THE DETERMINATION OF TREATMENT PLANS FOR VOLUMETRIC MODULATED ARC THERAPY

by

Pınar Dursun

B.S., Mathematical Engineering, İstanbul Technical University, 2006

M.S., Industrial Engineering, İstanbul Technical University, 2009

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Industrial Engineering

Boğaziçi University

2019

ACKNOWLEDGEMENTS

I know the words are not enough, still, I will try. İ. Kuban Altınel, my supervisor, is my most important guide in my long and difficult struggle to become a virtuous person and a good scientist. Even though my doctoral studies are over, I know that I have much more to learn from him. I am deeply grateful to him for supporting me in all matters, anytime and anywhere. It is a great honor for me to be one of his Ph.D. students.

I would like to express my sincere gratitude to Z. Caner Taşkın for being an inspire to me. Besides being a great supervisor, he always supported me like a friend. I am also grateful to him for answering my endless questions with his endless knowledge and patience.

I would like to thank Ethem Nezih Oral, Necati Aras, Serpil Sayın and Evrim Didem Güneş for taking part in my dissertation committee, and for their constructive comments and helpful suggestions. I am also very grateful to Hatice Bilge, Murat Okutan, Nazmiye Dönmez Kesen, Leyla Süncak, Canan Köksal and Uğur Akbaş from Istanbul University Institute of Oncology for their helpfulness and hospitality.

A very special thank my kind-hearted friends Gizem Aktaş, İpek Dursun, Oylum Şeker and Nefel Tellioğlu. Thank a lot for everything. I would like to thank Mustafa G. Baydoğan and Berk Orbay for their encouraging support and guidance. Ayşe Orbay Kaya, Deniz Orbay Kaya and Tolga Kaya, thank you for your precious friendship. And, Zafer Eren, thank you for being my sister. I am deeply grateful to Ümit Şenesen for his endless interest and support.

I would like to thank Umut Gündüz for being my spiritual supporter and comrade in this difficult and long journey.

Finally, I acknowledge that this dissertation was supported by Boğaziçi University Research Fund with grant number 11520-16A03D1 and by the Turkish Directorate of Strategy and Budget under the TAM Project number DPT2007K120610, which made it easier to undertake this research and gave me the opportunity to attend scientific conferences.

ABSTRACT

THE DETERMINATION OF TREATMENT PLANS FOR VOLUMETRIC MODULATED ARC THERAPY

Volumetric modulated arc therapy is the state-of-the-art technique for external radiation therapy treatment, where radiation can be delivered continuously during the rotation of the linear accelerator's gantry. This property makes this technique powerful in obtaining high conformal plans requiring short treatment times. However, the multileaf collimator system shapes the radiation beam continuously, thus the resulting apertures are interdependent due to leaf motion limitations, which makes treatment planning hard. In this thesis, we first propose two mixed integer linear programming formulations minimizing total radiation delivered to the patient subject to the geometrical and clinical requirements. Then, we develop exact solution algorithms that combine Benders decomposition with certain acceleration strategies and implement branch-and-price method where pricing subproblem is decomposable by rows of multileaf collimator and can be solved as a shortest path problem. We investigate their performance on a large set of test instances obtained from an anonymous real prostate cancer data. The computational results reveal that they are efficient and outperform a widely used commercial solver. In particular, branch-and-price implementation is capable to find optimal solutions for larger problem instances. However, they cannot provide realistic plans for real clinical problems because of their large size. In order to address this issue, we develop a two-phase column generation based heuristic that tunes the parameters of dose-volume requirements and yields an automated treatment planning environment, which does not require any human intervention. We test its performance on real prostate data sets and compare the quality of the generated plans with those obtained by a widely used commercial treatment planning system. Results show that it can obtain medically acceptable plans requiring significantly less radiation in reasonable computation times.

ÖZET

HACİMSEL YOĞUNLUK AYARLI ARK SAĞALTIMI PLANLARININ BELİRLENMESİ

Hacimsel yoğunluk ayarlı ark sağaltımı, doğrusal hızlandırıcı kızağı hasta etrafında dönerken ışının kesintisiz olarak gönderilebildiği dışsal radyasyon terapisinde yakın zamanda geliştirilen bir tekniktir. Bu özellik bu tekniği kısa sağaltım sürelerine gereksinim duyan yüksek uygunlukta planların elde edilmesinde güçlü kılmaktadır. Fakat, çok yapraklı yönlendirici radyasyon ışınını kesintisiz olarak biçimlendirir, bu nedenle elde edilen açıklıklar yaprak hareket kısıtlamaları nedeniyle birbirine bağımlıdır ve sağaltım planlaması zorlaşır. Bu tezde, ilk olarak geometrik kısıtlamaları ve sağaltıma ilişkin gereksinimleri sağlayarak hastaya iletilen toplam radyasyon miktarını en aza indiren iki karışık-tamsayılı doğrusal programlama gösterimi önerilmiştir. Daha sonra Benders ayrıştırma yönteminin belirli hızlandırma yaklaşımlarıyla birleştirildiği ve ederlendirme probleminin ayrıştırılarak en kısa yol problemi olarak çözüldüğü dal-eder algoritmaları geliştirilmiştir. Bu algoritmaların başarımları anonim bir prostat verisinden türetilmiş çok sayıda örnek üzerinde değerlendirildi. Bilgisayısal deneylerin sonuçları yaygın olarak kullanılan ticari bir eniyileme çözücüsünden daha iyi sonuçlar veren etkin algoritmalar olduklarını ortaya koymaktadır. Özellikle, dal-eder uygulaması daha büyük boyutlu problemler için eniyi çözümler elde edebilmektedir. Yine de klinik boyuttaki problemler için kabul edilebilir sağaltım planları elde etmek olanaklı değildir. Bu nedenle, doz-hacim gereksinimlerine ilişkin parametreleri ayarlayabilen ve karışma gerektirmeyen bir otomatik sağaltım ortamı sunan iki aşamalı sütun türetme temelli sezgisel bir algoritma geliştirilmiştir. Gerçek prostat verileri kullanılarak bu algoritma ile elde edilen planların kalitesi yaygın olarak kullanılan bir ticari sağaltım planlama dizgesince elde edilenlerle karşılaştırılmıştır. Karşılaştırma sonuçları sezgisel kullanıldığında daha az radyasyona gereksinim duyan, klinik olarak kabul edilebilir planlar elde edilebildiğini göstermektedir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

a_{ijk}	nonnegative continuous variable, radiation dose intensity of the jth beamlet of row i at control point k
b_i^e	binary variable, it is set to 1 if the feasible row arc z_i^e is selected
C_o	total number of partial volume constraints of OAR o
C_t	total number of partial volume constraints of TV t
$\mathbf{D}$	dose-influence matrix
$\overline{d}_{tc}$	the cth prescribed dose for TV t
d_v	nonnegative continuous variable, radiation dose absorbed by voxel v
g_{ijk}	nonnegative dual multiplier used in LSP
$\mathscr{I}$	index set of the beamlets having strictly positive effect on at least one voxel
K	total number of control points ($k = 1, ..., K$)
$\overline{K}$	subset of K
$\overline{\overline{K}}$	subset of $\overline{K}$
l_{ik}	nonnegative integer variable used to represent the position of the left leaf on row i at control point k
l_{ijk}	binary variable used to represent the position of the left leaf on row i at control point k
L^{mu}	lower bound on radiation dose intensity at a control point
L_t^{TV}	lower bound on the amount of radiation dose absorbed by a target voxel in TV t
m	total number of rows of an aperture ($i = 1, ..., m$)
mu_k	nonnegative continuous variable, radiation dose intensity at control point k
n	total number of columns of an aperture ($j = 1, ..., n$)
O	total number of OARs ($o = 1, ..., O$)
r_{ik}	nonnegative integer variable used to represent the position of the right leaf on row i at control point k

r_{ijk}	binary variable used to represent the position of the right leaf on row i at control point k
T	total number of TVs ($t = 1, ..., T$)
u_{ijk}	nonnegative dual multiplier used in LSP
U^{mu}	upper bound on radiation dose intensity at a control point
U_t^{TV}	upper bound on the amount of radiation dose absorbed by target a voxel in TV t
V	set of all voxels
V_o^{OAR}	set of voxels in OAR o
V^{OAR}	set of all voxels in all OARs
V_t^{TV}	set of voxels in TV t
V^{TV}	set of all voxels in all TVs
x_{tcv}	nonnegative continuous variable, the surplus of the value ξ_{tc}^{TV} by the dose received by voxel v in TV t
y_{ocv}	nonnegative continuous variable, the surplus of the value ξ_{oc}^{OAR} by the dose received by voxel v in OAR o
z_{ijk}	binary variable, 1 if the jth beamlet of row i at control point k is open, 0 otherwise (j=1,...,n)
Z_i	set of all feasible treatment row arcs for row i
$\mathscr{Z}$	index set of all beamlets at all control points
$\mathscr{Z}_0$	index set of closed beamlets
$\mathscr{Z}_1$	index set of open beamlets
$\mathscr{Z}^*$	index set of the beamlets that are associated with an MIS
α_{oc}^{OAR}	minimum ratio of voxels in OAR o that receive radiation at most the tolerance dose U_{oc}^{OAR}
α_{tc}^{TV}	minimum ratio of voxels in TV t that receive radiation at least the prescribed dose $\overline{d}_{tc}$
γ_{oc}^{OAR}	penalty cost for deviation in the cth partial volume constraints OAR o
γ_{tc}^{TV}	penalty cost for deviation in the cth partial volume constraints of TVs t

δ	maximum allowable distance (in beamlet) a leaf can move between two consecutive control points
Δ	set of extreme points of DSP
ϵ	a small number
η	represents the total radiation intensity
ξ_{tc}^{TV}	continuous variable, radiation dose absorbed by the ((1-α_{tc}^{TV})$\lvert V_t^{TV}\rvert$)th voxel in TV t receiving the lowest radiation
ξ_{oc}^{OAR}	continuous variable, radiation dose absorbed by the ((1-α_{oc}^{OAR})$\lvert V_o^{OAR}\rvert$)th voxel in OAR volume o receiving the highest radiation
Υ_k	fractionality of an aperture at control point k
ϕ_{tc}^{TV}	nonnegative continuous variable, deviation in the cth partial volume constraints of TVs t
ϕ_{oc}^{OAR}	nonnegative continuous variable, deviation in the cth partial volume constraints OAR o
Ω	set of extreme directions of DSP

LIST OF ACRONYMS/ABBREVIATIONS

3D-CRT	Three-Dimensional Conformal Radiation Therapy
AAA	Analytical Anisotropic Algorithm
AP	Alternative Problem
BAO	Beam Angle Optimization
BEV	Beam's Eye View
BIP	Binary Integer Programming
BP	Branch-and-price
CPU	Central Processing Unit
CT	Computed Tomography
CTV	Clinical Target Volume
CVaR	Conditional Value-at-Risk
DAO	Direct Aperture Optimization
DICOM	Digital Imaging and Communications in Medicine
DMLP	Dual Master Linear Problem
DNA	Deoxyribonucleic Acid
DPFSP	Dual Pure Feasibility Subproblem
DSP	Dual Subproblem
DVH	Dose-Volume Histogram
FMO	Fluence Map Optimization
GTV	Gross Tumor Volume
Gy	Gray
IGRT	Image Guided Radiation Therapy
IMAT	Intensity Modulated Arc Therapy
IMRT	Intensity Modulated Radiation Therapy
LB	Lower Bound
LD	Lagrangean Dual
LP	Linear Program
LPVMATP	Linear Programming Relaxation of VMATP

LSP	Lagrangean Subproblem
MILP	Mixed Integer Linear Programming
MIS	Minimal Infeasible System
MLC	Multileaf Collimator
MLP	Master Linear Problem
MLS	Multileaf Collimator Leaf Sequencing
MP	Master Problem
MU	Monitor Unit
MV	Megavolt
M-VMATP	Modified VMATP
OAR	Organ At Risk
PB	Penile Bulb
PBA	Pencil Beam Algorithm
PFSP	Pure Feasibility Subproblem
PSP	Pricing Subproblem
PTV	Planning Target Volume
RDSP	Reduced Dual Subproblem
RMLP	Restricted Master Linear Problem
RMP	Relaxed Master Problem
R-OAR	Rest of Organ At Risk
RVMATP	Relaxation of VMATP
SAD	Source-to-axis Distance
SP	Subproblem
SSD	Source-to-surface Distance
TPS	Treatment Planning System
TV	Target Volume
UB	Upper Bound
VaR	Value-at-Risk
VMAT	Volumetric Modulated Arc Therapy
VMATP	Volumetric Modulated Arc Therapy Planning

1. INTRODUCTION

Oncology is the medical practice dealing with cancerous tumors, including their origin, development, diagnosis, treatment, and prevention. The most common types of cancer treatments are surgery, chemotherapy, and radiation therapy. Surgery excises the tumor from the body if cancer has not metastasized or only small parts of the body are cancerous. Chemotherapy is the drug treatment where anti-cancer drugs are used to kill cancer cells. In radiation therapy high energy radiation is used to treat cancer. Each of these treatment methods may be applied alone or in combination.

Radiation therapy, or 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. However, if healthy cells are given high amount of radiation they may not repair themselves and other medical problems, such as organ destruction, may occur. 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, radiation beams are generated outside the patient and delivered to the tumor; on the other hand radiation sources 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. A *linear accelerator* (see Figure 1.1) is the most commonly used medical device, where the patient lies on a moveable treatment couch. The gantry of the linear accelerator can rotate around the patient and delivers high-energy beams from different angles by keeping the cancer volume on the target.

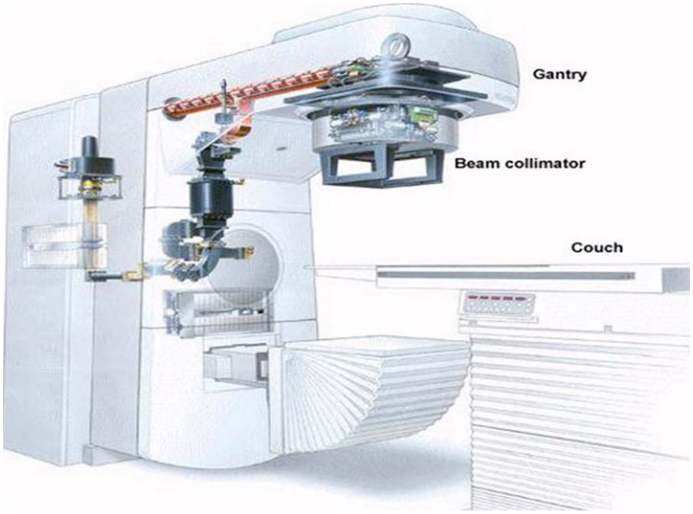

Figure 1.1. A linear accelerator [1].

IMRT and VMAT are two commonly used external-beam radiation therapy techniques. In both of them, the gantry of the linear accelerator is equipped with a *multileaf collimator* (MLC) system, 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 1.2). In this way, the conformity of dose distribution to the planning target volume (PTV), which is tumor plus some margin, and normal tissue sparing is much superior compared to earlier techniques [7]. However, IMRT and VMAT requires higher amount of radiation (in monitor units, MUs) to deliver a given fraction size compared with 3D-CRT [8,9]. The

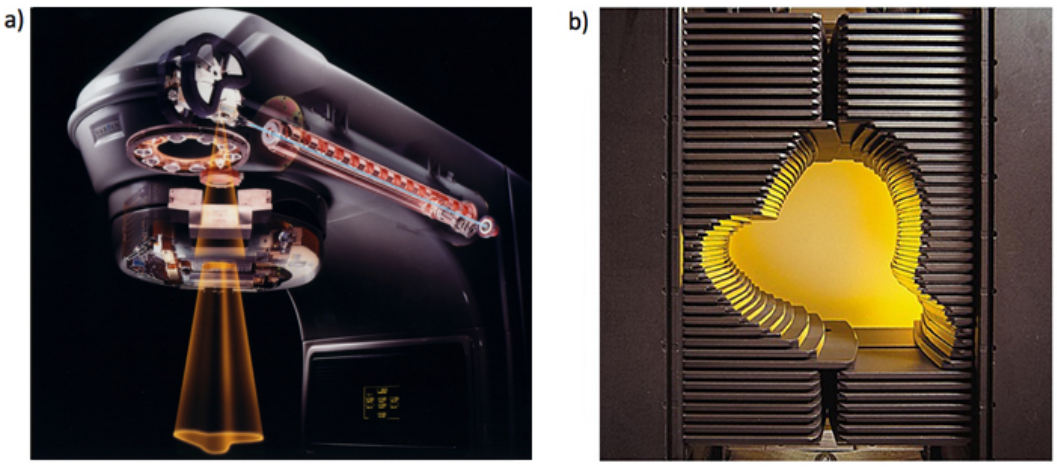

Figure 1.2. (a) A shaped beam [2], (b) A multileaf collimator system [3].

increase in MUs increases the risk of secondary radiation-induced malignancies [10]. Although IMRT has been used very extensively in radiation therapy since 1990s [11], VMAT is the state-of-the-art technology. In VMAT, the gantry of the linear accelerator rotates around the body along one or more arcs and delivers radiation continuously. The leaves of MLC system move and shape the beam, and also, dose rate and gantry speed can change simultaneously during the rotation of the gantry. These features of VMAT enable it to produce radiation therapy plans having high conformal dose distributions and requiring less radiation compared to IMRT [7, 8]. Also, radiation delivery times of the resulting plans become significantly shorter [12, 13]. On the other hand, there are typically only a few discrete angles (5-9) in IMRT plans [14] (see Figure 1.3). Furthermore, the linear accelerator stops delivering radiation while moving its gantry between different *beam angles* (or *control points*) 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 [11]. 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 [13].

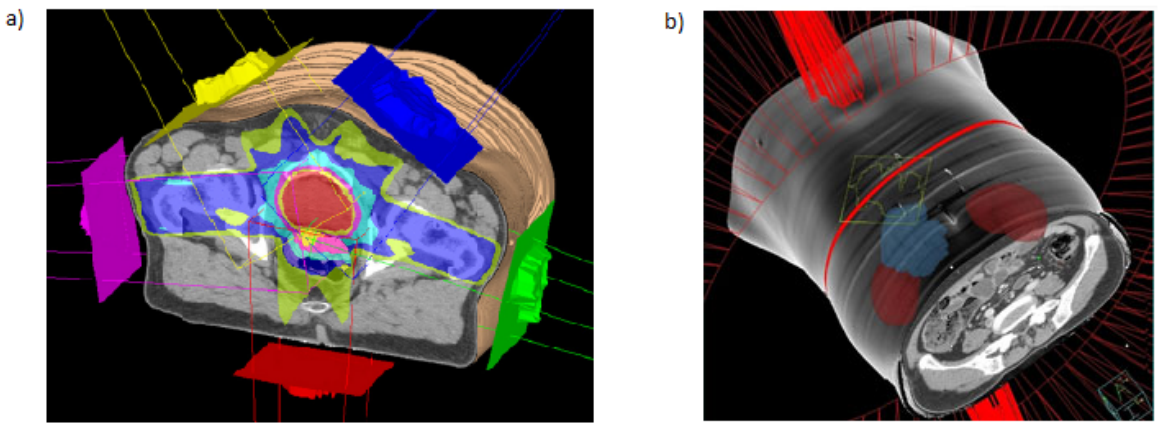

Figure 1.3. (a) IMRT [4], (b) VMAT [5].

The main advantage of VMAT is the ability to deliver radiation continuously, however, this causes a very thin slicing of the gantry's rotational arc at considerably

many control points in VMAT planning. As a consequence, adjacent control points become very close and that 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 radiation dose related dependencies introduce. These unique characteristic of the technique makes VMAT planning a challenging issue in radiation therapy compared to IMRT planning. As a result, most of the models in the literature are not comprehensive enough to take all aspects of VMAT treatment into account due to the increasing computational difficulty. 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 in order to obtain solvable relaxed formulations.

In order to close this gap, we develop two new mixed integer linear programming (MILP) formulations for VMAT planning that include all radiation dose requirements as hard constraints as well as mechanical limitations of the linear accelerator and MLC system. The models proposed so far, generally minimize dose deviations from the prescribed limits and represent two different treatment plans with similar dose deviations but different total MUs as equivalent; they just do not distinguish between the plans with respect to their radiation requirements. Also, 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 [15]. 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. To this end, we focus on finding the VMAT plans that are not only feasible with respect to the clinical prescriptions, but also require less radiation, by formulating the objective function of our MILP models to minimize the total radiation amount (in MUs) delivered to the patient.

Moreover, there is another gap in the literature of VMAT planning. The solution approaches proposed so far are heuristic algorithms, since the models underlying VMAT planning problem are large and hard to solve. To the best of our knowledge there is

not an exact solution algorithm for the determination of optimal VMAT plans. We believe it is important to focus on developing exact optimization algorithms and make progress in this research direction in order to reveal the potential of this technique better. In this dissertation, we develop two different exact solution approaches that we believe they will be pioneering ones in this field. The VMAT planning problem has a natural tendency to decompose into two interacting parts. One of them deals with the geometry of the equipment while the other determines the right amount of radiation dose of the treatment region. Based on this observation we propose an exact solution algorithm using Benders decomposition (in Chapter 5). The idea is to keep binary variables in the master problem and solve a linear programming subproblem to generate cuts. We improve the naive implementation of Benders decomposition by applying certain acceleration strategies. We test their performances on a large set of test instances derived from a real prostate cancer data set provided by Craft *et al.* [16, 17], and compare the computational results with those obtained by using a MILP solver. As given in Chapter 8 the improved Benders algorithms outperforms the MILP solver especially for large instances.

Afterward, we observe that reversing the order of decomposition and considering a subproblem including the binary variables may have been more advantageous since the problem itself can be decomposed into shortest path subproblems. In fact this gives birth to branch-and-price (BP) algorithms explained in Chapter 6. As can be observed from the computational results in Chapter 8 one of them performs significantly better than the best Benders decomposition algorithm, and can compute optimal treatment plans minimizing total radiation for considerably larger instances.

To the best of our knowledge, our exact solution algorithms are the first attempts to solve exactly a VMAT planning model in which all VMAT's treatment related constraints are forced to be satisfied. However, they are not capable of solving clinical problems including all structures. Finally, in Chapter 7 we develop a two-phase heuristic, which is based on column generation formulations developed in Chapter 6. We test the performance of the heuristic on real cancer patients data sets provided by Istanbul University Oncology Institute, which is one of the largest and oldest cancer centers in

Turkey, and make clinical comparisons with the plans obtained by the institute's staff. The computational results show that the new heuristic is capable of finding clinically acceptable plans with less MUs and does not need any intervention such as modifying the parameters of the plan and re-optimizing, which is the common practice in the radiation therapy planning departments.

The rest of this dissertation is organized as follows. In the next chapter, we describe some basic concepts that arise in external-beam radiation therapy planning in order to facilitate the follow-up of the forthcoming chapters. We provide the related literature review concentrating on the optimization methods for IMRT and VMAT planning in Chapter 3. In Chapter 4, we define the VMAT planning problem and present our mathematical formulations. We continue by explaining the exact solution algorithms in Chapter 5 and Chapter 6, and two-phase heuristic in Chapter 7. Chapter 8 presents the computational results for the Benders decomposition and BP algorithms and compare them with a MILP solver's. Also, we make clinical comparison of the plans obtained by our heuristic algorithm with the actual ones obtained in Istanbul University Oncology Institute. Finally, we give a brief summary to conclude the dissertation and point out the potential future research direction, in Chapter 9.

2. BASIC CONCEPTS

External-beam radiation therapy process starts with the determination of tumors and surrounding normal structures after the diagnosis of the patient with cancer. Then a treatment that satisfies radiation dose prescriptions as well as mechanical limitations of the linear accelerator and MLC system is planned by a medical physicist and/or an experienced dosimetrist. The plan is delivered in a specific number of identical sessions, which is called *fractionation* and the number of fractions mainly depends on the tumor type.

Treatment plans had been prepared manually until more sophisticated techniques were developed in parallel with technological advances. There are two main categories of planning approaches: forward and inverse planning. Forward planning is a trial and error approach where the parameters such as beam angles, MLC segments and radiation intensities are fixed and the dose distribution of the resulting plan is calculated. If treatment prescriptions are not satisfied, then the parameters are updated until a reasonable plan is obtained. Simulation is one of the methods used for forward planning. Nevertheless, this approach is inadequate to reflect the capabilities of the new advanced technologies. However, inverse planning is an automated planning approach that provides plans with better dose distributions and shorter treatment times as compared to forward planning. It requires optimization tools, hence, operations researchers and mathematical programmers are interested in the radiation treatment planning [11]. In Figure 2.1 the difference between these two planning approaches is illustrated [18]. We consider inverse planning approach in this dissertation.

From the point of operations research, a radiation therapy treatment plan actually answers the following questions: where to deliver radiation?; how to deliver radiation?; and how much radiation to deliver?. To answer these questions the radiation treatment planning problem should be formulated according to the solution technique that will be applied. In IMRT planning, there are a few number of beam angles, or control points, that irradiation occurs and they are usually determined by the experienced dosimetrist

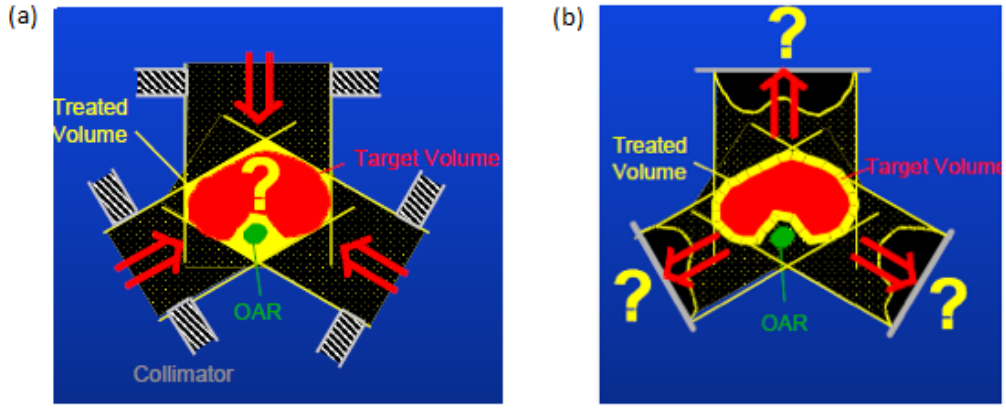

Figure 2.1. (a) Forward planning, (b) Inverse planning.

in advance. However, in VMAT planning, the continuous rotation of the gantry on a co-planar arc is generally discretized and it is assumed that radiation is delivered from a large number of equally spaced control points. (Since the couch of the linear accelerator can move, it is also possible to deliver radiation on a non-coplanar arc). Therefore, the answer to the first question is the location of the control points where the radiation delivery occurs.

The answers of the last two questions in IMRT planning are related to finding a *fluence map* at each one of the control points and to realizing them into a number of deliverable radiation beams. A fluence map is represented by a two-dimensional nonnegative matrix that gives the radiation intensity profile (see the left-most matrix on Figure 2.2). An opening where the radiation beam passes through is formed by the leaves of MLC at a control point and called as an *aperture*. A two-dimensional binary matrix is commonly used to represent an aperture, namely the opening is discretized into a number of *beamlets*. The number of rows of this matrix equals to the number of parallel leaf pairs on the MLC system. If a beamlet belongs to the open area, namely if it is exposed, than it takes value 1 and if it is blocked by the leaves of the MLC, then it is 0. In Figure 2.3 an aperture and its binary matrix representation are illustrated for an MLC system that has five leaf pairs (rows), and the leaf openings are decomposed into five columns. Due to the mechanical limitations of the linear accelerators, it is

only possible to deliver the radiation profile described by a fluence map using a number of apertures. In other words, in IMRT planning, a fluence map at a control point is realized by a weighted sum of a number of apertures (see Figure 2.2). The weight of an aperture represents the amount of radiation dose (in MU) that is delivered through this aperture. On the other hand, in VMAT planning, it is assumed that there is only one aperture at each control point. Therefore, to answer the second question, it is necessary to determine the shape of the apertures at all control points.

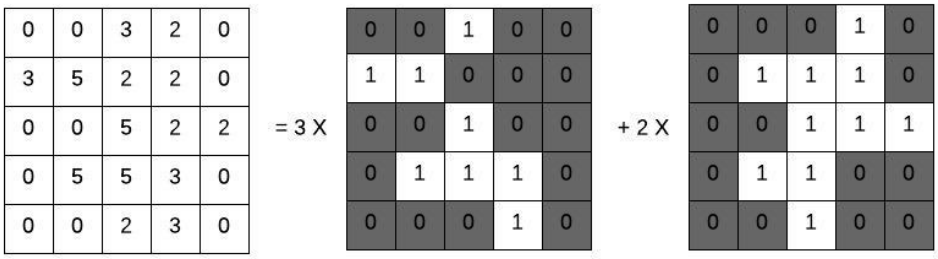

Figure 2.2. Decomposition of a fluence map into two apertures.

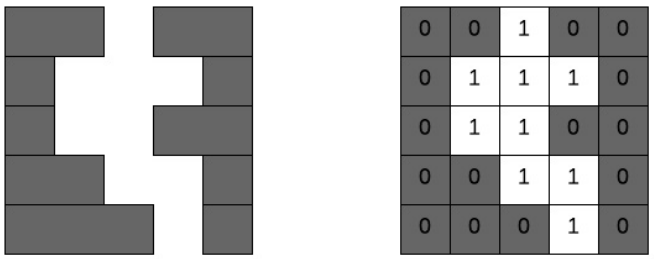

Figure 2.3. An aperture and its binary matrix representation.

There are some mechanical limitations of MLC systems that should be incorporated into the planning. For example, the leaves have to satisfy some properties depending on the type of the system. The most common one, almost all MLC systems must satisfy, is called the *consecutive ones property*. There can be at most one open beamlet chain in a row of an aperture, since MLC systems are made up of metal leaves. In Figure 2.4, the aperture on the left side satisfies the consecutive ones prop-

erty, however, the third row of the aperture on the right side violates the property since there are two open beamlet chains. Another property that some MLC systems must satisfy is *connectedness*, which requires that there is at most one open hole where the radiation passes through. The aperture on the left side on Figure 2.5 does not satisfy connectedness since there are two disjoint open holes. Finally, some MLC systems does

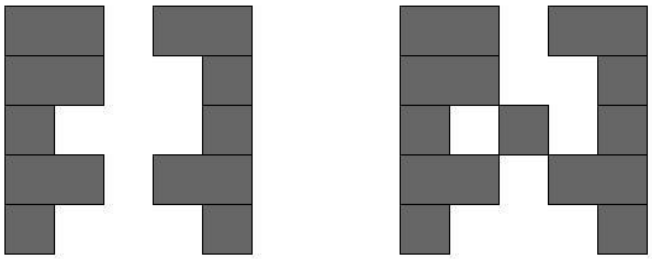

Figure 2.4. Consecutive ones property.

not allow the interdigitation of leaves, namely the left (or right) leaf at a row cannot touch the bottom or top right (or left) leaf. The third and fourth rows of the aperture on the right side on Figure 2.5 coincide; and that makes it infeasible according to MLC systems that does not allow interdigitation. We refer the reader to the study of Gören and Taşkın [19] for details of other possible properties of various MLC systems.

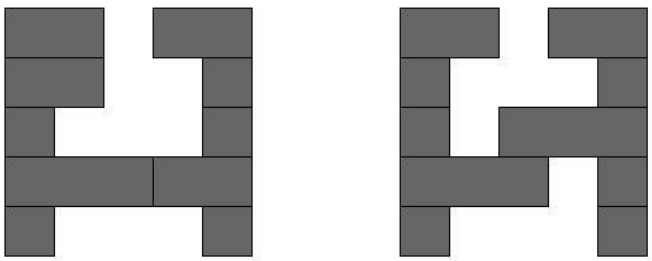

Figure 2.5. Connectedness and interdigitation.

In the dynamic version of IMRT the leaves of MLC system move and change the shape of the beam at a control point in order to obtain the desired fluence map. Similarly, in VMAT, during the rotation of the gantry the leaves also move and change the shape of the beam. However, there is a limitation on the speed of this movement.

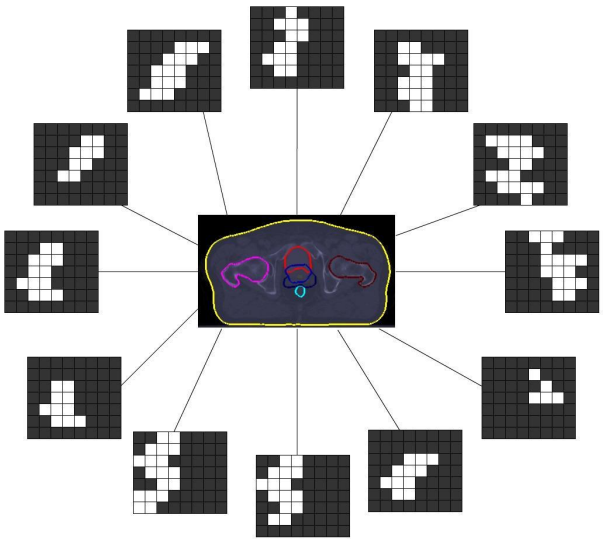

Figure 2.6. A sample VMAT treatment.

The maximum distance that a leaf can move per second depends on the technical characteristics of the linear accelerator. Similarly, the speed of the gantry and dose rate are limited from above and below, and there is a close relationship between them as well as MLC leaf movement. If the gantry rotates with high speed then the maximum radiation dose is less at a control point. Also, the apertures of neighboring control points are similar, since the leaves cannot move so much (or vice versa). Hence, at each control point the amount of MU has to be determined in order to answer the third question in VMAT planning. Figure 2.6 illustrates a simplified VMAT plan with few control points.

The geometric properties and mechanical limitations of the equipment used in the treatment are explained so far. The primary aim of the radiation therapy is to deliver enough radiation to tumor while protecting surrounding healthy tissues. For this purpose, the oncologist determines the location of the tumor and prescribes the radiation amounts that is delivered to the patient. There are three main volumes to be considered in the radiation therapy: gross tumor volume (GTV), clinical target volume (CTV), and planning target volume (PTV). The GTV is the primary tumor, which is visible and easily identifiable part of the malignant growth. The CTV contains the GTV and subclinical microscopic malignant lesions. The PTV surrounds the CTV and a

margin to account uncertainties in planning or delivery; it is considered in the treatment planning optimization [20] (see Figure 2.7). We use the terms PTV and target volume (TV) interchangeably in the rest of the dissertation. The oncologist contours the cancerous PTVs and surrounding organs at risk (OARs) on the computed tomography (CT) scans of the patient and prescribes the radiation doses best conforming to PTVs and OARs. There may be more than one PTV with different dose requirements as well as OAR depending on the cancer type and patient's anatomy.

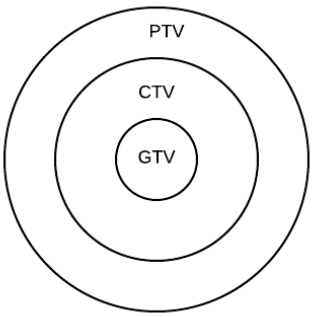

Figure 2.7. GTV, CTV, PTV.

The treatment prescriptions for a PTV require that the full or a partial volume of the PTV must absorb a predetermined amount of radiation. Also, there are tolerance dose limits for OARs. A specified partial volume of the organ must absorb below them. *Dose-volume histograms* (DVHs) are the most commonly used tools to evaluate the resulting dose distributions and the quality of a treatment plan. A DVH is a two-dimensional graph showing the fractional volume of a structure and the minimum dose absorbed by that volume. In Figure 2.8 DVHs illustrate the dose distributions of a PTV and OAR. For example, 60% of OAR absorbs at least 33 Gray (Gy) radiation. In other words, the maximum radiation dose that 40% of OAR absorbs is 33 Gy. Similarly, 100% of PTV absorbs around 71 Gy radiation.

In order to calculate dose distributions on the structures, the body of the patient is discretized into small cubes called *voxels* (see Figure 2.9 for an example) using CT scans. Moreover, dose calculation algorithms such as Pencil Beam Algorithm (PBA) [21] or Analytical Anisotropic Algorithm (AAA) [22] are used to calculate dose

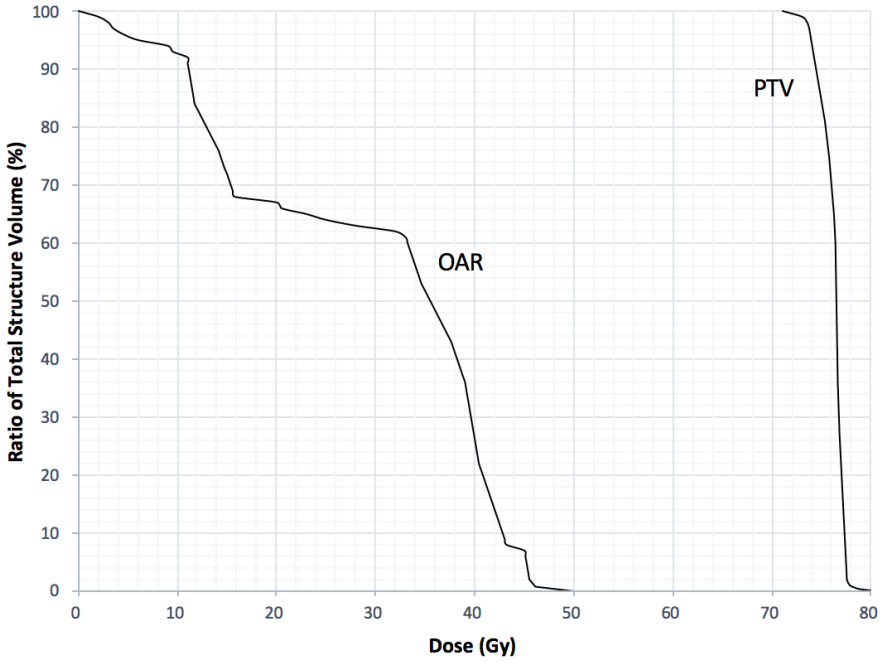

Figure 2.8. DVHs for a PTV and an OAR.

contribution of a beamlet to a voxel when it is delivered one unit of radiation from. Namely, they calculate the amount of absorbed radiation dose (Gy) per MU. These amounts are used as input in optimization models and called dose-influence matrices.

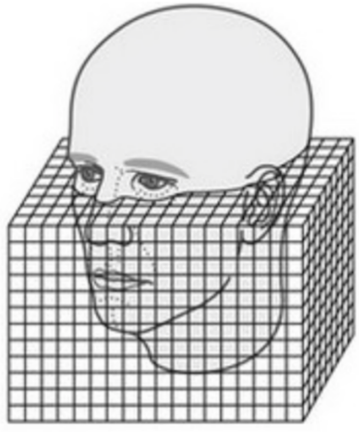

Figure 2.9. A voxel resolution of a head [6].

3. LITERATURE REVIEW

The studies in the literature developed for VMAT planning are mostly related to IMRT planning methods. Hence, in this chapter, we start by giving a brief explanation about IMRT planning phases referring to some representative publications, which makes easier to explain the algorithms developed for VMAT planning. Then, we continue by giving a detailed literature review on VMAT planning. Our aim is to develop a planning system which gives optimum treatment plans, thus we mostly take into account the studies trying to improve the treatment planning environment by operations research techniques.

3.1. Intensity Modulated Radiation Therapy Planning

There are three main phases in IMRT planning, which can be solved either sequentially or combining two of the phases. The first phase deals with the *beam angle optimization* (BAO) *problem* (or *geometry problem*): the number and orientation of beam angles (or control points) for irradiation are determined, which is mostly done by a medical physicist or dosimetrist in practice based on experience. There are also studies where a function is defined to determine the quality of a set of directions and this function is optimized in order to find the best set [23]. After determining the beam angles, a fluence map is obtained for each one of them in the second phase, which is called the *fluence map optimization* (FMO) *problem* (or *intensity problem*) [14, 24, 25]. As we mentioned in Chapter 2, a fluence map denotes the radiation intensity profile to be delivered through a given beam angle and can be represented by a two-dimensional nonnegative matrix. It is possible to formulate the FMO problem as a convex optimization problem; hence, it can be solved efficiently using one of the existing algorithms [26]. The third phase in IMRT planning is MLC *leaf sequencing* (MLS) *problem* (or *realization problem*), where a given fluence map is decomposed into a number of disjoint apertures and corresponding radiation intensities. In other words, a nonnegative matrix is re-expressed as a linear combination of binary matrices with positive weights. All of the binary matrices should satisfy the properties of the MLC system. During the de-

composition of a fluence map, total radiation delivery time (i.e. beam on time) and/or the total number of apertures (i.e. total machine setups) are minimized [7,27–33]. 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 [34].

Each one of these three phases can be handled independently and solved sequentially, however, there are also studies that consider two consecutive phases simultaneously. For instance, the first two phases, BAO and FMO, can be considered together and solved as a monolithic non-convex optimization problem to determine the beam angles and fluence maps simultaneously [35, 36]. There are also studies that directly optimize a number of apertures with intensities for each one of the determined beam angles. In other words, they solve the second and the third phases, FMO and MLS, simultaneously instead of finding a fluence map first and then decomposing it into a number of deliverable apertures [37–42]. This problem is called *direct aperture optimization* (DAO) problem. Column generation is one of the frequently used approaches where apertures are generated as new columns [38–42]. The general framework of the algorithms proposed in these studies is to start with an empty set of apertures and add apertures to the plan iteratively. The pricing subproblem yields the most promising feasible aperture in order to introduce it to the master problem, which determines optimum weights of the apertures generated and added to the treatment plan so far. In particular, Romeijn *et al.* [38] formulate a large-scale convex programming problem and solve their problem exactly, where they generate one or more promising apertures in each iteration by solving a network flow similar to the one in [30]. They modify the network model in order to make it possible to solve the DAO problem where the MLC system requires connected apertures. Men *et al.* [40] consider the MLC systems that allow only rectangular apertures, and solve the pricing subproblem by a polynomial time algorithm similar to the ones in [30, 38]. We refer the interested readers to the comprehensive survey of Ehrgott *et al.* [11] for more details on IMRT planning.

3.2. Volumetric Modulated Arc Therapy Planning

VMAT is not the first radiation treatment technique that benefits from the flexibility a rotating gantry introduces in order to obtain treatment plans with higher quality. Yu [43] proposed the rotational IMRT called Intensity Modulated Arc Therapy (IMAT) in 1995; but the clinical implementations remained very limited until Otto suggested VMAT in 2008. In VMAT, the gantry speed and the dose rate as well as the beam shape can vary during rotation. The linear accelerator can deliver radiation continuously to the patient's body, and thus in treatment planning it is commonly assumed that there is a large number of equally spaced control points in order to discretize this continuous radiation delivery. At each control point there is only one aperture; however, the apertures at two adjacent control points are interconnected. This is because there are limitations on the motion of the MLC leaves during rotation. Thus, the VMAT planning problem cannot be decomposed into a number of subproblems that can be solved independently. As a result, designing a VMAT plan is significantly harder compared to IMRT planning. Even when the total time to complete a tour is fixed, the resulting problem is a large-scale nonconvex optimization problem. These characteristics make VMAT planning a challenging task, which requires much more computational effort than IMRT planning [16].

Studies on VMAT planning can be classified into two groups. The members of the first group use a two-step approach that, in the first step, determines an optimal IMRT plan consisting of a number of fluence maps at evenly spaced control points. Then these fluence maps are converted into a deliverable VMAT plan using an arc-sequencing method in the second step. On the other hand, the studies in the second group directly optimize the leaf positions and radiation intensities of the apertures and are called DAO methods similar to the ones one can face in the IMRT planning literature. Our solution methods given in Chapter 5 – Chapter 7 fall into this group. We explain the studies in these two groups separately in the following subsections.

3.2.1. Two-step Approaches

Two-step approaches convert an idealized IMRT plan consisting of fluence maps at both coarse [44–46] and dense [16,47,48] sampling of control points into a deliverable VMAT plan. In the first step, a FMO problem is solved and intensity profiles are obtained. In the second step, an arc-sequencing method is used to convert the fluence maps into feasible apertures that satisfy MLC leaf limitations. Hence, these two-step approaches are also called *arc-sequencing methods*.

In one of the earliest work, Luan *et al.* [44] solve a shortest path problem to find k deliverable arcs from a number of continuous intensity patterns for equally spaced control points (typically with 10°-spacing). Each one of the fluence maps is decomposed into a number of apertures which realize the corresponding intensity map. Then, by selecting exactly one aperture from the generated ones at each control point, a deliverable arc is constituted. Namely, a treatment arc consists of relatively small number of control points. Finally, they obtain k different treatment arcs, since IMAT is not flexible as VMAT and the realization of the fluence intensity maps requires more than one arc, which causes long treatment times. The algorithm proposed by Wang *et al.* [45] solves a shortest path problem similar to the one in [44], however, they generate a single-arc plan by displacing the generated apertures onto the neighbor control points. It is assumed in both of these studies that the MLC system allows leaf interdigitation. A similar mechanism is used in the arc-sequencing algorithm of Cao *et al.* [46] to obtain a single-arc plan, where they reduce the number of apertures per control point to 2-6. They optimize the apertures directly in an IMRT plan using a direct machine parameter optimization method. A simulated annealing-based algorithm is used as an arc sequencer to obtain deliverable arcs.

Note that converting an IMRT plan that consists of a small number of control points may cause a deterioration in the quality of the dose distribution of the resulting VMAT plan. Since VMAT planning problem has its own constraints on the MLC's movement, which must be considered during arc-sequencing. Some of the studies first obtain an "ideal" IMRT plan including a large number of control points, and then

coarsen this plan to reduce delivery time by maintaining dose distribution quality. Craft *et al.* [16] propose an algorithm called VMERGE, that obtains a fine sample IMRT plan for 180 equally spaced beam angles in the first step by solving a convex multicriteria optimization problem. However, this ideal plan is obtained by disregarding treatment time. It is observed that the fluence maps of neighbor beam angles are similar, thus in the second step, the ideal plan is transformed into a deliverable VMAT plan by merging similar fluence maps iteratively as long as the dose distribution quality is maintained. The resulting maps are sequenced and delivered over the corresponding arc segment. In short, their algorithm starts with a finely sampled plan, and this plan is coarsened to reduce the delivery time. Then, Salari *et al.* [47] propose an improved form of VMERGE algorithm where a merging problem is formulated as a discrete bi-criteria optimization problem using a network flow model. In another extended version of VMERGE, optimal partial-arc plans are generated automatically [48]. They use the same iterative fluence map merging and sequencing algorithm given in [16] to find a plan for each partial-arc and select the best one with minimum treatment time. This new algorithm is called PMERGE, and computational experiments show that the treatment time of a plan obtained by PMERGE is lower than the ones obtained by VMERGE. However, there may be a large number of partial-arcs and this may increase the computation time.

3.2.2. Direct Aperture Optimization Methods

The studies of the second group optimize the beam shapes (i.e. aperture shapes or leaf positions) and beam intensities at all control points simultaneously. Therefore, MLC constraints and delivery time are considered during plan optimization, which makes the problem harder to solve. The solution methods proposed in the literature are generally heuristic algorithms. One of the earliest algorithm is introduced by Earl *et al.* [49] in 2003 for IMAT technology, which starts with a number of apertures at equally spaced control points with 10°-spacing. Each one of the apertures fits to the beam's eye view (BEV) of the target seen from the linear accelerator at the corresponding control point. Then, simulated annealing method is used to optimize the leaf positions

and intensities of the initial apertures. Since, it is not possible to vary dose rate (or gantry speed) in IMAT, more than one overlapping arcs are required to realize a plan with acceptable dose distribution. The work by Otto [12] is the first study on VMAT planning, which is commercialized under trade name Rapid Arc (Varian Medical Systems, Palo Alto, CA, USA). The proposed method starts with a relatively coarse sampling of the control points and they are increased progressively according to a schedule. The aperture shape of a newly added control point is determined by linear interpolation of the existing apertures at adjacent control points, and its radiation intensity is calculated using a linear function of the adjacent intensities. Each time a new control point is added to the plan, a number of simulated annealing iterations are conducted. At each iteration, one of the existing control points is randomly selected, and the current dose intensity or the position of a leaf is changed. If the new aperture is feasible and there is an improvement in the objective function, then the new solution is accepted. Yan *et al.* [50] propose a similar heuristic algorithm that starts with a coarse sampling of the control points and uses a progressive sampling strategy to find the final VMAT plan. Bzdusek *et al.* [51] and Bedford [52] propose a three-step method, where they initially apply a two-step approach similar to the one explained in the previous section to find a good starting point for their DAO methods. Namely, they find initial apertures at the first two steps; then they refine them in the third step where aperture shapes and intensities are decision variables. The algorithm in [51], which is commercialized under trade name Pinnacle SmartArc (Philips Medical Systems, Madison, WI, USA), decompose a set of fluence maps obtained at equally spaced control points with 24°-spacing into a number of apertures. Then, for each one of the control points only 2 apertures are selected and distributed over the arc. They refine the resulting arc by a local gradient based algorithm at the third step. Christiansen *et al.* [53] modify this algorithm in order to make the continuous aperture dose calculation possible. In Chapter 7 we introduce a two-phase heuristic that has similarities with these three-step approaches. In the first phase we find an initial treatment arc in two steps, where in the first step instead of solving a standard FMO problem we solve a linear programming model based on one of our optimization models. It finds a number of fluence maps with additional properties (e.g. the intensities of beamlets are bounded

from above by the maximum deliverable radiation intensity at the corresponding control point) for a subset of predefined control points. Then, in the second step we perform an arc sequencing heuristic to obtain apertures from these fluence maps. In the second phase of our algorithm we improve this initial treatment arc using column generation.

There are other studies in the literature that use column generation method in their DAO heuristic algorithms. Men *et al.* [54] formulate a large-scale convex programming model in which the cost function consists of quadratic one-sided voxel-based penalties and a penalty-based soft constraint for the maximum dose rate variation limitation. They start with an empty set of apertures and generate one aperture for an unoccupied control point at each iteration, which is compatible with the previously generated ones with respect to the maximum leaf motion speed. Then, the dose intensities of all generated apertures are optimized in the master problem by means of the gradient projection method [55]. They do not consider the dose rate limitation at control points, which is unrealistic according to capabilities of the existing linear accelerators, and Peng *et al.* [13] improve this solution approach in their new column generation based greedy heuristic that also takes into account dose rate and gantry speed limitations. In a recent study, Mahnam *et al.* [56] develop a large-scale nonlinear integer programming model that has a quadratic voxel-based least square penalty function as an objective function similar to the one in [13]. They also propose a column generation based heuristic that generates a set of sequential apertures as a new column by solving the pricing subproblems formulated as shortest path problems. They assume that the MLC system has only consecutive ones property, hence the apertures that form a partial arc can be decomposed into rows and can be handled independently. Namely, they find as many partial row arcs as the number of rows in the MLC system; then their union yields the aperture set in the partial arc. Then, they integrate DVH criteria into their column generation algorithm in [57]. We use a similar approach in our BP algorithms and also column generation heuristic given in Chapter 6 and Chapter 7, and formulate the pricing subproblems as network optimization problems on acyclic networks using the key point of decomposing the partial treatment arcs into rows and generating them separately. However, we generate rows of a full treatment arc and do not need any post-optimization.

Papp and Unkelbach [26] enforce unidirectional leaf motion over an arc segment, and determine the apertures by solving a sequence of convex optimization problems. They assume that gantry speed and dose rate are constant during rotation. Peng *et al.* [58] also propose a heuristic approach to solve VMAT with constant gantry speed and dose rate. On the other hand, Hoegele *et al.* [59] optimize leaf motion by utilizing a priori knowledge about the type of the leaf motion pattern during the radiation delivery. We also assume that the gantry rotates around the patient at a constant speed, however, there is not such an assumption on dose rate.

The studies of Gozbasi [60], Akartunali *et al.* [15], and Song *et al.* [61] are the first works that formulate MILP models for the VMAT planning problem in which an aperture and radiation intensity are optimized at each control point subject to a part of the clinical requirements. In [60] and [61] some of the treatment related constraints are relaxed and they are tried to be satisfied in the objective function (i.e. by minimizing total deviation from the prescribed doses or minimizing the weighted sum of the average dose on critical structures, etc.). On the other hand, Akartunali *et al.* [15] 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 target voxels that absorbs at least the prescribed amount of radiation. They make the first step towards the development of exact methods, however, they finally suggest heuristics to obtain good feasible treatment plans, which are clinically acceptable as well. We develop two MILP formulations for VMAT planning, which are explained in Chapter 4. They consider all mechanical limitations of the linear accelerator and MLC system as well as dose requirements of treatment. They are also different form the formulations introduced to the literature with respect to the objective function as well as the definition of the MLC leaves and the corresponding constraints. We develop algorithms (in Chapter 5 and Chapter 6) to solve one of these comprehensive VMAT planning models, which are the first exact solution algorithms proposed to the literature to the best of our knowledge.

For more detail about rotational therapy planning we recommend the studies of Unkelbach *et al.* [62] that reviews the mathematical optimization methods used in

VMAT planning and Cedric and Tang [63] that reviews mainly IMAT studies from a clinical point of view. Also, there is a recent comprehensive review of Breedveld *et al.* [64] that describes the use of multi-criteria optimization and decision-making methods in radiation therapy as well as clinical details of treatment. Finally, in this dissertation we consider co-planar treatment arc as the studies reviewed so far. However there are also studies in the literature that optimize VMAT plans for non-coplanar geometries obtained by couch rotation [65, 66].

3.2.3. Problem Sizes

The mathematical models proposed for VMAT planning have been relatively simple until the last few years. Typically, they do not include dose distribution restrictions as hard constraints. These constraints are forced to be satisfied by means of penalty terms added to the objective function. It is also highlighted in [15] that these studies can solve clinical size problems since they use such an objective function to reach feasibility. Thus, a plan obtained by solving such a model is not guaranteed to be clinically acceptable unless the value of the corresponding objective terms become zero. [13, 54, 56, 61] are examples of such studies, where it is possible to consider the instances with more structures and large number of voxels. However, the resulting plans do not guarantee the satisfaction of dose-volume restrictions.

Men *et al.* [54] test their algorithm on ten clinical cases with a beamlet size of 1 $\times$ 1 cm^2 and voxel size of 2.5 mm^3. However, they indicate that for unspecified tissues outside the TV and OARs they increase the voxel size in each dimension by a factor of two to reduce the optimization problem size. Total number of voxels varies in each cases and ranges from 28 931 to 74 438. Peng *et al.* [13] test their algorithm, which is an extension of the one proposed in [54], on 5 real prostate cancer data sets. They also use a down-sampled voxel grid: they select one grid point for every two voxels along each one of the three dimensions in critical structures, and one grid point for every four voxels along each one of the three dimensions in unspecified tissues. The resulting data sets has a total number of voxels varying between 9 602 and 13 769. They also increase voxel sizes and use a lower resolution ($4 \times 4 \times 2.5$ mm^3), and use

1×1 cm^2 beamlets. Song *et al.* [61] use two data sets provided by the open source platform CERR on MATLAB. For prostate case they consider TV and 4 OARs. They sample one voxel out of every two voxels in OAR structures to reduce total number of voxels, however they do not report the actual voxel numbers. Finally, Mahnam *et al.* [56] use CORT prostate dataset provided by Craft *et al.* [17], which we use also in the computational experiments for the algorithms provided in Chapter 5 and Chapter 6, and apply a clustering algorithm to sample down the voxels. They indicate that 5% of OAR voxels and 15% of target voxels are included in their optimization model. They consider both of the TVs and 4 OARs (rectum, bladder, left and right femoral heads). As a result, there are approximately 3 500 voxels in their experiments. On the other hand, the mathematical model of Akartunalı *et al.* [15], which is the closest one to ours, since they introduce all treatment requirements except the partial volume constraints of TVs to their mathematical model as hard constraints. They maximize total number of voxels absorbing radiation at least the prescribed amount. They are not able to access to an in-house dose deposition coefficient calculation software, and they generate their test instances by themselves. There are 33 instances differing from each other according to the total number of voxels, MLC dimensions, and other parameters (voxel numbers are varying between 216 and 15 625, but the maximum number of control points is 16 in these instances). They also generate 7 extra large instances in order to test one of their Guided Variable Neighborhood Search heuristic. Their integer programming based exact algorithms are not able to solve these extra large instances. To give more detail, there is only one instance with 180 control points and a MLC size of 10×10 with 6 750 voxels. All other instances have either less voxels and control points, or the dimension of the MLC system is small.

4. PROBLEM DEFINITON AND FORMULATIONS[1]

In this chapter we explain two different MILP models we have developed for VMAT planning. They are called as VMATP-1 and VMATP-2, respectively. The proposed models directly optimize the aperture shape and dose intensity at each control point while satisfying dose prescriptions and mechanical limitations of the linear accelerator and the MLC system. The objective is to minimize total radiation intensity during treatment in both models. The main difference between them stems from the definition of the decision variables related to the leaf pairs of the MLC system. This distinction also requires other modifications in the mathematical models. First we start by explaining VMATP-1 step by step and then continue by giving the differences of VMATP-2.

A VMAT plan must satisfy both radiation therapy dose prescriptions and mechanical limitations of the linear accelerator and the MLC system. Our first model VMATP-1 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-1 determines the aperture shape and the amount of radiation to be delivered at each of the control points. Common parameters and decision variables used to formulate both models are summarized in Table 4.1 and Table 4.2, respectively. We list the additional variables of VMATP-1 in Table 4.3.

A two-dimensional $m \times n$ matrix represents an aperture at a control point. The number of MLC leaf pairs, and thus the number of rows is m and the number of columns is n. We introduce a number of nonnegative integer variables and binary variables to form each one of these matrices. For each row i at control point k two nonnegative integer variables l_{ik} and r_{ik} define positions of the left and right leaves, respectively. There are also n binary variables for each row i at control point k, and

[1] An earlier version of this chapter appears in [67].

Table 4.1. Common parameters of VMATP-1 and VMATP-2.

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).
c	Index for a partial volume constraint of OAR o (c=1,...,C_o) or TV t (c=1,...,C_t).
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 on the amount of radiation dose absorbed by a target voxel in TV t (in Gy).
U_t^{TV}	Upper bound on the amount of radiation dose absorbed by target voxel in TV t (in Gy).
U_{oc}^{OAR}	Tolerance radiation dose amount of the cth partial volume constraint of OAR volume o (in Gy).
$\overline{d}_{tc}$	The cth prescribed dose for TV t (in Gy).
D_{ijkv}	Dose influence matrix (in Gy/MU).
δ	The maximum allowable distance (in beamlet) that a leaf can move between two consecutive control points.
α_{tc}^{TV}	The minimum ratio of voxels in TV t that receive radiation at least the prescribed dose $\overline{d}_{tc}$.
α_{oc}^{OAR}	The minimum ratio of voxels in OAR volume o that receive radiation at most the tolerance dose U_{oc}^{OAR}.
L^{mu}	Lower bound on radiation dose intensity at a control point (in MU).
U^{mu}	Upper bound on radiation dose intensity at a control point (in MU).

Table 4.2. Common decision variables of VMATP-1 and VMATP-2.

Variable	**Definition**
z_{ijk}	Binary variable, 1 if the jth beamlet of row i at control point k is open, 0 otherwise (j=1,...,n).
mu_k	Nonnegative continuous variable, radiation dose intensity (in MU) at control point k.
d_v	Nonnegative continuous variable, the total amount of radiation dose absorbed by voxel v (in Gy).
a_{ijk}	Nonnegative continuous variable, radiation dose intensity (in MU) delivered from the jth beamlet of row i at control point k.
ξ_{tc}^{TV}	Continuous variable used in constraint (4.21), the radiation dose absorbed by the $((1\text{-}\alpha_{tc}^{TV})\lvert V_t^{TV}\rvert)$th voxel in TV t receiving the lowest radiation.
ξ_{oc}^{OAR}	Continuous variable used in constraint (4.26), the radiation dose absorbed by the $((1\text{-}\alpha_{oc}^{OAR})\lvert V_o^{OAR}\rvert)$th voxel in OAR volume o receiving the highest radiation.
x_{tcv}	Nonnegative continuous variable for the surplus of the value ξ_{tc}^{TV} by the dose received by voxel v in TV t.
y_{ocv}	Nonnegative continuous variable for the surplus of the value ξ_{oc}^{OAR} by the dose received by voxel v in OAR o.

Table 4.3. Additional variables of VMATP-1.

Variable	**Definition**
l_{ik}	Nonnegative integer variable, the position of the left leaf (i.e. the rightmost beamlet closed by the left leaf on row i at control point k).
r_{ik}	Nonnegative integer variable, the position of the right leaf (i.e. the leftmost beamlet closed by the right leaf on row i at control point k).

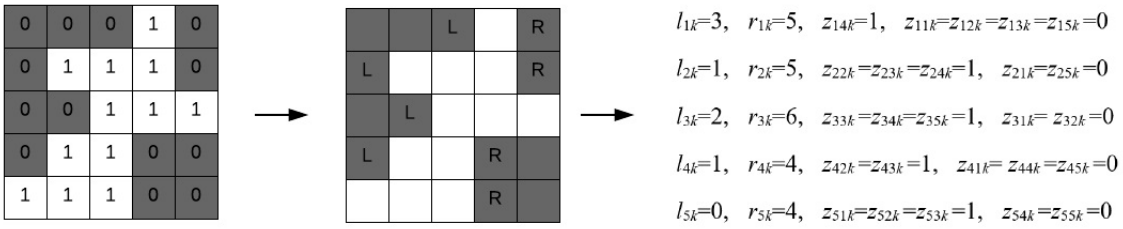

Figure 4.1. An aperture and its decision variables.

the binary variable z_{ijk} represents the beamlet j at this row and takes value of 1 if it is open. Only the beamlets between the leaf pairs are open. In Figure 4.1 an aperture consisting of five leaf pairs ($m = 5$) and five columns ($n = 5$) at control point k is illustrated with corresponding decision variables. Note that at row 3 the right leaf is at its home position and takes value $n + 1 = 6$, and at row 5 the left leaf is at its home position and takes value 0.

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 introduce the following constraints similar to the studies both in VMAT planning (e.g. [15]) and IMRT planning (e.g. [29]) in order to satisfy it:

$$r_{ik} - l_{ik} \geq 1 \qquad i = 1, \ldots, m; k = 1, \ldots, K \tag{4.1}$$

$$r_{ik} - j z_{ijk} \geq 1 \qquad i = 1, \ldots, m; j = 1, \ldots, n; k = 1, \ldots, K \tag{4.2}$$

$$(n + 1 - j) z_{ijk} + l_{ik} \leq n \qquad i = 1, \ldots, m; j = 1, \ldots, n; k = 1, \ldots, K \tag{4.3}$$

$$r_{ik} - l_{ik} - \sum_{j=1}^{n} z_{ijk} = 1 \qquad i = 1, \ldots, m; k = 1, \ldots, K \tag{4.4}$$

$$\mathbf{l} \in \mathbb{Z}_{+}^{m \times K}; \mathbf{r} \in \mathbb{Z}_{+}^{m \times K}; \tag{4.5}$$

$$\mathbf{z} \in \{0, 1\}^{m \times n \times K}. \tag{4.6}$$

For a given row i at control point k, Constraint (4.1) prevents the left and right leaves from overlapping. Constraint (4.2)– Constraint (4.4) 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 $n+1$. Note that we remove constraint (4.1) in chapters 6 and 7, since it is noticed that this constraint is redundant.

Another mechanical limitation of the MLC system, which is generally taken into account in VMAT studies (e.g. [15, 61]), 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 similarities:

$$l_{i(k+1)} - l_{ik} \leq \delta \qquad i = 1, \ldots, m; k = 1, \ldots, K-1 \tag{4.7}$$

$$l_{ik} - l_{i(k+1)} \leq \delta \qquad i = 1, \ldots, m; k = 1, \ldots, K-1 \tag{4.8}$$

$$r_{i(k+1)} - r_{ik} \leq \delta \qquad i = 1, \ldots, m; k = 1, \ldots, K-1 \tag{4.9}$$

$$r_{ik} - r_{i(k+1)} \leq \delta \qquad i = 1, \ldots, m; k = 1, \ldots, K-1. \tag{4.10}$$

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 (4.1)–(4.10) 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 (4.7)–(4.10). In addition to the aperture shape, VMATP-1 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 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 the 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

$$mu_k \geq L^{mu} \qquad k = 1, \ldots, K \tag{4.11}$$

$$mu_k \leq U^{mu} \qquad k = 1, \ldots, K \tag{4.12}$$

$$\mathbf{mu} \in \mathbb{R}^K_+, \tag{4.13}$$

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

A VMAT 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 forces 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

$$d_v - \sum_{k=1}^{K} \sum_{i=1}^{m} \sum_{j=1}^{n} D_{ijkv} z_{ijk} mu_k = 0 \qquad v \in V = V^{TV} \cup V^{OAR}. \tag{4.14}$$

Note that V is the set of all voxels, namely it is the union of all TVs (V^{TV}) and OARs (V^{OAR}). Note also that (4.14) includes nonlinear terms created by the product of binary variables **z** with the continuous variables **mu**. We use the linearization method introduced by Mccormick in 1976 [68], which eventually forms the convex envelop of general bilinear terms, to linearize constraint (4.14). We introduce auxiliary variable a_{ijk} for each beamlet to represent its radiation intensity and obtain

$$d_v - \sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n} D_{ijkv}a_{ijk} = 0 \quad v \in V = V^{TV} \cup V^{OAR} \tag{4.15}$$

$$a_{ijk} \leq U^{mu} z_{ijk} \qquad i = 1,\dots,m; j = 1,\dots,n; k = 1,\dots,K \tag{4.16}$$

$$a_{ijk} \geq mu_k - U^{mu}(1 - z_{ijk}) \qquad i = 1,\dots,m; j = 1,\dots,n; k = 1,\dots,K \tag{4.17}$$

$$a_{ijk} \leq mu_k \qquad i = 1,\dots,m; j = 1,\dots,n; k = 1,\dots,K \tag{4.18}$$

$$\mathbf{d} \in \mathbb{R}_{+}^{|V|} \tag{4.19}$$

$$\mathbf{a} \in \mathbb{R}_{+}^{m\times n\times K}. \tag{4.20}$$

The radiation passes through only the open beamlets, thus Constraint (4.16) – Constraint (4.18) force a_{ijk} to take value of mu_k if associated beamlet is open, and 0 otherwise.

Now it is possible to include the clinical requirements using the total absorbed radiation dose amounts of voxels. Similar to [14] and [60], we use Conditional Value-at-Risk (CVaR) approach, which was originally developed by Rockafellar *et al.* in 2000 [69] for portfolio optimization, to formulate partial volume constraints. For each TV t the following partial volume constraints are introduced:

$$\xi_{tc}^{TV} - \frac{1}{(1-\alpha_{tc}^{TV})|V_t^{TV}|}\sum_{v\in V_t^{TV}} x_{tcv} \geq \overline{d}_{tc} \quad t = 1,\dots,T; c = 1,\dots,C_t \tag{4.21}$$

$$x_{tcv} \geq \xi_{tc}^{TV} - d_v \qquad t = 1,\dots,T; c = 1,\dots,C_t; v \in V_t^{TV} \tag{4.22}$$

$$\mathbf{x} \in \mathbb{R}_{+}^{\sum_{t=1}^{T} C_t|V_t^{TV}|}; \boldsymbol{\xi}^{TV} \in \mathbb{R}^{\sum_{t=1}^{T} C_t}. \tag{4.23}$$

The average dose of the (1-α_{tc}^{TV})$|V_t^{TV}|$ voxels receiving the lowest dose in TV t, namely the *lower mean tail dose at level* α_{tc}^{TV} is forced to be at least the prescription dose. In other words, at least $\alpha_{tc}^{TV}|V_t^{TV}|$ voxels absorb radiation more than or equal to $\overline{d}_{tc}$. Note that, there may be more than one partial volume restriction for a TV (or an OAR), hence we introduce c index to the model that indicates the cth partial volume constraint. Furthermore, there are full volume constraints for each TV:

$$d_v \geq L_t^{TV} \qquad t = 1, \ldots, T; v \in V_t^{TV} \tag{4.24}$$

$$d_v \leq U_t^{TV} \qquad t = 1, \ldots, T; v \in V_t^{TV}, \tag{4.25}$$

which 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:

$$\xi_{oc}^{OAR} + \frac{1}{(1-\alpha_{oc}^{OAR})|V_o^{OAR}|} \sum_{v \in V_o^{OAR}} y_{ocv} \leq U_{oc}^{OAR} \quad o = 1, \ldots, O; c = 1, \ldots, C_o \tag{4.26}$$

$$y_{ocv} \geq d_v - \xi_{oc}^{OAR} \qquad o = 1, \ldots, O; c = 1, \ldots, C_o; \quad v \in V_o^{OAR} \tag{4.27}$$

$$\mathbf{y} \in \mathbb{R}_+^{\sum_{o=1}^{O} C_o |V_o^{OAR}|}; \boldsymbol{\xi}^{OAR} \in \mathbb{R}^{\sum_{o=1}^{O} C_o}. \tag{4.28}$$

The average dose of the (1-α_{oc}^{OAR})$|V_o^{OAR}|$ voxels absorbing the highest doses in OAR o, namely the *upper mean tail dose at level* α_{oc}^{OAR} is forced to be at most its tolerance dose limit U_{oc}^{OAR}. To give more detail about CVaR approach as discussed in [24], continuous variable ξ_{oc}^{OAR} in constraint (4.26) is a bound on the upper value-at-risk (VaR) at level α_{oc}^{OAR}, which is the smallest dose level with the property that no more than $100(1-\alpha_{oc}^{OAR})\%$ of OAR o receives a larger dose. Also, the left hand side of constraint (4.26) is the upper α_{oc}^{OAR}-CVaR, which is the mean of all doses that exceed the upper α_{oc}^{OAR}-VaR. The variable y_{ocv} is the surplus of the value ξ_{oc}^{OAR} by the dose received by voxel v in OAR o. Furthermore, if constraint (4.26) is satisfied as an equality in an optimal solution then ξ_{oc}^{OAR} equals to the VaR corresponding to that constraint. For

more detail about CVaR method we refer the reader to the study of Romeijn *et al.* [14] where it is applied for developing a linear-programming-based approach to solve FMO problem in IMRT planning. Finally, the objective function

$$\min \sum_{k=1}^{K} mu_k \tag{4.29}$$

minimizes total radiation intensity (in MU) the patient receives during his/her treatment. VMATP-1 finds an optimal plan minimizing total dose intensity among all feasible treatment plans.

We have explained VMATP-1 model so far and continue by explaining the second model, which we call VMATP-2. The parameters in Table 4.1 and decision variables in Table 4.2 are used to formulate VMATP-2. There are also additional decision variables to define the position of the leaves of MLC, which are summarized Table 4.4.

Table 4.4. Additional variables of VMATP-2.

Variable	**Definition**
l_{ijk}	Binary variable used to represent the position of the left leaf; it is set to 1 if jth beamlet is the rightmost closed one on row i at control point k.
r_{ijk}	Binary variable used to represent the position of the right leaf; it is set to 1 if jth beamlet is the leftmost closed one on row i at control point k.

Similar to [60, 61], we introduce two binary variables for each beamlet on a given row; l_{ijk} variable is related to the left leaf and r_{ijk} is related to the right leaf. For a given row i at control point k exactly one l_{ijk} variable takes value 1. Similarly, exactly one r_{ijk} variable is forced to be 1. For example, the left leaf on the first row of the aperture illustrated in Figure 4.1 blocks the first 3 beamlets, namely the rightmost closed beamlet is the third one. Therefore, only l_{13k} equals to 1 and the remaining ones are set to 0 ($l_{11k} = l_{12k} = l_{14k} = l_{15k} = 0$).

VMATP-2:

$$\min \sum_{k=1}^{K} mu_k \tag{4.29}$$

s.t.

$$(4.6), (4.11) - (4.13), (4.15) - (4.28),$$

$$\sum_{j=0}^{n} l_{ijk} = 1 \qquad i = 1, \ldots, m; k = 1, \ldots, K \tag{4.30}$$

$$\sum_{j=1}^{n+1} r_{ijk} = 1 \qquad i = 1, \ldots, m; k = 1, \ldots, K \tag{4.31}$$

$$\sum_{p=0}^{j} r_{i(p+1)k} - \sum_{p=0}^{j} l_{ipk} \leq 0 \qquad i = 1, \ldots, m; j = 0, \ldots, n; k = 1, \ldots, K \tag{4.32}$$

$$l_{ij(k+1)} - \sum_{p=max(0,j-\delta)}^{min(n,j+\delta)} l_{ipk} \leq 0 \qquad i = 1, \ldots, m; j = 0, \ldots, n; \quad k = 1, \ldots, K-1 \tag{4.33}$$

$$r_{ij(k+1)} - \sum_{p=max(1,j-\delta)}^{min(n+1,j+\delta)} r_{ipk} \leq 0 \qquad i = 1, \ldots, m; j = 1, \ldots, n+1; \quad k = 1, \ldots, K-1 \tag{4.34}$$

$$z_{ijk} - \sum_{p=0}^{j-1} l_{ipk} + \sum_{p=1}^{j} r_{ipk} = 0 \qquad i = 1, \ldots, m; j = 1, \ldots, n; k = 1, \ldots, K \tag{4.35}$$

$$\mathbf{l} \in \{0,1\}^{m \times (n+1) \times K};$$

$$\mathbf{r} \in \{0,1\}^{m \times (n+1) \times K}. \tag{4.36}$$

Constraint (4.30) satisfies that there is exactly one rightmost closed beamlet, which defines the position of the left leaf. The similar constraint for the right leaf is (4.31) and there can be exactly one leftmost closed beamlet. In order to prevent the overlapping of the leaf pairs (4.32) is introduced to the model. Constraint (4.33) and Constraint (4.34) limit the leaf motion during rotation: a leaf can move at most δ beamlets. As

in VMATP-1 model, we need to enforce z_{ijk} variables to be 1 if the corresponding beamlets are open and (4.35) satisfies this requirement. It also satisfies the consecutive ones property of the apertures. The last constraint (4.36) are the binary restrictions for the new variables.

Observe that VMATP formulations are different according to the geometric part of the problem where the apertures are determined at each control point and the leaf motion limitations are controlled. The remaining part, which finds radiation intensities and satisfies the clinical requirements are exactly the same. In VMATP-1, we define two nonnegative integer variables in order to determine the positions of the leaves (i.e. one for each of the left and right leaf). However, a binary variable is introduced for each one of the beamlet and also for the home positions of the leaves in VMATP-2 formulation. Thus, in VMATP-1 total number of nonnegative integer variables to define the position of the leaves is $2 \times m \times K$, on the other hand, there are $2 \times (n+1) \times m \times K$ binary variables in VMATP-2. Moreover, total number of constraints to satisfy the leaf motion limitations in VMATP-2 is $\frac{n+1}{2}$ times larger than the ones in VMATP-1. Observe that there are $4 \times m \times (K-1)$ such constraints in VMATP-1 and this number increases to $2 \times (n+1) \times m \times (K-1)$ in VMATP-2. Also, we observe that Constraint (4.1) is redundant in VMATP-1 and removed in Chapter 6. Thus, there are also $m \times K$ additional constraints in VMATP-2. As shown in the computational experiments in Section 8.2, where we evaluate the formulations on a large number of test instances, VMATP-1 performs better than VMATP-2 especially for large instances, which is not surprising. As the size of the test instances increases total number of these decision variables and constraints remain the same. However, the problems becomes easily intractable. Also, as explained in Chapter 6 in detail, the geometry part of the problem is decomposed into m subproblems and solved as shortest path problems. It can be observed that defining nonnegative integer variables for the left and right leaves is more suitable to formulate the geometry part as a network model.

In particular, the definition of the positions of the leaves using nonnegative integer variables is a new approach in literature. Namely, VMATP-1 also differs from the existing formulations with respect to these decision variables and associated constraints.

5. SOLUTION METHODS: BENDERS DECOMPOSITION ALGORITHMS[2]

Benders decomposition was proposed by Benders in 1962 [71], 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 [72].

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 linear program (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 VMATP-1. Because, only the variables that form apertures are integer valued. In this chapter, we use Benders decomposition and develop efficient solution algorithms after improving its naive form by means of computational strategies.

5.1. Benders Reformulation

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. Using this observation we decompose the original problem into a relaxed master problem and a subproblem. The relaxed master problem produces

[2] An earlier version of this chapter appears in [70].

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

$$\min \sum_{k=1}^{K} mu_k \tag{4.29}$$

s.t.

$$d_v - \sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n} D_{ijkv}a_{ijk} = 0 \qquad v \in V = V^{TV} \cup V^{OAR} \qquad (\pi_v) \tag{4.15}$$

$$a_{ijk} \leq U^{mu}\hat{z}_{ijk} \qquad i = 1,\dots m; j = 1,\dots,n; \; k = 1,\dots K \qquad (\beta^1_{ijk}) \tag{5.1}$$

$$a_{ijk} \geq mu_k - U^{mu}(1-\hat{z}_{ijk}) \qquad i = 1,\dots m; j = 1,\dots,n; \; k = 1,\dots K \qquad (\beta^2_{ijk}) \tag{5.2}$$

$$a_{ijk} \leq mu_k \qquad i = 1,\dots m; j = 1,\dots,n; \; k = 1,\dots K \qquad (\beta^3_{ijk}) \tag{4.18}$$

$$\xi^{TV}_{tc} - \frac{1}{(1-\alpha^{TV}_{tc})|V^{TV}_t|}\sum_{v\in V^{TV}_t} x_{tcv} \geq \overline{d}_{tc} \qquad t = 1,\dots T; c = 1,\dots,C_t \qquad (\theta^1_{tc}) \tag{4.21}$$

$$x_{tcv} \geq \xi^{TV}_{tc} - d_v \qquad t = 1,\dots,T; c = 1,\dots,C_t; \; v \in V^{TV}_t \qquad (\tau^1_{tcv}) \tag{4.22}$$

$$d_v \geq L^{TV}_t \qquad t = 1,\dots,T; v \in V^{TV}_t \qquad (\epsilon^1_{tv}) \tag{4.24}$$

$$d_v \leq U^{TV}_t \qquad t = 1,\dots,T; v \in V^{TV}_t \qquad (\epsilon^2_{tv}) \tag{4.25}$$

$$\xi_{oc}^{OAR} + \frac{1}{(1-\alpha_{oc}^{OAR})|V_o^{OAR}|} \sum_{v \in V_o^{OAR}} y_{ocv} \leq U_{oc}^{OAR} \quad o = 1, \dots O; \quad c = 1, \dots, C_o \quad (\theta_{oc}^2) \quad (4.26)$$

$$y_{ocv} \geq d_v - \xi_{oc}^{OAR} \quad o = 1, \dots, O; c = 1, \dots, C_o; \quad v \in V_o^{OAR} \quad (\tau_{ocv}^2) \quad (4.27)$$

$$mu_k \geq L^{mu} \quad k = 1, \dots, K \quad (\mu_k^1) \quad (4.11)$$

$$mu_k \leq U^{mu} \quad k = 1, \dots, K \quad (\mu_k^2) \quad (4.12)$$

$(4.13), (4.19) - (4.20), (4.23), (4.28),$

and

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

$$\max \sum_{k=1}^{K} \sum_{i=1}^{m} \sum_{j=1}^{n} U^{mu}(-\hat{z}_{ijk}\beta_{ijk}^1 + (\hat{z}_{ijk} - 1)\beta_{ijk}^2) + \sum_{t=1}^{T} \sum_{c=1}^{C_t} \theta_{tc}^1 \overline{d}_{tc} \quad (5.3)$$
$$+ \sum_{t=1}^{T} \sum_{v \in V_t^{TV}} (L_t^{TV}\epsilon_{tv}^1 - U_t^{TV}\epsilon_{tv}^2) - \sum_{o=1}^{O} \sum_{c=1}^{C_o} \theta_{oc}^2 U_{oc}^{OAR} + \sum_{k=1}^{K} (L^{mu}\mu_k^1 - U^{mu}\mu_k^2)$$

s.t.

$$\pi_v + \tau_{tcv}^1 + \epsilon_{tv}^1 - \epsilon_{tv}^2 \leq 0 \quad t = 1, \dots, T; c = 1, \dots, C_t; \quad v \in V_t^{TV} \quad (d_v) \quad (5.4)$$

$$\pi_v - \tau_{ocv}^2 \leq 0 \quad o = 1, \dots, O; c = 1, \dots, C_o; \quad v \in V_o^{OAR} \quad (d_v) \quad (5.5)$$

$$-\sum_{v \in V} D_{ijkv}\pi_v - \beta_{ijk}^1 + \beta_{ijk}^2 - \beta_{ijk}^3 \leq 0 \quad i = 1, \dots, m; j = 1, \dots, n; \quad k = 1, \dots, K \quad (a_{ijk}) \quad (5.6)$$

$$-\sum_{i=1}^{m} \sum_{j=1}^{n} (\beta_{ijk}^2 - \beta_{ijk}^3) + \mu_k^1 - \mu_k^2 \leq 1 \quad k = 1, \dots, K \quad (mu_k) \quad (5.7)$$

$$-\theta_{oc}^2 + \sum_{v \in V_o^{OAR}} \tau_{ocv}^2 = 0 \quad o = 1, \dots, O; c = 1, \dots, C_o \quad (\xi_{oc}^{OAR}) \quad (5.8)$$

$$\theta_{tc}^1 - \sum_{v \in V_t^{TV}} \tau_{tcv}^1 = 0 \quad t = 1, \dots, T; c = 1, \dots, C_t \quad (\xi_{tc}^{TV}) \quad (5.9)$$

$$-\frac{1}{(1-\alpha_{oc}^{OAR})|V_o^{OAR}|}\theta_{oc}^2+\tau_{ocv}^2\le 0 \qquad o=1,\dots,O; c=1,\dots,C_o; \quad v\in V_o^{OAR} \qquad (y_{ocv}) \quad (5.10)$$

$$-\frac{1}{(1-\alpha_{tc}^{TV})|V_t^{TV}|}\theta_{tc}^1+\tau_{tcv}^1\le 0 \qquad t=1,\dots,T; c=1,\dots,C_t; \quad v\in V_t^{TV} \qquad (x_{tcv}) \quad (5.11)$$

$$\boldsymbol{\pi}\in\mathbb{R}^{|V|}; \boldsymbol{\beta}^1\in\mathbb{R}_+^{m\times n\times K}; \boldsymbol{\beta}^2\in\mathbb{R}_+^{m\times n\times K};$$
$$\boldsymbol{\beta}^3\in\mathbb{R}_+^{m\times n\times K}; \boldsymbol{\theta}^1\in\mathbb{R}_+^{\sum_{t=1}^T C_t}; \boldsymbol{\theta}^2\in\mathbb{R}_+^{\sum_{o=1}^O C_o};$$
$$\boldsymbol{\tau}^1\in\mathbb{R}_+^{\sum_{t=1}^T C_t|V_t^{TV}|}; \boldsymbol{\tau}^2\in\mathbb{R}_+^{\sum_{o=1}^O C_o|V_o^{OAR}|};$$
$$\boldsymbol{\epsilon}^1\in\mathbb{R}_+^{|V^{TV}|}; \boldsymbol{\epsilon}^2\in\mathbb{R}_+^{|V^{TV}|}; \boldsymbol{\mu}^1\in\mathbb{R}_+^K; \boldsymbol{\mu}^2\in\mathbb{R}_+^K. \quad (5.12)$$

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

$$f(\boldsymbol{\beta}^1,\boldsymbol{\beta}^2,\boldsymbol{\theta}^1,\boldsymbol{\theta}^2,\boldsymbol{\epsilon}^1,\boldsymbol{\epsilon}^2,\boldsymbol{\mu}^1,\boldsymbol{\mu}^2)=\sum_{k=1}^K\sum_{i=1}^m\sum_{j=1}^n U^{mu}(-z_{ijk}\beta_{ijk}^1+(z_{ijk}-1)\beta_{ijk}^2)+$$
$$\sum_{t=1}^T\sum_{c=1}^{C_t}\theta_{tc}^1\overline{d}_{tc}+\sum_{t=1}^T\sum_{v\in V_t^{TV}}(L_t^{TV}\epsilon_{tv}^1-U_t^{TV}\epsilon_{tv}^2)-\sum_{o=1}^O\sum_{c=1}^{C_o}\theta_{oc}^2U_{oc}^{OAR}+\sum_{k=1}^K(L^{mu}\mu_k^1-U^{mu}\mu_k^2)$$

and the Benders reformulation of VMATP-1 becomes

$$\min \eta \quad (5.13)$$

s.t.

$(4.1)-(4.10)$,

$$f(\boldsymbol{\beta}^1,\boldsymbol{\beta}^2,\boldsymbol{\theta}^1,\boldsymbol{\theta}^2,\boldsymbol{\epsilon}^1,\boldsymbol{\epsilon}^2,\boldsymbol{\mu}^1,\boldsymbol{\mu}^2)\le\eta \qquad \boldsymbol{\beta}^1,\boldsymbol{\beta}^2,\boldsymbol{\theta}^1,\boldsymbol{\theta}^2,\boldsymbol{\epsilon}^1,\boldsymbol{\epsilon}^2,\boldsymbol{\mu}^1,\boldsymbol{\mu}^2\in\Delta \quad (5.14)$$

$$f(\boldsymbol{\beta}^1,\boldsymbol{\beta}^2,\boldsymbol{\theta}^1,\boldsymbol{\theta}^2,\boldsymbol{\epsilon}^1,\boldsymbol{\epsilon}^2,\boldsymbol{\mu}^1,\boldsymbol{\mu}^2)\le 0 \qquad \boldsymbol{\beta}^1,\boldsymbol{\beta}^2,\boldsymbol{\theta}^1,\boldsymbol{\theta}^2,\boldsymbol{\epsilon}^1,\boldsymbol{\epsilon}^2,\boldsymbol{\mu}^1,\boldsymbol{\mu}^2\in\Omega \quad (5.15)$$

$$\eta\ge 0. \quad (5.16)$$

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 (LB) of η is set to 0. Constraint (4.1)–Constraint (4.10) determine a feasible aperture shape for each control point. Constraint (5.14) are Benders optimality cuts and Constraint (5.15) are Benders feasibility cuts and they all represent the subproblem. In the naive form of Benders decomposition, all Benders cuts are relaxed initially and the resulting relaxed master problem (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 stronger 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 LB 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.

5.2. Algorithmic and Modeling Improvements

5.2.1. Valid Inequalities

In the Benders reformulation the objective function (4.29) is removed since it belongs to the subproblem. Also, initial LB 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 becomes 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

$$\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n} z_{ijk}D_{ijkv}U^{mu} \geq L_t^{TV} \qquad t = 1,\ldots,T;\;\; v \in V_t \tag{5.17}$$

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:

$$a_{ijk} \leq U^{mu} z_{ijk} \qquad i = 1,\ldots,m; j = 1,\ldots,n; k = 1,\ldots,K \tag{5.18}$$

$$a_{ijk} \leq mu_k \qquad i = 1,\ldots,m; j = 1,\ldots,n; k = 1,\ldots,K \tag{5.19}$$

$$\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ijk}D_{ijkv} \geq L_t^{TV} \quad t = 1,\ldots,T; v \in V_t \tag{5.20}$$

$$\eta \geq \sum_{k=1}^{K} mu_k. \tag{5.21}$$

Note that constraints (5.18) and (5.19) are similar to the linearization constraints (4.16) and (4.18) in the VMATP, however (4.17) is relaxed. The addition of inequalities (5.20) 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 (5.21) is valid, and it improves the LB 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 (5.17) anymore, since it is replaced by (5.20), which is tighter. These extensions make the master problem harder to solve. However, according to our preliminary observations, they significantly improve the LB and performance

of the Benders decomposition algorithm as a consequence. Thus, in the final form of the method we add Constraint (5.18)–Constraint(5.21) 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.

5.2.2. Strong Benders Cuts

Stronger Benders cuts may improve the LB 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 [73]. 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 [74] 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 (5.3) 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 $\mathscr{Z}$ be the index set of all beamlets at all control points, namely the set of all (i, j, k) index combinations. Also let $\mathscr{Z}_0 \subseteq \mathscr{Z}$ be the index set of beamlets where $\hat{z}_{ijk} = 0$ and $\mathscr{Z}_1 \subseteq \mathscr{Z}$ 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 $\boldsymbol{\beta}^3$. Namely, we determine new values of $\beta^1_{ijk}, (i, j, k) \in \mathscr{Z}_0$, and $\beta^2_{ijk}, (i, j, k) \in \mathscr{Z}_1$ by

solving the following reduced DSP (RDSP)

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

$$\max \sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n}(-\beta^1_{ijk} - \beta^2_{ijk}) \tag{5.22}$$

s.t.

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

$$k = 1,\ldots,K \tag{5.23}$$

$$-\sum_{i=1}^{m}\sum_{j=1}^{n}(\beta^2_{ijk} - \beta^3_{ijk}) + \hat{\mu}^1_k - \hat{\mu}^2_k \leq 1 \qquad k = 1,\ldots,K \tag{5.24}$$

$$\beta^1_{ijk} = \hat{\beta}^1_{ijk} \qquad (i,j,k) \in \mathscr{Z}_1 \tag{5.25}$$

$$\beta^2_{ijk} = \hat{\beta}^2_{ijk} \qquad (i,j,k) \in \mathscr{Z}_0 \tag{5.26}$$

$$\boldsymbol{\beta}^1 \in \mathbb{R}_+^{mn|K|}; \boldsymbol{\beta}^2 \in \mathbb{R}_+^{mn|K|}; \boldsymbol{\beta}^3 \in \mathbb{R}_+^{mnK}. \tag{5.27}$$

In other words, we lift some of the **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 (as shown in Appendix A), since none of the cuts derived from an alternative optimal solution dominates (or is stronger than) this resulting one [74, 75]. It is worth noting that, in these studies, after setting permanent 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 (5.24) do not allow such decomposition. There exist other studies in the literature considering the use of strong cuts in Benders decomposition [76, 77].

5.2.3. 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 [78] and stronger combinatorial cuts [79, 80]. Ac-

cording to this approach it is possible to determine unbounded directions of a problem using an alternative polyhedron that is bounded. Fischetti *et al.* [78] 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 $(\hat{\mathbf{z}}, \hat{\eta})$:

$$\sum_{k=1}^{K} mu_k \leq \hat{\eta} \qquad (\pi_0) \qquad (5.28)$$

$$(4.11) - (4.13), (4.15), (5.1) - (5.2), (4.18) - (4.28),$$

where π_0 is the dual variable associated with (5.28). Observe that if $(\hat{\mathbf{z}}, \hat{\eta})$ is feasible for PFSP, then it is optimal for VMATP-1 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 $(\hat{\mathbf{z}}, \hat{\eta})$:

$$\max \sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n} U^{mu}(-\hat{z}_{ijk}\beta^1_{ijk} + (\hat{z}_{ijk} - 1)\beta^2_{ijk}) + \sum_{t=1}^{T}\sum_{c=1}^{C_t} \theta^1_{tc}\overline{d}_{tc} + \sum_{t=1}^{T}\sum_{v \in V_t^{TV}} (L_t^{TV}\epsilon^1_{tv} -$$

$$U_t^{TV}\epsilon^2_{tv}) - \sum_{o=1}^{O}\sum_{c=1}^{C_o} \theta^2_{oc}U^{OAR}_{oc} + \sum_{k=1}^{K}(L^{mu}\mu^1_k - U^{mu}\mu^2_k) - \pi_0\hat{\eta} \qquad (5.29)$$

s.t.

$$-\sum_{i=1}^{m}\sum_{j=1}^{n}(\beta^2_{ijk} - \beta^3_{ijk}) + \mu^1_k - \mu^2_k - \pi_0 \leq 0 \qquad k = 1, \ldots, K \qquad (5.30)$$

$$\pi_0 \in \mathbb{R}_+ \qquad (5.31)$$

$$(5.4) - (5.6), (5.8) - (5.12).$$

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

$$\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n} U^{mu}(-z_{ijk}\hat{\beta}^1_{ijk} + (z_{ijk}-1)\hat{\beta}^2_{ijk}) + \sum_{t=1}^{T}\sum_{c=1}^{C_t}\hat{\theta}^1_{tc}\overline{d}_{tc} + \sum_{t=1}^{T}\sum_{v\in V_t^{TV}}(L_t^{TV}\hat{\epsilon}^1_{tv} -$$
$$U_t^{TV}\hat{\epsilon}^2_{tv}) - \sum_{o=1}^{O}\sum_{c=1}^{C_o}\hat{\theta}^2_{oc}U^{OAR}_{oc} + \sum_{k=1}^{K}(L^{mu}\hat{\mu}^1_k - U^{mu}\hat{\mu}^2_k) - \hat{\pi}_0\eta \leq 0. \tag{5.32}$$

Furthermore, the unbounded objective function is set to 1 for normalization as done by Gleeson and Ryan [81], and

$$\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n} U^{mu}(-\hat{z}_{ijk}\beta^1_{ijk} + (\hat{z}_{ijk}-1)\beta^2_{ijk}) + \sum_{t=1}^{T}\sum_{c=1}^{C_t}\theta^1_{tc}\overline{d}_{tc} + \sum_{t=1}^{T}\sum_{v\in V_t^{TV}}(L_t^{TV}\epsilon^1_{tv} -$$
$$U_t^{TV}\epsilon^2_{tv}) - \sum_{o=1}^{O}\sum_{c=1}^{C_o}\theta^2_{oc}U^{OAR}_{oc} + \sum_{k=1}^{K}(L^{mu}\mu^1_k - U^{mu}\mu^2_k) - \pi_0\hat{\eta} = 1 \tag{5.33}$$
$$(5.4)-(5.6), (5.8)-(5.12), (5.30)-(5.31)$$

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

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

$$\min\ \pi_0 \tag{5.34}$$

s.t.

$$(5.4)-(5.6), (5.8)-(5.12), (5.30)-(5.31), (5.33)$$

minimizes π_0 over this polyhedron and we solve AP instead of DSP in Benders iterations to generate Benders cuts. Fischetti *et al.* [78] 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 (5.30) and (5.33) 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 (5.7) remains feasible in ad-

dition 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.

5.2.4. Combinatorial Benders Cut

Combinatorial Benders decomposition is an extension of traditional Benders decomposition method, where the problem is again decomposed into a master integer program and a linear programming subproblem. Rahmaniani *et al.* [72] 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 (5.35) 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 (4.21)–(4.22) associated with a TV, (4.26)–(4.27) associated with an OAR, or both. For these cases, to repair infeasibility, 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 $\mathscr{I} \subseteq \mathscr{Z}$ be the index set of the beamlets having strictly positive effect on at least one voxel, namely $\mathscr{I} = \{(i, j, k) : D_{ijkv} > 0, v \in V\}$. Hence, we can add the combinatorial

cut

$$\sum_{\substack{\hat{z}_{ijk}=0 \\ (i,j,k)\in\mathscr{I}}} z_{ijk} + \sum_{\substack{\hat{z}_{ijk}=1 \\ (i,j,k)\in\mathscr{I}}} (1 - z_{ijk}) \geq 1. \tag{5.35}$$

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 Çevik [80], we find a minimal infeasible system (MIS) of the subproblem when an infeasible solution is detected. Gleeson and Ryan [81] 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 $\mathscr{Z}^* \subseteq \mathscr{Z}$ be the index set of the beamlets that are associated with the MIS corresponding to $\hat{\mathbf{z}}$. The cut (5.35) is revised so that it only has $\mathbf{z}$ variables in $\mathscr{I} \cap \mathscr{Z}^*$:

$$\sum_{\substack{\hat{z}_{ijk}=0 \\ (i,j,k)\in\mathscr{I}\cap\mathscr{Z}^*}} z_{ijk} + \sum_{\substack{\hat{z}_{ijk}=1 \\ (i,j,k)\in\mathscr{I}\cap\mathscr{Z}^*}} (1 - z_{ijk}) \geq 1. \tag{5.36}$$

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 (5.36). The resulting Benders algorithm including the improvement strategies explained so far is given in Figure 5.1 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 [77, 80]. 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 benefiting from 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 in the implementation of the new cut selection strategy explained in Section 5.2.3. 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 (UB) 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 LB and the UB 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 LB 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 LB 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 UB 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 UB. 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.

5.2.5. A Relaxation of the Model

According to the results of the algorithm obtained by implementing the improvement strategies explained so far we can say that the LB is not strong. In order to alleviate this problem we strengthen the LB using a Lagrangean relaxation approach. We dualize the complicating constraints (4.16) and (4.17) in VMATP-1 with nonnegative multipliers $\mathbf{u} \in \mathbb{R}_+^{m \times n \times K}$ and $\mathbf{g} \in \mathbb{R}_+^{m \times n \times K}$ to obtain the *Lagrangean subproblem*

(LSP)

LSP($\hat{\mathbf{u}}, \hat{\mathbf{g}}$):

$$\min \sum_{k=1}^{K} mu_k + \sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n}(\hat{u}_{ijk}(a_{ijk} - U^{mu}z_{ijk}) + \hat{g}_{ijk}(-U^{mu} - a_{ijk} + mu_k + U^{mu}z_{ijk})) \tag{5.37}$$

s.t.

$$(4.1) - (4.13), (4.15), (4.18) - (4.28).$$

It defines a valid dual bound on VMATP-1 for given $\hat{\mathbf{u}}$ and $\hat{\mathbf{g}}$ vectors. In general the best dual bound is obtained by solving the *Lagrangean dual problem* (LD): $\max_{\mathbf{u},\mathbf{g}\geq 0} \text{LSP}(\mathbf{u}, \mathbf{g})$. LD is a max-min problem and one of the most popular method to solve this problem is the subgradient algorithm [82], in which at each iteration dual multipliers $\mathbf{u}$ and $\mathbf{g}$ are updated and the resulting LSP($\hat{\mathbf{u}}, \hat{\mathbf{g}}$) problem is solved. According to our preliminary analysis LSP($\mathbf{0}, \mathbf{0}$) provides very strong lower bounds. Therefore, we just solve VMATP-1 after relaxing constraints (4.16) and (4.17), which is clearly equivalent to LSP($\mathbf{0}, \mathbf{0}$), and use the optimal value as a LB. We note that in this case, it is possible to also remove geometry constraints (4.1)-(4.10) from LSP($\mathbf{0}, \mathbf{0}$) problem since they do not have any contribution to the objective function. As a result, we obtain the following relaxation of VMATP-1:

RVMATP:

$$\min \sum_{k=1}^{K} mu_k \tag{4.29}$$

s.t.

$$(4.11) - (4.13), (4.15), (4.18) - (4.28).$$

Note that RVMATP is an LP model. As we also discuss in Section 8.3, the LB 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 LB 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-1 (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 optimality cut (5.14), 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 5.1 by appending the steps remaining outside the dotted frames.

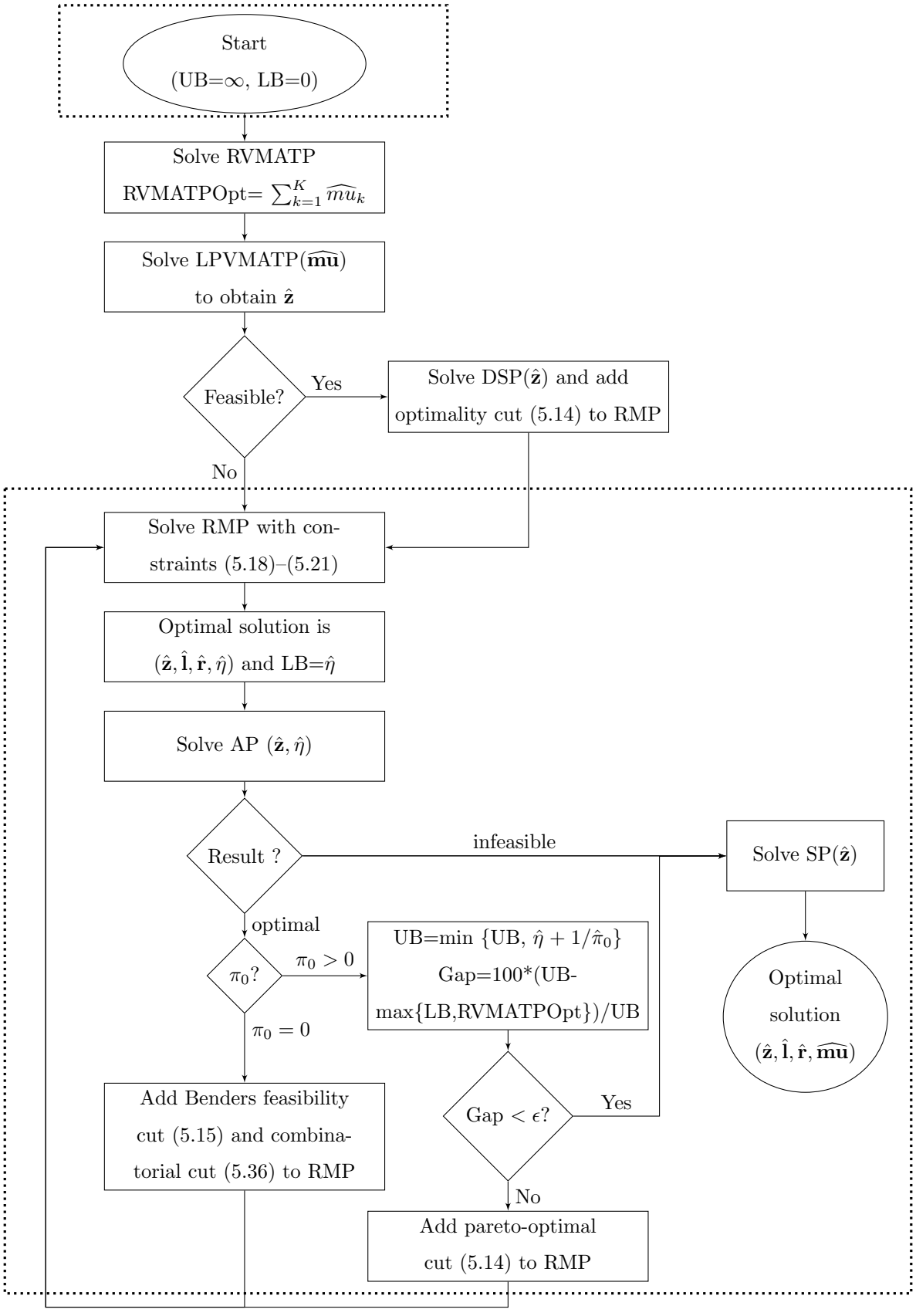

Figure 5.1. Improved Benders decomposition algorithms.

6. SOLUTION METHODS: BRANCH-AND-PRICE ALGORITHMS[3]

In this chapter we develop branch-and-price (BP) algorithms to solve VMATP-1 model. BP is the adaptation of column generation for the exact solution of integer programming problems. At each node of the branch-and-bound tree, column generation is used to solve linear programming relaxation of the reformulation. It is successfully applied to different integer programming problems such as routing, scheduling, and set partitioning problems. Efficiency of the method depends heavily on the problem structure and it is implementation dependent. There is a number of algorithmic issues that occur during implementation, and the proposed algorithms for solving these issues require problem specific solution approaches. In Lübbecke [84] a general framework of the method and common algorithmic issues that practitioners may encounter are explained in detail. In addition, Vanderbeck [85] and Desaulniers [86] present a number of different types of problems that BP methods have been applied. To the best of our knowledge, we apply BP method to solve VMAT planning problem for the first time. In the following sections we provide implementation details of the method as well as solution approaches for the problems that we encounter.

6.1. Column Generation Formulations

Optimal solution of VMATP-1 model yields a feasible VMAT plan with minimum MUs consisting of a feasible treatment arc (i.e. K sequential apertures, each for one of the control points, satisfying the consecutive ones property and leaf motion limitations) and radiation intensity mu_k delivered to the patient body through the corresponding aperture at control point k. Figure 6.1 illustrates a treatment arc consisting of only three equally spaced control points.

[3]An earlier version of this chapter appears in [83].

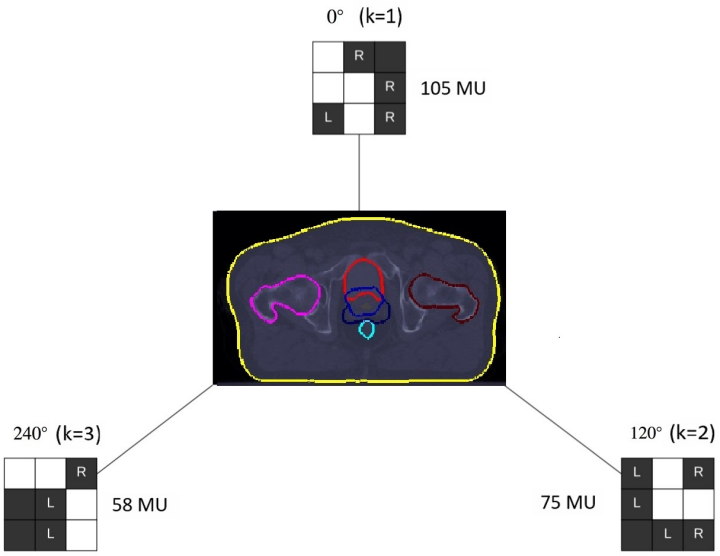

Figure 6.1. A treatment arc consisting of 3 control points, 3 rows and 3 columns.

We observe that it is possible to reformulate VMATP-1 in such a way that each feasible treatment arc is considered as a column. Let $Z = \{\mathbf{z}^1, \mathbf{z}^2, \ldots, \mathbf{z}^e, \ldots, \mathbf{z}^{|Z|}\}$ be the bounded set of all feasible *treatment arcs* (i.e. $\mathbf{z}^e = \{\hat{z}^e_{ijk} \in \{0,1\},\ i = 1, \ldots, m; j = 1, \ldots, n; k = 1, \ldots, K\}$). When we only consider the consecutive ones property and if we assume that there are no leaf motion limitations between consecutive control points, then total number of feasible treatment arcs $|Z|$ equals to $\left(\frac{1}{2}(n+1)(n+2)\right)^{Km}$, which is very large. Row i of the MLC system must satisfy consecutive ones property at control point k and k+1, and also must satisfy the maximum leaf motion limitations. Thus, row i at control points k and k+1 are dependent. However, there is no dependency between row i and other rows at any control point k. As a result, the rows of a treatment arc are independent. We can decompose a treatment arc into m *treatment row arcs*, and it is possible to consider each feasible treatment row arc for each row i as a column. Let $Z_i = \{\mathbf{z}^1_i, \mathbf{z}^2_i, \ldots, \mathbf{z}^e_i, \ldots, \mathbf{z}^{|Z_i|}_i\}$ be the bounded set of all feasible treatment row arcs for row i satisfying consecutive ones property and leaf motion limitations ($\mathbf{z}^e_i = \{\hat{z}^e_{ijk} \in \{0,1\},\ j = 1, \ldots, n; k = 1, \ldots, K\}$). As known from integer programming theory it is possible to express a bounded set $Z_i = \{\mathbf{z}_i \in \{0,1\}^{nK}, \mathbf{A} \in \mathbb{R}^{p \times nK}, \mathbf{h} \in \mathbb{R}^p : \mathbf{Az} \leq \mathbf{h}\}$ equivalently as $\{\mathbf{b}_i \in \{0,1\}^{|Z_i|} : \sum_{e=1}^{|Z_i|} b^e_i \mathbf{z}^e_i,\ \sum_{e=1}^{|Z_i|} b^e_i = 1\}$, where $\mathbf{z}^e_i \in \{0,1\}^{nK}$ $e = 1, \ldots |Z_i|$ are the feasible solutions of Z_i. Here, binary variable b^e_i indicates whether the feasible row arc $\mathbf{z}^e_i$ is selected ($b^e_i = 1$) or not ($b^e_i = 0$). Thus, it is possible to

represent binary variable z_{ijk} as $\sum_{e=1}^{|Z_i|} b_i^e \hat{z}_{ijk}^e$. Then, the resulting treatment row arc based reformulation of VMATP-1, in other words, the master problem (MP) can be written as

MP:

$$\min \sum_{k=1}^{K} mu_k \tag{4.29}$$

s.t.

$(4.11) - (4.13), (4.15), (4.18) - (4.28),$

$$\sum_{e=1}^{|Z_i|} b_i^e = 1 \qquad i = 1, \ldots, m \tag{6.1}$$

$$-a_{ijk} + U^{mu} \sum_{e=1}^{|Z_i|} b_i^e \hat{z}_{ijk}^e \geq 0 \qquad i = 1, \ldots, m; j = 1, \ldots, n; k = 1, \ldots, K \tag{6.2}$$

$$a_{ijk} - mu_k - U^{mu} \sum_{e=1}^{|Z_i|} b_i^e \hat{z}_{ijk}^e \geq -U^{mu} \qquad i = 1, \ldots, m; j = 1, \ldots, n; k = 1, \ldots, K \tag{6.3}$$

$$\mathbf{b} \in \{0,1\}^{\sum_{i=1}^{m} |Z_i|}, \tag{6.4}$$

where the convexity constraints (6.1) ensure that exactly one feasible treatment row arc is selected for each row i. Note that constraints (4.16) and (4.17) are replaced with constraints (6.2) and (6.3) as explained above in detail. Therefore, constraints (6.2), (6.3) and (4.18) guarantee that the radiation intensity of each beamlet a_{ijk} equals mu_k when this beamlet is open, and 0 if it is closed. We solve the linear programming relaxation of MP (MLP) by column generation. Moreover, we introduce one nonnegative artificial variable (ϕ_{tc}^{TV} or ϕ_{oc}^{OAR}) for each one of the constraints (4.21) and (4.26) to allow deviations. We penalize positive deviations in the objective function. Then the resulting modified master linear problem, which we continue to call as MLP, becomes

MLP:

$$\min \sum_{k=1}^{K} mu_k + \sum_{t=1}^{T}\sum_{c=1}^{C_t} \gamma_{tc}^{TV}\phi_{tc}^{TV} + \sum_{o=1}^{O}\sum_{c=1}^{C_o} \gamma_{oc}^{OAR}\phi_{oc}^{OAR} \tag{6.5}$$

s.t.

$$(4.11)-(4.13),(4.15),(4.18)-(4.20),$$

$$(4.22)-(4.25),(4.27)-(4.28),(6.2)-(6.3),$$

$$\sum_{e=1}^{|Z_i|} b_i^e = 1 \qquad i = 1,\dots m \qquad (\lambda_i) \tag{6.1}$$

$$\xi_{tc}^{TV} - \frac{1}{(1-\alpha_{tc}^{TV})|V_t^{TV}|}\sum_{v\in V_t^{TV}} x_{tcv} + \phi_{tc}^{TV} \geq \overline{d}_{tc} \qquad t = 1,\dots T;$$

$$c = 1,\dots,C_t \tag{6.6}$$

$$\xi_{oc}^{OAR} + \frac{1}{(1-\alpha_{oc}^{OAR})|V_o^{OAR}|}\sum_{v\in V_o^{OAR}} y_{ocv} - \phi_{oc}^{OAR} \leq U_{oc}^{OAR} \quad o = 1,\dots O;$$

$$c = 1,\dots C_o \tag{6.7}$$

$$\phi^{\mathbf{TV}} \in \mathbb{R}_+^{\sum_{t=1}^{T} C_t}, \tag{6.8}$$

$$\phi^{\mathbf{OAR}} \in \mathbb{R}_+^{\sum_{o=1}^{O} C_o}, \tag{6.9}$$

$$\mathbf{b} \in \mathbb{R}_+^{\sum_{i=1}^{m} |Z_i|}, \tag{6.10}$$

where γ_{tc}^{TV} and γ_{oc}^{OAR} are large penalty costs for deviations in the cth partial volume constraints of TVs t and OAR o, respectively.

We use same dual variables $\boldsymbol{\mu^1}, \boldsymbol{\mu^2}, \boldsymbol{\pi}, \boldsymbol{\beta^3}, \boldsymbol{\tau^1}, \boldsymbol{\epsilon^1}, \boldsymbol{\epsilon^2}, \boldsymbol{\tau^2}, \boldsymbol{\beta^1}, \boldsymbol{\beta^2}, \boldsymbol{\theta^1}, \boldsymbol{\theta^2}$ for the constraints (4.11), (4.12), (4.15), (4.18), (4.22), (4.24), (4.25), (4.27), (6.2), (6.3), (6.6), (6.7), respectively, which are also used to formulate DSP model in Section 5.1. Also, dual variables $\boldsymbol{\lambda}$ are used for constraint (6.1) to obtain following dual MLP (DMLP)

DMLP:

$$\max \sum_{i=1}^{m} \lambda_i - U^{mu} \sum_{k=1}^{K} \sum_{i=1}^{m} \sum_{j=1}^{n} \beta_{ijk}^2 + \sum_{t=1}^{T} \sum_{c=1}^{C_t} \theta_{tc}^1 \overline{d}_{tc} + \sum_{t=1}^{T} \sum_{v \in V_t^{TV}} (L_t^{TV} \epsilon_{tv}^1 - U_t^{TV} \epsilon_{tv}^2)$$
$$- \sum_{o=1}^{O} \sum_{c=1}^{C_o} \theta_{oc}^2 U_{oc}^{OAR} + \sum_{k=1}^{K} (L^{mu} \mu_k^1 - U^{mu} \mu_k^2) \tag{6.11}$$

s.t.

$$\lambda_i + U^{mu} \sum_{j=1}^{n} \sum_{k=1}^{K} (\beta_{ijk}^1 - \beta_{ijk}^2) \hat{z}_{ijk}^e \leq 0 \quad i = 1, \dots m; e = 1, ..., |Z_i| \quad (b_i^e) \tag{6.12}$$

$$\theta_{tc}^1 \leq \gamma_{tc}^{TV} \quad t = 1, \dots T; c = 1, \dots, C_t \tag{6.13}$$

$$\theta_{oc}^2 \leq \gamma_{oc}^{OAR} \quad o = 1, \dots O; c = 1, \dots, C_o \tag{6.14}$$

$$\lambda \in \mathbb{R}^m \tag{6.15}$$

$(5.4) - (5.12)$.

Note that if we do not consider leaf motion limitations then total number of feasible treatment row arcs for each row is $\left(\frac{1}{2}(n+1)(n+2)\right)^K$. Hence, total number of feasible treatment row arcs is $m\left(\frac{1}{2}(n+1)(n+2)\right)^K$, which is still very large, and the reformulated problem is not tractable due to the exponential number of columns. We can solve MLP by column generation starting with a restricted MLP (RMLP) model, which includes a subset of feasible row arcs Z_i^0 for each row. We iteratively search for new promising row arcs (columns for RMLP) by solving m pricing subproblems (PSPs). Then, we change the new columns with negative reduced cost with the current ones.

RMLP:

$$\min \sum_{k=1}^{K} mu_k + \sum_{t=1}^{T} \sum_{c=1}^{C_t} \gamma_{tc}^{TV} \phi_{tc}^{TV} + \sum_{o=1}^{O} \sum_{c=1}^{C_o} \gamma_{oc}^{OAR} \phi_{oc}^{OAR} \tag{6.5}$$

s.t.

$(4.11) - (4.13), (4.15), (4.18) - (4.20),$

$(4.22) - (4.25), (4.27) - (4.28), (6.6) - (6.9),$

$$\sum_{e=1}^{|Z_i^0|} b_i^e = 1 \qquad i = 1, \ldots, m \tag{6.16}$$

$$-a_{ijk} + U^{mu} \sum_{e=1}^{|Z_i^0|} b_i^e \hat{z}_{ijk}^e \geq 0 \qquad i = 1, \ldots, m; j = 1, \ldots, n; k = 1, \ldots, K \tag{6.17}$$

$$a_{ijk} - mu_k - U^{mu} \sum_{e=1}^{|Z_i^0|} b_i^e \hat{z}_{ijk}^e \geq -U^{mu} \quad i = 1, \ldots, m; j = 1, \ldots, n; k = 1, \ldots, K \tag{6.18}$$

$$\mathbf{b} \in \mathbb{R}_+^{\sum_{i=1}^m |Z_i^0|}. \tag{6.19}$$

Let $\hat{\lambda}_i$, $\hat{\beta}_{ijk}^1$ and $\hat{\beta}_{ijk}^2$ be an optimal dual solution associated with constraints (6.16)–(6.18). Then we have m subproblems (pricing subproblems (PSPs)) one for each row:

PSP_i:

$$\min \; -\hat{\lambda}_i - U^{mu} \sum_{k=1}^{K} \sum_{j=1}^{n} (\hat{\beta}_{ijk}^1 - \hat{\beta}_{ijk}^2) z_{ijk} \tag{6.20}$$

s.t.

$$r_{ik} - j z_{ijk} \geq 1 \qquad j = 1, \ldots, n; \;\; k = 1, \ldots, K \tag{6.21}$$

$$(n + 1 - j) z_{ijk} + l_{ik} \leq n \qquad j = 1, \ldots, n; \;\; k = 1, \ldots, K \tag{6.22}$$

$$r_{ik} - l_{ik} - \sum_{j=1}^{n} z_{ijk} = 1 \qquad k = 1, \ldots, K \tag{6.23}$$

$$l_{i(k+1)} - l_{ik} \leq \delta \qquad k = 1, \ldots, K-1 \tag{6.24}$$

$$l_{ik} - l_{i(k+1)} \leq \delta \qquad k = 1, \ldots, K-1 \tag{6.25}$$

$$r_{i(k+1)} - r_{ik} \leq \delta \qquad k = 1, \ldots, K-1 \tag{6.26}$$

$$r_{ik} - r_{i(k+1)} \leq \delta \qquad k = 1, \ldots, K-1 \tag{6.27}$$

$$\mathbf{l} \in \mathbb{Z}_+^K; \;\; \mathbf{r} \in \mathbb{Z}_+^K; \;\; \mathbf{z} \in \{0,1\}^{n \times K}. \tag{6.28}$$

Note that Constraint (6.21)–Constraint (6.28) are similar to Constraint (4.2)–Constraint (4.10) in Chapter 4, and generate a feasible treatment row arc for row i. Note also that there is not a constraint similar to Constraint (4.1), which is omitted because of being

redundant. Observe that all feasible treatment row arcs for each row i must satisfy constraint (6.12) in DMLP. When we start with RMLP that includes a subset of treatment row arcs (columns) for each row i, then the corresponding dual problem includes only the constraints corresponding to these columns. The remaining constraints associated with the columns which are not generated are relaxed at an iteration. The objective function (6.20) checks whether there is a violated constraint of row i and finds the one with the maximum violation. In our algorithm, we solve the modified version of the master problem (MP), which includes the new decision variables to allow deviation in CVaR constraints, by BP method. At each node of the branch-and-bound tree column generation is used to solve MLP. Each time after solving RMLP, we solve m PSPs separately and introduce a new treatment row arc of row i to RMLP only if its reduced cost is negative. On the other hand, if optimal objective value of each $\mathrm{PSP_i}$ yields nonnegative reduced cost, then the column generation iterations stop. Simply, we are at an optimal solution of the MLP. If all b_i^e variables are integer at the optimality (and also deviations are zero) then we have also an optimal solution of MP and VMATP-1. However, if at least one of them is fractional then we continue with branching and solve the modified restricted models by column generation at the new branches.

6.2. Generating Columns by Solving Shortest Path Problems

We observe in our preliminary experiments that solving PSPs by using a commercial MIP solver is inefficient. The variation of computation time between iterations is high, and it may take too long to generate a column at some iterations. Thus we formulate the pricing subproblems as shortest path problems similar to [56], [60] and [30]. We explain the shortest path problem formulation of PSP on a small example. Figure 6.2 illustrates the network representation of $\mathrm{PSP_1}$ for the first row of the problem given in Figure 6.1. There are only three control points ($K = 3$) and three beamlets in a row ($n = 3$). Note that the home positions of the leaves are $j = 0$ and $j = 4$ for the left and right leaves, respectively. The beamlets that are blocked by the leaves are dark gray, and open beamlets are shown as white rectangles. There are two additional nodes in the figure: start and finish nodes. For each one of the control points there are

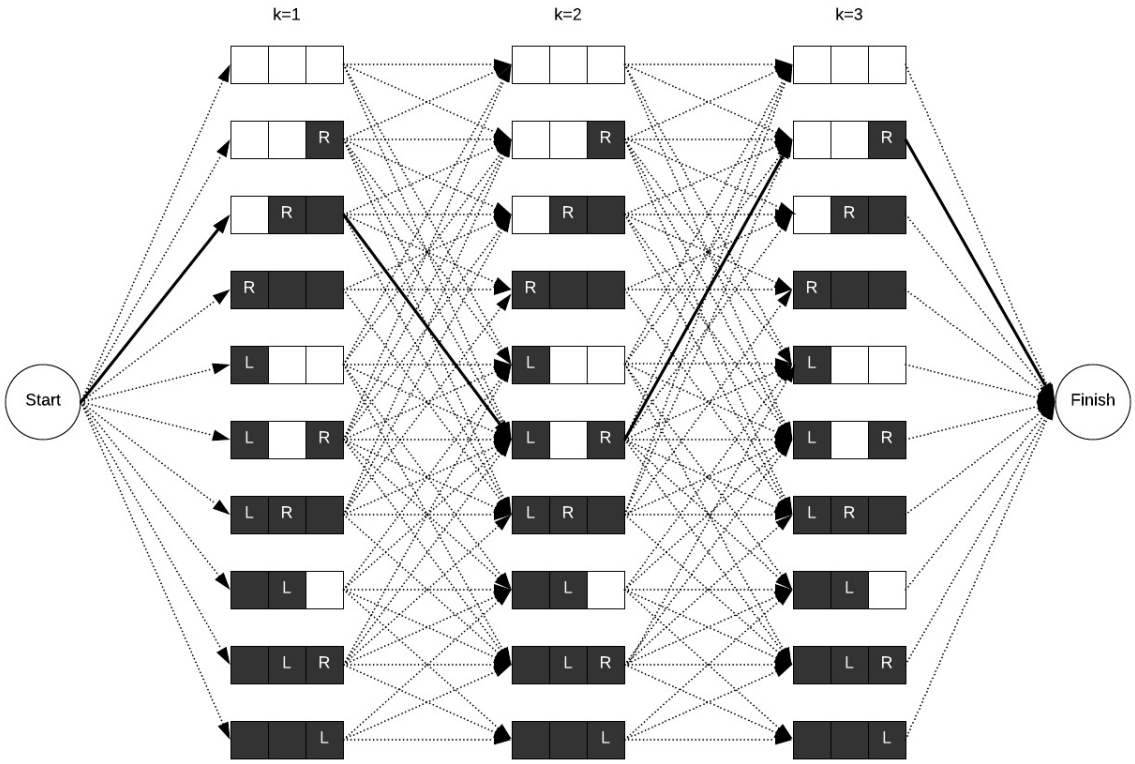

Figure 6.2. Network representation of PSP_1 for the first row of the treatment arc given in Figure 6.1 ($K = 3$, $n = 3$).

$\frac{(n+1)(n+2)}{2} = 10$ different leaf configurations. For example, the leaf configurations at the top of the figure represent that both leaves are at their home positions and all beamlets are open. Observe that at a control point there are four different combinations of the leaves for closing all beamlets, since the left and the right leaves may be adjacent in four different ways. Moreover, it is assumed that the leaves can move at most one beamlet between consecutive control points (i.e. $\delta = 1$). The leaf configurations at each one of the control points represent the nodes. Also, an arc between two nodes at two adjacent control points indicate that the maximum leaf movement limitations are satisfied, namely these two consecutive leaf configurations are compatible. If the new position of each one of the left and right leaves at the next control point is in its allowable range then there is an arc. Thus, the arcs in the graph represent feasible movements. Observe that the direction of the arcs point the rotation direction of the

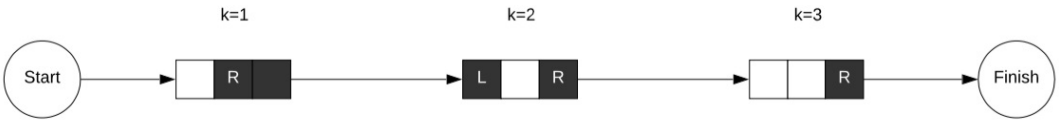

Figure 6.3. The treatment row arc obtained in Figure 6.2.

gantry and there are no arcs between any two nodes at the same control point. The costs of an arc that connects two nodes at control points k and $k+1$ is computed as $-U^{mu}\sum_{j=l+1}^{j=r-1}(\hat{\beta}^1_{ijk}-\hat{\beta}^2_{ijk})$. The cost of an arc that connects a node at the last control point K and the finish node is $-U^{mu}\sum_{j=l+1}^{j=r-1}(\hat{\beta}^1_{ijK}-\hat{\beta}^2_{ijK})$, and the cost of an arc between the start node and a node at the first control point is zero. Our aim is to find a path from the start node to the finish node with minimum cost. To sum up, we obtain a directed, acyclic and layered graph consisting of K layers which correspond to K control points. In each layer there are $\frac{(n+1)(n+2)}{2}$ nodes. If we assume that there are no leaf motion limitations between adjacent control points then the total number of arcs in the graph will be $\frac{(K-1)}{4}(n+1)^2(n+2)^2+(n+1)(n+2)$. We solve this problem using dynamic programming and the UB for the complexity of the algorithm is $O(Kn^4)$, which is a polynomial. The optimal solution of this problem yields one of the treatment row arcs with minimum reduced cost and if its reduced cost is negative, then we add this resulting column to RMLP. For example, the shortest path given in Figure 6.2 with solid lines indicates the treatment row arc illustrated in Figure 6.3. At the first control point the left leaf is at its home position and the right leaf blocks the second and third beamlets. Only the first beamlet is open. During the movement of the gantry from the first control point to the second one, the left leaf moves one beamlet to the right and blocks the first beamlet. On the other hand, the right leaf moves one beamlet to the right and opens the second beamlet. Finally, during the travel of the gantry from the second control point to the last one, the left leaf returns to its home position and stops blocking the first beamlet. However, the right beamlet does not move, and only the last beamlet is blocked. The reduced cost of this shortest path is equal to $-\hat{\lambda_1}-U^{mu}\left((\hat{\beta}^1_{111}-\hat{\beta}^2_{111})+(\hat{\beta}^1_{122}-\hat{\beta}^2_{122})+(\hat{\beta}^1_{113}-\hat{\beta}^2_{113}+\hat{\beta}^1_{123}-\hat{\beta}^2_{123})\right)$.

The study of Boland *et al.* [30] is one of the leading papers that uses network models in radiation therapy planning. They use their network model in their column generation approach to solve MLS problem, which is the last phase of IMRT planning. They decompose a given fluence map into a number of feasible apertures with radiation intensities. They design their network in such a way that each layer corresponds to a leaf pair. Therefore, there are as many layers as the number of rows of the fluence map. At each layer the nodes represent the potential positions of the left and right leaves satisfying the consecutive ones property. There are arcs between two nodes at two adjacent layers if the leaf configurations at these rows satisfy interdigitation constraints. Hence, a path in the network corresponds to an aperture, which is feasible with respect to the consecutive ones property and interdigitation constraints. Mahnam *et al.* [56] use a similar approach to generate a set of sequential apertures in VMAT planning. They consider that a full treatment arc (i.e. a 360°-arc) consists of a number of sequential partial arcs with the same length (i.e. there are 18 20°-arcs in a full arc). Also, each of these partial arcs includes a number of equally spaced apertures (i.e. 10 apertures with 2°-spacing at a 20°-arc). They can generate a partial arc, row by row, using a network model since they consider only the consecutive ones property. In their network model, the number of layers equals to the number of apertures in a partial arc and the nodes represents the leaf configurations, which is similar to the model of Boland *et al.* [30]. On the other hand, there is an arc between two nodes at two consecutive layers if the corresponding leaf configurations satisfy leaf motion limitations. After solving m subproblems by a shortest path algorithm, they take the union of the resulting partial row arcs to obtain a partial arc. Also, they need to join a number of partial arcs to obtain a full treatment arc, which necessitates a post-optimization (i.e. the intersection points of adjacent partial arcs may be incompatible due to leaf motion limitations). In our study, pricing subproblem generates a full row arc, which is feasible with respect to leaf motion limitations. Hence, we do not need any post-processing operation. Another difference is that the union of m row arcs yields a feasible full treatment arc. Finally, our network design is similar to the one proposed in Gözbaşı [60], which is used to solve VMAT planning problem in a two-stage heuristic approach. They generate a feasible full treatment arc in the first stage where the costs

of the arcs are calculated using a beamlet scoring algorithm. In the second stage, they find radiation intensity of each one of the apertures in the treatment arc.

6.3. Branching

At the root node of the branch-and-bound tree if the optimal values of all b_i^e variables are integral then we are at an optimal solution of the problem. Otherwise, we apply branching and solve the resulting restricted linear programming model at each one of the branch-and-bound tree nodes. It is important to find a branching strategy that prevents regenerating columns that are previously prohibited. Also, the columns generated so far must be divided into two groups and it must be possible to modify the PSP so that generating infeasible columns due to the branching constraints is prevented. It is known that applying the ordinary variable branching (dichotomized branching on a b_i^e variable with fractional value) is not efficient [87]. Instead, we branch on the original variables of VMATP-1; a beamlet with a fractional $\hat{z}_{ijk}$ value. Observe that if the optimal solution of MLP at a node is not integer then for at least one row i there must be at least two fractional b_i^e variables. Thus, there must be at least one beamlet with fractional $\hat{z}_{ijk}$ value. Observe that when there are two fractional variables b_i^1 and b_i^2 in the current solution with columns $\mathbf{z}_i^1$ and $\mathbf{z}_i^2$, respectively, then there must be at least one beamlet (i, j, k) having value 1 in exactly one of these columns. As a result, $\hat{z}_{ijk} = b_i^1 \hat{z}_{ijk}^1 + b_i^2 \hat{z}_{ijk}^2$ becomes fractional. In our branching rule we choose one of these fractional beamlets as the branching variable using a simple search mechanism. For each control point k we first calculate the following ratio: $\frac{\sum_{(i,j)} \hat{z}_{ijk}}{\sum_{\substack{(i,j) \\ \hat{z}_{ijk}>0}} 1}$. Observe that if there is at least one fractional beamlet, then this ratio is strictly between 0 and 1, and the corresponding aperture becomes fractional. Note that if all beamlets of an aperture at a control point take only 0 or 1 value, then we do not consider this control point. For each control point k with fractional aperture, we calculate the value $\Upsilon_k = \widehat{mu}_k \left(\frac{\sum_{(i,j)} \hat{z}_{ijk}}{\sum_{\substack{(i,j) \\ \hat{z}_{ijk}>0}} 1} \right)$. Then we select the one with the highest Υ_k value and branch on the beamlet at this control point that is fractional and closest to 1. Thus, we seek a beamlet belonging to an aperture having high radiation intensity and low fractionality. Let us denote the selected beamlet by z_{ijk}. We then obtain two child nodes: at one

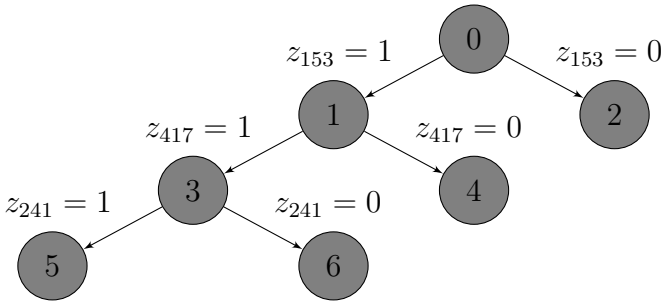

Figure 6.4. Branching rule.

of them beamlet (i, j, k) is open and at the other one it is closed. We illustrate the branching rule in Figure 6.4.

In the first child node, we remove the respective arcs in the PSP_i's network model that connect the nodes at control point $k-1$ to the nodes at control point k where the beamlet (i, j, k) is closed. Also, we set b_i^e variables to 0 if the value of beamlet (i, j, k) is 0 in the corresponding column. In the second child node, we remove all arcs that connect the nodes at control point $k-1$ to the nodes at control point k where the beamlet (i, j, k) is open. Also we set all b_i^e variables to 0 if beamlet (i, j, k) takes value 1 in the corresponding columns. Note that, at each branching we use only one beamlet (i, j, k) that belongs to one row (ith row), thus we partition only the columns associated with row i. Furthermore, at each node the PSPs are modified taking into account all branching decisions leading to the current node.

6.4. Initial Set of Columns

We generate m initial columns at the beginning of the BP algorithm by solving the following model

$$\min \sum_{k=1}^{K} m u_k \tag{4.29}$$

s.t.

$$(4.2) - (4.13), (4.16), (4.18), (4.20),$$

$$\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ijk} D_{ijkv} \geq L_t^{TV} \qquad t = 1, \ldots, T; \;\; v \in V_t^{TV}, \tag{5.20}$$

which includes all geometry constraints and valid inequalities (4.16) and (4.18). We observe that if a treatment arc consists of K apertures unable to deliver enough radiation to satisfy lower limits of voxels in each TV, then this treatment arc cannot yield a feasible solution for VMATP-1. In other words, a treatment arc must satisfy that each target voxel absorbs radiation at least the prescribed lower limit (i.e. L_t^{TV}). Thus, we generate the initial set of m columns by considering this observation and avoiding such infeasible treatment arcs. Optimal solution of the model given above yields K aperture shapes, namely all z_{ijk} values. Hence, for each row i we obtain a treatment row arc and use them to construct initial RMLP model. The initial treatment arc can yield a feasible solution for VMATP-1, namely the resulting RMLP may provide a solution that satisfies all prescription radiation doses. As a result, we obtain an UB since we start with only one column for each row and all b_i^1 values are 1 ($i = 1, ..., m$). However, this is not always the case, and to resolve this problem we add one artificial variable (ϕ_{tc}^{TV}, ϕ_{oc}^{OAR}) for each one of the constraints (6.6) and (6.7), and penalize them in the objective function. We keep these artificial variables at all nodes of the branch-and-bound tree and the infeasibility of RMLP is almost completely removed. Note that there are also full volume constraints in RMLP associated with target voxels (constraints (4.24) and (4.25)). The initial columns generated by the formulation given above guarantee that constraints (4.24) are satisfied at root node. However, RMLP may not satisfy constraints (4.25), or at successor nodes it may be infeasible due to the branching constraints that cause absence of relevant columns to satisfy constraints (4.24). This case is rarely encountered, yet we resort to Farkas Pricing as explained by Lübbecke [84] in detail. Also, the details of initialization using artificial variables is explained by Vanderbeck [85].

6.5. Lower Bounds

Column generation methods often suffer from the *tailing off effect*: at initial iterations a near optimal solution is reached quickly but in the following iterations the improvement in the objective value becomes very small and the algorithm terminates in very long time [88]. We also experience this effect; it takes very long to prove

optimality and terminate column generation iterations at each node of the branch-and-bound tree. As a remedy, we adapt and solve at each node the RVMATP model given in Section 5.2.5, which is a relaxation of VMATP-1 model and does not include geometry constraints. If it is possible, we update the LB of the current search node. Note that, since we branch on z_{ijk} variables and in each branch we set one of them to 0 or 1, it is possible to adjust RVMATP by setting each one of the a_{ijk} variables associated with the branching decisions that leads to the current node to either mu_k or 0. Moreover, during the column generation iterations, we update the LB if the optimal value of RMLP and sum of the optimal values of all PSPs is larger than the current LB. Finally, we use depth-first search as node selection strategy. The flow of the resulting algorithm, which we call Branch-and-price (BP) Algorithm 1, is given in Figure 6.5.

6.6. Algorithmic Improvements

We modify BP Algorithm 1 and obtain two enhancements, which we call BP Algorithm 2 and BP Algorithm 3. In BP Algorithm 2, at root node before branching, we solve the resulting restricted MP including columns generated so far as a MILP model. We update UB if the resulting solution is better than the incumbent.

According to the preliminary experiments we observe that the necessary time to solve the model given in Section 6.4, which generates the initial columns, increases as the size of the problem becomes larger. Moreover, it becomes impossible to solve it optimally within the given time limit. Thus, we simplify this part of the algorithm in BP Algorithm 3. The initial columns are generated by solving a different model that consists of only the geometry constraints and a different objective function:

$$\max \sum_{k=1}^{K} \sum_{i=1}^{m} \sum_{j=1}^{n} z_{ijk} \tag{6.29}$$

s.t.

$(4.2) - (4.10)$.

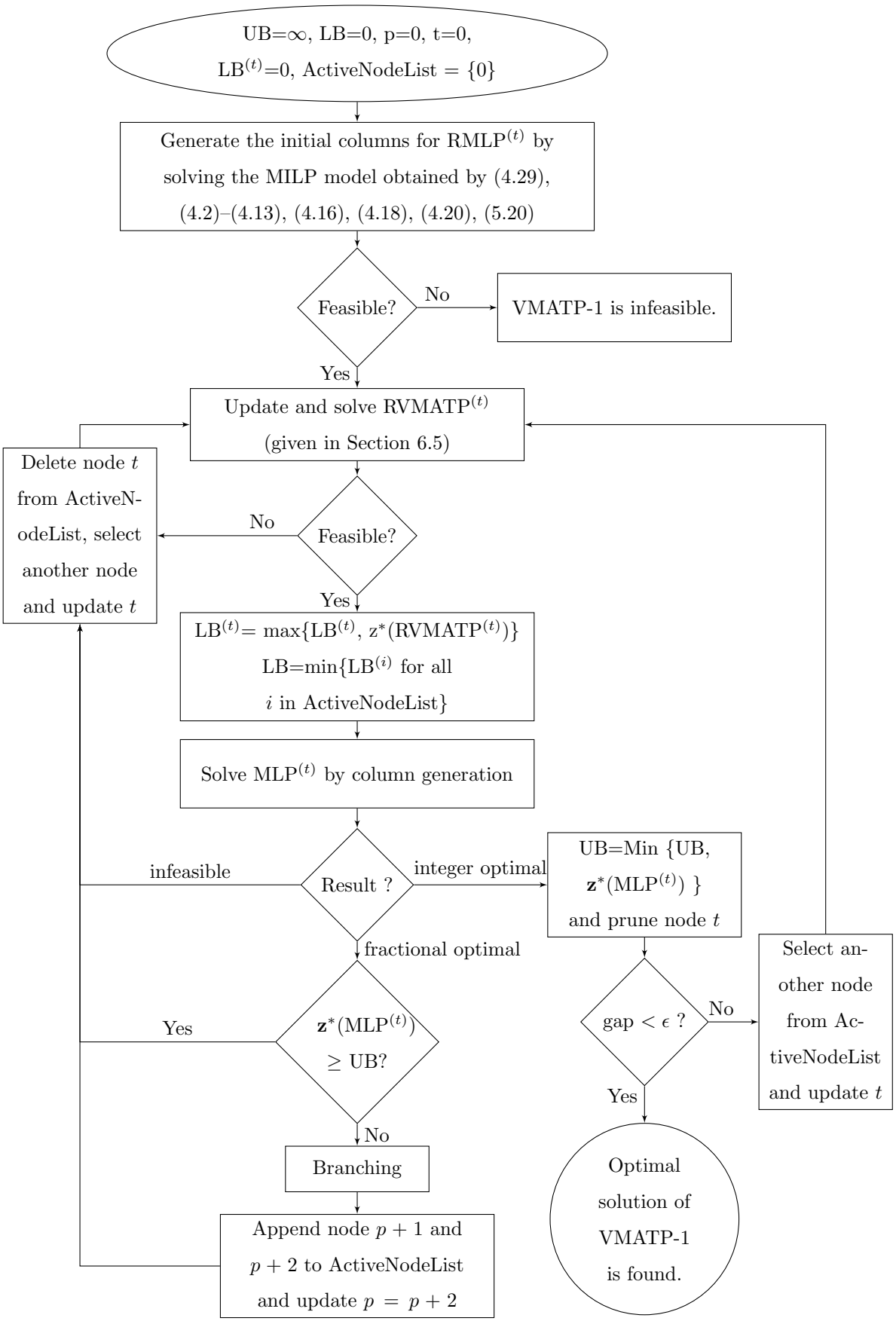

Figure 6.5. Branch-and-Price Algorithm 1.

An optimal solution of this model gives a full treatment arc with maximum total number of open beamlets that satisfies geometry constraints. Similar to BP Algorithm 2, a better feasible solution for VMATP-1 is sought at the root node. Instead of generating all promising columns before branching, each subproblem is solved only once and the resulting restricted MP is solved as a MILP model.

7. TWO-PHASE HEURISTIC [4]

VMATP-1 model optimizes simultaneously aperture shapes and radiation intensities at control points. It includes CVaR constraints for partial volume restrictions of all structures, and all treatment dose prescriptions are satisfied by the constraints as a result. In short, VMATP-1 seeks an optimal VMAT plan with minimum MUs, which is capable to handle many aspects of the complex decision process behind VMAT planning. However, this makes it computationally very difficult to solve exactly in order to generate optimal plans for realistic clinical cases with many OARs. In this chapter we try to address this issue and propose an efficient two-phase heuristic using the algorithmic ideas, such as column generation, we employ in Chapter 6 for the development of a BP exact solution algorithm.

In the first phase, we generate an initial full treatment arc using a two-step approach and calibrate the right hand side values of the CVaR constraints, simultaneously. In the second phase, we improve the initial treatment plan obtained in the first phase using column generation that we explain all necessary derivations in Chapter 6 such as the reformulation of VMATP-1, and pricing subproblem (PSP) formulation and solution by a shortest path algorithm. We test our heuristic on real prostate cancer patient cases provided by Istanbul University Oncology Institute. The results of the computational experiments and clinical comparisons are provided in Section 8.5.

7.1. Phase 1: Initial Column Generation

At the beginning there is exactly one column (one row arc) for each row i in the initial column pool to formulate RMLP. Their union yields a full treatment arc consisting of K apertures (one aperture per control point). We apply a simple heuristic consisting of two steps to generate these initial columns. At the first step a number of fluence maps with additional properties are generated by solving a linear programming model, and in the second step using a simple fluence map conversion algorithm a full

[4] An earlier version of this chapter is under revision at Physics in Medicine and Biology.

treatment arc, which yields a column per each row, is constructed. These two steps constitute the first phase of our two-phase heuristic. This initial full arc is modified in the second phase during the column generation iterations in order to obtain a better treatment plan, so it is very important to start with good columns.

7.1.1. Step 1: Fluence Map Generation

The LP model that we solve in the first step is derived from VMATP-1 and includes only a subset of the original control points and constraints. Note that, as in this study, K is generally taken 180 in VMAT planning studies (i.e. 180 equally spaced beam angles with 2°-spacing). The LP model, namely the modified VMATP-1 (M-VMATP) includes 45 control points with 8°-spacing. We let $\overline{K}$ denote this subset of control points. We introduce one artificial variable for each CVaR constraint of all OARs to M-VMATP, and penalize the positive deviations in the objective function. We allow these deviations not only because M-VMATP includes only a subset of control points but also because at the beginning CVaR constraints consisting of the original tolerance doses are very tight. Thus, finding a treatment plan which satisfies all treatment dose prescriptions is not easy. M-VMATP is formulated as follows:

M-VMATP:

$$\min \sum_{k\in\overline{K}} mu_k + \sum_{o=1}^{O}\sum_{c=1}^{C_o} \gamma_{oc}^{OAR}\phi_{oc}^{OAR} \tag{7.1}$$

s.t.

$(4.19), (4.21)-(4.25),$

$(4.27)-(4.28), (6.7), (6.9)$

$$a_{ijk} \le mu_k \qquad i=1,\dots,m;\ \ j=1,\dots,n;\ \ k\in\overline{K} \tag{7.2}$$

$$d_v - \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k\in\overline{K}} D_{ijkv}a_{ijk} = 0 \qquad v\in V \tag{7.3}$$

$$mu_k \ge L^{mu} \qquad k\in\overline{K} \tag{7.4}$$

$$mu_k \le U^{mu} \qquad k\in\overline{K} \tag{7.5}$$

$$\mathbf{mu} \in \mathbb{R}_{+}^{|\overline{K}|};\ \ \mathbf{a} \in \mathbb{R}_{+}^{m \times n \times |\overline{K}|}. \tag{7.6}$$

Solving this modified model has similarities with FMO in IMRT planning. M-VMATP finds a number of fluence maps for some of the control points in $\overline{K}$, which we denote as $\overline{\overline{K}}$. Notice that $\overline{\overline{K}} \subseteq \overline{K} \subset K$. However, $\overline{\overline{K}}$, the set of control points with positive intensities, is not determined in advance; M-VMATP model finds it. Also, the intensities of the beamlets of a fluence map are bounded from above with radiation intensity at this control point (mu_k). Observe that if there is no beamlet with positive intensity at control point k then mu_k value will be 0, since total radiation intensity is minimized in the objective function. Furthermore, in this step, in addition to generating fluence maps we also tune the parameters U_{oc}^{OAR} of CVaR constraints of OARs to obtain more reasonable feasible treatment plans. In a loop, we increase the tolerance dose U_{oc}^{OAR} by a small value if the difference between the radiation dose that the corresponding volume of OAR o receives and the original tolerance dose is at least a certain amount, and resolve M-VMATP model until there remains no parameter to increase. At the end of this loop, if there is a CVaR constraint with positive deviation we increase its right hand side parameter by the amount of deviation. We use the resulting tuned parameters throughout the entire algorithm.

7.1.2. Tuning of CVaR Constraints

The main advantage of using CVaR constraints is that it allows modeling of dose-volume requirements as linear inequalities. However, we observe that CVaR constraints with original parameters are very conservative and it is challenging to apply this approach in radiation therapy planning, which is also indicated in [24] and [89]. Let us consider a CVaR constraint defined for an OAR. It forces the average dose in the upper tail of the dose distribution of the OAR to be at most its tolerance dose. However, it is sufficient that the left end (i.e. VaR) of the tail does not exceed this tolerance dose [89]. In order to alleviate this problem we tune the right hand side values, i.e. U_{oc}^{OAR} of the CVaR constraints of OARs. We change these parameters in such a way that the resulting ones continue to produce treatment plans satisfying clinical prescrip-

tion doses. In particular, we solve M-VMATP and check whether the VaR values are too small than the corresponding bounds; if they are, we update the right hand sides of the constraints. In each iteration we do this operation for all OARs, after that we resolve M-VMATP. We continue until there is no CVaR constraint that we can update. The pseudo code of this parameter tuning procedure is given in Figure 7.1. Note that it is not appropriate to use the same right hand side values for all patients due to the anatomical differences between them. Thus, it is convenient to use such an adaptive procedure for tuning the right hand side values for each patient.

ϵ_1, ϵ_2, counter $= 0$
while true **do**
  update and solve M-VMATP
  for each o in OARs and c in C_o **do**
    if $\xi_{oc}^{OAR} < U_{oc}^{OAR} - \epsilon_1$ **then**
      $U_{oc}^{OAR} \leftarrow U_{oc}^{OAR} + \epsilon_2$
      counter $\leftarrow$ counter+1
    end if
  end for
  if counter $= 0$ **then**
    break
  else
    counter $= 0$
  end if
end while
for each o in OARs and c in C_o **do**
  if $\phi_{oc}^{OAR} > 0$ **then**
    $U_{oc}^{OAR} \leftarrow U_{oc}^{OAR} + \phi_{oc}^{OAR}$
  end if
end for

Figure 7.1. CVaR parameter tuning.

7.1.3. Step 2: Conversion Algorithm

In the second step we derive a number of apertures (at most three) from each one of the fluence maps obtained in the first step using a conversion algorithm. Clearly, in a fluence map, the beamlets with positive radiation intensity do not have to satisfy the consecutive ones property, also their intensities may differ from each other. We assume that all beamlets with positive intensity are open. Our conversion algorithm seeks at most three feasible apertures to cover the open beamlets as much as possible. If all open beamlets of all rows are consecutive in a fluence map at control point k, then we generate only one aperture and fix it at k. If there are at most two open beamlet chains at each row than we generate two different apertures and sequence them on to control points k and $k+2$. Otherwise, if there are rows with more than two open beamlet chains than we generate three different apertures and sequence them on to control points on to k, $k+1$, and $k+2$. There are two important details in the generation of these apertures. First, an aperture must be compatible with the fixed ones at the adjacent control points. Namely, the leaf motion limitations must be satisfied. If a row of an aperture is not compatible with the adjacent ones we close all beamlets in this row. The second point is that we first fix the aperture at k, and then $k+2$. If there is a third aperture, finally we fix it at $k+1$. After sequencing all apertures on to a subset of K, we fill the missing control points in such a way that the number of open beamlets in the resulting arc is maximum. Namely, we open all the beamlets as long as they are compatible with the ones at fixed apertures. Thus, we obtain a full treatment arc (consisting of K sequential apertures) to construct RMLP. The pseudo code of this conversion algorithm is given in Figure 7.2.

The first two rows of Figure 7.3 illustrates the conversion of a fluence map at the fifth control point (gantry angle 8°) into three apertures and their sequencing. The fluence map is decomposed into three apertures. The rows of the aperture at k=5 are the first consecutive beamlet chains from the left part of the fluence map. If there is more than one open beamlet chain at any row then we need another aperture to complete the fluence map. The aperture at control point 7 consists of the first open beamlet chains of the rows from the right. Finally, at control point 6, there is

Input: a number of fluence maps

Output: a full treatment arc

for fluence map at control point $k \in \overline{\overline{K}}$ **do**

  for row $i = 1, \ldots, m$ **do**

    if all beamlets with positive intensity are consecutive **then**

      generate a row with one open beamlet chain and if it is compatible with the previous control points in K then fix it at row i at control point k, else close all beamlets

    else

      generate a row including the first open beamlet chain from the left and if it is compatible with the previous control points in K then fix it at row i at control point k, else close all beamlets; and generate a row including the first open beamlet chain from the right and if it is compatible with previous control points in K then fix it at row i at control point $k + 2$, else close all beamlets

    end if

  end for

end for

for fluence map at control point $k \in \overline{\overline{K}}$ **do**

  for row $i = 1, \ldots, m$ **do**

    if there are more than two positive beamlet chains **then**

      generate a row including the second open beamlet chain from the left and if it is compatible with control points k and $k + 2$ then fix it at row i of control point $k + 1$

    end if

  end for

  if at least one row is fixed at control point $k + 1$ **then**

    find compatible apertures for control point $k + 1$ with minimum number of open beamlets

  end if

end for

find compatible apertures with maximum open beamlets for the control points without any fixed rows

Figure 7.2. Conversion Algorithm.

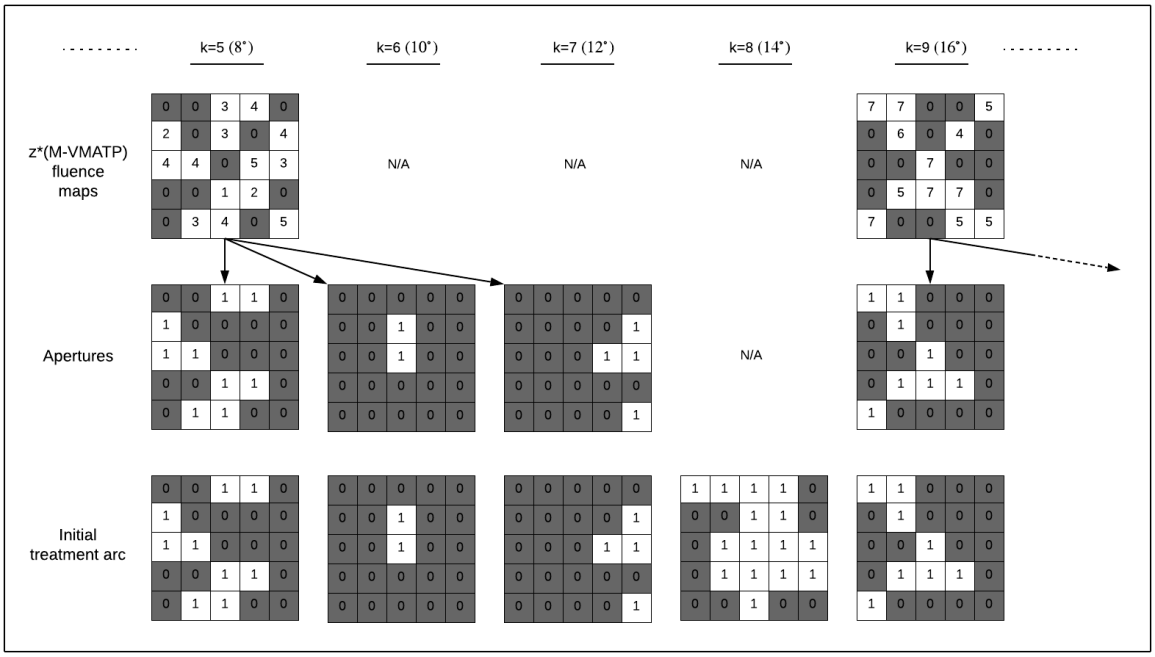

Figure 7.3. Initial treatment arc generation.

an aperture with one beamlet at the second row, since there are three open beamlet chains at the second row of the fluence map. Also, there is an open beamlet at the third row in order to make the apertures at control points 5 and 6 compatible (i.e. satisfying leaf motion limitations). At the last row of the figure, the part of the full treatment arc between 8° and 16° is shown where $\delta = 2$.

7.2. Phase 2: Improvement of the Existing Treatment Plan by Column Generation

Generally, as in the root node of the BP algorithms in Section 6, in each iteration of the column generation method a new promising column is obtained and added to the master problem, which is then solved with a larger column pool. At the optimality of MLP, if all decision variables associated with the columns are integer then an optimal solution of the original integer programming problem is obtained. However, this is a rare situation and generally optimal values of these variables are fractional, which does not provide a feasible solution for the original model.

In order to resolve this problem, at the beginning we generate only one column for each row i for the initial column pool. Hence, there is only one b_i^1 variable for each row and at the optimality of the initial RMLP they are set to 1. Namely, the first optimal solution yields a full treatment arc that is feasible in terms of geometry constraints of the MLC system and the linear accelerator. Moreover, each time a new promising row arc is generated we replace the current one with this new arc to ensure that there remains exactly one column for each row i in the column pool (i.e. the size of the column pool for each row $|Z_i^0| = 1$). Note that replacing an existing column with a new one may worsen the objective function value. The reason is that the new column is not guaranteed to improve the objective function value in the absence of the existing set of columns. We employ a specialized column generation strategy to ensure that the objective function value does not worsen in subsequent iterations of our algorithm. We first observe that often a subset of control points has positive radiation intensity (i.e. $mu_k > 0$) in an optimal solution of RMLP, whereas radiation intensity values associated with the remaining control points are zero. Based on this observation, our column generation strategy ensures that leaf positions corresponding to control points having $mu_k > 0$ are kept constant between successive iterations. This strategy ensures that the previous solution stays feasible with respect to the new set of columns, and therefore the objective function value does not deteriorate in successive iterations. For this reason, each time after solving RLMP, we modify each one of m PSPs in such a way that the leaf positions at the control points with positive radiation intensities remain the same as the positions in the previously generated column. We fix the leaf positions at the corresponding control points by removing the other nodes representing the other leaf configurations when we solve the associated PSP. In each iteration we update the additional constraints that are necessary for fixing the leaf positions.

To summarize, the new two-phase heuristic starts by generating initial columns to construct RMLP in the first phase that includes two steps: solving M-VMATP to obtain a number of fluence maps, and generating a full treatment arc from these fluence maps by applying the conversion algorithm. Initially there is only one column for each row i at RMLP, and they are improved during the column generation iterations in the second phase. The flow diagram of the resulting heuristic is given in Figure 7.4.

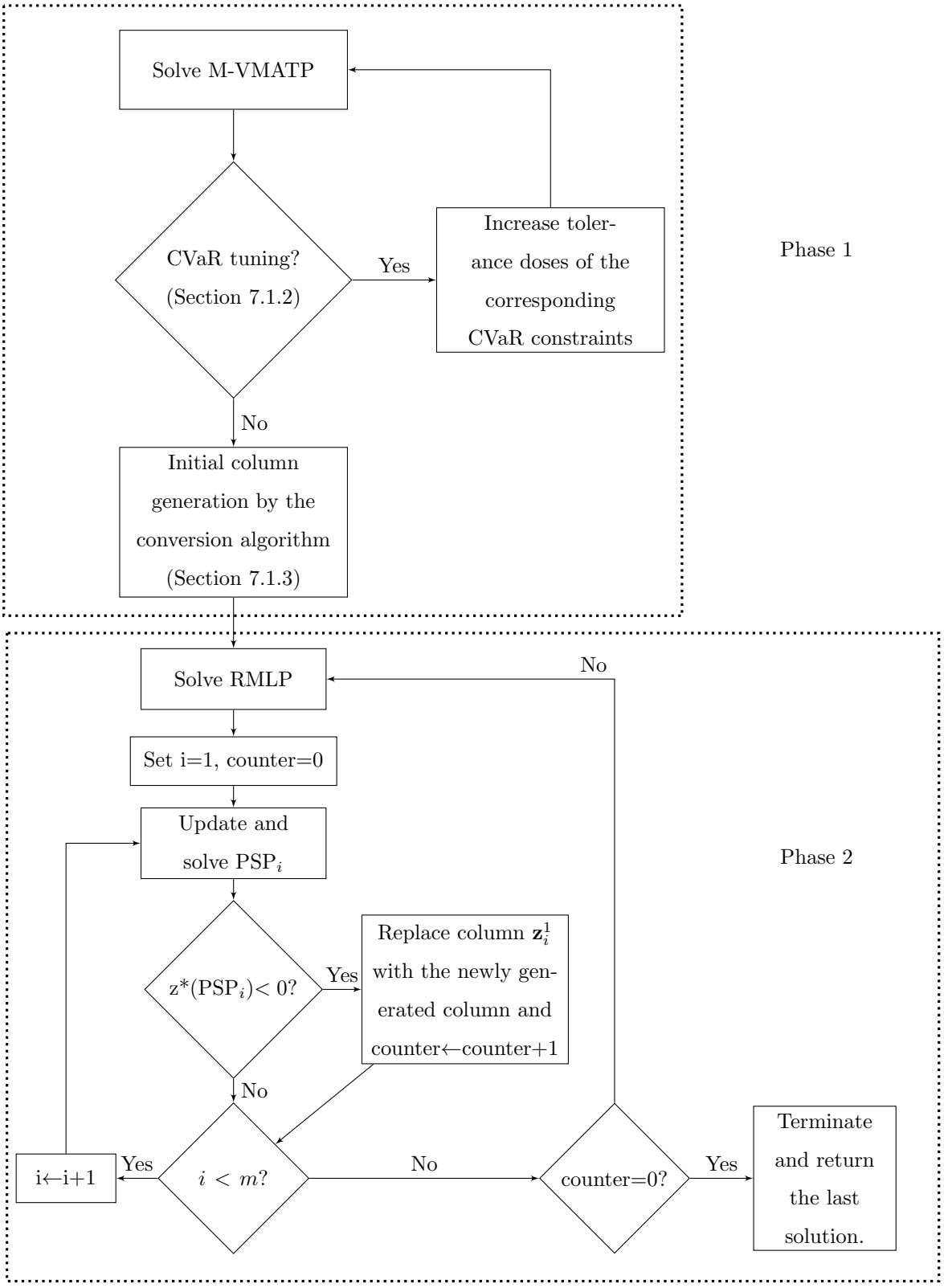

Figure 7.4. Flow diagram of the VMAT planning heuristic.

8. COMPUTATIONAL EXPERIMENTS

In this chapter we give the computational experiments for the evaluation of VMATP models as well as the performance of the algorithms that we have explained in the previous chapters. As can be seen easily there are two different test environments. The first one, is used to compare VMATP models introduced in Chapter 4, and to test the performance of Benders decomposition and BP algorithms explained respectively in Chapter 5 and Chapter 6. The second test environment includes nine real prostate data sets used in the computational experiments for the two-phase column generation based heuristic.

8.1. Test Environments

8.1.1. The First Test Environment

We use a real data set belonging to an anonymous prostate cancer patient provided by Craft *et al.* [90] in common optimization for radiation therapy (CORT) datasets [17] in order to evaluate VMATP formulations, and to test Benders decomposition and BP algorithms, respectively. In this section we define the test bed and the treatment parameters that we use in the computational experiments.

The dose influence matrices for 180 equally spaced control points (i.e. $K = 180$) in units of Gy per MU are provided in the original prostate data set. There are 690,373 voxels of size 3 mm^3 containing 9 defined structures and the remaining part of the body. The defined structures include 2 planning target volumes (PTV68 and PTV56) with different prescription doses and 5 OARs (rectum, bladder, penile bulb (PB), left femoral head, right femoral head). Since the original problem is intractable due to its size, we reduce the size of the problem by randomly selecting a number of voxels from each one of the structures as listed in Table 8.1 and Table 8.2. Also, we reduce total number of structures by assuming that there is only one OAR (union of five OARs) and we consider only PTV68. Note that Table 8.1 and Table 8.2 only contain number of

voxels in the structures we consider in the computational experiments. As an example, 600 voxels are selected from PTV68 to form an instance with 1301 voxels, and the remaining ones are selected from the OARs (120 voxels from rectum, 240 voxels from bladder, 101 voxels from penile bulb, and 120 voxels from each one of the femoral heads). We derive 18 data sets with different total voxel sizes, where each data set has 5 instances generated using a different random number sequence (i.e. there are 90 instances in total). Initially, we generate the first 45 instances (including 9 data sets with 22 voxels to 1301 voxels) and use them to evaluate two VMATP formulations. Then we increase the size of the data sets to 11 (i.e. 55 instances in total) while testing the performance of the Benders decomposition algorithms. Finally, we enlarge the test bed by adding 7 new data sets including new, larger 35 instances, and use them to test the performance of the BP algorithms. We divide the data sets into four groups: small (with 22-220 voxels), medium (with 660-1701 voxels), large (with 1901-2901 voxels), and very large (with 3401 and 4501 voxels). Note that we randomly select voxels from the OARs, and they do not belong to any intersection of the structures.

Table 8.1. Small and medium data sets.

Structure	Voxel	Small					Medium					
PTV68	6770	10	20	30	40	100	300	400	500	600	700	800
Rectum	1764	2	4	6	8	20	60	80	100	120	140	160
Bladder	11596	4	8	12	16	40	120	160	200	240	280	320
PB	101	2	4	6	8	20	60	80	100	101	101	101
Lt F	5857	2	4	6	8	20	60	80	100	120	140	160
Rt F	5974	2	4	6	8	20	60	80	100	120	140	160
TOTAL	32062	22	44	66	88	220	660	880	1100	1301	1501	1701

In the real data set there are 25 404 beamlets with size 1 cm^2. In VMAT planning, the continuous dose delivery is discretized over a finite number of control points and it is assumed that radiation delivery only occurs at the control points. This assumption is reasonable when there is a large number of control points with typically 2°-spacing [54, 56]. Moreover, Otto [12] indicates that poor sampling of control points and MLC leaf positions can degrade the plan accuracy. Thus, this results in unacceptable dosimetric

error. Hence, beamlet grid and control points are kept as they are in the original data set. There are no dose absorption values for the beamlets that do not belong to beam's eye view (BEV) at control points; hence we assume that they are closed during the rotation. Also, we assume that the MLC system has 13 rows and 16 columns ($m = 13, n = 16$), since this size is enough to cover all beamlets having dose absorption information. The data sets and a detailed description file are publicly available on our website [91].

Table 8.2. Large and very large data sets.

Structure	Voxel	Large					Very Large	
PTV68	6770	900	1000	1100	1300	1500	1650	2200
Rectum	1764	180	200	220	240	260	330	440
Bladder	11596	360	400	440	480	520	660	880
PB	101	101	101	101	101	101	101	101
Lt F	5857	180	200	220	240	260	330	440
Rt F	5974	180	200	220	240	260	330	440
TOTAL	32062	1901	2101	2301	2601	2901	3401	4501

In all computational experiments we assume that there is only one co-planar arc in VMAT treatment and the gantry completes a tour in 3 minutes with constant speed. The maximum dose rate of a typical linear accelerator is approximately 600 MUs per minute, which we also use in our experiments. There are 180 control points, and thus at each one of them the gantry delivers radiation for at most 1 second. As a result the maximum radiation dose intensity U^{mu} is set to 10 MUs. A leaf can approximately move 2.5 cm per second at maximum, thus we assume that the value of δ is 2, namely a leaf can move at most two beamlets between consecutive control points. There are one PTV and one OAR in the experiments and each one has only one partial volume constraint. Namely, T and O parameters, and also C_t and C_o parameters equal to 1. α_{11}^{OAR} and α_{11}^{TV} are set to 0.40 and 0.95, respectively. It is assumed that there are 34 fractions in the treatment and in each one the prescribed dose for PTV68 $\overline{d}_{11}$ is 2 Gy (i.e. 68 Gy in total). Upper and lower bounds on the amount of radiation dose absorbed by a voxel in PTV68, U_1^{TV} and L_1^{TV}, are set to 2.14 Gy and 1.9 Gy

per fraction, respectively (i.e. total values for 34 fractions are 72.76 Gy and 64.6 Gy). Finally the tolerance dose for OAR U_{11}^{OAR} is set to 1.47 Gy per fraction (i.e. 50 Gy in total).

8.1.2. The Second Test Environment

We generate VMAT plans by solving our two-phase column generation based heuristic algorithm, which is given in Chapter 7, for nine prostate cancer patients treated in Istanbul University Oncology Institute. The institute is one of the largest and oldest cancer research centers in Turkey, and an average of 5,000 new patients apply annually, and 60,000 patients are called for follow-up and control. Every day around 120-150 patients undergo radiotherapy and 90-100 patients receive chemotherapy. They provided us a number of CT images with 2.5 mm spacing and a radiation therapy (RT) structure set of each one of the patients in digital imaging and communications in medicine (DICOM) format [92]. There are two PTVs with different prescription doses (75.6 Gy and 56 Gy in 36 fractions, respectively) and 5 OARs (rectum, bladder, penile bulb, left and right femoral heads) in each case (i.e. $T = 2$ and $O = 5$).

Istanbul University Oncology Institute currently uses a commercial software called Varian Eclipse Treatment Planning System (TPS) v.15.6 [93] that uses AAA algorithm [22], which is embedded into the TPS, to calculate the dose-influence matrices. It is not possible to export these matrices, hence we compute them for a 6 megavolt (MV) photon energy using an open-source radiation TPS called matRad [94]. We set the voxel resolution to 5 mm^3 and bixel resolution to 1 cm^2 during the DICOM import in matRad. We list total number of voxels in each structure of all patients in Table 8.3. It uses a singular value decomposed pencil beam algorithm to accomplish photon dose calculation [95]. The couch angle is selected 0° for all patients and dose-influence matrices are computed for 180 evenly spaced control points with 2°-spacing are computed. Then, we scale the dose-influence matrices in such a way that the absorbed dose of 1 cGy/MU is delivered at 100 cm source-to-axis distance (SAD) at 5 cm depth with field size 10 cm $\times$ 10 cm. Thus, we divide original dose-influence matrices by the parameter 100 to obtain Gy/MU values. In other words, they are scaled such that a

weight of 1 is equivalent to 100 MU. We also validate this scaling parameter on a water equivalent phantom provided by Istanbul University Oncology Institute. In Appendix B dose calculation steps in matRad and the details of validation process are provided. Note that, matRad's dose calculation is restricted to the projection of the target onto the BEV at each control point. Thus, for each case, we determine the smallest MLC system size (i.e. value of m and n in Table 8.3) that includes all beamlets in the matrices provided by matRad.

Table 8.3. Properties of the prostate cancer data sets.

			Number of voxels of size 5 mm^3							
Patient	**m**	**n**	**PTV75.6**	**R-PTV56**	**Rectum**	**Bladder**	**PB**	**Lt F**	**Rt F**	**TOTAL**
P1	11	13	890	676	436	2836	58	1559	1476	7931
P2	9	13	1127	230	743	1567	20	923	953	5563
P3	9	13	1000	218	886	1915	53	1728	1791	7591
P4	9	13	889	407	736	2026	51	1653	1610	7372
P5	8	13	1056	346	632	755	90	2025	1845	6749
P6	12	9	1198	383	1035	4649	43	1300	1255	9863
P7	8	11	606	157	402	2911	67	1428	1495	7066
P8	9	9	699	213	757	4814	34	1827	1774	10118
P9	10	13	1971	219	753	1209	36	1630	1684	7502

Istanbul University Oncology Institute uses Varian's RapidArc technology to deliver VMAT plans. The MLC system of the linear accelerator consists of 120 leaves, which are 0.5 cm thick at the isocenter for the central 20 cm, and 1 cm in the outer 2×10 cm. The maximum leaf speed is 2.5 cm/second and the dose rate can be 0-600 MU/minute. We set the associated parameters in our algorithm alignment with these properties of the system. We assume that the gantry rotates at a constant rate and completes a tour in 6 minutes. There is single co-planar treatment arc, thus delivery time of each plan is 6 minutes. The gantry moves from one control point to the consecutive one in 2 seconds, thus we set $\delta = 5$ beamlets. Moreover, the maximum radiation intensity is 20 MUs since it is assumed that the radiation delivery lasts 2 seconds at a control point.

Table 8.4 lists all structure dose constraints used for prostate radiation therapy optimization at Istanbul University Oncology Institute in accordance with the recommendation of Buyyounouski *et al.* [96]. There are two partial volume constraints for each one of rectum and bladder (i.e. C_o parameter is 2 for each one of these OARs), and for the remaining structures there is only one constraint (i.e. C_t parameter for target volumes and C_o parameter for the remaining OARs is set to 1). We should note that there are conflicting dose-volume constraint suggestions for the penile bulb (PB) in the literature. First, Buyyounouski *et al.* [96] determine the dose prescription for PB as $D_{\%90} \leq 15$ Gy based on [97], where they study the effect of dose restrictions for erectile tissues on prostate coverage and rectal sparing and are able to limit the PB $D_{\%90}$ to 15 Gy in 80% of men. Emami [98] and Roach III *et al.* [99] explain that it is prudent to keep the mean dose to entire or 95% of the volume at most 50 Gy, respectively, to avoid erectile dysfunction. It is stated in [99] that it may also be careful to limit the $D_{\%70}$ and $D_{\%90}$ to 70 Gy and 50 Gy, respectively, without compromising planning target volume coverage. On the other hand as reported in [100], PB dose is not associated with erectile dysfunction. Moreover, the oncologists and medical physicists at Istanbul University Oncology Institute indicate that the dose prescription $D_{\%90} \leq 15$ Gy for PB is very tight, and it is possible to approve a treatment plan when it does not satisfy this constraint unless the deviation from 15 Gy is excessive. We use the constraints given in Table 8.4 and aim to deliver 75.6 Gy in 36 fractions (2.1 Gy per fraction). Table 8.4 also includes the corresponding ratios of all volumes used to formulate VMATP-1 model. Also, the values of $\overline{d}_{tc}$ and U_{oc}^{OAR} are set to the dose amounts prescribed for one fraction. As for the remaining parameters, lower and upper bound dose limits for PTV75.6 are selected as to 84 Gy (2.334 Gy per fraction) and 67 Gy (1.861 Gy per fraction), respectively. Also, we set the upper bound dose limit for PTV56 to 72 Gy (2 Gy per fraction).

Finally, a TV t may invade an OAR o, namely there may be a set of voxels in an OAR o which also belong to a TV. In such a case the overlap region is considered to belong to the TV t and the CVaR constraint is reformulated for the rest of OAR o. However, the entire OAR must satisfy the associated dose constraint according to the clinical guidelines, so we adjust α_{oc}^{OAR} parameter in the CVaR constraint as described

Table 8.4. Dose-volume constraints used at Istanbul University Oncology Institute.

Structure	$\mathbf{D_{x\%}}^*$	**Dose (in 36 fractions)**	**Ratio of volume**
PTV75.6	$D_{95\%}$	75.6 Gy	$\alpha_{11}^{PTV} = 0.95$
R-PTV56	$D_{95\%}$	56 Gy	$\alpha_{21}^{PTV} = 0.95$
Rectum	$D_{35\%}$	$\leq$ 40 Gy	$\alpha_{11}^{OAR} = 0.65$
	$D_{17\%}$	$\leq$ 65 Gy	$\alpha_{12}^{OAR} = 0.83$
Bladder	$D_{50\%}$	$\leq$ 40 Gy	$\alpha_{21}^{OAR} = 0.50$
	$D_{25\%}$	$\leq$ 65 Gy	$\alpha_{22}^{OAR} = 0.75$
PB	$D_{90\%}$	$\leq$ 15 Gy	$\alpha_{31}^{OAR} = 0.10$
Lt F	$D_{10\%}$	$\leq$ 50 Gy	$\alpha_{41}^{OAR} = 0.90$
Rt F	$D_{10\%}$	$\leq$ 50 Gy	$\alpha_{51}^{OAR} = 0.90$

$\mathbf{D_{x\%}}^*$: the minimum dose received by x% of the structure.

in [24] to meet this requirement. Let R-OAR stands for the rest of the corresponding OAR, then we use

$$\alpha_{oc}^{R-OAR} = 1 - \frac{(1-\alpha_{oc}^{OAR})|V_o^{OAR}| - |V_o^{OAR} \setminus V_o^{R-OAR}|}{|V_o^{R-OAR}|},$$

instead of α_{oc}^{OAR}. Clearly, if the set of voxels belonging also a TV is not empty, i.e. $|V_o^{OAR} \setminus V_o^{R-OAR}| \neq 0$, then $\alpha_{oc}^{R-OAR} > \alpha_{oc}^{OAR}$ (i.e. the resulting ratio α_{oc}^{R-OAR} yields a tighter constraint). Also, if $\alpha_{oc}^{R-OAR}|V_o^{R-OAR}|$ voxels satisfy the dose constraint in R-OAR then $\alpha_{oc}^{OAR}|V_o^{OAR}|$ voxels of the entire OAR also satisfy the constraint. Similarly, if a voxel belongs to more than one TV, it is considered only in the one with the highest prescription dose. Data sets provided by Istanbul University Oncology Institute include the rest of a structure if there is an intersection with this structure and a PTV. The rest of the structure is obtained by subtracting all voxels in the PTV with a margin of 2-3 mm.

8.2. Evaluation of the Formulations

We implement VMATP-1 and VMATP-2 models in Python 2.7 programming language [101] and use Gurobi 6.5 solver [102] running on a computer with Windows

Server 2012 R2 Standard 64-bit PC having 2.00 GHz Intel Xeon CPU, 46 GB RAM. We set time limit to 1800 seconds for both models and only one thread is used in computations.

Table 8.5. Summary of the computational results for VMATP formulations.

	VMATP-1				VMATP-2			
SAMPLE	GAP	CPU	S/T	O/T	GAP	CPU	S/T	O/T
22	0.00	61.8	5/5	5/5	0.00	377.5	5/5	5/5
44	0.00	65.9	5/5	5/5	0.00	407.9	5/5	5/5
66	0.04	501.0	5/5	4/5	0.02	897.5	5/5	4/5
88	0.00	118.1	5/5	5/5	0.00	646.1	5/5	5/5
220	0.00	1224.0	5/5	4/5	0.00	1036.6	5/5	4/5
660	20.00	1656.3	4/5	4/5	100	1800	0/5	0/5
880	60.03	1800	2/5	0/5	100	1800	0/5	0/5
1100	80.00	1675.7	1/5	1/5	100	1800	0/5	0/5
1301	100	1800	0/5	0/5	100	1800	0/5	0/5
Avg/Sum	**29.90**	**989.2**	**32/45**	**28/45**	**44.45**	**1174.0**	**25/45**	**23/45**

In Table 8.5 we give the summary of the computational results that includes average optimality gaps (%), central processing unit (CPU) times (seconds), total number of instances that the corresponding model can find a feasible solution (S/T) and can solve optimally (O/T) out of total instances. There are 9 data sets including 45 instances with at most 1301 voxels. Note that 0 is a valid LB for the objective function, total MUs of the treatment, since the amount of radiation intensity at each control point is nonnegative. Similarly, 1800 MUs is a valid UB since the radiation intensity at a control point can be at most 10 MUs. Whenever a method can find neither a feasible solution nor a LB for an instance, we calculate the optimality gap as 100% using these bounds. Detailed results including lower and upper bounds (in MU), optimality gap (%) and CPU time (in second) of each instance can be found in Table 8.6.

According to the average results, both models can provide an UB for all instances in small size data sets (with 22-220 voxels) and they can solve 23 out of 25 instances

optimally in 1800 seconds. For the larger data sets VMATP-1 outperforms VMATP-2 in all performance measures. It can find an UB for 7 out of 20 instances (i.e. it solves 5 out of 7 instances optimally), however VMATP-2 cannot provide a feasible solution for any of these instances. Since our aim is to increase the problem sizes and develop an algorithm that can solve clinical size of problems, we develop solution algorithms for VMATP-1 model.

8.3. Computational Results for Benders Decomposition Algorithms

We implement all Benders decomposition algorithms and VMATP-1 model in Python 2.7 programming language [101] and use Gurobi 6.5 solver [102] running on a computer with Windows Server 2012 R2 Standard 64-bit PC with 2.00 GHz Intel Xeon CPU, 46 GB RAM. We solve VMATP-1 model by Gurobi, naive Benders algorithm, and the two improved Benders algorithms for all instances in small and medium size data sets. We set the CPU time limit to 3600 seconds in all experiments and execute all algorithms on one thread in order to keep the conditions the same and to be able to compare the performances of them. We change the default method for the RVMATP and AP models in improved Benders algorithms and solve them by the barrier algorithm [103]. Also we set the "MIPFocus" parameter value of the master model in all Benders algorithms to 3 to focus on the bound. We execute Gurobi solver with the default settings and do not perform any parameter tuning while solving VMATP-1.

Table 8.7 summarizes the computational results; it includes the average optimality gap (%) and the average CPU time (in second) of five instances of each size. Similar to the results in Table 8.5 in previous section, the column with title "S/T" and "O/T" show the number of instances that the corresponding method finds a feasible solution and solves optimally out of total instances, respectively. Also, we set the optimality gap to 100% whenever a method can find neither a feasible solution nor a LB for an instance. In addition, whenever a method cannot provide an UB for a test instance we calculate the optimality gap by setting its UB to 1800. Note that the UB on the objective value is 1800 MUs since total number of control points is 180 and it is possible to send 10 MUs radiation at each one of them. Detailed results including bounds,

optimality gap and CPU time of each instance can be found in Table 8.8.

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 LB remains at zero level, which results in 100% optimality gap. 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 problem size 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 LB for the remaining three instances. On the other hand, improved Benders algorithms can find feasible solutions with small average optimality gaps (3.12% and 3.23%, respectively) for all instances, which indicates that a high-quality plan is found for each one of them. Furthermore, for only one of them (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 LB 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 instances the LB 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 optimally within the time limit. Nevertheless, we can conclude that Improved Benders Algorithm 2 is capable of finding good treatment plans even for large problem instances.

Table 8.6. Detailed results for VMATP formulations.

	VMATP-1				VMATP-2					VMATP-1				VMATP-2			
INSTANCE	LB	UB	GAP	CPU	LB	UB	GAP	CPU	INSTANCE	LB	UB	GAP	CPU	LB	UB	GAP	CPU
22-1	236.550	236.552	0.00	74.2	236.550	236.556	0.00	540.3	**660-1**	237.666	237.668	0.00	1736.0	N/A	N/A	N/A*	1800
22-2	231.148	231.148	0.00	68.2	231.148	231.148	0.00	303.6	**660-2**	237.666	237.668	0.00	1797.0	N/A	N/A	N/A*	1800
22-3	232.605	232.606	0.00	35.3	232.605	232.609	0.00	343.7	**660-3**	237.349	237.354	0.00	1234.2	N/A	N/A	N/A*	1800
22-4	233.700	233.700	0.00	69.8	233.700	233.702	0.00	384.5	**660-4**	238.339	238.342	0.00	1714.2	N/A	N/A	N/A*	1800
22-5	234.742	234.746	0.00	61.5	234.742	234.761	0.00	315.5	**660-5**	N/A	N/A	N/A*	1800	N/A	N/A	N/A*	1800
44-1	232.912	232.912	0.00	60.9	232.912	232.934	0.00	300.0	**880-1**	N/A	N/A	N/A*	1800	N/A	N/A	N/A*	1800
44-2	236.830	236.835	0.00	112.8	236.830	236.830	0.00	712.3	**880-2**	237.412	237.485	0.03	1800	N/A	N/A	N/A*	1800
44-3	232.912	232.912	0.00	62.6	232.912	232.934	0.00	299.6	**880-3**	238.357	238.663	0.13	1800	N/A	N/A	N/A*	1800
44-4	236.838	236.840	0.00	35.7	236.838	236.840	0.00	434.8	**880-4**	N/A	N/A	N/A*	1800	N/A	N/A	N/A*	1800
44-5	236.429	236.433	0.00	57.6	236.429	236.440	0.00	292.8	**880-5**	N/A	N/A	N/A*	1800	N/A	N/A	N/A*	1800
66-1	236.830	236.843	0.00	83.2	236.830	236.831	0.00	684.5	**1100-1**	N/A	N/A	N/A*	1800	N/A	N/A	N/A*	1800
66-2	237.597	237.612	0.00	361.6	237.597	237.620	0.00	799.9	**1100-2**	237.714	237.722	0.00	1178.6	N/A	N/A	N/A*	1800
66-3	236.451	236.451	0.00	45.6	236.451	236.458	0.00	350.4	**1100-3**	N/A	N/A	N/A*	1800	N/A	N/A	N/A*	1800
66-4	234.940	234.940	0.00	214.9	234.940	234.961	0.00	852.8	**1100-4**	N/A	N/A	N/A*	1800	N/A	N/A	N/A*	1800
66-5	237.731	238.128	0.17	1800	237.732	237.922	0.08	1800	**1100-5**	N/A	N/A	N/A*	1800	N/A	N/A	N/A*	1800
88-1	236.830	236.830	0.00	93.4	236.830	236.831	0.00	890.0	**1301-1**	N/A	N/A	N/A*	1800	N/A	N/A	N/A*	1800
88-2	237.050	237.050	0.00	92.4	237.050	237.058	0.00	361.0	**1301-2**	N/A	N/A	N/A*	1800	N/A	N/A	N/A*	1800
88-3	236.645	236.667	0.00	138.9	236.645	236.655	0.00	676.9	**1301-3**	N/A	N/A	N/A*	1800	N/A	N/A	N/A*	1800
88-4	237.048	237.048	0.00	150.1	237.048	237.067	0.00	765.9	**1301-4**	N/A	N/A	N/A*	1800	N/A	N/A	N/A*	1800
88-5	235.866	235.883	0.00	116.1	235.866	235.868	0.00	536.6	**1301-5**	N/A	N/A	N/A*	1800	N/A	N/A	N/A*	1800
220-1	236.864	236.864	0.00	970.3	236.864	236.886	0.00	921.8									
220-2	237.067	237.086	0.00	1675.6	237.067	237.067	0.00	695.3									
220-3	237.924	237.947	0.00	1023.8	237.924	237.924	0.00	798.9									
220-4	237.028	237.032	0.00	650.5	237.028	237.028	0.00	967.1									
220-5	237.871	237.899	0.01	1800	237.871	237.899	0.01	1800									

Note 1: LB and UB are in MU, GAP % and CPU in seconds. **Note 2:** cells marked with * are accepted as 100% in Table 8.5

Table 8.7. Summary of the computational results for Gurobi solver and Benders decomposition algorithms.

	Gurobi				Naive Benders Alg.				Impr. Benders Alg. 1				Impr. Benders Alg. 2			
SAMPLE	GAP	CPU	S/T	O/T	GAP	CPU	S/T	O/T	GAP	CPU	S/T	O/T	GAP	CPU	S/T	O/T
22	0.00	62.4	5/5	5/5	100	3600	4/5	0/5	0.00	1502.3	5/5	4/5	0.00	112.2	5/5	5/5
44	0.00	63.1	5/5	5/5	100	3600	5/5	0/5	0.00	245.6	5/5	5/5	0.00	627.8	5/5	5/5
66	0.01	809.8	5/5	4/5	100	3600	3/5	0/5	2.33	2330.8	5/5	2/5	0.14	1884.2	5/5	3/5
88	0.00	104.8	5/5	5/5	100	3600	1/5	0/5	0.00	243.4	5/5	5/5	0.00	826.9	5/5	4/5
220	0.00	1455.0	5/5	4/5	100	3600	2/5	0/5	0.61	1661.9	5/5	3/5	0.02	1187.1	5/5	4/5
660	20.00	2437.2	4/5	4/5	100	3600	2/5	0/5	6.12	3333.7	5/5	2/5	0.22	3585.9	5/5	1/5
880	66.67	3067.1	2/5	2/5	100	3600	0/5	0/5	3.12	3600	5/5	0/5	3.23	3600	5/5	0/5
1100	40.00	2690.6	3/5	3/5	100	3600	0/5	0/5	4.64	3600	5/5	0/5	0.96	3600	5/5	0/5
1301	40.00	2930.6	3/5	3/5	100	3600	1/5	0/5	3.21	3600	5/5	0/5	0.34	3600	5/5	0/5
1501	60.00	2861.8	2/5	2/5	100	3600	0/5	0/5	19.38	3600	4/5	0/5	18.43	3600	4/5	0/5
1701	40.00	2286.3	3/5	3/5	100	3600	0/5	0/5	13.59	3600	5/5	0/5	0.49	3600	5/5	0/5
Avg/Sum	**24.25**	**1706.3**	**42/55**	**40/55**	**100**	**3600**	**18/55**	**0/55**	**4.82**	**2483.4**	**54/55**	**21/55**	**2.17**	**2384.0**	**54/55**	**22/55**

Table 8.8: Detailed results for Benders decomposition algorithms.

	Gurobi				Naive Benders Alg.				Impr. Benders Alg. 1				Impr. Benders Alg. 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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	N/A	N/A*	3600	235.866	235.866	0.00	192.2	235.866	235.867	0.00	82.0

Table 8.8: Detailed results for Benders decomposition algorithms (cont.).

	Gurobi				Naive Benders Alg.				Impr. Benders Alg. 1				Impr. Benders Alg. 2			
INSTANCE	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU
220-1	236.864	236.864	0.00	804.8	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	N/A	N/A*	3600	237.847	237.847	0.00	2991.8	237.847	239.098	0.52	3600
660-3	237.349	237.351	0.00	1273.8	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	N/A	N/A*	3600	229.358	238.886	3.99	3600	237.801	238.182	0.16	3600

Table 8.8: Detailed results for Benders decomposition algorithms (cont.).

	Gurobi				Naive Benders Alg.				Impr. Benders Alg. 1				Impr. Benders Alg. 2			
INSTANCE	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU
1301-1	237.785	237.797	0.00	1952.3	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	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	0.000	N/A	N/A*	3600	232.831	238.824	2.51	3600	237.677	238.754	0.45	3600

Note 1: LB and UB are in MU, GAP % and CPU in seconds.

Note 2: cells marked with * are accepted as 100%, and ** are calculated as 100(1800-LB)/1800% in Table 8.7.

8.4. Computational Results for Branch-and-price Algorithms

We implement BP algorithms and Improved Benders Algorithm 2, which is the best performing algorithm of Chapter 5 as shown in Section 8.3, in Python 2.7 programming language [101] and use Gurobi 8.0 as the MILP solver [104]. All tests are carried out on a 64-bit PC with 3.20 GHz Intel(R) Core(TM) i5-6500 CPU and 8 GB of RAM. We solve VMATP-1 model by Gurobi 8.0 [104], BP algorithms and Improved Benders Algorithm 2 using all instances in all data sets (i.e. 90 instances in total) in order to compare the performance of the proposed BP algorithms with the others. We set 3600 seconds as CPU time limit and use one thread in all executions of all algorithms. In the BP algorithms, RVMATP model is solved using the barrier method [103] at the root node and then its method is changed to dual simplex in the descendant search nodes. We solve RMLP using primal simplex in order to warm start from the last basis after adding a new column. Also, there is a threshold on the reduced cost for the new generated columns. If the reduced cost is not below -0.05 we do not add the corresponding column to RMLP. We keep the parameter tuning of Improved Benders Algorithm 2 as explained at the beginning of Section 8.3. We do not perform any other parameter tuning for the Gurobi solver and keep parameters at their default settings.

We give the summary of the computational results that includes average optimality gaps (%), CPU times (seconds), total number of instances that the corresponding method can find a feasible solution (S/T) and can solve optimally (O/T) out of total instances in Table 8.9. We calculate the optimality gap of an instance as 100% whenever an algorithm cannot provide lower and upper bounds. Also, we accept the UB 1800 MUs in the optimality gap calculation when only a LB is provided. In Table 8.10 the computational results that include lower and upper bounds for each one of the instances are provided.

We partition data sets into four groups: small (with 22-220 voxels), medium (with 660-1701 voxels), large (with 1901-2901 voxels), and very large (with 3401 and 4501 voxels). The results on small data sets (with at most 220 voxels) show that Gurobi, Improved Benders Algorithm 2, BP Algorithm 2 and BP Algorithm 3 can solve almost

all of the instances optimally in short CPU times. BP Algorithm 1 cannot find a feasible solution for one instance having size 220 within time limit. Gurobi performs better than all of the BP algorithms and Improved Benders Algorithm 2 with respect to the average optimality gaps in all small data sets and also with respect to the average CPU times in data sets with 22, 44, and 66 voxels. In particular, Gurobi can solve all instances optimally except 66-5 and 220-5. BP algorithms and Improved Benders Algorithm 2 cannot solve these instances and also some other instances optimally.

As the size of the problem increases, Gurobi starts failing to solve some instances within time limit. It cannot provide neither an UB nor a LB for 9 out of 30 instances of medium size (having total number of voxels between 660-1701), and can only provide a LB for 2 out of 30 instances. It solves 18 out of the remaining 19 medium size instances optimally in relatively longer CPU times. On the other hand, all of the new BP algorithms perform better than Gurobi in both performance measures (only the average optimality gaps of data sets with 660 and 1100 voxels are worse in BP Algorithm 1, and also the average optimality gap of data set with 660 voxels is slightly worse in BP Algorithm 2). They can find a feasible solution for all of the medium size instances. Moreover, BP Algorithm 1, BP Algorithm 2, and BP Algorithm 3 can respectively solve 20, 26, and 29 instances optimally. BP Algorithm 2 cannot solve only 4 instances (660-1, 660-2, 1301-1 and 1301-2), and BP Algorithm 3 cannot solve only one instance (1301-2) optimally. Also, the resulting optimality gaps of these instances are very small (at most 0.02%). Improved Benders Algorithm 2 can also find a feasible solution for all of the medium size instances except 1701-2. The optimality gaps are below 1% in almost all cases (except 660-3, 880-1 and 1100-5), however the number of instances that it can solve optimally is only three. In particular, the smallest average CPU times and optimality gaps are obtained by BP Algorithm 3. As a result, BP Algorithm 3 outperforms the other ones for medium size instances.

The results of large problems (with 1901-2901 voxels) are also similar to the results of medium size problems. Gurobi can solve only 2 out of 25 instances optimally within time limit, and for 22 instances it can neither provide an UB nor a LB, for the remaining one instance it can only provide a LB. On the other hand, BP Algorithm 3 solves all

of the instances optimally within less than half of the time limit. The other two BP algorithms can also provide an UB for almost all of the instances. BP Algorithm 1 and BP Algorithm 2 cannot provide a feasible solution respectively for 3 and 1 instance within time limit. The number of instances that these algorithms can solve optimally is 17 and 20, respectively. Also, the optimality gaps of the problems, which are not solved optimally, are below 1% in almost all instances. Improved Benders Algorithm 2 can find a feasible solution for 20 instances, but it can solve only one of them optimally. Finally, BP Algorithm 3 outperforms other algorithms in both performance measures with significant differences.

In addition to these three groups of data sets, we generate and solve two larger data sets with 3401 and 4501 voxels to be able to make the difference between algorithms clearer and observe the limits of BP Algorithm 3. Gurobi fails to find lower and upper bounds for all instances. BP Algorithm 1 and BP Algorithm 2 both find a feasible solution for 5 and 4 out of 10 instances, and they can solve 3 and 2 instances optimally, respectively. On the other hand, BP Algorithm 3 solves 8 instances to optimally. It can only provide a LB for each one of the remaining two instances. Improved Benders Algorithm 2 can solve 4 out of 10 instances with small optimality gaps (below 1%), but it cannot solve any of the instances optimally.

In overall, BP Algorithm 3 can solve 83 out of 90 instances optimally, and for the remaining 5 instances it can provide very small optimality gaps. (i.e. below 0.1% except instance 22-5). However, Gurobi fails to provide upper and lower bounds for almost half of the instances (i.e. 41 out of 90) within time limit. Also, it cannot provide an UB for other 3 instances. It only solves 43 instances optimally and 3 instances with small optimality gaps. If we check the average CPU times and optimality gaps, BP Algorithm 3 outperforms all other methods in both performance measures in almost all data sets (except with 22, 44 and 66 voxels). The last row of Table 8.9 shows that the minimum average optimality gap of all instances is 1.93%, the minimum average CPU time is 1021.6 seconds, the maximum total number of instances with a feasible solution is 88, and the maximum number of instances that are solved optimally is 83, which are all obtained by BP Algorithm 3.

Table 8.9. Summary of the computational results of BP algorithms.

	Gurobi				BP Algorithm 1				BP Algorithm 2				BP Algorithm 3				Impr. Benders Alg. 2			
SAMPLE	GAP	CPU	S/T	O/T	GAP	CPU	S/T	O/T	GAP	CPU	S/T	O/T	GAP	CPU	S/T	O/T	GAP	CPU	S/T	O/T
22	0.00	24.2	5/5	5/5	0.00	60.6	5/5	5/5	0.00	40.7	5/5	5/5	0.03	1110.8	5/5	4/5	0.00	90.4	5/5	5/5
44	0.00	23.2	5/5	5/5	0.00	60.1	5/5	5/5	0.00	74.6	5/5	5/5	0.00	36.6	5/5	5/5	0.00	138.0	5/5	5/5
66	0.01	753.8	5/5	4/5	0.39	2176.2	5/5	2/5	0.03	2180.6	5/5	2/5	0.03	2010.4	5/5	3/5	0.16	1610.2	5/5	3/5
88	0.00	79.2	5/5	5/5	0.00	64.2	5/5	5/5	0.00	76.1	5/5	5/5	0.00	44.9	5/5	5/5	0.00	130.5	5/5	5/5
220	0.00	1233.7	5/5	4/5	17.36	800.9	4/5	4/5	0.01	810.9	5/5	4/5	0.00	792.4	5/5	4/5	0.03	1983.7	5/5	3/5
660	0.00	1914.4	5/5	5/5	1.20	1897.4	5/5	2/5	0.01	1759.6	5/5	3/5	0.00	222.7	5/5	5/5	2.04	3151.8	5/5	1/5
880	20.00	2931.7	4/5	3/5	0.00	927.8	5/5	5/5	0.00	649.5	5/5	5/5	0.00	187.1	5/5	5/5	3.07	3093.8	5/5	1/5
1100	0.00	2536.6	5/5	5/5	4.48	1352.7	5/5	4/5	0.00	824.1	5/5	5/5	0.00	381.9	5/5	5/5	0.34	3259.4	5/5	1/5
1301	37.35	2908.2	3/5	3/5	0.53	1907.3	5/5	3/5	0.01	2035.3	5/5	3/5	0.00	1200.8	5/5	4/5	0.26	3600	5/5	0/5
1501	97.36	3600	0/5	0/5	0.00	889.0	5/5	5/5	0.00	952.1	5/5	5/5	0.00	613.1	5/5	5/5	0.26	3600	5/5	0/5
1701	60.00	3205.8	2/5	2/5	0.67	3044.7	5/5	1/5	0.00	1666.4	5/5	5/5	0.00	938.1	5/5	5/5	17.55	3600	4/5	0/5
1901	77.36	3390.5	1/5	1/5	0.00	1867.5	5/5	5/5	0.00	1959.3	5/5	5/5	0.00	910.1	5/5	5/5	34.88	3600	3/5	0/5
2101	80.00	3587.2	1/5	1/5	0.17	2729.2	5/5	3/5	0.00	2615.1	5/5	5/5	0.00	1132.1	5/5	5/5	0.30	3600	5/5	0/5
2301	100	3600	0/5	0/5	34.72	2643.5	3/5	3/5	0.01	2476.0	5/5	3/5	0.00	800.0	5/5	5/5	34.87	3600	3/5	0/5
2601	100	3600	0/5	0/5	0.16	2551.5	5/5	4/5	0.00	2705.5	5/5	5/5	0.00	1209.0	5/5	5/5	1.17	3404.6	5/5	1/5
2901	100	3600	0/5	0/5	17.78	3172.6	4/5	2/5	20.43	3356.6	4/5	2/5	0.00	1441.3	5/5	5/5	17.62	3600	4/5	0/5
3401	100	3600	0/5	0/5	0.75	3011.8	5/5	3/5	20.75	3179.9	4/5	2/5	0.00	2041.7	5/5	5/5	17.73	3600	4/5	0/5
4501	100	3600	0/5	0/5	100	3600	0/5	0/5	100	3600	0/5	0/5	34.71	3316.9	3/5	3/5	86.78	3600	0/5	0/5
Avg/Sum	**48.45**	**2454.9**	**46/90**	**43/90**	**9.90**	**1819.8**	**81/90**	**61/90**	**7.85**	**1720.1**	**83/90**	**69/90**	**1.93**	**1021.6**	**88/90**	**83/90**	**12.06**	**2736.8**	**78/90**	**25/90**

Table 8.10: Detailed computational results of BP Algorithms.

	Gurobi				BP Algorithm 1				BP Algorithm 2				BP Algorithm 3				Impr. Benders Alg. 2			
INSTANCE	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU
22-1	236.550	236.550	0.00	34.0	236.550	236.550	0.00	35.6	236.550	236.550	0.00	36.8	236.550	236.550	0.00	1875.4	236.550	236.572	0.00	130.8
22-2	231.148	231.148	0.00	16.3	231.148	231.148	0.00	32.4	231.148	231.148	0.00	35.3	231.148	231.148	0.00	30.5	231.148	231.152	0.00	129.2
22-3	232.605	232.606	0.00	15.4	232.605	232.622	0.00	40.8	232.605	232.627	0.00	51.0	232.605	232.626	0.00	24.9	232.605	232.626	0.00	94.0
22-4	233.700	233.702	0.00	26.5	233.700	233.700	0.00	177.0	233.700	233.700	0.00	60.7	233.700	233.700	0.00	23.0	233.700	233.709	0.00	69.7
22-5	234.742	234.760	0.00	28.6	234.742	234.751	0.00	17.2	234.742	234.751	0.00	19.5	234.742	235.107	0.16	3600	234.742	234.753	0.00	28.4
44-1	232.912	232.913	0.00	14.0	232.912	232.916	0.00	37.5	232.912	232.916	0.00	48.4	232.912	232.912	0.00	32.2	232.912	232.929	0.00	172.1
44-2	236.830	236.847	0.00	38.4	236.830	236.830	0.00	39.2	236.830	236.830	0.00	50.0	236.830	236.830	0.00	47.4	236.830	236.840	0.00	144.0
44-3	232.903	232.903	0.00	19.6	232.903	232.903	0.00	34.8	232.903	232.903	0.00	38.8	232.903	232.903	0.00	29.5	232.903	232.911	0.00	61.1
44-4	236.838	236.855	0.00	19.4	236.838	236.838	0.00	143.6	236.838	236.838	0.00	175.0	236.838	236.852	0.00	45.7	236.838	236.846	0.00	91.7
44-5	236.429	236.429	0.00	24.7	236.429	236.429	0.00	45.5	236.429	236.429	0.00	60.8	236.429	236.429	0.00	28.4	236.429	236.432	0.00	221.3
66-1	236.830	236.830	0.00	29.0	236.830	236.831	0.00	33.4	236.830	236.831	0.00	41.4	236.830	236.830	0.00	53.1	236.830	236.831	0.00	666.6
66-2	237.597	237.614	0.00	45.2	237.597	237.634	0.02	3600	237.597	237.631	0.01	3600	237.597	237.688	0.04	3600	237.597	237.617	0.00	142.7
66-3	236.451	236.451	0.00	13.5	236.451	236.451	0.00	47.9	236.451	236.456	0.00	61.3	236.451	236.451	0.00	50.3	236.451	236.465	0.00	41.7
66-4	234.940	234.956	0.00	81.1	234.940	234.970	0.01	3600	234.940	234.970	0.01	3600	234.940	234.940	0.00	2748.5	234.940	236.322	0.58	3600
66-5	237.732	237.874	0.06	3600	237.730	242.330	1.90	3600	237.730	237.993	0.11	3600	237.730	237.949	0.09	3600	237.730	238.228	0.21	3600
88-1	236.830	236.830	0.00	46.4	236.830	236.834	0.00	89.3	236.830	236.837	0.00	110.0	236.830	236.830	0.00	35.7	236.830	236.845	0.00	157.1
88-2	237.050	237.067	0.00	43.9	237.050	237.050	0.00	89.1	237.050	237.050	0.00	103.8	237.050	237.050	0.00	60.7	237.050	237.058	0.00	207.3
88-3	236.645	236.647	0.00	75.2	236.645	236.645	0.00	61.9	236.645	236.656	0.00	86.1	236.645	236.645	0.00	41.0	236.645	236.655	0.00	52.4
88-4	237.048	237.048	0.00	158.8	237.048	237.048	0.00	46.8	237.048	237.048	0.00	46.9	237.048	237.048	0.00	42.4	237.048	237.055	0.00	199.3
88-5	235.866	235.866	0.00	71.9	235.866	235.867	0.00	33.8	235.866	235.867	0.00	33.6	235.866	235.866	0.00	44.7	235.866	235.867	0.00	36.5

Table 8.10: Detailed computational results of BP Algorithms (cont.).

	Gurobi				BP Algorithm 1				BP Algorithm 2				BP Algorithm 3				Impr. Benders Alg. 2			
INSTANCE	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU
220-1	236.864	236.874	0.00	335.5	236.864	236.864	0.00	81.2	236.864	236.864	0.00	81.1	236.864	236.864	0.00	59.9	236.864	236.951	0.04	3600
220-2	237.067	237.073	0.00	550.1	237.067	237.067	0.00	83.0	237.067	237.067	0.00	82.9	237.067	237.067	0.00	145.7	237.067	237.090	0.00	184.5
220-3	237.924	237.924	0.00	536.8	237.924	237.924	0.00	160.0	237.924	237.924	0.00	210.2	237.924	237.924	0.00	81.0	237.924	237.935	0.00	2267.1
220-4	237.028	237.028	0.00	1146.3	237.028	237.028	0.00	80.3	237.028	237.028	0.00	80.4	237.028	237.028	0.00	75.3	237.028	237.049	0.00	267.0
220-5	237.872	237.898	0.01	3600	237.870	N/A	N/A**	3600	237.870	238.001	0.06	3600	237.870	237.926	0.02	3600	237.870	238.105	0.10	3600
660-1	237.666	237.678	0.00	2341.4	237.666	237.701	0.01	1442.2	237.666	237.701	0.01	3600	237.666	237.666	0.00	267.3	237.666	237.683	0.00	1359.0
660-2	237.847	237.866	0.00	1842.8	237.847	241.372	1.46	3600	237.847	237.900	0.02	3600	237.847	237.847	0.00	135.8	237.847	238.532	0.29	3600
660-3	237.349	237.349	0.00	2262.6	237.349	237.349	0.00	462.3	237.349	237.349	0.00	462.4	237.349	237.349	0.00	139.6	237.349	262.662	9.64	3600
660-4	238.339	238.345	0.00	1585.2	238.339	249.641	4.53	3600	238.339	238.361	0.00	753.4	238.339	238.339	0.00	196.4	238.339	238.409	0.03	3600
660-5	237.869	237.869	0.00	1539.9	237.869	237.869	0.00	382.4	237.869	237.869	0.00	382.0	237.869	237.869	0.00	374.1	237.869	238.410	0.23	3600
880-1	238.177	238.190	0.00	2457.7	238.177	238.179	0.00	869.1	238.177	238.186	0.00	1011.2	238.177	238.177	0.00	174.0	238.177	279.394	14.75	3600
880-2	N/A	N/A	N/A*	3600	237.412	237.412	0.00	2383.6	237.412	237.412	0.00	723.1	237.412	237.412	0.00	199.8	237.412	238.093	0.29	3600
880-3	238.357	238.385	0.01	3600	238.357	238.358	0.00	433.6	238.357	238.358	0.00	561.1	238.357	238.358	0.00	210.9	238.357	238.806	0.19	3600
880-4	237.214	237.238	0.00	2551.7	237.214	237.214	0.00	388.3	237.214	237.214	0.00	388.7	237.214	237.214	0.00	202.3	237.214	237.227	0.00	1069.2
880-5	238.006	238.022	0.00	2449.0	238.006	238.017	0.00	564.5	238.006	238.017	0.00	563.4	238.006	238.006	0.00	148.4	238.006	238.302	0.12	3600
1100-1	238.082	238.101	0.00	2290.0	238.082	306.808	22.40	3600	238.082	238.082	0.00	1132.6	238.082	238.082	0.00	286.7	238.082	239.057	0.41	3600
1100-2	237.714	237.717	0.00	1599.2	237.714	237.714	0.00	1123.2	237.714	237.714	0.00	952.6	237.714	237.714	0.00	604.3	237.714	237.730	0.00	1897.2
1100-3	237.165	237.170	0.00	2623.5	237.165	237.165	0.00	851.3	237.165	237.165	0.00	640.6	237.165	237.165	0.00	431.4	237.165	237.493	0.14	3600
1100-4	237.321	237.321	0.00	2795.7	237.321	237.321	0.00	555.4	237.321	237.321	0.00	551.5	237.321	237.321	0.00	344.8	237.321	237.447	0.05	3600
1100-5	237.801	237.802	0.00	3374.6	237.801	237.801	0.00	633.7	237.801	237.801	0.00	843.3	237.801	237.801	0.00	242.1	237.801	240.459	1.11	3600

Table 8.10: Detailed computational results of BP Algorithms (cont.).

	Gurobi				BP Algorithm 1				BP Algorithm 2				BP Algorithm 3				Impr. Benders Alg. 2			
INSTANCE	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU
1301-1	237.785	237.796	0.00	1964.5	237.785	240.423	1.10	3600	237.785	237.824	0.02	3600	237.785	237.785	0.00	1159.8	237.785	239.284	0.63	3600
1301-2	N/A	N/A	N/A*	3600	238.405	242.142	1.54	3600	238.405	238.437	0.01	3600	238.405	238.430	0.01	3600	238.405	238.792	0.16	3600
1301-3	237.901	237.917	0.00	2700.2	237.901	237.916	0.00	878.2	237.901	237.903	0.00	1040.2	237.901	237.903	0.00	400.9	237.901	238.464	0.24	3600
1301-4	238.138	N/A	N/A**	3600	238.137	238.151	0.00	629.3	238.137	238.151	0.00	897.5	238.137	238.151	0.00	607.7	238.137	238.534	0.17	3600
1301-5	237.386	237.386	0.00	2676.3	237.386	237.386	0.00	829.1	237.386	237.386	0.00	1038.5	237.386	237.386	0.00	235.4	237.386	237.654	0.11	3600
1501-1	N/A	N/A	N/A*	3600	237.770	237.777	0.00	938.1	237.770	237.777	0.00	839.4	237.770	237.770	0.00	736.5	237.770	238.253	0.20	3600
1501-2	N/A	N/A	N/A*	3600	237.459	237.459	0.00	715.9	237.459	237.459	0.00	1161.3	237.459	237.459	0.00	472.9	237.459	238.068	0.26	3600
1501-3	N/A	N/A	N/A*	3600	237.925	237.925	0.00	1325.6	237.925	237.925	0.00	1446.2	237.925	237.925	0.00	438.4	237.925	238.263	0.14	3600
1501-4	N/A	N/A	N/A*	3600	237.925	237.925	0.00	701.2	237.925	237.925	0.00	636.3	237.925	237.925	0.00	751.4	237.925	238.347	0.18	3600
1501-5	237.665	N/A	N/A**	3600	237.665	237.665	0.00	764.2	237.665	237.665	0.00	677.3	237.665	237.665	0.00	666.0	237.665	238.894	0.51	3600
1701-1	N/A	N/A	N/A*	3600	238.008	241.879	1.60	3600	238.008	238.025	0.00	2815.2	238.008	238.008	0.00	604.1	238.008	238.239	0.10	3600
1701-2	N/A	N/A	N/A*	3600	237.823	237.842	0.00	823.3	237.823	237.842	0.00	1335.5	237.823	237.823	0.00	814.2	237.823	N/A	N/A**	3600
1701-3	238.105	238.105	0.00	2348.4	238.105	239.348	0.52	3600	238.105	238.105	0.00	1142.3	238.105	238.105	0.00	732.2	238.105	238.239	0.06	3600
1701-4	237.896	237.899	0.00	2880.3	237.896	239.334	0.60	3600	237.896	237.896	0.00	1246.2	237.896	237.896	0.00	1096.1	237.896	238.715	0.34	3600
1701-5	N/A	N/A	N/A*	3600	237.677	239.182	0.63	3600	237.677	237.677	0.00	1792.8	237.677	237.677	0.00	1443.9	237.677	238.823	0.48	3600
1901-1	N/A	N/A	N/A*	3600	237.693	237.706	0.00	1522.6	237.693	237.706	0.00	2632.7	237.693	237.701	0.00	1242.6	237.693	N/A	N/A**	3600
1901-2	237.709	N/A	N/A**	3600	237.709	237.709	0.00	890.5	237.709	237.709	0.00	888.5	237.709	237.709	0.00	599.6	237.709	239.087	0.58	3600
1901-3	237.829	237.830	0.00	2552.3	237.829	237.836	0.00	2022.6	237.829	237.836	0.00	2635.2	237.829	237.829	0.00	1265.9	237.829	238.211	0.16	3600
1901-4	N/A	N/A	N/A*	3600	237.662	237.678	0.00	2577.5	237.662	237.678	0.00	1938.8	237.662	237.662	0.00	712.0	237.662	N/A	N/A**	3600
1901-5	N/A	N/A	N/A*	3600	238.216	238.216	0.00	2324.4	238.216	238.216	0.00	1701.1	238.216	238.216	0.00	730.5	238.216	238.360	0.06	3600

Table 8.10: Detailed computational results of BP Algorithms (cont.).

	Gurobi				BP Algorithm 1				BP Algorithm 2				BP Algorithm 3				Impr. Benders Alg. 2			
INSTANCE	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU
2101-1	N/A	N/A	N/A*	3600	237.786	237.787	0.00	3475.3	237.786	237.787	0.00	1846.6	237.786	237.787	0.00	1150.4	237.786	238.496	0.30	3600
2101-2	N/A	N/A	N/A*	3600	237.622	237.622	0.00	1763.4	237.622	237.622	0.00	2507.2	237.622	237.622	0.00	944.0	237.622	238.534	0.38	3600
2101-3	N/A	N/A	N/A*	3600	237.770	237.775	0.00	1207.4	237.770	237.775	0.00	2383.0	237.770	237.770	0.00	1069.2	237.770	239.031	0.53	3600
2101-4	N/A	N/A	N/A*	3600	237.812	239.124	0.55	3600	237.812	237.828	0.00	2739.7	237.812	237.812	0.00	1238.4	237.812	238.371	0.23	3600
2101-5	237.858	237.858	0.00	3536.2	237.858	238.544	0.29	3600	237.858	237.860	0.00	3599.0	237.858	237.860	0.00	1258.3	237.858	237.969	0.05	3600
2301-1	N/A	N/A	N/A*	3600	237.970	N/A	N/A**	3600	237.970	237.985	0.00	2818.1	237.970	237.970	0.00	1100.1	237.970	N/A	N/A**	3600
2301-2	N/A	N/A	N/A*	3600	238.049	238.053	0.00	2814.4	238.049	238.114	0.03	3600	238.049	238.049	0.00	658.0	238.049	N/A	N/A**	3600
2301-3	N/A	N/A	N/A*	3600	237.679	N/A	N/A**	3600	237.679	237.716	0.02	3600	237.679	237.679	0.00	770.5	237.679	238.814	0.48	3600
2301-4	N/A	N/A	N/A*	3600	237.593	237.593	0.00	2161.4	237.593	237.593	0.00	1323.2	237.593	237.593	0.00	660.2	237.593	238.072	0.20	3600
2301-5	N/A	N/A	N/A*	3600	238.164	238.164	0.00	1041.9	238.164	238.164	0.00	1038.9	238.164	238.164	0.00	811.1	238.164	238.471	0.13	3600
2601-1	N/A	N/A	N/A*	3600	237.826	239.782	0.82	3600	237.826	237.827	0.00	3175.5	237.826	237.827	0.00	1124.2	237.826	238.532	0.30	3600
2601-2	N/A	N/A	N/A*	3600	238.026	238.026	0.00	1695.7	238.026	238.026	0.00	1789.2	238.026	238.026	0.00	1463.3	238.026	238.797	0.32	3600
2601-3	N/A	N/A	N/A*	3600	237.783	237.783	0.00	3104.5	237.783	237.783	0.00	2498.2	237.783	237.783	0.00	1027.2	237.783	237.803	0.00	2623.2
2601-4	N/A	N/A	N/A*	3600	237.980	237.980	0.00	1414.6	237.980	237.980	0.00	2775.9	237.980	237.980	0.00	1020.6	237.980	248.841	4.37	3600
2601-5	N/A	N/A	N/A*	3600	237.942	237.942	0.00	2942.8	237.942	237.942	0.00	3288.7	237.942	237.942	0.00	1409.6	237.942	239.973	0.85	3600
2901-1	N/A	N/A	N/A*	3600	237.766	241.732	1.64	3600	237.766	241.732	1.64	3600	237.766	237.766	0.00	1319.4	237.766	238.497	0.31	3600
2901-2	N/A	N/A	N/A*	3600	238.103	238.103	0.00	3301.2	238.103	238.103	0.00	2699.7	238.103	238.103	0.00	1212.9	238.103	238.829	0.30	3600
2901-3	N/A	N/A	N/A*	3600	237.758	237.758	0.00	1762.0	237.758	237.758	0.00	3283.4	237.758	237.758	0.00	1547.2	237.758	238.662	0.38	3600
2901-4	N/A	N/A	N/A*	3600	237.959	N/A	N/A**	3600	237.959	N/A	N/A**	3600	237.959	237.959	0.00	1500.1	237.959	N/A	N/A**	3600
2901-5	N/A	N/A	N/A*	3600	237.819	238.980	0.49	3600	237.819	238.980	0.49	3600	237.819	237.823	0.00	1626.7	237.819	238.570	0.31	3600

Table 8.10: Detailed computational results of BP Algorithms (cont.).

	Gurobi				BP Algorithm 1				BP Algorithm 2				BP Algorithm 3				Impr. Benders Alg. 2			
INSTANCE	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU	LB	UB	GAP	CPU
3401-1	N/A	N/A	N/A*	3600	237.703	244.550	2.80	3600	237.703	244.550	2.80	3600	237.703	237.703	0.00	1837.2	237.703	238.801	0.46	3600
3401-2	N/A	N/A	N/A*	3600	238.198	238.198	0.00	2432.2	238.198	238.198	0.00	2433.3	238.198	238.198	0.00	2479.5	238.198	239.270	0.45	3600
3401-3	N/A	N/A	N/A*	3600	238.048	238.048	0.00	2633.1	238.048	238.048	0.00	2666.2	238.048	238.048	0.00	1920.0	238.048	238.767	0.30	3600
3401-4	N/A	N/A	N/A*	3600	237.789	237.794	0.00	2793.6	237.789	N/A	N/A**	3600	237.789	237.789	0.00	1903.9	237.789	N/A	N/A**	3600
3401-5	N/A	N/A	N/A*	3600	237.969	240.198	0.93	3600	237.969	240.198	0.93	3600	237.969	237.969	0.00	2067.7	237.969	239.490	0.64	3600
4501-1	N/A	N/A	N/A*	3600	N/A	N/A	N/A*	3600	N/A	N/A	N/A*	3600	237.861	237.861	0.00	3206.4	237.861	N/A	N/A**	3600
4501-2	N/A	N/A	N/A*	3600	N/A	N/A	N/A*	3600	N/A	N/A	N/A*	3600	238.001	238.001	0.00	3020.4	238.001	N/A	N/A**	3600
4501-3	N/A	N/A	N/A*	3600	N/A	N/A	N/A*	3600	237.943	N/A	N/A**	3600	237.943	N/A	N/A**	3600	237.943	N/A	N/A**	3600
4501-4	N/A	N/A	N/A*	3600	N/A	N/A	N/A*	3600	N/A	N/A	N/A*	3600	237.888	N/A	N/A**	3600	237.888	N/A	N/A**	3600
4501-5	N/A	N/A	N/A*	3600	N/A	N/A	N/A*	3600	N/A	N/A	N/A*	3600	238.039	238.040	0.00	3157.7	238.039	N/A	N/A**	3600

Note 1: LB and UB are in MU, GAP % and CPU in seconds.

Note 2: cells marked with * are accepted as 100%, and ** are calculated as 100(1800-LB)/1800% in Table 8.9

8.5. Computational Results for the Two-Phase Heuristic

We implement the heuristic in Python 2.7 programming language [101] and use Gurobi 8.0 as the MILP solver [104]. All tests are carried out on a 64-bit PC with 3.20 GHz Intel(R) Core(TM) i5-6500 CPU and 8 GB of RAM. We set the number of threads of Gurobi solver to 4. In CVaR parameter tuning operation we set ϵ_1 to 0.10 and ϵ_2 to 0.03. We keep the parameter tuning of column generation as in BP algorithms, which are explained at the beginning of Section 8.4.

In Istanbul University Oncology Institute, the VMAT plans of all patients are optimized using two full arcs on older versions of Eclipse TPS (v.8.9 and v.15.1, [93]) using 6 MV photon beams. In Table 8.11 total MUs and dosimetric results of all plans are provided. According to these results all VMAT plans satisfy all dose-volume constraints given in Table 8.4 (except the plan of patient 6, since it does not satisfy the first dose-volume constraint of rectum, which requires $D_{\%35} \leq 40$ Gy). Total radiation dose (sum of MUs of two arcs) of plans varies between 570 and 743 MUs with average 633.9 MUs. Table 8.12 provides dosimetric results of VMAT plans obtained by our column generation based heuristic algorithm. Almost all plans are optimized within 20 minutes (1200 seconds) with an average of 1020 seconds. It takes a little longer to optimize plan 3 and plan 6 (1227 and 1782 seconds, respectively). We first note that total radiation intensity decreases in almost all plans significantly (except plan 5). The amount of radiation dose varies between 366 and 689 MUs with an average of 494.4 MUs. The maximum reduction occurs for the plan of patient 7, which is 363 MUs (48.9%). The average decrease of all plans is approximately 139.5 MUs and the average percentage of reduction is 22.0%. By assuming that the dose-influence matrices obtained by AAA algorithm [22] and by the singular value decomposed pencil beam algorithm [95] used in matRad are sufficiently close, we can say that our proposed model and solution algorithm can find high quality plans requiring less radiation.

We observe that the first dose-volume constraint of rectum is not satisfied in the plans of three patients (patient 1, 6 and 8). However, the maximum deviation

Table 8.11. Dosimetric results of the VMAT plans obtained by Eclipse.

		PTV75.6	**R-PTV56**	**Rectum**		**Bladder**		**PB**	**Lt F**	**Rt F**
Patient	**MU**	$\mathbf{D_{95\%}}$	$\mathbf{D_{95\%}}$	$\mathbf{D_{35\%}}$	$\mathbf{D_{17\%}}$	$\mathbf{D_{50\%}}$	$\mathbf{D_{25\%}}$	$\mathbf{D_{90\%}}$	$\mathbf{D_{10\%}}$	$\mathbf{D_{10\%}}$
P1	591	76.6	56.9	39.9	50.9	32.9	42.6	10.7	26.5	25.8
P2	575	75.8	56.2	38.9	52.9	31.7	57.3	10.5	38.1	37.0
P3	743	76.4	56.5	34.2	48.7	22.0	57.9	8.6	22.0	18.4
P4	659	76.8	57.1	37.2	54.1	30.7	55.3	14.1	35.0	34.2
P5	661	75.9	56.5	38.4	56.7	37.7	63.5	6.0	37.8	37.8
P6	570	76.1	56.8	41.2	53.7	16.2	26.4	14.4	36.2	32.9
P7	667	76.2	56.9	37.7	48.8	5.5	14.9	5.3	24.8	27.5
P8	625	76.4	56.6	33.8	49.1	4.5	16.8	11.1	26.0	30.8
P9	614	76.0	56.7	38.8	55.2	30.6	62.8	8.0	34.9	39.4
Avg.	633.9									

Table 8.12. Dosimetric results of the VMAT plans obtained by two-phase heuristic.

		PTV75.6	**R-PTV56**	**Rectum**		**Bladder**		**PB**	**Lt F**	**Rt F**
Patient	**MU**	$\mathbf{D_{95\%}}$	$\mathbf{D_{95\%}}$	$\mathbf{D_{35\%}}$	$\mathbf{D_{17\%}}$	$\mathbf{D_{50\%}}$	$\mathbf{D_{25\%}}$	$\mathbf{D_{90\%}}$	$\mathbf{D_{10\%}}$	$\mathbf{D_{10\%}}$
P1	526	76.7	57.3	42.0	51.0	27.2	46.3	24.3	40.9	43.9
P2	561	77.2	56.8	38.1	53.7	36.5	63.3	25.1	34.3	33.5
P3	380	76.4	56.9	36.1	50.8	34.7	64.6	12.1	42.3	37.3
P4	523	77.1	57.2	37.2	48.9	34.2	64.3	14.2	45.1	48.0
P5	689	76.2	58.2	38.9	57.5	39.1	64.6	6.7	41.1	43.5
P6	424	77.8	57.3	41.0	57.8	7.4	27.8	20.4	46.2	45.8
P7	366	76.2	57.1	38.3	50.9	3.3	16.9	7.4	35.1	36.1
P8	509	77.8	58.0	42.4	61.3	6.4	25.8	27.4	26.5	22.5
P9	472	75.9	57.1	34.9	54.0	37.8	64.5	12.8	46.7	46.8
Avg.	494.4									

from the tolerance dose, which is 40 Gy, is 2.4 Gy. Note also that if we decrease the radiation intensities at all control points by the same ratio without violating partial volume constraints of PTVs, then the resulting deviations will be less. For example, in plan 6 it is possible to reduce the radiation intensities at all control points to 97.73% of the original intensities. Therefore, 95% of R-PTV56 will receive 56 Gy and 95% of PTV75.6 will receive 76 Gy. The resulting plan almost satisfies all dose constraints of rectum ($D_{\%35}$ will be 40.07 Gy). By this way total MUs of the plan also decreases by around 9.6 MUs. Similarly, we can adjust plan 1 and plan 8, and reduce the $D_{\%35}$ of rectum to 41.4 Gy and 41.2 Gy, respectively. We should be careful when we are shifting the plans, since it will reduce also the minimum dose to PTVs and increase the risk of occurring cold spots. Moreover, in four out of nine plans (patient 1, 2, 6, and 8) $D_{\%90}$ of PB is more than 15 Gy (the maximum is 27.4 Gy), which are acceptable according to the oncologists and medical physicists at Istanbul University Oncology Institute and also other dose prescription recommendations (for example to limit $D_{\%70}$ and $D_{\%90}$ of PB to 70 Gy and 50 Gy, respectively) in the literature. Thus, we can say that the heuristic is capable to obtain high-quality VMAT plans with significantly fewer MUs in clinically reasonable times. In general, DVHs are used to evaluate the quality of a treatment plan. For a given structure, a DVH specifies the percentage of its volume that absorbs at least a certain amount of dose. We calculate DVHs of PTVs and OARs and compare them to the clinical guidelines and also to the ones obtained in the institute. We provide DVHs of all patients obtained by our algorithm and by Eclipse in Figure 8.1 – Figure 8.18 sequentially.

Finally, VMAT plans are made by the experienced dosimetrists in treatment planning departments. The planning process involves various manual interventions such as adapting planning objectives and constraints according to the individual anatomy of the patient. For example, shape and size of the tumor(s), and location of organs at risk are some of the anatomical properties that play an important role in the manual adjustments of the parameters, which influence the plan quality. Namely, the dosimetrists try to guide the treatment planning system towards a favorable plan by modifying optimization parameters. Thus, this manual process necessitates additional optimization steps and extra time, and also the quality of the final plan depends on the skills and

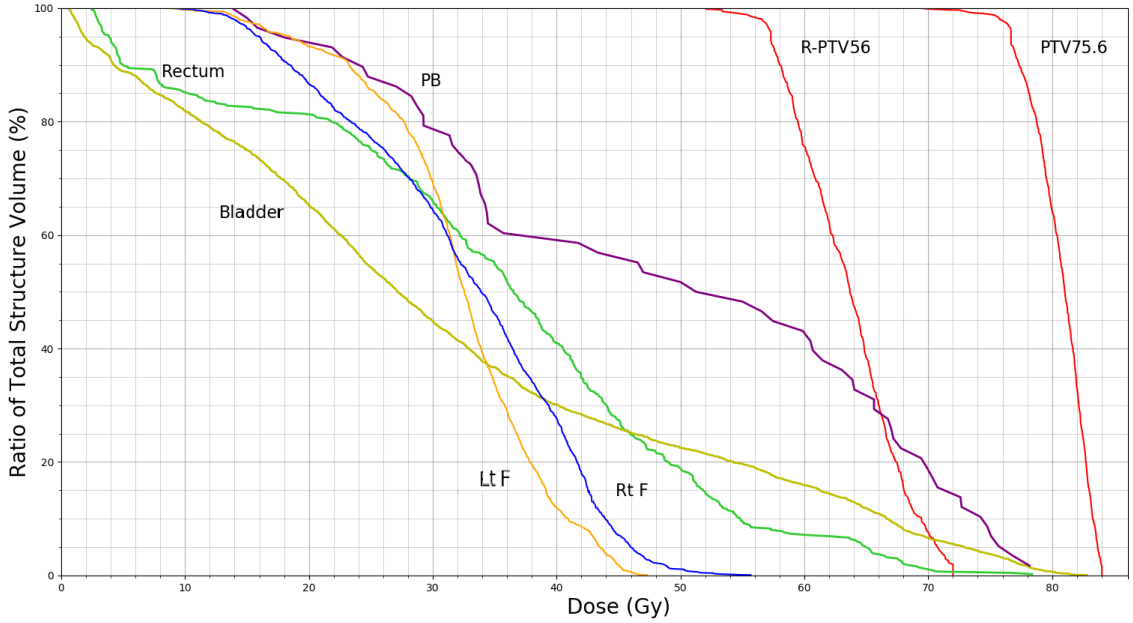

Figure 8.1. DVHs of the plan of patient 1 obtained by two-phase heuristic.

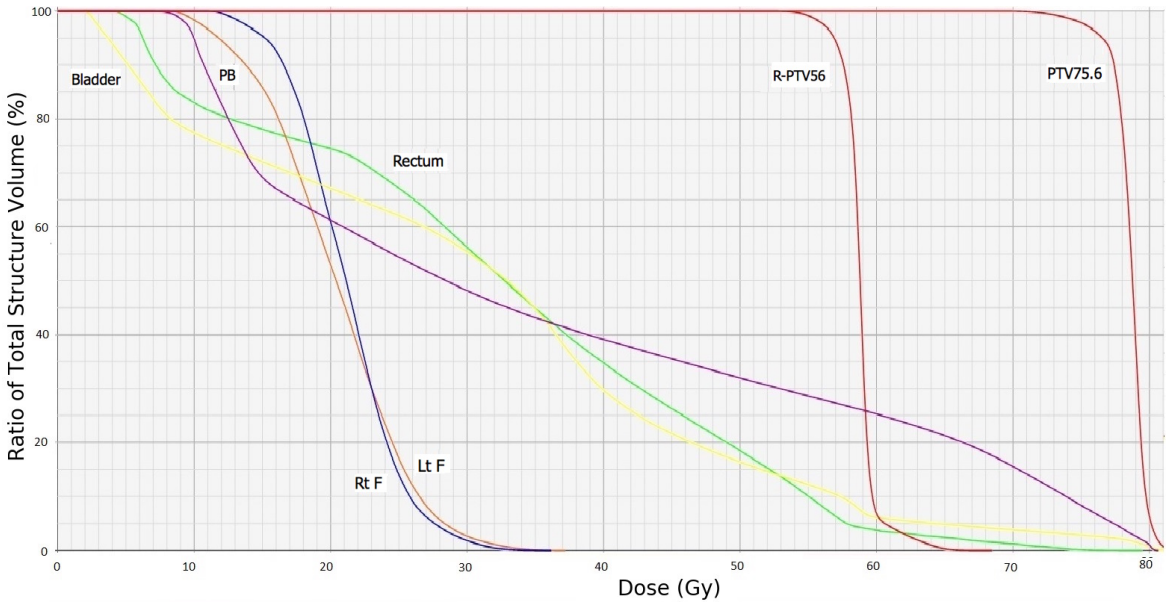

Figure 8.2. DVHs of the plan of patient 1 obtained by Eclipse v.15.1.

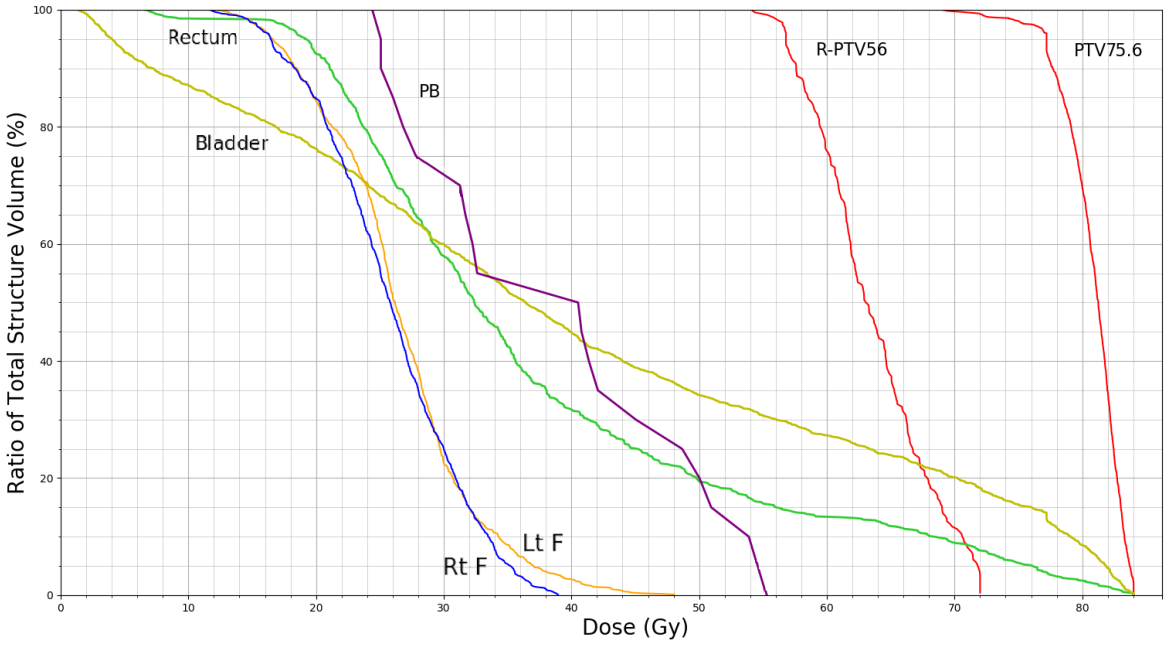

Figure 8.3. DVHs of the plan of patient 2 obtained by two-phase heuristic.

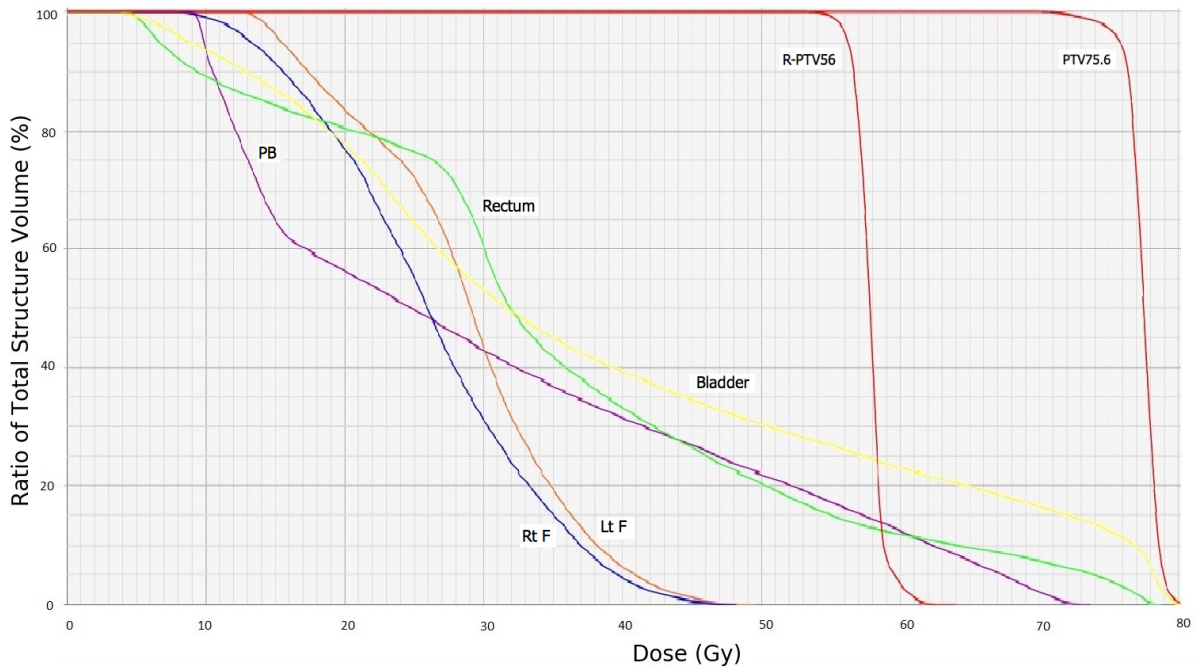

Figure 8.4. DVHs of the plan of patient 2 obtained by Eclipse v.15.1.

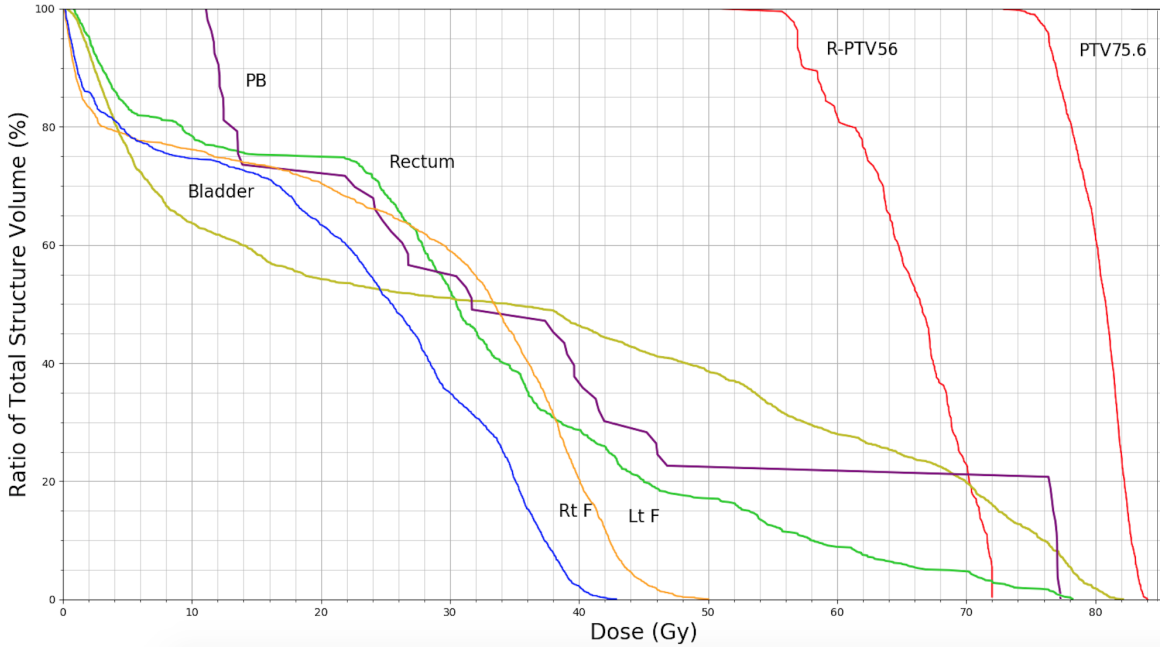

Figure 8.5. DVHs of the plan of patient 3 obtained by two-phase heuristic.

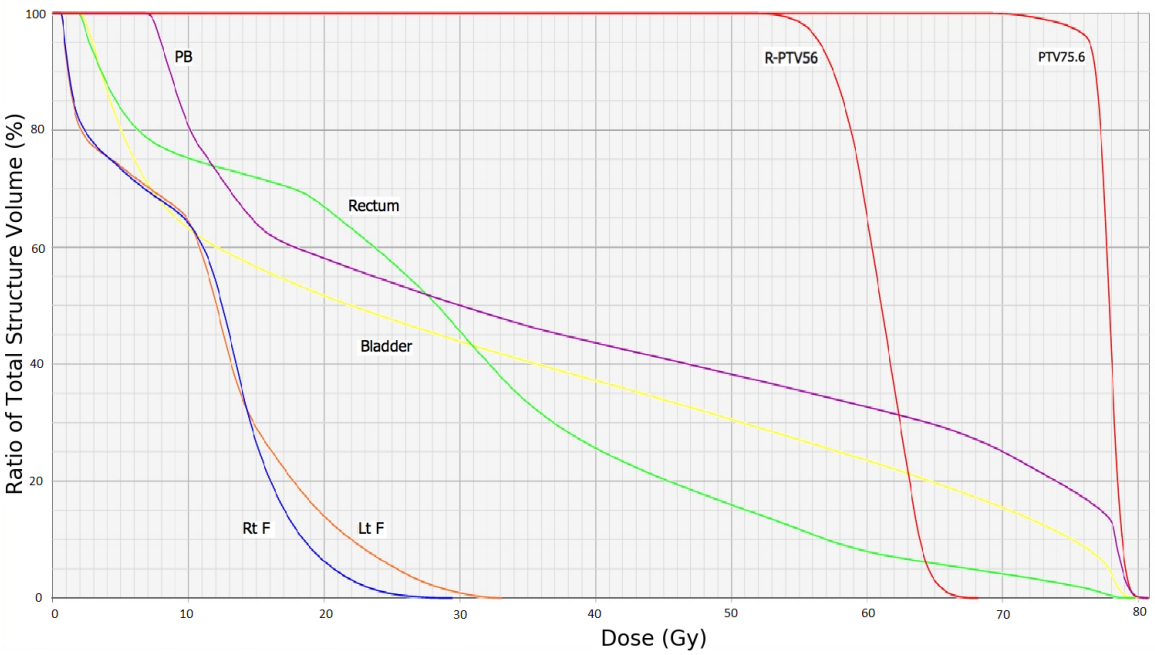

Figure 8.6. DVHs of the plan of patient 3 obtained by Eclipse v.15.1.

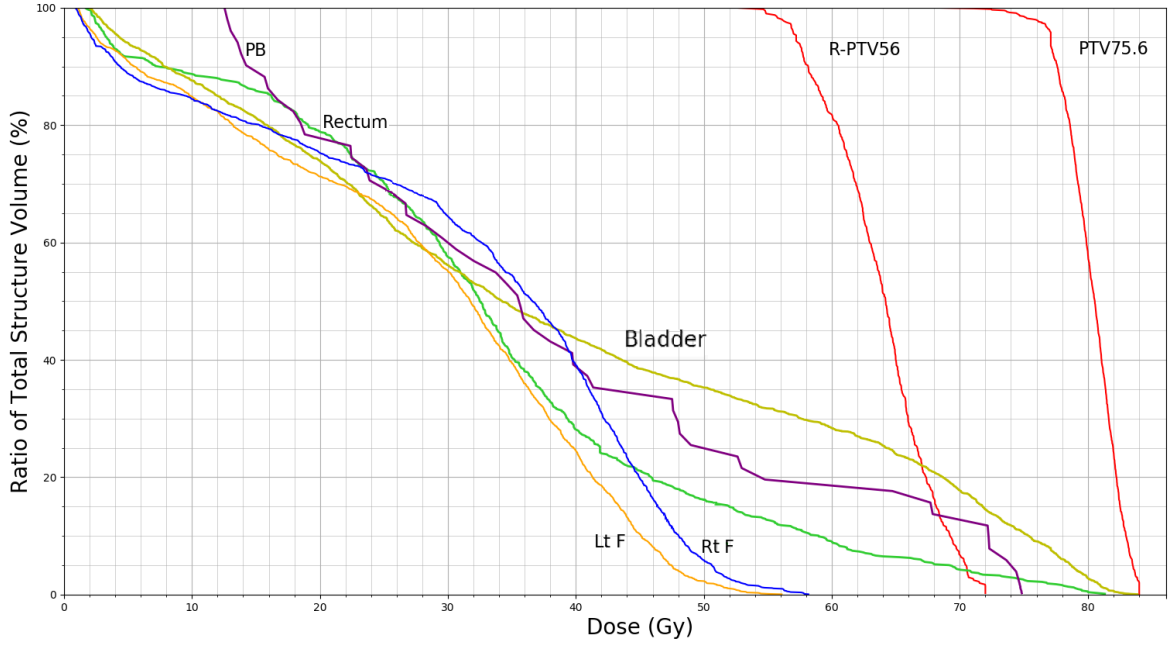

Figure 8.7. DVHs of the plan of patient 4 obtained by two-phase heuristic.

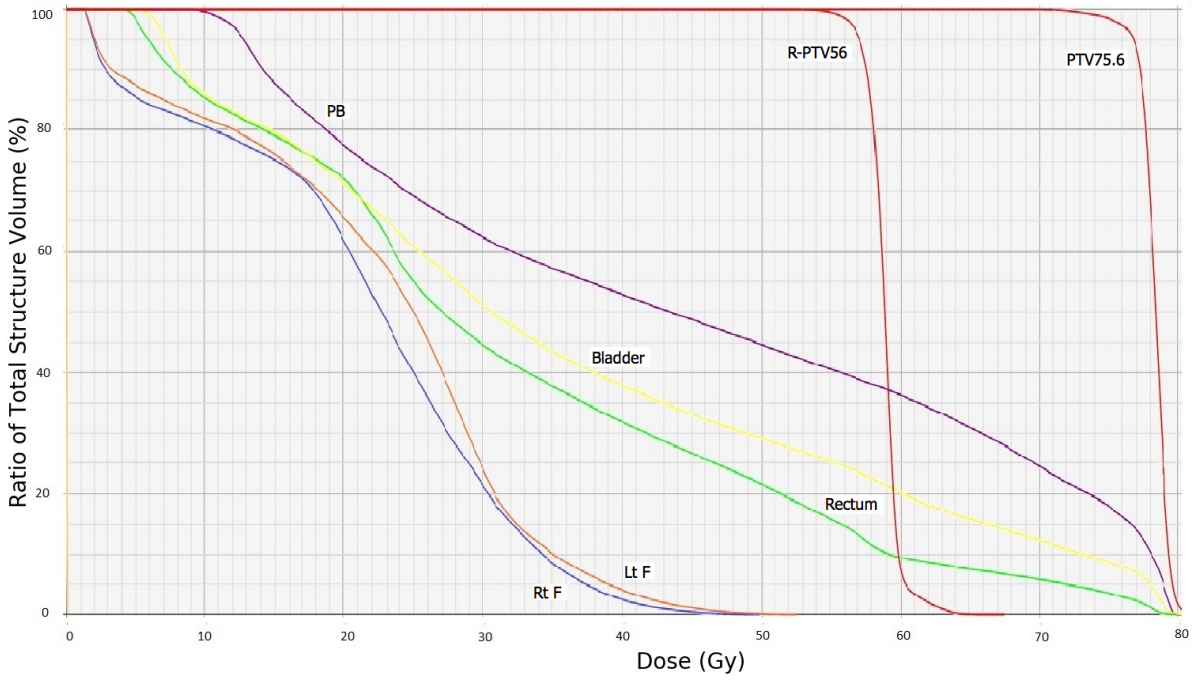

Figure 8.8. DVHs of the plan of patient 4 obtained by Eclipse v.15.1.

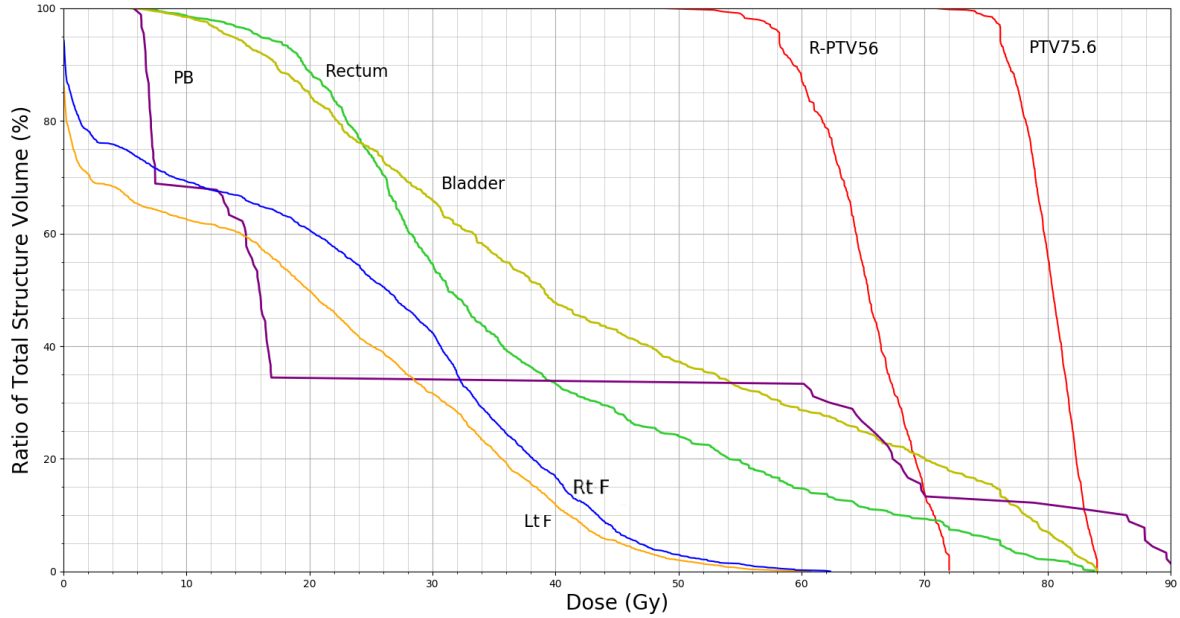

Figure 8.9. DVHs of the plan of patient 5 obtained by two-phase heuristic.

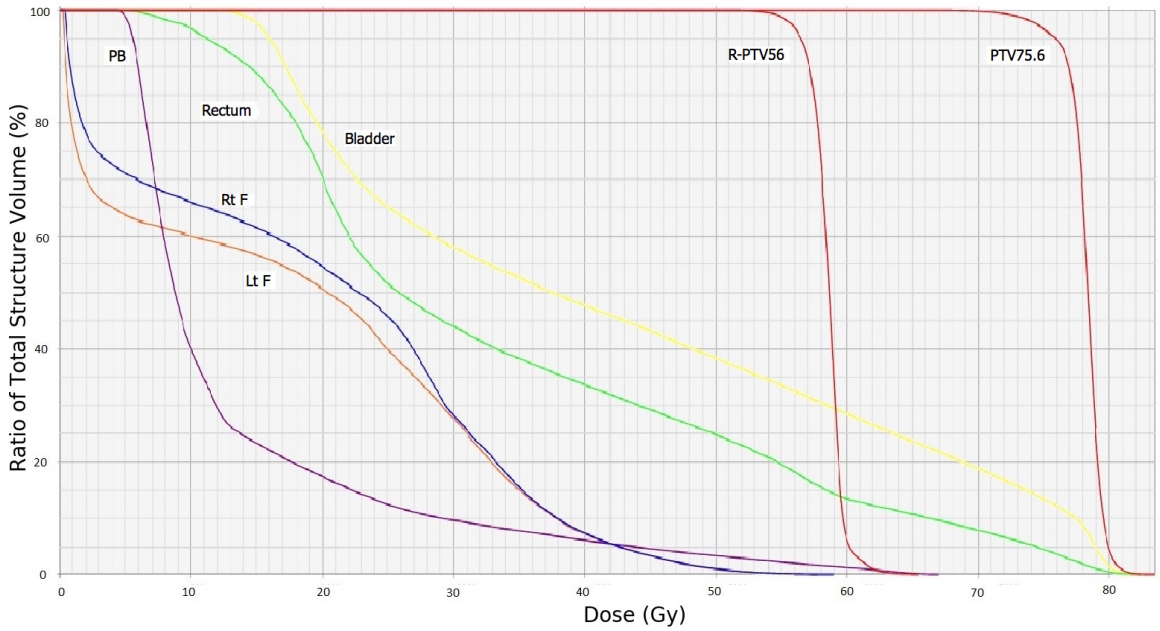

Figure 8.10. DVHs of the plan of patient 5 obtained by Eclipse v.15.1.

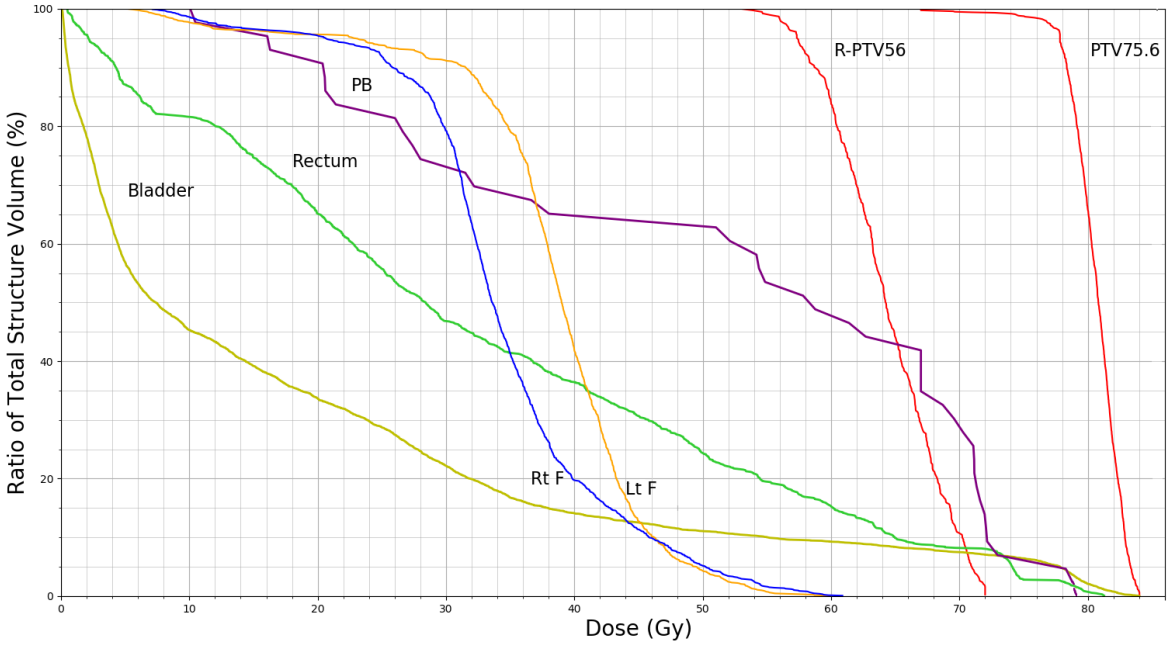

Figure 8.11. DVHs of the plan of patient 6 obtained by two-phase heuristic.

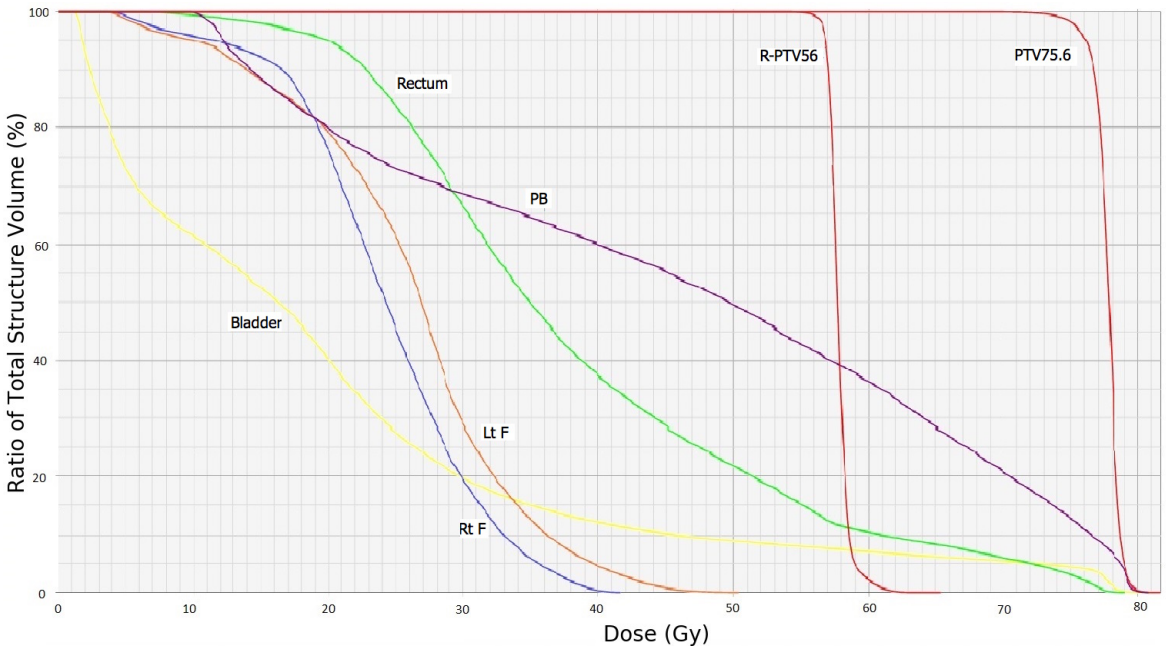

Figure 8.12. DVHs of the plan of patient 6 obtained by Eclipse v.15.1.

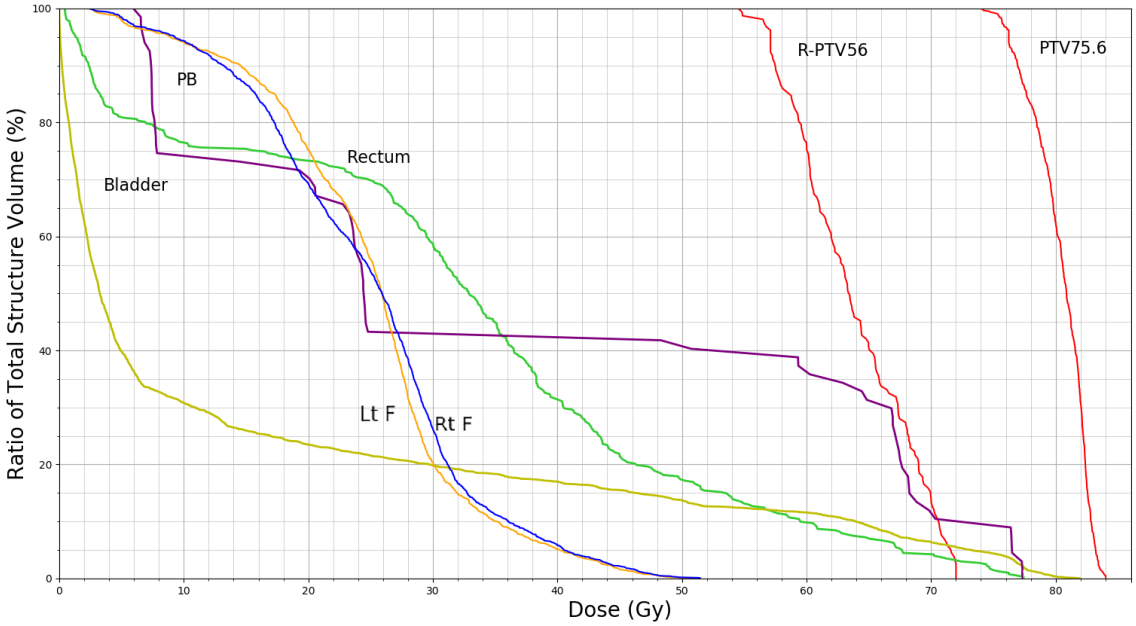

Figure 8.13. DVHs of the plan of patient 7 obtained by two-phase heuristic.

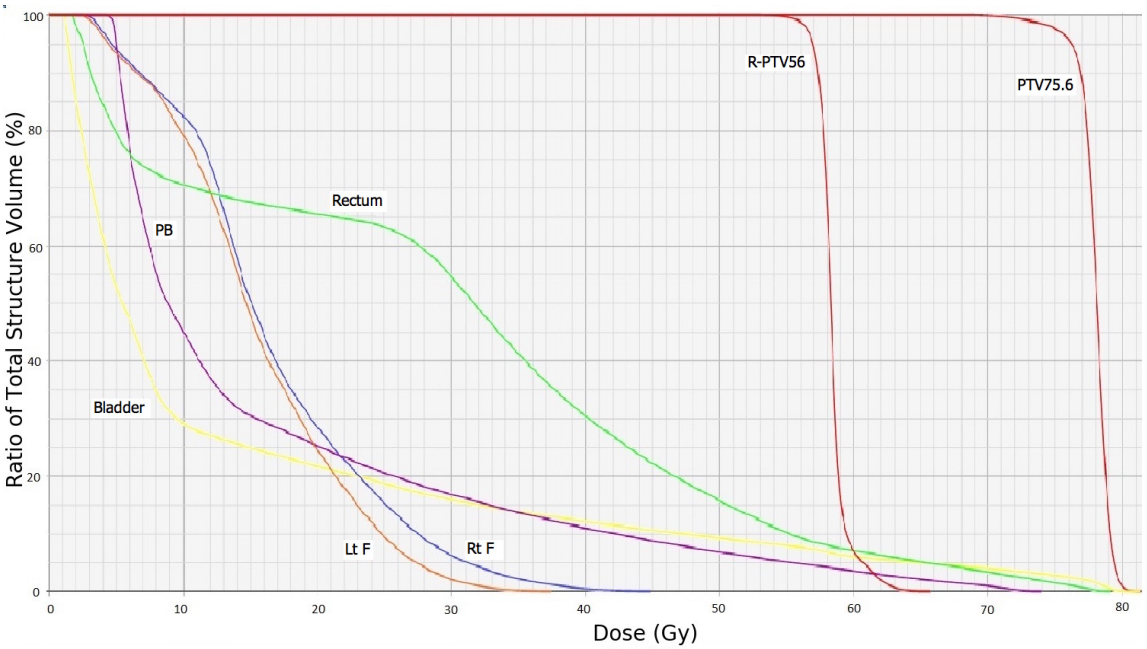

Figure 8.14. DVHs of the plan of patient 7 obtained by Eclipse v.15.1.

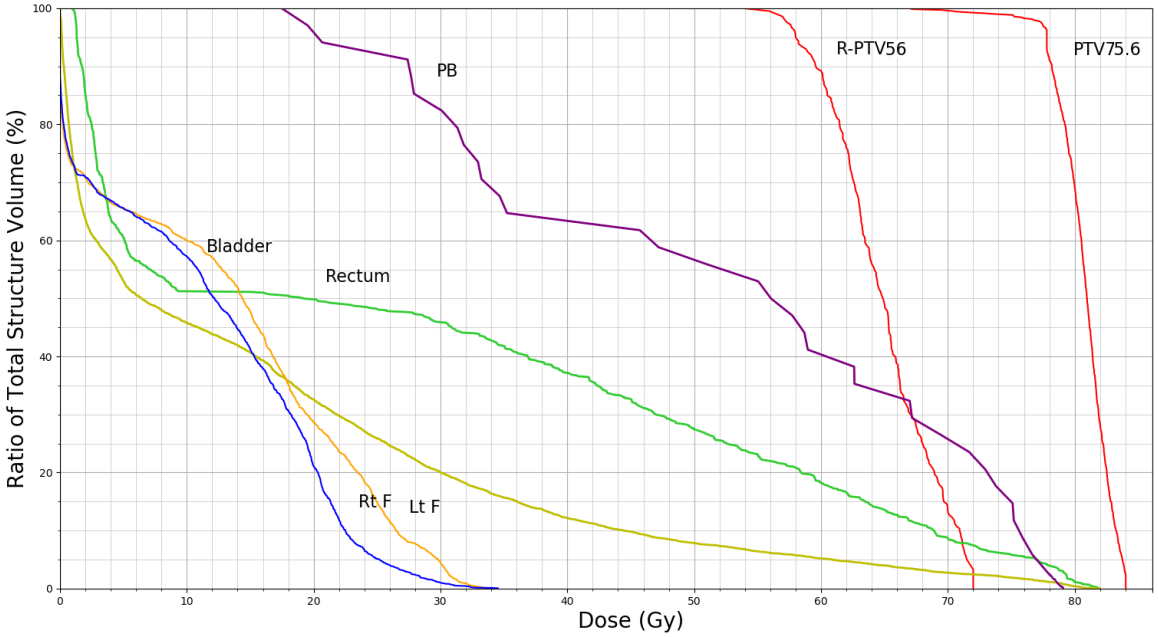

Figure 8.15. DVHs of the plan of patient 8 obtained by two-phase heuristic.

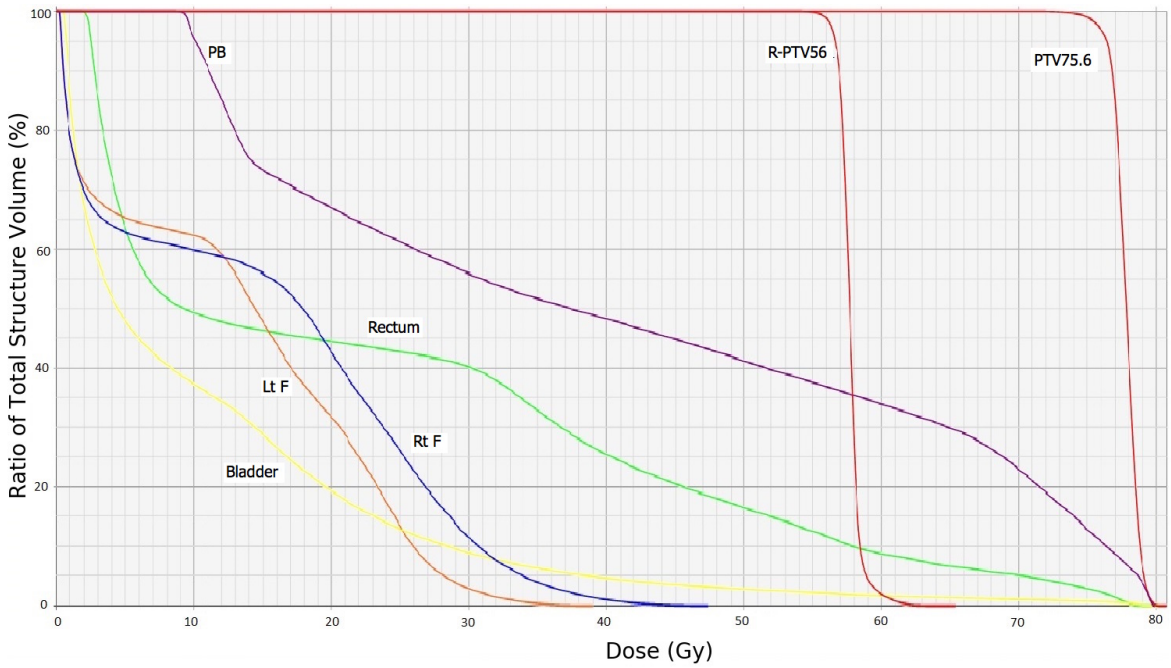

Figure 8.16. DVHs of the plan of patient 8 obtained by Eclipse v.15.1.

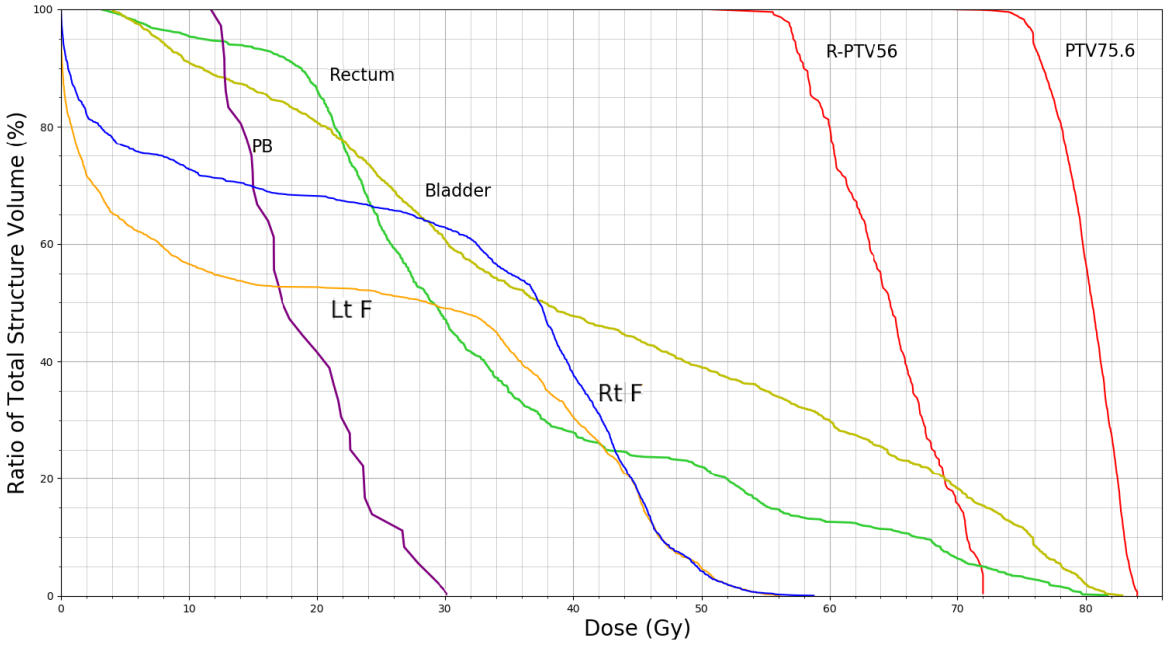

Figure 8.17. DVHs of the plan of patient 9 obtained by two-phase heuristic.

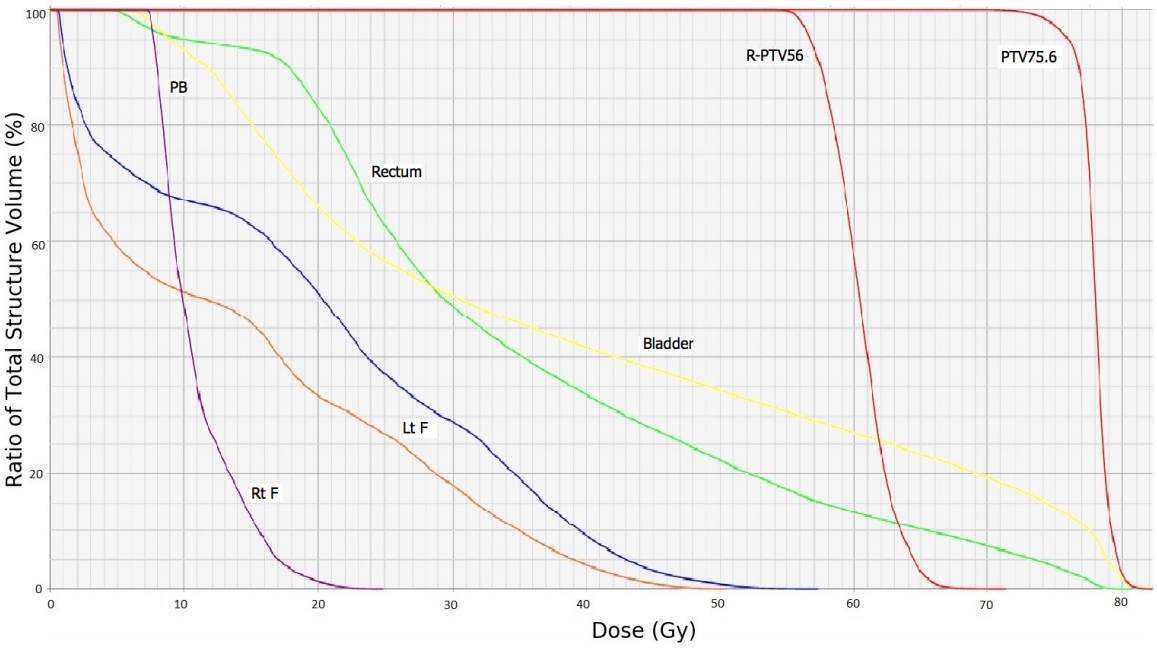

Figure 8.18. DVHs of the plan of patient 9 obtained by Eclipse v.15.1.

experience of the dosimetrist as well as the complexity of the case and time allocated for planning. The VMAT plans whose dosimetric results are shown in Table 8.11 are obtained and adjusted by an experienced dosimetrist via a manual process as explained above. In the original data sets there are some additional structures (e.g. a subset of rectum) for which the dosimetrist defines additional constraints (in 7 out of 9 plans) to ensure that the related received radiation amounts fall into approvable limits. However, we do not use such an additional structure and/or additional dose-constraint in our algorithm, which automatically adjusts parameters for each patient and does not require any expert guidance.

8.5.1. The Effect of Initial Columns

We analyze the effect of starting with initial columns generated from the fluence maps obtained by solving M-VMATP model. We generate VMAT plans for all patients using two different set of initial columns: the columns generated from a full treatment arc with maximum number of open beamlets and from a randomly generated full treatment arc. These new initial columns are also feasible with respect to the MLC constraints (i.e. satisfying the consecutive ones property and the leaf motion limitations). Also, random columns are generated from a treatment plan satisfying the full volume constraints of all target voxels, which is obtained by solving a model including all geometric constrains and also full volume constraints. In Table 8.13 and Table 8.14 we give the dosimetric results of the plans obtained by starting with columns having maximum number of open beamlets and with randomly generated ones, respectively. According to the results, none of the plans obtained using new initial columns are clinically acceptable. The average CPU time and total radiation decrease to 548.4 seconds and 347.2 MUs in the plans initial columns with maximum number of open beamlets. The average of total radiation of the plans with randomly generated columns slightly decreases to 469.9 MUs, however the average CPU time increases. These results show that starting with initial columns generated from the fluence maps improves the performance of the algorithm in terms of clinical dose requirements.

Table 8.13. Dosimetric results for the initial columns with maximum open beamlets.

			PTV75.6	**R-PTV56**	**Rectum**		**Bladder**		**PB**	**Lt F**	**Rt F**
Ptnt	**CPU**	**MU**	$\mathbf{D_{95\%}}$	$\mathbf{D_{95\%}}$	$\mathbf{D_{35\%}}$	$\mathbf{D_{17\%}}$	$\mathbf{D_{50\%}}$	$\mathbf{D_{25\%}}$	$\mathbf{D_{90\%}}$	$\mathbf{D_{10\%}}$	$\mathbf{D_{10\%}}$
P1	569	346	71.7	65.2	62.0	67.7	39.7	50.7	31.2	48.9	48.4
P2	621	331	69.6	63.3	57.7	64.6	41.0	68.1	42.1	53.6	55.5
P3	427	343	71.3	66.4	50.6	66.6	34.3	68.4	22.1	55.2	54.0
P4	545	341	71.8	63.6	59.2	66.8	35.5	65.6	34.1	52.0	50.2
P5	491	355	74.1	59.6	62.7	70.0	64.6	72.2	6.5	46.2	49.2
P6	646	370	74.0	66.0	42.7	58.9	30.3	43.2	25.0	48.7	49.4
P7	286	361	74.2	62.9	49.5	61.2	3.2	14.7	7.3	59.2	47.3
P8	284	356	71.6	64.1	35.8	64.1	4.3	40.9	38.6	48.2	47.6
P9	1067	322	68.2	64.9	53.0	65.6	49.4	65.3	19.2	48.1	45.8
Avg.	548.4	347.2									

Table 8.14. Dosimetric results for randomly generated initial columns.

			PTV75.6	**R-PTV56**	**Rectum**		**Bladder**		**PB**	**Lt F**	**Rt F**
Ptnt	**CPU**	**MU**	$\mathbf{D_{95\%}}$	$\mathbf{D_{95\%}}$	$\mathbf{D_{35\%}}$	$\mathbf{D_{17\%}}$	$\mathbf{D_{50\%}}$	$\mathbf{D_{25\%}}$	$\mathbf{D_{90\%}}$	$\mathbf{D_{10\%}}$	$\mathbf{D_{10\%}}$
P1	1555	466	75.4	56.8	50.6	57.6	31.5	54.0	26.3	41.0	44.0
P2	1580	488	72.4	56.7	40.9	53.2	30.3	64.4	29.5	52.0	53.3
P3	1369	474	76.5	56.9	35.3	49.5	17.4	64.6	12.2	45.2	57.6
P4	1385	477	76.5	56.8	46.3	53.7	32.8	63.5	14.0	43.1	49.3
P5	1505	494	76.2	56.7	51.5	61.5	50.7	67.4	6.5	45.8	34.8
P6	2559	427	77.5	58.1	42.9	58.6	9.1	28.2	19.1	48.0	42.7
P7	868	423	76.4	57.1	43.4	51.2	3.1	13.1	6.2	35.7	47.3
P8	1214	475	77.9	57.3	42.3	61.2	4.6	23.0	19.1	33.3	43.1
P9	1829	505	73.5	57.7	42.9	54.3	39.6	63.1	13.4	45.3	45.9
Avg.	1540.4	469.9									

8.5.2. The Effect of Parameter Tuning

We repeat the computational experiments for all patients without CVaR parameter tuning operation. In Table 8.15, we provide the dosimetric results, and CPU time and MUs of the plans. According to the results the average CPU time decreases by 130.4 seconds as compared the original ones provided in Table 8.12, which is consistent with our expectation. However, CPU time of the plans for patient 1 and patient 8 increases. Since, the necessary time for the second phase (the column generation phase) differ when CVaR tuning operation is not applied. The average MUs of the plans increases by 76.4 MUs, which is also expected, since CVaR constraints are tighter. All partial volume constraints of OARs in all plans are satisfied (due to tighter CVaR constraints). However, there are 3 plans (patient 2, patient 4 and patient 5) where the partial volume constraints of PTV75.6 are not satisfied. To sum, tuning the right hand side parameters of CVaR constraints makes the resulting plans more reasonable without compromising OAR's partial volume constraints. Also, all PTV constraints are satisfied with fewer MUs (there is a decrease around 13.3%). In detail, the average necessary time for tuning operation is 138.5 seconds, which is not very long.

Table 8.15. Dosimetric results of the VMAT plans without CVaR tuning operation.

			PTV75.6	**R-PTV56**	**Rectum**		**Bladder**		**PB**	**Lt F**	**Rt F**
Ptnt	**CPU**	**MU**	$\mathbf{D_{95\%}}$	$\mathbf{D_{95\%}}$	$\mathbf{D_{35\%}}$	$\mathbf{D_{17\%}}$	$\mathbf{D_{50\%}}$	$\mathbf{D_{25\%}}$	$\mathbf{D_{90\%}}$	$\mathbf{D_{10\%}}$	$\mathbf{D_{10\%}}$
P1	1232	633	76.6	57.8	35.1	45.6	21.2	41.6	20.1	39.7	47.2
P2	556	623	74.3	56.7	28.4	46.1	26.3	55.7	23.8	43.8	40.3
P3	854	513	75.8	56.8	26.4	41.9	13.0	46.2	5.9	42.8	38.4
P4	832	683	75.0	57.5	28.2	44.5	19.6	46.8	8.1	38.1	36.6
P5	713	708	74.7	58.6	33.4	54.7	37.8	62.5	5.0	38.9	42.9
P6	1405	452	77.5	57.7	26.3	44.1	8.6	23.1	19.1	46.8	47.2
P7	658	439	76.2	56.5	25.6	41.1	3.1	12.9	4.6	40.6	45.3
P8	744	526	78.0	56.8	18.4	45.3	4.0	18.4	15.5	47.0	35.3
P9	1014	560	76.2	58.6	27.8	48.6	29.6	56.6	9.2	45.7	46.5
Avg.	889.8	570.8									

8.5.3. Comparing the Performance of Two-Phase Heuristic with Exact Solution Algorithms

We solve all instances, which are explained in the first test environment section (Section 8.1.1) and used in the computational experiments of the exact solution algorithms (in Chapter 5 and Chapter 6), by two-phase heuristic. The first phase of the heuristic tunes the tolerance doses of OARs in CVaR constraints, thus we perform the computational experiments for both cases: without tuning operation and with tuning operation. In Table 8.16 we give the average results of all samples. Note that we do not obtain a LB when we solve an instance by two-phase heuristic, thus we use the best available LB (i.e. the maximum of LBs obtained by Gurobi, Benders and BP algorithms) in order to calculate optimality gaps. In Table 8.17 we provide the best LB for each instance. Note also that when we perform tuning operation the tolerance dose of OAR may be increased, namely the problem may be simplified, thus the LB of the resulting model may be decreased. Nevertheless, we give the average approximated optimality gaps in Table 8.16 using the same best LB. In Table 8.17, for each instance, we provide UB and CPU time obtained by the heuristic without and with tuning operation. There are also optimality GAPs for the heuristic without tuning operation. According to the results CPU times of all instances remarkably decrease. In both cases, the average CPU time is around three minutes (182.6 and 170.8 seconds for the computational experiments without tuning and with tuning operation, respectively). Two-phase heuristic can find a feasible solution with small optimality gap for all instances in both cases, however they can solve only 3 and 8 out of 90 instances optimally. We should note that partial volume constraint of OAR is satisfied for each instance when tuning operation is applied ($D^{OAR}_{\%60}$ ranges from 1.50 Gy and 23.6 Gy, where the treatment prescriptions is $D^{OAR}_{\%60} = 50$ Gy).

Table 8.16. Summary of the computational results of CORT dataset.

	Without CVaR tuning				With CVaR tuning			
SAMPLE	GAP	CPU	S/T	O/T	GAP*	CPU	S/T	O**/T
22	3.09	30.1	5/5	0/5	2.86	28.9	5/5	0/5
44	0.30	29.0	5/5	0/5	0.46	29.3	5/5	0/5
66	4.69	31.6	5/5	0/5	4.75	39.7	5/5	0/5
88	3.61	49.7	5/5	0/5	2.38	47.7	5/5	0/5
220	4.01	31.8	5/5	0/5	2.48	39.6	5/5	0/5
660	0.04	53.6	5/5	0/5	0.05	56.1	5/5	1/5
880	0.07	66.8	5/5	1/5	0.09	68.0	5/5	1/5
1100	0.06	79.0	5/5	1/5	0.06	90.3	5/5	2/5
1301	0.16	96.6	5/5	0/5	0.08	112.8	5/5	0/5
1501	0.18	120.6	5/5	0/5	0.12	127.1	5/5	1/5
1701	0.18	131.6	5/5	0/5	0.07	154.8	5/5	1/5
1901	0.15	184.2	5/5	0/5	0.12	154.0	5/5	0/5
2101	0.05	203.0	5/5	0/5	0.03	170.5	5/5	1/5
2301	0.09	247.1	5/5	0/5	0.04	230.2	5/5	0/5
2601	0.15	288.9	5/5	0/5	0.21	241.8	5/5	0/5
2901	0.12	346.1	5/5	0/5	0.08	325.1	5/5	0/5
3401	0.05	484.9	5/5	1/5	0.06	417.7	5/5	1/5
4501	0.10	812.5	5/5	0/5	0.07	741.2	5/5	0/5
Avg/Sum	**0.95**	**182.6**	**55/55**	**3/55**	**0.78**	**170.8**	**55/55**	**8/55**

Note: Cells marked with * are calculated using the best LB, thus they are approximated GAPs. Also, in the cells marked with ** the same best LBs are considered.

Table 8.17: Detailed computational results of CORT dataset.

		Without CVaR tuning			With CVaR tuning					Without CVaR tuning			With CVaR tuning		
SAMPLE	Best LB	UB	CPU	GAP	UB	CPU	GAP	SAMPLE	Best LB	UB	CPU	GAP	UB	CPU	GAP
22-1	236.550	261.782	41.3	9.64	261.268	30.9	N/A	**1501-1**	237.770	237.852	130.7	0.03	237.842	109.5	N/A
22-2	231.148	236.162	35.6	2.12	235.847	32.3	N/A	**1501-2**	237.459	237.742	127.7	0.12	237.571	122.1	N/A
22-3	232.605	235.726	28.9	1.32	235.115	22.1	N/A	**1501-3**	237.925	238.319	107.3	0.17	238.319	120.7	N/A
22-4	233.700	234.225	24.4	0.22	234.087	27.0	N/A	**1501-4**	237.925	237.954	106.5	0.01	237.937	108.4	N/A
22-5	234.742	239.921	20.2	2.16	238.591	32.0	N/A	**1501-5**	237.665	239.030	130.9	0.57	238.447	175.1	N/A
44-1	232.912	233.357	25.3	0.19	233.261	27.4	N/A	**1701-1**	238.008	238.057	122.0	0.02	238.020	139.0	N/A
44-2	236.830	237.200	43.1	0.16	239.563	39.2	N/A	**1701-2**	237.823	238.041	124.1	0.09	238.041	139.4	N/A
44-3	232.903	233.145	21.6	0.10	233.145	24.1	N/A	**1701-3**	238.105	239.775	155.2	0.70	238.544	205.6	N/A
44-4	236.838	238.672	29.8	0.77	238.318	27.7	N/A	**1701-4**	237.896	238.029	129.5	0.06	237.972	145.0	N/A
44-5	236.429	237.149	25.0	0.30	237.149	28.0	N/A	**1701-5**	237.677	237.739	127.1	0.03	237.738	144.9	N/A
66-1	236.830	268.682	30.9	11.85	268.682	34.3	N/A	**1901-1**	237.693	237.811	197.6	0.05	237.779	161.1	N/A
66-2	237.597	239.140	39.8	0.64	239.140	44.7	N/A	**1901-2**	237.709	238.530	189.0	0.34	238.469	153.9	N/A
66-3	236.451	248.569	21.8	4.88	248.569	24.5	N/A	**1901-3**	237.829	238.089	181.7	0.11	238.052	177.0	N/A
66-4	234.940	241.028	30.2	2.53	241.760	55.4	N/A	**1901-4**	237.662	238.058	187.3	0.17	237.935	145.3	N/A
66-5	237.732	246.528	35.4	3.57	246.528	39.5	N/A	**1901-5**	238.216	238.414	165.3	0.08	238.350	132.8	N/A

Table 8.17: Detailed computational results of CORT dataset (cont.).

		Without CVaR tuning			With CVaR tuning					Without CVaR tuning			With CVaR tuning		
SAMPLE	Best LB	UB	CPU	GAP	UB	CPU	GAP	SAMPLE	Best LB	UB	CPU	GAP	UB	CPU	GAP
88-1	236.830	240.209	60.7	1.41	237.799	49.8	N/A	**2101-1**	237.786	237.867	189.5	0.03	237.867	168.2	N/A
88-2	237.050	240.299	35.5	1.35	240.299	39.3	N/A	**2101-2**	237.622	237.748	198.9	0.05	237.713	180.1	N/A
88-3	236.645	243.248	21.6	2.71	241.777	24.3	N/A	**2101-3**	237.770	237.912	215.9	0.06	237.886	175.9	N/A
88-4	237.048	269.830	53.1	12.15	256.683	43.7	N/A	**2101-4**	237.812	237.937	228.7	0.05	237.827	166.1	N/A
88-5	235.866	236.861	77.5	0.42	236.724	81.2	N/A	**2101-5**	237.858	238.009	182.0	0.06	237.955	161.9	N/A
220-1	236.864	237.209	31.5	0.15	237.209	35.4	N/A	**2301-1**	237.970	238.100	240.2	0.05	238.008	222.7	N/A
220-2	237.067	237.945	27.5	0.37	237.522	30.7	N/A	**2301-2**	238.049	238.152	237.3	0.04	238.154	229.0	N/A
220-3	237.924	267.428	35.1	11.03	246.596	59.0	N/A	**2301-3**	237.679	237.991	250.5	0.13	237.819	257.1	N/A
220-4	237.028	249.490	35.2	4.99	249.490	39.7	N/A	**2301-4**	237.593	237.904	225.3	0.13	237.763	210.6	N/A
220-5	237.872	246.575	29.7	3.53	246.575	33.3	N/A	**2301-5**	238.164	238.322	282.1	0.07	238.210	231.3	N/A
660-1	237.666	237.693	49.0	0.01	237.693	55.4	N/A	**2601-1**	237.826	237.956	273.9	0.05	237.882	247.6	N/A
660-2	237.847	237.920	45.1	0.03	237.920	51.1	N/A	**2601-2**	238.026	238.091	296.7	0.03	238.417	244.5	N/A
660-3	237.349	237.414	46.5	0.03	237.414	55.9	N/A	**2601-3**	237.783	239.028	304.2	0.52	239.724	241.1	N/A
660-4	238.339	238.455	56.0	0.05	238.343	46.8	N/A	**2601-4**	237.980	238.107	277.6	0.05	238.092	238.1	N/A
660-5	237.869	238.019	71.5	0.06	238.293	71.2	N/A	**2601-5**	237.942	238.120	292.2	0.07	238.013	237.6	N/A

Table 8.17: Detailed computational results of CORT dataset (cont.).

		Without CVaR tuning			With CVaR tuning					Without CVaR tuning			With CVaR tuning		
SAMPLE	Best LB	UB	CPU	GAP	UB	CPU	GAP	SAMPLE	Best LB	UB	CPU	GAP	UB	CPU	GAP
880-1	238.177	238.481	71.2	0.13	238.481	75.5	N/A	**2901-1**	237.766	237.853	333.1	0.04	237.839	304.4	N/A
880-2	237.412	237.431	55.8	0.01	237.428	55.7	N/A	**2901-2**	238.103	238.370	330.3	0.11	238.211	300.6	N/A
880-3	238.357	238.673	91.9	0.13	238.875	78.6	N/A	**2901-3**	237.758	237.904	281.4	0.06	237.817	260.9	N/A
880-4	237.214	237.322	58.9	0.05	237.322	67.0	N/A	**2901-4**	237.959	238.131	345.3	0.07	238.362	351.4	N/A
880-5	238.006	238.086	56.0	0.03	238.086	63.3	N/A	**2901-5**	237.819	238.574	440.7	0.32	238.122	408.5	N/A
1100-1	238.082	238.086	71.2	0.00	238.086	81.2	N/A	**3401-1**	237.703	237.710	432.0	0.00	237.710	387.0	N/A
1100-2	237.714	237.991	107.2	0.12	237.991	122.7	N/A	**3401-2**	238.198	238.271	481.3	0.03	238.255	409.2	N/A
1100-3	237.165	237.461	80.7	0.12	237.461	93.0	N/A	**3401-3**	238.048	238.192	489.7	0.06	238.192	416.7	N/A
1100-4	237.321	237.405	70.8	0.04	237.405	81.2	N/A	**3401-4**	237.789	238.132	500.8	0.14	238.021	424.5	N/A
1100-5	237.801	237.831	65.2	0.01	237.809	73.2	N/A	**3401-5**	237.969	238.035	520.6	0.03	238.272	451.0	N/A
1301-1	237.785	238.743	99.0	0.40	238.101	137.5	N/A	**4501-1**	237.861	238.133	960.3	0.11	238.016	853.4	N/A
1301-2	238.405	238.574	106.2	0.07	238.574	120.4	N/A	**4501-2**	238.001	238.120	712.0	0.05	238.071	653.6	N/A
1301-3	237.901	238.119	89.6	0.09	237.979	103.2	N/A	**4501-3**	237.943	238.179	781.2	0.10	238.212	765.6	N/A
1301-4	238.138	238.674	99.6	0.22	238.556	99.8	N/A	**4501-4**	237.888	238.263	745.5	0.16	238.082	709.9	N/A
1301-5	237.386	237.419	88.6	0.01	237.415	103.2	N/A	**4501-5**	238.039	238.253	863.4	0.09	238.202	723.5	N/A

9. CONCLUSIONS

In this dissertation we studied volumetric modulated arc therapy (VMAT) planning, which is an important but difficult problem in cancer treatment. There are four main parts including two mixed integer linear programming formulations for VMAT planning, which minimize total amount of radiation delivered to the patient subject to geometric and clinical requirements, two exact solution methods in order to find optimal VMAT plans and one heuristic to generate plans for clinical size of problems.

In VMAT technique, radiation can be delivered continuously, and the leaves of the MLC system can move and shape the beam during the rotation of the gantry. Therefore, it is possible to obtain high conformal plans in terms of dose distributions requiring less treatment time, which makes the technique one of the widely applied method in external radiation therapy treatment. However, finding high quality VMAT plans is a challenging issue. The apertures composed by the multileaf collimator (MLC) leaves are interdependent, since there is a leaf motion limitation depending on the mechanical properties of the equipment. Namely, the apertures at two adjacent control points in a VMAT plan must be compatible. This makes VMAT planning problem impossible to decompose into independent smaller problems; it must be considered as a whole in contrast with the preceding technology intensity modulated radiation therapy (IMRT). It is challenging to develop good formulations and efficient methods that solve the problem exactly and find good treatment plans. For these reasons the formulations proposed in the literature are not comprehensive enough to include all aspects of the method. They generally relax the dose requirements and try to satisfy them in the objective function by solving a heuristic method. To the best of our knowledge, our mixed integer linear programming models are the first ones in the literature that take into account all requirements related to treatment as well as mechanical properties of the equipment. The formulations differ from each other with respect to the definitions of the leaf positions and each of them includes partial dose-volume requirements as Conditional Value-at-Risk (CVaR) constraints. Moreover, IMRT and VMAT techniques are capable to find high conformal radiation therapy

plans, however they increase the total radiation sent to patient's body during the treatment as compared to the previous techniques, which increase the integral body dose and the risk of secondary malignancy. Thus, the objective of the formulations is to find a solution that delivers as little radiation as possible to the patient, which is new in the VMAT planning literature.

The problem has mainly two parts to decide: the positions of the leaves and the amount of radiation intensity at each one of the control points. Using this observation we decompose the problem into two subproblems in order to develop Benders decomposition algorithms. In the master problem the positions of the leaves are obtained and they are given to the subproblem as input where the dose intensities are determined subject to the clinical requirements. We modify the naive form of the method by applying a number of acceleration strategies and obtained two improved Benders algorithms. In the BP algorithms we reformulate the problem in reverse and introduce each feasible treatment row arc for each MLC row as a variable of the reformulated model. We solve the linear programming relaxation of the reformulated model using column generation at each node of the branch-and-bound tree. For each pricing subproblem, a network model was developed and solved using dynamic programming in polynomial time. We test the performance of the exact solution algorithms on a large set of test instances derived from an anonymous prostate dataset [17]. Note that there are other studies that use the same dataset in VMAT planning within a different settings [56, 105, 106]. They all provide treatment plans satisfying different set of constraints and minimizing or maximizing different objective functions, which makes them incomparable. According to the computational results, Benders algorithms and BP algorithms outperform Gurobi solver especially for large instances. In particular, BP algorithms are more efficient than the improved Benders algorithms. We should also note that it is possible to solve real size problems including only one target volume and one OAR with the current version of our algorithms. For the first time, however, the exact solution algorithms have been proposed to solve a comprehensive mixed linear integer programming model for the VMAT planning problem. Although the problem involves the challenges to be overcome, such an attempt is important and valuable in that it demonstrates these difficulties and creates ground for the future contributions

that may further improve VMAT treatment.

Finally, we propose a two-phase column-generation heuristic, which produces treatment plans in a single call without any human intervention. This is in contrast with the commonly used software systems, which often require multiple iterations of modifications in parameters and re-run. Our heuristic can find high quality VMAT plans for problems with clinically adequate voxel and bixel (beamlet) resolution. In the first phase of the algorithm we generate an initial plan by solving a relaxed model, which is derived from the original model and gives a number of fluence maps. Then we convert these fluence maps into deliverable apertures and sequence them on an arc by applying a simple sequencing operation. In the second phase, we improve the initial solution by column generation iterations. Use of CVaR constraints is not widespread in VMAT planning due to their conservatism. The proposed heuristic includes an automated strategy to tune the parameters of these constraints in order to make them usable without degrading quality of plans. We test our algorithm on nine real prostate patient data and compare the resulting VMAT plans with the ones obtained by an expert dosimetrist on Eclipse [93] in one of the major oncology institutes of Turkey. Our model includes dose-volume constraints of all critical organs and two planning target volumes, parallel to clinical application. The results show that our heuristic is capable to find treatment plans of high quality with respect to clinical dose-volume criteria and requiring fewer MUs in clinically acceptable time.

Potential future research directions include extending the proposed algorithms in order to involve some other properties such as connectedness and disallowing interdigitation of the leaves that may be imposed by some of the MLC systems. Also, we assume that the gantry of the linear accelerator rotates at a constant speed, which can be relaxed by introducing additional variables and linearizing the potential resulting nonlinearities. Finally, the formulations and algorithms may be adapted for other radiation therapy modalities such as Tomotherapy and CyberKnife as in [15], and also for other new technologies such as the intensity modulated proton therapy whose optimization demands very large data sets since it is highly sensitive to uncertainties, or four-dimensional radiation therapy that includes the temporal changes in the patient's

anatomy while planning the treatment.

REFERENCES

1. de Araújo Montagno, E. and R. M. E. Sabbatini, *Radiosurgery*, 1997, `http://www.cerebromente.org.br/n02/tecnologia/radiocirurg.htm`, accessed at May 2019.

2. Varian, *Inside a Varian linear accelerator*, 2019, `https://www.varian.com/fi/about-varian/newsroom/image-gallery/inside-varian-linear-accelerator`, accessed at May 2019.

3. Varian, *The Evaluation of Intensity Modulated Radiation Therapy (IMRT)*, 1999, `http://media.corporate-ir.net/media_files/nys/var/annual/10.htm`, accessed at May 2019.

4. SIMBALLC, *IMRT-What is Intensity-Modulated Radiation Therapy*, 2013, `http://www.simballc.org/imrt.html`, accessed at May 2019.

5. Center, V. M. C., *VCU Massey Cancer Center introduces safer, more effective form of radiation therapy*, 2011, `https://www.massey.vcu.edu/about/blog/2011/massey_cancer_center_introduces_radiation_therapy/`, accessed at May 2019.

6. Pocket Dentistry, *Cone beam computed tomography (CBCT)*, 2015, `https://pocketdentistry.com/13-cone-beam-computed-tomography-cbct/`, accessed at May 2019.

7. Cambazard, H., E. O'Mahony and B. O'Sullivan, "A shortest path-based approach to the multileaf collimator sequencing problem", *Discrete Applied Mathematics*, Vol. 160, No. 1, pp. 81–99, 2012.

8. Teoh, M., C. H. Clark, K. Wood, S. Whitaker and A. Nisbet, "Volumetric modulated arc therapy: a review of current literature and clinical use in practice", *The*

British Journal of Radiology, Vol. 84, pp. 967–996, 2011.

9. Palma, D., E. Vollans, K. James, S. Nakano, V. Moiseenko, R. Shaffer, M. McKenzie, J. Morris and K. Otto, "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*, Vol. 72, No. 4, pp. 996–1001, 2008.

10. Hall, E. J. and C.-S. Wuu, "Radiation-induced second cancers: the impact of 3D-CRT and IMRT", *International Journal of Radiation Oncology* Biology* Physics*, Vol. 56, No. 1, pp. 83–88, 2003.

11. Ehrgott, M., Ç. Güler, H. W. Hamacher and L. Shao, "Mathematical optimization in intensity modulated radiation therapy", *Annals of Operations Research*, Vol. 175, No. 1, pp. 309–365, 2010.

12. Otto, K., "Volumetric modulated arc therapy: IMRT in a single gantry arc", *Medical Physics*, Vol. 35, No. 1, pp. 310–317, 2008.

13. Peng, F., X. Jia, X. Gu, M. A. Epelman, H. E. Romeijn and S. B. Jiang, "A new column-generation-based algorithm for VMAT treatment plan optimization", *Physics in Medicine & Biology*, Vol. 57, No. 14, pp. 4569–4588, 2012.

14. Romeijn, H. E., R. K. Ahuja, J. F. Dempsey and A. Kumar, "A new linear programming approach to radiation therapy treatment planning problems", *Operations Research*, Vol. 54, No. 2, pp. 201–216, 2006.

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

16. Craft, D., D. McQuaid, J. Wala, W. Chen, E. Salari and T. Bortfeld, "Multicri-

teria VMAT optimization", *Medical Physics*, Vol. 39, No. 2, pp. 686–696, 2012.

17. Craft, D., M. Bangert, T. Long, D. Papp and J. Unkelbach, *Supporting material for: "Shared data for IMRT optimization research: the CORT dataset"*, 2014, `http://gigadb.org/dataset/100110`, accessed at May 2019.

18. Bortfeld, T., *IMRT Optimization Based on Physical Criteria*, `https://www.aapm.org/meetings/03SS/Presentations/Bortfield.pdf`, accessed at May 2019.

19. Gören, M. and Z. C. Taşkın, "A column generation approach for evaluating delivery efficiencies of collimator technologies in IMRT treatment planning", *Physics in Medicine & Biology*, Vol. 60, No. 5, p. 1989, 2015.

20. Burnet, N. G., S. J. Thomas, K. E. Burton and S. J. Jefferies, "Defining the tumour and target volumes for radiotherapy", *Cancer Imaging*, Vol. 4, No. 2, p. 153, 2004.

21. Ahnesjö, A., M. Saxner and A. Trepp, "A pencil beam model for photon dose calculation", *Medical Physics*, Vol. 19, No. 2, pp. 263–273, 1992.

22. Sievinen, J., W. Ulmer and W. Kaissl, "AAA photon dose calculation model in Eclipse", *Palo Alto (CA): Varian Medical Systems*, Vol. 118, p. 2894, 2005.

23. Ehrgott, M., A. Holder and J. Reese, "Beam selection in radiotherapy design", *Linear Algebra and its Applications*, Vol. 428, No. 5, pp. 1272–1312, 2008.

24. Romeijn, H. E., R. K. Ahuja, J. F. Dempsey, A. Kumar and J. G. Li, "A novel linear programming approach to fluence map optimization for intensity modulated radiation therapy treatment planning", *Physics in Medicine & Biology*, Vol. 48, No. 21, p. 3521, 2003.

25. Mahmoudzadeh, H., T. G. Purdie and T. C. Chan, "Constraint generation meth-

ods for robust optimization in radiation therapy", *Operations Research for Health Care*, Vol. 8, pp. 85–90, 2016.

26. Papp, D. and J. Unkelbach, "Direct leaf trajectory optimization for volumetric modulated arc therapy planning with sliding window delivery", *Medical Physics*, Vol. 41, No. 1, p. 011701, 2014.

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

28. Ernst, A. T., V. H. Mak and L. R. Mason, "An exact method for the minimum cardinality problem in the treatment planning of intensity-modulated radiotherapy", *INFORMS Journal on Computing*, Vol. 21, No. 4, pp. 562–574, 2009.

29. Mason, L. R., V. H. Mak-Hau and A. T. Ernst, "An exact method for minimizing the total treatment time in intensity-modulated radiotherapy", *Journal of the Operational Research Society*, Vol. 63, No. 10, pp. 1447–1456, 2012.

30. Boland, N., H. W. Hamacher and F. Lenzen, "Minimizing beam-on time in cancer radiation treatment using multileaf collimators", *Networks*, Vol. 43, No. 4, pp. 226–240, 2004.

31. Guta, B., *Subgradient Optimization Methods in Integer Programming with an Application to a Radiation Therapy Problem*, Ph.D. Thesis, Teknishe Universitat Kaiserlautern, Kaiserlautern, 2003.

32. Taşkın, Z. C., J. C. Smith, H. E. Romeijn and J. F. Dempsey, "Optimal multileaf collimator leaf sequencing in IMRT treatment planning", *Operations Research*, Vol. 58, No. 3, pp. 674–690, 2010.

33. Mason, L. R., V. H. Mak-Hau and A. T. Ernst, "A parallel optimisation approach

for the realisation problem in intensity modulated radiotherapy treatment planning", *Computational optimization and applications*, Vol. 60, No. 2, pp. 441–477, 2015.

34. Baatar, D., H. W. Hamacher, M. Ehrgott and G. J. Woeginger, "Decomposition of integer matrices and multileaf collimator sequencing", *Discrete Applied Mathematics*, Vol. 152, No. 1, pp. 6–34, 2005.

35. Lee, E. K., T. Fox and I. Crocker, "Integer programming applied to intensity-modulated radiation therapy treatment planning", *Annals of Operations Research*, Vol. 119, No. 1-4, pp. 165–181, 2003.

36. Bertsimas, D., V. Cacchiani, D. Craft and O. Nohadani, "A hybrid approach to beam angle optimization in intensity-modulated radiation therapy", *Computers & Operations Research*, Vol. 40, No. 9, pp. 2187–2197, 2013.

37. Shepard, D., M. Earl, X. Li, S. Naqvi and C. Yu, "Direct aperture optimization: a turnkey solution for step-and-shoot IMRT", *Medical Physics*, Vol. 29, No. 6, pp. 1007–1018, 2002.

38. Romeijn, H. E., R. K. Ahuja, J. F. Dempsey and A. Kumar, "A column generation approach to radiation therapy treatment planning using aperture modulation", *SIAM Journal on Optimization*, Vol. 15, No. 3, pp. 838–862, 2005.

39. Preciado-Walters, F., M. P. Langer, R. L. Rardin and V. Thai, "Column generation for IMRT cancer therapy optimization with implementable segments", *Annals of Operations Research*, Vol. 148, No. 1, pp. 65–79, 2006.

40. Men, C., H. E. Romeijn, Z. C. Taşkın and J. F. Dempsey, "An exact approach to direct aperture optimization in IMRT treatment planning", *Physics in Medicine & Biology*, Vol. 52, No. 24, pp. 7333–7352, 2007.

41. Carlsson, F., "Combining segment generation with direct step-and-shoot opti-

mization in intensity-modulated radiation therapy", *Medical Physics*, Vol. 35, No. 9, pp. 3828–3838, 2008.

42. Salari, E. and J. Unkelbach, "A column-generation-based method for multi-criteria direct aperture optimization", *Physics in Medicine & Biology*, Vol. 58, No. 3, pp. 621–639, 2013.

43. Yu, C. X., "Intensity-modulated arc therapy with dynamic multileaf collimation: an alternative to tomotherapy", *Physics in Medicine & Biology*, Vol. 40, No. 9, pp. 1435–1449, 1995.

44. Luan, S., C. Wang, D. Cao, D. Z. Chen, D. M. Shepard and X. Y. Cedric, "Leaf-sequencing for intensity-modulated arc therapy using graph algorithms", *Medical Physics*, Vol. 35, No. 1, pp. 61–69, 2008.

45. Wang, C., S. Luan, G. Tang, D. Z. Chen, M. A. Earl and X. Y. Cedric, "Arc-modulated radiation therapy (AMRT): a single-arc form of intensity-modulated arc therapy", *Physics in Medicine & Biology*, Vol. 53, No. 22, pp. 6291–6303, 2008.

46. Cao, D., M. K. Afghan, J. Ye, F. Chen and D. M. Shepard, "A generalized inverse planning tool for volumetric-modulated arc therapy", *Physics in Medicine & Biology*, Vol. 54, No. 21, pp. 6725–6738, 2009.

47. Salari, E., J. Wala and D. Craft, "Exploring trade-offs between VMAT dose quality and delivery efficiency using a network optimization approach", *Physics in Medicine & Biology*, Vol. 57, No. 17, pp. 5587–5600, 2012.

48. Wala, J., E. Salari, W. Chen and D. Craft, "Optimal partial-arcs in VMAT treatment planning", *Physics in Medicine & Biology*, Vol. 57, No. 18, pp. 5861–5874, 2012.

49. Earl, M., D. Shepard, S. Naqvi, X. Li and C. Yu, "Inverse planning for intensity-

modulated arc therapy using direct aperture optimization", *Physics in Medicine & Biology*, Vol. 48, No. 8, pp. 1075–1089, 2003.

50. Yan, H., J.-R. Dai and Y.-X. Li, "A fast optimization approach for treatment planning of volumetric modulated arc therapy", *Radiation Oncology*, Vol. 13, No. 1, p. 101, 2018.

51. Bzdusek, K., H. Friberger, K. Eriksson, B. Hårdemark, D. Robinson and M. Kaus, "Development and evaluation of an efficient approach to volumetric arc therapy planning", *Medical Physics*, Vol. 36, No. 6, pp. 2328–2339, 2009.

52. Bedford, J. L., "Treatment planning for volumetric modulated arc therapy", *Medical Physics*, Vol. 36, No. 11, pp. 5128–5138, 2009.

53. Christiansen, E., E. Heath and T. Xu, "Continuous aperture dose calculation and optimization for volumetric modulated arc therapy", *Physics in Medicine & Biology*, Vol. 63, No. 21, p. 21NT01, 2018.

54. Men, C., H. E. Romeijn, X. Jia and S. B. Jiang, "Ultrafast treatment plan optimization for volumetric modulated arc therapy (VMAT)", *Medical Physics*, Vol. 37, No. 11, pp. 5787–5791, 2010.

55. Rosen, J. B., "The gradient projection method for nonlinear programming. Part I. Linear constraints", *Journal of the society for industrial and applied mathematics*, Vol. 8, No. 1, pp. 181–217, 1960.

56. Mahnam, M., M. Gendreau, N. Lahrichi and L.-M. Rousseau, "Simultaneous delivery time and aperture shape optimization for the volumetric-modulated arc therapy (VMAT) treatment planning problem", *Physics in Medicine & Biology*, Vol. 62, No. 14, pp. 5589–5611, 2017.

57. Mahnam, M., M. Gendreau, N. Lahrichi and L.-M. Rousseau, "Integrating DVH criteria into a column generation algorithm for VMAT treatment planning",

Physics in Medicine & Biology, Vol. 64, No. 8, p. 085008, 2019.

58. Peng, F., S. B. Jiang, H. E. Romeijn and M. A. Epelman, "VMATc: VMAT with constant gantry speed and dose rate", *Physics in Medicine & Biology*, Vol. 60, No. 7, p. 2955, 2015.

59. Hoegele, W., R. Loeschel, N. Merkle and P. Zygmanski, "An efficient inverse radiotherapy planning method for VMAT using quadratic programming optimization", *Medical Physics*, Vol. 39, No. 1, pp. 444–454, 2012.

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

61. Song, J., Z. Shi, B. Sun and L. Shi, "Treatment Planning for Volumetric-Modulated Arc Therapy: Model and Heuristic Algorithms", *IEEE Transactions on Automation Science and Engineering*, Vol. 12, No. 1, pp. 116–126, 2015.

62. Unkelbach, J., T. Bortfeld, D. Craft, M. Alber, M. Bangert, R. Bokrantz, D. Chen, R. Li, L. Xing, C. Men *et al.*, "Optimization approaches to volumetric modulated arc therapy planning", *Medical Physics*, Vol. 42, No. 3, pp. 1367–1377, 2015.

63. Cedric, X. Y. and G. Tang, "Intensity-modulated arc therapy: principles, technologies and clinical implementation", *Physics in Medicine & Biology*, Vol. 56, No. 5, p. R31, 2011.

64. Breedveld, S., D. Craft, R. van Haveren and B. Heijmen, "Multi-criteria optimisation and decision-making in radiotherapy", *European Journal of Operational Research*, 2018.

65. Smyth, G., J. C. Bamber, P. M. Evans and J. L. Bedford, "Trajectory optimization for dynamic couch rotation during volumetric modulated arc radiotherapy", *Physics in Medicine & Biology*, Vol. 58, No. 22, pp. 8163–8177, 2013.

66. Lyu, Q., Y. Y. Victoria, D. Ruan, R. Neph, D. O'Connor and K. Sheng, "A novel optimization framework for VMAT with dynamic gantry couch rotation", *Physics in Medicine & Biology*, 2018.

67. Dursun, P., Z. C. Taşkın and İ. K. Altınel, "Mathematical Models for Optimal Volumetric Modulated Arc Therapy (VMAT) Treatment Planning", *Procedia Computer Science (Proceedings of International Conference on Health and Social Care Information Systems and Technologies, HCist, Porto)*, Vol. 100, pp. 644–651, 2016.

68. McCormick, G. P., "Computability of global solutions to factorable nonconvex programs: Part I- Convex underestimating problems", *Mathematical Programming*, Vol. 10, No. 1, pp. 147–175, 1976.

69. Rockafellar, R. T., S. Uryasev *et al.*, "Optimization of conditional value-at-risk", *Journal of Risk*, Vol. 2, pp. 21–42, 2000.

70. Dursun, P., Z. C. Taşkın and İ. K. Altınel, "The determination of optimal treatment plans for Volumetric Modulated Arc Therapy (VMAT)", *European Journal of Operational Research*, Vol. 272, No. 1, pp. 372–388, 2019.

71. Benders, J. F., "Partitioning procedures for solving mixed-variables programming problems", *Numerische mathematik*, Vol. 4, No. 1, pp. 238–252, 1962.

72. Rahmaniani, R., T. G. Crainic, M. Gendreau and W. Rei, "The Benders decomposition algorithm: A literature review", *European Journal of Operational Research*, Vol. 259, No. 3, pp. 801–817, 2017.

73. Magnanti, T. L. and R. T. Wong, "Accelerating Benders decomposition: Algorithmic enhancement and model selection criteria", *Operations Research*, Vol. 29, No. 3, pp. 464–484, 1981.

74. Van Roy, T. J., "A cross decomposition algorithm for capacitated facility loca-

tion", *Operations Research*, Vol. 34, No. 1, pp. 145–163, 1986.

75. Üster, H. and H. Agrahari, "A Benders decomposition approach for a distribution network design problem with consolidation and capacity considerations", *Operations Research Letters*, Vol. 39, No. 2, pp. 138–143, 2011.

76. Adulyasak, Y., J.-F. Cordeau and R. Jans, "Benders Decomposition for Production Routing Under Demand Uncertainty", *Operations Research*, Vol. 63, No. 4, pp. 851–867, 2015.

77. Lin, S., *Benders Decomposition and an IP-Based Heuristic for Selecting IMRT Treatment Beam Angles*, Master's Thesis, The University of Texas at Austin, 2014.

78. Fischetti, M., D. Salvagnin and A. Zanette, "A note on the selection of Benders' cuts", *Mathematical Programming*, Vol. 124, No. 1-2, pp. 175–182, 2010.

79. Codato, G. and M. Fischetti, "Combinatorial Benders' cuts for mixed-integer linear programming", *Operations Research*, Vol. 54, No. 4, pp. 756–766, 2006.

80. Taşkın, Z. C. and M. Cevik, "Combinatorial Benders cuts for decomposing IMRT fluence maps using rectangular apertures", *Computers & Operations Research*, Vol. 40, No. 9, pp. 2178–2186, 2013.

81. Gleeson, J. and J. Ryan, "Identifying minimally infeasible subsystems of inequalities", *ORSA Journal on Computing*, Vol. 2, No. 1, pp. 61–63, 1990.

82. Held, M., P. Wolfe and H. P. Crowder, "Validation of subgradient optimization", *Mathematical Programming*, Vol. 6, No. 1, pp. 62–88, 1974.

83. Dursun, P., Z. C. Taşkın and İ. K. Altınel, "Using Branch-and-Price to Determine Optimal Treatment Plans for Volumetric Modulated Arc Therapy (VMAT)", *Computers & Operations Research*, Vol. 110, pp. 1–17, 2019.

84. Lübbecke, M. E., "Column generation", *Wiley Encyclopedia of Operations Research and Management Science*, 2011.

85. Vanderbeck, F., *Decomposition and column generation for integer programs*, Ph.D. Thesis, Université catholique de Louvain, 1994.

86. Desaulniers, G., J. Desrosiers and M. M. Solomon, *Column Generation*, Vol. 5, Springer Science & Business Media, 2006.

87. Savelsbergh, M., "A branch-and-price algorithm for the generalized assignment problem", *Operations Research*, Vol. 45, No. 6, pp. 831–841, 1997.

88. Lübbecke, M. E. and J. Desrosiers, "Selected topics in column generation", *Operations Research*, Vol. 53, No. 6, pp. 1007–1023, 2005.

89. Hindi, H., "A tutorial on optimization methods for cancer radiation treatment planning", *American Control Conference (ACC), 2013*, pp. 6804–6816, IEEE, 2013.

90. Craft, D., M. Bangert, T. Long, D. Papp and J. Unkelbach, "Shared data for intensity modulated radiation therapy (IMRT) optimization research: the CORT dataset", *GigaScience*, Vol. 3, No. 1, p. 37, 2014.

91. Taşkın, Z. C., *VMAT Data Sets*, 2019, `http://www.ie.boun.edu.tr/~taskin/data/vmat`, accessed at May 2019.

92. Pianykh, O. S., *Digital imaging and communications in medicine (DICOM): a practical introduction and survival guide*, Springer Science & Business Media, 2009.

93. Varian, *Eclipse Treatment Planning System*, 2019, `https://www.varian.com/oncology/products/software/treatment-planning/eclipse-treatment-planning-system`, accessed at May 2019.

94. Wieser, H.-P., E. Cisternas, N. Wahl, S. Ulrich, A. Stadler, H. Mescher, L.-R. Müller, T. Klinge, H. Gabrys, L. Burigo, A. Mairani, S. Ecker, B. Ackermann, M. Ellerbrock, K. Parodi, O. Jäkel and M. Bangert, "Development of the open-source dose calculation and optimization toolkit matRad", *Medical Physics*, Vol. 44, No. 6, pp. 2556–2568, 2017.

95. Bortfeld, T., W. Schlegel and B. Rhein, "Decomposition of pencil beam kernels for fast dose calculations in three-dimensional treatment planning", *Medical Physics*, Vol. 20, No. 2, pp. 311–318, 1993.

96. Buyyounouski, M. K., E. M. Horwitz, R. A. Price, S. J. Feigenberg and A. Pollack, "Prostate IMRT", *Image-Guided IMRT*, pp. 391–410, Springer, 2006.

97. Buyyounouski, M. K., E. M. Horwitz, R. A. Price, A. L. Hanlon, R. G. Uzzo and A. Pollack, "Intensity-modulated radiotherapy with MRI simulation to reduce doses received by erectile tissue during prostate cancer treatment", *International Journal of Radiation Oncology Biology Physics*, Vol. 58, No. 3, pp. 743–749, 2004.

98. Emami, B., "Tolerance of normal tissue to therapeutic radiation", *Reports of radiotherapy and Oncology*, Vol. 1, No. 1, 2013.

99. Roach III, M., J. Nam, G. Gagliardi, I. El Naqa, J. O. Deasy and L. B. Marks, "Radiation dose–volume effects and the penile bulb", *International Journal of Radiation Oncology Biology Physics*, Vol. 76, No. 3, pp. S130–S134, 2010.

100. Tøndel, H., J.-Å. Lund, S. Lydersen, A. D. Wanderås, B. Y. Aksnessæther, C. A. Jensen, S. Kaasa and A. Solberg, "Dose to penile bulb is not associated with erectile dysfunction 18 months post radiotherapy: A secondary analysis of a randomized trial", *Clinical and Translational Radiation Oncology*, Vol. 13, pp. 50–56, 2018.

101. Python, *Python 2.7.11 documentation*, 2015, `https://docs.python.org/release/2.7.11/`, accessed at May 2019.

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

103. McShane, K. A., C. L. Monma and D. Shanno, "An implementation of a primal-dual interior point method for linear programming", *ORSA Journal on computing*, Vol. 1, No. 2, pp. 70–83, 1989.

104. Gurobi, O., *Gurobi optimizer reference manual version 8.0*, 2018, `https://www.gurobi.com/documentation/8.0/refman.pdf`, accessed at May 2019.

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

106. Balvert, M. and D. Craft, "Fast approximate delivery of fluence maps for IMRT and VMAT", *Physics in Medicine & Biology*, Vol. 62, No. 4, p. 1225, 2017.

107. MATLAB, "Version 8.5", *The MathWorks Inc*, 2015.

APPENDIX A: STRONG BENDERS CUT

Let's $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}^1, \hat{\boldsymbol{\mu}}^2, \bar{\boldsymbol{\beta}}^1, \bar{\boldsymbol{\beta}}^2, \bar{\boldsymbol{\beta}}^3)$ be the optimal solution of RDSP$(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}^1, \hat{\boldsymbol{\mu}}^2, \hat{\boldsymbol{\beta}}^1, \hat{\boldsymbol{\beta}}^2)$. Note that $\bar{\boldsymbol{\beta}}^1$ for $(i,j,k) \in \mathscr{Z}_1$ equals to $\hat{\boldsymbol{\beta}}^1$, and $\bar{\boldsymbol{\beta}}^2$ for $(i,j,k) \in \mathscr{Z}_0$ equals to $\hat{\boldsymbol{\beta}}^2$. Namely, only $\bar{\boldsymbol{\beta}}^1$ for $(i,j,k) \in \mathscr{Z}_0$, $\bar{\boldsymbol{\beta}}^2$ for $(i,j,k) \in \mathscr{Z}_1$ and $\bar{\boldsymbol{\beta}}^3$ for all (i,j,k) may be different from the optimal solution of RDSP. Let's assume that the Benders optimality cuts $f(\bar{\boldsymbol{\beta}}^1, \bar{\boldsymbol{\beta}}^2, \hat{\boldsymbol{\theta}}^1, \hat{\boldsymbol{\theta}}^2, \hat{\boldsymbol{\epsilon}}^1, \hat{\boldsymbol{\epsilon}}^2, \hat{\boldsymbol{\mu}}^1, \hat{\boldsymbol{\mu}}^2) \leq \eta$ and $f(\hat{\boldsymbol{\beta}}^1, \hat{\boldsymbol{\beta}}^2, \hat{\boldsymbol{\theta}}^1, \hat{\boldsymbol{\theta}}^2, \hat{\boldsymbol{\epsilon}}^1, \hat{\boldsymbol{\epsilon}}^2, \hat{\boldsymbol{\mu}}^1, \hat{\boldsymbol{\mu}}^2) \leq \eta$ are derived from the optimal solution of RDSP and DSP, respectively. There are two cases:

(i) $\bar{\boldsymbol{\beta}}^1 \leq \hat{\boldsymbol{\beta}}^1$ for all $(i,j,k) \in \mathscr{Z}_0$ and $\bar{\boldsymbol{\beta}}^2 \leq \hat{\boldsymbol{\beta}}^2$ for all $(i,j,k) \in \mathscr{Z}_1$.
In this case, it is trivial that the following inequality

$$f(\hat{\boldsymbol{\beta}}^1, \hat{\boldsymbol{\beta}}^2, \hat{\boldsymbol{\theta}}^1, \hat{\boldsymbol{\theta}}^2, \hat{\boldsymbol{\epsilon}}^1, \hat{\boldsymbol{\epsilon}}^2, \hat{\boldsymbol{\mu}}^1, \hat{\boldsymbol{\mu}}^2) \leq f(\bar{\boldsymbol{\beta}}^1, \bar{\boldsymbol{\beta}}^2, \hat{\boldsymbol{\theta}}^1, \hat{\boldsymbol{\theta}}^2, \hat{\boldsymbol{\epsilon}}^1, \hat{\boldsymbol{\epsilon}}^2, \hat{\boldsymbol{\mu}}^1, \hat{\boldsymbol{\mu}}^2) \tag{A.1}$$

is satisfied for all master solutions ($\mathbf{z}$). Since,
$-U^{mu} \sum_{(i,j,k)\in Z_0} z_{ijk}\hat{\beta}^1_{ijk} + U^{mu} \sum_{(i,j,k)\in Z_1} (z_{ijk}-1)\hat{\beta}^2_{ijk} \leq -U^{mu} \sum_{(i,j,k)\in Z_0} z_{ijk}\bar{\beta}^1_{ijk} + U^{mu} \sum_{(i,j,k)\in Z_1} (z_{ijk}-1)\bar{\beta}^2_{ijk}$ is ensured when there are z_{ijk} variables take value 1 in $(i,j,k) \in \mathscr{Z}_0$ and/or take value 0 in $(i,j,k) \in \mathscr{Z}_1$. Note that, the remaining parts of the functions where dual variables take the same value are exactly same in both cuts.

(ii) For at least one $(i,j,k) \in \mathscr{Z}_0$ $\bar{\beta}^1_{ijk} > \hat{\beta}^1_{ijk}$ and/or for at least one $(i,j,k) \in \mathscr{Z}_1$ $\bar{\beta}^2_{ijk} > \hat{\beta}^2_{ijk}$.
Hence, the inequality (A.1) may not be satisfied for some of the master solutions in this case, but $f(\hat{\boldsymbol{\beta}}^1, \hat{\boldsymbol{\beta}}^2, \hat{\boldsymbol{\theta}}^1, \hat{\boldsymbol{\theta}}^2, \hat{\boldsymbol{\epsilon}}^1, \hat{\boldsymbol{\epsilon}}^2, \hat{\boldsymbol{\mu}}^1, \hat{\boldsymbol{\mu}}^2) \leq \eta$ can not dominate the new cut $f(\bar{\boldsymbol{\beta}}^1, \bar{\boldsymbol{\beta}}^2, \hat{\boldsymbol{\theta}}^1, \hat{\boldsymbol{\theta}}^2, \hat{\boldsymbol{\epsilon}}^1, \hat{\boldsymbol{\epsilon}}^2, \hat{\boldsymbol{\mu}}^1, \hat{\boldsymbol{\mu}}^2) \leq \eta$. Because, there must be at least one other $(i,j,k) \in \mathscr{Z}_0$ or $(i,j,k) \in \mathscr{Z}_1$ that satisfies $\bar{\beta}^1_{ijk} < \hat{\beta}^1_{ijk}$ or $\bar{\beta}^2_{ijk} < \hat{\beta}^2_{ijk}$. Since the objective function of RDSP minimizes the sum of $\boldsymbol{\beta}^1$ and $\boldsymbol{\beta}^2$. Therefore, there are other master solutions that satisfy $f(\hat{\boldsymbol{\beta}}^1, \hat{\boldsymbol{\beta}}^2, \hat{\boldsymbol{\theta}}^1, \hat{\boldsymbol{\theta}}^2, \hat{\boldsymbol{\epsilon}}^1, \hat{\boldsymbol{\epsilon}}^2, \hat{\boldsymbol{\mu}}^1, \hat{\boldsymbol{\mu}}^2) < f(\bar{\boldsymbol{\beta}}^1, \bar{\boldsymbol{\beta}}^2, \hat{\boldsymbol{\theta}}^1, \hat{\boldsymbol{\theta}}^2, \hat{\boldsymbol{\epsilon}}^1, \hat{\boldsymbol{\epsilon}}^2, \hat{\boldsymbol{\mu}}^1, \hat{\boldsymbol{\mu}}^2)$ is ensured. We can conclude that the new cut $f(\bar{\boldsymbol{\beta}}^1, \bar{\boldsymbol{\beta}}^2, \hat{\boldsymbol{\theta}}^1, \hat{\boldsymbol{\theta}}^2, \hat{\boldsymbol{\epsilon}}^1, \hat{\boldsymbol{\epsilon}}^2, \hat{\boldsymbol{\mu}}^1, \hat{\boldsymbol{\mu}}^2)) \leq \eta$ is not dominated, namely it is strong (or pareto-optimal among the cuts of alternative optimal solutions of DSP($\hat{\mathbf{z}}$).

APPENDIX B: DOSE CALCUTATION BY matRad

matRad [94] is a multi-modality radiation treatment planning system written in MATLAB [107], and supports IMRT planning for photons, scanned protons, and scanned carbon ions at clinically adequate resolution. It is freely available online and has been developed to contribute to educational and research activities.

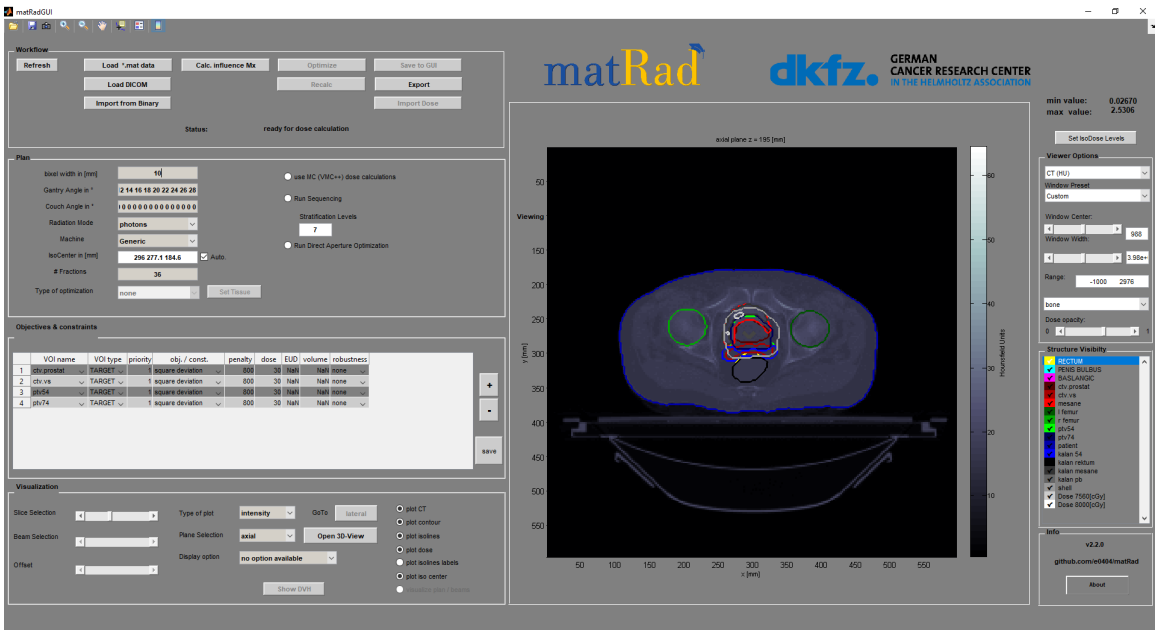

Figure B.1. matRadGUI.

The first step is to import DICOM images with radiation therapy (RT) structure files into the matRadGUI interface shown in Figure B.1. After selecting the voxel resolution, they must be converted into a .mat file, which can be opened in madRadGUI. This .mat file should be opened, and then, the couch angle, the beam angles to calculate dose-influence matrices are selected. It is possible to enter more than one beam angle (e.g. 0 2 4 6). Also, PTV and OARs must be introduced and the corresponding objective functions with priority weights and other parameters such as gantry location, energy type etc. must be specified. Finally, dose influence matrices, which include dose contributions to each voxel from each beamlet/pencil beam at unit intensity, can be calculated and saved as .mat file.

We scale the dose-influence matrices in such a way that the absorbed dose of 1 cGy/MU (i.e. 0.01 Gy/MU) is delivered at 100 cm SAD at 5 cm depth with field size 10 cm $\times$ 10 cm similar to the calibration used at Istanbul University Oncology Institute. We use a solid water phantom CT data which is provided by the institute. We execute a simulation using this phantom, in which we define a 10 cm x 10 cm target volume at 100 cm SAD whose center is passing through the isoline. Also we contour a small volume at 5 cm depth at the center. Then, we validate in Eclipse that 0.01 Gy is absorbed by this volume when 1 MU radiation is delivered. The necessary setup is illustrated in Figure B.2. (SSD stands for source-to-surface distance). We calculate

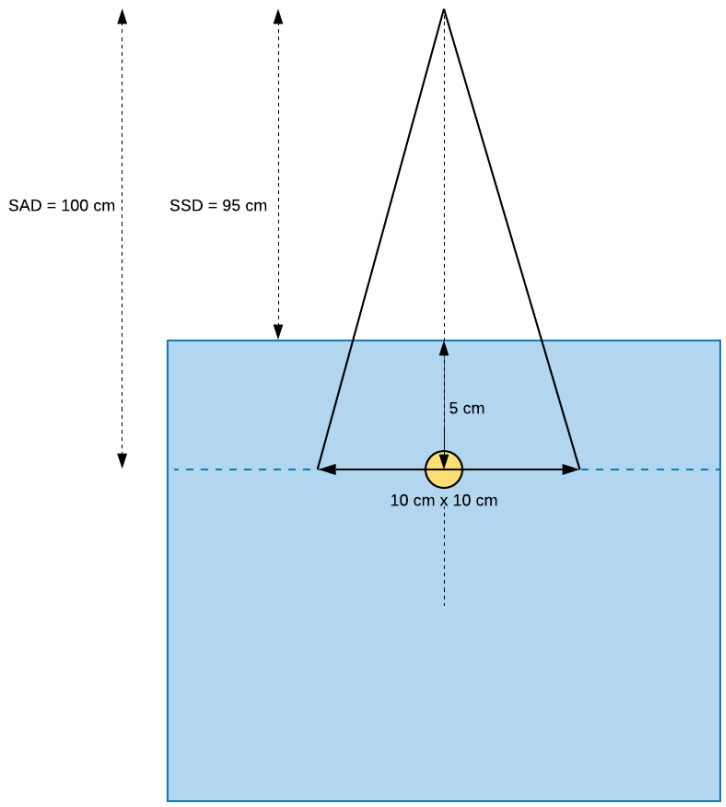

Figure B.2. SAD setup.

dose-influence matrices for this volume defined on phantom in matRad also. For each one of the voxels, especially for the ones at the center, these values are very close to 1 (i.e. 1 Gy), which means 100 MUs radiation is delivered. Thus, we divide the original dose-influence matrices 100 to obtain Gy/MU values.

In order to validate this scaling parameter, we also check the depth dose curves obtained by Eclipse (Figure B.3) and also matRad (Figure B.4) and observe that they are consistent.

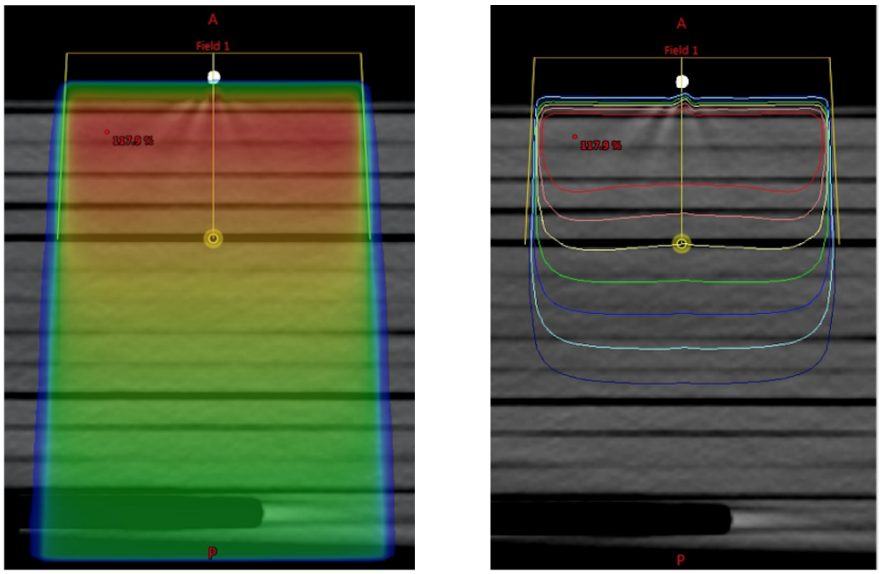

Figure B.3. Depth dose curves obtained in Eclipse.

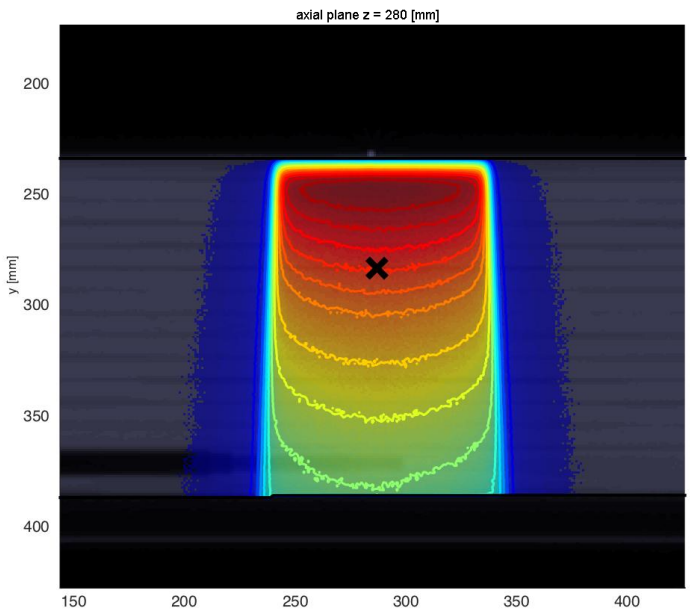

Figure B.4. Depth dose curves obtained in matRad.