DEPARTMENT OF OPERATIONS RESEARCH
UNIVERSITY OF AARHUS

Working Paper no. 2002/1

A bicriteria stochastic programming model for capacity expansion in telecommunications

Morten Riis
Jørn Lodahl

ISSN 1600-8987

Department of Mathematical Sciences
Telephone: +45 8942 1111
E-mail: institut@imf.au.dk

Building 530, Ny Munkegade
DK-8000 Aarhus C, Denmark
URL: www.imf.au.dk

A bicriteria stochastic programming model for capacity expansion in telecommunications

MORTEN RIIS*
Department of Operations Research
University of Aarhus, building 530
Ny Munkegade
DK - 8000 Århus C
Denmark

JØRN LODAHL
Sonofon
Skelagervej 1
DK - 9100 Ålborg
Denmark

January 18, 2002

Abstract

We consider capacity expansion of a telecommunications network in the face of uncertain future demand and potential future failures of network components. The problem is formulated as a bicriteria stochastic program with recourse in which the total cost of the capacity expansion and the probability of future capacity requirements to be violated are simultaneously minimized. Assuming the existence of a finite number of possible future states of the world, an algorithm for the problem is elaborated. The algorithm determines all non-dominated solutions to the problem by a reduced feasible region method, solving a sequence of restricted subproblems by a cutting plane procedure. Computational results are reported for three different problem instances, one of which is a real-life problem faced by Sonofon, a Danish communications network operator.

Keywords: Capacity Expansion; Telecommunications; Stochastic Programming; Integer Programming; Bicriteria Optimization.

1 Introduction

Capacity expansion problems is an important class of problems arising in many contexts. Uncertainty is almost always an inherent feature of the system being modeled, and the importance of taking due account of this uncertainty when formulating the problem as an optimization problem, is well-recognized. This paper is concerned with capacity expansion of a telecommunications network in an uncertain environment. The uncertainties

*Corresponding author. Email: riis@imf.au.dk

facing the network operator are assumed to be twofold. Some arises due to the inherent uncertainty involved in the assessment of future demand and some is due to the potential future failure of nodes or edges in the network. Previous studies have mainly dealt with these issues separately. Capacity expansion problems with uncertain future demand have been considered by authors such as Dempster, Medova and Thompson [5], Medova [9], Riis and Andersen [13], Riis, Skriver and Lodahl [15], and Sen, Doverspike and Cosares [17]. Capacity expansion problems including potential future failures in the network have been considered in the framework of survivable network design by e.g. Dahl and Stoer [4] and Rios, Marianov and Gutierrez [16]. The emphasis in most of these studies has been on minimization of the expected cost of capacity installments. Sen, Doverspike and Cosares [17], however, used another approach in which capacity expansion was planned so as to minimize the expected number of unserved requests subject to a budget constraint. Dempster, Medova and Thompson [5] and Medova [9] use chance-constrained programming to solve the capacity expansion problem subject to constraints limiting the blocking probabilities at different network levels, whereas the remaining capacity expansion models fall in the general category of two-stage stochastic programs with recourse. (This terminology, though, is usually not used in connection with survivable network design.) The basic assumption underlying the two-stage stochastic programming model is that decisions can be split into two groups, a group of first-stage decisions which must be taken with only distributional information on the uncertainties of the model, and a group of second-stage decisions which may be postponed until uncertainty has been revealed. In the case of capacity expansion problems, the first stage corresponds to the planning of capacity installments and the second stage corresponds to routing of traffic in the network once actual demand has been observed and a failure has possibly occured. For a general introduction to the field of stochastic programming, we refer to the textbooks by Birge and Louveaux [3], Kall and Wallace [8], and Prékopa [12].

The demand input for the long-term capacity expansion planning models under consideration here is a set of capacity requirements between node-pairs, needed to maintain a prescribed Grade-of-Service (GoS). (See e.g. Dempster, Medova and Thompson [5], Medova [9], or Riis and Andersen [13] for related discussions.) In the case of node or edge failures it is required that a certain fraction of these capacity requirements are available to uphold the GoS for all node-pairs. Since the prescribed GoS, as well as the fraction of capacity requirements to be available in case of failures, are selected somewhat arbitrarily, however, refusing to waive these requirements under any circumstances, may not make sense in a cost minimization framework. In other words, we may obtain a considerable decrease in the optimal cost by relaxing the requirements for a few critical failure states. Moreover, such a cost reduction would be of major interest if the probability of the critical failures to actually supervene is very small. To illuminate the trade-off between total cost of the capacity expansion and the probability of GoS requirements to be violated, we formulate a bicriteria model for capacity expansion in which these two objectives are simultaneously minimized. A general bicriteria problem takes the form

$$\begin{aligned} \min\ z_1 &= f_1(x) \\ \min\ z_2 &= f_2(x) \\ \text{s.t. } x &\in S. \end{aligned} \tag{1}$$

Since, in general, we cannot expect to obtain a solution $\bar{x} \in S$ which minimizes both objectives over S, it is not immediately clear what an "optimal" solution of problem (1) should be. The relevant concept in this respect is that of efficient solutions, defined next. Let the feasible region in criterion space be

$$\mathcal{Z} = \left\{ (z_1, z_2) \in \mathbb{R}^2 \,\middle|\, \exists x \in S : z_1 = f_1(x),\ z_2 = f_2(x) \right\}.$$

Definition 1. A criterion vector $(z_1, z_2) \in \mathcal{Z}$ is dominated if there exists $x \in S$ such that $f_1(x) \leq z_1$ and $f_2(x) \leq z_2$ with at least one inequality being strict. Otherwise (z_1, z_2) is a non-dominated criterion vector.

Definition 2. A solution vector $x \in S$ is efficient if $\big(f_1(x), f_2(x)\big)$ is a non-dominated criterion vector. Otherwise x is inefficient.

For a basic introduction to multicriteria optimization we refer to Steuer [18].

This paper is organized as follows. In Section 2 we develop a bicriteria stochastic integer programming formulation of the problem. Next, in Section 3 we elaborate two algorithms. The first algorithm determines all non-dominated solutions to the bicriteria problem by a reduced feasible region method. The second algorithm is used to solve a series of restricted subproblems arising during the course of the first algorithm. In Section 4 we present the results of some computational experiments performed on three different problem instances. One of these is a real-life problem provided by Sonofon, a Danish mobile communications network operator, whereas the other two are modified instances of real-life problems previously studied in e.g. Bienstock and Günlük [2], Günlük [6], and Riis and Andersen [13]. Finally, in Section 5 we give some concluding remarks.

2 Problem Formulation

The network is modeled as a connected undirected graph $G = (V, E)$, where V denotes the set of nodes (switches) and E denotes the set of edges (circuit groups). Also a set K

of point-to-point pairs of nodes between which demand is to be routed is given. We shall think of the future state of the world as a random event ω and denote by $(\Omega, \mathcal{F}, P)$ the underlying probability space. Associated with ω is a specific failure state (possibly no failure) reducing the set of functional nodes and edges to $V(\omega)$ and $E(\omega)$, respectively, and a set of point-to-point demands $D_k(\omega)$ between functional node-pairs $k \in K(\omega)$, that are to be routed in the network $G(\omega) = (V(\omega), E(\omega))$. We assume that routing of traffic is restricted to a set of prespecified routes $\mathcal{P}$. Given a state of the world ω, the set of functional routes between node-pairs $k \in K(\omega)$ is denoted by $\mathcal{P}_k(\omega)$ and the set of functional routes which use the edge $\{i,j\} \in E(\omega)$ is denoted by $\mathcal{Q}_{ij}(\omega)$. Finally, we let $\mathcal{P}(\omega) = \bigcup_{k \in K(\omega)} \mathcal{P}_k(\omega)$ denote the set of all functional routes given the state of the world ω. The existing capacity on an edge $\{i,j\} \in E$ is denoted by C_{ij}. Additional capacity on the edge may be installed in multiples of a fixed batch size. In particular, a facility providing a capacity of λ may be installed at a unit cost of c_{ij}. By rescaling demand and existing capacity, we may assume that $\lambda = 1$. The bicriteria stochastic programming model may now be formulated as

$$\begin{aligned} \min\, z_1 &= \sum_{\{i,j\} \in E} c_{ij} x_{ij} \\ \min\, z_2 &= P\big(\omega \in \Omega : \phi(x, \omega) > 0\big) \\ \text{s.t. } & x \in \mathbb{Z}_+^{|E|}. \end{aligned} \tag{2}$$

Here x_{ij} denotes the number of facilities to be installed on edge $\{i,j\} \in E$. The first objective z_1 is the total cost of the capacity expansion while the second objective z_2 is the probability of capacity requirements to be violated. Hence, as the function $\phi(\cdot, \omega)$, we can use any function which is less or equal to zero if sufficient capacity is installed and strictly greater than zero otherwise. One possibility is to use the following definition,

$$\begin{aligned} \phi(x, \omega) := \min & \sum_{k \in K(\omega)} t_k \\ \text{s.t.} & \sum_{p \in \mathcal{P}_k(\omega)} f_p + t_k = \rho_k(\omega) D_k(\omega), \quad k \in K(\omega), \\ & \sum_{p \in \mathcal{Q}_{ij}(\omega)} f_p \le C_{ij} + x_{ij}, \quad \{i,j\} \in E(\omega), \\ & f_p \ge 0, \quad p \in \mathcal{P}(\omega), \end{aligned} \tag{3}$$

where f_p denotes the amount of capacity allocated to route $p \in \mathcal{P}(\omega)$, and $\rho_k(\omega)$ is the fraction of capacity requirements between node-pairs $k \in K(\omega)$ that should be available in state $\omega \in \Omega$.

Remark 1. Modeling the actual process of real-time call-by-call routing within a long-term planning model such as problem (2)-(3) is obviously not viable. As briefly pointed out in Section 1, the demand inputs for this problem is a set of capacity requirements for each point-to-point pair, needed to maintain a prescribed Grade-of-Service (GoS). (See e.g. Dempster, Medova and Thompson [5] and Medova [9] who consider ATM-based broadband integrated services digital networks (B-ISDN) and determine the capacity requirements for each point-to-point pair as the effective bandwidth requirements needed to ensure that a set of blocking probabilities are not exceeded.) In Sen, Doverspike and Cosares [17] the approximation of real-time routing by a static model similar to problem (3) was validated using simulation, and the results are encouraging.

Remark 2. The assumption that routing of demand is restricted to a set of prespecified routes is a common one, employed also by e.g. Sen, Doverspike and Cosares [17], Dempster, Medova and Thompson [5], and Medova [9]. The assumption may be justified by the fact that most static real-time routing algorithms implemented in switch software choose routes from a limited set, allowing us to simply enumerate the routes of interest.

Remark 3. As pointed out in Section 1, problem (2) fits in the general framework of two-stage stochastic programs with recourse. The first stage includes decisions on capacity expansion x which must be taken before the future state of the world is known. Once uncertainty is revealed, the second-stage decision, consisting of allocation of capacity to routes f, is settled. One might argue that a three-stage formulation of the problem would more accurately capture the actual alternating process of decisions and observations of random outcomes. In the first stage, as before, the decision on capacity expansion is taken. In the second stage, an actual outcome of random demand is observed and the capacity is allocated accordingly to routes. Finally, in the third stage, a failure possibly occurs and capacity may be reallocated among routes. Clearly, however, the second-stage decisions in such a formulation is of no importance for the capacity planning model, and hence the second and third stages may be joined to obtain the two-stage formulation (2). See also Riis, Skriver and Lodahl [15] for a related discussion.

Remark 4. Ignoring the first objective z_1, problem (2) turns into a special case of the so-called minimum risk problem considered by Riis and Schultz [14]. In this paper structural properties of the problem are investigated and, in particular, the authors establish lower semicontinuity of the objective z_2 with respect to x. Using this fact, problem (2) is obviously well-defined. Riis and Schultz also elaborated an algorithm for the minimum risk problem, and the seminal idea of their approach is a corner stone in the solution procedure for problem (2) presented in Section 3.

3 Solution Procedure

For practical purposes we need to make the following assumption.

(A1) The distribution P of ω is discrete and has finite support, say $\Omega = \{\omega^1, \ldots, \omega^S\}$ with corresponding probabilities $P\big(\{\omega^1\}\big) = \pi^1, \ldots, P\big(\{\omega^S\}\big) = \pi^S$.

From now on we shall refer to a possible future state of the world $\omega^s \in \Omega$ as a scenario, and for notational convenience we let $D_k^s = D_k(\omega^s)$ for all $k \in K(\omega^s)$ and define V^s, E^s, K^s, $\mathcal{P}^s$, $\mathcal{P}_k^s$, $\mathcal{Q}_{ij}^s$, and ρ_k^s accordingly for $k \in K(\omega^s)$ and $\{i,j\} \in E(\omega^s)$.

Employing Assumption (A1), problem (2)-(3) may be restated as

$$\begin{aligned} \min\ z_1 &= \sum_{\{i,j\}\in E} c_{ij} x_{ij} \\ \min\ z_2 &= \sum_{s=1}^{S} \pi^s \psi(x, \omega^s) \\ \text{s.t. } & x \in \mathbb{Z}_+^{|E|}, \end{aligned} \tag{4}$$

where $\psi : \mathbb{Z}_+^{|E|} \times \Omega \mapsto \mathbb{B}$ is an indicator function defined by

$$\psi(x, \omega) := \begin{cases} 1 & \text{if} \quad \phi(x, \omega) > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

3.1 Finding all Non-Dominated Solutions

To determine all non-dominated solutions of problem (4), we observe that the second objective z_2 can only take on a finite number of values, say $p_1, \ldots, p_n$ in any solution to the problem. Hence to obtain all non-dominated solutions to the bicriteria problem, we may simply solve the following problem for all possible values p,

$$\begin{aligned} \min\ z_1 &= \sum_{\{i,j\}\in E} c_{ij} x_{ij} \\ \text{s.t. }\ z_2 &= \sum_{s=1}^{S} \pi^s \psi(x, \omega^s) \le p, \\ & x \in \mathbb{Z}_+^{|E|}. \end{aligned} \tag{6}$$

We shall refer to problem (6) as the p-restricted problem.

Remark 5. Note that problem (6) is feasible for all values of p since z_2 can be made arbitrarily small (equal to zero) by installing sufficient capacity. Hence an optimal solution can always be found in Step 2 of Algorithm 1.

Algorithm 1

Step 1 (*Initialization*) Let $0 = p_1 < \cdots < p_n$ be the possible values of $\sum_{s \in J} \pi^s$, $J \subseteq \{1, \ldots, S\}$ to be considered. Set $\mathcal{L} = \emptyset$ and $i = 1$.

Step 2 (*Solve problem*) Solve the p-restricted problem (6) with $p = p_i$ and let (x^i, z_1^i, z_2^i) be an optimal solution vector.

Step 3 (*Update list*) If $z_1^i < z_1^{i-1}$ then set $\mathcal{L} = \mathcal{L} \cup \{(x^i, z_1^i, z_2^i)\}$.

Step 4 (*Termination*) If $i = n$ then stop. Otherwise set $i = i+1$ and go to Step 2.

Remark 6. In general the number n may be very large ($n \leq 2^S$). For practical purposes, however, it will often be sufficient to consider a modest number of possible values for p in Algorithm 1. This happens for two reasons. First of all the network operator is most likely to accept only very small values of p and hence all values p_i exceeding some maximum acceptable level may be discarded beforehand. Secondly, the number of possible values is reduced if any of the scenario probabilities are equal. In particular, if the distribution P is uniform on Ω, i.e. $\pi^1 = \cdots = \pi^S$, we have $n \leq S+1$. Uniform scenario probabilities are often used in practical studies for example when scenarios are generated by sampling.

Proposition 1. *At termination of Algorithm 1, the set* $\{(z_1, z_2) \mid (x, z_1, z_2) \in \mathcal{L}\}$ *is the set of non-dominated criterion vectors.*

Proof. First note that $z_1^1 \geq \cdots \geq z_1^n$ since $p_1 < \cdots < p_n$. Now, assume that in some iteration i of the algorithm we have $z_1^i < z_1^{i-1}$ but (z_1^i, z_2^i) is a dominated criterion vector, i.e. there exists a solution (x, z_1, z_2) to problem (4) such that $z_1 \leq z_1^i$ and $z_2 \leq z_2^i$ with at least one inequality being strict. Now, we must have $z_2 < z_2^i$, since $z_1 < z_1^i$ contradicts optimality of (x^i, z_1^i, z_2^i) in problem (6) with $p = p_i$. This means that $z_2 = p_j$ for some $j < i$. But this contradicts optimality of (x^j, z_1^j, z_2^j) in problem (6) with $p = p_j$ since $z_1^j \geq z_1^{i-1} > z_1^i = z_1$ by assumption. Hence we see that only solutions for which the criterion vector is non-dominated are put into the list $\mathcal{L}$.

To see that the set $\{(z_1, z_2) \mid (x, z_1, z_2) \in \mathcal{L}\}$ contains all non-dominated criterion vectors, we assume that, in some iteration i of the algorithm, the solution (x^i, z_1^i, z_2^i) is not put into the list $\mathcal{L}$, i.e. we have $z_1^i = z_1^{i-1}$. If $z_2^i > z_2^{i-1}$ the criterion vector (z_1^i, z_2^i) is dominated by (z_1^{i-1}, z_2^{i-1}). Assume on the contrary that $z_2^i \leq z_2^{i-1}$. Then we have $z_2^i = p_j$ for some $j < i$. Now, we obviously have that $z_2^j \leq z_2^i$, and since (x^i, z_1^i, z_2^i) is a feasible solution for problem (6) with $p = p_j$ we must also have $z_1^j \leq z_1^i$. Thus the criterion vector (z_1^i, z_2^i) is either equal to or dominated by (z_1^j, z_2^j) and hence the solution (x^i, z_1^i, z_2^i) can be excluded from the list $\mathcal{L}$ with no loss of non-dominated solutions. The result follows, since all possible values of z_2 are considered during the course of the algorithm. $\square$

Remark 7. Clearly, several efficient solutions may correspond to the same non-dominated criterion vector. Hence if one should want the set $\{x \mid (x, z_1, z_2) \in \mathcal{L}\}$ to contain all efficient solution vectors, it is necessary, in Step 2 of Algorithm 1, to determine *all* optimal solution vectors to problem (6) in each iteration i where $z_1^i < z_1^{i-1}$.

3.2 Solving the p-Restricted Problems

The question remains how to efficiently solve problem (6) in Step 2 of Algorithm 1. The approach we suggest here is based on the seminal idea presented by Riis and Schultz [14]. The idea is for each scenario $\omega \in \Omega$ to replace the indicator function $\psi(\cdot, \omega)$ by a binary variable and a number of cutting planes derived through linear programming duality. In particular, for any feasible solution x, the binary variable θ^s representing $\psi(x, \omega^s)$ should be equal to one if and only if $\phi(x, \omega^s) > 0$, where we recall that

$$\begin{aligned} \phi(x, \omega^s) := \min & \sum_{k \in K^s} t_k \\ \text{s.t.} & \sum_{p \in \mathcal{P}_k^s} f_p + t_k = \rho_k^s D_k^s, \quad k \in K^s, \\ & \sum_{p \in \mathcal{Q}_{ij}^s} f_p \leq C_{ij} + x_{ij}, \quad \{i,j\} \in E^s, \\ & f_p \geq 0, \quad p \in \mathcal{P}^s. \end{aligned} \tag{7}$$

Consider now the dual of problem (7). Letting $M > 0$ be some upper bound on the optimal value of this problem and denoting by $\mathcal{D}^s$ the set of extreme points of the feasible region, it is easily seen that problem (6) is equivalent to the following problem,

$$\begin{aligned} \min\ & z_1 = \sum_{\{i,j\} \in E} c_{ij} x_{ij} \\ \text{s.t.}\ & z_2 = \sum_{s=1}^{S} \pi^s \theta^s \leq p, \\ & \sum_{k \in K^s} \rho_k^s D_k^s u_k - \sum_{\{i,j\} \in E^s} (C_{ij} + x_{ij}) v_{ij} \leq M \theta^s, \quad (u, v) \in \mathcal{D}^s,\ s = 1, \ldots, S, \\ & x \in \mathbb{Z}_+^{|E|},\ \theta \in \mathbb{B}^{|S|}. \end{aligned}$$

Remark 8. We note that problem (7) is always feasible and bounded and hence the same things go for its dual. Thus an optimal solution to the problems always exists and their optimal values are equal. Moreover, an upper bound on the optimal value of the problems, obtained letting $f_p = 0$ for $p \in \mathcal{P}^s$ and $s = 1, \ldots, S$, is

$$M := \max_{s \in \{1, \ldots, S\}} \sum_{k \in K^s} \rho_k^s D_k^s.$$

Remark 9. The cutting planes described above may be seen as generalizations of the metric inequalities originally introduced by Iri [7] and Onaga and Kakusho [11]. These inequalities have been employed as valid inequalities in cutting plane procedures for the capacitated network design problem in a stochastic setting by Riis and Andersen [13] and in a deterministic setting by e.g. Bienstock et al. [1].

The algorithm progresses by sequentially solving a master problem and adding violated cutting planes generated through the solution of subproblems (7). Hence for some subsets $\mathcal{E}^s \subseteq \mathcal{D}^s$ of the dual extreme points, we define the master problem as the following relaxation in which only some of the cutting planes are included,

$$\begin{aligned} \min\ z = & \sum_{\{i,j\}\in E} c_{ij}x_{ij} \\ \text{s.t.}\ & \sum_{s=1}^{S} \pi^s\theta^s \leq p, \\ & \sum_{k\in K^s} \rho_k^s D_k^s u_k - \sum_{\{i,j\}\in E^s} (C_{ij}+x_{ij})v_{ij} \leq M\theta^s, \quad (u,v)\in\mathcal{E}^s,\ s=1,\dots,S, \\ & x\in\mathbb{Z}_+^{|E|},\ \theta\in\mathbb{B}^{|S|}. \end{aligned} \tag{8}$$

Algorithm 2

Step 1 (*Initialization*) Set $\nu = 0$ and let $\mathcal{E}^s \subseteq \mathcal{D}^s$ for $s = 1,\dots,S$ be subsets of dual extreme points for which the corresponding cutting plane is included in the initial master problem.

Step 2 (*Solve master problem*) Solve the current master problem (8) and let (x_ν, θ_ν) be an optimal solution vector with optimal value z_ν.

Step 3 (*Solve subproblems*) Solve the second-stage problem (7) corresponding to all scenarios for which $\theta_\nu^s = 0$. Consider the following situations:

1. If $\phi(x_\nu, \omega^s) = 0$ for all of these scenarios, then the current solution x_ν is optimal for problem (6) with $(z_1, z_2) = (z_\nu, \sum_{s=1}^{S} \pi^s\theta_\nu^s)$.
2. If $\phi(x_\nu, \omega^s) > 0$ for some of the scenarios, say for $s \in \mathcal{S} \subseteq \{1,\dots,S\}$, then a dual extreme point $(u^s, v^s) \in \mathcal{D}^s$ with positive objective value is identified for each $s \in \mathcal{S}$ and the corresponding violated optimality cuts are added to the master. Set $\mathcal{E}^s = \mathcal{E}^s \cup \{(u^s, v^s)\}$ for $s \in \mathcal{S}$ and let $\nu = \nu + 1$; go to Step 2.

Remark 10. Recall that Algorithm 1 involved the solution of problem (6) for a sequence of increasing values of p. The cutting planes generated while solving the first of these problems remain valid when p is changed. Hence in Step 1 of Algorithm 2, we may let the sets $\mathcal{E}^s$, $s = 1, \ldots, S$ consist of the dual extreme points generated in previous runs of Algorithm 2 (or some subset thereof). This strategy of retaining cutting planes from previous runs resulted in remarkable time savings in the overall solution time for problem (4).

Remark 11. MirHassani et al. [10] considered a capacity expansion problem arising in supply chain network planning and solved the problem by a Benders decomposition approach, similar in many ways to the solution procedure described above. The authors observed that solutions to the master problem, in particular in early iterations, performed very poorly, since the master problem tends to minimize the amount of capacity installed, whereas "good" solutions in the second stage require substantial amounts of capacity to be installed. To circumvent this problem several enhancements of the master problem were considered. One such enhancement was to include some scenario in the master problem, thus making it more representative of the second-stage subproblems. Using this expanded formulation, MirHassani et al. observed a considerable improvement in overall solution time. In our setting, however, the expanded formulation performed very poorly. This is not too surprising, since the expanded master problem in this case is a capacitated network design problem with additional constraints, and several studies such as e.g. Bienstock et al. [1] and Riis and Andersen [13] have shown that projecting out the flow variables f is an efficient solution approach for such problems. Also, when the strategy of retaining cuts from previous runs is employed, the lack of consistency between the master problem and the second-stage subproblems is only significant in early iterations of the first run ($p = 0$) and hence does not outweigh the increased effort required to solve an expanded master problem.

Proposition 2. *Algorithm 2 terminates with an optimal solution in a finite number of iterations.*

Proof. First of all note that the optimal value of the master problem in any iteration is a lower bound on the optimal value of problem (6), since the master problem is a relaxation. Now, suppose that in some iteration ν for some scenario $s \in \{1, \ldots, S\}$ we have $0 = \theta_\nu^s < \psi(x_\nu, \omega^s) = 1$. In that case a violated cutting plane, cutting off the current solution (x_ν, θ_ν), is identified in Step 3 and the algorithm proceeds. Since the number of dual extreme points is finite, this can only happen a finite number of times and we will eventually have

$$\psi(x_\nu, \omega^s) \leq \theta_\nu^s, \quad s = 1, \ldots, S,$$

and hence

$$\sum_{s=1}^{S} \pi^s \psi(x_\nu, \omega^s) \leq \sum_{s=1}^{S} \pi^s \theta_\nu^s \leq p.$$

At this point the current solution x_ν is feasible in problem (6) and hence optimal. The corresponding criterion vector is $(z_1, z_2) = \big(z_\nu, \sum_{s=1}^{S} \pi^s \theta_\nu^s\big)$. □

3.3 Valid Inequalities for the p-Restricted Problems

The algorithm for the p-restricted problem (6), proposed in the previous section, solves a sequence of master problems which are all integer programming problems. Clearly, though, one should not put too much effort into finding optimal integer solutions for these problems in early iterations, since solutions are cut off anyway as more cutting planes are added. Hence, rather than solving the integer master problem (8) to optimality in each iteration, we chose to work with a relaxation of the problem and strengthen the formulation using valid inequalities.

From now on we shall refer by the linear relaxation of the master problem (8) to the corresponding problem in which integer requirements on the capacity variables have been relaxed. (Hence we speak of a linear relaxation, even though the artificial variables θ are still restricted to binaries.) Starting from this relaxation, we add cutting planes defining the indicator functions as described in the previous section. These cutting planes, however, should not only be used to define the indicator functions but also as valid inequalities for the convex hull of feasible integer solutions. In particular, since the feasible region of the dual of problem (7) is a rational polyhedron, we may assume that the extreme points $(u, v) \in \mathcal{D}^s$ are integral — this can be achieved by scaling. Hence the cutting planes derived in the previous section may be strengthened by rounding. Applying this approach, we arrive at what we shall refer to as the strengthened linear relaxation of the master problem,

$$\begin{aligned}
\min\ & z = \sum_{\{i,j\}\in E} c_{ij} x_{ij} \\
\text{s.t.}\ & \sum_{s=1}^{S} \pi^s \theta^s \leq p, \\
& \sum_{\{i,j\}\in E^s} x_{ij} v_{ij} + M\theta^s \geq \Big\lceil \sum_{k\in K^s} \rho_k^s D_k^s u_k - \sum_{\{i,j\}\in E^s} C_{ij} v_{ij} \Big\rceil, \quad (u,v) \in \mathcal{E}^s,\ s = 1,\ldots,S, \\
& x \in \mathbb{R}_+^{|E|},\ \theta \in \mathbb{B}^{|S|}.
\end{aligned} \tag{9}$$

Clearly, solving problem (9) rather than problem (8) may not produce an integer solution at termination of Algorithm 2. Therefore, to obtain an optimal integer solution to problem (6), this approach should be combined with some branching procedure. This could be done simply by explicitly reintroducing the integer requirements on capacity variables in the master problem and proceeding with Algorithm 2. Alternatively, the solution procedure could be incorporated in a more extensive branch-and-cut approach similar to the one described in Riis and Andersen [13]. As the computational experiments presented in Section 4 indicate, though, the integrality gap is very small when the strengthened linear relaxation of the master problem is put to use. Hence we consider it a viable approach to restrict attention to the strengthened linear relaxation of the master problem and subsequently establish near-optimal integer solutions by some heuristic.

Remark 12. Once again we note that the cutting planes used in the strengthened linear relaxation of the master problem may be seen as a generalization of the class of integral metric inequalities which are valid inequalities for the capacitated network design problem, discussed in a stochastic setting by Riis and Andersen [13] and in a deterministic setting by e.g. Bienstock et al. [1], Bienstock and Günlük [2], and Günlük [6]. In a similar manner we may derive generalizations of other classes of valid inequalities discussed in these papers such as e.g. partition inequalities and mixed integer rounding inequalities.

4 Computational Experiments

The solution procedure for the bicriteria problem (4) described in the previous section was implemented in C++ using procedures from the callable library of CPLEX 6.6 to solve the linear subproblems (7) and (mixed-) integer master problems (8) or (9). A series of computational experiments was carried out using three different problem instances. In this section we give some brief implementational details, describe the three problem instances and finally we report results of some preliminary computational experiments.

4.1 Implementational Details

As briefly discussed in Section 3.3 we chose to relax the integer requirements on capacity variables in the master problem at the start of each run. Starting from this relaxation we proceeded with Algorithm 2 until no more cuts could be identified. If the current solution at this point was not integral we explicitly reintroduced the integer requirements on x in the master problem and proceeded with Algorithm 2 until no more cuts could be identified. Using this approach it turned out that in general only very few additional iterations upon reintroduction of the integer requirements on x were necessary before an optimal integer solution of the current p-restricted problem was achieved. Hence we did not find the effort of generating cuts during branching in a more extensive branch-and-cut approach worthwhile.

Cutting planes for the master problem were generated through the solution of subproblems (7). To obtain the generalized integral metric inequalities described in Section 3.3 we used a heuristic. This heuristic simply divides all coefficients of the cut by the smallest positive coefficient. If the resulting coefficients are integral, a generalized integral metric inequality is obtained by rounding up the right-hand side.

As pointed out in Section 3.2, we achieved a considerable reduction in overall solution time by keeping cuts from previous runs in the master problem when the value of p was updated. To control the size of the master problem, however, it was necessary to temporarily remove "old" cuts. A cut $ax + M\theta^s \geq b$ was considered to be inactive if the corresponding binary variable θ^s was equal to 0 in the current solution and the relative slack $(ax + M\theta^s - b)/b$ was larger than 10%. A cut which had been inactive for more than 10 iterations was temporarily removed from the master problem and stored in a cut-pool. The cut-pool on the other hand was searched at regular intervals, and any violated cuts were returned to the master problem. The definition of inactive cuts and the number of iterations to keep an inactive cut in the master problem were chosen somewhat arbitrarily, so as to keep the size of the master problem manageable, while limiting the number of movements in and out of the cut-pool.

Even though the additional number of iterations required upon reintroduction of the integer requirements on x in the master problem was small, the CPU time required for these additional iterations turned out to be substantial, at least for the larger instances. Hence, for some problems, it may not be practicable to search for an optimal integer solution in this fashion, and the need arises for good heuristics providing upper bounds on the optimal solution. We propose a simple heuristic based on sequential rounding. The heuristic starts from the optimal solution $\bar{x}$ of the strengthened linear relaxation. The index $\{i, j\}$, for which $c_{ij}(\lceil \bar{x}_{ij} \rceil - \bar{x}_{ij})$ is minimal, is identified and the constraint $x_{ij} \geq \lceil \bar{x}_{ij} \rceil$ is added to the problem. The heuristic proceeds by alternatingly solving the problem, checking for violated cutting planes, and rounding up variables until a feasible integer solution is obtained.

4.2 Problem Instances

The first problem instance is a real-life telecommunications network provided by Sonofon, Denmark. The network is a complete network on 7 nodes and hence has 21 edges. The remaining two problem instances are modified versions of two real-life instances previously studied in e.g. Bienstock and Günlük [2], Günlük [6], and Riis and Andersen [13]. In the original versions of these instances two different types of facilities (i.e. low-capacity and high-capacity) were available for installation. The cost exhibited a high degree of economies to scale and hence we chose to use only the high-capacity facilities for our experiments, in order to fit the instances into the framework of the present paper. The first of these instances is a network representing the Atlanta area, containing 15 nodes and 22 edges. The second instance is a denser network representing the New York area. This network contains 16 nodes and 49 edges and has no existing capacity on the edges. All three instances have fully dense traffic matrices.

For each network we performed a series of experiments with varying number of scenarios. We considered only one type of failure, namely failure of a single edge. Moreover, we randomly generated a number of outcomes of future point-to-point demands assuming some uncertainty in the overall demand level captured in a parameter μ as well as some regional (node dependent) fluctuations captured in parameters ρ_i $(i \in V)$. The demand between nodes i and j under scenario s was calculated as

$$D_{ij}^s = \mu^s \rho_i^s \rho_j^s D_{ij},$$

where D_{ij} is the expected demand between nodes i and j and the random parameters μ^s and ρ_i^s $(i \in V)$ are sampled from uniform distributions,

$$\begin{aligned} &\mu^s \sim U(0.9, 1.1), \\ &\rho_i^s \sim U(0.9, 1.1), \quad \forall\, i \in V. \end{aligned}$$

For all situations with no failure, the network was required to fulfill the capacity requirements for each point-to-point pair no matter the level of demand. Hence for a scenario s representing a situation with no failure, the binary variable was excluded from the cutting planes (i.e. set to zero) and $\rho_k^s = 1$ for all $k \in K^s$. Likewise, a scenario s corresponding to some failure situation was considered as violating whenever the capacity requirement was not fulfilled for some point-to-point pair, i.e. $\rho_k^s = 1$ for all $k \in K^s$.

Assuming that the probability of failure is equal for all edges in the network (which was the case at least for the Sonofon problem), we used uniform scenario probabilities for the failure situations $\pi^s = 1/S$ for $s = 1, \ldots, S$, where $S = |E| \cdot d$ for a problem instance with $|E|$ edges in the network and d possible values of future point-to-point demands. The assumption is justified in situations where typical failures mainly occur at the end-points of connections. Such failures include breakdowns of electronic equipment as well as human errors during configuration of switches. Since the backbone network is normally well-protected (e.g. carried along highways, railroads or high-voltage transmission lines), such failures are often more likely than damage to the actual connection. Situations with no failure were not treated as actual scenarios cf. the discussion above, and hence no probabilities were assigned to them. Note that using these scenario probabilities, the parameter p in problem (6) denotes the *conditional* probability of capacity requirements to be violated given that a failure occurs — or in other words, the fraction of failure situations for which the capacity requirements are violated.

4.3 Computational Results

The first series of experiments was conducted in order to examine the quality of the approximations provided by the strengthened linear relaxation and the upper bounding heuristic. The first run was performed on the Atlanta problem assuming that demand is deterministic ($d = 1$). All values of p ranging from 0 to 1 were considered in the computations. Figure 1 shows the optimal objective values resulting from the IP-formulation, the linear relaxation, the strengthened linear relaxation and the upper bounding heuristic, respectively.

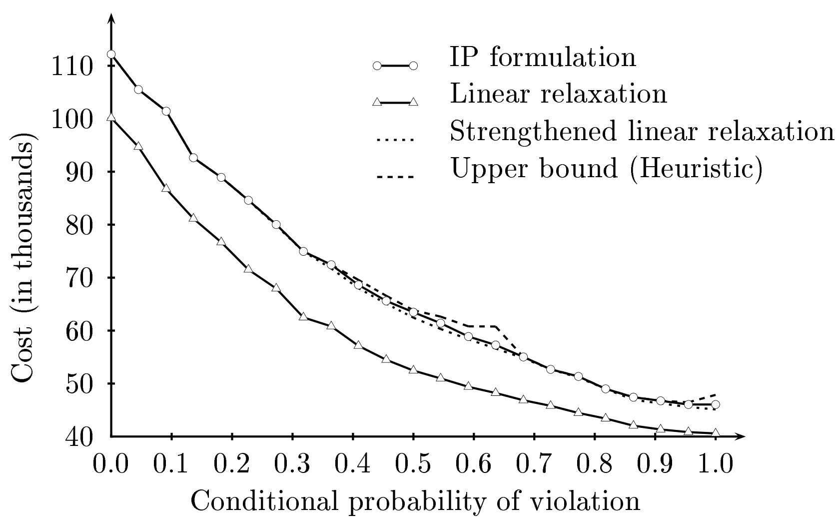

Figure 1: Atlanta problem ($d = 1$)

For this instance the integrality gaps were substantial, ranging from 10% to 20% for different values of p. Evidently, though, we see that the strengthened linear relaxation as well as the upper bounding heuristic performed extremely well. The integrality gap was closed by the strengthened linear relaxation for 8 of 26 problems, and the remaining gap was very small ($< 2\%$) in all other cases. Moreover, the upper bound was 5.7% off in the worst case, and the optimal integer solution was found by the heuristic for 14 different values of p.

The second run was performed on the Sonofon problem. Once again we considered all values of p ranging from 0 to 1 assuming that demand is deterministic. For this instance the integrality gaps were quite small and hence, in Figure 2, we plotted for each value of p the optimal objective value resulting from the linear relaxation (LR), the strengthened linear relaxation (SLR), and the upper bounding heuristic (UBH), *relative* to the objective value of the optimal integer solution.

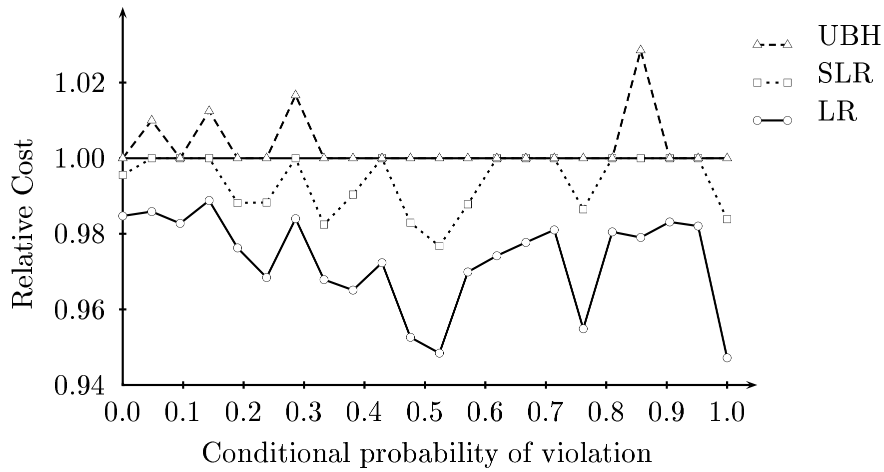

Figure 2: Sonofon problem ($d = 1$)

Once again we see that the strengthened linear relaxation as well as the upper bounding heuristic performed very well. In fact the integrality gap was closed by the strengthened linear relaxation for 12 of 22 problems and more than halved in most other cases, and the upper bounding heuristic found the optimal integer solution in all but four cases.

Finally, we performed one more run on the Sonofon problem, this time generating 5 possible outcomes of future demand. Considering values of p ranging from 0 to 0.25, we obtained the results shown in Figure 3.

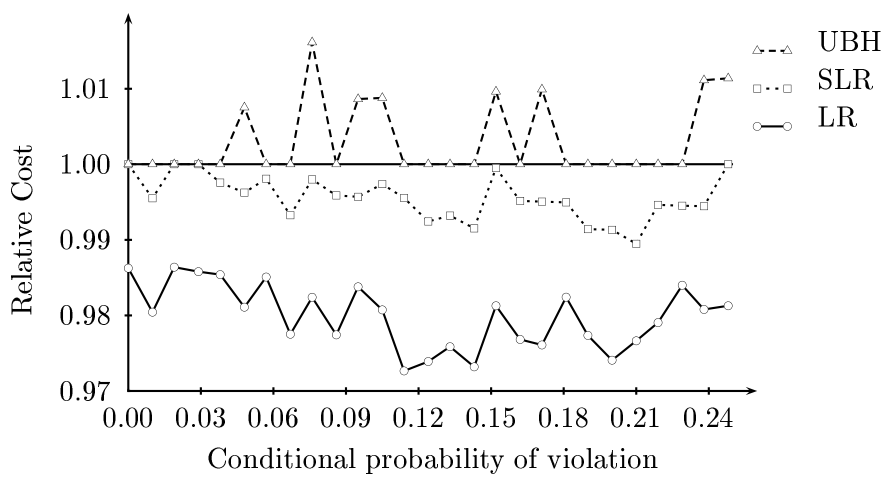

Figure 3: Sonofon problem ($d = 5$)

We see that the integrality gaps are somewhat smaller than what was seen for the deterministic problem. The integrality gap remaining from the strengthened linear relaxation was less than 1% for all different values of p and the optimal integer solution was found by the upper bounding heuristic for 19 of 27 problems.

The second series of experiments was conducted in order to test the practicability of the solution procedure. For each of the three instances we solved a series of problems with varying number of scenarios. In all cases the maximum acceptable level of p was set to 10%. Table 1 to Table 3 give the results. We report the CPU time required to solve the bicriteria stochastic programming problem using the IP-formulation as well as the strengthened linear relaxation. Also, the number of p-restricted problems to be solved during computation is reported. For illustration we also report these figures when only values of p less than or equal to 5% are considered. All CPU times are reported as minutes:seconds. Computations were stopped after three hours of CPU time and in this case the last value of p for which the p-restricted problem was being solved is given in brackets. Note once again that situations with no failure are not counted as actual scenarios since they do not give rise to any binary variables in the master problem, but still subproblems have to be solved to generate cuts ensuring feasibility for these situations.

Table 1: Sonofon problems

Number of scenarios	Maximum value of p	Number of problems	CPU time IP formulation	CPU time Relaxation
$21 \cdot 1$	0.05	2	0:01	0:01
	0.10	3	0:01	0:01
$21 \cdot 5$	0.05	6	0:11	0:04
	0.10	11	1:03	0:19
$21 \cdot 10$	0.05	11	2:35	0:58
	0.10	22	92:36	25:03

Table 2: Atlanta problems

Number of scenarios	Maximum value of p	Number of problems	CPU time IP formulation	CPU time Relaxation
$22 \cdot 1$	0.05	2	0:07	0:06
	0.10	3	0:10	0:09
$22 \cdot 5$	0.05	6	0:53	0:41
	0.10	12	5:25	1:52
$22 \cdot 10$	0.05	12	25:33	5:44
	0.10	23	180:00 (0.077)	107:55

Table 3: New York problems

Number of scenarios	Maximum value of p	Number of problems	CPU time IP formulation	CPU time Relaxation
$49 \cdot 1$	0.05	3	135:37	6:02
	0.10	5	180:00 (0.061)	12:25
$49 \cdot 5$	0.05	13	180:00 (0.000)	180:00 (0.037)
	0.10	25	– = –	– = –
$49 \cdot 10$	0.05	25	180:00 (0.000)	180:00 (0.010)
	0.10	50	– = –	– = –

As expected we see that computation time increases drastically with the number of scenarios as well as the size of the network. The increase in CPU time, when a larger number of possible outcomes of future demand is generated, is partly explained by the fact that a larger number of second-stage multicommodity flow problems have to be solved in each iteration. More important, however, was the increased effort required to solve the master problems as the number of binary variables increase. When a network containing a larger number of edges is considered, not only the number of scenarios (and hence the number of second-stage problems and the number of binary variables in the master problem) increases, but also the number of first-stage variables. Hence the problem complexity is heavily dependent on the number of edges in the network, as illustrated by the difference in CPU time for the New York problem compared to the two smaller instances. Finally, we observed that the p-restricted problems were increasingly difficult to solve, as the value of p was increased. Hence for the New York network the algorithm was only practicable for the case of deterministic demand, unless only very small values of p were considered.

5 Conclusions

The capacity expansion model considered in this paper incorporates uncertainty concerning future demand as well as potential future failures of network components. The problem was formulated as a bicriteria stochastic program in which the total cost as well as the probability of capacity requirements to be violated are minimized. We have proposed a simple reduced feasible region method (Algorithm 1) which was shown to determine all non-dominated solutions of the problem (Proposition 1) by solving a finite number of so-called p-restricted problems. The p-restricted problems determine the minimum cost such that at most a fraction of p of failure situations result in violations of the capacity requirements. To solve the p-restricted problems we elaborated a cutting-plane procedure (Algorithm 2), and it was shown that the algorithm terminates in a finite number of iterations (Proposition 2). The solution procedure has been succesfully implemented and our preliminary computational experiments indicate that the method is certainly practicable, at least for moderate size networks or when only small values of p are allowed.

References

[1] D. Bienstock, S. Chopra, O. Günlük, and C.-Y. Tsai. Minimum cost capacity installation for multicommodity network flows. *Mathematical Programming*, 81:177–200, 1998.

[2] D. Bienstock and O. Günlük. Capacitated network design. Polyhedral structure and computation. *INFORMS Journal on Computing*, 8(3):243–259, 1996.

[3] J.R. Birge and F.V. Louveaux. *Introduction to Stochastic Programming.* Springer-Verlag, New York, 1997.

[4] G. Dahl and M. Stoer. A cutting plane algorithm for multicommodity survivable network design problems. *INFORMS Journal on Computing*, 10(1):1–11, 1998.

[5] M.A.H. Dempster, E.A. Medova, and R.T. Thompson. A stochastic programming approach to network planning. *Teletraffic Contributions for the Information Age. Proceedings of the 15th International Teletraffic Congress - ITC 15*, 1:329–339, 1997.

[6] O. Günlük. A branch-and-cut algorithm for capacitated network design. *Mathematical Programming*, 86:17–39, 1999.

[7] M. Iri. On an extension of the max-flow min-cut theorem to multicommodity flows. *Journal of the Operations Research Society of Japan*, 13:129–135, 1971.

[8] P. Kall and S.W. Wallace. *Stochastic Programming.* John Wiley and Sons, Chichester, UK, 1994.

[9] E.A. Medova. Chance-constrained stochastic programming for integrated services network management. *Annals of Operations Research*, 81:213–229, 1998.

[10] S.A. MirHassani, C. Lucas, G. Mitra, E. Messina, and C.A. Poojari. Computational solution of capacity planning models under uncertainty. *Parallel Computing*, 26:511–538, 2000.

[11] K. Onaga and O. Kakusho. On feasibility conditions of multicommodity flows in networks. *IEEE Transactions in Circuit Theory CT-18*, 4:425–429, 1971.

[12] A. Prékopa. *Stochastic Programming.* Kluwer Academic Publishers, Dordrecht, Netherlands, 1995.

[13] M. Riis and K.A. Andersen. Capacitated network design with uncertain demand. To appear in *INFORMS Journal on Computing*, 2000.

[14] M. Riis and R. Schultz. Applying the minimum risk criterion in stochastic recourse programs. Working Paper 2001/3, University of Aarhus, Department of Operations Research, 2001. Available at http://home.imf.au.dk/riis/.

[15] M. Riis, A.J.V. Skriver, and J. Lodahl. Network planning in telecommunications: A stochastic programming approach. Working Paper 2001/2, University of Aarhus, Department of Operations Research, 2001. Available at http://home.imf.au.dk/riis/.

[16] M. Rios, V. Marianov, and M. Gutierrez. Survivable capacitated network design problem: new formulation and Lagrangean relaxation. *Journal of the Operational Research Society*, 51:574–582, 2000.

[17] S. Sen, R.D. Doverspike, and S. Cosares. Network planning with random demand. *Telecommunication Systems*, 3:11–30, 1994.

[18] R.E. Steuer. *Multiple criteria optimization: Theory, computation and application.* Wiley, New York, 1986.