The Determination of Optimal Treatment Plans for Volumetric Modulated Arc Therapy (VMAT)

Pınar Dursun[a], Z. Caner Taşkın[a], İ. Kuban Altınel[a,*]

[a] *Department of Industrial Engineering, Boğaziçi University, 34342, Bebek, İstanbul, Turkey*

Abstract

The success of radiation therapy depends on the ability to deliver the proper amount of radiation to cancerous cells while protecting healthy tissues. As a natural consequence, any new treatment technology improves quality standards concerning primarily this issue. Similar to the widely used Intensity Modulated Radiation Therapy (IMRT), the radiation resource is outside of the patient's body and the beam is shaped by a multi-leaf collimator mounted on the linear accelerator's head during the state-of-the-art Volumetric Modulated Arc Therapy (VMAT) as well. However, unlike IMRT, the gantry of the accelerator may rotate along one or more arcs and deliver radiation continuously. This property makes VMAT powerful in obtaining high conformal plans in terms of dose distribution; but the apertures are interdependent and optimal treatment planning problem cannot be decomposed into simpler independent subproblems as a consequence. In this work, we consider optimal treatment planning problem for VMAT. First, we formulate a mixed-integer linear program minimizing total radiation dose intensity subject to clinical requirements embedded within the constraints. Then, we develop efficient solution procedures combining Benders decomposition with certain acceleration strategies. We investigate their performance on a large set of test instances obtained from an anonymous real prostate cancer data.

Keywords: `Integer programming; Benders decomposition; radiation therapy; VMAT; algorithms`

1. Introduction

Radiation therapy (radiotherapy) sends high-energy particles or waves on to cancerous tissues in order to damage the deoxyribonucleic acid (DNA) of cancer cells, which destroys their ability to reproduce. Radiation can also harm healthy cells, which can repair themselves unless they are exposed to doses beyond their tolerance limits. Hence, the success of the treatment depends on the ability to deliver the proper amount of radiation to the malignant region while sparing healthy tissues so that they are exposed a minimal amount of radiation.

External-beam radiation and internal radiation therapy (brachytherapy) are two modes of radiotherapy. In the former one, a machine delivers radiation to the patient from outside the body; on the other hand radiation sources

[*] Corresponding author. Phone: + 90 (212) 359 6407, Fax: + 90 (212) 265 1800

Email addresses: `pinar.dursun@boun.edu.tr` (Pınar Dursun), `caner.taskin@boun.edu.tr` (Z. Caner Taşkın), `altinel@boun.edu.tr` (İ. Kuban Altınel)

like implants or liquids are placed inside the patient's body in the latter. Three-Dimensional Conformal Radiation Therapy (3D-CRT), Image Guided Radiation Therapy (IGRT), Intensity Modulated Radiation Therapy (IMRT), Tomotherapy, and Volumetric Modulated Arc Therapy (VMAT) are being tested and applied forms of external-beam radiation therapy.

External-beam radiation therapy process starts with the determination of tumors and surrounding normal structures after the diagnosis of the patient with cancer. The oncologist contours the cancerous target volumes (TVs) and surrounding organs at risk (OARs) on the computed tomography (CT) scans of the patient and prescribes the radiation doses best conforming to TVs and OARs. There may be more than one TV with different dose requirements as well as OAR depending on the cancer type and patient's anatomy. The treatment to be applied to the patient is planned by medical physicists.

1.1. IMRT and VMAT

In both IMRT and VMAT, a *linear accelerator* (see Figure 1) rotates around the patient's body and sends high-energy beams from different angles by keeping the cancer volume on the target. The gantry of the linear accelerator is generally equipped with a *multileaf collimator* (MLC), which consists of a number of parallel metal leaf pairs. The leaves can move horizontally and shape the opening that the radiation beam passes through. Namely, they can block some fraction of the beam (see Figure 2). In this way, the conformity of dose distribution to the planning target volume (tumor plus some margin) and normal tissue sparing is much superior compared to earlier techniques (Cambazard et al., 2012). Hovewer, IMRT and VMAT requires higher amount of radiation (in monitor units, MUs) to deliver a given fraction size compared with 3D-CRT (Teoh et al., 2011; Palma et al., 2008). The increase in MU causes higher integral body dose, which increases the risk of secondary malignancy (Hall and Wuu, 2003). Although IMRT has been used very extensively in radiation therapy since 1990s (Ehrgott et al., 2010), VMAT is the state-of-the-art technology. In VMAT, not only the MLC's leaves can move without stopping, but also the radiation delivery is continuous during the rotation of the gantry. Thus, high conformal dose distributions are obtained by generating and delivering less radiation (Teoh et al., 2011), and treatment times become significantly shorter (Otto, 2008; Peng et al., 2012). On the other hand, there are typically only a few discrete angles (5-9) in IMRT plans (Romeijn et al., 2006) (see Figure 3). Furthermore, the linear accelerator stops delivering radiation while moving its gantry between different beam angles in both dynamic (*sliding window technique*) and static (*step-and-shoot technique*) types of IMRT, and during the change of MLC shapes at a beam angle in the latter one (Ehrgott et al., 2010). In addition to the clinical benefits of delivering less radiation to the patient, there are several other advantages of short treatments. The discomfort of patients and the risk of negative effects that may result from patient movements decrease. Also, it is possible to treat more people since resource utilization becomes more efficient (Peng et al., 2012).

Treatment planning processes for IMRT and VMAT have similarities, which is not surprising since they are closely related technologies. There are three main phases in IMRT planning. The first phase deals with the *beam angle optimization* (BAO), where the number and orientation of the beam angles for irradiation are determined. The second

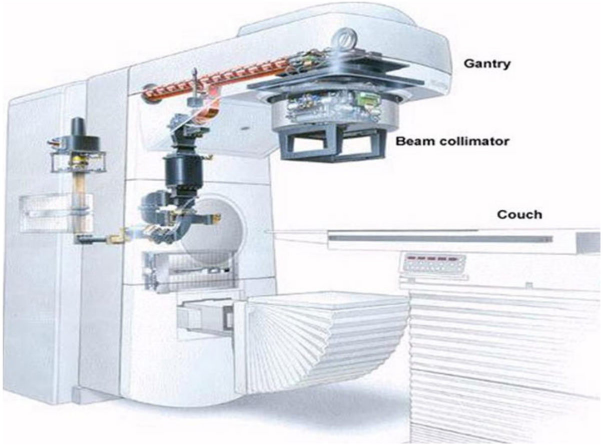


Figure 1: A linear accelerator (de Araújo Montagno and Sabbatini, 1997)

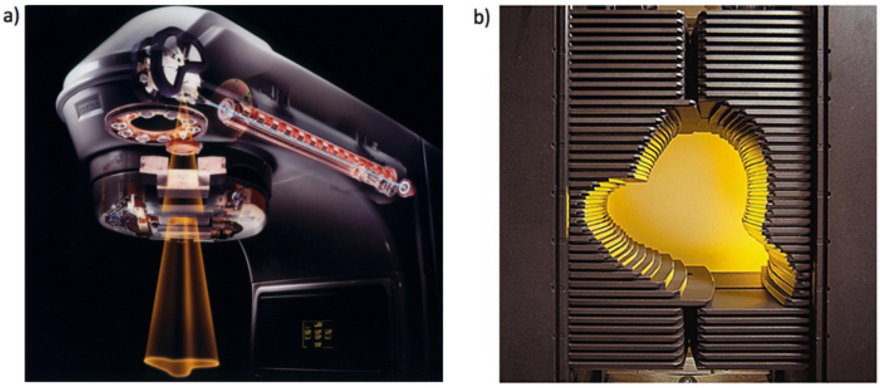


Figure 2: (a) A linear accelerator (Varian, 2017a) (b) A multileaf collimator system (Varian, 2017b).

phase is the *fluence map optimization* (FMO) (or *intensity problem*), in which a *fluence map* is obtained for each of the beam angles determined in the first phase. A fluence map is represented by a two-dimensional nonnegative integer matrix that gives the radiation intensity profile. The third phase is the *MLC leaf sequencing* (MLS) (or *realization problem*): a given fluence map is decomposed into a number of disjoint MLC openings, called *apertures*, and their radiation intensities. In other words, deliverable radiation beams are obtained in the last phase and the union of these beams realizes the corresponding fluence maps. Each of the three phases of IMRT planning can be dealt independently and solved sequentially, or the integration of two consecutive phases can be considered simultaneously. The possibility of handling the three phases independently makes IMRT planning relatively simpler.

Unlike IMRT, there is a large number of evenly spaced discrete beam angles on a co-planar arc (i.e. trajectory) of the gantry in VMAT, since the radiation delivery is continuous. The term *control point* is often used interchangeably with beam angle since aperture shape, dose rate, and gantry speed are controlled at each one of the specified beam

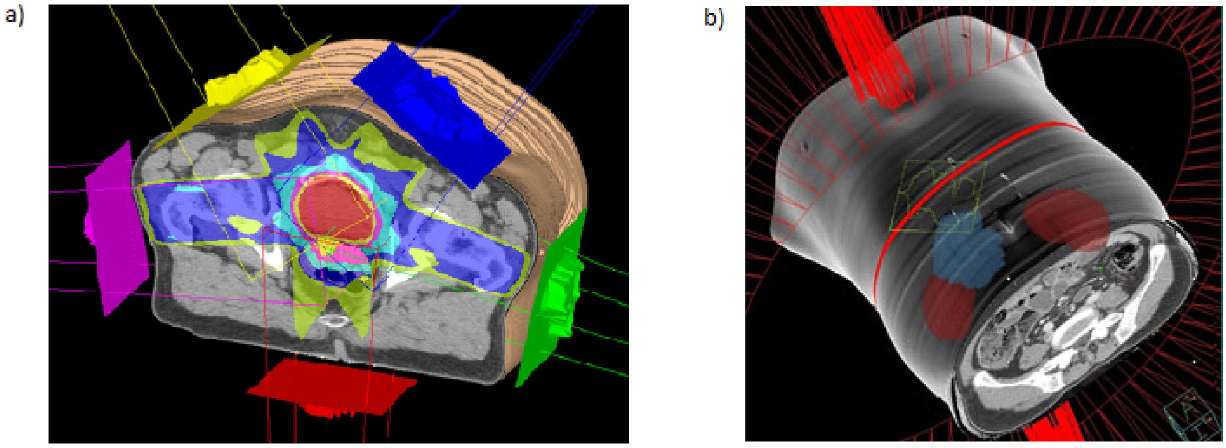
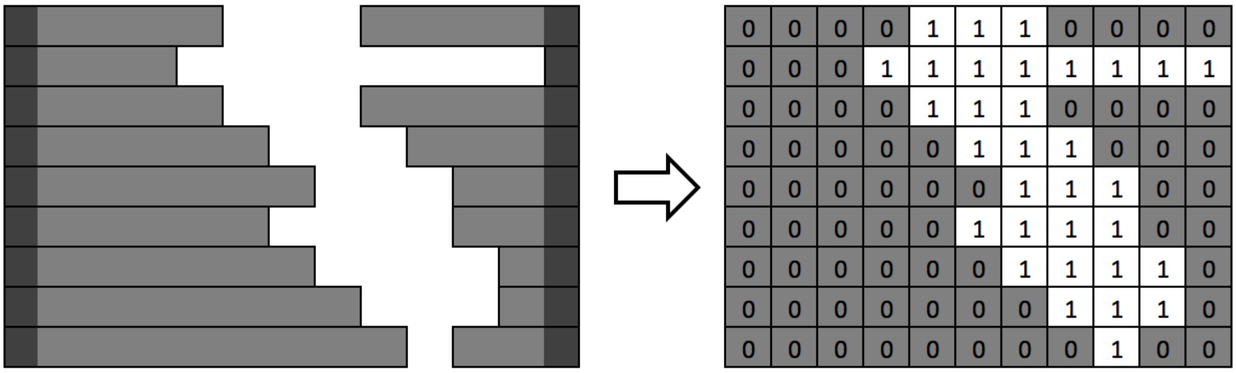


Figure 3: (a) IMRT (SIMBALLC, 2013) (b) VMAT (VCU Massey Cancer Center, 2017)

angles. An aperture can be represented by means of a binary matrix. Each entry of the matrix is called a *beamlet*. In Figure 4, an example of an aperture and its binary matrix representation is illustrated. Note that the two dark columns on the left figure are the home positions of the leaves and if a beamlet is open then the corresponding entry of the matrix is 1; it is 0 otherwise. Each aperture must satisfy the *consecutive ones property* of the MLC: there can be at most one open beamlet chain in a row of an aperture. During the rotation of the gantry the leaves can move and change the shape of the aperture. However, this movement is limited and depends on the speed of the gantry. The apertures of neighboring control points are similar and cannot be handled independently because of the limitations on the movement; and it is not possible to decompose VMAT planning problem into independent subproblems. As a result, designing a VMAT treatment plan is significantly harder since total number of control points is large and the adjacent ones are interconnected. Even if the delivery time is fixed, the resulting problem is a large-scale nonconvex optimization problem, which makes VMAT planning a challenging goal that requires much more computational effort than IMRT planning (Craft et al., 2012).

A treatment plan must satisfy the treatment related requirements, such as the dose prescriptions, and the mechanical limitations of the linear accelerator and MLC system. In order to calculate dose distributions on the structures, the body of the patient is discretized into small cubes called *voxels* (see Figure 5 for an example) using CT scans, and the amount of radiation each of these voxels absorbs during the treatment is calculated. Absorbed radiation amounts can be obtained if, how many Gray (Gy) a voxel absorbs when it is exposed to one MU of radiation from a beamlet at a control point, is known. It is possible to determine these amounts using by one of the dose calculation algorithms such as Pencil Beam Algorithm (PBA) or Analytical Anisotropic Algorithm (AAA) (Oelkfe U., 2006). In this study we use the dose influence matrix (**D**) of a real prostate cancer case, which is obtained from Craft et al. (2014). The entries of **D** are in Gy per MU (Gy/MU).



0	0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	1	1	1	1	1
0	0	0	0	1	1	1	0	0	0	0
0	0	0	0	0	1	1	1	0	0	0
0	0	0	0	0	0	1	1	1	0	0
0	0	0	0	0	1	1	1	1	0	0
0	0	0	0	0	0	1	1	1	1	0
0	0	0	0	0	0	0	1	1	1	0
0	0	0	0	0	0	0	0	1	0	0

Figure 4: An aperture and its binary matrix representation

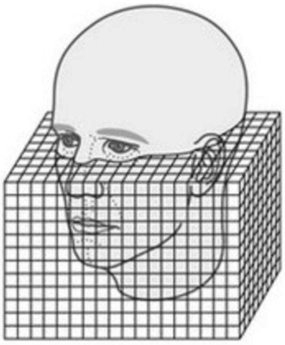


Figure 5: A voxel resolution for a head (Pocket Dentistry, 2015)

1.2. Motivation and Contributions

As mentioned previously, the main advantage of VMAT is its ability to deliver radiation continuously, which causes a very thin slicing of the gantry's rotational arc at many beam angles increasing the number of control points considerably. As a consequence, adjacent ones become very close, and this makes them interdependent with respect to the movement of MLC leaves. Then, the determination of radiation dose, gantry speed and their control become harder; and this directly effects the structure of the related mathematical optimization models. First of all, the number of decision variables increases not only for dealing with the controllability issues, but also for linearizing the nonlinearities that the dependencies introduce. Besides, the few existing mathematical optimization formulations either do not include hard constraints for many of the radiation dose related dependencies, or these constraints are relaxed so that the relaxed formulations are solvable. An exception is the work by Akartunalı et al. (2015), which is explained in the next section. We aim to close this gap by including constraints that model radiation dose related interdependencies in our mixed-integer linear programming (MILP) formulation. In other words, instead of relaxing the constraints associated with the dose distributions and penalizing the deviations from the dose limits (i.e. thresholds) in the objective function, we formulate and include them in the MILP formulation as hard constraints, which are validated by the oncologists and medical physicists affiliated with Istanbul University Oncology Institute (Bilge et al., 2017). The resulting formulation is new, also in terms of decision variables required to define the mechanical constraints of the MLC system.

It is known that treatment delivery time and total MU need are low in VMAT plans compared with those of IMRT. It is also known that decreasing total MU of a plan decreases the risk of secondary radiation-induced malignancies and also the total body dose due to radiation leakage. To the best of our knowledge there is no work that takes into account total MUs for VMAT plans. The models proposed so far generally minimize dose deviations from the prescribed limits and treat two different treatment plans with similar dose deviations but with different total MUs as being equivalent; they just do not distinguish between the feasible plans with respect to their radiation requirements. Therefore, it is not possible to benefit from VMAT's whole potential in radiation treatment if one of these models is used to determine optimal treatment plans. We also consider this point and formulate the objective function of our MILP model to minimize the total radiation amount (in MUs) delivered during a fraction.

When the total deviation of a treatment plan is not zero, it is not guaranteed that the resulting plan is feasible according to the dose constraints (Akartunalı et al., 2015). In reality, in the planning departments, each time the resulting treatment plan is checked by the planner and the oncologist, if it is not feasible then the plan is re-optimized after modifying the weights penalizing the deviations in the objective. This re-optimizing process continues until a feasible plan is obtained; and this clearly increases the treatment planning time for patients.

The solution of our new MILP formulation, which we call VMAT Planning (VMATP) model, is computationally challenging as we have shown in our earlier work (Dursun et al., 2016). However, it produces VMAT plans with small MUs satisfying dosimetric quality levels, and this is desired very much in real clinic applications (Bilge et al., 2017). To the best of our knowledge, our study is the first attempt to solve exactly a model in which all VMAT's treatment related constraints are forced to be satisfied. Our experiments show that solving our VMATP as a single MILP formulation quickly becomes intractable as the problem size increases. Benders decomposition is an exact large-scale optimization method that has been used to improve solvability of a wide range of difficult optimization problems (Rahmaniani et al., 2017). A large problem is partitioned into smaller problems each considering only a subset of variables and constraints. Since computational difficulty of solving optimization problems increases significantly with the number of variables and constraints, solving these smaller problems iteratively can be more efficient than solving a single large-scale problem (Taşkın, 2010). We observe that VMATP consists of two main types of decision variables: geometry variables modeling aperture shapes, and dose variables determining the amount of radiation to be delivered through each aperture. This model structure provides a basis for our Benders decomposition approach, which iteratively solves a master problem to find optimal aperture shapes and a subproblem to calculate optimal doses. In this paper, we have used Benders decomposition and developed an efficient solution algorithm after improving its naive form by means of computational strategies.

The rest of the paper is organized as follows. Next section summarizes the related literature concentrating on optimization methods for IMRT and VMAT planning. In Section 3, we give the mathematical formulation of the new MILP formulation VMATP, in detail, which is followed by Section 4 where we develop a Benders decomposition method and explain its improvements. In Section 6, we explain the results of the application of the improved Benders decomposition method using real prostate cancer data, and assess the efficiency of the new algorithms. Finally, we conclude the paper in Section 7.

2. Related Works

Beam angle optimization (BAO), which is the first phase in IMRT planning, is often done manually by medical physicists based on their experience. There are also studies where a function is defined to determine the quality of a set of directions and this function is optimized to find the best one (Ehrgott et al., 2008). After determining the beam angles (i.e. control points) it is possible to solve the fluence map optimization (FMO) problem, the second phase, as a convex optimization model; it can be solved efficiently using one of the existing algorithms for convex optimization (Papp and Unkelbach, 2014). The third phase, namely MLC leaf sequencing (MLS) problem is closely related with VMAT treatment planning. The fluence map at a control point is decomposed into a number of apertures with intensities, namely a nonnegative integer matrix is re-expressed as a linear combination of binary matrices with positive integer weights. All of the binary matrices should satisfy the properties of the MLC system as the consecutive ones property. During the decomposition of the fluence map, total delivery time (i.e. beam on time) and/or the total number of apertures (i.e. total machine setups) are minimized (Baatar et al., 2007; Ernst et al., 2009; Mason et al., 2012; Boland et al., 2004; Guta, 2003; Taşkın et al., 2010; Cambazard et al., 2012; Mason et al., 2015). The problem of minimizing total delivery time, which is the time that the radiation delivery is on, consists of the minimization of the sum of the individual intensities determined for each aperture, and it is polynomially solvable. However, in the cardinality problem the total number of apertures is minimized and this problem is shown to be strongly NP-hard (Baatar et al., 2005). There are also studies in the literature that integrates BAO and FMO and solve a monolithic non-convex optimization problem to determine the beam angles and fluence maps simultaneously (Lee et al., 2003; Bertsimas et al., 2013). *Direct aperture optimization* (DAO) is the name given to the integration of the last two phases FMO and MLS. Romeijn et al. (2005); Men et al. (2007); Carlsson (2008); Salari and Unkelbach (2013); Preciado-Walters et al. (2006) directly finds an optimal number of apertures with intensities at each one of the given beam angles instead of finding a fluence map and then solving the realization problem sequentially. We refer the interested reader to the excellent survey of Ehrgott et al. (2010) for more details on IMRT planning.

Rotational arc therapy was first introduced in 1995 to the literature with name Intensity Modulated Arc Therapy (IMAT) (Yu, 1995). This method had been stagnant until VMAT technique appeared in the study of Otto (2008). VMAT is a single-arc IMAT. It is more flexible since there are additional degrees of freedom: gantry speed and dose rate are also variables to be optimized. There are mainly two categories of studies dealing with VMAT treatment planning. The members of the first group use a two-phase approach that generally starts by first identifying an optimal IMRT treatment plan, which is comprised of fluence intensity maps at evenly spaced control points. Then this plan is converted into a deliverable VMAT plan with an arc-sequencing approach (Cameron, 2005; Wang et al., 2008; Luan et al., 2008; Shepard et al., 2007; Cao et al., 2009; Bzdusek et al., 2009). Converting an IMRT plan, where the fluence maps are optimized at a small number of control points, may cause deviation in dose distribution quality of the resulting VMAT plan, since VMAT planning problem has its own constraints on MLC movement, and during the conversion these constraints must be satisfied. However, Craft et al. (2012), Salari et al. (2012), Wala et al. (2012) first obtain a fine sample IMRT plan with a large number of equally spaced control points, which is then coarsened to reduce the delivery time by maintaining the dose distribution's quality requirements.

The studies of the second category directly optimize the leaf positions and intensities of the apertures at control points, instead of converting a set of optimized fluence intensity maps; they are called DAO methods similar to the ones one can face in the IMRT planning literature. The algorithms proposed in the studies of Earl et al. (2003), Otto (2008), and Zhang et al. (2010) start with a relatively coarse sampling of the control points, and heuristic methods are used to find the final treatment plan. Men et al. (2010) formulate a large-scale convex programming problem and solve it by a column generation heuristic that adds aperture shapes at control points one by one while taking into account their compatibility with the previously added ones. Peng et al. (2012) extend their formulation and approach considering some additional physical restrictions. Papp and Unkelbach (2014) enforce unidirectional leaf motion over an arc segment, and determine the apertures by solving a sequence of convex optimization problems. In a recent study Mahnam et al. (2017) develop a heuristic for the VMAT planing problem where a full treatment arc is decomposed into a number of partial arcs with the same length, and a set of apertures for each partial arc is generated as new columns. They formulate the pricing problem as a shortest path problem and solve it using a standard shortest path algorithm. They also assume that the MLC leaves move unidirectionally from left to right and analyze two leaf-motion strategies. Gozbasi (2010), Akartunalı et al. (2015), and Song et al. (2015) formulate MILP models for the VMAT planning problem in which an aperture and radiation intensity are optimized at each control point subject to the clinical requirements. Akartunalı et al. (2015) embed the treatment requirements, except the partial volume constraints of TVs, to their mathematical model as hard constraints, and they try to maximize total number of voxels that absorbs at least the prescribed amount of radiation. Moreover, although they make the first step towards the development of exact methods, they finally suggest heuristics to obtain good feasible treatment plans, which are clinically acceptable as well. Finally, in our recent work Dursun et al. (2016) we propose two new MILP formulations for the VMAT planning problem. They essentially differentiate in how the aperture related constraints are formulated.

3. Mathematical Formulation

A VMAT treatment plan must satisfy both radiation therapy dose prescriptions and mechanical limitations of the linear accelerator and MLC system. Our VMAT planning model VMATP, whose preliminary version can be found in (Dursun et al., 2016), consists of the constraints related to these requirements and minimizes the total radiation dose delivered during the treatment. First, we discretize continuous radiation delivery by assuming that there is a large number of evenly spaced control points (i.e. 180) on a co-planar rotational arc. VMATP determines the aperture shape and the amount of radiation to be delivered at each of the control points. Parameters and decision variables used to formulate VMATP are summarized in Table 1 and Table 2, respectively. We introduce integer and binary decision variables to the model to represent an aperture as a binary matrix. For each row i of an aperture at control point k, l_{ik} and r_{ik} are nonnegative integer variables that define positions of the left and right leaves respectively. Also, for each beamlet j of row i at control point k there is a binary variable z_{ijk}; it is set to 1 if this beamlet is open, 0 otherwise. In Figure 6 the second row of the aperture given at Figure 4 and the corresponding decision variables are illustrated. The

Table 1: Parameters of VMATP

Parameter	Definition
i	Index for an MLC row (i=1,...,m).
j	Index for an MLC column (j=0,...,n+1), 0 and n+1 are home positions of the left and the right leaves, respectively.
k	Index for a control point (k=1,...,K).
t	Index for a target volume (TV) (t=1,...,T).
o	Index for an organ at risk (OAR) volume (o=1,...,O).
v	Index for a voxel in a volume.
V_t^{TV}	Set of voxels in TV t.
V^{TV}	Set of all voxels in all TVs, $V^{TV} = \bigcup_{t=1}^{T} V_t^{TV}$.
V_o^{OAR}	Set of voxels in OAR volume o.
V^{OAR}	Set of all voxels in all OAR volumes, $V^{OAR} = \bigcup_{o=1}^{O} V_o^{OAR}$.
V	Set of all voxels, $V = V^{TV} \cup V^{OAR}$.
L_t^{TV}	Lower bound for total absorbed radiation dose amount of a target voxel in TV t (in Gy).
U_t^{TV}	Upper bound for total absorbed radiation dose amount of a target voxel in TV t (in Gy).
U_o^{OAR}	Tolerance radiation dose amount of OAR volume o (in Gy).
$\overline{d}_t$	The prescribed dose for TV t (in Gy).
D_{ijkv}	Dose influence matrix (in Gy/MU).
δ	The maximum allowable distance (in beamlet(s)) that a leaf can move between two consecutive control points.
α_t^{TV}	The minimum ratio of voxels in TV t that receive radiation at least the prescribed dose $\overline{d}_t$.
α_o^{OAR}	The minimum ratio of voxels in OAR volume o that receive radiation at most the tolerance dose U_o^{OAR}.
L^{mu}	Lower bound of dose intensity at a control point (in MU).
U^{mu}	Upper bound of dose intensity at a control point (in MU).

Table 2: Variables of VMATP

Variable	Definition
l_{ik}	Nonnegative integer variable that represents the position of the left leaf (i.e. the last (rightmost) closed beamlet on the left side of row i at control point k).
r_{ik}	Nonnegative integer variable that represents the position of the right leaf (i.e. the first (leftmost) closed beamlet on the right side of row i at control point k).
z_{ijk}	Binary variable, whose value is 1 if the jth beamlet of row i at control point k is open, 0 otherwise (j=1,...,n).
mu_k	Nonnegative continuous variable, which stands for radiation dose intensity (in MU) at control point k.
d_v	Nonnegative continuous variable, which represents total absorbed dose by voxel v (in Gy).
a_{ijk}	Nonnegative continuous variable used for linearization.
ξ_t^{TV}	Continuous variable, which represents absorbed radiation amount by the ((1-α_t^{TV})|V_t^{TV}|)th voxel in TV t receiving the lowest radiation. The use of the variable is described in detail in constraint (19).
ξ_o^{OAR}	Continuous variable, which represents absorbed radiation amount by the ((1-α_o^{OAR})|V_o^{OAR}|)th voxel in OAR volume o receiving the highest radiation. The use of the variable is described in detail in constraint (24).
x_{tv}	Artificial variable for voxel v in TV t.
y_{ov}	Artificial variable for voxel v in OAR volume o.

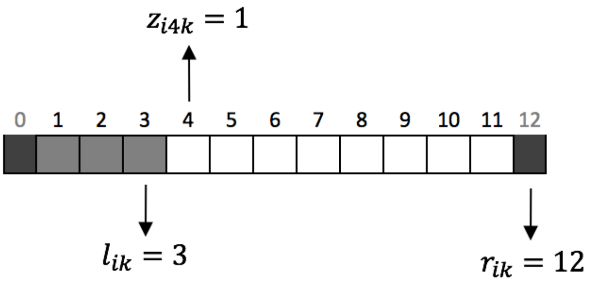


Figure 6: Row i of the aperture at control point k with decision variable values

left leaf blocks the first three beamlets and takes the value of 3, and the right leaf is at its home position so it takes the value of $n + 1 = 11 + 1 = 12$.

The first mechanical constraint is associated with the MLC system. In a row of an aperture there can be at most one open beamlet chain, which is called consecutive ones property that must be satisfied by almost all MLC systems. We only consider this property, and refer the reader to the study of Gören and Taşkın (2015) for detailed information about other mechanical properties of different MLC systems. Similar to the studies both in VMAT planning (e.g. (Akartunalı et al., 2015)) and IMRT planning (e.g. Mason et al. (2012)), to ensure the consecutive ones property we introduce the following constraints:

$$r_{ik} - l_{ik} \geq 1 \qquad i = 1, \ldots, m; \;\; k = 1, \ldots, K \tag{1}$$

$$r_{ik} - jz_{ijk} \geq 1 \qquad i = 1, \ldots, m; \;\; j = 1, \ldots, n; \;\; k = 1, \ldots, K \tag{2}$$

$$(n + 1 - j)z_{ijk} + l_{ik} \leq n \qquad i = 1, \ldots, m; \;\; j = 1, \ldots, n; \;\; k = 1, \ldots, K \tag{3}$$

$$r_{ik} - l_{ik} - \sum_{j=1}^{n} z_{ijk} = 1 \qquad i = 1, \ldots, m; \;\; k = 1, \ldots, K \tag{4}$$

$$\mathbf{l} \in \mathbb{Z}_+^{m \times K}; \mathbf{r} \in \mathbb{Z}_+^{m \times K}; \mathbf{z} \in \{0, 1\}^{m \times n \times K}. \tag{5}$$

For a given row i at control point k, constraint (1) prevents the left and right leaves from overlapping. Constraints (2)–(4) force all z_{ijk} variables associated with the open beamlets between the left and right leaves to be 1. Also, as a consequence of these constraints, the left leaf can be between 0 and n and the right leaf can be between 1 and $n + 1$.

Another mechanical limitation of the MLC system, which is generally taken into account in VMAT studies (e.g. (Akartunalı et al., 2015; Song et al., 2015), is that during the rotation of the gantry, between two adjacent control points of the arc, a leaf cannot move more than a certain distance, depending on the speed of the gantry. Namely, the aperture shapes at two adjacent control points must be similar. We introduce the following constraints to formulate similarites:

$$l_{i(k+1)} - l_{ik} \leq \delta \qquad i = 1, \ldots, m;\ \ k = 1, \ldots, K - 1 \tag{6}$$

$$l_{ik} - l_{i(k+1)} \leq \delta \qquad i = 1, \ldots, m;\ \ k = 1, \ldots, K - 1 \tag{7}$$

$$r_{i(k+1)} - r_{ik} \leq \delta \qquad i = 1, \ldots, m;\ \ k = 1, \ldots, K - 1 \tag{8}$$

$$r_{ik} - r_{i(k+1)} \leq \delta \qquad i = 1, \ldots, m;\ \ k = 1, \ldots, K - 1. \tag{9}$$

These constraints restrict the leaves to move no more than δ beamlets between control points k and $k+1$. To sum up, as the speed of the gantry increases the amount of δ decreases and the apertures at the adjacent control points become similar.

We have explained the geometry constraints (1)–(9) that generate a feasible aperture for each control point so far. Now, we continue by introducing radiation delivery and treatment constraints. During the rotation of the gantry, the linear accelerator delivers radiation continuously to the patient's body through the aperture formed by the MLC. We assume that the radiation delivery is realized at the control points only and lasts for a certain time. This is reasonable, because not only the effect of radiation but also the apertures at adjacent control points are similar due to the similarity constraints (6)–(9). In addition to the aperture shape, VMATP determines the radiation dose intensity at each control point. Note that there is a relation between the dose rate of the linear accelerator and radiation dose intensity. The dose rate is in monitor unit (MU) per unit time, and the dose intensity at a control point is a function of the dose rate and gantry rotation speed (i.e. if the gantry is slow then it is possible to deliver more radiation). Dose rate and intensity may change at control points. However, they must be within mechanical limits of the linear accelerator, which also depends on the rotation speed. Also, we assume that the speed of the gantry is constant. We introduce a nonnegative continuous variable mu_k to represent the radiation dose intensity at each control point k. We also introduce constraints

$$mu_k \geq L^{mu} \qquad k = 1, \ldots, K \tag{10}$$

$$mu_k \leq U^{mu} \qquad k = 1, \ldots, K \tag{11}$$

$$\mathbf{mu} \in \mathbb{R}_{+}^{K}, \tag{12}$$

where parameters L^{mu} and U^{mu} are calculated using dose rate limits and gantry speed.

A VMAT treatment plan should also satisfy the clinical requirements, which are prescribed by the oncologists, depending on the tumor's type and patient's anatomy. Generally, two types of constraints are defined for a given target: *partial volume constraints* and *full volume constraints*. For an OAR, only partial volume constraints are prescribed. For example, a partial volume constraint defined for a TV must satisfy that at least 95% of the volume must absorb radiation at least as the prescribed dose. The coverage rate becomes 100% in a full volume constraint: 100% of the volume must absorb radiation within the prescribed bounds. The body of the patient is discretized into voxels in order to be able to formulate these restrictions. The amount of radiation (d_v) absorbed by each voxel v is calculated using equality

$$d_v - \sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n} D_{ijkv} z_{ijk} mu_k = 0 \qquad v \in V = V^{TV} \cup V^{OAR}. \tag{13}$$

Note that V is the set of all voxels, namely it is union of all TVs (V^{TV}) and OARs (V^{OAR}). Note also that (13) includes nonlinear terms created by the product of binary variables **z** with the continuous variables **mu**. We use the linearization method by McCormick (1976), which eventually forms the convex envelop of general bilinear terms, to linearize constraint (13). We introduce auxiliary variable a_{ijk} for each beamlet and obtain

$$d_v - \sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n} D_{ijkv} a_{ijk} = 0 \qquad v \in V = V^{TV} \cup V^{OAR} \tag{14}$$

$$a_{ijk} \leq U^{mu} z_{ijk} \qquad i = 1, \ldots, m; \;\; j = 1, \ldots, n; \;\; k = 1, \ldots, K \tag{15}$$

$$a_{ijk} \geq mu_k - U^{mu}(1 - z_{ijk}) \qquad i = 1, \ldots, m; \;\; j = 1, \ldots, n; \;\; k = 1, \ldots, K \tag{16}$$

$$a_{ijk} \leq mu_k \qquad i = 1, \ldots, m; \;\; j = 1, \ldots, n; \;\; k = 1, \ldots, K \tag{17}$$

$$\mathbf{d} \in \mathbb{R}_+^{|V|}; \mathbf{a} \in \mathbb{R}_+^{m \times n \times K}. \tag{18}$$

Now it is possible to include the clinical requirements using the total absorbed radiation dose amounts of voxels. Similar to Romeijn et al. (2006) and Gozbasi (2010), we use Conditional Value At Risk (CVaR) approach, which is popular in risk management (Sarykalin et al., 2008), to formulate partial-volume constraints. For each TV t the following partial volume constraints are introduced:

$$\xi_t^{TV} - \frac{1}{(1 - \alpha_t^{TV})|V_t^{TV}|} \sum_{v \in V_t^{TV}} x_{tv} \geq \overline{d}_t \qquad t = 1, \ldots, T \tag{19}$$

$$x_{tv} \geq \xi_t^{TV} - d_v \qquad t = 1, \ldots, T; \;\; v \in V_t^{TV} \tag{20}$$

$$\mathbf{x} \in \mathbb{R}_+^{|V^{TV}|}; \boldsymbol{\xi}^{TV} \in \mathbb{R}^T. \tag{21}$$

The average dose of the (1-α_t^{TV})$|V_t^{TV}|$ voxels receiving the lowest dose in TV t, namely the *lower mean tail dose at level* α_t^{TV} is forced to be at least the prescription dose. In other words, at least $\alpha_t^{TV}|V_t^{TV}|$ voxels absorb radiation more than or equal to $\overline{d}_t$. Furthermore, there are full volume constraints for each TV:

$$d_v \geq L_t^{TV} \qquad t = 1, \ldots, T; \;\; v \in V_t^{TV} \tag{22}$$

$$d_v \leq U_t^{TV} \qquad t = 1, \ldots, T; \;\; v \in V_t^{TV}. \tag{23}$$

Full volume constraints (22) and (23) ensure that each voxel in TV t receives radiation within its prescribed limits.

There are only partial volume constraints for OAR volumes in VMATP. Similar to the ones defined for TVs we introduce the following inequalities for each OAR:

$$\xi_o^{OAR} + \frac{1}{(1-\alpha_o^{OAR})|V_o^{OAR}|} \sum_{v \in V_o^{OAR}} y_{ov} \leq U_o^{OAR} \qquad o = 1, \ldots, O \tag{24}$$

$$y_{ov} \geq d_v - \xi_o^{OAR} \qquad o = 1, \ldots, O; \;\; v \in V_o^{OAR} \tag{25}$$

$$\mathbf{y} \in \mathbb{R}_+^{|V^{OAR}|}; \boldsymbol{\xi}^{OAR} \in \mathbb{R}^O. \tag{26}$$

The average dose of the (1-α_o^{OAR})$|V_o^{OAR}|$ voxels absorbing the highest doses in OAR o, namely the *upper mean tail dose at level* α_o^{OAR} is forced to be at most its tolerance dose limit U_o^{OAR}. For details about CVaR method we refer the reader to the study of Romeijn et al. (2006) where it is applied for developing a linear-programming-based approach to solve FMO problem in IMRT treatment planning. Finally, the objective function

$$\min \sum_{k \in K} mu_k \tag{27}$$

minimizes total radiation intensity (in MU) the patient receives during his/her treatment. VMATP finds an optimal plan minimizing total dose intensity among all feasible treatment plans.

Finally, we should point out that the formulations by Gozbasi (2010), Akartunalı et al. (2015), and Song et al. (2015) have relations with the new formulation given above. They all provide treatment plans for VMAT; but they are all different and incomparable since they are not based on a common problem definition.

4. Benders Decomposition

Benders decomposition was proposed by Benders (1962), and has been widely used in the solution of large-scale mathematical optimization problems. It is particularly effective for solving problems having a subset of variables that are *complicating* in the sense that the problem becomes significantly easier to solve if such complicating variables are fixed. Its ability to exploit the structure of the problem and distribute the overall computational work are key facts behind the many successful applications of Benders decomposition (Rahmaniani et al., 2017).

The methodological frame of Benders decomposition is quite direct. Once the mathematical formulation is obtained and the set of complicating variables is determined, the formulation is first projected onto the subspace of the complicating variables. This alternative formulation, which contains only the complicating variables, is known as the *Benders reformulation* of the original problem, and it can be constructed by means of the extreme points and directions of the (dual) formulation obtained by fixing the complicating variables. Since the number of extreme points and rays of a polyhedron is exponential in general, the resulting Benders reformulation has a huge set of constraints and its solution, when possible, is computationally very demanding. Therefore, a relaxation strategy based on the dynamic constraint generation is applied. At each iteration a relaxation of Benders reformulation is obtained by including a subset of the inequalities, namely the *relaxed master problem* (RMP), and solved. Then, whether the optimal solution of the current restricted master is also optimal with respect to the rest of the inequalities that are not explicitly considered yet is tested by solving a (dual) subproblem. In case of an unfavorable response, a Benders cut is generated and added to the restricted master. This cut is an *optimality cut*, if the subproblem has a finite optimum solution, since it is constructed using the optimal solution of the subproblem. Otherwise, the subproblem is unbounded and a *feasibility cut* is constructed using an extreme direction proving that it is unbounded. At every iteration a new inequality is added to the restricted master problem until the addition of a new cut is not possible; an optimal solution of the restricted master is also optimal for Benders reformulation.

Benders decomposition was initially proposed for MILPs, which becomes a linear program (LP) after fixing the integer variables (i.e. complicating variables). Then, it is possible to use standard duality theory to generate the optimality and feasibility cuts. In fact, the nature of the radiotherapy is very suitable from this perspective since the variables used to shape the apertures in order to determine the *geometry* of the beam, are integer valued and the variables used to determine the prescribed dose requirements are continuous. Once the geometry variables are fixed, the geometry of the apertures are set and the resulting LP can be solved to determine optimal beam intensities subject to *dose* inequalities. As can be observed, this partitioning strategy of the variables is also possible for our MILP formulation for VMAT planning, i.e. VMATP. We should point out that a similar partitioning strategy is also used by Taşkın (2010) for the MLS problem in IMRT planning, where the MLC can only form rectangular apertures.

4.1. Benders Reformulation of VMATP

We identify the binary integer variables $\mathbf{z}$, which represent the beamlets of the apertures, as the complicating variables in our model. If they are fixed, namely if we know the shape of each aperture at each control point, the dose constraints do not include integer variables (LP). Using this observation we decompose the original problem into a relaxed master problem and a subproblem. The relaxed master problem produces a feasible aperture at each control point; and the subproblem calculates the optimum intensity for each one of them, namely the optimum radiation dose that the linear accelerator delivers at each control point while considering the feasibility of the treatment plan with respect to the clinical requirements.

Given a vector $\hat{\mathbf{z}}$ that denotes values assigned to $\mathbf{z}$ variables, the subproblem SP($\hat{\mathbf{z}}$) and its dual DSP($\hat{\mathbf{z}}$) can be formulated as

SP ($\hat{\mathbf{z}}$):

$$\min \sum_{k \in K} mu_k \tag{27}$$

s.t.

$$d_v - \sum_{k=1}^{K} \sum_{i=1}^{m} \sum_{j=1}^{n} D_{ijkv} a_{ijk} = 0 \qquad v \in V = V^{TV} \cup V^{OAR} \qquad : \pi_v \tag{14}$$

$$a_{ijk} \leq U^{mu} \hat{z}_{ijk} \qquad i = 1, \ldots, m;\ \ j = 1, \ldots, n;\ \ k = 1, \ldots, K \qquad : \beta^1_{ijk} \tag{15}$$

$$a_{ijk} \geq mu_k - U^{mu}(1 - \hat{z}_{ijk}) \qquad i = 1, \ldots, m;\ \ j = 1, \ldots, n;\ \ k = 1, \ldots, K \qquad : \beta^2_{ijk} \tag{16}$$

$$a_{ijk} \leq mu_k \qquad i = 1, \ldots, m;\ \ j = 1, \ldots, n;\ \ k = 1, \ldots, K \qquad : \beta^3_{ijk} \tag{17}$$

$$\xi^{TV}_t - \frac{1}{(1 - \alpha^{TV}_t)|V^{TV}_t|} \sum_{v \in V^{TV}_t} x_{tv} \geq \overline{d}_t \qquad t = 1, \ldots, T \qquad : \theta^1_t \tag{19}$$

$$x_{tv} \geq \xi^{TV}_t - d_v \qquad t = 1, \ldots, T;\ \ v \in V^{TV}_t \qquad : \tau^1_{tv} \tag{20}$$

$$d_v \geq L^{TV}_t \qquad t = 1, \ldots, T;\ \ v \in V^{TV}_t \qquad : \epsilon^1_{tv} \tag{22}$$

$$d_v \leq U^{TV}_t \qquad t = 1, \ldots, T;\ \ v \in V^{TV}_t \qquad : \epsilon^2_{tv} \tag{23}$$

$$\xi^{OAR}_o + \frac{1}{(1 - \alpha^{OAR}_o)|V^{OAR}_o|} \sum_{v \in V^{OAR}_o} y_{ov} \leq U^{OAR}_o \qquad o = 1, \ldots, O \qquad : \theta^2_o \tag{24}$$

$$y_{ov} \geq d_v - \xi^{OAR}_o \qquad o = 1, \ldots, O;\ \ v \in V^{OAR}_o \qquad : \tau^2_{ov} \tag{25}$$

$$mu_k \geq L^{mu} \qquad k = 1, \ldots, K \qquad : \mu^1_k \tag{10}$$

$$mu_k \leq U^{mu} \qquad k = 1, \ldots, K \qquad : \mu^2_k \tag{11}$$

(12), (18), (21), (26),

and

DSP ($\hat{\boldsymbol{z}}$):

$$
\begin{aligned}
\max \; & \sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n} U^{mu}(-\hat{z}_{ijk}\beta^1_{ijk} + (\hat{z}_{ijk}-1)\beta^2_{ijk}) - \sum_{o=1}^{O} \theta^2_o U^{OAR}_o + \\
& \sum_{t=1}^{T} \theta^1_t \overline{d}_t + \sum_{t=1}^{T}\sum_{v\in V_t^{TV}} (L_t^{TV}\epsilon^1_{tv} - U_t^{TV}\epsilon^2_{tv}) + \sum_{k=1}^{K}(L^{mu}\mu^1_k - U^{mu}\mu^2_k) \qquad (28)
\end{aligned}
$$

s.t.

$$
\begin{aligned}
& \pi_v + \tau^1_{tv} + \epsilon^1_{tv} - \epsilon^2_{tv} \le 0 && t = 1,\dots,T;\ v \in V_t^{TV} && : d_v && (29)\\
& \pi_v - \tau^2_{ov} \le 0 && o = 1,\dots,O;\ v \in V_o^{OAR} && : d_v && (30)\\
& -\sum_{v\in V} D_{ijkv}\pi_v - \beta^1_{ijk} + \beta^2_{ijk} - \beta^3_{ijk} \le 0 && i = 1,\dots,m;\ j = 1,\dots,n; && && \\
& && k = 1,\dots,K && : a_{ijk} && (31)\\
& -\sum_{i=1}^{m}\sum_{j=1}^{n}(\beta^2_{ijk} - \beta^3_{ijk}) + \mu^1_k - \mu^2_k \le 1 && k = 1,\dots,K && : mu_k && (32)\\
& -\theta^2_o + \sum_{v\in V_o^{OAR}} \tau^2_{ov} = 0 && o = 1,\dots,O && : \xi^{OAR}_o && (33)\\
& \theta^1_t - \sum_{v\in V_t^{TV}} \tau^1_{tv} = 0 && t = 1,\dots,T && : \xi^{TV}_t && (34)\\
& -\frac{1}{(1-\alpha^{OAR}_o)|V^{OAR}_o|}\theta^2_o + \tau^2_{ov} \le 0 && o = 1,\dots,O;\ v \in V_o^{OAR} && : y_{ov} && (35)\\
& -\frac{1}{(1-\alpha^{TV}_t)|V^{TV}_t|}\theta^1_t + \tau^1_{tv} \le 0 && t = 1,\dots,T;\ v \in V_t^{TV} && : x_{tv} && (36)\\
& \boldsymbol{\pi} \in \mathbb{R}^{|V|}; \boldsymbol{\beta}^1 \in \mathbb{R}_+^{m\times n\times K}; \boldsymbol{\beta}^2 \in \mathbb{R}_+^{m\times n\times K}; \boldsymbol{\beta}^3 \in \mathbb{R}_+^{m\times n\times K}; && && && (37)\\
& \boldsymbol{\theta}^1 \in \mathbb{R}_+^{T}; \boldsymbol{\theta}^2 \in \mathbb{R}_+^{O}; \boldsymbol{\tau}^1 \in \mathbb{R}_+^{|V^{TV}|}; \boldsymbol{\tau}^2 \in \mathbb{R}_+^{|V^{OAR}|}; && && && \\
& \boldsymbol{\epsilon}^1 \in \mathbb{R}_+^{|V^{TV}|}; \boldsymbol{\epsilon}^2 \in \mathbb{R}_+^{|V^{TV}|}; \boldsymbol{\mu}^1 \in \mathbb{R}_+^{K}; \boldsymbol{\mu}^2 \in \mathbb{R}_+^{K}. && && &&
\end{aligned}
$$

Extreme points and extreme directions of the dual polyhedron are used to construct Benders reformulation of the original problem. Suppose that Δ and Ω denote the set of extreme points and the set of extreme directions of the dual polyhedron, respectively. We further define $f(\boldsymbol{\beta}^1,\boldsymbol{\beta}^2,\boldsymbol{\theta}^1,\boldsymbol{\theta}^2,\boldsymbol{\epsilon}^1,\boldsymbol{\epsilon}^2,\boldsymbol{\mu}^1,\boldsymbol{\mu}^2) = \sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n} U^{mu}(-z_{ijk}\beta^1_{ijk} + (z_{ijk}-1)\beta^2_{ijk}) - \sum_{o=1}^{O}\theta^2_o U^{OAR}_o + \sum_{t=1}^{T}\theta^1_t\overline{d}_t + \sum_{t=1}^{T}\sum_{v\in V_t^{TV}}(L_t^{TV}\epsilon^1_{tv} - U_t^{TV}\epsilon^2_{tv}) + \sum_{k=1}^{K}(L^{mu}\mu^1_k - U^{mu}\mu^2_k)$ and the Benders reformulation of VMATP becomes

$$\min \eta \tag{38}$$

s.t.

(1) − (9) and

$$f(\boldsymbol{\beta}^1, \boldsymbol{\beta}^2, \boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \boldsymbol{\epsilon}^1, \boldsymbol{\epsilon}^2, \boldsymbol{\mu}^1, \boldsymbol{\mu}^2) \leq \eta \qquad \boldsymbol{\beta}^1, \boldsymbol{\beta}^2, \boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \boldsymbol{\epsilon}^1, \boldsymbol{\epsilon}^2, \boldsymbol{\mu}^1, \boldsymbol{\mu}^2 \in \Delta \tag{39}$$

$$f(\boldsymbol{\beta}^1, \boldsymbol{\beta}^2, \boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \boldsymbol{\epsilon}^1, \boldsymbol{\epsilon}^2, \boldsymbol{\mu}^1, \boldsymbol{\mu}^2) \leq 0 \qquad \boldsymbol{\beta}^1, \boldsymbol{\beta}^2, \boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \boldsymbol{\epsilon}^1, \boldsymbol{\epsilon}^2, \boldsymbol{\mu}^1, \boldsymbol{\mu}^2 \in \Omega \tag{40}$$

$$\eta \geq 0. \tag{41}$$

We introduce a new variable η representing the total radiation intensity, which is the objective function of the subproblem. Since $\mathbf{0}$ is a feasible solution of the dual problem, the lower bound of η is set to 0. Constraints (1)–(9) determine a feasible aperture shape for each control point. Constraints (39) are Benders optimality cuts and constraints (40) are Benders feasibility cuts and they all represent the subproblem. In the naive form of Benders decomposition, initially all Benders cuts are relaxed and the resulting RMP is solved iteratively. In each iteration either an optimality cut or feasibility cut is added to the RMP, which is re-solved until the stopping condition is satisfied.

Our preliminary results show that the naive form is inferior according to the computation time and solution quality. The most important reason of the time consumption is that in each iteration RMP is solved from scratch after adding a new inequality (i.e. a new Benders cut). Even though solving RMP optimally and generating a cut for the optimal solution may yield strong cuts, solution time increases as the number of Benders cuts, and thus the size of RMP, increases. Another drawback of the naive implementation is that the lower bound improves very slowly. A feasible solution for the whole problem may not be obtained within a reasonable amount of time, since the number of feasible RMP solutions, namely feasible MLC combinations according to aperture shape (i.e. geometry) constraints, is very large.

5. Algorithmic and Modeling Improvements

5.1. Valid Inequalities

In the Benders reformulation the objective function (27) is removed since it belongs to the subproblem. Also, initial lower bound of the master objective value is set to zero since $\mathbf{0}$ is a trivial feasible solution of the dual problem. This causes a large optimality gap at the beginning, which slowly gets smaller as Benders cuts are added. To address this issue, we aim to discard some of the master solutions that are infeasible for the whole problem. We observe that, if a master solution (an aperture per control point) does not have enough capacity to deliver enough radiation such that each voxel of TV t absorbs at least L_t^{TV} amount of radiation, this solution cannot be feasible for the whole problem.

Hence, we can eliminate such solutions at the beginning by adding inequalities

$$\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n} z_{ijk} D_{ijkv} U^{mu} \geq L_t^{TV} \qquad t = 1, \ldots, T;\ \ v \in V_t \tag{42}$$

to the RMP. Recall that the parameter U^{mu} is the maximum radiation intensity that linear accelerator can deliver at a control point. However, according to our preliminary experiments, we note that the improvement due to these valid inequalities is not significant. Thus, we introduce to RMP new surrogate decision variables (a continuous variable a per beamlet and a continuous variable mu per control point), and related constraints similar to those in the whole problem. As a result, we add the following inequalities instead:

$$a_{ijk} \leq U^{mu} z_{ijk} \qquad i = 1, \ldots, m;\ \ j = 1, \ldots, n;\ \ k = 1, \ldots, K \tag{43}$$

$$a_{ijk} \leq mu_k \qquad i = 1, \ldots, m;\ \ j = 1, \ldots, n;\ \ k = 1, \ldots, K \tag{44}$$

$$\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ijk} D_{ijkv} \geq L_t^{TV} \qquad t = 1, \ldots, T;\ \ v \in V_t \tag{45}$$

$$\eta \geq \sum_{k=1}^{K} mu_k. \tag{46}$$

Note that constraints (43) and (44) are similar to the linearization constraints (15) and (17) in the VMATP, however (16) is relaxed. The addition of inequalities (45) to RMP guarantees that in any master solution each target voxel absorbs radiation no less than the prescribed lower bound. Benders optimality cuts ensure that η is at least as large as the objective function value of DSP for a given master solution, namely the minimum total radiation dose intensity in a feasible treatment. Constraint (46) is valid, and it improves lower bound effectively, since the minimum total radiation dose is found considering only target voxels in this extended master problem, and this amount can be at most the minimum total radiation dose calculated by solving DSP. Finally, we do not have to add constraint set (42) anymore, since it is replaced by (45), which is tighter. These extensions make the master problem harder to solve. However, according to our preliminary observations, they significantly improve the lower bound and performance of the Benders decomposition algorithm as a consequence. Thus, in the final form of the method we add constraints (43)–(46) to the master problem. These inequalities contain some information about the original objective function that we project out, and cuts some of the master solutions that are not feasible for the whole treatment.

5.2. Strong Benders Cuts

Stronger Benders cuts may improve the lower bound faster and help for the rapid convergence to optimality. For the optimization problem $\min_{\mathbf{y}\in Y, w\in\mathbb{R}}\{w : f(\mathbf{u}) + \mathbf{y}g(\mathbf{u}) \leq w, \mathbf{u} \in U\}$ the cut $w \geq f(\mathbf{u}_1) + \mathbf{y}g(\mathbf{u}_1)$ (is stronger than) and dominates the cut $w \geq f(\mathbf{u}_2) + \mathbf{y}g(\mathbf{u}_2)$, if $f(\mathbf{u}_1) + \mathbf{y}g(\mathbf{u}_1) \geq f(\mathbf{u}_2) + \mathbf{y}g(\mathbf{u}_2), \mathbf{y} \in Y$ and there is at least one $\mathbf{y} \in Y$ which makes this inequality strict. A cut is called *strong* or *pareto-optimal* if it is not dominated by any other cut (Magnanti and Wong, 1981). Note that it is possible to generate multiple Benders optimality cuts for a given master problem solution, because DSP may have alternative optimal solutions. Van Roy (1986) indicates that a cut derived from a particular dual optimal solution is strong if it is not dominated by a cut derived from any other dual optimal solution, and presents a two-phase approach to strengthen a Benders cut. We apply this approach to our problem. Observe that given a master solution $\hat{\mathbf{z}}$, the value of dual variable β^1_{ijk} with zero coefficient does not have any impact on the optimum objective value of DSP. Hence, we can modify β^1_{ijk} without changing the value of the objective function (28) when $\hat{z}_{ijk} = 0$. We can modify β^2_{ijk} similarly when $\hat{z}_{ijk} = 1$. Note that feasibility must be maintained during these modifications. Let Z be the index set of all beamlets at all control points, namely the set of all (i, j, k) index combinations. Also let $Z_0 \subseteq Z$ be the index set of beamlets where $\hat{z}_{ijk} = 0$ and $Z_1 \subseteq Z$ be the index set of beamlets where $\hat{z}_{ijk} = 1$ in the master solution $\hat{\mathbf{z}}$. First, we solve DSP and find an optimal dual solution. Then, dual variables are fixed at their optimal values except $\boldsymbol{\beta}^1$ and $\boldsymbol{\beta}^2$ with zero coefficients in the optimal objective, and $\boldsymbol{\beta}^3$. In other words, we determine new values of $\beta^1_{ijk}, (i, j, k) \in Z_0$, and $\beta^2_{ijk}, (i, j, k) \in Z_1$ by solving the following reduced DSP

RDSP $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}^1, \hat{\boldsymbol{\mu}}^2, \hat{\boldsymbol{\beta}}^1, \hat{\boldsymbol{\beta}}^2)$:

$$\max \sum_{k=1}^{K} \sum_{i=1}^{m} \sum_{j=1}^{n} (-\beta^1_{ijk} - \beta^2_{ijk}) \tag{47}$$

s.t.

$$-\sum_{v \in V} D_{ijkv} \hat{\pi}_v - \beta^1_{ijk} + \beta^2_{ijk} - \beta^3_{ijk} \leq 0 \qquad i = 1, \ldots, m; j = 1, \ldots, n; k = 1, \ldots, K \tag{48}$$

$$-\sum_{i=1}^{m} \sum_{j=1}^{n} (\beta^2_{ijk} - \beta^3_{ijk}) + \hat{\mu}^1_k - \hat{\mu}^2_k \leq 1 \qquad k = 1, \ldots, K \tag{49}$$

$$\beta^1_{ijk} = \hat{\beta}^1_{ijk} \qquad (i, j, k) \in Z_1 \tag{50}$$

$$\beta^2_{ijk} = \hat{\beta}^2_{ijk} \qquad (i, j, k) \in Z_0 \tag{51}$$

$$\boldsymbol{\beta}^1 \in \mathbb{R}^{mn|K|}_+; \boldsymbol{\beta}^2 \in \mathbb{R}^{mn|K|}_+; \boldsymbol{\beta}^3 \in \mathbb{R}^{mnK}_+. \tag{52}$$

In other words, we lift some of the $\mathbf{z}$ variables in the associated Benders cut without changing the objective function of DSP or violating the feasibility. Therefore, we obtain a strong Benders cut, since none of the cuts derived from an alternative optimal solution dominates (or is stronger than) this resulting one (Van Roy, 1986; Üster and Agrahari, 2011). It is worth noting that, in these studies, after setting unmodifiable dual variables to their optimal values, the remaining problem can be decomposed into subproblems and solved efficiently. Unfortunately, this is not possible in our case. Constraints (49) do not allow such decomposition. There exist other studies in the literature considering the use of strong cuts in Benders decomposition (Adulyasak et al., 2015; Lin, 2014).

5.3. Minimal Infeasible Subsystems and New Benders Cut Selection Strategy

We observe that it can take a long time to generate a feasibility cut during the initial iterations for large problem instances. There is a relatively new approach in the literature for generating Benders cuts (Fischetti et al., 2010) and stronger combinatorial cuts (Codato and Fischetti, 2006; Taşkın and Cevik, 2013). According to this approach it is possible to determine unbounded directions of a problem using an alternative polyhedron that is bounded. Fischetti et al. (2010) show that Benders subproblem can be converted into a pure feasibility problem, and that it is possible to obtain both feasibility and optimality cuts solving an alternative problem derived from this extended subproblem. Given a master solution $(\hat{\mathbf{z}}, \hat{\eta})$, the pure feasibility subproblem (PFSP) becomes

PFSP $(\hat{\mathbf{z}}, \hat{\eta})$:

$$\sum_{k \in K} mu_k \leq \hat{\eta} \qquad : \pi_0 \tag{53}$$

$$(10) - (12), (14) - (26),$$

where π_0 is the dual variable associated with (53). Observe that if $(\hat{\mathbf{z}}, \hat{\eta})$ is feasible for PFSP, then it is optimal for VMATP problem. Thus, a violated cut can be generated if and only if PFSP is infeasible, or equivalently, if its dual problem is unbounded. The dual of PFSP (DPFSP) can be written as

DPFSP $(\hat{\mathbf{z}}, \hat{\eta})$:

$$\max \sum_{k=1}^{K} \sum_{i=1}^{m} \sum_{j=1}^{n} U^{mu}(-\hat{z}_{ijk}\beta^1_{ijk} + (\hat{z}_{ijk} - 1)\beta^2_{ijk}) - \sum_{o=1}^{O} \theta^2_o U^{OAR}_o + \tag{54}$$

$$\sum_{t=1}^{T} \theta^1_t \overline{d}_t + \sum_{t=1}^{T} \sum_{v \in V^{TV}_t} (L^{TV}_t \epsilon^1_{tv} - U^{TV}_t \epsilon^2_{tv}) + \sum_{k=1}^{K} (L^{mu}\mu^1_k - U^{mu}\mu^2_k) - \pi_0 \hat{\eta}$$

s.t.

$$-\sum_{i=1}^{m} \sum_{j=1}^{n} (\beta^2_{ijk} - \beta^3_{ijk}) + \mu^1_k - \mu^2_k - \pi_0 \leq 0 \qquad k = 1, \ldots, K \tag{55}$$

$$\pi_0 \in \mathbb{R}_+ \tag{56}$$

$$(29) - (31), \;\; (33) - (37).$$

Note that $\mathbf{0}$ is the trivial solution of DPFSP. Therefore, for a given master solution $(\hat{\mathbf{z}}, \hat{\eta})$ if PFSP is infeasible, then associated DPFSP is unbounded. Given a ray $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\beta}}^1, \hat{\boldsymbol{\beta}}^2, \hat{\boldsymbol{\beta}}^3, \hat{\boldsymbol{\theta}}^1, \hat{\boldsymbol{\theta}}^2, \hat{\boldsymbol{\tau}}^1, \hat{\boldsymbol{\tau}}^2, \hat{\boldsymbol{\epsilon}}^1, \hat{\boldsymbol{\epsilon}}^2, \hat{\boldsymbol{\mu}}^1, \hat{\boldsymbol{\mu}}^2, \hat{\pi}_0)$ of DPFSP the associated cut is:

$$\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n} U^{mu}(-z_{ijk}\hat{\beta}^1_{ijk} + (z_{ijk}-1)\hat{\beta}^2_{ijk}) - \sum_{o=1}^{O}\hat{\theta}^2_o U^{OAR}_o + \\ \sum_{t=1}^{T}\hat{\theta}^1_t \overline{d}_t + \sum_{t=1}^{T}\sum_{v\in V^{TV}_t}(L^{TV}_t\hat{\epsilon}^1_{tv} - U^{TV}_t\hat{\epsilon}^2_{tv}) + \sum_{k=1}^{K}(L^{mu}\hat{\mu}^1_k - U^{mu}\hat{\mu}^2_k) - \hat{\pi}_0\eta \leq 0. \tag{57}$$

Furthermore, the unbounded objective function is set to 1 for normalization as done by Gleeson and Ryan (1990), and

$$\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n} U^{mu}(-\hat{z}_{ijk}\beta^1_{ijk} + (\hat{z}_{ijk}-1)\beta^2_{ijk}) - \sum_{o=1}^{O}\theta^2_o U^{OAR}_o + \\ \sum_{t=1}^{T}\theta^1_t \overline{d}_t + \sum_{t=1}^{T}\sum_{v\in V^{TV}_t}(L^{TV}_t\epsilon^1_{tv} - U^{TV}_t\epsilon^2_{tv}) + \sum_{k=1}^{K}(L^{mu}\mu^1_k - U^{mu}\mu^2_k) - \pi_0\hat{\eta} = 1 \tag{58}$$

$$(29) - (31),\ \ (33) - (37),\ \ (55) - (56)$$

is the resulting alternative polyhedron. The alternative problem (AP)

AP $(\hat{\mathbf{z}}, \hat{\eta})$:

$$\min \pi_0 \tag{59}$$

s.t.

$$(29) - (31),\ \ (33) - (37),\ \ (55) - (56), (58)$$

minimizes π_0 over this polyhedron and we solve AP instead of DSP in Benders iterations to generate Benders cuts. Fischetti et al. (2010) state that when the objective of this problem is to minimize only π_0 then the original Benders' dual problem (DSP) arises. They also state that a feasibility cut or an optimality cut is generated depending on the optimal value of π_0: $\hat{\pi}_0 = 0$ implies a feasibility cut since DSP($\hat{\mathbf{z}}$) is unbounded. Observe that an optimal solution $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\beta}}^1, \hat{\boldsymbol{\beta}}^2, \hat{\boldsymbol{\beta}}^3, \hat{\boldsymbol{\theta}}^1, \hat{\boldsymbol{\theta}}^2, \hat{\boldsymbol{\tau}}^1, \hat{\boldsymbol{\tau}}^2, \hat{\boldsymbol{\epsilon}}^1, \hat{\boldsymbol{\epsilon}}^2, \hat{\boldsymbol{\mu}}^1, \hat{\boldsymbol{\mu}}^2, \hat{\pi}_0 = 0)$ of AP that satisfies constraints (55) and (58) provides an unbounded direction for DSP($\hat{\mathbf{z}}$). It can be shown that for any $\lambda > 0$, $(\lambda\hat{\boldsymbol{\pi}}, \lambda\hat{\boldsymbol{\beta}}^1, \lambda\hat{\boldsymbol{\beta}}^2, \lambda\hat{\boldsymbol{\beta}}^3, \lambda\hat{\boldsymbol{\theta}}^1, \lambda\hat{\boldsymbol{\theta}}^2, \lambda\hat{\boldsymbol{\tau}}^1, \lambda\hat{\boldsymbol{\tau}}^2, \lambda\hat{\boldsymbol{\epsilon}}^1, \lambda\hat{\boldsymbol{\epsilon}}^2, \lambda\hat{\boldsymbol{\mu}}^1, \lambda\hat{\boldsymbol{\mu}}^2)$ remains feasible for DSP($\hat{\mathbf{z}}$) (since constraint (32) remains feasible in addition to other constraints of DSP($\hat{\mathbf{z}}$)) and objective value becomes λ. If $\hat{\pi}_0 > 0$ then $(\hat{\boldsymbol{\pi}}/\hat{\pi}_0, \hat{\boldsymbol{\beta}}^1/\hat{\pi}_0, \hat{\boldsymbol{\beta}}^2/\hat{\pi}_0, \hat{\boldsymbol{\beta}}^3/\hat{\pi}_0, \hat{\boldsymbol{\theta}}^1/\hat{\pi}_0, \hat{\boldsymbol{\theta}}^2/\hat{\pi}_0, \hat{\boldsymbol{\tau}}^1/\hat{\pi}_0, \hat{\boldsymbol{\tau}}^2/\hat{\pi}_0, \hat{\boldsymbol{\epsilon}}^1/\hat{\pi}_0, \hat{\boldsymbol{\epsilon}}^2/\hat{\pi}_0, \hat{\boldsymbol{\mu}}^1/\hat{\pi}_0, \hat{\boldsymbol{\mu}}^2/\hat{\pi}_0)$ is an optimal solution for DSP($\hat{\mathbf{z}}$) with optimal objective value $1/\hat{\pi}_0 + \hat{\eta}$. Observe that we can derive a feasible solution for DSP($\hat{\mathbf{z}}$) from each one of the feasible solutions of AP($\hat{\mathbf{z}}, \hat{\eta}$) where $\bar{\pi}_0 > 0$ dividing this solution by $\bar{\pi}_0$. The optimal (minimum) objective value of AP($\hat{\mathbf{z}}, \hat{\eta}$) is $\hat{\pi}_0$, hence we reach an optimal solution with maximum objective value of DSP($\hat{\mathbf{z}}$). Additionally, we can solve RDSP using this optimal solution and generate pareto-optimal cuts.

5.4. Combinatorial Benders Cut

Combinatorial Benders decomposition is an extension of traditional Benders decomposition method, where the problem is again decomposed into an integer programming master problem and a linear programming subproblem. Rahmaniani et al. (2017) explain the difference between the two methods and state that combinatorial Benders decomposition does not use the dual information to generate cuts. The master problem is a binary integer programming problem (BIP) and when the subproblem is infeasible a combinatorial Benders cut similar to (60) is derived and used as a feasibility cut.

Assume that for a given feasible master solution $\hat{\mathbf{z}}$, it is not possible to find a feasible treatment, which means the subproblem is infeasible. In this case, another valid inequality may be generated according to the following observation: the subproblem may be infeasible with respect to partial-volume constraints (19)–(20) associated with a TV, OAR (24)–(25), or both. For these cases, to repair infeasiblity, we should do at least one of the following: open at least one of the closed beamlets, close at least one of the open beamlets, or both. Furthermore, the candidate beamlet ($\hat{z}_{ijk}$) to open or close must have positive effect on at least one voxel. Namely, the entries of the $\mathbf{D}$ matrix must be "strictly" positive for at least one v (otherwise, they will be all zero for a specific combination of i, j, k and hence can be removed). Let $I \subseteq Z$ be the index set of the beamlets having strictly positive effect on at least one voxel, namely $I = \{(i, j, k) : D_{ijkv} > 0, v \in V\}$. Hence, we can add the combinatorial cut

$$\sum_{\substack{\hat{z}_{ijk}=0 \\ (i,j,k)\in I}} z_{ijk} + \sum_{\substack{\hat{z}_{ijk}=1 \\ (i,j,k)\in I}} (1 - z_{ijk}) \geq 1. \tag{60}$$

to the RMP each time an infeasible solution is obtained.

This cut is not tight according to our preliminary results obtained on random samples. Thus, as in the study of Taşkın and Cevik (2013), we find a minimal infeasible system (MIS) of the subproblem when an infeasible solution is detected. Gleeson and Ryan (1990) show that there is one-to-one correspondence between MISs of an infeasible linear system and the supports of vertices of the related alternative polyhedron. Thus, solving AP instead of the original dual problem not only provides Benders cuts, but also detects an MIS each time π_0 is found to be zero. Let $Z^* \subseteq Z$ be the index set of the beamlets that are associated with the MIS corresponding to $\hat{\mathbf{z}}$. The cut (60) is revised so that it only has $\mathbf{z}$ variables in $I \cap Z^*$:

$$\sum_{\substack{\hat{z}_{ijk}=0 \\ (i,j,k)\in I\cap Z^*}} z_{ijk} + \sum_{\substack{\hat{z}_{ijk}=1 \\ (i,j,k)\in I\cap Z^*}} (1 - z_{ijk}) \geq 1. \tag{61}$$

In the final version of our Benders decomposition algorithm, each time a Benders feasibility cut is added to the master problem we also add a constraint of type (61). The resulting Benders algorithm including the improvement strategies explained so far is given in Figure 7 within the dotted frames. We refer to this algorithm as Improved Benders Algorithm 1.

In addition to these strategies, we also use a single branch-and-bound tree, which has received widespread attention in the literature recently (Lin, 2014; Taşkın and Cevik, 2013). Even though it is not proved theoretically that using this strategy outperforms the naive form, practical results reveal its superiority. In the naive form, each time a Benders cut is added to RMP it is solved from scratch. This makes Benders decomposition more and more expensive as the number of cuts increases. Instead, we solve RMP using only one branch and bound tree using the solver's callback mechanism. In our implementation each time a new incumbent is found a new Benders cut is generated and added to RMP or otherwise the incumbent is accepted.

We observe an important difference between in the implementation of the new cut selection strategy explained in Section 5.3. In the naive form of Benders decomposition, if RMP returns a solution $(\hat{\mathbf{z}}, \hat{\eta})$ which is found to be feasible for SP, an optimality cut is added to RMP and the upper bound of the entire algorithm is updated. Thus, if the same solution is chosen by RMP for the second time with the updated objective value $(\hat{\mathbf{z}}, \bar{\eta})$, the lower bound and the upper bound of the problem are equal. The reason is that RMP is solved to optimality in each iteration and its optimal objective value always provides a lower bound for the whole problem. Therefore, when PFSP becomes feasible, AP becomes infeasible, the optimality gap becomes zero and the algorithm stops. On the other hand, in the callback implementation when an incumbent solution $(\hat{\mathbf{z}}, \hat{\eta})$ is obtained for the first time, which is found to be feasible for SP also, similarly an optimality cut is added to RMP. However, an incumbent solution does not provide a lower bound for the whole problem, if it is not optimal, as in the naive implementation; but if it is returned one more time, it is certain that the current upper bound in the branch and bound is higher than the objective value of this solution. Otherwise, the associated search node of the branch and bound tree would have been pruned. Re-obtaining an incumbent solution means that the callback can accept it and update the upper bound. In summary if PFSP is feasible, AP is infeasible, then the algorithm does not stop and continues until the optimality gap falls below a certain level.

5.5. A Relaxation of VMATP

According to the results of the algorithm obtained by implementing the improvement strategies explained so far we can say that the lower bound is not strong. To rectify this, we relax VMATP model by removing the geometry constraints and solve the resulting relaxation (RVMATP)

RVMATP:

$$\min \sum_{k \in K} mu_k \tag{27}$$

$$(10) - (12); \;\; (14); \;\; (17) - (26).$$

Note that RVMATP is an LP model. As we also discuss in Section 6, the lower bound obtained solving this relaxation is remarkably stronger and improves the optimality gap. However, since we relax the geometry constraints, it is not possible to obtain the exact information about the aperture shapes. Hence, the lower bound obtained by this relaxed model can only be used to calculate the optimality gap. Nevertheless, the optimal solution of RVMATP gives the radiation dose intensity at each of the control points, given these radiation intensities we can try to determine a feasible solution for the LP relaxation of VMATP (LPVMATP). If LPVMATP is feasible for the given radiation intensities, we have enough information about the aperture shapes (i.e. values for $\hat{\mathbf{z}}$ variables) to generate a cut. Notice that these $\hat{\mathbf{z}}$ variables can be fractional; but still given fractional $\hat{\mathbf{z}}$ values, we solve DSP to obtain an optimality cut (39), which we add to RMP at the beginning of the callback implementation. The fractional $\hat{\mathbf{z}}$ vector changes the objective function of DSP only and gives another extreme point in its feasible region. The optimality cut obtained using this extreme point is valid for the LP relaxation of RMP, thus it is also valid for RMP. We call the resulting algorithm as Improved Benders Algorithm 2, which we illustrate in Figure 7 by appending the steps remaining outside the dotted frames.

6. Computational Results

6.1. Test Bed

We use a real data set belonging to an anonymous prostate patient from the Common Optimization for Radiation Therapy (CORT) datasets provided by Craft et al. (2014). The dataset contains beams for 180 equispaced co-planar beam angles (control points). There are 13 rows and 16 columns in the MLC. The size of each beamlet is 1 cm^2, and there are 25,404 beamlets in total. Furthermore, there are 9 different volumes: 2 TVs (PTV 68 and PTV 56), 5 OARs (bladder, left femoral head, right femoral head, penile bulb, and rectum), and 2 other tissues (prostate bed and lymph), which are not taken into consideration in the experiments, since they are covered by target and irradiated. The prescription doses of the TVs are different; the highest prescription dose target is called PTV 68 and the lowest prescription dose target is called PTV 56. The total number of voxels in the patient is 690,373 and the voxels are of size 3x3x3 mm^3. The size of the original data is very large and none of the optimization models and algorithms we discuss in this paper is able to find feasible solutions. Therefore, we reduce the size of the problem. To this end we only change the size of the patient's body and keep other parameters such as the total number of control points, the total number of beamlets, etc. same as in the original data. We assume that there is only one TV (PTV 68) and one OAR. Instead of considering only one of them we combine all OARs into a single volume. Another approach for reducing the size of the problem is to use down-sampled voxels. We randomly select a number of voxels from each of the structures as given in Table 3 and obtain the sampled data sets used in the computational experiments.

Table 3: Number of voxels in the complete data set and reduced data sets

REGION	Complete Data	v-22	v-44	v-66	v-88	v-220	v-660	v-880	v-1100	v-1301	v-1501	v-1701
PTV 68	6770	10	20	30	40	100	300	400	500	600	700	800
Bladder	11596	4	8	12	16	40	120	160	200	240	280	320
Left f. h.	5857	2	4	6	8	20	60	80	100	120	140	160
Right f. h.	5974	2	4	6	8	20	60	80	100	120	140	160
Penile b.	101	2	4	6	8	20	60	80	100	101	101	101
Rectum	1764	2	4	6	8	20	60	80	100	120	140	160
TOTAL	32062	22	44	66	88	220	660	880	1100	1301	1501	1701

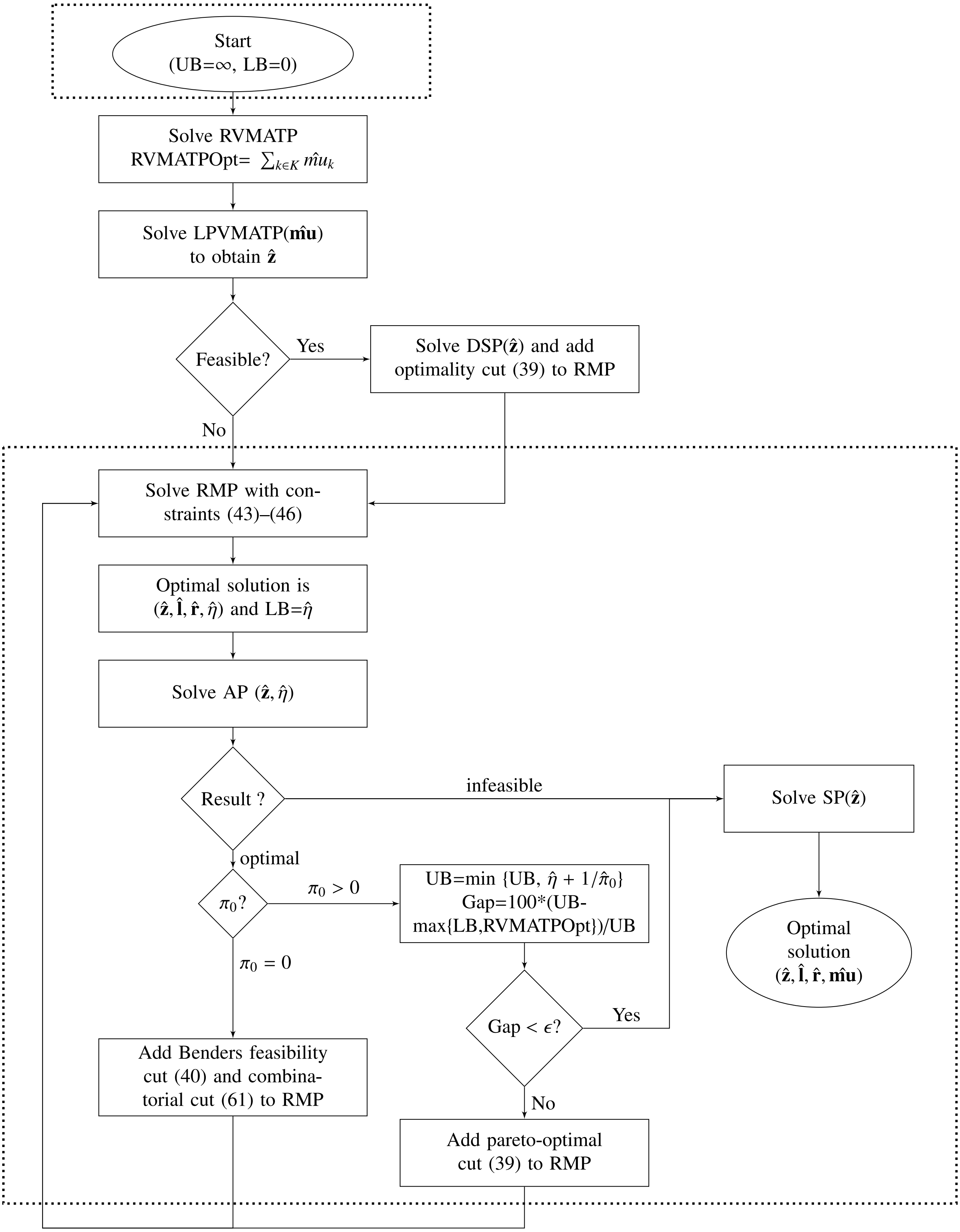

Start
(UB=∞, LB=0)
Solve RVMATP
RVMATPOpt= $\sum_{k\in K}\hat{mu}_k$
Solve LPVMATP($\hat{\mathbf{mu}}$)
to obtain $\hat{\mathbf{z}}$
Feasible?
Yes
Solve DSP($\hat{\mathbf{z}}$) and add optimality cut (39) to RMP
No
Solve RMP with constraints (43)–(46)
Optimal solution is ($\hat{\mathbf{z}}, \hat{\mathbf{l}}, \hat{\mathbf{r}}, \hat{\eta}$) and LB=$\hat{\eta}$
Solve AP ($\hat{\mathbf{z}}, \hat{\eta}$)
Result ?
infeasible
Solve SP($\hat{\mathbf{z}}$)
Optimal solution
($\hat{\mathbf{z}}, \hat{\mathbf{l}}, \hat{\mathbf{r}}, \hat{\mathbf{mu}}$)
optimal
π_0?
$\pi_0 > 0$
UB=min {UB, $\hat{\eta} + 1/\hat{\pi}_0$}
Gap=100*(UB-max{LB,RVMATPOpt})/UB
$\pi_0 = 0$
Gap < ϵ?
Yes
No
Add Benders feasibility cut (40) and combinatorial cut (61) to RMP
Add pareto-optimal cut (39) to RMP

Figure 7: Improved Benders decomposition algorithms

There are 11 different data sets with different names. For example, the data set v-22 consists of two structures and there are 22 voxels: 10 voxels are randomly selected from PTV 68 and 12 voxels are randomly selected from the OARs as shown in the third column of Table 3. Note that we select the voxels from the regions of the OARs that do not intersect with PTVs. There are 180 control points and 25,404 beamlets as in the original data. We generate 5 different instances for each size, which makes 55 instances in total, to be used in the computational experiments.

6.2. Computational Experiments

We implemented all algorithms and optimization models in Python 2.7 programming language (Python, 2015) and used Gurobi 6.5 solver (Gurobi, 2016) running on a computer with Windows Server 2012 R2 Standard 64-bit PC with 2.00 GHz Intel Xeon CPU, 46 GB RAM. For each size of data sets we generated 5 different samples from the real data, and solved VMATP by Gurobi, naive Benders algorithm, and the two improved Benders algorithms. We report CPU time and optimality gap for each run. We set a limit of 3600 seconds on the running time and executed all algorithms on one thread. We changed the default method for the RVMATP and AP models in improved Benders algorithms and solved them by barrier algorithm. Also we set the "MIPFocus" parameter value of the master model in all Benders algorithms to 3 to focus on the bound. We executed Gurobi solver with the default settings and did not perform any parameter tuning while solving VMATP.

We assume that between two consecutive control points a leaf can move at most 2 beamlets and set $\delta = 2$. α_1^{OAR} and α_1^{TV} are set to be 0.40 and 0.95, respectively. We also set the number of fractions of radiation therapy 34, where for one fraction $\overline{d}_1$, U_1^{TV}, L_1^{TV}, and U_1^{OAR} are set to 2 Gy, 2.14 Gy, 1.9 Gy, and 1.47 Gy respectively (i.e. total values for 34 fractions are 68 Gy, 72.76 Gy, 64.6 Gy, and 50 Gy). We also assume that the gantry speed is constant and let a tour be completed in 3 minutes. The maximum dose rate of a linear accelerator is generally 600 MU/min, therefore dose rate can be at most 10 MU/second, and dose intensity can be at most 10 MU at a control point.

Table 4 summarizes the computational results; it includes the average optimality gap (%) and the average CPU time (second) of five instances of each size. Also, the column with title "S/T" shows the number of instances that the corresponding method finds a feasible solution out of five instances. It is not possible to calculate optimality gap whenever a method cannot find a lower bound and/or an upper bound for a test instance. In order to be able to calculate average optimality gaps we assume that the lower bound of the objective value is 0. This is reasonable because the radiation intensity at a control point is within (0, 10). We accept the optimality gaps of such instances be 100%. Similarly, the upper bound on the objective value is 1800 since total number of control points is 180. Thus, whenever a method cannot provide an upper bound for a test instance we calculate the optimality gap by setting its upper bound to 1800. Detailed results including bounds, optimality gap (%) and CPU time (second) of each instance can be found in Appendix 1. According to the results naive Benders decomposition fails in both performance measures compared to others. It can only find a feasible solution with high total radiation for some instances. For all instances the lower bound remains at zero level, which means an optimality gap of 100%. On the other hand, Gurobi outperforms naive and improved Benders algorithms in both performance measures when the size of instances are small (i.e. total number of voxels is less than or equal to 220). Note that the difference between the average optimality gaps obtained by Gurobi and Improved Benders Algorithm 2 is not significant. As the size of the problem increases Gurobi cannot find a feasible solution for some of the instances within the given time limit. For example, it can compute a feasible solution only two out of five instances having 880 voxels to optimality, but it can neither find a feasible solution nor a lower bound for the remaining three instances. On the other hand, improved Benders algorithms can find feasible solutions for all instances with small average optimality gaps (3.12% and 3.23%, respectively), which indicates that a high-quality plan is found for each instance. Furthermore, for only one instance (out of five) with size 1501 voxels, the improved Benders algorithms cannot find a feasible solution.

When we compare improved Benders algorithms, we observe that finding a better lower bound by solving the relaxation (RVMATP) and also introducing the initial optimality cut derived from an optimal solution of LPVMATP improves the performance of Benders algorithm. The CPU times are similar and neither one outperforms the other. However, optimality gaps decrease in almost all problems in the Improved Benders Algorithm 2. For instance, the average gap is 13.59% for the problem having 1701 voxels and decreases down to 0.49%. The reason is that in almost all problems the lower bound is close to the optimal objective value in the Improved Benders Algorithm 2. Also, it can provide feasible solutions that are very close to the optimal value for almost all large problems, but still it cannot solve them to optimality within the time limit. Nevertheless, we can conclude that Improved Benders Algorithm 2 is capable of finding good treatment plans even for larger problem instances. There are also other studies that use CORT datasets in VMAT planning within a different settings such as (Mahnam et al., 2017; Balvert et al., 2017; Balvert and Craft, 2017)). They all provide treatment plans satisfying different set of constraints and minimizing or maximizing different objective functions, which makes them incomparable.

Table 4: Average results

	Gurobi			Naive Benders A.			Impr. Benders A. 1			Impr. Benders A. 2		
# OF VOXELS	GAP	CPU	S/T*	GAP	CPU	S/T*	GAP	CPU	S/T*	GAP	CPU	S/T*
22	0.00	62.4	5/5	100	3600	4/5	0.00	1502.3	5/5	0.00	112.2	5/5
44	0.00	63.1	5/5	100	3600	5/5	0.00	245.6	5/5	0.00	627.8	5/5
66	0.01	809.8	5/5	100	3600	3/5	2.33	2330.8	5/5	0.14	1884.2	5/5
88	0.00	104.8	5/5	100	3600	1/5	0.00	243.4	5/5	0.00	826.9	5/5
220	0.00	1455.0	5/5	100	3600	2/5	0.61	1661.9	5/5	0.02	1187.1	5/5
660	20.00	2437.2	4/5	100	3600	2/5	6.12	3333.7	5/5	0.22	3585.9	5/5
880	66.67	3067.1	2/5	100	3600	0/5	3.12	3600	5/5	3.23	3600	5/5
1100	40.00	2690.6	3/5	100	3600	0/5	4.64	3600	5/5	0.96	3600	5/5
1301	40.00	2930.6	3/5	100	3600	1/5	3.21	3600	5/5	0.34	3600	5/5
1501	60.00	2861.8	2/5	100	3600	0/5	19.38	3600	4/5	18.43	3600	4/5
1701	40.00	2286.3	3/5	100	3600	0/5	13.59	3600	5/5	0.49	3600	5/5

7. Conclusions

VMAT treatment planning is an important but difficult issue in cancer treatment. It is challenging to develop good formulations and efficient methods that solve the problem exactly and find good treatment plans. To this end we first formulate a MILP model, which is new in a couple of aspects. First of all the decision variables to formulate an aperture and structure of the corresponding technical constraints are new. Also, it has all partial-volume and full-volume constraints, and does not try to find a feasible treatment by minimizing total overdose of OAR or underdose of TV. Instead, our objective is to find a solution that delivers as little radiation as possible to the patient. To the best of our knowledge, this objective has not been used before in any VMAT planning. Furthermore, we propose two efficient Benders decomposition algorithms for the exact solution of the model. A real clinical data is used to compare Gurobi with the naive and improved Benders algorithms regarding solution time and optimality gap. Since, it is not possible to solve the problem either by Gurobi or Benders decomposition algorithms using original size of data, we coarse it by randomly sampling voxels in all volumes. We generated a data set consisting of 55 instances, 5 for each of 11 different sample sizes. Our improved Benders algorithms are able to find feasible solutions almost for all instances and Improved Benders Algorithm 2 provides smaller optimality gap than Improved Benders Algorithm 1 in all instances (with the notable exception of 880-5), within the time limit of 3600 seconds. However, Gurobi can neither provide a finite lower bound nor find a feasible solution in many instances. Our results point out that VMAT planning is a promising new research area for the application of mathematical optimization techniques and there is possibility to develop exact methods which can determine optimal treatment plans. The problems are large scale, which makes the use of the new decomposition strategies and their contributions with heuristic approaches as potential future research.

Acknowledgements

This research has been partially supported by the Boğaziçi University Research Fund Grant No 11520-16A03D1. The authors would like to thank very kindly to Prof. Hatice Bilge, Asst. Prof. Murat Okutan and Prof. Ethem Nezih Oral from Istanbul University Oncology Institute for their guidance. We would also like to thank two anonymous reviewers for their constructive comments.

Appendix 1. Detailed Computational Results

	Gurobi				Naive Benders A.				Impr. Benders A. 1				Impr. Benders A. 2			
INSTANCE	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU
22-1	236.550	236.550	0.00	92.7	0.00	N/A	N/A*	3600	236.545	236.589	0.02	3600	236.550	236.550	0.00	85.3
22-2	231.148	231.148	0.00	48.8	0.00	523.342	100	3600	231.148	231.148	0.00	100.1	231.148	231.150	0.00	79.8
22-3	232.605	232.606	0.00	52.1	0.00	429.773	100	3600	232.605	232.605	0.00	2002.7	232.605	232.611	0.00	259.9
22-4	233.700	233.700	0.00	52.8	0.00	535.893	100	3600	233.700	233.700	0.00	1650.5	233.700	233.705	0.00	77.5
22-5	234.742	234.742	0.00	65.6	0.00	401.074	100	3600	234.742	234.742	0.00	157.9	234.742	234.750	0.00	58.8
44-1	232.912	232.912	0.00	49.0	0.00	414.703	100	3600	232.912	232.912	0.00	127.5	232.912	232.930	0.00	136.8
44-2	236.830	236.843	0.00	105.5	0.00	403.192	100	3600	236.830	236.830	0.00	671.3	236.830	236.830	0.00	2658.4
44-3	232.912	232.912	0.00	49.9	0.00	414.703	100	3600	232.912	232.912	0.00	134.7	232.912	232.930	0.00	137.2
44-4	236.838	236.838	0.00	63.6	0.00	452.829	100	3600	236.838	236.838	0.00	126.2	236.838	236.841	0.00	82.8
44-5	236.429	236.429	0.00	47.4	0.00	563.134	100	3600	236.429	236.429	0.00	168.4	236.429	236.436	0.00	123.6
66-1	236.830	236.830	0.00	110.2	0.00	N/A	N/A*	3600	236.824	236.834	0.00	681.2	236.830	236.835	0.00	1648.0
66-2	237.597	237.610	0.00	77.4	0.00	1022.734	100	3600	234.865	237.616	1.16	3600	237.597	237.620	0.00	454.0
66-3	236.451	236.451	0.00	42.1	0.00	845.769	100	3600	236.451	236.451	0.00	172.7	236.451	236.458	0.00	118.9
66-4	234.940	234.942	0.00	219.2	0.00	N/A	N/A*	3600	223.150	235.888	5.40	3600	234.940	236.192	0.53	3600
66-5	237.732	237.871	0.06	3600	0.00	946.373	100	3600	226.006	238.085	5.07	3600	237.730	238.085	0.15	3600
88-1	236.830	236.836	0.00	101.7	0.00	579.613	100	3600	236.830	236.830	0.00	321.2	236.830	236.869	0.02	3600
88-2	237.050	237.050	0.00	59.9	0.00	N/A	N/A*	3600	237.050	237.050	0.00	201.2	237.050	237.061	0.00	136.8
88-3	236.645	236.663	0.00	181.8	0.00	N/A	N/A*	3600	236.645	236.645	0.00	313.4	236.645	236.661	0.00	167.0
88-4	237.048	237.051	0.00	123.7	0.00	N/A	N/A*	3600	237.048	237.048	0.00	189.1	237.048	237.065	0.00	148.9
88-5	235.866	235.866	0.00	56.8	0.00	N/A	N/A*	3600	235.866	235.866	0.00	192.2	235.866	235.867	0.00	82.0
220-1	236.864	236.864	0.00	804.8	0.00	562.247	100	3600	236.781	236.954	0.07	3600	236.864	236.877	0.00	301.6
220-2	237.067	237.072	0.00	621.4	0.00	N/A	N/A*	3600	237.066	237.067	0.00	286.2	237.067	237.068	0.00	401.5
220-3	237.924	237.924	0.00	1315.0	0.00	N/A	N/A*	3600	237.924	237.924	0.00	503.0	237.924	237.936	0.00	1352.0
220-4	237.028	237.036	0.00	934.1	0.00	468.199	100	3600	237.028	237.028	0.00	320.2	237.028	237.042	0.00	280.4
220-5	237.871	237.899	0.01	3600	0.00	N/A	N/A*	3600	230.946	238.052	2.99	3600	237.870	238.094	0.09	3600
660-1	237.666	237.681	0.00	2468.7	0.00	N/A	N/A*	3600	237.665	237.687	0.00	2876.9	237.666	237.968	0.13	3600
660-2	237.847	237.847	0.00	3142.0	0.00	N/A	N/A*	3600	237.847	237.847	0.00	2991.8	237.847	239.098	0.52	3600

	Gurobi				Naive Benders A.				Impr. Benders A. 1				Impr. Benders A. 2			
INSTANCE	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU
660-3	237.349	237.351	0.00	1273.8	0.00	477.336	100	3600	228.640	304.016	24.79	3600	237.349	237.918	0.24	3600
660-4	238.339	238.341	0.00	1701.3	0.00	505.473	100	3600	227.578	239.377	4.93	3600	238.339	238.818	0.20	3600
660-5	N/A	N/A	N/A*	3600	0.00	N/A	N/A*	3600	236.574	238.602	0.85	3600	237.869	237.882	0.00	3529.7
880-1	N/A	N/A	N/A*	3600	0.00	N/A	N/A*	3600	228.095	241.029	5.37	3600	238.177	239.618	0.60	3600
880-2	237.412	237.435	0.00	2098.3	0.00	N/A	N/A*	3600	236.412	238.197	0.75	3600	237.412	238.288	0.37	3600
880-3	N/A	N/A	N/A*	3600	0.00	N/A	N/A*	3600	233.330	239.691	2.65	3600	238.357	238.462	0.04	3600
880-4	N/A	N/A	N/A*	3600	0.00	N/A	N/A*	3600	233.209	237.652	1.87	3600	237.214	237.878	0.28	3600
880-5	238.006	238.017	0.00	1904.6	0.00	N/A	N/A*	3600	228.188	240.092	4.96	3600	238.006	279.477	14.84	3600
1100-1	N/A	N/A	N/A*	3600	0.00	N/A	N/A*	3600	228.976	256.198	10.63	3600	238.082	239.495	0.59	3600
1100-2	237.714	237.733	0.00	1181.6	0.00	N/A	N/A*	3600	228.976	240.046	4.61	3600	237.714	246.745	3.66	3600
1100-3	237.165	237.165	0.00	3402.8	0.00	N/A	N/A*	3600	234.575	237.455	1.21	3600	237.165	237.469	0.13	3600
1100-4	N/A	N/A	N/A*	3600	0.00	N/A	N/A*	3600	231.437	237.964	2.74	3600	237.321	237.907	0.25	3600
1100-5	237.801	237.801	0.00	1668.6	0.00	N/A	N/A*	3600	229.358	238.886	3.99	3600	237.801	238.182	0.16	3600
1301-1	237.785	237.797	0.00	1952.3	0.00	N/A	N/A*	3600	227.114	238.257	4.68	3600	237.785	238.507	0.30	3600
1301-2	N/A	N/A	N/A*	3600	0.00	546.717	100	3600	230.157	239.793	4.02	3600	238.405	239.530	0.47	3600
1301-3	237.901	237.912	0.00	2409.6	0.00	N/A	N/A*	3600	234.391	238.672	1.79	3600	237.901	238.880	0.41	3600
1301-4	N/A	N/A	N/A*	3600	0.00	N/A	N/A*	3600	233.119	239.026	2.47	3600	238.137	238.875	0.31	3600
1301-5	237.386	237.389	0.00	3091.2	0.00	N/A	N/A*	3600	230.483	237.840	3.09	3600	237.386	237.824	0.18	3600
1501-1	N/A	N/A	N/A*	3600	0.00	N/A	N/A*	3600	227.807	N/A	N/A**	3600	237.770	249.340	4.64	3600
1501-2	N/A	N/A	N/A*	3600	0.00	N/A	N/A*	3600	235.473	238.039	1.08	3600	237.459	237.629	0.07	3600
1501-3	N/A	N/A	N/A*	3600	0.00	N/A	N/A*	3600	232.667	238.891	2.61	3600	237.925	N/A	N/A**	3600
1501-4	237.925	237.925	0.00	1178.8	0.00	N/A	N/A*	3600	230.816	238.846	3.36	3600	237.925	238.817	0.37	3600
1501-5	237.665	237.687	0.00	2330.2	0.00	N/A	N/A*	3600	232.797	238.808	2.52	3600	237.665	238.351	0.29	3600
1701-1	N/A	N/A	N/A*	3600	0.00	N/A	N/A*	3600	230.406	238.928	3.57	3600	238.008	238.892	0.37	3600
1701-2	N/A	N/A	N/A*	3600	0.00	N/A	N/A*	3600	229.234	529.396	56.7	3600	237.823	238.260	0.18	3600
1701-3	238.105	238.123	0.00	1440.9	0.00	N/A	N/A*	3600	232.533	238.855	2.65	3600	238.105	239.137	0.43	3600
1701-4	237.896	237.908	0.00	1183.3	0.00	N/A	N/A*	3600	232.905	238.972	2.54	3600	237.896	240.371	1.03	3600
1701-5	237.677	237.677	0.00	1607.6	0.00	N/A	N/A*	3600	232.831	238.824	2.51	3600	237.677	238.754	0.45	3600

Note: *Calculated as 100% since LB and UB are unavailable **Calculated as (100(1800-LB)/1800)% since UB is unavailable

References

Adulyasak, Y., Cordeau, J.-F., Jans, R., 2015. Benders decomposition for production routing under demand uncertainty. Operations Research 63 (4), 851–867.

Akartunalı, K., Mak-Hau, V., Tran, T., 2015. A unified mixed-integer programming model for simultaneous fluence weight and aperture optimization in VMAT, Tomotherapy, and Cyberknife. Computers & Operations Research 56, 134–150.

Baatar, D., Boland, N., Brand, S., Stuckey, P. J., 2007. Minimum cardinality matrix decomposition into consecutive-ones matrices: CP and IP approaches. In: International Conference on Integration of Artificial Intelligence (AI) and Operations Research (OR) Techniques in Constraint Programming. Springer, pp. 1–15.

Baatar, D., Hamacher, H. W., Ehrgott, M., Woeginger, G. J., 2005. Decomposition of integer matrices and multileaf collimator sequencing. Discrete Applied Mathematics 152 (1), 6–34.

Balvert, M., Craft, D., 2017. Fast approximate delivery of fluence maps for IMRT and VMAT. Physics in medicine and biology 62 (4), 1225.

Balvert, M., et al., 2017. Improving the quality, efficiency and robustness of radiation therapy planning and delivery through mathematical optimization. Tech. rep., Tilburg University, School of Economics and Management.

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

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.

Bilge, H., Okutan, M., Oral, E. N., 2017. private communications. Oncology Institute, İstanbul University.

Boland, N., Hamacher, H. W., Lenzen, F., 2004. Minimizing beam-on time in cancer radiation treatment using multileaf collimators. Networks 43 (4), 226–240.

Bzdusek, K., Friberger, H., Eriksson, K., Hårdemark, B., Robinson, D., Kaus, M., 2009. Development and evaluation of an efficient approach to volumetric arc therapy planning. Medical physics 36 (6), 2328–2339.

Cambazard, H., O'Mahony, E., O'Sullivan, B., 2012. A shortest path-based approach to the multileaf collimator sequencing problem. Discrete Applied Mathematics 160 (1), 81–99.

Cameron, C., 2005. Sweeping-window arc therapy: an implementation of rotational IMRT with automatic beam-weight calculation. Physics in medicine and biology 50 (18), 4317–4336.

Cao, D., Afghan, M. K., Ye, J., Chen, F., Shepard, D. M., 2009. A generalized inverse planning tool for volumetric-modulated arc therapy. Physics in medicine and biology 54 (21), 6725–6738.

Carlsson, F., 2008. Combining segment generation with direct step-and-shoot optimization in intensity-modulated radiation therapy. Medical physics 35 (9), 3828–3838.

Codato, G., Fischetti, M., 2006. Combinatorial Benders' cuts for mixed-integer linear programming. Operations Research 54 (4), 756–766.

Craft, D., Bangert, M., Long, T., Papp, D., Unkelbach, J., 2014. Shared data for intensity modulated radiation therapy (IMRT) optimization research: the CORT dataset. GigaScience 3 (1), 1.

Craft, D., McQuaid, D., Wala, J., Chen, W., Salari, E., Bortfeld, T., 2012. Multicriteria VMAT optimization. Medical physics 39 (2), 686–696.

de Araújo Montagno, E., Sabbatini, R. M. E., 1997. Radiosurgery. Accessed: 2018-01-23.
URL http://www.cerebromente.org.br/n02/tecnologia/radiocirurg_i.htm

Dursun, P., Taşkın, Z. C., Altınel, İ. K., 2016. Mathematical models for optimal volumetric modulated arc therapy (VMAT) treatment planning. In: Procedia Computer Science (Proceedings of International Conference on Health and Social Care Information Systems and Technologies, HCist, Porto). Vol. 100. pp. 644–651.

Earl, M., Shepard, D., Naqvi, S., Li, X., Yu, C., 2003. Inverse planning for intensity-modulated arc therapy using direct aperture optimization. Physics in medicine and biology 48 (8), 1075–1089.

Ehrgott, M., Güler, Ç., Hamacher, H. W., Shao, L., 2010. Mathematical optimization in intensity modulated radiation therapy. Annals of Operations Research 175 (1), 309–365.

Ehrgott, M., Holder, A., Reese, J., 2008. Beam selection in radiotherapy design. Linear Algebra and its Applications 428 (5), 1272–1312.

Ernst, A. T., Mak, V. H., Mason, L. R., 2009. An exact method for the minimum cardinality problem in the treatment planning of intensity-modulated radiotherapy. INFORMS Journal on Computing 21 (4), 562–574.

Fischetti, M., Salvagnin, D., Zanette, A., 2010. A note on the selection of Benders' cuts. Mathematical Programming 124 (1-2), 175–182.

Gleeson, J., Ryan, J., 1990. Identifying minimally infeasible subsystems of inequalities. ORSA Journal on Computing 2 (1), 61–63.

Gören, M., Taşkın, Z. C., 2015. A column generation approach for evaluating delivery efficiencies of collimator technologies in IMRT treatment planning. Physics in medicine and biology 60 (5), 1989.

Gozbasi, H. O., 2010. Optimization approaches for planning external beam radiotherapy. Ph.D. thesis, Georgia Institute of Technology.

Gurobi, O., 2016. Gurobi optimizer reference manual version 6.5. Accessed: 2018-01-23.
URL `https://www.gurobi.com/documentation/6.5/refman/index.html`

Guta, B., 2003. Subgradient optimization methods in integer programming with an application to a radiation therapy problem. Ph.D. thesis, Teknishe Universitat Kaiserlautern, Kaiserlautern.

Hall, E. J., Wuu, C.-S., 2003. Radiation-induced second cancers: the impact of 3D-CRT and IMRT. International Journal of Radiation Oncology* Biology* Physics 56 (1), 83–88.

Lee, E. K., Fox, T., Crocker, I., 2003. Integer programming applied to intensity-modulated radiation therapy treatment planning. Annals of Operations Research 119 (1-4), 165–181.

Lin, S., 2014. Benders decomposition and an IP-based heuristic for selecting IMRT treatment beam angles. Master's thesis, The University of Texas at Austin.

Luan, S., Wang, C., Cao, D., Chen, D. Z., Shepard, D. M., Cedric, X. Y., 2008. Leaf-sequencing for intensity-modulated arc therapy using graph algorithms. Medical physics 35 (1), 61–69.

Magnanti, T. L., Wong, R. T., 1981. Accelerating Benders decomposition: Algorithmic enhancement and model selection criteria. Operations Research 29 (3), 464–484.

Mahnam, M., Gendreau, M., Lahrichi, N., Rousseau, L.-M., 2017. Simultaneous delivery time and aperture shape optimization for the volumetric-modulated arc therapy (VMAT) treatment planning problem. Physics in Medicine & Biology 62 (14), 5589–5611.

Mason, L. R., Mak-Hau, V. H., Ernst, A. T., 2012. An exact method for minimizing the total treatment time in intensity-modulated radiotherapy. Journal of the Operational Research Society 63 (10), 1447–1456.

Mason, L. R., Mak-Hau, V. H., Ernst, A. T., 2015. A parallel optimisation approach for the realisation problem in intensity modulated radiotherapy treatment planning. Computational optimization and applications 60 (2), 441–477.

McCormick, G. P., 1976. Computability of global solutions to factorable nonconvex programs: Part I convex underestimating problems. Mathematical programming 10 (1), 147–175.

Men, C., Romeijn, H. E., Jia, X., Jiang, S. B., 2010. Ultrafast treatment plan optimization for volumetric modulated arc therapy (VMAT). Medical Physics 37 (11), 5787–5791.

Men, C., Romeijn, H. E., Taşkın, Z. C., Dempsey, J. F., 2007. An exact approach to direct aperture optimization in IMRT treatment planning. Physics in medicine and biology 52 (24), 7333–7352.

Oelkfe U., S. C., 2006. Dose calculation algorithms. In: Schlegel W., Bortfeld T., G. A. (Ed.), New Technologies in Radiation Oncology. Springer, pp. 187–196.

Otto, K., 2008. Volumetric modulated arc therapy: IMRT in a single gantry arc. Medical physics 35 (1), 310–317.

Palma, D., Vollans, E., James, K., Nakano, S., Moiseenko, V., Shaffer, R., McKenzie, M., Morris, J., Otto, K., 2008. Volumetric modulated arc therapy for delivery of prostate radiotherapy: comparison with intensity-modulated radiotherapy and three-dimensional conformal radiotherapy. International Journal of Radiation Oncology* Biology* Physics 72 (4), 996–1001.

Papp, D., Unkelbach, J., 2014. Direct leaf trajectory optimization for volumetric modulated arc therapy planning with sliding window delivery. Medical physics 41 (1), 011701.

Peng, F., Jia, X., Gu, X., Epelman, M. A., Romeijn, H. E., Jiang, S. B., 2012. A new column-generation-based algorithm for VMAT treatment plan optimization. Physics in medicine and biology 57 (14), 4569–4588.

Pocket Dentistry, 2015. Cone beam computed tomography (CBCT). Accessed: 2018-01-23.
URL https://pocketdentistry.com/13-cone-beam-computed-tomography-cbct/

Preciado-Walters, F., Langer, M. P., Rardin, R. L., Thai, V., 2006. Column generation for IMRT cancer therapy optimization with implementable segments. Annals of Operations Research 148 (1), 65–79.

Python, 2015. Python 2.7.11 documentation. Accessed: 2018-01-23.
URL https://docs.python.org/release/2.7.11/

Rahmaniani, R., Crainic, T. G., Gendreau, M., Rei, W., 2017. The Benders decomposition algorithm: A literature review. European Journal of Operational Research 259 (3), 801–817.

Romeijn, H. E., Ahuja, R. K., Dempsey, J. F., Kumar, A., 2005. A column generation approach to radiation therapy treatment planning using aperture modulation. SIAM Journal on Optimization 15 (3), 838–862.

Romeijn, H. E., Ahuja, R. K., Dempsey, J. F., Kumar, A., 2006. A new linear programming approach to radiation therapy treatment planning problems. Operations Research 54 (2), 201–216.

Salari, E., Unkelbach, J., 2013. A column-generation-based method for multi-criteria direct aperture optimization. Physics in medicine and biology 58 (3), 621–639.

Salari, E., Wala, J., Craft, D., 2012. Exploring trade-offs between VMAT dose quality and delivery efficiency using a network optimization approach. Physics in medicine and biology 57 (17), 5587–5600.

Sarykalin, S., Serraino, G., Uryasev, S., 2008. Value-at-risk vs. conditional value-at-risk in risk management and optimization. In: State-of-the-Art Decision-Making Tools in the Information-Intensive Age. Informs, pp. 270–294.

Shepard, D., Cao, D., Afghan, M., Earl, M., 2007. An arc-sequencing algorithm for intensity modulated arc therapy. Medical physics 34 (2), 464–470.

SIMBALLC, 2013. IMRT-What is Intensity-Modulated Radiation Therapy. Accessed: 2018-01-23.
URL http://www.simballc.org/imrt.html

Song, J., Shi, Z., Sun, B., Shi, L., 2015. Treatment planning for volumetric-modulated arc therapy: Model and heuristic algorithms. IEEE Transactions on Automation Science and Engineering 12 (1), 116–126.

Taşkın, Z. C., 2010. Benders decomposition. In: Cochran, J. J. (Ed.), Encyclopedia of Operations Research and Management Science. Wiley.

Taşkın, Z. C., Smith, J. C., Romeijn, H. E., Dempsey, J. F., 2010. Optimal multileaf collimator leaf sequencing in IMRT treatment planning. Operations Research 58 (3), 674–690.

Taşkın, Z. C., Cevik, M., 2013. Combinatorial Benders cuts for decomposing IMRT fluence maps using rectangular apertures. Computers & Operations Research 40 (9), 2178–2186.

Teoh, M., Clark, C. H., Wood, K., Whitaker, S., Nisbet, A., 2011. Volumetric modulated arc therapy: a review of current literature and clinical use in practice. The British Journal of Radiology 84, 967–996.

Üster, H., Agrahari, H., 2011. A Benders decomposition approach for a distribution network design problem with consolidation and capacity considerations. Operations Research Letters 39 (2), 138–143.

Van Roy, T. J., 1986. A cross decomposition algorithm for capacitated facility location. Operations Research 34 (1), 145–163.

Varian, 2017a. Accessed: 2018-01-23.
URL https://www.varian.com/fi/about-varian/newsroom/image-gallery/inside-varian-linear-accelerator

Varian, 2017b. Accessed: 2018-01-23.
URL http://newsroom.varian.com/imagegallery?cat=2473

VCU Massey Cancer Center, 2017. VCU Massey Cancer Center introduces safer, more effective form of radiation therapy. Accessed: 2018-01-23.
URL https://www.massey.vcu.edu/about/blog/2011/massey_cancer_center_introduces_radiation_therapy/

Wala, J., Salari, E., Chen, W., Craft, D., 2012. Optimal partial-arcs in VMAT treatment planning. Physics in medicine and biology 57 (18), 5861–5874.

Wang, C., Luan, S., Tang, G., Chen, D. Z., Earl, M. A., Cedric, X. Y., 2008. Arc-modulated radiation therapy (AMRT): a single-arc form of intensity-modulated arc therapy. Physics in medicine and biology 53 (22), 6291–6303.

Yu, C. X., 1995. Intensity-modulated arc therapy with dynamic multileaf collimation: an alternative to tomotherapy. Physics in medicine and biology 40 (9), 1435–1449.

Zhang, P., Happersett, L., Yang, Y., Yamada, Y., Mageras, G., Hunt, M., 2010. Optimization of collimator trajectory in volumetric modulated arc therapy: development and evaluation for paraspinal SBRT. International Journal of Radiation Oncology* Biology* Physics 77 (2), 591–599.