathbf{x}}_{1_{[1]}}\right)$ is the smallest achieved optimal value and $\hat{\mathbf{x}}_{1_{[1]}}$ is its relevant optimal solution, $f\left(\hat{\mathbf{x}}_{1_{[2]}}\right)$ is the second smallest optimal value and $\hat{\mathbf{x}}_{1_{[2]}}$ is its relevant optimal solution, and so on.

3 Pick for example the second smallest optimal value and let it be the lower bound.

4 Find $\min\left\{f\left(\hat{\mathbf{x}}_{2_l}\right) \middle| \hat{\mathbf{x}}_{2_l}\ isfeasible\right\}$ and let it be the upper bound. Also, let $\hat{\mathbf{x}}_{Optimal}$ be its relevant optimal solution.

5 Compute the gap between the upper and lower bounds.

In the above algorithm, we assumed that the objective function has been given explicitly. As stated before, problem (A.6) in paper A includes an expected value objective function and a chance constraint together. Hence, the algorithm which is explained in section 3.2.3 of paper A, combines the two SAA algorithms which are discussed above.

1.3.3 Majorization

The SAA problem associated with a chance constraint problem is usually a combinatorial problem which is difficult to be solved. Sometimes the chance constraint is simplified as a non-probabilistic constraint. Of course, when there is no chance constraint, the stochastic problem could be solved much easier. For example, assume a stochastic problem with a chance constraint defined by a linear inequality as

$$\begin{aligned} &\min_{\mathbf{x}} \ f(\mathbf{x}) \\ &\text{subject to:} \\ &\begin{cases} P\left(\mathbf{x}^T\boldsymbol{\xi} \leq b\right) \geq 1-\alpha_0, \\ \mathbf{x} \in \mathcal{X}, \end{cases} \end{aligned} \tag{19}$$

where b is a fixed value. In addition, assume that the random vector $\boldsymbol{\xi}$ has a multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix Σ. In this setting, the chance constraint defines a convex set, and it can be transformed to the following constraint:

$$\boldsymbol{\mu}^T\mathbf{x} + \phi^{-1}(1-\alpha_0)\sqrt{\mathbf{x}^T\Sigma\mathbf{x}} \leq b. \tag{20}$$

Thus, under the assumption of normality, the stochastic problem (19) is simplified and solved easier.

Now, consider a situation where ξ_i for any $i = 1, 2, \ldots, n$ is not normally distributed, but there is another variable, say ζ_i, which is normally distributed and majorizes ξ_i as it is defined below.

Definition 9. *A random variable X majorizes (dominates) another random variable Y, denoted $X \succeq Y$ if*

$$P(X \leq y) \leq P(Y \leq y), \forall y \in \mathbb{R}.$$

It is common to say that the random variable X is *"more diffused"* than the random variable Y, if $X \succeq Y$.

One way to bound the optimal value of problem (19) from above is to replace the ξ_i's with more diffused normally distributed random variables, ζ_i's. Clearly, this leads to a stochastic problem with a constraint similar to inequality (20) which is easier to handle. The majorization theorem, which implies this method for finding an upper bound on the optimal value, is as follows:

Theorem 8. *Let $x_1, x_2, \ldots, x_n$ and b be some (deterministic) known values. Assume that the random variables $\xi_i, i = 1, 2, \ldots, n$, are independent and ζ_i's be a similar collection of random variables which majorize the variables ξ_i's, i.e., $\zeta_i \succeq \xi_i$ for every*

i. Then,

$$P\left(\sum_{i=1}^{n} x_i\zeta_i \leq b\right) \leq P\left(\sum_{i=1}^{n} x_i\xi_i \leq b\right).$$

Proof. This is proved by induction. Assume that $\zeta_1 \succeq \xi_1$. Based on the definition of majorization, we have $P(x_1\zeta_1 \leq b) \leq P(x_1\xi_1 \leq b)$. Assume that $P\left(\sum_{i=1}^{n-1} x_i\zeta_i \leq b\right) \leq P\left(\sum_{i=1}^{n-1} x_i\xi_i \leq b\right)$. We need to show that $P(\sum_{i=1}^{n} x_i\zeta_i \leq b) \leq P(\sum_{i=1}^{n} x_i\xi_i \leq b)$. To do so, we have

$$\begin{aligned}
P\left(\sum_{i=1}^{n} x_i\xi_i \leq b\right) &= \int P\left(x_n\xi_n \leq b - \sum_{i=1}^{n-1} x_iy_i\right) dF_{\xi_1,\ldots,\xi_n}(y_1,\ldots,y_{k-1}) \\
&\geq \int P\left(x_n\zeta_n \leq b - \sum_{i=1}^{n-1} x_iy_i\right) dF_{\xi_1,\ldots,\xi_n}(y_1,\ldots,y_{k-1}) \\
&= P\left(x_n\zeta_n + \sum_{i=1}^{n-1} x_i\xi_i \leq b\right) \\
&= \int P\left(\sum_{i=1}^{n-1} x_i\xi_i \leq b - x_ny_n\right) dF_{\zeta_n}(y_n) \\
&\geq \int P\left(\sum_{i=1}^{n-1} x_i\zeta_i \leq b - x_ny_n\right) dF_{\zeta_n}(y_n) \\
&= P\left(\sum_{i=1}^{n} x_i\zeta_i \leq b\right). \qquad \square
\end{aligned}$$

According to the above theorem, the set $\mathcal{C}_1 = \{\mathbf{x} \in \mathcal{X} \,|P(\sum_{i=1}^{n} x_i\zeta_i \leq b) \geq 1 - \alpha_0\}$ is contained in the set $\mathcal{C}_2 = \{\mathbf{x} \in \mathcal{X} \,|P(\sum_{i=1}^{n} x_i\xi_i \leq b) \geq 1 - \alpha_0\}$. Therefore, the optimal value of problem $\min_{\mathbf{x}\in\mathcal{C}_1} f(\mathbf{x})$ is an upper bound for the optimal value of problem $\min_{\mathbf{x}\in\mathcal{C}_2} f(\mathbf{x})$. Since we have assumed that the ζ_i's are normally distributed random variables, it is not difficult to obtain the upper bound. A similar discussion on bounding a chance constraint problem from above has been given by Ben-Tal et al. (2009, pp. 105-109).

Likewise, we can bound the optimal value of problem (19) from below by replacing the ξ_i's with less diffused normally distributed random variables. This leads to an algorithm which defines an upper and a lower bound for the optimal value of the stochastic problem (19).

Finally, consider a situation where ξ_i for any $i = 1, 2, \ldots, n$ is not normally distributed and neither is there a normally distributed random variable which majorizes ξ_i, but there is a normally distributed random variable, say ζ_i, which γ-majorizes ξ_i as it is defined below.

Definition 10. *Fix $0 \leq \gamma \leq 1$, a random variable X γ-majorizes (γ-dominates) another random variable Y, denoted $X \succeq_\gamma Y$ if*

$$P(X \leq y) \leq P(Y \leq y) + \gamma, \forall y \in \mathbb{R}.$$

Using the above definition, our γ-majorization theorem is as follows:

Theorem 9. *Let $x_1, x_2, \ldots, x_n$ and b be some (deterministic) known values. Assume that the random variables $\xi_i, i = 1, 2, \ldots, n$, are independent random variables and ζ_i's be similar collection of random variables which γ_i-majorize the variables ξ_i's, i.e., $\zeta_i \succeq_{\gamma_i} \xi_i$ for every i. Then,*

$$P\left(\sum_{i=1}^{n} x_i \zeta_i \leq b\right) \leq P\left(\sum_{i=1}^{n} x_i \xi_i \leq b\right) + \sum_{i=1}^{n} \gamma_i.$$

The proof is similar to the proof of theorem 8, and it is also given in section 3.2.4 of paper A. According to the above theorem, the set

$$\mathcal{C}_1 = \left\{ \mathbf{x} \in \mathcal{X} \,\middle|\, P\left(\sum_{i=1}^{n} x_i \zeta_i \leq b\right) \geq 1 - \alpha_0 + \sum_{i=1}^{n} \gamma_i \right\}$$

is contained in the set

$$\mathcal{C}_2 = \left\{ \mathbf{x} \in \mathcal{X} \,\middle|\, P\left(\sum_{i=1}^{n} x_i \xi_i \leq b\right) \geq 1 - \alpha_0 \right\}.$$

Therefore, the optimal value of the problem $\min_{\mathbf{x} \in \mathcal{C}_1} f(\mathbf{x})$ is an upper bound for the optimal value of problem (19). If $\sum_{i=1}^{n} \gamma_i$ is close to zero such that $\sum_{i=1}^{n} \gamma_i \ll \alpha_0$, we can be hopeful to attain a good upper bound for the problem.

Likewise, a lower bound on the optimal value of problem (19) is obtained by replacing the ξ_i's with some normally distributed random variables, ζ_i's, where $\xi_i \succeq_{\gamma_i} \zeta_i$ for every i.

With the above discussion, to be able to find an upper bound and a lower bound on the optimal value of problem (19), we should first find

- normally distributed random variables, e.g., $\zeta_{1_i} \sim N\left(\mu_{1_i}, \sigma_{1_i}^2\right), i = 1, 2, \ldots, n$, that are γ-majorized by the random variables of the problem, i.e., $\xi_i \succeq_{\gamma_i} \zeta_{1_i}, \forall i$, and

- normally distributed random variables, e.g., $\zeta_{2_i} \sim N\left(\mu_{2_i}, \sigma_{2_i}^2\right), i = 1, 2, \ldots, n$, that γ-majorize the random variables of the problem, i.e., $\zeta_{2_i} \succeq_{\gamma_i} \xi_i, \forall i$.

One way to find such random variables with small fixed γ_i values is to solve the following two optimization problems for $i = 1, 2, \ldots, n$:

$$\begin{aligned}
&\max_{\mu_{1_i}, \sigma_{1_i}} \ \mu_{1_i} \\
&\text{subject to:} \\
&\begin{cases} P\left(\xi_i \leq y\right) \leq P\left(\frac{\zeta_{1_i} - \mu_{1_i}}{\sigma_{1_i}} \leq \frac{y - \mu_{1_i}}{\sigma_{1_i}}\right) + \gamma_i, \forall y \in \mathbb{R}, \\ \sigma_{1_i} \geq 0, \end{cases}
\end{aligned}$$

and

$$\begin{aligned}
&\min_{\mu_{2_i}, \sigma_{2_i}} \ \mu_{2_i} \\
&\text{subject to:} \\
&\begin{cases} P\left(\frac{\zeta_{2_i} - \mu_{2_i}}{\sigma_{2_i}} \leq \frac{y - \mu_{2_i}}{\sigma_{2_i}}\right) \leq P\left(\xi_i \leq y\right) + \gamma_i, \forall y \in \mathbb{R}, \\ \sigma_{2_i} \geq 0. \end{cases}
\end{aligned}$$

In the application, we use a finite subset of $\mathbb{R}$ that has been observed or generated from the known specific distribution of ξ_i. This reduces the above two optimization problems to simple LP problems.

The constraint in problem (A.6) can also be simplified as inequality (20) by substituting the random variables with normally distributed random variables that γ-majorize and are γ-majorized by the random variables of the problem. Note that in problem (A.6), there is also an expected value objective function. Hence, the algorithm which is explained in section 3.2.4 of paper A, combines the γ-majorization trick with the SAA algorithm.

To sum up, in paper A, we present two different algorithms. One combines the SAA algorithm for solving a stochastic problem with an expected value objective function and the SAA algorithm for solving a chance constraint problem. The other combines the SAA algorithm for solving a stochastic problem with an expected value objective function and the γ-majorization trick. We compare the results of both methods with respect to two things. One is the estimated optimality gap, and the other is the time of solving the problem. We will see that the γ-majorization algorithm gives a larger gap but runs faster than the SAA algorithm.

2 Statistical modeling

Consider a dataset consisting of a binary response random variable and perhaps some other variables. A possible approach to analyze a binary response variable is to construct a statistical model. A *statistical model* is a mathematical representation of the relationship between the response variable and

some other variables, together with a measure of uncertainty. The aim of this section is explaining different statistical models which are used in analyzing such data with an emphasis on the correlated (in the sense explained later) binary response data.

The statistical models for *correlated binary data* are divided into two somewhat different categories including cluster-specific models, such as mixed-effects models and population-averaged models, such as beta-binomial models. We explain some examples of these models in sections 2.2.1-2.2.3. Correlated data may exhibit overdispersion. Hence, we also explain this phenomenon and its potential causes in this section.

This chapter is divided into two parts. First, some basic models for analyzing binary data are studied. Then, some models for analyzing correlated binary data are discussed.

2.1 Short introduction to the logistic regression model

Assume that $Z_1, Z_2, \ldots, Z_N$ are N binary random variables such that $Z_i \in \{0, 1\}, \forall i$. Let $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ci})$ denote c explanatory variables observed for $i = 1, 2, \ldots, N$. Assume that the Z_is depend on the values of c explanatory variables. An appropriate distribution for Z_i is the Bernoulli distribution with probability of success equal $p_i = P(Z_i = 1 | \mathbf{x}_i)$. It is clear that the response variables, Z_is, are not identically distributed since the p_is differ from one data point, $\mathbf{x}_i$, to another, but they are assumed to be independent conditional on the values of the explanatory variables. A statistical model can be defined by a logistic transformation or logit of a success probability, $\text{logit}(p_i) = \log(p_i/(1-p_i))$, linearly related to the c explanatory variables. This is called a *logistic regression model*. A more general model than a linear logistic regression model allows a nonlinear predictor expression and is called *logistic-nonlinear* regression model and is formulated as

$$\log\left(\frac{p_i}{1-p_i}\right) = g(\mathbf{x}_i, \boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ is a vector of unknown parameters of the model.

A logistic-nonlinear regression model can also be defined with *grouped binary data* (also called *clustered binary data*). To do so, suppose that there are k groups of binary data each having a common value of $\mathbf{x}$, labeled with $\mathbf{x}_j, j = 1, 2, \ldots, k$. Let Y_j denote the number of successes out of n_j independent trials, for $j = 1, 2, \ldots, k$, where the logistic transform of the corresponding probability of success, p_j, is to be modeled as a combination of the values of c explanatory variables, i.e., $\text{logit}(p_j) = g(\mathbf{x}_j, \boldsymbol{\theta})$. The appropriate distribution for the response variables in this case is the binomial distribution, and $Y_1, Y_2, \ldots, Y_k$, conditional on the values of the explanatory variables, are

assumed to be independent.

A logistic-nonlinear regression model is a member of a class of models called *generalized nonlinear models*. Nonlinear models have been studied by many authors, see, e.g., Jennrich (1969); Seber and Wild (1989) and Zhanga et al. (2005).

The parameters of the model can be estimated with standard methods, such as maximum likelihood method and Markov Chain Monte Carlo method through a Bayesian analysis. There are also standard R packages to fit some well-known generalized nonlinear models, see, e.g., the `gnm` package in R by Turner and Firth (2005). In paper B, section 3.1, we define a logistic-nonlinear regression model on data of one specific airport and use a Bayesian approach to estimate the parameters of the model.

As mentioned above, one vital assumption in a logistic regression model is that $Z_1, Z_2, \ldots, Z_N$ are assumed to be independent conditional on the values of the explanatory variables, or in other words Y_j for $j = 1, 2, \ldots, k$ is the number of successes out of n_j independent trials. Sometimes the binary observations in the same group tend to exhibit intracluster correlation. In this case, standard models such as logistic-nonlinear regression models, that ignore this dependence, are inadequate to represent the observed data.

For example, in paper B, an informal graphical test assesses the lack of fit of the defined logistic-nonlinear regression model. This graphical test represents the clustering pattern of the data which suggests the existence of such an intracluster correlation.

Since the logistic-nonlinear regression model is inadequate for modeling the correlated binary data, we need to consider other statistical models for analyzing such data. Some statistical models, which are appropriate for analyzing the correlated binary data, are studied in the next section.

2.2 Correlated binary data and Overdispersion problem

In a logistic regression model, it is assumed that the logistic transformation of the response probability of success depends only on the values of the explanatory variables, and the number of successes are assumed to have a binomial distribution. Sometimes, the assumptions of the logistic regression model are not valid. Therefore, the logistic regression model is not appropriate and probably causes some problems.

To explain the above, let the individual binary random variables that make up Y_j, which was defined in section 2.1, be denoted with $Z_{j1}, Z_{j2}, \ldots, Z_{jn_j}$. Hence $Y_j = \sum_{l=1}^{n_j} Z_{jl}$, for $j = 1, 2, \ldots, k$. Now, assume that Z_{jl} may not be independent of $Z_{jl'}$ when $l \neq l'$. In this case, the assumption of the binomial distribution for Y_j is not valid and a standard logistic regression model seems inappropriate. Suppose that the correlation between Z_{jl} and $Z_{jl'}$ equals ρ. The

mean and variance of Y_j conditional on n_j and $\mathbf{x}_j$ are

$$\begin{aligned} E\left(Y_j \left| n_j, \mathbf{x}_j\right.\right) &= \sum_{l=1}^{n_j} E\left(Z_{jl}\right) = n_j p_j, \\ Var\left(Y_j \left| n_j, \mathbf{x}_j\right.\right) &= \sum_{l=1}^{n_j} Var\left(Z_{jl}\right) + 2 \sum_{1 \le l < l' \le n_j} Cov\left(Z_{jl}, Z_{jl'}\right) \\ &= n_j p_j \left(1 - p_j\right) \left(1 + \left(n_j - 1\right) \rho\right). \end{aligned}$$

It is clear that if $\rho = 0$, there is no correlation between the binary random variables conditional on $\mathbf{x}_j$ and Y_j has a binomial distribution, but when there is a positive correlation, the variance of Y_j exceeds the variance of a binomial random variable. This phenomenon is called *overdispersion* or *extra binomial variation*. In general, overdispersion means that the observed variation in a variable is greater than what is expected from a model.

Above, we explained how the invalidity of the assumption of independence, conditional on the values of the explanatory variables, between the binary responses causes the overdispersion problem. The problem of overdispersion may also happen due to some other reasons (Collett, 2002, p. 196). One reason may be misspecification of the predictor expression, e.g., when a linear predictor is assumed while it should be nonlinear. It may also be caused by the existence of outliers. Modification or omission of these data points may help to alleviate this problem. In paper B, section 3.1.5, we consider this potential reason and find that although the model fit is improved by omission of a few data points, the observed lack of fit still persist and is annoying. Finally, the overdispersion problem may happen when the response probabilities of successes vary over groups of binary random variables with similar values of the explanatory variables.

In fact, the last mentioned reason above is related to the correlation between the binary random variables. This means that the existence of positive correlation, conditional on the values of the explanatory variables, leads to the variation of the probabilities of successes over groups of binary random variables with similar values of the explanatory variables, and vice versa. Since the two reasons are explaining the same thing, it is expected that they lead to the same statistical model (Collett, 2002, pp. 197-199).

To face the problem of overdispersion, we may look for some other explanatory variables which can explain the observed extra binomial variation. However, it is more realistic to define a model which postulate a source of extra binomial variation (Williams, 1982).

The problem of overdispersion for the binomial data has been extensively studied so far, see, e.g., Hinde and Demetrio (1998) and Collett (2002). The models which are suggested for these types of data are classified in two large

categories of *conditional* (also known as *cluster-specific*) versus *marginal* (also known as *population-averaged*) models. In the first one, a cluster effect is included in the model and in the second one, it is not included. When the research interest focuses on the changes between clusters, a model with cluster effect is more appropriate, see Neuhaus et al.'s discussion (Neuhaus et al., 1991) for comparing the two types of models. Beta-binomial, mixed-effects, mixture and hidden Markov models are some examples of models which are used for solving the overdispersion problem. In the following, we discuss some of these models.

2.2.1 Beta-binomial model

As mentioned previously, one reason for the occurrence of overdispersion is that the probabilities of successes vary over groups of binary random variables with similar values of the explanatory variables. In the following, we make some assumptions about this variation which lead to a general model for overdispersed data. This general model has been suggested by Williams (1982). Afterward, we explain the beta-binomial model which is a special case of Williams's general model.

Suppose that $\mathbf{x}_j$, n_j and Y_j for $j = 1, 2, \ldots, k$ are defined as before. Now, assume that the corresponding probability of success, p_j, varies around μ_j with variance $\delta\mu_j\left(1-\mu_j\right)$, where the logistic transform of μ_j is a function of the explanatory variables, i.e., $\text{logit}\left(\mu_j\right) = g\left(\mathbf{x}_j, \boldsymbol{\theta}\right)$. In this case, the mean and variance of Y_j conditional on n_j and $\mathbf{x}_j$ are

$$\begin{aligned} E\left(Y_j|n_j,\mathbf{x}_j\right) &= E\left[E\left(Y_j|n_j,p_j\right)\right] = n_j\mu_j, \\ Var\left(Y_j|n_j,\mathbf{x}_j\right) &= Var\left(E\left[Y_j|n_j,p_j\right]\right) + E\left[Var\left(Y_j|n_j,p_j\right)\right] \\ &= Var\left(n_jp_j\right) + E\left[n_jp_j\left(1-p_j\right)\right] \\ &= n_j\mu_j\left(1-\mu_j\right)\left(1+\left(n_j-1\right)\delta\right). \end{aligned} \tag{21}$$

It is clear that if $\delta = 0$, the probability of success, conditional on $\mathbf{x}_j$, does not vary and Y_j has a binomial distribution, but when δ is greater than zero, the variance of Y_j exceeds the variance of a binomial random variable and that is why the overdispersion happens. So, an appropriate statistical model for the overdispersed data is defined as follows:

$$\begin{aligned} &\left(Y_j\,|n_j,p_j\right) \sim Bin\left(n_j,p_j\right), \\ &E\left(p_j\,|\mathbf{x}_j\right) = \mu_j, \\ &Var\left(p_j\,|\mathbf{x}_j\right) = \delta\mu_j\left(1-\mu_j\right), \\ &\log\left(\frac{\mu_j}{1-\mu_j}\right) = g\left(\mathbf{x}_j,\boldsymbol{\theta}\right). \end{aligned}$$

We call this model *Williams's general model*. In particular, we can assume p_j as a random variable with a specific distribution function, e.g., a beta distribution. This model was first introduced by Williams (1975) and is called the *beta-binomial model*. The beta-binomial model as a special case of Williams's general model is defined as follows:

$$\begin{aligned}
&\left(Y_j \left| n_j, p_j \right.\right) \sim Bin\left(n_j, p_j\right), \\
&\left(p_j \left| \mathbf{x}_j \right.\right) \sim Beta\left(\mu_j\tau, \left(1-\mu_j\right)\tau\right), \\
&\log\left(\frac{\mu_j}{1-\mu_j}\right) = g\left(\mathbf{x}_j, \boldsymbol{\theta}\right),
\end{aligned}$$

where *Beta* denotes the beta distribution, and the density function of p_j given $\mathbf{x}_j$ is

$$f\left(p_j \left| \mathbf{x}_j \right.\right) = \frac{1}{B\left(\mu_j\tau, \left(1-\mu_j\right)\tau\right)} p_j^{\mu_j\tau-1}\left(1-p_j\right)^{\left(1-\mu_j\right)\tau-1},$$

where B is the Beta function.

Note that in this model, p_j given $\mathbf{x}_j$ follows a beta distribution with mean μ_j and variance $\mu_j\left(1-\mu_j\right)/(\tau+1)$. Thus, the mean and variance of Y_j conditional on n_j and $\mathbf{x}_j$ are the same as (21), where δ is substituted with $1/(\tau+1)$.

To estimate the parameters of the model, we need to obtain the likelihood function of the parameters $\boldsymbol{\theta}$ and τ. With the assumption of independence between $Y_1, Y_2, \ldots, Y_k$, conditional on the values of the explanatory variables, the likelihood function of the observations equals

$$L\left(\boldsymbol{\theta}, \tau \left| y_1, y_2, \ldots, y_k \right.\right) = \prod_{j=1}^{k} \int_0^1 \binom{n_j}{y_j} \frac{\left(\exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}\right)\right)\right)^{y_j}}{\left(1+\exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}\right)\right)\right)^{n_j}} f\left(p_j \left| \mathbf{x}_j \right.\right) dp_j,$$

where $f\left(p_j \left| \mathbf{x}_j \right.\right)$ is defined as before.

Since the beta distribution is the conjugate distribution of the binomial, the likelihood function is simplified as

$$L\left(\boldsymbol{\theta}, \tau \left| y_1, y_2, \ldots, y_k \right.\right) = \prod_{j=1}^{k} \binom{n_j}{y_j} \frac{B\left(\mu_j\tau + y_j, \left(1-\mu_j\right)\tau + n_j - y_j\right)}{B\left(\mu_j\tau, \left(1-\mu_j\right)\tau\right)}. \tag{22}$$

Clearly, this likelihood function is computationally convenient.

In section 3.2 of paper B, we define a beta-binomial logistic-nonlinear regression model. Since this is a nonlinear model, we could not use the standard R packages for obtaining the maximum likelihood estimates. Instead, we have made an R code to estimate the parameters of the model through a Bayesian approach.

In this section, we explained how the variation of the probabilities of

successes, over groups of binary random variables with similar values of the explanatory variables, leads to the overdispersion. Since the probabilities of successes vary, the assumptions of the binomial distribution are not met. One way to deal with this issue is to use a beta-binomial distribution instead of a binomial distribution. Another way to deal with this issue is to use models with random effects or hidden random variables. The reason is that the observed extra variation may be caused by some explanatory variables that are not recorded, and therefore a model with random effects or hidden random variables, which considers the unrecorded variables as well, seems appropriate. In the following, we explain these types of models to modeling overdispersed data.

2.2.2 Mixed-effects model

The overdispersion problem reveals the fact that a relatively large part of the existent variation in the binary data is not explained by the model. Such a variation may happen due to a certain number of explanatory variables which are not recorded. A random effect can be employed to alleviate this problem. That means, mixed-effects models can be used to model overdispersed data.

As before, suppose that there are k groups of binary data each having a common explanatory variable labeled with $\mathbf{x}_j, j = 1, 2, \ldots, k$. Likewise, let Y_j be the number of successes out of n_j trials in the jth group. Now, assume that the logistic transform of the probability of success, p_j, is a function of c explanatory variables and some other variables which are not recorded. Lets assume that $U_{1j}, U_{2j}, \ldots, U_{c'j}$ are these unknown random variables and the correct model for p_j is

$$\log\left(\frac{p_j}{1-p_j}\right) = g\left(\mathbf{x}_j, \boldsymbol{\theta}\right) + \beta_1 u_{1j} + \beta_2 u_{2j} + \ldots + \beta_{c'} u_{c'j},$$

where $u_{1j}, u_{2j}, \ldots, u_{c'j}$ are the realizations of the unrecorded random variables and $\beta_1, \beta_2, \ldots, \beta_{c'}$, beside $\boldsymbol{\theta}$, are the unknown parameters of the model. There are therefore two sources of variability between the groups. One is the variability which is explained by the values of the explanatory variables, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$, and the other is the variability caused by some unknown random variables. Such an unknown variation is interpreted as overdispersion. We can combine the effect of these variables in a single variable and consider it as a random effect with common mean and variance in our logistic regression model.

The use of random effects for modeling overdispersion was first consid-

ered by Pierce and Sands (1975). The suggested model is defined as follows:

$$\begin{aligned}
&\left(Y_j \left| n_j, p_j \right.\right) \sim Bin\left(n_j, p_j\right), \\
&\log\left(\frac{p_j}{1-p_j}\right) = g\left(\mathbf{x}_j, \boldsymbol{\theta}\right) + \gamma_j,
\end{aligned}$$

where γ_j is a random effect corresponding to the jth group.

We need to assume a probability distribution for the random effects. For example, we assume that $\gamma_1, \gamma_2, \ldots, \gamma_k$ are the realizations of k independent normally distributed random variables with mean zero and variance σ_γ^2. Note that since $g\left(\mathbf{x}_j, \boldsymbol{\theta}\right)$ usually involves an intercept part, it is appropriate to assume that the mean of γ_j equals zero. In this case, the parameter which has to be estimated is σ_γ. We can also substitute $\gamma_j = \sigma_\gamma z_j$ in this model, where z_j is a realization of the standard normal distribution.

To estimate the parameters of the model, we can use the maximum likelihood method. With the assumption of independence between $Y_1, Y_2, \ldots, Y_k$, conditional on the values of the explanatory variables, the likelihood function of the observations equals

$$L\left(\boldsymbol{\theta}, \sigma_\gamma, z_1, z_2, \ldots, z_k \left| y_1, y_2, \ldots, y_k \right.\right) = \prod_{j=1}^{k} \binom{n_j}{y_j} \frac{\left(\exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}\right) + \sigma_\gamma z_j\right)\right)^{y_j}}{\left(1 + \exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}\right) + \sigma_\gamma z_j\right)\right)^{n_j}}.$$

To obtain the marginal likelihood of the parameters $\boldsymbol{\theta}$ and σ_γ, we integrate the above likelihood function with respect to the distribution of the standard normal random variables which is

$$L\left(\boldsymbol{\theta}, \sigma_\gamma \left| y_1, y_2, \ldots, y_k \right.\right) = \prod_{j=1}^{k} \int_{-\infty}^{\infty} \binom{n_j}{y_j} \frac{\left(\exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}\right) + \sigma_\gamma z_j\right)\right)^{y_j}}{\left(1 + \exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}\right) + \sigma_\gamma z_j\right)\right)^{n_j}} \sigma\phi\left(\sigma_\gamma z_j\right) dz_j, \quad (23)$$

where ϕ is the standard normal density.

The likelihood function (23) does not have a closed form expression as the likelihood function of the beta-binomial model, (22). Hence, maximizing the likelihood function is a bit complicated, and it should be approximated with numerical methods. For example, we can approximate the above integral by using the Gauss-Hermite formula as follows:

$$\int_{-\infty}^{\infty} f\left(s\right) e^{-s^2} ds \approx \sum_{r=1}^{m} w_r f\left(s_r\right)$$

where the w_r and s_r values are given in standard tables, see, e.g., Abramowitz and Stegun (1972).

Therefore, the marginal likelihood of the parameters is

$$L\left(\boldsymbol{\theta}, \sigma_\gamma \left| y_1, y_2, \ldots, y_k \right.\right) \approx \pi^{-k/2} \prod_{j=1}^{k} \binom{n_j}{y_j} \sum_{r=1}^{m} w_r \frac{\left(\exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}\right) + \sigma_\gamma s_r \sqrt{2}\right)\right)^{y_j}}{\left(1 + \exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}\right) + \sigma_\gamma s_r \sqrt{2}\right)\right)^{n_j}}.$$

As mentioned above, the likelihood function (23) is more complicated to be computed than the likelihood function (22). However, the approaches for dealing with the overdispersion issue relate to one another. In the following, we discuss how both of the above defined models consider a distribution function for the logistic transformation of the success probabilities.

A comparison between the mixed-effects model and the beta-binomial model In the above defined mixed-effects model, it is assumed that $\gamma_1, \gamma_2, \ldots, \gamma_k$ are k independent normally distributed random variables with mean zero and variance σ_γ^2. Clearly, $\text{logit}\left(p_j\right)$ for $j = 1, 2, \ldots, k$ can be considered as normally distributed random variables with mean $g\left(\mathbf{x}_j, \boldsymbol{\theta}\right)$ and variance σ_γ^2.

Now, lets look at the beta-binomial model which was explained in section 2.2.1. Also in this model, $\text{logit}\left(p_j\right)$ for $j = 1, 2, \ldots, k$ can be considered as a random variable since p_j is assumed to be a random variable. In the following, we show what the distribution of $\text{logit}\left(p_j\right)$ in a beta-binomial model looks like, and we compare it with the assumed distribution of $\text{logit}\left(p_j\right)$ in the defined mixed-effects model.

In a beta-binomial model, p_j given $\mathbf{x}_j$ has a beta distribution with parameters a_j and b_j, where

$$\begin{aligned} a_j &= \frac{\exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}\right)\right)\tau}{1 + \exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}\right)\right)}, \\ b_j &= \frac{\tau}{1 + \exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}\right)\right)}. \end{aligned}$$

It can be shown that if $X \sim Beta\left(d_1, d_2\right)$, then $d_2 X / \left(d_1\left(1 - X\right)\right) \sim F\left(2d_1, 2d_2\right)$, where F is representing an F-distribution. Therefore,

$$\frac{1}{\exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}\right)\right)} \cdot \frac{p_j}{1 - p_j} \left| \mathbf{x}_j \right. \sim F\left(2a_j, 2b_j\right).$$

It is also known that if $X \sim F\left(n, m\right)$, then $1/2 \log\left(X\right) \sim FisherZ\left(n, m\right)$, where FisherZ is representing a Fisher's z-distribution. Therefore,

$$\frac{-1}{2} g\left(\mathbf{x}_j, \boldsymbol{\theta}\right) + \frac{1}{2} \log\left(\frac{p_j}{1 - p_j}\right) \left| \mathbf{x}_j \right. \sim FisherZ\left(2a_j, 2b_j\right).$$

Consequently, the mean of logit (p_j) is $g(\mathbf{x}_j, \boldsymbol{\theta})$ plus twice the mean of a Fisher's z-distribution with parameters $2a_j$ and $2b_j$. Also, the variance of logit (p_j) is four times the variance of a Fisher's z-distribution with parameters $2a_j$ and $2b_j$. Hence, the mean and variance of logit (p_j) depends on the explanatory variables.

In the mixed-effects model, logit (p_j) for $j = 1, 2, \ldots, k$ had a common variance, σ_γ^2, while in a beta-binomial model, it is seen that the variance of logit (p_j) for $j = 1, 2, \ldots, k$ differ according to the values of the explanatory variables.

Summed up briefly, both the beta-binomial and the mixed-effects models define a distribution function for the logistic transformation of the probabilities. With the difference that in a beta-binomial model the variances of logit (p_j), for $j = 1, 2, \ldots, k$ differ according to the values of the explanatory variables rather than being the same.

In paper B, we prefer the beta-binomial model to modeling the overdispersed data. However, in section 4 of this paper, it is suggested to use a mixed-effects beta-binomial logistic regression model to prevent removing some identified outliers. These outliers make the variation of the observed data much larger than what is expected from a beta-binomial model. Therefore, it is suggested to employ a random effect to prevent the omission of the outliers.

2.2.3 Mixture model

As discussed previously, overdispersion may be explained with the variation of the probabilities of successes given the explanatory variables, where such variation is due to a certain number of variables not having been recorded. Since these variables are not recorded, they are called *hidden* or *latent* variables. It can be assumed that the number of successes, conditional on the values of the latent variables and the explanatory variables, have the binomial distribution. This leads to a mixture model.

Suppose that $\mathbf{x}_j$, n_j and Y_j for $j = 1, 2, \ldots, k$ are all defined as before. Now, assume that there is a latent variable, denoted U_j, for each j which is categorically distributed with $\mathcal{M}$ categories. The random variable Y_j, conditional on the values of the explanatory variables, $\mathbf{x}_j$, and the value of the latent variable, $U_j = m$, has a binomial distribution with probability of success equal to p_{jm}. So, the mixture model is defined as follows:

$$
\begin{aligned}
&\left(Y_j \,|\, U_j = m\right) \sim Bin\left(n_j, p_{jm}\right), \\
&U_j \sim Cat(\mathcal{M}, \boldsymbol{\alpha}), \\
&\log\left(\frac{p_{jm}}{1 - p_{jm}}\right) = g\left(\mathbf{x}_j, \boldsymbol{\theta}_m\right),
\end{aligned}
$$

where *Cat* is the categorical distribution with parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_{\mathcal{M}})$ and $\sum_{i=1}^{\mathcal{M}} \alpha_i = 1$.

Above, we have formulated a *finite-mixture model* for modeling overdispersed data. This means that the distribution function of Y_j given n_j and $\mathbf{x}_j$ is the sum of a finite mixture of components as follows:

$$f\left(y_j \left| n_j, \mathbf{x}_j\right.\right) = \sum_{i=1}^{\mathcal{M}} \alpha_i \binom{n_j}{y_j} \frac{\left(exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}_m\right)\right)\right)^{y_j}}{\left(1 + exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}_m\right)\right)\right)^{n_j}}.$$

We can also define an *infinite-mixture model*, where there is an infinite set of component distributions. The beta-binomial model, which was defined in section 2.2.1, is an example of infinite-mixture models, where the distribution function of Y_j given n_j and $\mathbf{x}_j$ is

$$f\left(y_j \left| n_j, \mathbf{x}_j\right.\right) = \int_0^1 \binom{n_j}{y_j} \frac{\left(\exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}\right)\right)\right)^{y_j}}{\left(1 + \exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}\right)\right)\right)^{n_j}} f\left(p_j \left| \mathbf{x}_j\right.\right) dp_j,$$

where $f\left(p_j \left| \mathbf{x}_j\right.\right)$ is the density function of a beta distribution.

One well-known finite-mixture model for modeling overdispersion is the *zero-inflated* model. This model is handling zero-inflated data, where an excess of zeros is present in the data. In this model, the zeros and positive data are modeled as separate populations. Vieira et al. (2000) first applied such a model to a dataset to modeling extra binomial variation.

To define a zero-inflated model, let the latent variable be a binary variable. When the latent variable is one, the random variable Y_j, conditional on the values of the explanatory variables, follows a binomial distribution, and otherwise it equals zero. So, the zero-inflated model is defined as follows:

$$\begin{aligned}
&\left(Y_j \left| U_j = 1\right.\right) \sim Bin\left(n_j, p_j\right), \\
&\left(Y_j \left| U_j = 0\right.\right) = 0, \\
&U_j \sim Bernoulli(\alpha), \\
&\log\left(\frac{p_j}{1 - p_j}\right) = g\left(\mathbf{x}_j, \boldsymbol{\theta}\right).
\end{aligned}$$

In this case, the distribution of Y_j, conditional on the latent variable, u_j, is given by

$$f\left(y_j \left| u_j\right.\right) = \left(\binom{n_j}{y_j} \frac{\left(\exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}\right)\right)\right)^{y_j}}{\left(1 + \exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}\right)\right)\right)^{n_j}}\right)^{u_j} I\left(y_j = 0\right)^{\left(1 - u_j\right)}.$$

With the assumption of independence between $Y_1, Y_2, \ldots, Y_k$, conditional on the values of the explanatory variables, and independence between $U_1, U_2, \ldots, U_k$,

the likelihood function of the observations equals

$$\begin{aligned} L\left(\boldsymbol{\theta}, \alpha \left| y_1, y_2, \ldots, y_k, u_1, u_2 \ldots, u_k \right.\right) &= \prod_{j=1}^{k} f\left(y_j \left| u_j \right.\right) \alpha^{u_j} \left(1-\alpha\right)^{\left(1-u_j\right)} \\ &= \prod_{j=1}^{k} \left(\binom{n_j}{y_j} \frac{\left(\exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}\right)\right)\right)^{y_j}}{\left(1+\exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}\right)\right)\right)^{n_j}} \alpha \right)^{u_j} \times \\ &\quad \left(I\left(y_j = 0\right)\left(1-\alpha\right)\right)^{\left(1-u_j\right)}. \end{aligned} \tag{24}$$

If we sum over the values of U_j, then the marginal likelihood of the parameters $\boldsymbol{\theta}$ and α based on the observed data equals

$$L\left(\boldsymbol{\theta}, \alpha \left| y_1, y_2, \ldots, y_k \right.\right) = \prod_{j=1}^{k} \left[\binom{n_j}{y_j} \frac{\left(\exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}\right)\right)\right)^{y_j}}{\left(1+\exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}\right)\right)\right)^{n_j}} \alpha + I\left(y_j = 0\right)\left(1-\alpha\right) \right]. \tag{25}$$

The maximum likelihood estimates of the parameters, $\boldsymbol{\theta}$ and α, can be obtained by using the Expectation-Maximization (EM) algorithm. This algorithm is suited to problems including missing or hidden values. Since the u_j values are hidden, it seems appropriate to use this algorithm. The goal is to estimate the unknown parameters, $\boldsymbol{\theta}$ and α, by maximizing the likelihood of the incomplete data (including only observed values), i.e., (25). While in the EM algorithm, the likelihood of the complete data (including observed and hidden values) is maximized, i.e., (24). This is because maximizing the likelihood of the complete data is easier than the incomplete data. The obtained maximum likelihood parameter estimates, by the EM algorithm, typically converge to the true maximum likelihood estimates.

The algorithm is started with some initial values for the parameters, $\boldsymbol{\theta}^{(0)}$ and $\alpha^{(0)}$, and calculates

$$\sum_{j=1}^{k} E\left[\log\left(L\left(\boldsymbol{\theta}, \alpha \left| y_j, u_j \right.\right)\right) \left| \boldsymbol{\theta}^{(0)}, \alpha^{(0)}, y_j \right.\right].$$

To do so, we need to calculate

$$\begin{aligned} E\left[U_j \left| \boldsymbol{\theta}^{(0)}, \alpha^{(0)}, y_j \right.\right] &= P\left(U_j = 1 \left| \boldsymbol{\theta}^{(0)}, \alpha^{(0)}, y_j \right.\right) \\ &= \frac{L\left(\boldsymbol{\theta}^{(0)}, U_j = 1 \left| y_j \right.\right) P\left(U_j = 1 \left| \alpha^{(0)} \right.\right)}{L\left(\boldsymbol{\theta}^{(0)}, \alpha^{(0)} \left| y_j \right.\right)} \\ &= \frac{\alpha^{(0)}}{\alpha^{(0)} + \left(1-\alpha^{(0)}\right)\left(1+exp\left(g\left(\mathbf{x}_j, \boldsymbol{\theta}^{(0)}\right)\right)\right)^{n_j}} I\left(y_j = 0\right) + \end{aligned}$$

$$I\left(y_j \neq 0\right).$$

The above gives estimates for the u_j values. In the next step, the loglikelihood of the complete data is maximized, where the u_j values are substituted with the obtained estimates from the previous step. The above two steps are repeated several times until a stopping rule criterion is satisfied. The EM algorithm is robust with respect to the choice of starting values and will generally always converge.

As mentioned before, in paper B, we use a beta-binomial model to modeling the overdispersed data. However, in section 4 of this paper, a zero-inflated beta-binomial logistic-nonlinear regression model is suggested to modeling the overdispersion without removing the outliers.

We stated that the statistical models for modeling overdispersion belong to two large categories of marginal and conditional models. The beta-binomial model is a marginal model since it does not define any cluster effect. On the contrary, the other discussed models are all conditional models since they include a cluster effect. For example, in a mixed-effects model, random effects corresponding to each cluster are defined, and in mixture models, categorical/binary latent variables corresponding to each cluster are defined.

References

M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs and mathematical tables*. U.S. Government Printing Office, Washington, D.C., 1972.

T. Achterberg, T. Koch, and A. Martin. Branching rules revisited. *Operations Research*, 33:42–54, 2005.

M. Y. An. Log-concave probability distributions: theory and statistical testing. `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1933`, 1995.

M. Avriel. r-convex functions. *Mathematical Programming*, 2:309–323, 1972.

A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton University Press, Princeton and Oxford, 2009.

D. Bertsimas and J. N. Tsitsiklis. *Introduction to linear optimization*. Athena Scientific, 1997.

J. R. Birge and F. Louveaux. *Introduction to stochastic programming*. Springer-Verlag, New York, 2 edition, 2011.

C. Borell. Convex measures on locally convex spaces. *Arkiv for Matematik*, 12: 239–252, 1974.

S. Boyd and A. Mutapcic. Stochastic subgradient methods. http://see.stanford.edu/materials/lsocoee364b/04-stoch_subgrad_notes.pdf, 2006.

S. Boyd and J. Park. Subgradient methods. http://web.stanford.edu/class/ee364b/lectures/subgrad_method_notes.pdf, 2007.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, New York, 2004.

H.J. Brascamp and E.H. Lieb. On extensions of the brunn-minkowski and prekopa-leindler theorems including inequalities for log concave functions, and with an application to the diffusion equations. *Journal of Functional Analysis*, 22:366–389, 1976.

A. Charnes and W. W. Cooper. Chance constrained programming. *Management Science*, 6:73–79, 1959.

D. Collett. *Modelling binary data*. Chapman and Hall, 2002.

G. B. Dantzig. Linear programming under uncertainty. *Management Science*, 1:197–206, 1955.

G. B. Dantzig and M. N. Thapa. *Linear programming 1: Introduction*. Springer-Verlag, New York, 1997.

D. Dentcheva. *Optimization models with probabilistic constraints in Probabilistic and randomized methods for design under uncertainty*. Springer London, 2006.

D. Dentcheva, A. Prekopa, and A. Ruszczyński. Concavity and efficient points of discrete distributions in probabilistic programming. *Mathematical Programming*, 89:55–77, 2000.

M. Elshahed, H. Zeineldin, and M. Elmarsfawy. A bivariate chance constraint of wind sources for multi-objective dispatching. *Smart Grid and Renewable Energy*, 4:325–332, 2013.

C. Giuseppe and D. Fabrizio. *Probabilistic and randomized methods for design under uncertainty*. Springer-Verlag, 2006.

R. Henrion and A. Moller. Optimization of a continuous distillation process under random inflow rate. *Computers and Mathematics with Applications*, 45: 247–262, 2003.

J. Hinde and C. G. B. Demetrio. overdispersion: models and estimation. *Computational Statistics and Data Analysis*, 27:151–170, 1998.

R. I. Jennrich. Asymptotic properties of nonlinear least squares estimators. *The Annals of Mathematical Statistics*, 40:633–643, 1969.

A. J. King and S. W. Wallace. *Modeling with stochastic programming*. Springer, 2012.

A. J. Kleywegt, A. Shapiro, and T. Homem-De-Mello. The sample average approximation method for stochastic discrete optimization. *SIAM J.OPTIM*, 12:479–502, 2001.

H. W. Kuhn and A. W. Tucker. Nonlinear programming. *Proceedings of 2nd Berkeley Symposium. Berkeley: University of California Press*, pages 481–492, 1951.

K. Lange. *Optimization*. Springer, New York, 2004.

M. A. Lejeune and A. Ruszczynski. An efficient trajectory method for probabilistic inventory-production-distribution problems. *Operations Research*, 55:378–394, 2007.

J. Luedtke and S. Ahmed. A sample approximation approach for optimization with probabilistic constraints. *SIAM Journal on Optimization*, 19:674–699, 2008.

W. Mak, D. Morton, and K. Wood. Monte carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters*, 24:47–56, 1999.

J. E. Mitchell. Branch-and-cut algorithms for combinatorial optimization problems. *Handbook of Applied Optimization*, pages 65–77, 2002.

J. M. Neuhaus, J. D. Kalbfleisch, and W. W. Hauck. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review*, 59:25–35, 1991.

B. K. Pagnoncelli, S. Ahmed, and A. Shapiro. Sample average approximation method for chance constrained programming: theory and applications. *J Optim Theory Appl*, 142:399–416, 2009.

D. A. Pierce and B. R. Sands. Extra-bernoulli variation in binary data. *Technical Report 46, Department of Statistics, Oregon State Universit*, 1975.

F. A. Potra and S. J. Wright. Interior-point methods. *Journal of Computational and Applied Mathematics*, 124:281–302, 2000.

A. Prekopa. Logarithmic concave measures with application to stochastic programming. *Acta Scientiarum Mathematicarum*, 32:301–316, 1971.

Y. Rinott. On convexity of measures. *Annals of Probability*, 4:1020–1026, 1976.

A. Ruszczynski and A. Shapiro. Monte-carlo sampling methods. *Handbooks in OR and MS*, 10:353–425, 2003.

G. A. F. Seber and C. J. Wild. *Nonlinear regression*. Wiley, New York, 1989.

A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on stochastic programming modelling and theory*. MPS-SIAM serios on optimization, Philadelphia, 2009.

N. Z. Shor. *Nondifferentiable optimization and polynomial problems*. Springer Science and Business Media, 1998.

H. Turner and D. Firth. Generalized nonlinear models in r: An overview of the gnm package. `http://cran.r-project.org/web/packages/gnm/index.html`, 2005.

A. M. C. Vieira, J. P. Hinde, and C. G. B. Demetrio. Zero-inflated proportion data models applied to a biological control assay. *Journal of Applied Statistics*, 27:373–389, 2000.

A. Wald. Statistical decision functions which minimize the maximum risk. *The Annals of Mathematics*, 46:265–280, 1945.

S. W. Wallace and W. T. Ziemba. *Applications of stochastic programming*. MPS-SIAM Series on Optimization, 2005.

Q. Wang, Y. Guan, and J. Wang. A chance-constrained two-stage stochastic program for unit commitment with uncertain wind power output. *IEEE Transactions on power systems*, 27:206–215, 2012.

D. A. Williams. The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, 31:949–952, 1975.

D. A. Williams. Extra-binomial variation in logistic linear models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31:144–148, 1982.

M. H. Wright. The interior-point revolution in optimization: History, recent developments, and lasting consequences. *Bulletin of the American Mathematical Society*, 42:39–56, 2004.

N. Zhanga, B. Weib, and J. Linc. Generalized nonlinear models and variance function estimation. *Computational Statistics and Data Analysis*, 48:549–570, 2005.

ISSN (online): 2246-1248
ISBN (online): 978-87-7112-359-3

AALBORG UNIVERSITY PRESS