THE REPORT COMMITTEE FOR SIFENG LIN
CERTIFIES THAT THIS IS THE APPROVED VERSION OF THE FOLLOWING REPORT:

Benders Decomposition and an IP-Based Heuristic for Selecting IMRT Treatment Beam Angles

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

Jonathan F. Bard

Anant Balakrishnan

Benders Decomposition and an IP-based Heuristic for Selecting IMRT Treatment Beam Angles

by

SIFENG LIN, B.E.

REPORT

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN ENGINEERING

THE UNIVERSITY OF TEXAS AT AUSTIN

DECEMBER 2014

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Jonathan Bard for his support and guidance in carrying out this report. Also, I also wish to thank my second reader Prof. Anant Balakrishnan for his suggestions. Special thanks to Guven Kaya and Prof. Gino Lim for their help with data sets and insightful discussions.

Abstract

Benders Decomposition and an IP-based Heuristic for Selecting IMRT Treatment Beam Angles

Sifeng Lin, M.S.E.

The University of Texas at Austin, 2014

Supervisor: Jonathan F. Bard

To optimize the beam angle and fluence map in Intensity Modulated Radiation Therapy (IMRT) planning, we apply Benders decomposition as well as develop a two-stage integer programming-based heuristic. Benders decomposition is first implemented in the traditional manner by iteratively solving the restricted master problem, and then identifying and adding the violated Benders cut. We also implemented Benders decomposition using the "lazy constraint" feature included in CPLEX. In contrast, our two-stage heuristic first seeks to find a good solution by iteratively eliminating the least used angles in the linear programming relaxation solution until the size of the formulation is manageable. In the second stage of the heuristic, the solution is improved by applying local branching. The various methods were tested on real patient data in order to investigate their effectiveness and runtime characteristics. The results indicated that implementing Benders using the lazy constraint usually led to better feasible solutions than the traditional approach. Moreover, the LP rounding heuristic was seen to generate high-quality solutions within a short amount of time, with further improvement obtained with the local branching search.

Table of Contents

List of Tables vii

List of Figures viii

1. Introduction 1

2. Literature Review 3

3. Problem Formulation 5

4. Benders Decomposition 9

    4.1 Benders reformulation 10

    4.2 Implementation of Benders decomposition 12

5. Heuristics 17

    5.1. LP rounding heuristic 17

    5.2. Local branching 17

6. Computational Results 21

7. Summary and Conclusions 27

8. Reference 28

List of Tables

Table 1. Data set summary.....21

Table 2. Solution time for 12-angle problems 22

Table 3. Comparing different solution methods 23

Table 4. Comparing MIP formulation of instances P1-36, P2-36, and P3-36 23

Table 5. Comparison between traditional and modern Benders decomposition ...24

List of Figures

Figure 1. Procedure for traditional Benders decomposition 14

Figure 2. Flowchart for updating the incumbent solution in modern Benders decomposition 15

Figure 3. LP rounding heuristic 17

Figure 4. Procedure for local branching 20

Figure 5. Dose volume histogram for LP-rounding heuristic solutions.....26

1. Introduction

Intensity modulated radiation therapy (IMRT) has gained popularity among oncologists because of its demonstrated capability to deliver higher doses of radiation to tumors while doing a better job at sparing healthy tissue than other forms of radiation treatments. Lim et al. (2012) summarize the general treatment procedure as follows: the patient lies still on a special couch and is irradiated by photon beams generated by a linear accelerator (LINAC). The movable arm of the LINAC, called a gantry, can rotate in a plan perpendicular to the couch to deliver radiation to the desired location of the body. For any given angle, the computer controlled multi-leaf collimator in the LINAC can adjust the radiation beams to match the shape of the tumor. An open radiation field is fractionated into hundreds of subfields called beamlets. Each beamlet is assigned its own level of intensity, and the set of beamlets carried in each beam angle is referred to as a fluence map. For more details about the procedure and equipment, see Lim et al. (2008).

Before designing an IMRT plan, physicians use various scanning procedures to capture the geometry of the tumor area. The area is classified into two types of volumes: planning target volume (PTV) and organs-at-risk (OARs). The PTV represents the area of the cancer, and OARs are healthy organs or tissues of the body. An IMRT plan should simultaneously deliver the desired amount of radiation to the tumor while limiting the amount to the healthy organs.

PTV and the involved OARs are divided into three-dimensional treatment cubes called voxels (Lim et al. 2007). The total dose that each voxel receives is defined by the weighted sum of the beamlet dose delivered to the voxel. In IMRT plans, the regions of high dose and low dose are called hot spots and low spots, respectively. For voxels in PTV, we want to control both the hot and cold spots to guarantee the desired treatment effect. For voxels in OARs, we only control hot spots to spare the healthy organs. Both hot and cold spot control can be modeled by enforcing hard constraints or penalizing deviations from the desired dose.

The purpose of this study is to explore solution techniques that can help planners optimize the angle selection and fluence intensity maps. Various mixed-integer programming (MIP) formulations have been proposed to solve this problem (e.g., Lee et al. 2000, Lim et al. 2007, and Yarmand et al. 2013). Because solving the MIPs with commercial software has proven difficult, if not impossible, researchers have proposed various heuristics. Decomposition methods such as Benders decomposition and Lagrangian relaxation, which have been successfully applied to various MIP models, have not been applied to IMRT planning to the best of our knowledge. In this study, we show that Benders decomposition may find better solutions than a standard MIP solver. In addition, we develop a two-stage heuristic that uses a standard MIP solver (i) to construct a good initial solution in the first stage and (ii) to implement local branching (Fischetti and Lodi 2003) in the second stage. The first stage of the heuristic reduces the solution space of the original problem by iteratively eliminating unpromising angles identified in the linear programming solution until the remaining problem is easy to solve. In the second stage, the solution is used as a starting point for local branching. Our results show that the LP rounding heuristic is fast and can generate good solutions, which may be further improved by local branching in the second stage.

The rest of the report is organized as follows. Section 2 summarizes the related literature. Section 3 introduces the problem formulation and Section 4 discusses the Benders decomposition. In Section 5, we develop an LP rounding heuristic and couple it with local branching to improve the solution. Section 6 presents the computational results and Section 7 concludes the report.

2. Literature Review

IMRT is a popular technique for irradiating tumors. Its application and efficacy have been extensively discussed by the medical research community (e.g., Milano et al. 2006 and Zelefsky et al. 2000). The comparison of IMRT with another popular therapy, proton therapy, is also well studied (e.g., Chang et al. 2005 and Mock et al. 2004).

An IMRT plan entails both the angle at which the dose is delivered and the intensity of the fluence map for the angles selected. Accordingly, researchers and practitioners face two interrelated problems: the angle optimization problem and the fluence intensity optimization problem. Since the 1970s, researchers have developed various methods to solve the latter. Depending on the penalty function for the undesired dose delivered to voxels in PTV or OARs, the resulting model may be a linear or nonlinear (usually quadratic) mathematical program. Ehrgott et al. (2008) present an extensive survey. When the objective function is linear or can be linearized, the program can be solved with standard commercial software (Lim et al. 2007, Romeijn et al. 2003, Saka et al. 2011, Saka et al. 2014). Otherwise, researchers have developed their own the problem-specific algorithms. For example, Spirou and Chui (1998) developed a gradient inverse planning algorithm to determine the intensity-modulated beams.

When the penalty function is linear, the angle selection and fluence intensity map optimization problems can be modeled as a MIP. Lee et al. (2000) proposed a MIP formulation for generating LINAC radiosurgery treatment plans. Yarmand et al. (2013) first solved a linear program to find an ideal, albeit infeasible, plan in which all candidate beams can be used. They then developed a MIP to find the plan with the desired quality while minimizing the number of beams. Lim et al. (2007) proposed an optimization framework to automatically design radiation treatment plans. Their work includes (1) the determination of the beam's eye view for each potential angle, (2) the generation of the corresponding dose matrices for each beam from each angle, (3) the development of three models to optimize the beam angles, beam weights, and wedge orientations, respectively, (4) techniques to solve the optimization models, and (5) techniques to

control the dose-volume histograms associated with the OARs. Unable to solve the MIPs exactly, they investigated different ways to both reduce the model size and to obtain high quality feasible solutions.

To reduce the number of potential beams, Yarmand et al. (2013) added neighborhood cuts to their MIP so that the selection contains at most one or a few of the beams in every set of adjacent beams. Lim et al. (2008) developed an LP-based iterative beam angle elimination algorithm to obtain promising beam angles with optimized fluence maps. Lim et al. (2014) examined the strengths and weaknesses of six optimization methods for selecting beam angles: branch and bound, simulated annealing (SA), genetic algorithms (GA), nested partitions (NP), branch and prune (BP), and local neighborhood search (LNS). They concluded that it is more effective to apply hybrid methods that first find a good feasible solution using SA, GA, NP, or BP, and then use the resulting solution as a starting point for LNS to arrive at a local optimum.

3. Problem Formulation

Devising an IMRT plan involves selecting a subset of angles and designing the associated fluence map to apply the desired dose to the planning target volume without damaging the healthy organs. Accordingly, we need to penalize both hot and cold spots for voxels in PTV and penalize the hot spots for voxels in OARs. The following notation is used in the developments.

Indices and sets

A	set of candidate beam angles; $a \in A$
O	set of OARs; $i \in O$
T	set of voxels in PTV; $v \in T$
S_i	set of voxels in organ i; $i \in O$
V	set of all voxels; $v \in T$
B_a	set of beamlets in angle a; $b \in B_a$

Parameters

η	maximum number of treatment angles in a treatment plan
d_{vb}	dose contribution to voxel v from beamlet b
U_v	upper bound on the dose applied to voxel $v \in T$
L_v	lower bound on the dose applied to voxel $v \in T$
θ_U	hot spot control parameter on voxels in PTV
θ_L	cold spot control parameter on voxels in PTV
ϕ	hot spot control parameter on voxels in OARs
λ_t^+	penalty coefficient for hot spots in PTV
λ_t^-	penalty coefficient for cold spots in PTV
λ_s	penalty coefficient for hot spots in OARs
M_{ab}	maximum intensity of beam $b \in B_a$

Decision variables

ψ_a 1 if angle a is selected, 0 otherwise

w_{ab} intensity of beamlet $b \in B_a$ for angle a

D_v total dose applied to voxel $v \in V$

With slight abuse of notation, we use D_T to denote the vector of dose values for voxels in PTV, D_{S_i} for the vector of dose values applied to organ i, and D for the vector of dose values for all voxels. Now, given the total dose D_v applied to each voxel $v \in V$, Lim et al. (2007) use the following penalty function

$$f(D) = \lambda_t^+ \left\| \left(D_T - \theta_U e^T \right)^+ \right\|^\infty + \lambda_t^+ \left\| \left(\theta_L e^T - D_T \right)^+ \right\|^\infty + \sum_{i \in O} \lambda_s \left\| \left(D_{S_i} - \phi e^{S_i} \right)^+ \right\|^1 / |S_i|,$$

where $\| \mathbf{x} \|^\infty$ and $\|\mathbf{x}\|^1$ respectively denote the infinity and 1-norm of vector $\mathbf{x}$, $(y)^+$ equals $\max\{y, 0\}$, and e^T and e^{S_i} are vectors of 1's with dimensions $|T|$ and $|S_i|$, respectively. The first and second terms in $f(D)$ control the hot and cold spots for voxels in PTV by penalizing the maximum excess dose and maximum shortage of dose, respectively; the third term controls the hot spot for voxels in OARs by penalizing the dose that is more than ϕ. The full model is:

$$\text{Minimize} \quad f(D) \tag{1a}$$

$$\text{subject to} \quad D_v = \sum_{a \in A} \sum_{b \in B_a} d_{vab} w_{ab} \qquad \forall v \in V \tag{1b}$$

$$\sum_{a \in A} \psi_a \le \eta \tag{1c}$$

$$0 \le w_{ab} \le M_{ab} \psi_a \qquad \forall a \in A, b \in B_a \tag{1d}$$

$$L_v \le D_v \le U_v \qquad \forall v \in T \tag{1e}$$

$$\psi_a \in \{0,1\} \qquad \forall a \in A \tag{1f}$$

The objective function (1a) minimizes the total penalty from the radiation doses applied to both PTV and the neighboring organs. Equalities (1b) specify the total dose each voxel receives, which is the weighted sum of the individual doses. Constraints (1c) limit the number of angles that can be used to the maximum specified. Constraints (1d)

guarantee the non-negativity of the beamlet intensity and ensure that a beamlet can only carry a positive dose if the angle is selected. Refer to Lim et al. (2007) for more detail on how to refine the value of M_{ab}, the maximum intensity of beamlet $b \in B_a$. Finally, bounds are placed on the dose applied to PTV in (1e) and the angle selection variable, ψ_a, is defined to be binary in (1f).

To solve the problem as an integer (linear) programming, we need to linearize the objective function. Define $y_v = \left\| \left(D_{S_i} - \phi e^T \right)^+ \right\|^1$ for each $v \in S_i$, $z^+ = \left\| \left(D_T - \theta_U e^T \right)^+ \right\|^\infty$, and $z^- = \left\| \left(\theta_L e^T - D_T \right)^+ \right\|^\infty$. Model (1) can then be reformulated as follows.

Minimize $\lambda_t^+ z^+ + \lambda_t^- z^- + \sum_{i \in O} \sum_{v \in S_i} \lambda_v y_v / |S|$ (2a)

subject to $z^+ \geq D_v - \theta_U \quad \forall v \in T$ (2b)

$z^- \geq \theta_L - D_v \quad \forall v \in T$ (2c)

$y_v \geq \sum_{a \in A} \sum_{b \in B_a} d_{vb} \omega_{ab} - \phi \quad \forall v \in S_i, i \in O$ (2d)

$L_v \leq D_v \leq U_v \quad \forall v \in T$ (2e)

$D_v = \sum_{a \in A} \sum_{b \in B_a} d_{vb} \omega_{ab} \quad \forall v \in T$ (2f)

$\sum_{a \in A} \psi_a \leq \eta$ (2g)

$0 \leq \omega_{ab} \leq M_{ab} \psi_a \quad \forall a \in A, b \in B_a$ (2h)

$\psi_a \in \{0,1\} \quad \forall a \in A$ (2i)

$y_v \geq 0 \quad \forall v \in S_i, i \in O$ (2j)

$z^+, z^- \geq 0$ (2k)

Because we minimize the objective function, (2a) is equivalent to (1a) when constraints (2b), (2c), and (2d) are enforced. Note that we do not remove variables D_v

for $v \in T$ since keeping them improves computational performance. Although the number of variables increases slightly, the constraint matrix is sparser.

We can reduce the number of variables and constraints in model (2) in the following ways.

- In the third term in (2a), we only penalize the excess dose applied to the voxels in OAR, i.e., the dose beyond ϕ. Thus, if some beamlet $b \in B_a$ does not affect the voxels in PTV, i.e., $d_{vb} = 0$ for all $v \in T$, we must have an optimal solution with $w_{ab} = 0$. Making $w_{ab} > 0$ only increases the penalty associated with the voxels in OAR. Therefore, we can exclude these variables from model (2).
- For any two voxels $v_1, v_2 \in T$, if $d_{v_1 b} \geq d_{v_2 b}$ for all $b \in B_a$ and $a \in A$, we have $D_{v_1} - \theta_U \geq D_{v_2} - \theta_U$ and $\theta_L - D_{v_1} \leq \theta_L - D_{v_2}$. Therefore, constraints $z^+ \geq D_{v_1} - \theta_U$ and $z^- \geq \theta_L - D_{v_2}$ are both redundant and can be removed.
- If $\sum_{a \in A} \sum_{b \in B_a} d_{vb} M_{ab} - \phi \leq 0$ for some $v \in S_i, i \in O$, constraint $y_v \geq \sum_{a \in A} \sum_{b \in B_a} d_{vb} w_{ab} - \phi$ is redundant. Accordingly, we can exclude both the constraint and variable y_v from the model.

4. Benders Decomposition

Benders decomposition is an algorithm for solving MIPs that has been widely applied since 1960s. It is best suited for models of the form: $\min\{cx + dy : Ax + By \geq b, x \in \Re_+^n, y \in \mathbb{Z}_+^p\}$, where p is relatively small and when fixed, the constraints $Ax \geq b - By$ divide into independent subsets in the x variables. An integer master problem is set up that is equivalent to the original problem when all its constraints are included. None of those constraints are known at the outset, though, so they are generated iteratively one or two at a time by solving the dual of the LP subproblems that result when y is fixed. Each subproblem provides an *optimality cut* and perhaps a *feasibility cut* which are added to the *restricted* master problem. Convergence is finite but may require many iterations. Cordeau et al. (2001) applied Benders to simultaneously solve the aircraft routing and crew scheduling problems, while Binato et al. (2001) used it to solve power transmission network design problems. Costa (2005) gives an extensive survey on applications to fixed-charge network design problems. In this section, we apply Benders decomposition to model (2).

4.1 Benders reformulation

For fixed values of ψ_a for all $a \in A$, and after a few substitutions, (2a) – (2k) reduces to an LP whose constraints are given below. Their corresponding dual variables are defined on the right.

$$z^+ - \sum_{a\in A}\sum_{b\in B_a} d_{vb}\omega_{ab} \geq -\theta_U \qquad \forall v \in T \qquad \mu_v^+$$

$$z^- + \sum_{a\in A}\sum_{b\in B_a} d_{vb}\omega_{ab} \geq \theta_L \qquad \forall v \in T \qquad \mu_v^-$$

$$y_v - \sum_{a\in A}\sum_{b\in B_a} d_{vb}\omega_{ab} \geq -\phi \qquad \forall v \in S_i, i \in O \qquad \mu_v$$

$$\sum_{a\in A}\sum_{b\in B_a} d_{vb}\omega_{ab} \geq L_v \qquad \forall v \in T \qquad \sigma_v^-$$

$$-\sum_{a\in A}\sum_{b\in B_a} d_{vb}\omega_{ab} \geq -U_v \qquad \forall v \in T \qquad \sigma_v^+$$

$$-\omega_{ab} \geq -M_{ab}\psi_a \qquad \forall a \in A, b \in B_a \qquad \pi_{ab}$$

Model (3) is the corresponding dual formulation of the LP after fixing the angle selection vector ψ.

$$h(\psi) = \text{Maximize} \sum_{v\in T}\left(-\theta_U\mu_v^+ + \theta_L\mu_v^- + L_v\sigma_v^- - U_v\sigma_v^+\right) - \sum_{i\in O}\sum_{v\in S_i}\phi\mu_v - \sum_{a\in A}\sum_{b\in B_a} M_{ab}\psi_a\pi_{ab} \quad (3a)$$

subject to

$$\sum_{v\in T}\mu_v^+ \leq \lambda_t^+ \quad (3b)$$

$$\sum_{v\in T}\mu_v^- \leq \lambda_t^- \quad (3c)$$

$$0 \leq \mu_v \leq \lambda_S / |S_i| \qquad \forall i \in O, v \in S_i \quad (3d)$$

$$-\pi_{ab} + \sum_{v\in T} d_{va}\left(-\sigma_v^+ + \sigma_v^- - \mu_v^+ + \mu_v^-\right) - \sum_{i\in O}\sum_{v\in S_i} d_{vb}\mu_v \leq 0 \quad \forall a \in A, b \in B_a \quad (3f)$$

$$\mu_v^+, \mu_v^- \geq 0 \qquad \forall v \in T \quad (3g)$$

$$\pi_{ab} \geq 0 \qquad \forall a \in A, b \in B_a \quad (3h)$$

Let Q be the set of extreme points for the feasible region (3b) – (3h). For any extreme point $q \in Q$, with $q = (\mu_v^+, \mu_v^-, \sigma_v^-, \sigma_v^+, \mu_v, \pi_{ab})$, we define the linear function $g(\psi, q)$ as

$$g(\psi, q) = \sum_{v \in T} \left(-\theta_U \mu_v^+ + \theta_L \mu_v^- + L_v \sigma_v^- - U_v \sigma_v^+ \right) - \sum_{i \in O} \sum_{v \in S_i} \phi \mu_v - \sum_{a \in A} \sum_{b \in B_a} M_{ab} \psi_a \pi_{ab} .$$

Since $g(\psi, q)$ is the objective function of model (3), we can then denote model (3) as $g(\psi) = \text{Max}\{g(\psi, q): q \in Q\}$.

Lemma 1: Let q^1 and q^2 be two solutions to model (3) and assume that all their components are the same except π_{ab}. Let π_{ab}^1 and π_{ab}^2 be the corresponding components of q^1 and q^2, respectively. If $\pi_{ab}^1 \geq \pi_{ab}^2$, then constraint $g(\psi, q^2) \leq W$ dominates $g(\psi, q^1) \leq W$.

Proof: Since $\pi_{ab}^1 \geq \pi_{ab}^2$, we have $g(\psi, q^1) \leq g(\psi, q^2)$, which proves the result. ■

Lemma 1 indicates that we would like to have a value of π_{ab} that is as small as possible to get a stronger Benders cut. When $\psi_a > 0$, optimality of (3) ensures that the corresponding constraints (3f) are binding for all $b \in B_a$ (otherwise, we can decrease the objective function and maintain feasibility by decreasing π_{ab}), which means that we cannot decrease π_{ab} without affecting the feasibility of the solution. When $\psi_a = 0$, the optimal solution may have $\pi_{ab} > \sum_{v \in T} d_{va} \left(-\sigma_v^+ + \sigma_v^- - \mu_v^+ + \mu_v^- \right) - \sum_{i \in O} \sum_{v \in S_i} d_{vb} \mu_v$. In this case, we can state the value of π_{ab} as follows:

$$\pi_{ab} = \max\{0, \sum_{v \in T} d_{vb} \left(-\sigma_v^+ + \sigma_v^- - \mu_v^+ + \mu_v^- \right) - \sum_{i \in O} \sum_{v \in S_i} d_{vb} \mu_v \}$$

Benders optimality cuts can be slightly strengthened when a lower bound for the original problem is known. Given any extreme point $q \in Q$, where $q = (\mu_v^+, \mu_v^-, \sigma_v^-, \sigma_v^+, \mu_v, \pi_{ab})$, let $c_a = \sum_{b \in B_a} M_{ab} \pi_{ab}$ and $b = \sum_{v \in T} (-\theta_U \mu_v^+ + \theta_L \mu_v^- + L_v \sigma_v^- - U_v \sigma_v^+) - \sum_{i \in O} \sum_{v \in S_i} \phi \mu_v$ for all $a \in A$. The corresponding Benders feasibility cut is $W + \sum_{a \in A} c_a \psi_a \geq b$. If W_1 is a lower bound on

the optimal objective function W^*, i.e., $W^* \geq W_1$, then we can replace coefficient c_a with $c_a^1 = \min\{c_a, b - W_1\}$.

Let Ψ denote the set of angle combinations that give an unbounded objective function value in model (3), i.e., $h(\bar{\psi}) = \infty$ for all $\bar{\psi} \in \Psi$. We can add the feasibility constraint $\sum_{a \in A} \bar{\psi}_a \psi_a \leq \eta - 1$ to prevent solution $\bar{\psi}$ from being selected. The original problem can be reformulated as:

$$\text{Minimize} \quad W \tag{4a}$$

$$\text{subject to} \quad g(\psi, q) \leq W \qquad \forall q \in Q \tag{4b}$$

$$\sum_{a \in A} \bar{\psi}_a \psi_a \leq \eta - 1 \qquad \forall \bar{\psi} \in \Psi \tag{4c}$$

$$\sum_{a \in A} \psi_a \leq \eta \tag{4d}$$

$$\psi_a \in \{0,1\} \qquad \forall a \in A \tag{4e}$$

4.2 Implementation of Benders decomposition

We start the algorithm with a restricted master problem and add constraints on the fly when they are indicated. The initial restricted master problem is as follows and has an initial solution $W = -\infty$.

$$\begin{aligned} &\text{Minimize} \quad W \\ &\text{subject to} \quad \sum_{a \in A} \psi_a \leq \eta \\ &\qquad\qquad \psi_a \in \{0,1\} \qquad \forall a \in A \end{aligned} \tag{5}$$

Adding a Benders cut to (5) has traditionally meant restarting the IP solver from scratch, which is computationally expensive. To improve performance, it is advantageous to start with a set of promising Benders cuts in the restricted master problem. McDaniel and Devine (1977) introduced the idea of relaxing the integrality requirements in the master problem and generating cuts from the fractional solution. Since these cuts are also defined by feasible solutions of model (3), they are valid for

model (4). Thus, we adopt a two-phase Benders decomposition: in the first phase, we relax the integrality constraint on the ψ_a variables in model (5) and apply Benders decomposition until its objective function is within 5% of the LP relaxation value of model (2); in the second phase, we start with the Benders cuts found in the first phase and continue in the traditional manner. Figure 1 describes the procedure of implementing the Benders decomposition in this traditional way. Step 1 of the procedure initializes the set Q^* and Ψ^*, Step 2 corresponds to the first phrase, and Steps 3 and 4 correspond to the second phrase.

Rubin (2011) proposed a more efficient approach to implementing Benders decomposition. Instead of starting branch and bound from scratch after each cut is added, the *modern* approach adds the cuts as "lazy" constraints in the MIP solver (e.g., CPLEX). Lazy constraints are a set of inequalities specified by the user that are required to define the feasible region of the model but are not part of the model when the solver is initiated. Instead, they are only checked when an integer feasible solution is identified, and any of those constraints that turn out to be violated are then included in the model currently being solved. Note that branch and bound is not restarted when violated lazy constraints are added. More discussion can be found in CPLEX (2011).

Essentially, the presence of lazy constraints requires a modification of the incumbent update procedure in branch and bound. At each node of the traditional search tree, an LP subproblem (note, this is not the Benders LP subproblem) is solved and one or more heuristics are typically applied to convert the fractional solution to a feasible (integer) solution. If a better feasible solution results, then the incumbent, i.e., the best feasible solution found so far, is updated. With lazy constraints, the solver must make sure that any candidate feasible solution satisfies all the lazy constraints before updating the incumbent. If there are no violations then the incumbent is updated. Otherwise, the solver adds the violated lazy constraints to the model being solved and does not update the incumbent. This logic ensures that branch and bound finds the optimal solution to the original model while only enforcing lazy constraints when violations are detected.

Procedure_traditional_Benders_decomposition

Step 1: Set of extreme points $Q^* = \varnothing$

Set of infeasible angle profiles $\Psi^* = \varnothing$

Step 2: Solve the LP relaxation of model (3) and denote the optimal objective function value as W^{LP}

Do

Solve $W^* = \min\{W:\ g(\psi, q) \le W,\ \forall q \in Q^*,\ (4d),\ 0 \le \psi_a \le 1,\ \forall\ a \in A\}$ and denote the optimal solution as ψ^*

Solve model (3) to get $h(\psi^*)$ and the corresponding solution q^*.

Put $\Psi^* \leftarrow \Psi^* \cup \{\psi^*\}$

While $W^* < 0.95\ W^{\mathrm{LP}}$

Step 3: Solve the restricted master problem

$W^* = \min\{W:\ g(\psi, q) \le W,\ \forall\ q \in Q^*,\ \sum_{a \in A} \bar{\psi}_a \psi_a \le \eta - 1,\ \forall\ \bar{\psi} \in \Psi^*,\ (4d),\ (4e)\}$

If problem is infeasible, then

terminate, the original problem is infeasible.

Else

Let the solution be ψ^*

Go to Step 4.

Step 4: Solve model (3) to get $h(\psi^*)$ and the corresponding solution q^*.

If $h(\psi^*) = \infty$

Put $\Psi^* \leftarrow \Psi^* \cup \{\psi^*\}$

Go to Step 3.

Else if $h(\psi^*) = W^*$

The optimal angle profile is ψ^* and terminate.

Else

Put $Q^* \leftarrow Q^* \cup \{q^*\}$

Go to Step 3.

Figure 1. Traditional Benders decomposition

We also adopt the two-phase approach in our implementation of the modern Benders decomposition. The same first phase (Step 2 in Figure 1) as in the traditional approach is used to find a set of promising Benders cuts. In the second phase, we start the MIP solver with this set of cuts and treat all other constraints (4b) and (4c) as lazy

constraints; however, because the number of these constraints is too large to be enforced explicitly as lazy constraints, a separation procedure is necessary to identify violations. Figure 2 depicts the flowchart for updating the incumbent. Given any candidate solution $\hat{\psi}$ and its objective function value $\hat{W}$ in the restricted master problem, model (3) with $\psi = \hat{\psi}$ is solved. If the resulting problem is unbounded, the feasibility cut corresponding to $\hat{\psi}$ is added to the current restricted master problem. If the resulting solution is bounded and $\hat{W} = h(\hat{\psi})$, the incumbent is updated as in traditional branch and bound. Otherwise, the indicated optimality cut is added to the current restricted master problem.

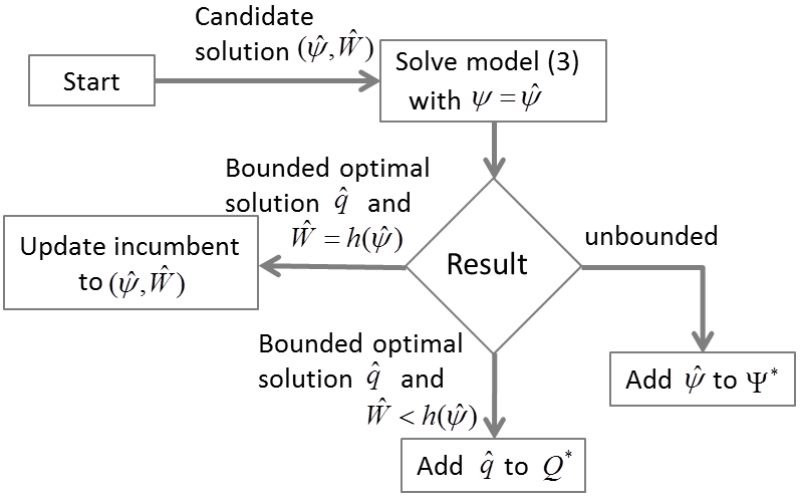


Figure 2. Logic for updating the incumbent solution in modern Benders decomposition

After solving the restricted master problem in modern Benders decomposition, the optimal solution satisfies all constraints (4b) and (4c), while only a subset of them typically needs to be included in (4). The modern approach makes better use of the existing information in the current search tree (Rubin 2011) because it exploits all the information gathered in previous runs rather than discarding it. A single search tree is used. When the modern approach identifies a violated Benders cut and applies it as a lazy constraint to the restricted master problem, a state-of-the-art solver like CPLEX should be able to resume the enumerations without reinitializing the search.

Another point to make is that the modern approach can generate many more Benders cuts than the traditional approach. The former generates a Benders cut whenever a candidate solution is found in the restricted master problem, while the latter only generates a Benders cut when the optimal solution of the restricted master problem is found. Nevertheless, adding Benders cuts that are derived from a non-optimal solution could be a double-edge sword: additional cuts increase the size of the restricted master problem and thus increase its difficulty, but they can also help to improve its lower bound (Rei 2009).

5. Heuristics

In this section, we present a two-stage optimization-based heuristic to solve model (1). In the first stage, a feasible solution is found by LP rounding. In the second stage, this solution is used as a starting point for local branching, which is designed to find improved solutions in the neighborhood of the incumbent.

5.1. LP rounding heuristic

The angle selection problem is difficult since there is an exponential number of angle combinations from which to choose. Several researchers have proposed heuristics aimed at eliminating unpromising choices. Lim et al. (2008), for example, developed an iterative scheme that removes angles based on scores derived from the LP solution of the original problem. Here, we discuss a heuristic that iteratively eliminates unpromising angles in the LP solution until the remaining problem is manageable as an IP. Specifically, we solve the LP relaxation and eliminate angles that are least used in the solution in each iteration. The process terminates when a given number of angles, denoted by U^*, are removed. At that point, the reduced IP is solved with only the remaining set of angles. Figure 3 describes the procedure.

Procedure_LP_rounding_heuristic

Input: U^* = maximum number of angles to eliminate

Step 1: $A^* = \varnothing$

While $|A^*| < U^*$

Solve the LP relaxation of model (2) with $\psi_a = 0$ for $a \in A^*$ and denote the solution by ψ^*.

Let $a^* = \arg\min_a \{\psi_a : a \notin A^*\}$

Put $A^* \leftarrow A^* \cup \{a^*\}$

Step 2: Solve the reduced IP with $\psi_a = 0$ for all $a \in A^*$

Figure 3. LP rounding heuristic

5.2. Local branching

To improve the quality of a given feasible solution, Fischetti and Lodi (2003) proposed the idea of local branching, which sets up an *outer* search tree that may be partially or

fully explored depending on the time available for the computations and the desired accuracy of the solution. They use a generic MIP solver as a black-box tool to explore the neighborhood of a given solution in hopes of finding better solutions. In this section, we describe the local branching heuristic we implemented to improve the solutions found by the rounding heuristic.

Lemma 2: If $\eta \leq |A|$, there exists an optimal solution to model (1) with exactly η angles chosen.

Proof: Assume there is an optimal solution ψ^* with $n < \eta$ angles. We can always find another solution with n angles in solution ψ^* and an additional $\eta - n$ angles whose beamlets do not carry any intensity. ■

As indicated in the proof of Lemma 2, any solution that uses n angles, with $n < \eta$, can possibly be improved by incorporating $\eta - n$ additional angles and so may not be a local optimum. Our computational experience confirms that local optima always contain η angles. Now, Fischetti and Lodi define the k-OPT neighborhood of $\bar{\psi}$ as the set of feasible angles that satisfy the additional local branching constraint

$$\Delta(\psi, \bar{\psi}) = \sum_{a \in A} (1 - \bar{\psi}_a)\psi_a \leq k \qquad (6)$$

Intuitively, constraint (6) permits at most k of the angles selected in $\bar{\psi}$ to be replaced by angles not in the solution.

Local branching can be used as a heuristic or as an exact method. Given the incumbent solution $\bar{\psi}$, we introduce an *outer* search tree that is constructed by enforcing $\Delta(\psi, \bar{\psi}) \leq k$ on the left branch and $\Delta(\psi, \bar{\psi}) \geq k+1$ on the right branch as in depth-first search. This procedure represents a high level partition of the solution space. Denote the set of left branch constraints as L and the set of right branch constraints as R. We start the procedure with $R = \varnothing$, $L = \varnothing$, and a feasible solution $\bar{\psi}$. At each iteration, we seek to find better solutions in the neighborhood of the incumbent ψ^* by solving model (2) with the existing sets L and R, and the local branching constraint $\Delta(\psi, \psi^*) \leq k$. To control the time spent on each subproblem, we impose a limit of τ hours on each.

Three situations may arise when solving a node: (i) if a better solution is found within time τ, then the incumbent is updated and used as the new starting point to search for better solutions; (ii) if the subproblem is solved to optimally within time τ but no better solution is found, then we expand the local branching neighborhood by putting $k \leftarrow k + 1$ and continue; and (iii) if the subproblem is not solved to optimality within time τ and no better solution is found, then we terminate the procedure and return the best solution found so far. In the latter case, the solution space of the current subproblem is too large to be fully explored within time τ. Larger values of k in the local branching constraint $\Delta(\psi, \psi^*) \leq k$ correspond to larger neighborhoods and result in more difficult instances. When the value of k is too large, the subproblem approaches the original problem and becomes difficult to solve optimally. In our case, we terminate the computations when the value of k reaches the threshold $k_{\max}$. The procedure is outlined in Figure 4.

Procedure_Local_branching

Input: Initial feasible solution ψ^0

Time limit to solve each subproblem τ

Minimum neighborhood parameter $k_{\min}$

Maximum neighborhood parameter $k_{\max}$

Output: Improved feasible solution ψ^*

Step 0: Iteration count $m = 1$

Set of right branch constraints $R = \varnothing$

Set of left branch constraints $L = \varnothing$

$\psi^* = \psi^0$

Step 1: $k = k_{min}$

Step 2: Solve model (2) with additional sets of constraints R and L and constraint $\Delta(\psi, \psi^{n-1}) \le k$, and set time limit to τ; denote the resulting solution, if any, as ψ^{n+1}

Step 3: If $v(\psi^*) > v(\psi^{n+1})$, then //better solution is found

Update $\psi^* = \psi^{n+1}$

Add constraint $\Delta(\psi, \psi^{n+1}) \ge k+1$ to L

Put $m \leftarrow m + 1$

Go to step 1.

Else if the problem is solved to optimality, then

//expand the neighborhood for better solutions

Put $k \leftarrow k + 1$

If $k < k_{max}$, then

Go to Step 2

Else

Terminate the procedure and return solution ψ^*.

Else // problem is not solved to optimality

Terminate the procedure and return ψ^*.

Figure 4. Local branching logic

6. Computational Results

Four data sets associated with prostate tumors were used to test the effectiveness of our algorithms. The details are summarized in Table 1. Instances PX-12 and PX-36, with X = 1, 2, 3, correspond to the same clinical case, but with a different number of candidate angles. Instances P1-36 and P2-36 both use 36 beam angles. Instance P2-36 has many more voxels for PTV than instance P1-36, although the number of voxels for OARs is slightly smaller. Instance P3-36 has the largest number of voxels for PTV but the smallest number of voxels for OARs. For instance P3-36, all other OARs except the bladder are removed for convenience.

Table 1. Data set summary

Measure	Instance					
	P1-12	P1-36	P2-12	P2-36	P3-12	P3-36
# of candidate angles	12	36	12	36	12	36
# of voxels for PTV	1,000	1,000	4,005	4,005	5,245	5,245
# of voxels for bladder (OAR)	10,603	10,603	7,850	7,850	0	0
# of voxels for rectum (OAR)	5,848	5,848	5,719	5,719	1,936	1,936
# of positive d_{vb}	1.27E7	1.95E7	1.10E7	3.31E7	2.88E6	8.72E6

The number of positive d_{vb} in Table 1 is roughly proportional to the density of the constraint matrix in model (1), and is thus a good indicator of problem instance difficulty. As shown in constraint (1b), the total dose D_v delivered to each voxel v is the weighted sum of d_{vb}, with beamlet intensity ω_{ab} as the weight. Thus, if the number of positive d_{vb} is reduced, the number of possible D_v values is also reduced. Accordingly, the number of positive d_{vb} can affect the feasible region of the dose value applied to each voxel.

All algorithms were implemented in JAVA and run under Ubuntu Linux on a Dell Poweredge T610 workstation with two 6-core hyperthreading 3.33-GHz Xeon processors and 24 GB of memory. CPLEX 12.4 was used as the MIP solver. In the computations, we followed the convention in Lim et al. (2007) and used the following parameter values for instances P1 and P2: $\theta_U = 1.05$, $\theta_L = 0.97$, $L_v = 0.94$ and $U_v = 1.15$ for all $v \in P$, and

$\phi = 0.3$. Since P3 instances have large PTV, we used the following parameter values to better control cold and hot spots: $\theta_U = 1.05$, $\theta_L = 0.97$, $L_v = 0.96$ and $U_v = 1.15$.

As an initial test, instances P1-12, P2-12, and P3-12, which have $|A| = 12$, were solved, first with CPLEX alone, and then with traditional Benders decomposition and modern Benders decomposition. All methods converged within 30 minutes giving optimal objective function values. The solution times presented in Table 2 demonstrate that a 12-angle problem can be solved quickly with standard commercial software. Both Benders decompositions, although not as efficient as CPLEX for instances P1-12 and P2-12, can still get optimal solutions within a reasonable amount of time.

Table 2. Solution time for 12-angle problems

Instance	Objective value	Time(min)		
		CPLEX	Traditional Benders	Modern Benders
P1-12	0.0306	3	7	8
P2-12	0.0121	20	22	28
P3-12	0.1148	23	14	10

For instances with more angles, CPLEX, Benders decomposition and the two stage heuristic were applied to find feasible solutions, although the former did not always converge. Also, because local branching subproblems with $k \geq 4$ are too difficult to solve to optimality, we used parameter values $k_{\min} = 1$, $k_{\max} = 3$ in the implementation. The subproblem time limit τ was set to 2 hours to control the total local branching time. Since 12-angle problems are usually well-solved, we used $U^* = 24$ in the LP-based heuristic.

Table 3 compares the solution times and solution quality of the different methods and Table 4 summarizes the characteristics of the MIP formulations. One key observation from the computations is that problem instances become more difficult as the number of voxels in PTV grows. For problems with small PTV (e.g., instance P1-36), CPLEX can solve model (1) directly; for problems with large PTV, CPLEX has a hard time closing the optimality gap (e.g., instance P3-36) although the results are only a few

percentage points from the best solution found. In contrast, the number of voxels in OARs does not have a noticeable impact.

Table 3. Comparison of different solution methods

Instance	Method	CPLEX	Modern Benders	Traditional Benders	LP-rounding heuristic	Two-stage heuristic
P1-36	Time (min)	300	300	300	5	173
	Obj. val.	0.03036	0.03036	0.03046	0.03048	0.03036
	% of best [a]	100.00%	100.00%	100.33%	100.40%	100.00%
P2-36	Time (min)	300	300	300	51	284
	Obj. val.	0.01183	0.01167	0.01170	0.01158	0.01158
	% of best [a]	102.16%	100.78%	101.04%	100.00%	100.00%
P3-36	Time (min)	300	300	300	71	280
	Obj. val.	0.12151	0. 11679	0.11830	0.11411	0.11411
	% of best [a]	106.49%	102.35%	103.67%	100.00%	100.00%

[a] % of best = current obj. val. / best obj. val. among all methods × 100%

Table 4. Dimensions of MIP formulation for instances P1-36, P2-36 and P3-36

Name	P1-36	P2-36	P3-36
Optimality gap [a]	2.84%	14.59%	74.41%
# of variables	20,364	25,098	9,891
# of constraints	22,327	33,071	20,344
LP relaxation time (sec)	115	1126	228

[a] Optimality gap = the mixed integer programming gap when CPLEX terminates

Another implication of the result is that the magnitude of d_{vb} plays an important role in determining the problem difficulty. As we can see, P3-36, which has only 1000 more voxels in PTV than P2-36, has a much larger optimality gap (74.41%) than P2-36 (14.59%), indicating that P3-36 is much more difficult to solve. This is probably due to the relative magnitude of d_{vb}. Because the doses per beamlet are much smaller in P3-36 than in P2-36, the bound provided by the LP relaxation for P3-36 is weaker, leading to a larger optimality gap. Although reducing d_{vb} for a subset of v and b can make the LP relaxation easier to solve, it also leads to a weaker relaxation. In summary, the results suggest that the number of voxels in PTV and the magnitude of d_{vb} are key indicators of problem difficulty.

Table 3 also indicates that modern Benders decomposition can find better feasible solutions than traditional Benders decomposition. Moreover, depending on the instance, the solution is either the same or better than the solution obtained with CPLEX. Note that neither Benders algorithms required feasibility cuts.

Table 5 compares the computational aspects of the two approaches. The results indicate that the modern approach generates many more optimality cuts than the traditional approach. That is, many more subproblems are solved to identify violated cuts, and thus many more feasible solutions are examined. However, the lower bound associated with the modern approach, which is obtained by solving the restricted master problem, is not as strong as the bound from the traditional approach. This implies that the Benders cuts generated in the modern approach are not as effective in improving the bounds. Given that our focus is on finding good solutions rather than closing the optimality gap, the modern approach is the better choice since it generates better feasible solutions.

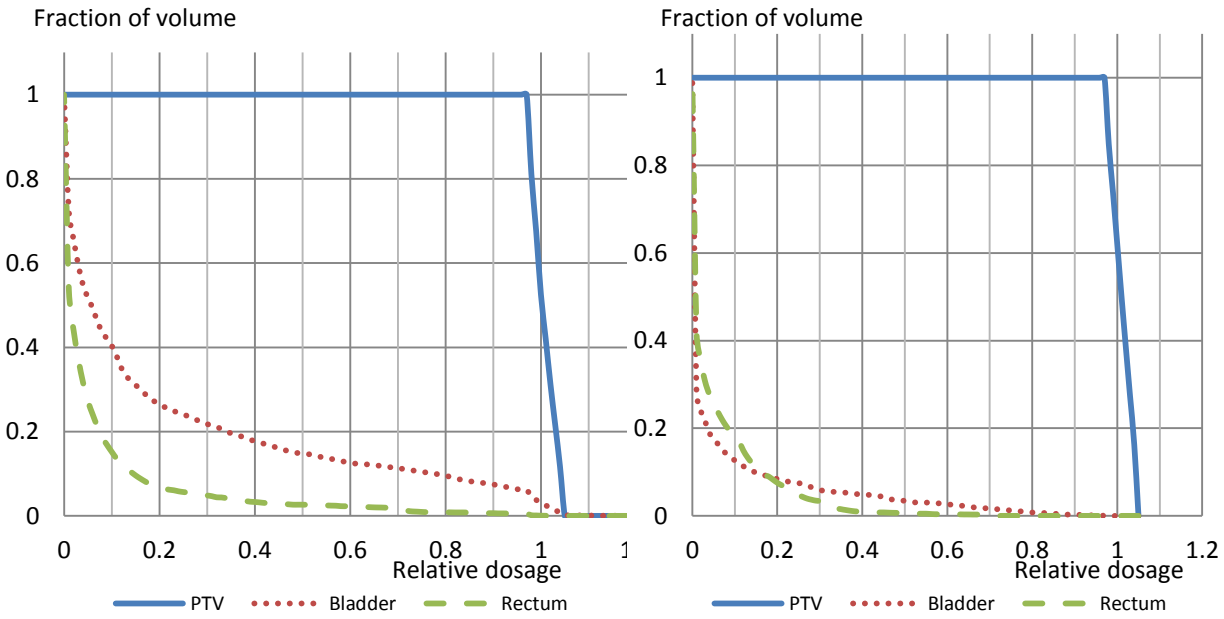
Table 5. Comparison between traditional and modern Benders decomposition

		Modern Benders		Traditional Benders	
Instance	First stage time (min)	Lower bound	# of Benders cuts	Lower bound	# of Benders cuts
P1-36	5	0.027851	11,824	0.028403	1,669
P2-36	140	0.009792	1,170	0.010164	750
P3-36	58	0.029561	6,791	0.036965	721

Table 3 also shows that the heuristic solutions obtained from the LP rounding heuristic are either close to (for instance P1-36) or better than (for instances P2-36 and P3-36) the solutions given by CPLEX and both Benders decompositions. Moreover, the solution times are only a small fraction of those of the latter methods. The two-stage heuristic, using local branching provided the best results with shorter runtimes. Nevertheless, the computations still took much longer than desired, especially for P2-36. As the problem size grows, especially with respect to the number of voxels in PTV, runtimes become excess for CPLEX, Benders decomposition, and local branching.

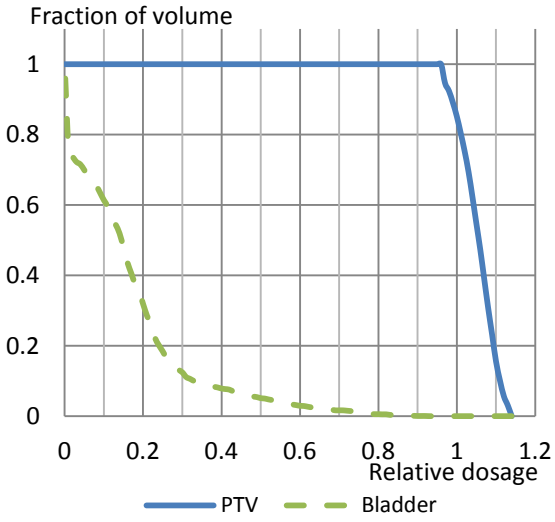
When a good solution is needed quickly, the best approach is to use the LP rounding heuristic.

A common way to graphically demonstrate the effectiveness of a treatment plan is by plotting the dose volume histogram (DVH) for PTV and each OAR. The horizon axis of the DVH represents the dose value, and the vertical axis represents the fraction of volume. The DVH contains one curve for PTV, and one for each of the OARs. Each point on the curve specifies the percentage of volume (in the corresponding OAR or PTV) that receives a dose greater than a give value. For example, point (0.3, 21.83%) on the curve for the bladder (OAR) in Figure 5(a) indicates that 21.83% percent of voxels in the bladder receives dose more than 0.3. Figure 5 shows the DVH for solutions obtained by the LP rounding heuristic for instances P1-36, P2-36 and P3-36. As we can see in Figure 5(a) for P1-36, the percentage of voxels in OARs, i.e., bladder and rectum, receiving more than $\phi = 0.3$ of the relative does is only 21.83 % and 4.85%, respectively. Also, the relative dose for all the voxels in PTV is between $\theta_L = 0.97$ and $\theta_U = 1.05$. Similarly, Figure 5(b) shows that only 5. 95% of the voxels in the bladder and 3.37% of the voxels in the rectum receive doses higher than $\phi = 0.3$ in the solution for instance P2-36. Also, the relative dose for all voxels in PTV is between $\theta_L = 0.97$ and $\theta_U = 1.05$. Since the number of voxels in the bladder is 0, Figure 5(c) only shows the DVH of PTV and rectum for instance P3-36. Figure 5(c) shows that 12.45% of the voxels in the rectum receive doses higher than $\phi = 0.3$. All voxels in PTV receive doses between 0.96 and 1.14, which is between the bounds $L_v = 0.96$ and $U_v = 1.15$, and 39.98% of them are between $\theta_L = 0.97$ and $\theta_U = 1.05$. This indicates that the majority of the radiation is delivered to PTV while largely sparing OARs.



(a). Instance P1-36

(b). Instance P2-36



(c). Instance P3-36

Figure 5. Dose volume histogram for LP-rounding heuristic solutions

7. Summary and Conclusions

This study explored the use of Benders decomposition and optimization-based heuristics to solve the beam angle and fluence map problem for IMRT treatment planning. The results showed that instances with 12 angles can all be solved quickly with any of the proposed methods. For the larger instances with 36 angles, Benders decomposition can generate good feasible solutions, at least with respect to CPLEX, but after 5 hours of computations large optimality gaps still remained. Comparatively speaking, we also found that modern Benders decomposition, which generates more optimality cuts than the traditional approach, can produce slightly better solutions within the same amount of time. The best results were obtained with the LP rounding heuristic in conjunction with local branching. When runtimes are critical, the best compromise is to use the LP heuristic by itself. For future researches, it is appealing to study how to strengthen the Benders feasibility cuts or identify better Benders cuts so as to improve the Benders lower bound. Besides, it may be worthwhile to apply other local search heuristic to improve the solution generated by LP-rounding.

8. References

Benders, J. F. (1962). Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik*, *4*(1), 238-252.

Binato, S., M. V. F. Pereira, and S. Granville (2001). A new Benders decomposition approach to solve power transmission network design problems. *IEEE Transactions on Power Systems, 16*(2), 235-240.

Ehrgott, M., Ç. Güler, H. W. Hamacher, and L. Shao (2008). Mathematical optimization in intensity modulated radiation therapy. *4OR*, *6*(3), 199-262.

Chang, J. Y., X. Zhang, X. Wang, Y. Kang, B. Riley, S. Bilton and J. D. Cox (2006). Significant reduction of normal tissue dose by proton radiotherapy compared with three-dimensional conformal or intensity-modulated radiation therapy in Stage I or Stage III non–small-cell lung cancer. *International Journal of Radiation Oncology, Biology, Physics*, *65*(4), 1087-1096.

Cordeau, J. F., G. Stojković, F. Soumis, and J. Desrosiers (2001). Benders decomposition for simultaneous aircraft routing and crew scheduling. *Transportation Science*, *35*(4), 375-388.

Costa, A. M. (2005). A survey on benders decomposition applied to fixed-charge network design problems. *Computers & Operations Research*, *32*(6), 1429-1450.

CPLEX, IBM ILOG. (2011). V12.4: User's Manual for CPLEX. *International Business Machines Corporation*, *46*(53), 157.

Fischetti, M. and A. Lodi (2003). Local branching. *Mathematical Programming*, *98*(1-3), 23-47.

McDaniel, D., and M. Devine (1977). A modified Benders' partitioning algorithm for mixed integer programming. *Management Science*, *24*(3), 312-319.

Lee, E. K., T. Fox, and I. Crocker (2000). Optimization of radiosurgery treatment planning via mixed integer programming. *Medical Physics*, *27*(5), 995-1004.

Lim, G. J., M. C. Ferris, S. J. Wright, D. M Shepard, and M. A. Earl (2007). An optimization framework for conformal radiation treatment planning. *INFORMS Journal on Computing*, *19*(3), 366-380.

Lim, G. J., and W. Cao (2012). A two-phase method for selecting IMRT treatment beam angles: branch-and-prune and local neighborhood search. *European Journal of Operational Research*, *217*(3), 609-618.

Lim, G. J., L. Kardar, and W. Cao (2014). A hybrid framework for optimizing beam angles in radiation therapy planning. *Annals of Operations Research*, *217*(1), 357-383.

Lim, G. J., J. Choi, and R. Mohan (2008). Iterative solution methods for beam angle and fluence map optimization in intensity modulated radiation therapy planning. *OR Spectrum*, *30*(2), 289-309.

Milano, M. T., M. C. Garofalo, S. J. Chmura, K. Farrey, C. Rash, R. Heimann, and A. B. Jani (2006). Intensity-modulated radiation therapy in the treatment of gastric cancer: early clinical outcome and dosimetric comparison with conventional techniques. *The British Journal of Radiology*, *79*(*942*), 497-503

Mock, U., D. Georg, J. Bogner, T. Auberger, and R. Pötter (2004). Treatment planning comparison of conventional, 3D conformal, and intensity-modulated photon (IMRT) and proton therapy for paranasal sinus carcinoma. *International Journal of Radiation Oncology Biology Physics*, *58*(1), 147-154.

Rei, W., J. F. Cordeau, M. Gendreau, and P. Soriano (2009). Accelerating Benders decomposition by local branching. *INFORMS Journal on Computing*, *21*(2), 333-345.

Romeijn, H. E., R. K. Ahuja, J. F. Dempsey, A. Kumar, and J. G. Li (2003). A novel linear programming approach to fluence map optimization for intensity modulated radiation therapy treatment planning. *Physics in Medicine and Biology*, *48*(21), 3521.

Rubin, P. October 9, 2011. OR in an OB World. http://orinanobworld.blogspot.com/2011/10/benders-decomposition-then-and-now.html

Spirou, S. V., and C. S. Chui (1998). A gradient inverse planning algorithm with dose-volume constraints. *Medical Physics*, *25*(3), 321-333.

Saka, B., R. L. Rardin, and M. P. Langer (2013). Biologically guided intensity modulated radiation therapy planning optimization with fraction-size dose constraints. *Journal of the Operational Research Society*, *65*(4), 557-571.

Saka, B., R. L. Rardin, M. P. Langer, and D. Dink (2011). Adaptive intensity modulated radiation therapy planning optimization with changing tumor geometry and fraction size limits. *IIE Transactions on Healthcare Systems Engineering*, *1*(4), 247-263.

Yarmand, H., B. Winey and D. Craft (2013). Guaranteed epsilon-optimal treatment plans with the minimum number of beams for stereotactic body radiation therapy. *Physics in Medicine and Biology*, *58*(17), 5931.

Zelefsky, M. J., Z. Fuks, L. Happersett, H. J. Lee, C. C. Ling, C. M. Burman, and S. A. Leibel, (2000). Clinical experience with intensity modulated radiation therapy (IMRT) in prostate cancer. *Radiotherapy and Oncology*, *55*(3), 241-249.