European Journal of Operational Research 000 (2016) 1–12

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Interfaces with Other Disciplines

Benders decomposition and an IP-based heuristic for selecting IMRT treatment beam angles

Sifeng Lin[a], Gino J. Lim[b], Jonathan F. Bard[a,*]

[a] Graduate Program in Operations Research & Industrial Engineering, The University of Texas, Austin, TX 78712, United States
[b] Department of Industrial Engineering, The University of Houston, Houston, TX 77204, United States

ARTICLE INFO

Article history:
Received 11 February 2015
Accepted 31 December 2015
Available online xxx

Keyword:
Intensity modulated radiation therapy (IMRT)
Radiation therapy
Benders decomposition
Local branching
Integer programming

ABSTRACT

In this paper, two Benders decomposition algorithms and a novel two-stage integer programming-based heuristic are presented to optimize the beam angle and fluence map in Intensity Modulated Radiation Therapy (IMRT) planning. Benders decomposition is first implemented in the traditional manner by iteratively solving the restricted master problem and then identifying and adding the violated Benders cuts. We also implemented Benders decomposition using the "lazy constraint" feature included in CPLEX. In contrast, the 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 to evaluate 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.

1. Introduction

Intensity modulated radiation therapy (IMRT) is one of the most commonly used procedures to treat cancer patients because of its demonstrated capability to deliver a highly conformal dose of radiation to tumors while minimizing damage to nearby healthy organs. During the procedure, the patient lies on a specially designed couch while a linear accelerator (LINAC) generates photon beams to irradiate the target volume. The movable arm, called the gantry, of the LINAC can rotate 360° in a plane perpendicular to the couch. In sequence, it stops at a predetermined set of angles to deliver the radiation. If need be the couch can also rotate. At each angle, various two-dimensional beam shapes are constructed by a computer controlled multi-leaf collimator (MLC) in the LINAC to conform to the beam's-eye-view of the target volume. An open radiation beam consists of hundreds of beamlets or pencil beams that are defined by the specification of the MLC. Each beamlet is assigned a radiation intensity level called a beamlet weight, which can be modulated (the result is a fluence intensity map). Because it can be time consuming to control one beamlet at a time for radiation delivery, subfields are often constructed and the same intensity is assigned so that the resulting dose delivery is close to the optimized fluence map. For more detail about the procedure and equipment, see Lim, Choi, and Mohan (2008).

Before designing an IMRT treatment plan, the geometry of the tumor region is captured using a medical imaging technique such as computed tomography or magnetic resonance imaging, depending on the type of tumor. The three-dimensional information obtained from the digital image representing the tumor as well as the surrounding healthy organs is then used for developing the treatment plan. Two types of volumes are typically considered at this stage. The first is the planning target volume (PTV), which consists of the gross tumor volume and a fixed margin around the tumor to account for treatment planning parameter and data uncertainties. Unfortunately, some healthy organs can be damaged by the radiation due to their proximity to the tumor. These are referred to as organs-at-risk (OARs), the second type of volume. The PTV corresponds to the area of the cancer while the OARs are healthy organs or tissues of the body. A well-designed IMRT plan should deliver the desired amount of radiation to the tumor while limiting the amount to the healthy organs.

For planning purposes, the PTV and the involved OARs are divided into three-dimensional treatment cubes called voxels (for

* Corresponding author. Tel.: +1 512 471 3076; fax: +1 512 471 8727.
E-mail addresses: sifenglin@utexas.edu (S. Lin), ginolim@central.uh.edu (G.J. Lim), jbard@mail.utexas.edu (J.F. Bard).

http://dx.doi.org/10.1016/j.ejor.2015.12.050
0377-2217/

more detail, see Lim, Ferris, Wright, Shepard, & Earl, 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, a hot spot is the region that receives a higher dose than the desired level, and a cold spot is the region receiving a dose below its desired level. For voxels in the PTV, we want to control both the hot spots and cold spots to guarantee the desired treatment effect. For voxels in OARs, we only control hot spots in an effort to spare the healthy organs. Both hot and cold spot control can be modeled by either enforcing hard constraints or by penalizing deviations from the desired dose.

The purpose of this study is to explore solution techniques that can help planners select the optimal radiation delivery angles and fluence intensity maps. Various mixed-integer programming (MIP) formulations have been proposed to solve this problem (e.g., see Aleman et al., 2013; Lee, Fox, & Crocker, 2000; Lim et al., 2007; 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 Lagrangian relaxation and Benders, which have been successfully applied in other areas, have not been applied to the treatment angle selection problem associated with 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 when the computation time is limited. In addition, we develop a two-stage heuristic that uses (i) a standard MIP solver to construct a good initial solution in the first stage, and (ii) local branching (Fischetti & Lodi, 2003) to improve the incumbent 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 feasible solutions, which may be further improved by local branching in the second stage.

The rest of the paper is organized as follows. In the next section, we summarize the related literature concentrating on algorithm methods for obtaining solutions. Section 3 introduces the problem formulation and Section 4 discusses our Benders implementations. 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 with a number of observations from the research.

2. Literature review

IMRT is currently the mainstream treatment modality worldwide for irradiating tumors. Its application and efficacy have been extensively discussed by the medical research community (e.g., Milano et al., 2006; Zelefsky et al., 2000).

An IMRT plan specifies the angles at which the doses are to be delivered, along with the fluence intensity map. Accordingly, planners face three interrelated problems: the beam angle optimization (BAO) problem, the fluence intensity optimization (FMO) problem, and the beam segmentation problem. Since the 1970s, researchers have developed various methods to solve FMO. Depending on the penalty function for the undesired dose delivered to voxels in the PTV or OARs, the resulting model may be a linear or nonlinear (usually quadratic) mathematical program. Ehrgott, Güler, Hamacher, and Shao (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; Lim et al., 2014; Romeijn, Ahuja, Dempsey, Kumar, & Li, 2003; Saka, Rardin, & Langer, 2013; Saka, Rardin, Langer, & Dink, 2011). 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.

The joint problem of angle selection and FMO entails selecting a subset of angles and specifying the fluence map for the selected angles. The BAO problem is a very large-scale optimization problem that can be formulated as mixed-integer linear program if the objective function is linear (Lee et al., 2000; Lim et al., 2007, 2008; Wang, Dai, & Hu, 2003; Yarmand, Winey, & Craft, 2013). Optimally selecting a small subset of angles from a 360 degree circumference using a MIP model, however, can be a daunting task and it is clinically not useful due to long computation times. Consequently, researchers have developed practical approaches using metaheuristics such as simulated annealing (SA) (Bortfeld & Schlegel, 1993), genetic algorithms (GA) (Ezzell, 1996; Lei & Li, 2009), nested partitions (NP) (Zhang, Shi, Meyer, Nazareth, & D'Souza, 2009), particle swarm (Li, Yao, Yao, & Chen, 2005), response surface (Aleman, Romeijn, & Dempsey, 2009), data mining (Price, Golden, Wasil, & Zhang, 2014), and neural networks (Rowbottom, Webb, & Oldham, 1999). When the FMO objective function is nonlinear, the problem is usually solved using neighborhood search (Aleman, Kumar, Ahuja, Romeijn, & Dempsey, 2008; Djajaputra, Wu, Wu, & Mohan, 2003; Mišić, Aleman, & Sharpe, 2010; Rowbottom, Nutting, & Webb, 2001). That is, a metaheuristic is applied to explore the neighborhood of a given angle selection solution, and then a series of FMO problems are solved to evaluate the quality of angles selected.

Researchers have also investigated other models and solution strategies to concurrently optimize the angle selection and fluence intensity map. Rocha, Dias, Ferreira, and Lopes (2013) used a pattern search method. Bertsimas, Cacchiani, Craft, and Nohadani (2013) proposed a nonconvex mathematical programming model that uses continuous variables to denote the beam angles. To solve the problem, they designed a hybrid heuristic that exploits gradient information to quickly find a local minimum and simulated annealing to search for the global minimum. Lee, Aleman, and Sharpe (2011) transformed the beam orientation optimization problem, whose objective is limited to only selecting the beam angles, into a set covering problem. Beam intensities were not part of the solution, only the requirement that each voxel be covered by at least a prespecified number of beam angles.

More recently, Lim, Kardar, and Cao (2014) examined the strengths and weaknesses of six optimization methods for selecting beam angles: branch and bound, SA, GA, NP, branch and prune (BP), and local neighborhood search (LNS). They concluded that it is more effective to apply hybrid approaches 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. Although some hybrid approaches can solve larger scale BAO problems to within a small percentage of the global optimal, they still require a substantial amount of time to converge.

3. Problem formulation

Devising an IMRT plan involves selecting a subset of angles and constructing 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 the PTV and penalize the hot spots for voxels in OARs. The following notation is used in the developments.

Indices and sets

a	index for a beam angle
A	set of candidate beam angles
i	index for OAR
O	set of OARs
v	index for a voxel

Please cite this article as: S. Lin et al., Benders decomposition and an IP-based heuristic for selecting IMRT treatment beam angles, European Journal of Operational Research (2016), http://dx.doi.org/10.1016/j.ejor.2015.12.050

T set of voxels in the PTV
S_i set of voxels in organ i; $i \in O$
V set of all voxels in the PTV and OARs
b index for a beamlet
B_a set of beamlets in angle a

Parameters

η maximum number of treatment angles in a treatment plan; $\eta < |A|$
d_{vb} dose deposition coefficient for voxel v and beamlet b (when beamlet b has unit intensity, its relative dose contribution to voxel v is d_{vb})
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 the PTV
θ_L cold spot control parameter on voxels in the PTV
ϕ_i hot spot control parameter on voxels in OAR i
λ_t^+ nonnegative penalty coefficient for hot spots in the PTV
λ_t^- nonnegative penalty coefficient for cold spots in the PTV
λ_s nonnegative penalty coefficient for hot spots in OARs
M_{ab} maximum intensity of beamlet $b \in B_a$

Decision variables

ψ_a 1 if angle a is selected, 0 otherwise
w_{ab} intensity of beamlet $b \in B_a$ for angle $a \in A$

Auxiliary variables

D_v total relative 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 the PTV, D_{S_i} for the vector of dose values applied to organ i, and $D = (D_T, (D_{S_i})_{i\in O}) = (D_v)$, that is, 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:

$$f(D) = \lambda_t^+ \|(D_T - \theta_U e^T)^+\|_\infty + \lambda_t^- \|(\theta_L e^T - D_T)^+\|_\infty + \sum_{i\in O} \lambda_s \|(D_{S_i} - \phi_i e^{S_i})^+\|_1 / |S_i|$$

where $\|\mathbf{x}\|_\infty$ and $\|\mathbf{x}\|_1$ respectively denote the infinity and 1-norm of vector $\mathbf{x}$, $(y)^+ \equiv \max\{y,0\}$, 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 the PTV by penalizing the maximum excess dose and maximum shortage of dose, respectively; the third term controls the hot spot for voxels in OAR i by penalizing a dose that is more than ϕ_i, a control parameter given by the oncologist. Because λ_t^+, λ_t^-, and λ_s are nonnegative parameters, $f(D)$ is a convex function of D. The full model is as follows.

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

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

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

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

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

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

The objective function (1a) minimizes the total penalty from the radiation doses applied to both the PTV and the neighboring organs. Constraint (1b) calculates the total dose that each voxel receives, which is the sum of beamlet intensities weighted by their corresponding dose deposition coefficient. Constraint (1c) limits the number of angles that can be used to the maximum specified. Next, constraint (1d) guarantees the nonnegativity of the beamlet intensity and ensure that a beamlet can only carry a positive intensity if the angle is selected. We 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 total dose applied to the PTV in (1e) and the angle selection variable, ψ_a, is defined to be binary in (1f).

To solve the problem, it is first necessary to create an equivalent linear formulation of model (1). To do so, define $y_v = (D_v - \phi_i)^+$ for each $v \in S_i$, $z^+ = \|(D_T - \theta_U e^T)^+\|_\infty$, and $z^- = \|(\theta_L e^T - D_T)^+\|_\infty$, and note that $f(D)$ in (1a) is a piecewise linear convex function. For such functions, a solution to model (1) can be found by solving the following mixed-integer linear program:

$$\text{minimize } \lambda_t^+ z^+ + \lambda_t^- z^- + \sum_{i\in O} \sum_{v\in S_i} \lambda_s y_v / |S_i| \tag{2a}$$

$$\text{subject to } z^+ \ge D_v - \theta_U \quad \forall v \in T \tag{2b}$$

$$z^- \ge \theta_L - D_v \quad \forall v \in T \tag{2c}$$

$$y_v \ge \sum_{a\in A} \sum_{b\in B_a} d_{vb} w_{ab} - \phi_i \quad \forall v \in S_i, i \in O \tag{2d}$$

$$L_v \le D_v \le U_v \quad \forall v \in T \tag{2e}$$

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

$$\sum_{a\in A} \psi_a \le \eta \tag{2g}$$

$$0 \le w_{ab} \le M_{ab} \psi_a \quad \forall a \in A, b \in B_a \tag{2h}$$

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

$$y_v \ge 0 \quad \forall v \in S_i, i \in O \tag{2j}$$

$$z^+, z^- \ge 0 \tag{2k}$$

Because we have a minimization objective, (2a) is equivalent to (1a) when constraints (2b)–(2d), (2j), and (2k) are enforced. Note that although it is possible to substitute out the variables D_v for all $v \in T$, keeping them in the formulations improved computational performance. The constraint matrix is sparser with them but the difference is insignificant. A more effective way to reduce the number of variables and constraints is as follows:

- In the third term in (2a), we only penalize the excess dose applied to the voxels in OAR, i.e., the dose beyond ϕ_i. Thus, if some beamlet $b \in B_a$ does not affect the voxels in the PTV, i.e., $d_{vb} = 0$ for all $v \in T$, then 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). In our data sets, this observation only eliminates a very small number of variables.
- For any two voxels $v_1, v_2 \in T$, if $d_{v_1 b} \ge d_{v_2 b}$ for all $b \in B_a$ and $a \in A$, then we have $D_{v_1} - \theta_U \ge D_{v_2} - \theta_U$; accordingly, $z^+ \ge D_{v_1} - \theta_U$ is implied by constraints $z^+ \ge D_{v_2} - \theta_U$, and is thus redundant. Similarly, we can show that constraint $z^- \ge \theta_L - D_{v_2}$ can be removed.

Please cite this article as: S. Lin et al., Benders decomposition and an IP-based heuristic for selecting IMRT treatment beam angles, European Journal of Operational Research (2016), http://dx.doi.org/10.1016/j.ejor.2015.12.050

- If $\sum_{a\in A}\sum_{b\in B_a} d_{vb}M_{ab} - \phi_i \le 0$ for some $v \in S_i, i \in O$, then the constraint $y_v \ge \sum_{a\in A}\sum_{b\in B_a} d_{vb}w_{ab} - \phi_i$ is redundant. Accordingly, we can exclude this constraint and corresponding 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: $\min\{cx + dy : Ax + By \ge b, x \in \mathbb{Z}^n_+, y \in \Re^p_+\}$, where n is relatively small and when fixed, the constraints $By \ge b - Ax$ divide into independent subsets in the y 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 x 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, Stojković, Soumis, and Desrosiers (2001) applied Benders to simultaneously solve the aircraft routing and crew scheduling problems, while Binato, Pereira, and Granville (2001) used it to solve power transmission network design problems. Costa (2005) gives an extensive survey on applications to fixed-charge network design problems, while Taşkın, Sasaki, Segawa, and Ando (2012) used Benders to solve the problem of decomposing an IMRT fluence map into deliverable apertures. 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) reduce to an LP whose constraints are given in the model below. Their corresponding dual variables are specified on the right,

$$\text{minimize} \quad \lambda_t^+ z^+ + \lambda_t^- z^- + \sum_{i\in O}\sum_{v\in S_i} \lambda_s y_v/|S_i|$$

$$\begin{aligned}
\text{subject to} \quad & z^+ - \sum_{a\in A}\sum_{b\in B_a} d_{vb}w_{ab} \ge -\theta_U && \forall v \in T && \mu_v^+\\
& z^- + \sum_{a\in A}\sum_{b\in B_a} d_{vb}w_{ab} \ge \theta_L && \forall v \in T && \mu_v^-\\
& y_v - \sum_{a\in A}\sum_{b\in B_a} d_{vb}w_{ab} \ge -\phi_i && \forall v \in S_i,\ i \in O && \mu_v\\
& \sum_{a\in A}\sum_{b\in B_a} d_{vb}w_{ab} \ge L_v && \forall v \in T && \sigma_v^-\\
& -\sum_{a\in A}\sum_{b\in B_a} d_{vb}w_{ab} \ge -U_v && \forall v \in T && \sigma_v^+\\
& -w_{ab} \ge -M_{ab}\psi_a && \forall a \in A, b \in B_a && \pi_{ab}\\
& w_{ab} \ge 0 && \forall a \in A, b \in B_a &&
\end{aligned}$$

For simplicity, we will use the vector ψ to denote the angle section variables ψ_a for all $a \in A$. Model (3) is the corresponding dual formulation of the LP after fixing the angle selection vector ψ.

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

$$\text{subject to} \sum_{v\in T}\mu_v^+ \le \lambda_t^+ \tag{3b}$$

$$\sum_{v\in T}\mu_v^- \le \lambda_t^- \tag{3c}$$

$$0 \le \mu_v \le \lambda_s/|S_i| \qquad \forall v \in S_i, i \in O \tag{3d}$$

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

$$\mu_v^+, \mu_v^-, \sigma_v^+, \sigma_v^- \ge 0 \quad \forall v \in T \tag{3g}$$

$$\mu_v \ge 0 \quad \forall v \in S_i,\ i \in O \tag{3h}$$

$$\pi_{ab} \ge 0 \quad \forall a \in A, b \in B_a \tag{3i}$$

Evidently, model (3) is feasible. To see this, consider that a solution with $\pi_{ab} = 0$ for all $a \in A$, $b \in B_a$, $\mu_v^+ = \mu_v^- = \sigma_v^+ = \sigma_v^- = 0$ for all $v \in T$, and $\mu_v = 0$ for all $v \in S_i, i \in O$ satisfies constraints (3b)–(3i). Thus, model (3) must have either an extreme point solution or an unbounded optimal value. For a given angle selection vector ψ, having an unbounded optimal value, i.e., $h(\psi) = \infty$, indicates that the treatment plan with angle selection vector ψ is not feasible. Let Ψ denote the set of angle selection vectors that give an unbounded objective function value in model (3), i.e., $h(\delta) = \infty$ for all $\delta \in \Psi$. To prevent selecting any $\delta \in \Psi$, the standard Benders decomposition requires applying feasibility cuts based on extreme rays of polyhedral defined by (3b)–(3h). Alternatively, we can apply the following *Benders feasibility cuts*.

$$\sum_{a\in A}\delta_a\psi_a + \sum_{a\in A}(1-\delta_a)(1-\psi_a) \le |A| - 1 \quad \forall \delta \in \Psi^*$$

which ensure that $h(\psi) < \infty$ when $\psi \notin \Psi$.

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

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

Note that $g(\psi, q)$ defines the objective function in (3a). Thus, if $\psi \notin \Psi$, the objective function of model (3) is equivalent to

$$\text{minimize} \max\{g(\psi, q) : q \in Q\}.$$

Hence, model (2) can be reformulated as follows:

$$\text{minimize} \, W \tag{4a}$$

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

$$\sum_{a\in A}\delta_a\psi_a + \sum_{a\in A}(1-\delta_a)(1-\psi_a) \le |A| - 1 \quad \forall \delta \in \Psi^* \tag{4c}$$

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

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

$$W \ge 0 \tag{4f}$$

In model (4), constraint (4b) corresponds to Benders' optimality cuts and constraint (4c) corresponds to the feasibility cuts.

Please cite this article as: S. Lin et al., Benders decomposition and an IP-based heuristic for selecting IMRT treatment beam angles, European Journal of Operational Research (2016), http://dx.doi.org/10.1016/j.ejor.2015.12.050

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

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

As demonstrated by Magnanti and Wong (1981), the performance of Benders decomposition can be improved by strengthening the Benders cuts. 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 optimality cut. When $\psi_a > 0$, optimality of (3) ensures that the corresponding constraints (3f) are binding for all $b \in B_a$ (otherwise, we could 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

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

In this case, we can state the value of π_{ab} as follows:

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

Benders optimality cuts can be slightly strengthened when a lower bound for model (4) is known. One possible way to generate a lower bound is by solving the LP relaxation of model (2), which is what we do. 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 $b = \sum_{v\in T}(-\theta_U\mu_v^+ + \theta_L\mu_v^- + L_v\sigma_v^- - U_v\sigma_v^+) - \sum_{i\in O}\sum_{v\in S_i}\phi_i\mu_v$ for all $a \in A$. As such, we have $g(\psi, q) = b - \sum_{a\in A} c_a\psi_a$ with corresponding Benders optimality cut given by

$$W + \sum_{a\in A} c_a\psi_a \geq b. \tag{5}$$

Lemma 2. If W_1 is a lower bound on the optimal objective function value W^* in model (4), i.e., $W^* \geq W_1$, then constraint (5) is equivalent to

$$W + \sum_{a\in A} c^1_a\psi_a \geq b \tag{6}$$

where $c^1_a = \min\{c_a, b - W_1\}$.

Proof. We will show that the set of solutions associated with (5) and (6) are the same. First we note that given $c^1_a \leq c_a$, any solution that satisfies constraint (6) also satisfies (5). To show the opposite, we define $A_1 = \{a \in A : b - W_1 \leq c_a\}$. Now consider a solution $(\hat{W}, \hat{\psi})$ that satisfies (5). If $\hat{\psi}_a = 0$ for all $a \in A_1$, then we must have

$$\hat{W} + \sum_{a\in A} c^1_a\hat{\psi}_a = \hat{W} + \sum_{a\in A} c_a\hat{\psi}_a \geq b.$$

Otherwise, we can find an angle $e \in A_1$ with $\hat{\psi}_e = 1$. Since $e \in A_1$, we have $c^1_e = b - W_1 \leq c_e$. Now, noting that model (4) is a minimization problem, we must have $\hat{W} \geq W^* \geq W_1$. Accordingly,

$$\hat{W} + \sum_{a\in A} c_a\hat{\psi}_a \geq \hat{W} + c_e \geq W_1 + b - W_1 = b.$$

Thus, any solution that satisfies constraint (5) also satisfies constraint (6), and vice versa, which means that they are equivalent. □

Since constraint (6) is stronger, we use (6) rather than (5) for Benders optimality cuts in the subsequent developments,.

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. Specifically, we initialize model (7) with $Q^* = \emptyset$ and $\Psi^* = \emptyset$ which gives an initial solution $W = -\infty$.

$$\text{minimize} \quad W \tag{7a}$$

$$\text{subject to } W \geq g(\psi, q) \quad \forall q \in Q^* \tag{7b}$$

$$\sum_{a\in A}\delta_a(1-\psi_a) + \sum_{a\in A}(1-\delta_a)\psi_a \leq |A| - 1 \quad \forall \delta \in \Psi^* \tag{7c}$$

$$\sum_{a\in A}\psi_a \leq \eta \tag{7d}$$

$$\psi_a \in \{0, 1\} \quad \forall a \in A \tag{7e}$$

$$W \geq 0 \tag{7f}$$

Adding a Benders cut to (7) 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. 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 approach: in the first phase, we relax the integrality constraint on the ψ_a variables in model (7) and apply Benders decomposition until its objective function is within 5% of the value obtained by solving the LP relaxation of model (2); in the second phase, we start with the Benders cuts found in the first phase and continue in the traditional manner. Fig. 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.

More recently, 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 a good integer feasible solution is identified, and any of those constraints that turn out to be violated are added to 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 branch and bound search tree, the LP relaxation at the current node (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.

Please cite this article as: S. Lin et al., Benders decomposition and an IP-based heuristic for selecting IMRT treatment beam angles, European Journal of Operational Research (2016), http://dx.doi.org/10.1016/j.ejor.2015.12.050

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 (2) and denote the optimal objective function value as W^{LP}

Do

Solve model (7) to get the optimal objective value W^* and the optimal solution as ψ^*

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

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

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

Else

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

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

Step 3: Solve LP relaxation of model (7) to get optimal objective value W^* and optimal solution ψ^*

If problem is infeasible, then

terminate, the original problem is infeasible.

Else

Go to Step 4.

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

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

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

Go to Step 3.

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

The optimal angle profile is ψ^*; terminate.

Else

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

Go to Step 3.

Fig. 1. Traditional Benders decomposition.

The procedure is similar to the branch-and-check algorithm of Thorsteinsson (2001) which also can be viewed an extension of traditional Benders when the feasible region contains nonlinear constraints. His algorithm solves a relaxation of the original problem by branch and bound but in the search tree, subproblems are solved to check the feasibility of solutions and correctness of the objective function value. The check is based on the cuts added to the relaxed master problem, which are similar to Benders optimality cuts.

We also adopt the two-phase approach in our implementation of the modern Benders decomposition. The same first phase (Step 2 in Fig. 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 there are generally too many constraints to explicitly enumerate and check, a separation procedure is necessary to identify violations. Fig. 2 depicts the flowchart for the separation procedure as well as the logic for updating the incumbent.

In particular, given any candidate solution $\hat{\psi}$ and its objective function value $\hat{W}$ in the restricted master problem, model (3) is solved with $\psi = \hat{\psi}$ to identify violated Benders cuts, if any. If the resulting problem is unbounded, then the feasibility cut corresponding to $\hat{\psi}$ is added to the current restricted master problem. This represents the separation procedure. If the resulting solution is bounded and $\hat{W} = h(\hat{\psi})$, then the incumbent is updated as in traditional branch and bound. Otherwise, the indicated optimality cut is added to the master problem.

After solving the restricted master problem in the modern Benders decomposition, the optimal solution satisfies all constraints (4b)–(4c) although 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 because it exploits

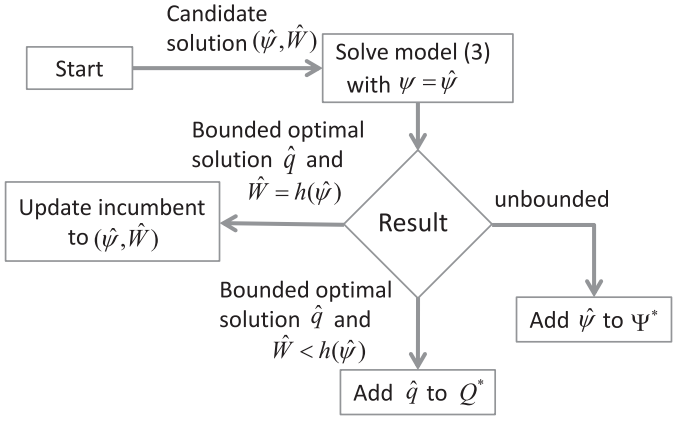


Fig. 2. Logic for updating the incumbent solution in modern Benders decomposition.

all the information gathered in previous runs rather than discarding it (Rubin, 2011). 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 is 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-edged sword: additional cuts increase the size of the problem and thus

Please cite this article as: S. Lin et al., Benders decomposition and an IP-based heuristic for selecting IMRT treatment beam angles, European Journal of Operational Research (2016), http://dx.doi.org/10.1016/j.ejor.2015.12.050

Procedure_LP_rounding_heuristic

Input: U^* = maximum number of angles to eliminate
Output: Feasible solution to model (1)
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 $a^* = \arg\min\{\psi_a^* : a \notin A^*\}$
Put $A^* \leftarrow A^* \cup \{a^*\}$
Step 2: Solve the reduced IP with $\psi_a = 0$ for all $a \in A^*$

Fig. 3. LP rounding heuristic.

increase its difficulty, but they can also help to improve its lower bound (Rei, Cordeau, Gendreau, & Soriano, 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. Thus, a natural heuristic is to eliminate some angles so that the problem size is manageable. 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 angles not likely to be in the optimal solution until the remaining problem is manageable as an IP. Specifically, we solve the LP relaxation and eliminate the angle whose variable take the smallest value in each iteration. Ties are broken arbitrarily. The process terminates when a given number of angles, denoted by U^*, is removed. At that point, the reduced IP is solved with only the remaining set of angles. Fig. 3 describes the procedure.

The size of the feasible region that remains after eliminating U^* angles depends on the magnitude of U^*. However, because the heuristic only eliminates the least used angles in the LP solution, it is likely that the reduced feasible region still contain good, if not the optimal, solution, especially when U^* is relatively small. Thus, by selecting an appropriate value for U^*, the heuristic is expected to generate a good solution within reasonable amount of time.

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 feasible solution in hopes of finding a better one. In this section, we describe the local branching heuristic we implemented to improve the solutions found by the rounding heuristic.

Let us start with the neighborhood used. One key observation of the solution space is that there exists an optimal solution to model (1) in which exactly η angles are chosen. To see this, assume there is an optimal solution ψ^* with $n < \eta$ angles. We can always find an equivalent solution with n angles in solution ψ^* and an additional $\eta - n$ angles whose beamlets carry zero intensity. Our computational experience confirms that local optima always contain η angles. Given a solution $\bar{\psi}$ with η angles selected, Fischetti and Lodi define the k-OPT neighborhood of $\bar{\psi}$ as the set of feasible angles that satisfy the additional local branching constraint

$$\Delta(\psi, \bar{\psi}) = \sum_{a \in A} (1 - \bar{\psi}_a)\psi_a \le k \tag{8}$$

Intuitively, constraint (8) permits at most k of the η 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 $\Delta(\psi, \bar{\psi}) \le k$ on the left branch and $\Delta(\psi, \bar{\psi}) \ge k+1$ 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 $R = \emptyset$, $L = \emptyset$, 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 additional sets of constraints $L = \{\Delta(\psi, \psi^*) \le k\}$ and the existing sets R. 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 the solution is no better than the incumbent, 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 $\Delta(\psi, \psi^*) \le k$ 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 Fig. 4. For simplicity, we will use the function $h(\psi)$ discussed in Section 4.1 to denote the objective function value given by solution angle ψ.

6. Computational results

Three data sets associated with prostate tumors were used to test the effectiveness of our algorithms. Parameter values 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. Instance P2-36 has many more voxels for the

Please cite this article as: S. Lin et al., Benders decomposition and an IP-based heuristic for selecting IMRT treatment beam angles, European Journal of Operational Research (2016), http://dx.doi.org/10.1016/j.ejor.2015.12.050

Procedure_Local_branching

Input: Initial feasible solution ψ^0

τ = time limit to solve each subproblem
$k_{\min}$ = minimum neighborhood parameter
$k_{\max}$ = maximum neighborhood parameter

Output: Improved feasible solution ψ^*

Step 0: Iteration count $m = 1$
Set of right branch constraints $R = \varnothing$
$\psi^* = \psi^0$

Step 1: set $k = k_{min}$

Step 2: Solve model (2) with additional sets of constraints R and $L = \{\Delta(\psi, \psi^*) \le k\}$; set time limit to τ and denote the resulting solution, if any, as ψ^m
If the problem is solved to optimality within τ, then
add constraint $\Delta(\psi, \psi^*) \ge k + 1$ to R

Step 3: If $h(\psi^*) > h(\psi^m)$, then
//better solution is found
Update $\psi^* = \psi^m$
Put $m \leftarrow m + 1$
Go to Step 1
Else if the problem is solved to optimality in Step 2
//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^*.

Fig. 4. Local branching logic.

Table 1
Summary of data sets.

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 the PTV	1000	1000	4005	4005	5245	5245
# of voxels for bladder (OAR)	10,603	10,603	7850	7850	–	–
# of voxels for rectum (OAR)	5848	5848	5719	5719	1936	1936
# of positive d_{vb}	1.27E7	1.95E7	1.10E7	3.31E7	2.88E6	8.72E6
# of beamlets[a]	1419	2875	2500	7486	885	2672

[a] Number of beamlets after problem reduction discussed in Section 3.

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 the PTV but the smallest number of voxels for OARs. For convenience, the bladder is the only OAR considered in instance P3-36. This is a well-studied test case that is known to be difficult to optimize within clinical treatment specifications even without the bladder.

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 indicated in constraints (1b), the total dose D_v delivered to each voxel v is the weighted sum of d_{vb} with beamlet intensity w_{ab} as the weight. Thus, if the number of positive d_{vb} is reduced, it is reasonable to expect that the solution space of D_v is also reduced. Accordingly, the number of positive d_{vb} can affect the feasible region of the dose value applied to each voxel. Table 2 lists the dose-volume requirements for all organs.

Table 2
Dose volume requirements.

Organ	Constraints
PTV	Prescription: 76 Gy
PTV	Volume receiving at least the prescription dose ≥ 95 percent
Rectum	Volume receiving doses higher than 60 Gy: ≤40 percent
Rectum	Volume receiving doses higher than 70 Gy: ≤25 percent
Bladder	Volume receiving doses higher than 70 Gy: ≤25 percent

All algorithms were implemented in JAVA and run under Ubuntu Linux on a Dell Poweredge T610 workstation with two 6-core hyperthreading 3.33-gigahertz Xeon processors and 24 gigabytes of memory. CPLEX 12.4 was used as the MIP solver. In the computations, we followed the convention in Lim et al. (2007)

Table 3
Equidistant-angle solutions and runtimes for 12-angle instances.

Instance	Equidistant-angle objective value	Optimal objective value	Time (minutes)		
			CPLEX	Traditional Benders	Modern Benders
P1-12	0.06049	0.03060	3	7	8
P2-12	0.07073	0.01210	20	22	28
P3-12	0.19763	0.11480	23	14	10

Table 4
Comparison of different solution methods.

Instance	Metric	Method					
		CPLEX	Modern Benders	Traditional Benders	LP-rounding heuristic	Two-stage heuristic	SA-LNS[a]
P1-36	Time (min)	300	300	300	5	173	106
	Obj. val.	0.03036	0.03036	0.03046	0.03048	0.03036	0.03044
	Percent of gap[b]	0	0	0.33	0.40	0	0.26
P2-36	Time (min)	300	300	300	51	284	298
	Obj. val.	0.01183	0.01167	0.01170	0.01158	0.01158	0.01151
	Percent of gap[b]	2.78	1.39	1.65	0.61	0.61	0
P3-36	Time (min)	300	300	300	71	280	19
	Obj. val.	0.12151	0.11679	0.11830	0.11411	0.11411	0.11411
	Percent of gap[b]	6.49	2.35	3.67	0	0	0

[a] The two phase framework discussed in Lim et al. (2014).
[b] 100 percent × (current obj. val./best obj. val. among all methods − 1).

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 $v \in T$. Because the P3 instances have a large PTV, we used the following parameter values to better control cold and hot spots: $\theta_U = 1.05$, $\theta_L = 0.97$, $L_v = 0.96$, $U_v = 1.15$ for all $v \in T$. In all instances, we used $\eta = 6$, $\phi_i = 0.3$ for all $i \in O$, $\lambda_t^+ = \lambda_t^- = \lambda_s = 1$. We chose these values to be consistent with our past work (Lim & Cao, 2012, Lim et al., 2008, Lim et al., 2014). With respect to η, the results given below confirm that we are able to obtain clinically acceptable treatment plans as defined by the dose volume histogram (DVH0 requirements when $\eta = 6$.

In our initial testing, we discovered that the largest problems that we could solve to optimality with CPLEX and Benders decomposition within a reasonable amount of time, say, 30 minutes, where those with 12 angles. This observation provided the motivation for the LP rounding heuristic which reduces the 36-angle problems to 12-angle problems and then calls CPLEX to find the corresponding integer solution. Table 3 highlights these results for P1-12, P2-12 and P3-12, which have $|A| = 12$. CPLEX and both Benders decomposition algorithms converged within 30 minutes giving optimal objective function values. The solution times confirm that a 12-angle problem can be solved quickly with commercial software. Although not as efficient as CPLEX for instances P1-12 and P2-12, either Benders approach can still find optimal solutions within 28 minutes.

As a point of comparison, we also ran these instances for the case where the six angles are equidistant from each other starting at 0 degree. The results in Table 3 show that the corresponding objective function values are far from optimal. With carefully selected angles, the optimal solution provides better objective functions since it can (i) control the hot and cold spots for the PTV and (ii) control the hot spots for OARs.

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 (we tried to warm-start CPLEX with the LP rounding solution but little improvement in solution quality was observed within 5 hours). 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. Finally, given that the 12-angle problems are usually well-solved, we used $U^* = 24$ in the LP-based heuristic.

Table 5
Statistics associated with the MIP formulation for 36-angle instances.

Statistics	P1-36	P2-36	P3-36
Optimality gap (percent)[a]	2.84	14.59	75.64
# of variables	20,364	25,098	9891
# of constraints	22,327	33,071	20,344
LP relaxation time (seconds)	115	1126	228

[a] Optimality gap = MIP gap when CPLEX terminates.

Table 4 compares the solution times and solution quality of the different methods, including the SA-LNS approach discussed in Lim et al. (2014). Table 5 summarizes the statistics associated with the MIP models. One key observation from the computations is that problem instances become more difficult as the number of voxels in the PTV grows. For problems with small PTV (e.g., instance P1-36), CPLEX can solve model (2) 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 the OARs does not have a noticeable impact.

Another implication of the results 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 the PTV than P2-36, has a much larger optimality gap (74.41 percent) than P2-36 (14.59 percent). 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 the PTV and the magnitude of d_{vb} are key indicators of problem difficulty.

Table 4 also indicates that modern Benders decomposition can find better feasible solutions than traditional Benders decomposition. Moreover, for instances P2-36 and P3-36, both methods generated a better feasible solution with a smaller optimality gap than CPLEX, while for P1-36 only modern Benders outperformed CPLEX.

Please cite this article as: S. Lin et al., Benders decomposition and an IP-based heuristic for selecting IMRT treatment beam angles, European Journal of Operational Research (2016), http://dx.doi.org/10.1016/j.ejor.2015.12.050

Table 6
Comparison between traditional and modern Benders decomposition.

Instance	First stage time (minutes)	Modern Benders		Traditional Benders	
		Opt. gap[a]	# of cuts	Opt. gap[a]	# of cuts
P1-36	5	8.26	11,824	6.75	1669
P2-36	140	16.09	1170	13.13	750
P3-36	58	74.69	6791	68.75	721

[a] Opt. gap = 100 percent × (Benders upper bound − Benders lower bound)/Benders lower bound.

It is also worth noting that neither Benders approach required feasibility cuts.

Table 6 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 optimality gap associated with the modern approach, obtained by solving the restricted master problem, is not as small as the optimality gap accompanying 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 feasible solutions rather than closing the optimality gap, the modern approach is the better choice.

Returning to Table 4, we see that the 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. By design, the LP rounding heuristic involves solving an IP formulation with 12 angles in Step 2. The corresponding solutions were seen to be better than those found by solving the problem with 12 equidistant angles, as reported in Table 3. This result suggests that the information obtained from the LP relaxation is helpful in selecting good candidate angles.

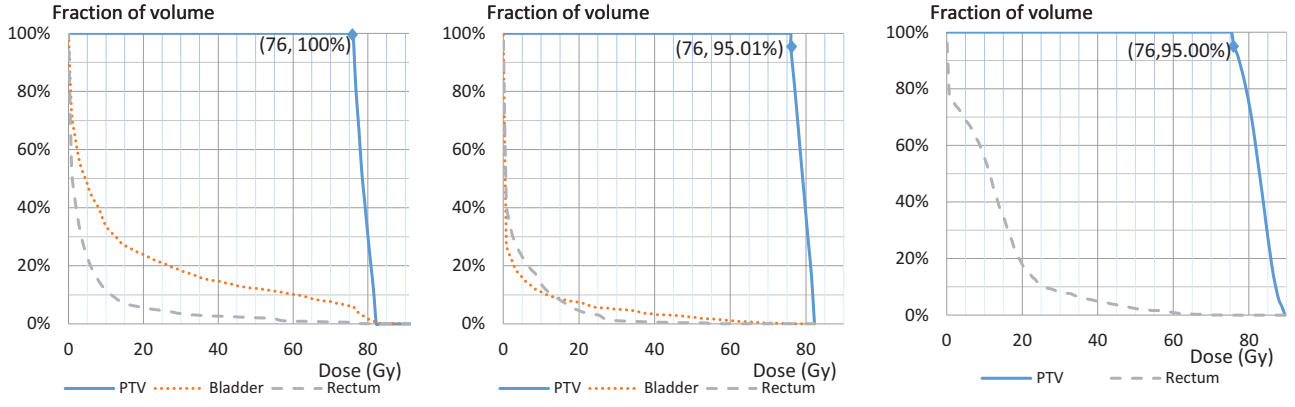
Of the newly proposed methods, the two-stage heuristic (LP-rounding with local branching) provided the best results in the least amount of time. Nevertheless, the computations still took much longer than desired, especially for P2-36. As the problem size grows, especially with respect to the number of voxels in the 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.

Table 4 also compares our methods with the two-stage SA-LNS approach discussed in Lim et al. (2014). Similar to our two-stage heuristic, SA-LNS first develops an initial solution but with simulated annealing, and then improves it using local neighborhood search. Their computational work demonstrated the general superiority of SA-LNS over the other methods tested, including branch and bound, genetic algorithms, and branch and prune.

Although our two-stage heuristic may require longer runtimes than SA-LNS for some instances, Table 3 shows that solution quality is comparable: LP rounding with local branching provides a better solution for instance P1-36, SA-LNS provides a better solution for P2-36, while both provide the same solution for P3-36. With respect to Benders, SA-LNS provides better solutions than the traditional approach for all three instances and better solutions than the modern approach for P2-36 and P3-36. One conclusion that can be drawn from these results is that specifically designed methods like our two-stage heuristic and SA-LNS are likely to outperform Benders decomposition, at least when it comes to developing IMRT treatment plans.

As previously mentioned, our LP-rounding heuristic without local branching can provide good solutions quickly. To further demonstrate the quality of the generated treatment plans, we have plotted the DVH for the PTV and each OAR obtained from the LP rounding heuristic in Fig. 5. The horizontal axis represents the dose value and the vertical axis represents the fraction of volume. The DVH contains one curve for the PTV and one for each of the OARs. Each point on the curve specifies the percentage of volume (in the corresponding OAR or the PTV) that receives a dose greater than a given value. For example, point (23.51, 21.83 percent) on the curve for the bladder (OAR) in Fig. 5a indicates that 21.83 percent of voxels in the bladder receive a dose more than 23.51 Gy. Note that the DVH plots for the other methods tested look nearly identical to those in Fig. 5 and are available from the authors.

Fig. 5 shows three DVHs corresponding to instances P1-36, P2-36 and P3-36. Table 7 summarizes the key metrics for the treatment plans. As can be seen in Fig. 5a and Table 7, all PTV voxels receive 76 Gy or higher and the dose for OARs P1-36 is well within the prescribed bounds: only 10.31 percent of the voxels in the bladder receive a dose higher than 70 Gy, and only 0.89 percent



(a) Instance P1-36 (b) Instance P2-36 (c) Instance P3-36

Fig. 5. Dose volume histograms for LP-rounding heuristic solutions.

Please cite this article as: S. Lin et al., Benders decomposition and an IP-based heuristic for selecting IMRT treatment beam angles, European Journal of Operational Research (2016), http://dx.doi.org/10.1016/j.ejor.2015.12.050

Table 7
Summary of key metrics for LP-rounding heuristic solutions.

	P1-36	P2-36	P3-36
Fraction of bladder with dose greater than 70 Gy	10.31 percent	0.25 percent	–
Fraction of rectum with dose greater than 60 Gy	0.89 percent	0 percent	0.88 percent
Fraction of the PTV with dose greater than 76 Gy	100 percent	95.01 percent	95.00 percent

of the voxels in the rectum receive a dose value higher than 60 Gy. Similarly, Fig. 5b shows that only 0.25 percent of the voxels in the bladder receive a dose higher than 70 Gy and none of the voxels in the rectum receive a dose higher than 60 Gy in the solution for instance P2-36. For the PTV, more than 95 percent of its voxels have a dose value higher than 76 Gy.

For P3-36, because the number of voxels in the bladder is 0, Fig. 5c only plots the DVH for the PTV and rectum. The figure shows that 0.88 percent of the voxels in the rectum receive doses higher than 60 Gy. Also, 95 percent of the voxels in the PTV receive a dose value more than 76 Gy. This indicates that the majority of the radiation is delivered to the PTV while largely sparing the OARs.

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; on the other hand, traditional Benders generates cuts with higher quality and thus produces a superior lower bound.

The best results were obtained with the LP rounding heuristic in conjunction with local branching. By itself, the former was seen to provide good solutions quickly. When runtimes are critical, the best compromise would be to just use the LP rounding heuristic. For future research, it might be beneficial to investigate ways of strengthening Benders feasibility cuts or to identify stronger Benders cuts. So doing would increase the lower bound provided by the restricted master problem and hence speed convergence. Since Benders decomposition has performed well on other types of problems, there is likely room for improvement in our implementation. In addition, it may be worthwhile experimenting with other local search procedures to improve the solutions generated by the LP rounding heuristic.

Acknowledgment

The authors are grateful for the valuable comments and suggestions made by the three anonymous referees.

References

Aleman, D. M., Ghaffari, H. R., Mišić, V. V., Sharpe, M. B., Ruschin, M., & Jaffray, D. A. (2013). Optimization methods for large-scale radiotherapy problems. In P. M. Pardalos, P. G. Georgiev, P. J. Papajorgji, & B. Neugaard (Eds.), *Systems analysis tools for better health care delivery* (pp. 1–20). New York: Springer.
Aleman, D. M., Kumar, A., Ahuja, R. K., Romeijn, H. E., & Dempsey, J. F. (2008). Neighborhood search approaches to beam orientation optimization in intensity modulated radiation therapy treatment planning. *Journal of Global Optimization, 42*(4), 587–607.
Aleman, D. M., Romeijn, H. E., & Dempsey, J. F. (2009). A response surface approach to beam orientation optimization in intensity-modulated radiation therapy treatment planning. *INFORMS Journal on Computing, 21*(1), 62–76.
Bertsimas, D., Cacchiani, V., Craft, D., & Nohadani, O. (2013). A hybrid approach to beam angle optimization in intensity-modulated radiation therapy. *Computers & Operations Research, 40*(9), 2187–2197.
Binato, S., Pereira, M. V. F., & Granville, S. (2001). A new Benders decomposition approach to solve power transmission network design problems. *IEEE Transactions on Power Systems, 16*(2), 235–240.
Bortfeld, T., & Schlegel, W. (1993). Optimization of beam orientations in radiation therapy: Some theoretical considerations. *Physics in Medicine and Biology, 38*(2), 291–304.
Cordeau, J.-F., Stojković, G., Soumis, F., & Desrosiers, J. (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.
IBM (2011), *IBM ILOG CPLEX V12.4: User's manual for CPLEX: Vol. 46* p. p. 157. International Business Machines Corporation.
Djajaputra, D., Wu, Q., Wu, Y., & Mohan, R. (2003). Algorithm and performance of a clinical IMRT beam-angle optimization system. *Physics in Medicine and Biology, 48*(19), 3191–3212.
Ehrgott, M., Güler, Ç., Hamacher, H. W., & Shao, L. (2008). Mathematical optimization in intensity modulated radiation therapy. *4OR, 6*(3), 199–262.
Ezzell, G. A. (1996). Genetic and geometric optimization of three-dimensional radiation therapy treatment planning. *Medical Physics, 23*(3), 293–305.
Fischetti, M., & Lodi, A. (2003). Local branching. *Mathematical Programming, 98*(1–3), 23–47.
Lee, C. H. J., Aleman, D. M., & Sharpe, M. B. (2011). A set cover approach to fast beam orientation optimization in intensity modulated radiation therapy for total marrow irradiation. *Physics in Medicine and Biology, 56*(17), 5679–5695.
Lee, E. K., Fox, T., & Crocker, I. (2000). Optimization of radiosurgery treatment planning via mixed integer programming. *Medical Physics, 27*(5), 995–1004.
Lei, J., & Li, Y. (2009). An approaching genetic algorithm for automatic beam angle selection in IMRT planning. *Computer Methods and Programs in Biomedicine, 93*(3), 257–265.
Li, Y., Yao, D., Yao, J., & Chen, W. (2005). A particle swarm optimization algorithm for beam angle selection in intensity-modulated radiotherapy planning. *Physics in Medicine and Biology, 50*(15), 3491–3514.
Lim, G. J., Ferris, M. C., Wright, S. J., Shepard, D. M., & Earl, M. A. (2007). An optimization framework for conformal radiation treatment planning. *INFORMS Journal on Computing, 19*(3), 366–380.
Lim, G. J., & Cao, W. (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., Choi, J., & Mohan, R. (2008). Iterative solution methods for beam angle and fluence map optimization in intensity modulated radiation therapy planning. *OR Spectrum, 30*(2), 289–309.
Lim, G. J., Kardar, L., & Cao, W. (2014). A hybrid framework for optimizing beam angles in radiation therapy planning. *Annals of Operations Research, 217*(1), 357–383.
Magnanti, T. L., & Wong, R. T. (1981). Accelerating Benders decomposition: Algorithmic enhancement and model selection criteria. *Operations Research, 29*(3), 464–484.
McDaniel, D., & Devine, M. (1977). A modified Benders' partitioning algorithm for mixed integer programming. *Management Science, 24*(3), 312–319.
Milano, M. T., Garofalo, M. C., Chmura, S. J., Farrey, K., Rash, C., Heimann, R., & Jani, A. B. (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.
Mišić, V. V., Aleman, D. M., & Sharpe, M. B. (2010). Neighborhood search approaches to non-coplanar beam orientation optimization for total marrow irradiation using IMRT. *European Journal of Operational Research, 205*(3), 522–527.
Rei, W., Cordeau, J.-F., Gendreau, M., & Soriano, P. (2009). Accelerating Benders decomposition by local branching. *INFORMS Journal on Computing, 21*(2), 333–345.
Price, S., Golden, B., Wasil, E., & Zhang, H. H. (2014). Data mining to aid beam angle selection for intensity-modulated radiation therapy. In *Proceedings of the 5th ACM conference on bioinformatics, computational biology, and health informatics, September 2014* (pp. 351–359). ACM.
Rocha, H., Dias, J. M., Ferreira, B. C., & Lopes, M. C. (2013). Beam angle optimization for intensity-modulated radiation therapy using a guided pattern search method. *Physics in Medicine and Biology, 58*(9), 2939.
Romeijn, H. E., Ahuja, R. K., Dempsey, J. F., Kumar, A., & Li, J. G. (2003). A novel linear programming approach to fluence map optimization for intensity modulated radiation therapy treatment planning. *Physics in Medicine and Biology, 48*(21), 3521–3542.
Rowbottom, C. G., Nutting, C. M., & Webb, S. (2001). Beam-orientation optimization of intensity-modulated radiotherapy: Clinical application to parotid gland tumours. *Radiotherapy and Oncology, 59*(2), 169–177.

Please cite this article as: S. Lin et al., Benders decomposition and an IP-based heuristic for selecting IMRT treatment beam angles, European Journal of Operational Research (2016), http://dx.doi.org/10.1016/j.ejor.2015.12.050

Rowbottom, C. G., Webb, S., & Oldham, M. (1999). Beam-orientation customization using an artificial neural network. *Physics in Medicine and Biology, 44*(9), 2251–2262.

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

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

Saka, B., Rardin, R. L., & Langer, M. P. (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., Rardin, R. L., Langer, M. P., & Dink, D. (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.

Taşkın, A. A., Sasaki, S., Segawa, K., & Ando, Y. (2012). Manifestation of topological protection in transport properties of epitaxial Bi 2 Se 3 thin films. *Physical Review Letters, 109*(6).

Thorsteinsson, E. S. (2001). Branch-and-check: A hybrid framework integrating mixed integer programming and constraint logic programming. In T. Walsh (Ed.), *Lecture notes in computer science,* Proceeding of the seventh international conference on principles and practice of constraint programming *(*CP 2001*)* (pp. 16–30). Springer, Germany.

Wang, C., Dai, J., & Hu, Y. (2003). Optimization of beam orientations and beam weights for conformal radiotherapy using mixed integer programming. *Physics in Medicine and Biology, 48*(24), 4065–4076.

Yarmand, H., Winey, B., & Craft, D. (2013). Guaranteed epsilon-optimal treatment plans with the minimum number of beams for stereotactic body radiation therapy. *Physics in Medicine and Biology, 58*(17), 5931–5944.

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

Zhang, H. H., Shi, L., Meyer, R., Nazareth, D., & D'Souza, W. (2009). Solving beam-angle selection and dose optimization simultaneously via high-throughput computing. *INFORMS Journal on Computing, 21*(3), 427–444.

Please cite this article as: S. Lin et al., Benders decomposition and an IP-based heuristic for selecting IMRT treatment beam angles, European Journal of Operational Research (2016), http://dx.doi.org/10.1016/j.ejor.2015.12.050