Robust Direct Aperture Optimization Methods for Cardiac Sparing in Left-Sided Breast Cancer Radiation Therapy

by

Danielle Ripsman

A thesis presented to the University of Waterloo
in fulfillment of the thesis requirement for the degree of
Master of Applied Science in
Management Sciences

Waterloo, Ontario, Canada, 2018

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

Supervisor:	Houra Mahmoudzadeh Assistant Professor, Department of Management Sciences, Faculty of Engineering, University of Waterloo
Other Members:	Fatma Gzara Associate Professor, Department of Management Sciences, Faculty of Engineering, University of Waterloo
	Fatih Safa Erenay Associate Professor, Department of Management Sciences, Faculty of Engineering, University of Waterloo

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Designing conformal and equipment-compatible radiation therapy plans is essential for ensuring high-quality treatment outcomes for cancer patients. Intensity modulated radiation therapy (IMRT) is a commonly-used method of radiation delivery for cancer patients, wherein beams of radiation are individually contoured to cover a patient's tumour cells while avoiding healthy cells and organs. In IMRT for left-sided breast cancer, the goal is to irradiate all cells in the breast tissue while avoiding the neighbouring, and extremely radiation-sensitive, heart cells. To add to the complexity of this treatment, the entire dose must be delivered while the patient is breathing, causing the location of the heart and target organs to move and deform unpredictably.

The search for a plan that is of the highest quality for a specified set of parameters is called treatment plan optimization. One method of treatment plan optimization that provides an optimal radiation distribution, even under the worst-case realization of a patient's motion uncertainty, uses a framework called robust optimization. A drawback of using this robust optimization framework, however, is that it does not immediately output physically deliverable IMRT plans. Rather, a subsequent, non-trivial post-processing phase must be applied to the output intensity distributions in order to generate an equipment-compatible plan; a process which can substantially degrade the treatment quality.

In this thesis, a holistic approach that combines enforcement of delivery constraints with robust optimization is introduced. The process for creating deliverable plans is called direct aperture optimization (DAO), and the combined model is called robust DAO (RDAO). Novel modelling strategies for integrating the DAO requirements into a robust framework are presented, leading to a large-scale difficult-to-solve mixed integer programming problem. To contend with the complexity of the problem, additional modelling approaches are suggested for improving solution efficiency. These approaches include a hybrid heuristic-optimization technique, which provides good quality, but non-optimal treatment plans. Clinicians may use the output of this hybrid technique as is, or apply it as a warm start for the RDAO model.

The models are implemented in C++ and CPLEX and results are presented, first using a one-dimensional phantom, and then a three-dimensional clinical patient dataset. While the full RDAO model is quite time-consuming to run, high-quality plans are ultimately produced. These plans are both clinically deliverable and mitigate the risk of underdosing a patient's cancerous cells under motion uncertainty, demonstrating their value over plans that did not account for motion uncertainty.

Acknowledgements

I would like to thank my supervisor, Dr. Houra Mahmoudzadeh for her support and guidance throughout this masters. Her dedication to her students and passion for operations research is infectious and something I strive to emulate. I could not have asked for a better, more engaged advisor.

I would also like to thank my readers, Dr. Fatma Gzara and Dr. Fatih Safa Erenay for their time and insight.

I would like to thank my colleagues, Akram, Paulo, Cynthia, Hannah and Daniel, for engaging in research discussion, commiserating and generally making me feel less disconnected from the world in the strange bubble that is post-graduate studies.

Finally, I would like to thank my family and friends for always supporting and encouraging my weird interests, as well as putting up with my periodic caffeine-fuelled disappearances. My mother Arlene and father Colin have always insisted I could do anything I set my mind to, and I am lucky that I had that kind of support growing up (my mother also said I would end up in graduate school, so yes mom, here I am, and yes, you are always right).

My brothers David and Ryan are the best and most supportive siblings I could hope for. And thanks to my Nana and Grandpa, Bubby and Zaidy for always providing the love, and hassling me about when I'll be home next. An added thanks to Nixi for sticking around and being a good, furry shoulder to lean on whenever I came home, for as long as she could.

Table of Contents

List of Tables

List of Figures

Abbreviations

4DCT four-dimensional computed tomography 10

CT computed tomography 4, 10, 41

CTV clinical target volume 4, 8, 9, 12, 16–19, 43, 44, 47, 50, 51, 53, 56–58, 61, 78

DAO direct aperture optimization 7, 13–16, 20–23, 30, 41, 42, 46, 48, 49, 57, 58, 63, 64

DVH dose volume histogram 12, 53, 56, 58, 61

FM fluence map 12, 53, 54, 61

FMO fluence map optimization 5–7, 13, 14, 16–18, 21, 28, 41, 46, 49, 51, 57

IMRT intensity modulated radiation therapy 1–5, 8, 10, 14, 17, 20, 42, 63

linac linear accelerator 2, 4, 20

LP linear program 30, 44, 76

MIP mixed integer programming 7, 14, 23, 36, 40, 41, 49, 61, 63, 64, 76

MLC multileaf collimator 3–5, 16, 20, 21, 26–28

OARs organs at risk 4, 12, 17

RDAO robust direct aperture optimization 14–16, 22, 63, 64

RFMO robust fluence map optimization 19, 20, 63

VMAT volumetric modulated radiation therapy 5, 64

Chapter 1

Introduction

Fundamental improvements in both patient imaging and radiation delivery technology have led to an increase in utilization of radiation therapy for treating cancer patients (Bernier et al., 2004). Today, radiation therapy is recommended for approximately 50% of all cancer patients either in a curative capacity or in an effort to achieve local tumour control (Baskar et al., 2012). Designing conformal and equipment-compatible radiation therapy plans is essential for ensuring high-quality treatment outcomes for these patients.

In intensity modulated radiation therapy (IMRT), a commonly used method of radiation delivery for cancer patients, beams of radiation are individually contoured to conform to the patient's tumour cells while avoiding healthy cells and organs. This level of customization is a huge step forward in the world of patient-tailored medicine, however, it comes at a significant cost in terms of time and planning complexity. Providing tools which automate and optimize key decision-making elements of this planning process for clinicians can lead, not only to better outcomes for patients, but also to improvements in the overall efficiency of the treatment process.

This thesis explores optimization and modelling methodologies for IMRT devices, with an emphasis on left-sided breast cancer treatments. Left-sided breast cancer patients require a unique set of considerations due to the complications that arise from the structural anatomy, cardiac radiation exposure and breathing motion uncertainty within their treatment region. In this chapter, the relevant background about the two major elements of the treatment planning process, i.e., the IMRT device and the breast cancer treatment region, are introduced. This introduction is followed by high-level overview of the math behind the planning process, the methods for evaluating plan quality and a review of recent and relevant literature. Finally, thesis contributions and structure are outlined.

1.1 Intensity Modulated Radiation Therapy

Intensity modulation radiation therapy is a form of external beam radiation therapy delivered using a device called a linear accelerator (linac). The linac shoots photon beams at the patient, leaving a continuous deposit of dose throughout the patient's tissue (Nguyen and Zietman, 2008). Each beam deposits the highest amount of energy at the surface layer of the patient, gradually losing energy as it travels towards (and then through) the target, or tumour. Through a combination of beam shapes and angles, a focused high-intensity centre is accumulated around the tumour, so that the maximum dose is concentrated within the tumour region, rather than along the external tissue (Vachani, 2018). This beam focal point effect is demonstrated in Figure 1.1.

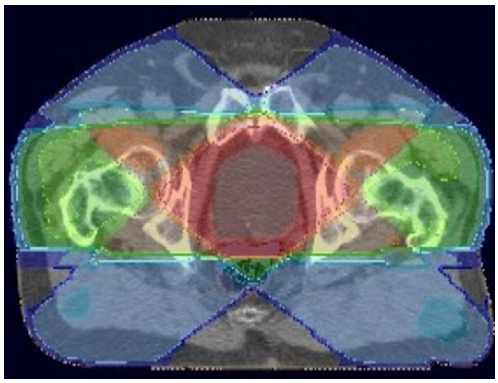


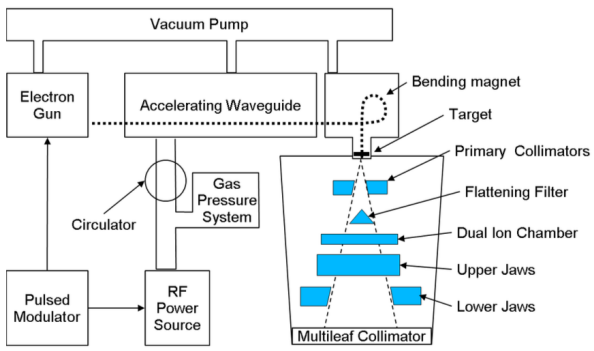
Figure 1.1: Six IMRT beams converge on a central target. The higher, central doses are visually signified by the warmer red and orange colours, while cooler blue and green colours are used to indicate the external, lower doses. Image taken from Vachani (2018).

While the physics behind the radiation source and photon acceleration chambers of a linac is well beyond the scope of this thesis, the key elements of the device are shown schematically in Figure 1.2a. These elements operate largely behind the scenes during treatment, whereas the patient-facing elements, such as the bench, the gantry and the collimator, are all visible during delivery, as in Figure 1.2b.

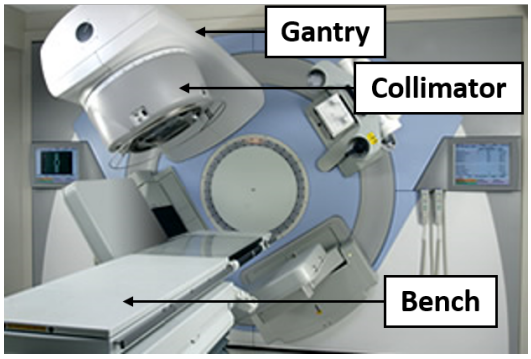
The *bench* is a horizontal bed, on which a patient lies to receive treatment. The placement of the patient is critical and various methodologies are used to ensure that the patient is placed accurately and remains completely still throughout the treatment delivery. These methods range from tattoos, to netting, to body molds. Whichever method is used, the end result (and assumption for this work) is a presumed deterministic knowledge of where the patient is located relative to the delivery device during treatment.

The *gantry* is the large arm which holds the delivery component of the device. The gantry has the ability to rotate 360° around the bench, allowing the treatment to be delivered from all angles.

The delivery component, which sits at the end of the gantry, is the *collimator*. The collimator is not only the point of exit for the beam, but in an IMRT-enabled linac, it



(a) General linac schematic from Milette (2008)



(b) Elekta Synergy® delivery unit from Princeton Neurological Surgery

Figure 1.2: Linac devices with IMRT delivery capabilities.

is often referred to as the multileaf collimator (MLC) since it houses pairs of radiation-absorbing tungsten bits of metal, called *leaves*. These leaves are attached to linear motors which extend inwards from either side of the mouth of the collimator, as in Figure 1.3. The leaves effectively block out portions of the initially rectangular beam, while the remaining negative space between the leaves form shapes called *apertures*.

In addition to the leaves, in a typical collimator, there are solid blocks of metal called jaws positioned above the leaves, as shown in Figure 1.2a. Once the leaf positions for a given aperture are chosen, these 4 solid sheets of metal can come in from each side of the MLC, providing further shielding for the tiny gaps between closed leaves. Like leaves, the jaws are attached to linear motors, and come in pairs in order to span the full mouth of the collimator. The final panel in Figure 1.3 shows how jaw placement works without interfering with the chosen beam-shape.

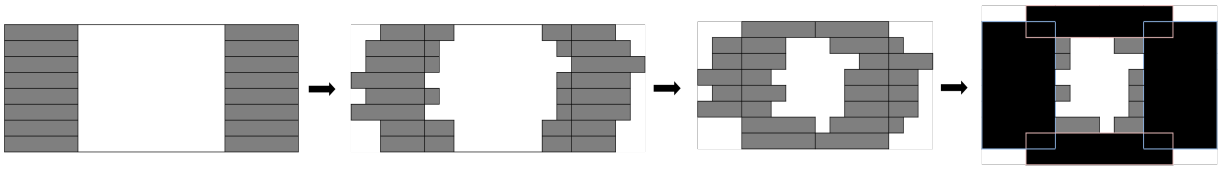


Figure 1.3: MLC leaves move towards the centre of the beam to form an aperture. Upon completion, the peripheral parts of the leaves are covered with upper and lower jaws.

Preparing a patient for IMRT treatment typically requires three steps: (1) imaging of the patient's target region, (2) organ delineation, and (3) treatment strategy creation (Romeijn and Dempsey, 2008).

Imaging is typically performed using a computed tomography (CT) scan, which uses radiation to create a 3D image of the patient (Milette, 2008). The delineation of the various organs, such as the clinical target volume (CTV) and organs at risk (OARs) within these images is typically performed by a radiation oncologist. Information from the delineation is used to determine a treatment strategy, which includes the prescribed dose, the maximum allowable dose to the OARs and the number of fractions over which a treatment plan is to be delivered. Fractions are the number of repeat sessions required for total radiation delivery, meaning the patients comes back to the clinic multiple times to receive their full treatment. According to the American Cancer Society (2018), the number of fractions typically ranges from 25-40, and the treatment time at each fraction is generally between 15 and 30 minutes.

1.1.1 Step-and-Shoot IMRT

One of the most straight-forward uses of the IMRT technology, and also the focus of this thesis, is designing *step-and-shoot* treatment plans. In step-and-shoot IMRT delivery, the linac is turned on to deliver a specified dose at each selected aperture shape and is shut off and re-oriented before delivering the next aperture (here, *aperture* refers to both the shape and the beam intensity delivered through that shape, as a convenient shorthand). A *plan* for step-and-shoot treatment consists of 3 pieces of information:

1. A set of gantry angles
2. The corresponding set of aperture shapes for each angle
3. Beam intensities (or equivalently, beam duration at specified dose rates) for each aperture

The primary drawback of using step-and-shoot planning is the requisite setup time prior to each aperture delivery, wherein the gantry rotates to the predetermined angles and/or the MLC is reshaped. Treatment time is an important consideration, seeing as clinics have queues for treating cancer patients, and longer treatment durations can result in fewer patients being treated. Long treatment times can also be physically uncomfortable for patients, seeing as the treatments occur daily and require immobilization for the duration of the delivery period. For these reasons, it is desirable to balance the quality of the treatment with the number of apertures required for its delivery.

Alternative IMRT-based delivery methods typically use continuous delivery techniques to speed up the treatment process. The *sliding window* method, for example, delivers

dose to the patient in a single, unidirectional sweeping motion at each angle. Like step-and-shoot delivery, sliding window delivery requires the beam to be turned off during angle reorientation, however within each angle, the leaf orientation is dynamic, so the treatment time may be quicker. A more extreme method of continuous delivery is called volumetric modulated radiation therapy (VMAT). In VMAT, the angle orientation and MLC orientation occur dynamically as the treatment is delivered over a 360° (or 720°, or 1080°, etc.) rotation of the gantry. The net result is that all angles are covered within each rotation, without ever turning off the beam.

One weakness of these continuous motion treatments is a loss in optimality, as extra restrictions are placed on leaf movement and sequencing. A second disadvantage, particularly for VMAT plans, is in the complexity of planning the treatments which can have orders of magnitude more apertures than conventional step-and-shoot plans, each of which must be pre-specified by the planner (Mahnam et al., 2017). Finally, not all clinics are equipped with the technology to deliver these continuous plans, whereas step-and-shoot enabled IMRT devices are more common.

In practice, the complications of continuous planning make step-and-shoot treatment a viable and widely applicable mode of radiation therapy delivery. Additionally, many advancements made within the field of step-and-shoot IMRT are readily applicable to improving their continuous counterparts.

1.1.2 Traditional Step-and-Shoot Planning Methodologies

As outlined in Section 1.1.1, designing step-and-shoot plans requires the selection of beam angles, apertures and intensities. Unfortunately, even with the simplifying assumptions, solving the global problem, i.e., choosing the optimal combinations of all three of these plan elements, becomes an intractably large combinatorial version of an already NP-hard problem (Sultan, 2006). These factors have conventionally led to the division of efforts across multiple smaller problems (Taşkın et al., 2010). More specifically, the planning process is often approached in a three-phase method as follows (Gladwish et al., 2007, Romeijn and Dempsey, 2008):

Phase 1. Beam-angles are selected

Phase 2. An idealized fluence map optimization (FMO) is solved

Phase 3. Apertures and intensities are chosen

Phase 1 or beam angle selection has been recognized to be very complex, often relying on empirical heuristics such as choosing equidistant angles rather than investigating every potential angle placement. In general, up to 9 angles are selected for a generalized target, although this number may vary considerably, depending on type of cancer being treated (Bortfeld and Schlegel, 1993, Jiang et al., 2005).

Phase 2 or FMO determines a set of beam intensities for each beam angle which creates an overall optimal shape for tumour conformity, with minimal harm to the surrounding organs. This is an idealized linear optimization, as it assumes that there can be multiple beam intensities at each of the given beam angles, without regard for creating feasible, uniform aperture patterns. The resulting output at a given beam angle, looks something like the first panel in Figure 1.4. While this optimization can provide a good abstraction of the problem and insight to planners, it requires further, non-trivial processing to be deliverable, while meeting the treatment goals. This is because it fails to account for practical mechanical limitations, most notably, the requirement for the plan to consist of a finite number of uniform apertures.

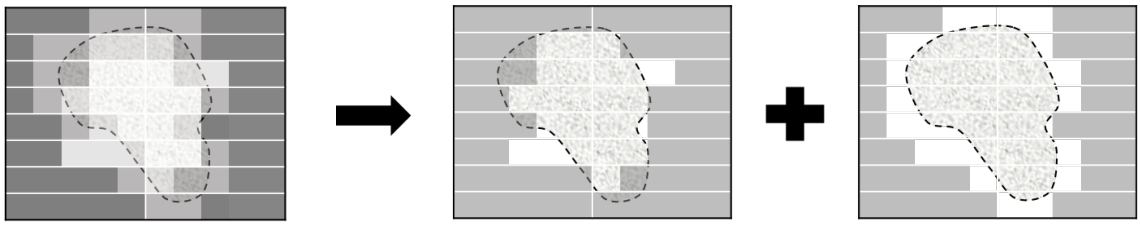


Figure 1.4: Simplified fluence map optimization divided into two uniform apertures.

The deliverability issues in Phase 2 are accounted for in Phase 3, which is the division of the fluence map into a finite number of uniform apertures. A simplified diagram of this process is captured in Figure 1.4. The image is rather deceptively simple, seeing as the division requires either secondary non-trivial optimization, or more commonly, an algorithmic step called leaf sequencing. Leaf sequencing consists of selecting both the aperture shapes (which are a product of leaf placement, leading to the name) and the intensities of the respective apertures. It can be implemented in a number of fashions, with fluence being discretized initially, followed by aperture selection, or both parts of the process being achieved simultaneously (Romeijn and Dempsey, 2008).

Regardless of how the decomposition is done, there is a potentially large loss of optimality between Phases 2 and 3 that comes about from running the FMO without full problem information available. Frequently, the aperture selection is done by selecting deliverable approximations of the FMO using rapid heuristics, which means that the delivered plan

degrades significantly post-optimization. Even when leaf sequencing is run to optimality as in Boland et al. (2004) and Taşkın et al. (2010), it can result in high numbers of apertures and dose homogeneity concerns (Salari et al., 2011).

1.1.3 Direct Aperture Optimization

Combining Phases 2 and 3 of the traditional step-and-shoot planning methodologies leads to the introduction of a method called direct aperture optimization (DAO). The DAO approach finds the globally optimal set of apertures without any intermediate stages. It does this by integrating the device-specific requirements along with the planning requirements into a single mixed-integer optimization problem.

DAO produces plans that are immediately deliverable, like the apertures shown in Figure 1.5, meaning the plans do not undergo further processing or degrade in subsequent phases. The drawback of using DAO is the introduction of new layers of complexity, to an already difficult large-scale problem. In place of a linear FMO problem and a heuristic or linearized leaf sequencing algorithm, DAO plans are the output of much more difficult large-scale mixed integer programming (MIP) models.

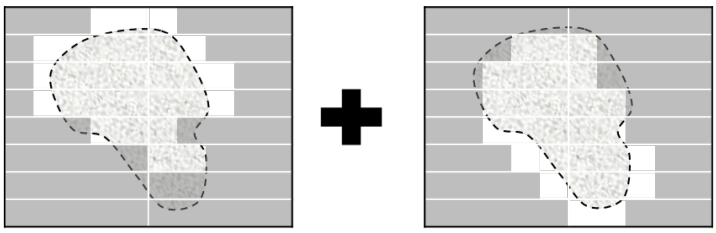
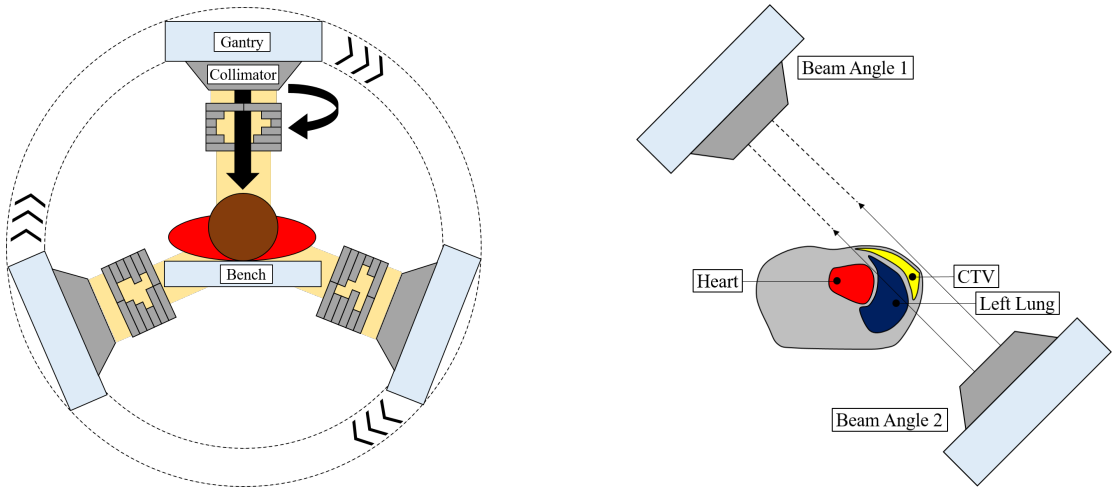


Figure 1.5: Simplified direct aperture optimization containing two apertures

1.2 Radiation Therapy for Left-Sided Breast Cancer

Breast cancer is the most frequently diagnosed cancer in Canadian women (Canadian Cancer Statistics Advisory Committee, 2018). It is presently the deadliest cancer to women in developing regions, while just recently dropping to the second deadliest in developed regions, after lung cancer (International Agency for Research on Cancer et al., 2012). Breast cancer has been shown to be a good candidate for adjuvant radiation therapy, as it is often detected in early stages, where a combination of breast-conserving surgery and



(a) General IMRT delivery setup. The gantry rotates the beam to a number of angles

(b) Axial cross-section of a breast cancer patient's chest cavity, in tangential IMRT setup

Figure 1.6: Contrasting IMRT delivery setups

radiation therapy has been show to be an equally effective alternative to a mastectomy (Canadian Cancer Statistics Advisory Committee, 2018, Miller et al., 2016).

Planning IMRT treatments for breast cancer patients differs from the planning in many other regions of the body. For example, in breast cancer IMRT the tumour is typically surgically removed prior to the treatment, to decrease risk of cancer recurrence. As a result, the CTV, which typically includes a carefully delineated tumour volume (often with a small uncertainty region around the tumour, to ensure coverage), in this case includes the entire breast volume, meaning the whole structure must be fully irradiated. The Phase 1 angle selection can also be largely simplified for breast cancer patients, seeing as effective plans can be designed with just a pair of beam angles that run tangential to the body. As shown in Figure 1.6, rather than choosing angles that surround the CTV as in Figure 1.6a, two tangential angles are chosen, as in Figure 1.6b, such that they encompass the entire target region, while keeping the bulk of the sensitive organs out of harms way (Kestin et al., 2000, Purdie et al., 2011).

In left-sided breast cancer, in particular, even with this tangential setup, parts of the heart are often still included in the treatment field. For this reason, apertures must be chosen carefully to minimize overdosages, increasing the risk level associated with the procedure (Wang et al., 2012). Studies have shown that even at low doses, radiation to the heart increases the risk of radiation-related heart disease; in correlation with volume of

heart tissue exposed (Darby et al., 2010, Hooning et al., 2007). For this reason, cardiac sparing must be made an explicit goal of left-sided breast cancer radiation therapy treatment planning, second only to CTV coverage.

The final characteristic of breast cancer treatment is its most challenging feature; the presence of unpredictable breathing motion throughout treatment. The motion occurs due to the expansion and contraction of a patient's lungs, leading to physical deformation and volumetric changes in both the CTV and heart throughout treatment (Quirk et al., 2014). These changes introduce potentially life-threatening uncertainty into the treatment process, both in terms of overdosing the heart, or under-dosing the CTV.

1.2.1 Conventional Methods for Addressing Breathing Motion

There are a number of strategies for mitigating the impact of breathing motion. The simplest and most conventional approach is the incorporation of a margin (sometimes referred to as the skin flash in breast cancer treatment planning) (Conroy et al., 2015, Keall et al., 2006, Purdie et al., 2011). The margin incorporates the full range of target motion into the CTV, meaning a much larger volume gets irradiated as if it were all part of the target. Margins provide the most conservative possible estimate in terms of guaranteeing target coverage, but this comes at the price of a higher risk of unnecessarily overdosing the sensitive organs (Conroy et al., 2015, Keall et al., 2006).

On the other end of the spectrum, there are methods such as breath hold, which use an active breathing control device in order to keep the patient at a specified inspiration level (typically 70-80% of max. inspiration) (Sixel et al., 2001, Wang et al., 2012). Keeping the patient near maximum inspiration should lead to an idealized treatment environment, as the heart is pushed the furthest distance away from the breast by the lungs, while motion is simultaneously inhibited. Although breath-hold methods are theoretically as close as possible to ideal treatment conditions on paper, they can be impractical for a number of reasons, ranging from physical difficulties with the device, as not all patients can tolerate the device, to extra time and medical resource requirements (Sixel et al., 2001).

1.2.2 Robust Optimization for Addressing Breathing Motion

Robust optimization can be used as a mathematical approach to immunize the treatment plan against a patient's breathing motion. Like a margin, this approach requires extra work at the planning stages. Unlike the margin, however, it is based off realizations of realistic motion scenarios, rather than an unrealistic aggregate of all possible scenarios.

Like with general IMRT planning, the input data for the robust methodology is delineated patient images, as captured using CT scanning technology. Unlike general IMRT, robust optimization uses four-dimensional computed tomography (4DCT), which collects a set of CT images are taken over time, rather than a single static image. For robust optimization for left-sided breast IMRT, this means capturing a patient's complete breathing cycle through a series of CT scans, resulting in a finite set of realizable breathing states. Using these images, a discrete number of breathing phases are defined. The time spent at each phase is then measured or estimated, and together, this information is used to define a base or *nominal* breathing pattern.

It is reasonable to expect that the patient's breathing pattern during treatment will deviate from this pattern, but it is unlikely to do so at an extreme, as a patient must still inhale and exhale throughout treatment (i.e., if a patient spent 50% of their nominal breathing pattern at exhale, spending 55% in that phase during the realized treatment is reasonable, but 95% is not). The maximum expected amount of deviation from this nominal breathing pattern defines the extreme points of the robust uncertainty set. Optimizing over this uncertainty set is far less conservative than placing the uncertainty around the location of the target region, itself, since the target is likely only in each of its extreme positions, even conservatively, only for a fraction of the patient's treatment time.

1.3 Mathematical Modelling of IMRT

Due to the inherent complexity of solving even the step-and-shoot problem, some fairly standard simplifications are used to make the problem more manageable and appropriate for mathematical modelling. The first is an abstraction of the beam into a grid of units called *beamlets* as depicted in Figure 1.7. The height of a beamlet is equal to the height of a leaf, whereas the width can be chosen based on the desired granularity of the solution, although it is typically on the order of 1×1 cm^2 (Romeijn and Dempsey, 2008).

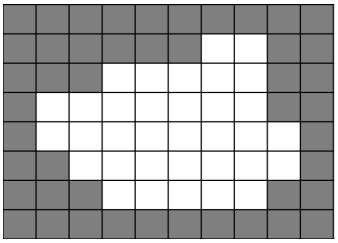


Figure 1.7: The beam is broken down into a 2D grid of units called beamlets.

For deterministic modelling, the dose delivered by each beamlet is assumed to be proportional to the total beamlet intensity or duration, and have no impact on it neighbouring beamlets. This means that a beamlet that is on for twice as long will give off twice the dose to the tissue it reaches. There is, realistically, a stochastic nature to this dose deposition; however, the effects have been found to be minimal enough to use the linear results with a fairly high level of confidence.

The target region contains a volume of human tissue that must also be transformed into a dataset for mathematical modelling. This transformation is done similarly to the discretization of the beam, in that tissue is divided into a 3D grid of units called voxels (volume pixels). The voxels segment the region, as shown in Figure 1.8, so that each structure may be defined by a specified voxel set. Any voxels with shared boundaries are allocated based on a priority queue; a process typically automated in imaging softwares.

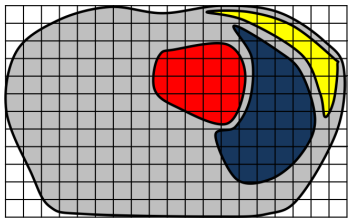


Figure 1.8: A 2D slice of the treatment region, divided into a grid of units called voxels.

This problem definition can be taken one step further by defining a relationship between the beam at a given angle and the voxels in the treatment region as a matrix. This matrix describes the dosimetric influence that each beamlet at a specified angle has on each voxel, and is therefore labeled a *dose influence matrix*. Figure 1.9 demonstrates the visual interpretation a single entry in a dose influence matrix.

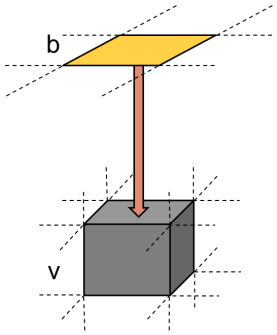


Figure 1.9: Dose received by voxel v when beamlet b is on, per unit time or beam intensity.

1.4 Depicting and Evaluating IMRT Plans

Once a treatment plan is generated, clinical metrics such as treatment time, CTV conformity, sparing of OARs and plan deliverability are used to assess plan quality. The dosage-based goals, such as sparing OARs and achieving high CTV conformity, can be summed up using a diagram called a dose volume histogram (DVH). Like the name implies, a DVH is a cumulative plot that has the percentage of the prescribed dose on the $x-$axis, and the percentage of volume receiving a specified dose on the y-axis. At a glance, this diagram tells physicians whether or not an adequate dose will be delivered to the CTV (e.g., will at least 99% of the CTV, by volume, receive at least 95% of the prescribed dose?). Similarly, a prohibitively high volume of dose to large sections of OARs, like the heart, can easily be detected. An ideal DVH would look like Figure 1.10.

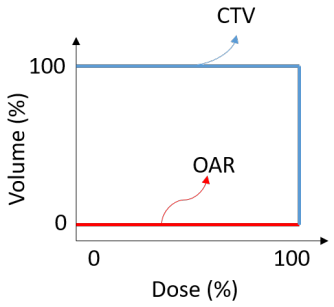


Figure 1.10: In an idealized DVH, 100% of the CTV receives exactly 100% of the prescribed dose while OARs receive 0%.

Plans may also be depicted from the beam's perspective, in terms of the intensity of dose delivered by each beamlet on a 2D grid called a fluence map (FM). In this thesis, these maps are depicted in greyscale, with lighter shades corresponding to higher dose intensities, as depicted in Figure 1.11. Both the DVH and FM diagrams are used in this thesis to gain insights into generated plans.

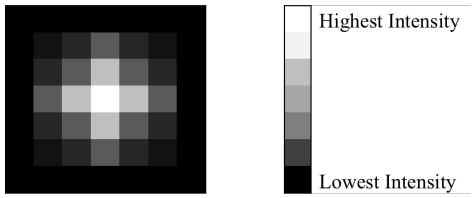


Figure 1.11: A perfectly symmetrical fluence map, concentrated on a central target.

1.5 Literature Review

The use of optimization to design radiation therapy treatment plans for cancer patients has become a classical problem in the field of operations research (Romeijn and Dempsey, 2008). While FMO has received the bulk of the attention in the radiation therapy literature over the years, the leaf sequencing phase has become a major topic of interest as well (Bortfeld, 2006).

Some authors have chosen to segment the fluence maps into discrete fluence levels initially, and then afterwards to apply sequencing algorithms (Gladwish et al., 2007, Kamath et al., 2003, Xia and Verhey, 1998). While others have found exact algorithms for leaf sequencing when the goal of minimizing beam-on-time (i.e., the total duration of radiation delivery) is selected (Langer et al., 2001, Siochi, 1999). A recent increase in efficiency in leaf sequencing with the goal of beam-on-time minimization has come from noting the similarity between the leaf sequencing component of the problem and polynomial-solvable network flow problems, which has been leveraged to produce polynomial solvable leaf-sequencing problems that minimize the total beam-on-time (Ahuja and Hamacher, 2005, Boland et al., 2004, Taşkın et al., 2010). A more realistic objective for the leaf sequencing stage of the problem is potentially minimizing the total apertures chosen or minimizing the total treatment time, however, these objectives have been shown to lead to NP-hard problems (Baatar et al., 2005).

The DAO approach was initially introduced using a simulated annealing algorithm to generate the uniform apertures (Shepard et al., 2002). This work was later augmented by a number of researchers who have built on the work using inexact, single-step solvers to generate direct aperture plans (Broderick et al., 2009, Li et al., 2003, Milette, 2008). While DAO remains much more difficult and time intensive to solve than FMO and its subsequent leaf-sequencing, progress has been made towards finding efficient, globally optimal solutions here as well, more specifically on problems where the goal is to minimize the beam-on-time of a treatment. Leveraging methods similar to the network models above in the subproblems, column generation approaches to solving the global problem have been introduced (Mahnam et al., 2017, Men et al., 2007, Romeijn et al., 2005, Salari and Unkelbach, 2013). While these innovations have been a remarkable step forward in DAO technology, the field remains in its infancy and further work must be done to incorporate the flexibility and advance methods of years worth of FMO research back into the DAO framework.

Robust optimization as a method for handling uncertainty in radiation therapy optimization for inter-fractional patient setup and organ motion uncertainty was proposed in

Chu et al. (2005). Intra-fractional uncertainties in pre-calculated features such as the dose influence matrix have also been addressed using robust optimization (Olafsson and Wright, 2006). The concept of applying robust optimization to mitigate intra-fractional uncertainty associated with patient breathing patterns was introduced in Chan et al. (2006). This work showed that a compromise could be provided between tumour elimination and organ sparing even under motion uncertainty. Further work by (Bortfeld et al., 2008, Chan et al., 2014, Mahmoudzadeh et al., 2013, 2015) have further solidified the robust model as a flexible and extensible model, that is amenable to including highly sophisticated objectives and constraints. Work in Mahmoudzadeh et al. (2016) has showed that the robust optimization can be solved very quickly using constraint generation methods, even with the additional variables on top of the large-scale FMO problem. Bortfeld et al. (2008) show that robust objectives can be formed in a similar manner to robust constraints, using an example of worst-case maximum heart dose in left-sided breast cancer optimization.

The impact of DAO on respiration motion in breast cancer has been examined using non-robust methodologies and a commercial planning software (Zhang et al., 2006). Similarly, Ahunbay et al. (2007) use a commercial planning system to generate DAO plans for breast cancer patients, but do not account for motion. In Ahunbay and Li (2007) motion is accounted for using a gating system, not robust, along with a commercial DAO planning system. Robust optimization in conjunction with DAO was used to mitigate the impact of the so called tongue-and-groove effect that results from IMRT devices (Salari et al., 2011). However, the uncertainty with regards to the tongue-and-groove effect differs quite a lot from breathing uncertainty. Solving a hybrid robust direct aperture optimization (RDAO) problem has been a topic of interest in recent years, however, due to the computational complexity of each of the two methods, to our knowledge, approaches are limited and their application to breast cancer, in particular, has yet to be examined.

1.6 Thesis Objectives

As DAO makes its way into off-the-shelf planning products and the field moves towards continuous delivery methods, finding new ways to incorporate clinical constraints and encourage scalability becomes increasingly important. To that end, this thesis aims to provide the following:

1. A unifying MIP framework to combine robust and DAO models
2. Fast, high-quality heuristic estimations for the difficult-to-solve DAO problem
3. Novel modelling techniques to increase solution efficiency

1.7 Thesis Outline

The remainder of the thesis is organized as follows, Chapter 2: Methodology and Modelling introduces the mathematical terminology used throughout this work, as well as explaining the mechanics of DAO and RDAO modelling. Chapter 3: Efficiency Improvement Techniques introduces the techniques designed to improve the solution time and bounds for solving the DAO and RDAO problem. The results of applying these models, first to a one-dimensional case, then a clinical case, are reported, and compared to their continuous counterparts in Chapter 4: Results. Finally, the thesis is concluded in Chapter 5: Conclusions, with a summary and recommendations for future work.

Chapter 2

Methodology and Modelling

This chapter describes the mathematical tools used to model the robust direct aperture optimization (RDAO) model. First, a fluence map optimization (FMO) planning approach and an existing robust framework are introduced. Then, our proposed direct aperture optimization (DAO) and angle-selection requirements are provided.

The chapter is broken down as follows: Section 2.1 introduces the application-specific notation used throughout this thesis. Section 2.2 introduces the FMO model, which serves as the base model for all subsequent formulations. Robust modelling strategies are introduced in Section 2.3, and the integration and construction of various direct aperture features is discussed in Section 2.4.

2.1 Notation

Any radiation therapy modelling constitutes two major components: the *beam of radiation* and the *region of interest*. The beam of radiation is modelled as a set of $b \in \mathcal{B}$ beamlets, where the index of beamlet b is dependent on both location within the multileaf collimator (MLC) and angle of the MLC.

The region of interest comprises all structures, $s \in \mathcal{S}$, within the patient's body that are exposed to the beam of radiation throughout the treatment. In the case of breast cancer, this set includes two elements: the target clinical target volume (CTV), T, and the heart, H. Each of these structures is broken up into a finite number of voxels, which may be denoted by $v \in \mathcal{V}_s$, for convenience. The prescribed level of dosage to each of the target voxels $v \in \mathcal{V}_T$ is denoted L_v.

The dose received by each voxel v when beamlet b is active is captured by the influence matrix $D_{v,b}$. For 4D planning techniques like robust optimization, the influence matrix must also take on a third dimension, which represents phase. The three dimensional parameter $D_{v,b,i}$ captures the influence that each beamlet b (at a specified angle) has on each voxel v during breathing phase i. Each phase $i \in \mathcal{I}$ also has an associated proportion p_i, which is the proportion of time that the patient spends in phase i, over the course of a complete breathing cycle.

Throughout this thesis, sets are denoted by calligraphic font, variables and constants are italicized and vectors are written in bolded text.

2.2 Nominal FMO Model

We first present the basic deterministic form of intensity modulated radiation therapy (IMRT) which uses FMO and assumes no motion uncertainty. The resulting, idealized plans consist of a set of beamlet intensities at each angle, which best conform to the shape of the target, without compromising the organs at risk (OARs).

Decision variable:

w_b represents the intensity of beamlet b.

Mathematical FMO model:

(**M-FMO**)

$$\min \sum_{s\in\{T,H\}} \frac{c_s}{|\mathcal{V}_s|} \sum_{v\in\mathcal{V}_s} \sum_{b\in\mathcal{B}} \sum_{i\in\mathcal{I}} p_i D_{v,b,i} w_b \tag{2.1}$$

$$\text{s.t.} \sum_{b\in\mathcal{B}} \sum_{i\in\mathcal{I}} p_i D_{b,v,i} w_b \geq L_v \qquad \forall v \in \mathcal{V}_T, \tag{2.2}$$

$$w_b \geq 0 \qquad \forall b \in \mathcal{B}, \tag{2.3}$$

where:

c_s is the objective weight for each structure s.

The objective of our **M-FMO** model is to minimize the weighted average of the expected dose to each structure. By setting the CTV weighting to 0, the objective can be

simplified to minimizing the average heart dose. The requirement of delivering the prescribed dose to every voxel in the CTV is ensured by constraints (2.2). Since this is the nominal model, the proportion of time spent in each breathing phase, p_i, is assumed to be known and without any uncertainty.

In addition to handling multiple breathing phases, the **M-FMO** model can handle simpler forms of the FMO problem that are often tackled in clinic. For static treatments, where organs are not moving, for example, the dose influence matrix has only two dimensions. This is a simple case of the above model, where there is only one phase, meaning $|\mathcal{I}| = 1$ and $p_1 = 1$.

2.3 Robust FMO Model

In general, robust optimization removes the assumption of parameter certainty. For this problem, it is realistic to relax this assumption of certainty regarding the proportion of time spent in each breathing phase, p_i. Robust optimization can then be used to ensure that even in the worst-case realization of this uncertainty, the CTV still receives its prescribed dosage.

Formulation-wise, the robust setup looks very similar to the **M-FMO** model above, but in constraints (2.2), the deterministic proportion value, $\mathbf{p}$ is substituted for an uncertain, $\tilde{\mathbf{p}}$ value. This $\tilde{\mathbf{p}}$ sums to 1, and is bounded by a set of upper and lower deviations from the nominal proportions, denoted as $\bar{\mathbf{p}}$ and $\underline{\mathbf{p}}$, respectively. Mathematically, these requirements can be written,

$$p_i - \underline{p}_i \leq \tilde{p}_i \leq p_i + \bar{p}_i \quad \forall i \in \mathcal{I}, \tag{2.4}$$

$$\sum_{i \in \mathcal{I}} \tilde{p}_i = 1. \tag{2.5}$$

It can also be affirmed that because $\tilde{p}_i$ is a proportion of time in a breathing phase, the following is also true,

$$0 \leq \tilde{p}_i \leq 1 \quad \forall i \in \mathcal{I}. \tag{2.6}$$

We denote the uncertainty set of $\tilde{\mathbf{p}}$ as $\mathcal{P}$, where,

$$\mathcal{P} = \{\tilde{\mathbf{p}} \in \mathbb{R}^{|\mathcal{I}|} | (2.4); (2.5); (2.6)\}. \tag{2.7}$$

While the above requirements limit the total number of possible realizations of $\tilde{\mathbf{p}}$, there remain infinite possible breathing pattern realizations (with the exception of the trivial case, in which $\underline{\mathbf{p}} = \bar{\mathbf{p}}$, since that uncertainty set only includes one element).

Subbing the new $\tilde{\mathbf{p}}$ into constraints (2.2), yields an infinite number of robust constraints of the form,

$$\sum_{b\in\mathcal{B}}\sum_{i\in\mathcal{I}} \tilde{p}_i D_{b,v,i} w_b \geq \quad L_v \quad \forall v \in \mathcal{V}_T,\ \forall \tilde{\mathbf{p}} \in \mathcal{P}. \tag{2.8}$$

Constraints (2.8) are intractable, but Chan et al. (2006) show that the equivalent robust counterpart of these constraints is both tractable and linear, at the expense of the introduction of more variables into the model. The robust counterpart is derived by reformulating the left hand-side of constraints (2.8) as a separate minimization subproblem for each $v \in \mathcal{V}_T$, as follows:

$$\begin{aligned} \min_{\hat{p}} \ & \sum_{i\in\mathcal{I}}\sum_{b\in\mathcal{B}} \hat{p}_i D^t_{b,v,i} w_b \\ \text{s.t.} \ & \sum_{i\in\mathcal{I}} \hat{p}_i = 1, \\ & (p_i - \underline{p}_i) \leq \hat{p}_i \leq (p_i + \bar{p}_i) \quad \forall i \in \mathcal{I}. \end{aligned} \tag{2.9}$$

The next step involves taking the finite dual of this subproblem, with dual variables $y_{i,v}$ and subbing it back into the original model. Details of this derivation, are given in Appendix A. The resultant robust fluence map optimization (RFMO) model has $|\mathcal{V}_T|\times(|\mathcal{I}|+1)$ new variables and $|\mathcal{V}_T| \times |\mathcal{I}|$ new constraints, in place of the $|\mathcal{V}_T|$ nominal CTV dose constraints (2.2), and is formulated as follows:

New decision variable:

$y_{i,v}$ is the subproblem dual variable for the worst-case realization of phase i for voxel v.

Mathematical RFMO model:

(**M-RFMO**)

$$\min \sum_{s\in\{T,H\}} \frac{c_s}{|\mathcal{V}_s|} \sum_{v\in\mathcal{V}_s} \sum_{b\in\mathcal{B}} \sum_{i\in\mathcal{I}} p_i D_{v,b,i} w_b$$

$$\text{s.t.} \sum_{i\in\mathcal{I}} \left[\underline{p}_i y_{0,v} - (\underline{p}_i + \bar{p}_i) y_{i,v} + (p_i - \underline{p}_i) \sum_{b\in\mathcal{B}} D_{b,v,i} w_b \right] \geq L_v \qquad \forall v \in \mathcal{V}_T, \quad (2.10)$$

$$\sum_{b\in\mathcal{B}} D_{b,v,i} w_b - y_{0,v} + y_{i,v} \geq 0 \qquad \forall i \in \mathcal{I}, \forall v \in \mathcal{V}_T, \quad (2.11)$$

$$y_{0,v} \text{ URS} \qquad \forall v \in \mathcal{V}_T, \quad (2.12)$$

$$y_{i,v} \geq 0 \qquad \forall i \in \mathcal{I}, \forall v \in \mathcal{V}_T, \quad (2.13)$$

$$w_b \geq 0 \qquad \forall b \in \mathcal{B}.$$

Constraints (2.10), (2.11), (2.12) and (2.13) accomplish the same goal as (2.8), while also being finite and linear. The objective for the **M-RFMO** model remains the same as for the **M-FMO** model, as we are more concerned with minimizing an expected dose to the heart, rather than the worst-case dose, seeing as a full treatment typically spans a number of fractions, or treatment sessions.

2.4 Proposed Robust DAO Model

The **M-FMO** and **M-RFMO** models introduced in the previous sections account for clinical requirements and delivery uncertainty, however, they do not accommodate the delivery limitations that come about from IMRT delivery equipment. As a result, the output plans, or fluence maps, are not deliverable by our linear accelerator (linac) device. This section introduces deliverability constraints, to ensure that our models output realistic plans. Deliverable plans are made up of a set of beams with uniform intensities, shaped by the delivery device's MLC leaves, called apertures.

The new DAO model will output deliverable plans with the following properties, that were missing from previous models:

- An optimal plan with a preselected number of apertures
- An allocation of the apertures to each beam angle

- The (feasible) placement of MLC leaves for each aperture
- The (uniform) beam intensity for each of the selected apertures

The rest of this section is structured as follow: the necessary steps for integrating the integer DAO constraints into the previously introduced models are covered in Section 2.4.1. Aperture uniformity requirements are then addressed in Section 2.4.2, followed by angle considerations in Section 2.4.3. Leaf placement is addressed in Section 2.4.4, while Section 2.4.5 addresses additional requirements and modelling extensions. The full **M-RDAO** model is then assembled in Section 2.4.6.

2.4.1 DAO-Specific Constants and Variables

With the introduction of DAO, the concepts of both relative beamlet location within the vectorized beam and individual aperture contribution gain much more importance than in the initial FMO modelling. This section addresses A) the necessary additional information for tracking beam layout, and B) a variable adjustment that is needed to integrate aperture separation into our models.

A) Beam Layout

As shown in Figure 2.1, there are varying levels of information about beamlet relationships available to the model. For **M-FMO** and **M-RFMO** models, location within the beam is unimportant for a given beamlet b, as the relevant dosage information is captured in the dose influence matrix and beamlet intensities are optimized independently. In terms of Figure 2.1, that means only layer (a), or beamlet index information, is required.

In DAO we need further information, such as which angle a beamlet belongs to, the dimensions of that angle and where a particular beamlet is situated relative to other beamlets within that angle, in order to generate deliverable apertures. Graphically, this is shown in Figure 2.1. Mathematically, we may define a set of angles, $\theta \in \Theta$, and their corresponding beamlets, within the existing vector of beamlets $b \in \mathcal{B}_\theta$. We may also introduce sets of row and column coordinates, $\mathcal{Q}_\theta$ and $\mathcal{K}_\theta$, respectively.

B) DAO-Adjusted Variables

In order to capture the more detailed, deliverable apertures, we also need to adjust the w_b decision variable used above. Now, rather than just being concerned with the intensity

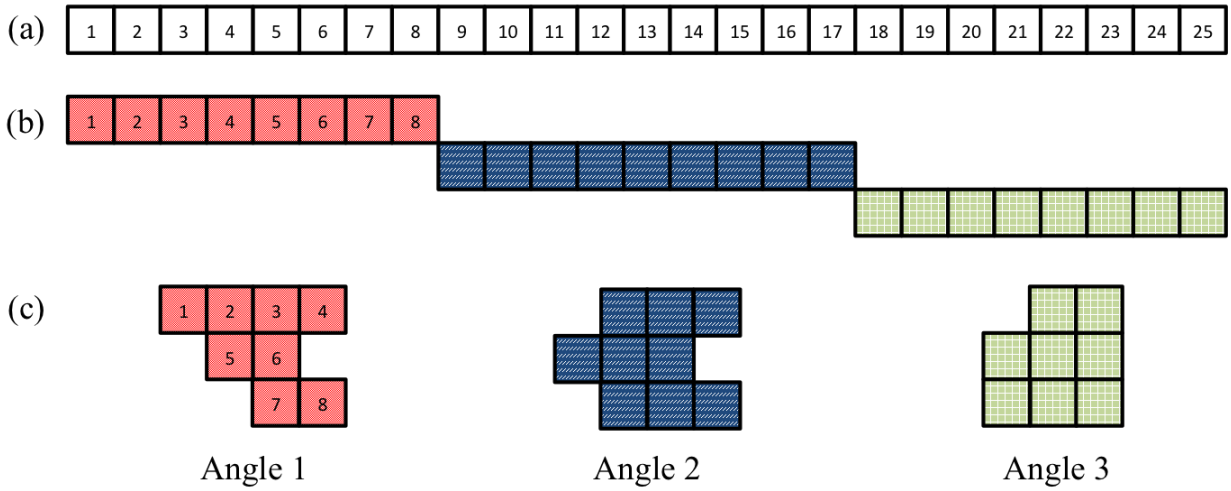


Figure 2.1: Different levels of detail required for the various DAO beamlet constraints. (a) is just the beamlet indices, (b) has beamlet indices as well as angle association and (c) maintains index, angle and row/column adjacency information. FMO methods only require (a), but DAO requires (b) and (c).

of each beamlet, we also must consider the aperture $a \in \mathcal{A}$ to which a beamlet intensity belongs. In this work, the total number of allowable apertures, $|\mathcal{A}|$, is chosen prior to optimization and will dictate the second dimension of a now two-dimensional $w_{b,a}$ decision variable. This new decision variable, however, does not fit with our **M-FMO** and **M-RFMO** models above, where each beamlet intensity was a single, independent decision. Therefore, in our DAO models, the w_b from the **M-FMO** and **M-RFMO** models is replaced with an updated w'_b, which is defined as a sum of the fluences across apertures as follows:

$$w'_b = \sum_{a \in \mathcal{A}} w_{b,a} \quad \forall b \in \mathcal{B}. \tag{2.14}$$

Using this update, all of the constraints introduced in the following sections can be added directly into the original models with no further effort, to create complete DAO and RDAO models.

2.4.2 Uniformity Constraints

The DAO methodology requires that apertures be of a uniform dose, which means that all beamlets that are active within a given aperture take on the same intensity value. This reflects reality, as any exposed beamlets in a given aperture allow the beam to come through for the same duration or dose-intensity.

Mathematically, this idea of uniformity relies on the concept of an *on* and *off* setting for each beamlet. This requirement may be enforced through a set of binary variables, denoted as x, which represent whether a single unit of the collimator is open (i.e., no leaf is blocking the radiation) and delivering a specified intensity, or shut (blocked by a leaf). This on-off decision is made on a per-beamlet basis, where each beamlet becomes defined by two variables: an intensity variable, w as in the previous models, and a setting variable, x.

When initially tasked with applying this uniformity constraint, it seemed natural to want to multiply the variables to achieve the on and off effect. This led to the following set of variables and constraints.

New decision variables:

$x_{b,a}$ indicates the active or inactive state of beamlet b in aperture a.

f_a is the uniform intensity (or fluence) of all active beamlets in aperture a.

And the following set of non-linear constraints:

$$\begin{aligned} w_{b,a} &= f_a x_{b,a} \quad \forall b \in \mathcal{B}, a \in \mathcal{A}, \\ x_{b,a} &\in \{0,1\} \quad \forall b \in \mathcal{B}, a \in \mathcal{A}. \end{aligned}$$

These constraints enforce all beamlets in aperture a to be 0 or have the same intensity, f_a. Unfortunately, introducing non-linearity into our mixed integer programming (MIP) problem makes it even more difficult to solve, so the above constraints were not suited to the problem at hand. Using the same variables we linearize the constraints, as follows:

$$w_{b,a} \leq M x_{b,a} \qquad \forall b \in \mathcal{B}, a \in \mathcal{A}, \tag{2.15}$$
$$w_{b,a} \leq f_a + M(1 - x_{b,a}) \qquad \forall b \in \mathcal{B}, a \in \mathcal{A}, \tag{2.16}$$
$$w_{b,a} \geq f_a - M(1 - x_{b,a}) \qquad \forall b \in \mathcal{B}, a \in \mathcal{A}, \tag{2.17}$$
$$f_a \geq 0 \qquad \forall a \in \mathcal{A}, \tag{2.18}$$
$$x_{b,a} \in \{0, 1\} \qquad \forall b \in \mathcal{B}, a \in \mathcal{A}, \tag{2.19}$$

where:

M is a very large number.

Constraints (2.16) and (2.17) force all $w_{b,a}$ to take on the intensity of f_a, if beamlet b in aperture a is open ($x_{b,a} = 1$), while (2.15) along with non-negativity constraint (2.3) from the original **M-FMO** model, force $w_{b,a}$ to zero, if the beamlet is closed ($x_{b,a} = 0$). Together these constraints enforce uniform apertures, by restricting intensities within each aperture to only two values:

$$w_{b,a} = \begin{cases} f_a & \text{if beamlet } b \text{ is active } (x_{b,a} = 1), \\ 0 & \text{if beamlet } b \text{ is inactive } (x_{b,a} = 0). \end{cases}$$

2.4.3 Aperture Selection

When applied to all apertures, the uniformity constraints above would allow beamlets within a single aperture to span over multiple angles, so long as they are all relegated to the same intensity value. Clearly, this is not possible since there is only one beam, so apertures can only include beamlets from a single beam angle.

One way to mitigate this problem is by pre-selecting the number of apertures allowed per angle. This means that if a plan with 6 apertures and 2 beam angles is desired, it may be arbitrarily decided before optimizing the treatment that 3 apertures will be delivered from each side, regardless of the relative complexity of the dose needed at either side.

Practically, the optimal allocation of beam angles per apertures is not known upfront, so making this decision can result in suboptimal solutions. For this reason, we chose to take a more flexible approach, which allows for the algorithm to choose how many apertures to allocate to each angle. This means, in the example above, rather than enforcing 3 apertures at each angle, the algorithm may choose to divide things as 2 and 4, or 5 and

1, depending on the relative benefit to the patient.

We model this angle-aperture allocation using the new variable:

$u_{a,\theta}$ indicates whether or not aperture a is on at angle θ.

Along with the constraints:

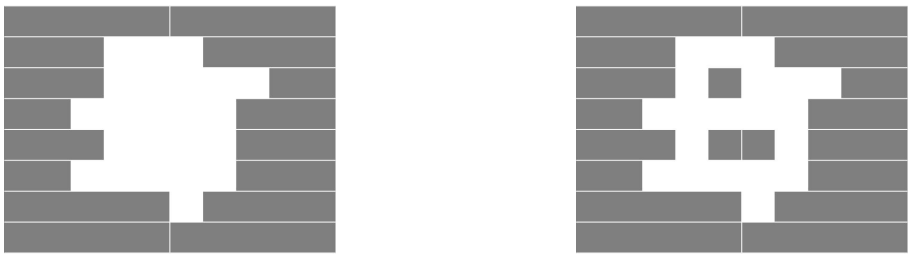
$$\sum_{b \in \mathcal{B}_\theta} x_{b,a} \leq |\mathcal{B}_\theta| u_{a,\theta} \qquad \forall a \in \mathcal{A}, \theta \in \Theta, \tag{2.20}$$

$$\sum_{\theta \in \Theta} u_{a,\theta} = 1 \qquad \forall a \in \mathcal{A}. \tag{2.21}$$

Constraints (2.20) force beamlets to be off unless the whole angle is active, and constraints (2.21) restrict the number of active angles to 1 per aperture.

2.4.4 Island Removal

The constraints above enforce the beam requirements, however, they fail to address the mechanical restrictions associated with beam modulation. Since the modulation is performed by sets of opposing, linearly extending tungsten leaves, the model must output beam setups that may be physically realized by this mechanism. Mathematically, this means constraining against breaks in the leaves, which result in detached sections called islands, shown in Figure 2.2.



(a) MLC leaves form a deliverable aperture (b) Undeliverable MLC setup with islands

Figure 2.2: Possible MLC realizations with the current DAO constraints

For the sake of exposition, we assume the beam is a $|\mathcal{Q}|$ row $\times$ $|\mathcal{K}|$ column rectangle at each angle and that the leaves may traverse the entire span of the collimator from both the

left and right sides. It is possible to implement non-uniform subsets of these constraints, i.e., non-rectangular collimator shapes or traversals and differently shaped angles, however, the notation becomes much messier.

We denote the variables using a dimension for each column, row, angle and aperture, for clarity. In reality, these variables only need two dimensions (beamlet index b and aperture a), since a beamlet index can be converted back and forth from the row, column, angle domain to beamlet location domain. The mapping looks as follows $[\mathcal{Q}, \mathcal{K}, \Theta] \rightarrow \mathcal{B}$, and uses the formula $b = \sum_{\theta'=0}^{\theta-1} |\mathcal{B}_{\theta'}| + |\mathcal{K}| \times (q-1) + k$, where $|\mathcal{B}_0| = 0$, an example of which is depicted graphically in Figure 2.3.

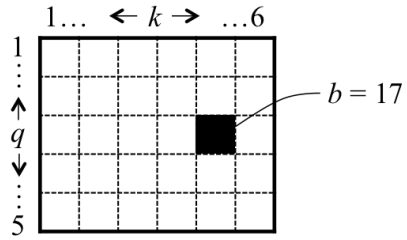


Figure 2.3: Mapping beamlets to location indices. Assuming that this is the first angle ($\theta = 1$), we have $b = |\mathcal{B}_0| + |\mathcal{K}| \times (q-1) + k = 0 + 6 \times (3-1) + 5 = 17$.

We next derive the island removal formulations for two potential types of MLC setup: in case A) we assume that collimator leaves cannot extend past the centre of the collimator, and in case B) we relax this assumption and formulate a more complex setup. Finally, in part C) we introduce and compare our relaxed method with an alternate relaxation approach.

A) Collimator Leaves Stop at the Centre

If we assume that leaves do not extend beyond the centre of the MLC, we can use the existing variables to enforce the extra feature, as follows:

$$x_{q,k+1,\theta,a} \geq x_{q,k,\theta,a} \qquad \forall k \in \{1, \cdots, \left\lfloor \frac{|\mathcal{K}|}{2} \right\rfloor - 1\}, q \in \mathcal{Q}, \theta \in \Theta, a \in \mathcal{A}, \tag{2.22}$$

$$x_{q,k,\theta,a} \geq x_{q,k+1,\theta,a} \qquad \forall k \in \{\left\lfloor \frac{|\mathcal{K}|}{2} \right\rfloor + 1, \cdots, |\mathcal{K}| - 1\}, q \in \mathcal{Q}, \theta \in \Theta, a \in \mathcal{A}. \tag{2.23}$$

Constraints (2.22) restrict the left fingers, while constraints (2.23) restrict the right fingers. They accomplish this restriction by ensuring that the binary variable closer to the centre of the collimator is always greater than the proceeding variable in its row. What this does, conceptually, is it requires all beamlets in a given row, on a given side to be active after any one beamlet is activated. So if a beamlet on the middle of the left side is on, the beamlet after it has to be ≥ 1, meaning it is also on. This prevents undeliverable, discontinuous leaves from occuring, as in Figure 2.2b.

B) Collimator Leaves May Pass the Centre

Realistically leaves can extend past the centre of the MLC, as in Figure 2.4.

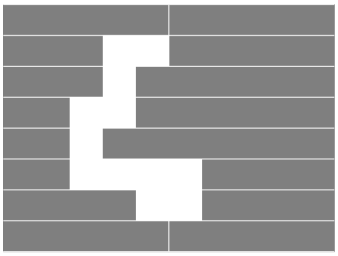


Figure 2.4: Right-handed collimator leaves (rows $q \in \{3, 4, 5\}$) extend past the centre

To restrict the leaves, the binary on-off constraints, $x_{q,k,\theta,a}$ from the previous section, along with two additional sets of binary variables, $l_{b,a}$ and $r_{b,a}$ can be used to represent the continuous leaves extended from the left and right side of the collimator, respectively. As with the uniformity constraint, every beamlet becomes further defined by four variables, $w_{b,a}, x_{b,a}, r_{b,a}$ and $l_{b,a}$.

The left leaf's open position (not extended over a beamlet) can be defined as $l_{b,a} = 1$, and similarly, an open right leaf is indicated by $r_{b,a} = 1$. If $l_{b,a} = 1$ and $r_{b,a} = 1$ the beamlet is open, meaning it is on, or $x_{b,a} = 1$. If either $l_{b,a}$ or $r_{b,a} = 0$, the beamlet is closed, and both sides cannot cover the same beamlet simultaneously, meaning they cannot both be 0 and $l_{b,a} + r_{b,a} \geq 1$.

The formal new variable definitions are as follows:

$l_{b,a}$ indicates whether a left leaf in aperture a is blocking beamlet b (0) or open (1).

$r_{b,a}$ indicates whether a right leaf in aperture a is blocking beamlet b (0) or open (1).

These constraints create continuous non-overlapping leaves:

$$l_{q,k+1,\theta,a} \geq l_{q,k,\theta,a} \qquad \forall k \in \mathcal{K}', q \in \mathcal{Q}, \theta \in \Theta, a \in \mathcal{A}, \tag{2.24}$$

$$r_{q,k,\theta,a} \geq r_{q,k+1,\theta,a} \qquad \forall k \in \mathcal{K}', q \in \mathcal{Q}, \theta \in \Theta, a \in \mathcal{A}, \tag{2.25}$$

$$x_{q,k,\theta,a} = -1 + l_{q,k,\theta,a} + r_{q,k,\theta,a} \qquad \forall k \in \mathcal{K}, q \in \mathcal{Q}, \theta \in \Theta, a \in \mathcal{A}, \tag{2.26}$$

$$l_{q,k,\theta,a}, r_{q,k,\theta,a} \in \{0,1\} \qquad \forall k \in \mathcal{K}, q \in \mathcal{Q}, \theta \in \Theta, a \in \mathcal{A}, \tag{2.27}$$

where:

$\mathcal{K}' = \{1, ..., |\mathcal{K}| - 1\}$, since we are enforcing pair-wise positions of beamlets, and have one degree of freedom per row.

As a result of these constraints,

$$x_{b,a} = \begin{cases} 1 & \text{if } l_{b,a} = 1 \cup r_{b,a} = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Constraints (2.24) force every beamlet to the right of an inactive collimator leaf on the left-extending leaf to also be inactive. Constraints (2.25) enforce the same constraint for opposing direction leaves. After making these leaf-sequencing choices, constraints (2.26) set the on-off state of a given beamlet based on whether or not it is blocked by the MLC leaves.

C) Comparison with an Alternate Past-Centre Approach

An alternative approach to the full MLC no islands constraint was proposed in a paper about FMO leaf sequencing by Boland et al. (2004). Their method uses a similar framework to our past-centre approach, except that their method requires an additional l and r variable per row. This extra variable represents the "off" position of each leaf, and it gives their formulation the latitude to reduce the number of leaf restriction constraints from $|\mathcal{K}| - 1$ per row, to a single constraint per row.

Rather than each $l_{b,a}$ and $r_{b,a}$ variable representing a beamlet, in this model, they each indicate whether or not a row has ended, as illustrated in Figure 2.5. Their associated constraints exploit the continuous nature of the MLC leaves by finding the index of the last covered beamlet from each leaf from both directions, allowing the same apertures as in the previous method to be defined.

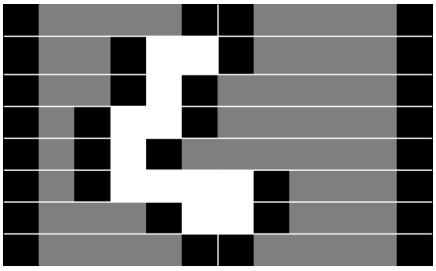


Figure 2.5: Same aperture as Figure 2.4, with modelling changes highlighted by darker beamlets. The column of darker beamlets at the far sides of the MLC are the "off" setting for each leaf. The inner darker beamlets are the active l and r variables

In Boland et al. (2004), the variable definitions are as follows:

$l_{b,a}$ indicates whether a left beamlet b in aperture a is the last in its row (1) or not (0).

$r_{b,a}$ indicates whether a right beamlet b in aperture a is the last in its row (1) or not (0).

These constraints are defined as follows:

$$\sum_{k \in \mathcal{K}''} l_{q,k-1,\theta,a} = 1 \qquad \forall q \in \mathcal{Q}, \theta \in \Theta, a \in \mathcal{A}, \tag{2.28}$$

$$\sum_{k \in \mathcal{K}''} r_{q,k,\theta,a} = 1 \qquad \forall q \in \mathcal{Q}, \theta \in \Theta, a \in \mathcal{A}, \tag{2.29}$$

$$x_{q,k,\theta,a} = \sum_{\iota=0}^{k-1} l_{q,k+\iota,\theta,a} - \sum_{\iota=1}^{k} r_{q,k+\iota,\theta,a} \qquad \forall k \in \mathcal{K}, q \in \mathcal{Q}, \theta \in \Theta, a \in \mathcal{A}, \tag{2.30}$$

where:

$\mathcal{K}'' = \{1, ..., |\mathcal{K}| + 1\}$, to account for the dummy "off" beamlets.

The difference in mechanisms of these two constraints can be observed by examining how a single row is enforced. Below is the second row ($q = 2$) of Figures 2.4 and 2.5, which has the 4^{th} and 5^{th} beamlets active, while all others are blocked.

	Proposed Island Removal										Boland et al. (2004) Island Removal											
k	1	2	3	4	5	6	7	8	9	10	0	1	2	3	4	5	6	7	8	9	10	11
L	0	0	0	1	1	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0	-
R	1	1	1	1	1	0	0	0	0	0	-	0	0	0	0	0	1	0	0	0	0	0
x	0	0	0	1	1	0	0	0	0	0	-	0	0	0	1	1	0	0	0	0	0	-

In practice, both types of constraints result in the same feasible set of x variables. Empirically, however, our proposed formulation (2.24 - 2.27) was found to work more effectively. This is likely due to the infeasibility introduced when attempting to propagate values in the linear program (LP) relaxation of the (2.28) and (2.29), which is not seen in the LP relaxation of (2.24) and (2.25). This could make it more difficult for the solver to find feasible solutions; an effect that is demonstrated by an example in Appendix B.

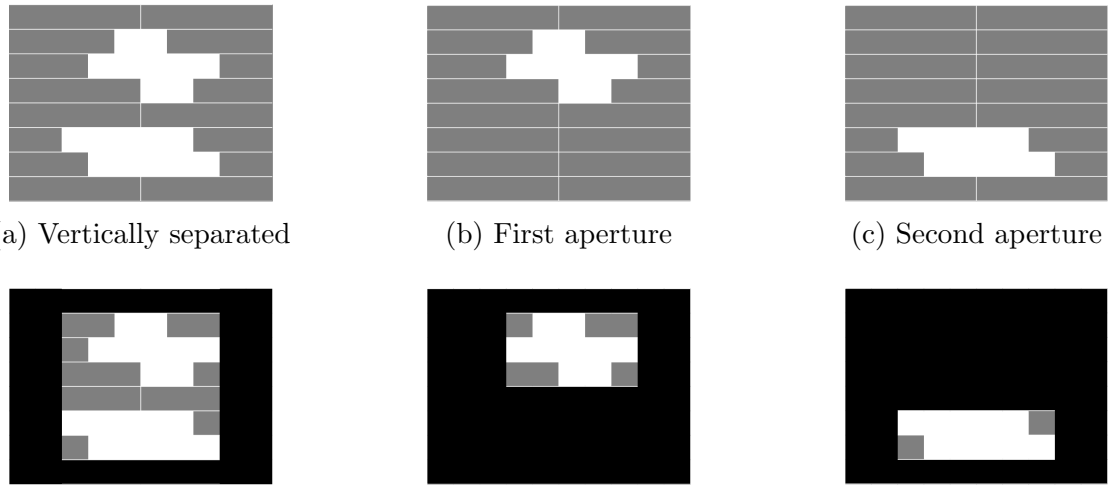
2.4.5 Extensions for Aperture Continuity

The previous sections ensure that our output DAO plans can be feasibly delivered by the IMRT equipment, however, in order to be integrated into clinical treatments, the plans may also have to meet additional clinical standards. These standards include restricting apertures shapes to eliminate forms of A) vertical and B) horizontal discontinuities.

A) No Vertical Breaks

Multiple separated groups of beamlets within each aperture can be undesirable, due to leakage between adjacent closed leaves. It may also be practically difficult to deliver, as some devices are mechanically unable to have opposing leaves within the same row meet, requiring pairwise leaves to have at least a small gap in between them, to avoid collisions. Luckily, these phenomena can be largely avoided by creating continuous segments that are mostly covered by jaws.

If a clinician was presented with multiple grouping of apertures, as in Figure 2.6, for example, they would typically, manually create separate apertures. This would increase the total number of apertures by one. Since the total number of allowed apertures are fixed, we can design our constraints to enforce the same principle by restricting each aperture to be vertically continuous using a per-row activation constraint, in a similar fashion to the continuous row constraints in the previous section. We require the introduction of $2 \times |\mathcal{Q}| \times |\Theta| \times |\mathcal{A}|$ new binary variables, denoted $\bar{\mathbf{j}}$ and $\underline{\mathbf{j}}$, to represent the upper and lower jaws, respectively (left and right jaws are not considered here, they can simply be



(a) Vertically separated (b) First aperture (c) Second aperture

(d) Black shows jaw block (e) New jaw block ap. 1 (f) New jaw block ap. 2

Figure 2.6: Deliverable aperture with clinically unacceptable vertical break, separated to two apertures. Images (d-f) use darker colour to show jaw coverage in (a-c).

calculated in post-processing stages). For convenience, we also add a summary variable **j**, which indicates whether or not a row is active.

The new binary variables are defined as follows:

$\bar{j}_{q,\theta,a}$ indicates if the upper jaw at angle θ, aperture a is blocking row q (0) or open (1).

$\underline{j}_{q,\theta,a}$ indicates if the lower jaw at angle θ, aperture a is blocking row q (0) or open (1).

$j_{q,\theta,a}$ indicates whether angle θ, aperture a, row q is blocked (0) or open (1).

Constraints:

$$j_{q,\theta,a} = -1 + \bar{j}_{q,\theta,a} + \underline{j}_{q,\theta,a} \qquad \forall q \in \mathcal{Q}, \theta \in \Theta, a \in \mathcal{A}, \tag{2.31}$$

$$j_{q,\theta,a} \leq \sum_{k \in \mathcal{K}} x_{q,k,\theta,a} \qquad \forall q \in \mathcal{Q}, \theta \in \Theta, a \in \mathcal{A}, \tag{2.32}$$

$$|\mathcal{K}| \times j_{q,\theta,a} \geq \sum_{k \in \mathcal{K}} x_{q,k,\theta,a} \qquad \forall q \in \mathcal{Q}, \theta \in \Theta, a \in \mathcal{A}, \tag{2.33}$$

$$\bar{j}_{q,\theta,a} \leq \bar{j}_{q+1,\theta,a} \qquad \forall q \in \mathcal{Q}', \theta \in \Theta, a \in \mathcal{A}, \tag{2.34}$$

$$\underline{j}_{q+1,\theta,a} \leq \underline{j}_{q,\theta,a} \qquad \forall q \in \mathcal{Q}', \theta \in \Theta, a \in \mathcal{A}, \tag{2.35}$$

$$j_{q,\theta,a}, \bar{j}_{q,\theta,a}, \underline{j}_{q,\theta,a} \in \{0,1\} \qquad \forall q \in \mathcal{Q}, \theta \in \Theta, a \in \mathcal{A}, \tag{2.36}$$

where:

$\mathcal{Q}' = \{1, ..., |\mathcal{Q}| - 1\}$, as once again, pairwise comparison has a degree of freedom.

Here, constraints (2.32) enforce at least one beamlet to be on in the active rows, while constraints (2.33) functions as an off switch if either jaw is blocking the row. Constraints (2.34) and (2.35) force consistent jaw motion.

B) No Horizontally Disconnected Rows

There is a similar incentive to avoid difficult, leakage-prone setups such as leaf collisions and disconnected rows. Leaf collisions may occur when the right leaves extend beyond adjacent row left leaves, as in Figure 2.7a. If the linear leaf motors become even the tiniest bit misaligned, these leaves may bump into each other during setup, so clinicians tend to choose plans that mitigate this risk.

(a) Leaf collision (b) Singleton beamlet

Figure 2.7: Examples of undesirable pairwise row behaviour

Similarly, disconnected rows are those that have no vertical beamlet connection between adjacent rows, leading to singleton beamlets and other horizontally separate segments, as

in Figure 2.7b. Since dose uncertainty primarily arises around the edges of leaves, plans with very small and separated sections will result in unwanted uncertainty . Works such as Boland et al. (2004) have introduced preventative measures for avoiding collisions, but their constraints still allow rows to become detached vertically.

To prevent all horizontal detachment, we propose a set of constraints that require all active, adjacent rows to share at least one active beamlet. Starting at the left side, this equation would be formulated as follows:

$$\sum_{\delta=1}^{k} x_{q,\delta,\theta,a} \leq \sum_{k\in\mathcal{K}} x_{q,k,\theta,a} - 1 + (1 - j_{q,\theta,a}) + (1 - j_{q-1,\theta,a}) + \sum_{\delta=1}^{k} x_{q-1,\delta,\theta,a}$$
$$\forall k \in \mathcal{K}, q \in \mathcal{Q}'', \theta \in \Theta, a \in \mathcal{A},$$

where:

$$\mathcal{Q}'' = \{2, \ldots, |\mathcal{Q}|\}.$$

The equation may be simplified to:

$$j_{q,\theta,a} + j_{q-1,\theta,a} - \sum_{\delta=k+1}^{|\mathcal{K}|} x_{q,\delta,\theta,a} \leq 1 + \sum_{\delta=1}^{k} x_{q-1,\delta,\theta,a} \quad \forall k \in \mathcal{K}, q \in \mathcal{Q}'', \theta \in \Theta, a \in \mathcal{A}. \tag{2.37}$$

Similarly, the right side would look as follows:

$$j_{q,\theta,a} + j_{q-1,\theta,a} - \sum_{\delta=1}^{|\mathcal{K}|-k} x_{q,\delta,\theta,a} \leq 1 + \sum_{\delta=|\mathcal{K}|-k+1}^{|\mathcal{K}|} x_{q-1,\delta,\theta,a} \quad \forall k \in \mathcal{K}, q \in \mathcal{Q}'', \theta \in \Theta, a \in \mathcal{A}. \tag{2.38}$$

The logic here is that any active row must have one active beamlet from its active neighbouring row, before it reaches the total number of active beamlets in that row. Because these constraints are enforced from both the left and right sides, and we have already restricted against gaps in the rows with the no-island constraints, there must be at least one shared beamlet between each set of neighbouring rows for the constraints to be satisfied, as desired. The additional j terms are the edge cases, since where one or both of the rows are off, the constraint must be relaxed.

2.4.6 Full Robust Direct Aperture Model

Putting it all together from this section, we have the following complete model.

Note, for the **M-DAO** model, just sub out the **M-RFMO** model constraints in the **M-RDAO** model, for the following:

(**M-DAO**)

$$\sum_{b\in\mathcal{B}}\sum_{i\in\mathcal{I}}\sum_{a\in\mathcal{A}} w_{b,a}p_i D_{b,v,i} \geq L_v \qquad \forall v \in \mathcal{V}_T.$$

(**M-RDAO**)

$$\min \sum_{s\in\{T,H\}} \frac{c_s}{|\mathcal{V}_s|} \sum_{v\in\mathcal{V}_s} \sum_{b\in\mathcal{B}} \sum_{i\in\mathcal{I}} \sum_{a\in\mathcal{A}} p_i D_{v,b,i} w_{b,a}$$

$$\text{s.t.} \sum_{i\in\mathcal{I}} \Big[\underline{p}_i y_{0,v} - (\underline{p}_i + \bar{p}_i) y_{i,v} + (p_i - \underline{p}_i) \sum_{b\in\mathcal{B}} \sum_{a\in\mathcal{A}} D_{b,v,i} w_{b,a} \Big] \geq L_v \qquad \forall v \in \mathcal{V}_T,$$

$$\sum_{b\in\mathcal{B}} \sum_{a\in\mathcal{A}} D_{b,v,i} w_{b,a} - y_{0,v} + y_{i,v} \geq 0 \qquad \forall i \in \mathcal{I}, \forall v \in \mathcal{V}_T,$$

$$w_{b,a} \leq M x_{b,a} \qquad \forall b \in \mathcal{B}, a \in \mathcal{A},$$

$$w_{b,a} \leq f_a + M(1 - x_{b,a}) \qquad \forall b \in \mathcal{B}, a \in \mathcal{A},$$

$$w_{b,a} \geq f_a - M(1 - x_{b,a}) \qquad \forall b \in \mathcal{B}, a \in \mathcal{A},$$

$$\sum_{b\in\mathcal{B}_\theta} x_{b,a} \leq |\mathcal{B}_\theta| u_{a,\theta} \qquad \forall a \in \mathcal{A}, \theta \in \Theta,$$

$$\sum_{\theta\in\Theta} u_{a,\theta} = 1 \qquad \forall a \in \mathcal{A},$$

$$l_{q,k+1,\theta,a} \geq l_{q,k,\theta,a} \qquad \forall k \in \mathcal{K}', q \in \mathcal{Q}, \theta \in \Theta, a \in \mathcal{A},$$

$$r_{q,k,\theta,a} \geq r_{q,k+1,\theta,a} \qquad \forall k \in \mathcal{K}', q \in \mathcal{Q}, \theta \in \Theta, a \in \mathcal{A},$$

$$x_{q,k,\theta,a} = -1 + l_{q,k,\theta,a} + r_{q,k,\theta,a} \qquad \forall k \in \mathcal{K}, q \in \mathcal{Q}, \theta \in \Theta, a \in \mathcal{A},$$

$$j_{q,\theta,a} = -1 + \bar{j}_{q,\theta,a} + \underline{j}_{q,\theta,a} \qquad \forall q \in \mathcal{Q}, \theta \in \Theta, a \in \mathcal{A},$$

$$j_{q,\theta,a} \leq \sum_{k\in\mathcal{K}} x_{q,k,\theta,a} \qquad \forall q \in \mathcal{Q}, \theta \in \Theta, a \in \mathcal{A},$$

$$|\mathcal{K}| \times j_{q,a,s} \geq \sum_{k\in\mathcal{K}} x_{q,k,\theta,a} \qquad \forall q \in \mathcal{Q}, \theta \in \Theta, a \in \mathcal{A},$$

$$\bar{j}_{q,\theta,a} \leq \bar{j}_{q+1,\theta,a} \qquad \forall q \in \mathcal{Q}', a \in \mathcal{A}, s \in S,$$

$$\underline{j}_{q+1,\theta,a} \leq \underline{j}_{q,\theta,a} \qquad \forall q \in \mathcal{Q}', a \in \mathcal{A}, s \in S,$$

$$j_{q,\theta,a} + j_{q-1,\theta,a} - \sum_{\delta=k+1}^{|\mathcal{K}|} x_{q,\delta,\theta,a} \leq 1 + \sum_{\delta=1}^{k} x_{q-1,\delta,\theta,a} \qquad \forall k \in \mathcal{K}, q \in \mathcal{Q}'', \theta \in \Theta, a \in \mathcal{A},$$

$$j_{q,\theta,a} + j_{q-1,\theta,a} - \sum_{\delta=1}^{|\mathcal{K}|-k} x_{q,\delta,\theta,a} \leq 1 + \sum_{\delta=|\mathcal{K}|-k+1}^{|\mathcal{K}|} x_{q-1,\delta,\theta,a} \qquad \forall k \in \mathcal{K}, q \in \mathcal{Q}'', a \in \mathcal{A}, s \in S,$$

$$w_{b,a} \geq 0 \qquad \forall b \in \mathcal{B}, a \in \mathcal{A},$$

$$f_a \geq 0 \qquad \forall a \in \mathcal{A},$$

$$y_{i,v} \geq 0 \qquad \forall i \in \mathcal{I}, \forall v \in \mathcal{V}_T,$$

$$y_{0,v} \text{ URS} \qquad \forall v \in \mathcal{V}_T,$$

$$x_{b,a} \in \{0,1\} \qquad \forall b \in \mathcal{B}, a \in \mathcal{A},$$

$$u_{a,\theta} \in \{0,1\} \qquad \forall a \in \mathcal{A}, \theta \in \Theta,$$

$$l_{q,k,\theta,a}, r_{q,k,\theta,a} \in \{0,1\} \qquad \forall k \in \mathcal{K}, q \in \mathcal{Q}, \theta \in \Theta, a \in \mathcal{A},$$

$$j_{q,\theta,a}, \bar{j}_{q,\theta,a}, \underline{j}_{q,\theta,a} \in \{0,1\} \qquad \forall q \in \mathcal{Q}, \theta \in \Theta, a \in \mathcal{A}.$$

Chapter 3

Efficiency Improvement Techniques

The **M-RDAO** and **M-DAO** models are extremely difficult-to-solve mixed integer programming (MIP) models. There are, however, approaches that simplify the model without reducing any of the plan quality. These approaches include symmetry elimination (Section 3.1), sampling (Section 3.2) and warm-starting techniques (Section 3.3).

3.1 Angle Symmetry Elimination

As explained in Section 2.4.3, the **M-RDAO** and **M-DAO** models take in the total number of apertures as input, but allow for freedom in terms of selecting the number of apertures allocated to each angle. This freedom leads to a larger decision space than another popular alternative, where the total number of apertures per angle is the input. In this section, methods for removing some of this redundancy are introduced, allowing our models to run more efficiently, despite the lack of preallocation.

In Section 3.1.1, we describe the symmetry that exists within our models and how it differs from the angle pre-allocated models. We then propose three methods for symmetry elimination and discuss the pros and cons of each. First we describe the naive intensity ordering method (Section 3.1.2), then the general angle-based ordering is introduced (Section 3.1.3) and finally, a hybrid increase-decreasing ordering approach, intended specifically for two-beam-angle problems, is outlined (Section 3.1.4). The three methods are compared in Section 3.1.5

3.1.1 Angle Symmetry

In the proposed models, if there are $|\mathcal{A}|$ apertures, there exists $|\mathcal{A}|!$ potential permutations of each plan. Angles are selected based on the binary variable $u_{a,\theta}$, which is defined by:

$$u_{a,\theta} = \begin{cases} 1 & \text{if angle } \theta \text{ is active in aperture } a, \\ 0 & \text{if angle } \theta \text{ is inactive in aperture } a. \end{cases}$$

The sum of $u_{a,\theta}$ is then restricted such that only one angle can be active per aperture. Once the value of $u_{a,\theta}$ have been chosen, the model behaves the same as the preallocated models in terms of order uncertainty.

In contrast, for a preallocated model, if the number of apertures per angle, n_θ, is selected, then there are $\prod_{\theta \in \Theta} n_\theta!$ permutations, i.e., there are $n_\theta!$ ways to organize the apertures within each angle.

To illustrate this behaviour with an example, in a case with 2 angles and 6 apertures, in the proposed model, each plan could be arranged $6! = 720$ different ways, between the aperture allocation and ordering flexibilities. If 3 apertures were pre-allocated per angle, each plan could only be arranged $3! \times 3! = 36$ different ways, which is clearly a much smaller decision space.

3.1.2 Naive Intensity Ordering

One simple way to reduce symmetry is to ensure that the total dose in each successive aperture is decreasing monotonically, as follows:

$$\sum_{b \in \mathcal{B}} w_{b,a} \geq \sum_{b \in \mathcal{B}} w_{b,a+1} \quad \forall a \in 1, \ldots, |\mathcal{A}| - 1.$$

These constraints order apertures as defined by each aperture's location in the set of apertures, $\mathcal{A}$. These constraints are agnostic to angle order and placement as it deals in absolute total intensities (i.e., the angle that comes up first will be that with the total highest intensity). As a result, the decision space is reduced, but there is a lack of information regarding angle choice meaning all permutations must be searched.

When this same constraint is applied within each angle of the preallocated constraint, however, it removes the symmetry, successfully eliminating all redundancies. Since it does not work that well for the proposed model, further symmetry reduction strategies are investigated.

3.1.3 General Angle-Based Ordering

Another method of forcing a sequence is to reduce the permutations of allowable activations. In these constraints, angles can only grow increasingly large with time, based on the location of the angle. Assuming angles have been labeled in some form of increasing order, the resulting plan will be output in increasing order of angle, numerically.

This method effectively takes the problem size of the proposed model down to the size of the (unreduced) preallocated model. It cannot, however, remove ordering redundancy, leaving the problem with the $\prod_{\theta\in\Theta}[\sum_{a\in\mathcal{A}} u_{a,\theta}]!$ possible arrangements of a single solution (the starts and ends of each allocation are not known, so the constraints in Section 3.1.2 cannot be applied to remove redundancy, as in the preallocated case). The ordering method is formulated as follows:

$$\sum_{\theta\in\Theta} \theta \times u_{a+1,\theta} \geq \sum_{\theta\in\Theta} \theta \times u_{a,\theta} \quad \forall a \in 1,\ldots,|\mathcal{A}|-1.$$

To illustrate how this works, the an example with five apertures ($|\mathcal{A}| = 5$) and three angles ($|\Theta| = 3$) is used.

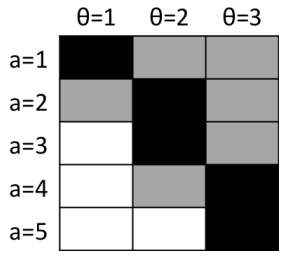


Figure 3.1: Allowable realizations of a u variable, given the label assignment ordering constraint. Binaries set to 1 are shaded in black, while allowable selections are grey.

Figure 3.1 shows an iterative use of these constraints. Starting with aperture 1, the u value for $a = 1$, can set any of the three angles equal to 1. If arbitrarily, $u_{11} = 1$ is selected, the set of constraints for $a = 2$ look as follows:

$$1 \times u_{2,1} + 2 \times u_{2,2} + 3 \times u_{2,3} \geq 1 \times u_{1,1} + 2 \times u_{1,2} + 3 \times u_{1,3}.$$

Subbing in the first aperture:

$$1 \times u_{2,1} + 2 \times u_{2,2} + 3 \times u_{2,3} \geq 1 \times (1) + 2 \times (0) + 3 \times (0).$$

Simplified:

$$1 \times u_{2,1} + 2 \times u_{2,2} + 3 \times u_{2,3} \geq 1.$$

Once again, any angle can be chosen for aperture two, so if 2 is chosen arbitrarily, the next aperture now gets restricted, as follows:

$$1 \times u_{3,1} + 2 \times u_{3,2} + 3 \times u_{3,3} \geq 1 \times (0) + 2 \times (1) + 3 \times (0).$$

This simplifies to:

$$1 \times u_{3,1} + 2 \times u_{3,2} + 3 \times u_{3,3} \geq 2.$$

Since only one angle may be chosen per aperture, it is clear from this constraint that angle 1 is no longer an option for aperture $a = 3$, since each angle must exceed its proceeding aperture. The constraints continue to propagate in this fashion.

3.1.4 Hybrid Increase-Decrease Ordering

One final method for sorting angles works by removing permutations in intensity and angle, simultaneously, meaning there are no permutations possible. This method has the benefits of both of the above methods in one, the caveat being that it does not scale as more angles are added. This does, however, make it very well suited to the application at hand, as tangential breast cancer radiation therapy only requires the two angles.

The constraints may be formulated as follows:

$$\sum_{b \in \mathcal{B}_1} w_{b,a} \geq \sum_{b \in \mathcal{B}_1} w_{b,a+1} \qquad \forall a \in 1, \ldots, |\mathcal{A}| - 1, \tag{3.1}$$

$$\sum_{b \in \mathcal{B}_2} w_{b,a} \leq \sum_{b \in \mathcal{B}_2} w_{b,a+1} \qquad \forall a \in 1, \ldots, |\mathcal{A}| - 1. \tag{3.2}$$

Constraints (3.1) specify that intensities in angle 1 ($\mathcal{B}_1$) must be ordered from greatest to least, while constraints (3.2) specify that angle 2 ($\mathcal{B}_2$) must be ordered from least to greatest. This eliminates all possible permutations, without capping the number of apertures per direction, since only one angle can be active at any given time, and any number of apertures generated can be sorted in angle 1-2 order, by increase then decrease, respectively. This effect is demonstrated visually in Figure 3.2.

Unfortunately, when it comes to scaling this method, it is not obvious how to add a third or higher number of angles, seeing as even if we know the desired angle order (e.g.,

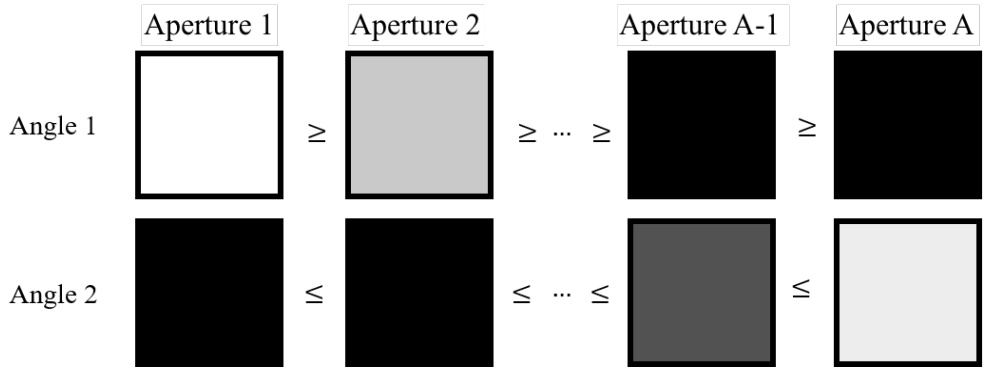


Figure 3.2: Visualization of angle/intensity sorting. Lighter colours represent higher beam intensities.

1-3-2), we do not know where each angle starts and do not want to restrict its intensity value, or starting point. This means leaving the flexibility for an angle to be set to 0, then (potentially) any intensity greater than 0, then 0 again, which does not seem to be feasible using sets of inequality constraints.

3.1.5 Ordering Method Summary

The above methods can each be useful in different scenarios, depending on the treatment setup and features desired. These properties include how many angles the constraint can handle, whether or not it restricts intensity-based permutations and angle-based permutations and whether or not it still holds in the linear relaxation of the MIP problem. We summarize the capabilities of each method in Table 3.1, below. Note that for the ordering constraints that use total intensity, maximum intensity should be equally valid, having no major impact on performance.

	Symmetry Breaking Method		
Properties	Label Assignment	Total Intensity	Increase-Decrease
Max Angles	Unlimited	Unlimited	2
Sorts Intensities	Yes	No	Yes
Sorts Angles	No	Yes	Yes
Can be relaxed	Yes	No	Yes

Table 3.1: Comparison of proposed symmetry reduction methods

Since only two angles are required, using the increase-decrease method makes the most sense as an addition to the **M-RDAO** model. Knowing that it still holds in a linear relaxation is useful as well, however, adding more constraints to a linear model is likely to slow down solution time, so it is less beneficial in a fluence map optimization (FMO) setting than a direct aperture optimization (DAO) problem, where it more effectively cuts down the decision space.

3.2 Sampling Techniques

Clinical datasets contain detailed computed tomography (CT) images of the patient's body. While these high-resolution scans are important for properly delineating organs they tend to be prohibitively large for planning purposes. Not only do the datasets overwhelm planning software at the highest resolutions, but large discrepancies between beamlet grid-resolution and voxel size also lead to multiple neighbouring voxels having roughly the same dose-influence values, introducing a fair amount of redundancy into the problem.

One way to reduce the problem size is by considering larger voxels, however the CT scan resolution cannot be changed in post processing. An approximated method which achieves nearly the same effect as granularity reduction is achieved by only considering every n^{th} voxel. More advanced sampling techniques such as varying aggression based on organ, or region within an organ, or clustering voxels based on similarity as implemented using k-means clustering in Mahnam et al. (2017), are also available, however, for the purpose of this study, the downsampling methods were kept simple. The convention observed in this work was an initial sampling at the granularity level set by clinics followed by more aggressive downsampling (i.e., increasing the value of n) if the problem was still too large to converge. The clinical sampling has previously been shown to yield roughly the same distribution as considering every voxel at the granularity extracted from the CT scans in Chan et al. (2014).

3.3 Warm Start Algorithm

Since the DAO methods are large MIP problems, they are very time consuming to solve. In addition to being difficult, even finding a feasible solution can be non-trivial and often at quite a large gap from optimality. For this reason, it is desirable to find good approximations for this problem, either as input to the larger model, or as a stand-alone result if the output happens to be clinically acceptable.

Generating fairly high-quality, feasible solutions can help, not only with the optimization process, but also to give context for pre-optimization decisions, such as putting a bound on the required number of apertures to achieve reasonably high quality, and even early tradeoff assessment between plan quality and critical organ dose. In order to achieve these objectives, a partially heuristic, partially optimization-based method for finding good, feasible, direct aperture plans is developed. The method requires three steps, each of which makes up a subsection of this section:

1. A version of the **M-FMO** or **M-RFMO** model, with efficiency improvements is run to get an initial z^*_{FMO} value (Section 3.3.1).
2. A modified **M-FMO** or **M-RFMO** model called the **M-WS** model is run to get a set of beamlet intensities, w^{minmax} (Section 3.3.2).
3. A gap-filling heuristic is run, to yield a feasible solution w^{ws} and objective function value z^{ws} (Section 3.3.3).

After the algorithm is introduced, some limitations of the formulation are discussed in Section 3.3.4.

3.3.1 Warm Start Step 1: Running the (R)FMO Model

The warm start algorithm starts by running the basic **M-FMO** or **M-RFMO** model, corresponding to the ultimate type of DAO plan desired by the user (i.e., **M-DAO** or **M-RDAO**). This can be achieved by using the models introduced in Sections 2.2 and 2.3, respectively. While these models are continuous and therefore run fairly quickly, they are being run as inputs for increasingly larger models. For this reason, it is desirable to ensure that each individual model converges as quickly as possible. In this section, a constraint generation method is proposed for speeding up each of the optimizations. It should be noted that the two constraint generation techniques are not intended to be implemented simultaneously. The first method, described in A) is tailored to the **M-FMO** model, while the second, in B) is specific to the **M-RFMO** model.

A) Constraint Generation for FMO

One of the major factors in the difficulty of planning intensity modulated radiation therapy (IMRT) plans is the size of the decision space. For the **M-FMO** model, a method of

addressing this issue, that differs from sampling, is the iterative addition of voxels. Rather than starting with the complete set of (optionally downsampled) voxels, the problem starts with a subset and optimizes over said subset, only bringing in additional constraints with large violations, with the goal of incrementally improving solutions.

We propose initiating the constraint generation with a master problem that only includes a constraint restricting the average dosage to the clinical target volume (CTV).

The master problem looks as follows:

$$\begin{aligned} \min \sum_{s\in\{T,H\}} \frac{c_s}{|\mathcal{V}_s|} \sum_{v\in\mathcal{V}_s} \sum_{b\in\mathcal{B}} \sum_{i\in\mathcal{I}} p_i D_{v,b,i} w_b \\ \text{s.t.} \sum_{i\in\mathcal{I}} \sum_{v\in\mathcal{V}_T} \sum_{b\in\mathcal{B}} \sum_{a\in\mathcal{A}} p_i D_{b,v,i} w_{b,a} \geq \sum_{v\in\mathcal{V}_T} L_v, \\ w_b \geq 0 \quad \forall b \in \mathcal{B}. \end{aligned}$$

At every iteration, one or a set number of voxel constraints may be added:

$$\sum_{i\in\mathcal{I}} \sum_{b\in\mathcal{B}} \sum_{a\in\mathcal{A}} p_i D_{b,v,i} w_{b,s} \geq L_v \quad \forall v \in \{\text{worst-case violated voxel from a previous iteration}\}.$$

The worst violation can be found in a simple subproblem, where the current plan's dose to each voxel is compared to that of the prescribed dose. When all voxels meet or exceed the prescribed dose, the problem is at optimality.

B) Constraint Generation for RFMO

The robust problem tends to take several hours for full-sized clinical problem and can take even longer with an objective change, as will be discussed in the upcoming section. For this reason, we propose an efficient constraint generation approach to solving the problem, which is based on Mahmoudzadeh et al. (2016). We propose a methodology that solves the same problem as the **M-RFMO** model, using the **M-FMO** framework along with a master-subproblem approach. In its essence, it satisfies the specification made in constraints (2.8), that force every realization of $\tilde{\mathbf{p}} \in \mathcal{P}$ to be accounted, using an iterative, rather than a transformative approach. The problem starts with a small subset of $\tilde{\mathbf{p}}$ realizations, and more constraints with different $\tilde{\mathbf{p}}$ values are only added as necessary.

In practice, this means that the master problem runs the optimization over a fixed subset of $\mathcal{P}^{sub} \in \mathcal{P}$, and at each iteration, we get a new, larger $\mathcal{P}^{sub}$, until, in the worst

case, all $|\mathcal{I}|!$ corner-point realizations of $\mathcal{P}$ are in the master problem, or in practice, none of the remaining realizations of $\tilde{\mathbf{p}}$ lead to a CTV dose violation.

The master problem is built on the **M-FMO** model, but with nominal CTV constraints (2.2) updated to represent robust corner points,

$$\begin{aligned} \min \sum_{s\in\{T,H\}} \frac{c_s}{|\mathcal{V}_s|} \sum_{v\in\mathcal{V}_s} \sum_{b\in\mathcal{B}} \sum_{i\in\mathcal{I}} p_i D_{v,b,i} w_b & \\ \text{s.t.} \sum_{b\in\mathcal{B}} \sum_{i\in\mathcal{I}} \tilde{p}_i D_{b,v,i} w_b \geq L_v \quad \forall v \in \mathcal{V}_T,\ \forall \tilde{\mathbf{p}} \in \mathcal{P}^{sub}, & \\ w_b \geq 0 \qquad\qquad \forall b \in \mathcal{B}. & \end{aligned} \tag{3.3}$$

Mahmoudzadeh et al. (2016) proves that the maximum violation, $\tilde{\mathbf{p}}$ calculation is separable for each voxel, and can be done using a subproblem (which is also linear). For this work, we propose an alternative, deterministic sorting approach as the subproblem, which finds the same worst-case $\tilde{\mathbf{p}}$ value, and maximum violation as Mahmoudzadeh et al. (2016), but without requiring a linear program (LP) in the subproblem.

We achieve this by algorithmically calculating a worst-case $\tilde{\mathbf{p}}$ at a given, potentially optimal w_b, at each voxel, v. This is doable, since each worst-case scenario is dependent on the sequencing of phases, not on the w_b values themselves. Conceptually, this method relies on the idea that the worst breathing realization that can occur is the one that puts the maximum allowable proportion of time in the phase that has the least payoff for CTV dose, given a specified set of beam intensities.

In order to find this sequencing, first total dose to CTV voxel per phase, $d_{i,v}$, is calculated,

$$d_{i,v} = \sum_{b\in\mathcal{B}} D_{b,v,i} w_b \quad \forall v \in \mathcal{V}_T. \tag{3.4}$$

Then, for each v, $d_{i,v}$ is ordered from least to greatest, in a variable which we will capture as $o_{i,v}$. The worst-case realization of $\tilde{\mathbf{p}}$ for each voxel is then determined by distributing the deviation based on ordering $o_{i,v}$, and ensuring that it sums to 0. Since the worst-case $\tilde{\mathbf{p}}$ is desired, the most upwards deviation is given to the phases with the least impact on the CTV for each voxel, and the least is given to the phases with the largest impact. The set of $|\mathcal{I}|!$ realizations of $\tilde{\mathbf{p}}$ sequence-dependent worst-case vectors turns out to be the entire set of corner points of the subproblem.

Similar to Mahmoudzadeh et al. (2016), we start the algorithm with a single set of $\mathcal{P}^{sub}$ values. However, for implementation purposes, we chose a $\tilde{\mathbf{p}}$ value for each voxel that corresponds to the worst-case $d_{i,v}$, if all values of w_b are equal to 1. This is unlikely

to reflect the real worst-case, but unlike using a nominal $\mathbf{p}$, which is guaranteed to not be an extreme point, these will account for potential worst-case sequencings. In further iterations, the worst-case constraints are added for all constraints that lead to a violation that is greater than 0.

The pseudocode for the worst-case calculation subproblem is as follows:

Algorithm 1 Worst-Case $\tilde{p}$ Generation Algorithm

▷ Inputs: dose-influence matrix (current voxel), beam fluences, prescribed dose
▷ (current voxel), nominal proportions, allowable upper and lower
▷ deviations from nominal proportions
▷ Outputs: worst case proportions ($\tilde{p}$), violation
▷ ———————————————————————————————————————
procedure WorstP_PerVox($D, w, L, p, \bar{p}, \underline{p}$)
    ▷ ———————— Sum and sort the dose to voxel per phase ————————
    $I \leftarrow |p|$ ▷ Number of breathing phases
    $B \leftarrow |w|$ ▷ Number of beamlets
    for $i \leftarrow 1$ to I **do** ▷ Calculate the dose delivered at each phase
        $d[i] \leftarrow 0$
        for $b \leftarrow 1$ to B **do**
            $d[i] += D[b][i] \times w[b]$
    $o[i] = \mathbf{sort}(d[i]$ **in** $i)$ ▷ Sort the breathing phases by ascending size of $d[i]$
    ▷ ———————— Prepare to allocate the worst-case uncertainty ————————
    room4excess $\leftarrow 0$
    dose2voxel $\leftarrow 0$
    for $i \leftarrow 1$ to I **do** ▷ Set up the allowable delta bounds
        room4excess += $\underline{p}[i]$
    for $i \leftarrow 1$ to I **do** ▷ Maximize proportion of time in lowest payoff phases
        $j \leftarrow o[i]$
        room4excess -= $\underline{p}[j]$ ▷ Largest upward delta bounded by lower delta
        pDelta = $\mathbf{min}\{\bar{p}[j]$, room4excess$\}$
        room4excess -= pDelta ▷ Ensures that everything sums to 0 in the end
        $\tilde{p}[j] \leftarrow p[j]$ + pDelta ▷ Final value is the change added to the nominal
        dose2voxel += $d[j] \times \tilde{p}[j]$
    violation $\leftarrow \mathbf{max}\{L-$ dose2voxel, $0\}$
    return $\tilde{p}$, violation

3.3.2 Warm Start Step 2: An Optimization Formulation

The novel optimization component of the warm start algorithm was inspired by the similarity between the segment uniformity constraints and a min-max optimization problem. It operates in a similar method to the FMO problem, except with a min-max objective. We introduce the warm start with the constraints from the **M-FMO** model, but it is easily extendable to the **M-RFMO** model as well. The constraint generation algorithms still apply to their corresponding warm start algorithms.

The warm start makes use of the DAO adjusted variable and a max intensity variable:

$w_{b,a}$ represents the intensity of beamlet b in aperture a.

$m_{\theta,a}$ takes on the max intensity value at angle θ in aperture a.

The model looks as follows:

(**M-WS**)

$$\min \sum_{\theta\in\Theta}\sum_{a\in\mathcal{A}} m_{\theta,a} + e_1 \sum_{b\in\mathcal{B}}\sum_{a\in\mathcal{A}} w_{a,b}$$

$$\text{s.t.} \sum_{s\in\{T,H\}} \frac{c_s}{|\mathcal{V}_s|} \sum_{v\in\mathcal{V}_s}\sum_{b\in\mathcal{B}}\sum_{i\in\mathcal{I}}\sum_{a\in\mathcal{A}} p_i D_{v,b,i} w_{b,a} \leq e_2 \left[z^*_{FMO}\right], \tag{3.5}$$

$$\sum_{b\in\mathcal{B}}\sum_{i\in\mathcal{I}}\sum_{a\in\mathcal{A}} p_i D_{b,v,i} w_{b,a} \geq L_v \qquad \forall v \in \mathcal{V}_T, \tag{3.6}$$

$$m_{\theta,a} \geq w_{b,a} \qquad \forall b \in \mathcal{B}_\theta, a \in \mathcal{A}, \theta \in \Theta, \tag{3.7}$$

$$w_{b,a} \geq 0 \qquad \forall b \in \mathcal{B}, a \in \mathcal{A},$$

where:

e_1 and e_2 are weighting parameters.

$[z^*_{FMO}]$ is the objective value of a previously run **M-FMO** model.

Constraints (3.7) set force variable $m_{\theta,a}$ to take on the maximum fluence per angle θ, per aperture a variable, which then gets minimized in the first term of the objective function. The secondary term, is the total beam intensity, which may optionally be penalized by setting $e_1 \geq 0$. This may help with conformity, but needs to be tuned based on the problem at hand.

Finally, the $[z^*_{FMO}]$ value in (3.5) is a product of a previous optimization that gets fed into the warm start optimization at hand. The weighting factor e_2 is also a tuneable parameter, but it must be greater than 1, to ensure the feasibility of the problem.

The robust version of the formulation above is identical, except with robust CTV dosage constraints, instead of the nominal ones in (3.6).

3.3.3 Warm Start Step 3: Heuristic Gap Filling

The heuristic for filling in the beam is fairly simple. It sets all w^{minmax} values to the same value and fills in any rows to obey island constraints. The pseudocode for the algorithm is as follows:

Algorithm 2 Warm Start Gap Filling Algorithm

▷ Inputs: beam fluences from warm-start algorithm part 1, row indices, column
▷ indices, angle indices, aperture indices
▷ Outputs: DAO compliant beamlet intensities (w^{ws})
▷ ---
procedure FILL_FLUENCE($w^{minmax}, \mathcal{Q}, \mathcal{K}, \Theta, \mathcal{A}$)
    ▷ First, fill in the rows of the generated beam intensities:
    for $a \leftarrow \mathcal{A}$ **do** ▷ For each aperture
        for $\theta \leftarrow \Theta$ **do** ▷ And each angle
            wMax $\leftarrow$ **max**$\{w^{minmax}[a][\theta]\}$ ▷ Find the largest intensity at this aperture
            for $q \leftarrow \mathcal{Q}$ **do** ▷ Go through each row
                ind $\leftarrow$ index$\{w^{minmax}[a][\theta][q] > 0\}$
                for $k \leftarrow$ **min**{ind} to **max**{ind} **do**
                    $w^{minmax}[a][\theta][q][k] \leftarrow$ wMax ▷ Fill in row
                    if Vertical **and** hasGap **do**
                        Fill in w^{minmax} of top of gap until 1 unit overlap with bottom
                        Fill in w^{minmax} vertical line at left-most aligned unit
    ▷ Next, sort w for symmetry-breaking constraint compatibility:
    ord1 = **sort**$_a\{w^{minmax}[a][\theta = 1][q]$ by **sum**$\{w^{minmax}[a][\theta = 1]\}$, descending$\}$ ▷ Sort
    ord2 = **sort**$_a\{w^{minmax}[a][\theta = 2][q]$ by **sum**$\{w^{minmax}[a][\theta = 2]\}$, ascending$\}$
    for i = 1 to |ord1| **do**
        w^{ws}[i][$\theta = 1$] = w^{minmax}[ord[i]][$\theta = 1$]
        w^{ws}[i+|ord1|][$\theta = 2$] = w^{minmax}[ord2[i]][$\theta = 2$]
    return w^{ws}

If a vertical fill in of the algorithm is desired Vertical is set to true. The algorithm naively prevents vertical gaps by filling in a vertical column between the two separate partitions, always choosing the left-most if there are multiple options. If the two segments don't align, a straight line of beamlets is turned on in the upper part of the segment until there is a single-beamlet of overlap.

There are more thorough ways of performing this vertical integration, but they require sub enumerations of different options, which starts bringing overhead into the model.

The objective function of the total warm start algorithm, z^{ws}, is calculated based on the output of this final heuristic, w^{ws}. The optimality gap of this warm-start algorithm can then be calculated by:

$$100 * (1 - z^*_{FMO}/z^{ws}).$$

3.3.4 Limitations of the Warm Start Method

While the warm start provides a good jumping-off point for future exploration there are some limitations, that are not present in the original **M-DAO** model.

For one, the angle flexibility is no longer available, since in order to define the min-max variable, $m_{\theta,a}$, a set number of angles must be made available per aperture. It is possible that the optimization chooses less than the total allowable angles, i.e., sets some angles to zero, and in that case, fewer angles may be needed in the final optimization (particularly if there is a low optimality gap). As a rule of thumb, we chose to $1/|\Theta|$ the number of warm apertures as desired in the final DAO optimization when running the warm start, since up to $|\Theta| \times |\mathcal{A}|$ apertures are output by the model, rather than the desired $|\mathcal{A}|$ apertures.

Another caveat of this warm start is the requirement for tuned parameters. Since there is a heuristic component of filling in gaps in the apertures, it is unclear how much we want to value uniformity over optimality, and at which point we will see diminishing returns in the optimality bound. At some point, the **M-WS** model tends towards the FMO solution, which will likely be much less uniform than competing options, and give a worse lower bound.

Chapter 4

Results

This chapter discusses the implementation and results of applying the proposed model to patient datasets. The computational infrastructure used is discussed in Section 4.1. Section 4.2 shows the results of the initial application of limited-feature versions of our models to a one-dimensional phantom dataset as a proof of concept. Next, a clinical patient dataset was used to run the fully-developed model and to explore the effects of the improved techniques for this large-scale application. The results of this exploration are presented in Section 4.3. Finally, computational results are reported in Section 4.4.

4.1 Computational Infrastructure

The initial modelling was done with a mix of MATLAB R2016b to generate dummy datasets and run the code, and AMPL and Gurobi were used to run the optimization. As soon as this proof of concept was working appropriately, the optimization portion was recoded to be run using a C++ and CPLEX combination and used to validate results, for the ease of future customization. All results reported in the following sections were obtained using the C++ and CPLEX 12.7.1 combination.

The one-dimensional and linear fluence map optimization (FMO) were run locally on a 2.6 GHz Intel Core i5 computer with 8GB of memory. Since direct aperture optimization (DAO) problems are complex mixed integer programming (MIP) problems which had to be run for many days, they were run on a single node of the Centre for Advanced Computing cluster at Queens University, with 2.2 GHz with 24 cores and 100 GB of memory.

4.2 One-Dimensional Proof of Concept Study

We tested our preliminary **M-RDAO** model using a one-dimensional phantom consisting of 151 voxels. Of the voxels, the first 50 make up the heart and the next 51 are occupied by the clinical target volume (CTV). The remaining voxels represent air external to the body. The voxels are each set to be 0.2 cm wide. The beam is also chosen to be a 1D array, with its 56 beamlets spanning several voxels each, at 0.5 cm. The use of the phantom allowed us to perform a simplified version of the study where movement is strictly lateral, the beam is a single beamlet in depth and only one row of voxels is used, in place of a full set of human organs. An even smaller-scale schematic of this toy setup is depicted in Figure 4.1. As the patient inhales, the lung forces the CTV and heart apart.

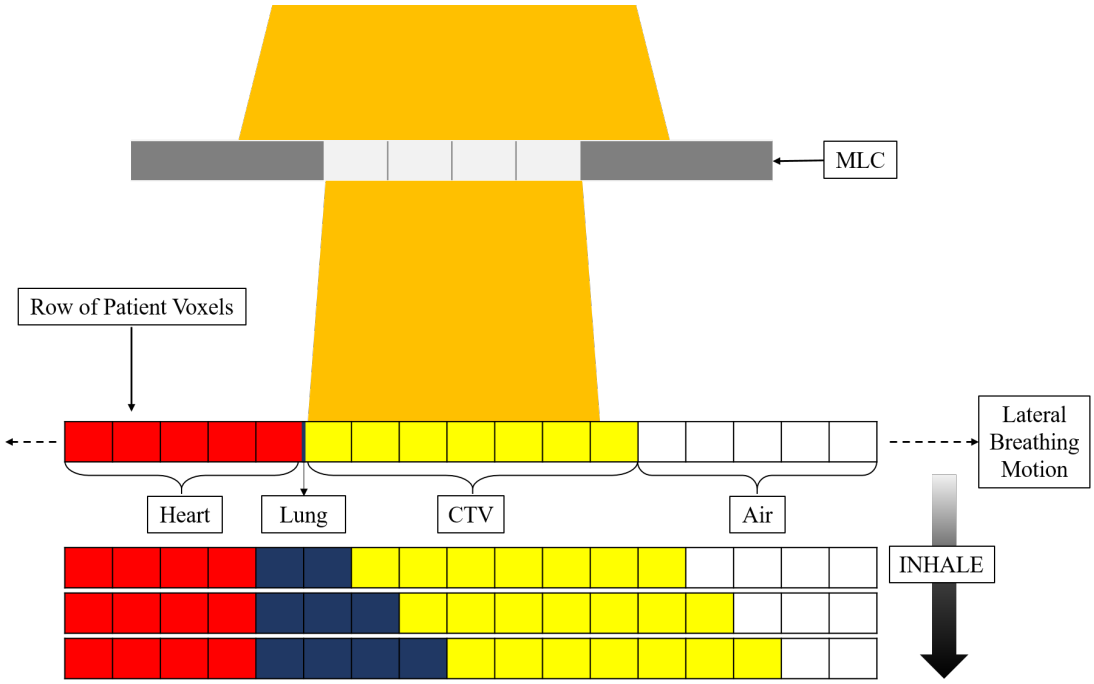


Figure 4.1: A simplified representation of the 1D phantom left-sided breast cancer case. The fully exhaled phase is shown directly underneath the beam, while subsequent inhale phases are shown underneath.

The case was run with $|\mathcal{I}| = 4$ phases. In each phase, the CTV moves forward a single beamlet-width, while the heart only moves one unit backwards in the second phase and then stays put. Since this phantom was only used to test the capabilities of the code, the

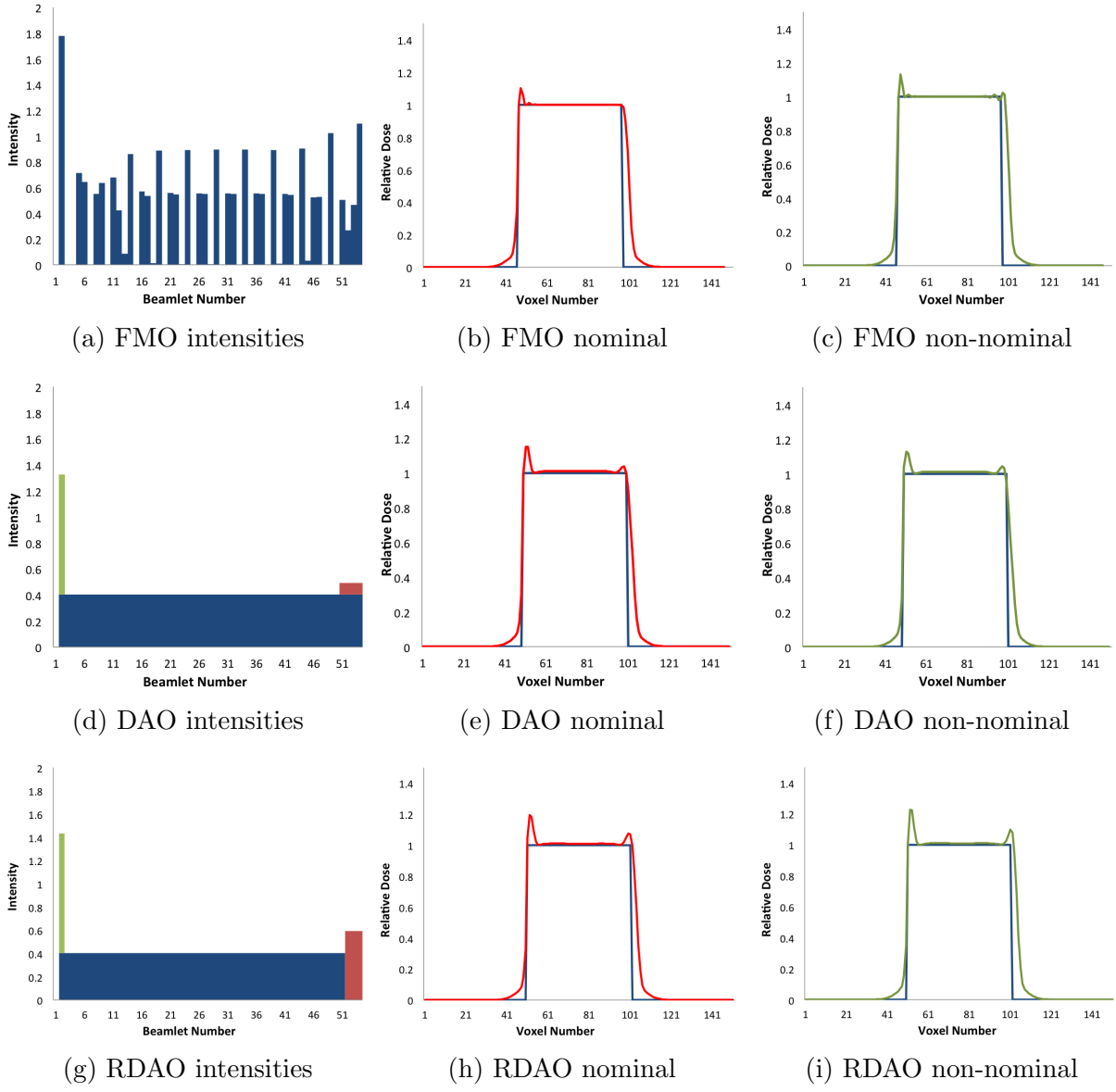
motion was kept simple by design. The nominal four-phase $\mathbf{p}$ vector, from exhale to inhale, was set to $\mathbf{p} = [0.5, 0.1667, 0.1667, 0.1667]$. The $\underline{\mathbf{p}}, \bar{\mathbf{p}}$ were set to ± 0.1 for each element.

The results of running the **M-FMO**, **M-DAO** and **M-RDAO** models on the above case are shown in Figure 4.2. The **M-RFMO** model was run as well, to address dosing problems in the nominal model, and is shown in Appendix C.1. Three apertures were used to generate the **M-DAO** and **M-RDAO** model cases in 4.2. Theses models were each also applied to the phantom with one and two apertures as well, the results of which are depicted in Appendices C.2 and C.3. The desired CTV dose is normalized to 1, for consistency in comparison across models.

The beamlet intensities, shown in Figure 4.2a are not uniform or deliverable, as is to be expected from a FMO. In contrast, the beam intensities in 4.2d and 4.2g are each deliverable in three distinct, uniform segments.

All three models perform very well in terms of nominal dose, with Figures 4.2b, 4.2e and 4.2h not dipping below the blue CTV dose reference line at all. In contrast, in the non-nominal realization of $\tilde{\mathbf{p}}$ results in a very non-uniform delivery of the plan from the **M-FMO** model towards the edges of the CTV, shown in Figure 4.2c. This fluctuation is significant enough that a fair amount of the CTV gets underdosed. To a lesser, but still significant effect, the **M-DAO** model also underdoses the CTV in sections towards the edge, as shown in Figure 4.2f. These issues are rectified in Figure 4.2i, which stays above the required dose, even under the non-nominal uncertainty realization.

We note that a primary difference between the robust and nominal plans is that robust plans provide a higher dosage to the uncertain edges of the CTV, while the regular **M-FMO** model tries to drive down this dosage, to conform as closely as possible to the nominal breathing pattern.



(a) FMO intensities (b) FMO nominal (c) FMO non-nominal

(d) DAO intensities (e) DAO nominal (f) DAO non-nominal

(g) RDAO intensities (h) RDAO nominal (i) RDAO non-nominal

Figure 4.2: The results of running the models on the 1D phantom. Figure (a),(d) and (g) show the chosen beamlet intensities. In the models with apertures, the intensities allocated to each aperture are indicated by colour. Figures (b),(e) and (h) show the prescribed dose in blue, along with the nominal realized dose in red. Figures (c),(f) and (i) show the prescribed dose in blue, along with the non-nominal realized breathing pattern in green.

4.3 Clinical Left-Sided Breast Cancer Case Study

In this section, the application of the models to a clinical patient dataset is discussed. In Section 4.3.1, a summary of the clinical data set is given. Next the plan evaluation tools introduced in Section 1.4 are applied to the large-scale (1x the clinical sampling rate) plans, with Section 4.3.2, showing the fluence map (FM) diagrams and Section 4.3.3, presenting and discussing the corresponding dose volume histogram (DVH) diagrams.

4.3.1 Clinical Data

The patient data comes from a 4D-CT dataset, which was provided by the Princess Margaret Cancer Centre, Toronto, Canada. The data considered contained $|\mathcal{I}| = 5$ breathing phases sorted from inhale to exhale. We define $\mathbf{p} = [0.5, 0.125, 0.125, 0.125, 0.125]$ and uncertainty set is defined using $\underline{\mathbf{p}} = \bar{\mathbf{p}} = \mathbf{0.1}$, i.e., the range around each value is ± 0.1.

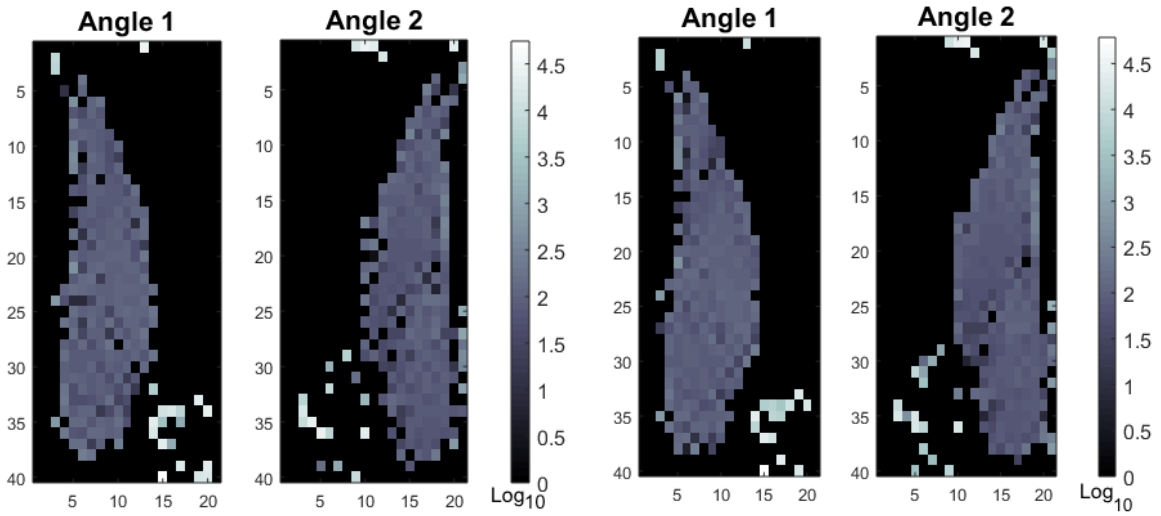
The treatment region consists of 33,592 CTV voxels and 42,342 heart voxels, (only 2,878 of which are exposed to the beam throughout treatment). The patient was prescribed a dose of 42.4 Gy for the full CTV. The voxel dimensions in the 4DCT scan are $1 \times 1 \times 2$ mm^3, but clinical voxels are typically $4 \times 4 \times 4$ mm^3, so we start sampling at a rate of $4 \times 4 \times 2 = 32$ voxels for every 1 CT voxel. This method was shown to work effectively in Chan et al. (2014), without an overall loss in plan quality.

The two tangential angles were pre-selected in clinic, and the dose-influence matrix for each angle, was pre-calculated using planning software and used as model input data for all models. The beamlet grid has a resolution of 0.5×0.5 cm^2 at each of these two angles. The beamlets are oriented on a 40×19 grid at each angle, for a total of $|\mathcal{B}|$ =1520 beamlets.

4.3.2 Fluence Map Visualizations

We began by running the continuous **M-FMO** and **M-RFMO** models on the large clinical dataset, yielding the the FM diagrams in Figure 4.3. Note, these plots show the log of the dose, in order to depict the whole treatment region - since the intensities vary greatly.

The plans look nearly identical in the images, however, their differing impacts will become clear when examining their DVHs in the next section. Both plans are non-uniform but can be seen to mostly cover the entire CTV region, with the exception of a few high-intensity outliers.

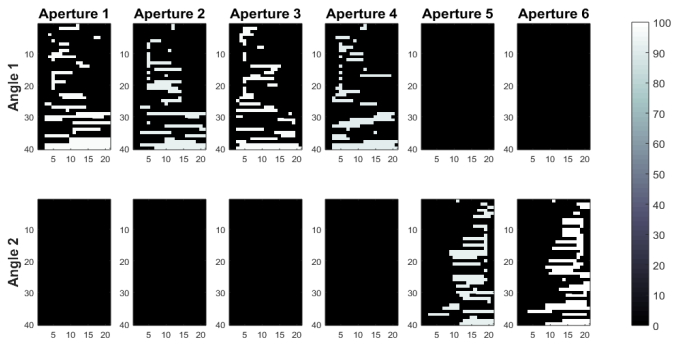


(a) FM of the FMO (b) FM of the RFMO

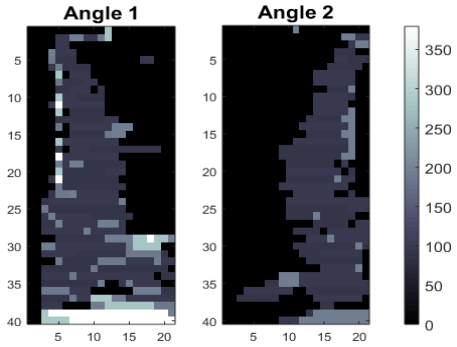
Figure 4.3: Depiction of the log-scale fluence map of the FMO and RFMO plan beamlet intensities, respectively. The darkest beamlets represent $b = 0$, while the lighter beamlets represent higher intensities.

The FM of the treatment of the **M-DAO** model is shown in Figure 4.4. In Figure 4.4a, each of the apertures and the selected angles are shown. In Figure 4.4b, the consolidation of these apertures into a single fluence map is depicted. The **M-DAO** was run without any symmetry breaking constraints or aperture continuity constraints, however, a small M bound in the uniformity constraint allowed it the model yield a gap of 7.46%.

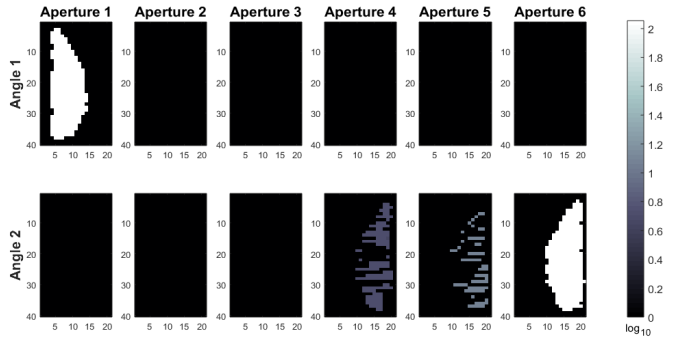
The **M-RDAO** model similarly resulted in a set of apertures and an aggregated fluence map. This problem was run with a non-restrictive M value, symmetry breaking constraints, and then both with and without aperture continuity constraints. Logarithmic base 10 plots were used to generate Figures 4.4c - 4.4f, since there was a large discrepancy between aperture intensities in the robust output, which makes it difficult to visualize all the apertures. In Figure 4.4c, the model chose not to fill the second and third aperture over its optimization period. A similar result is found in Figure 4.4e. These solver choices are likely related to the rather large optimality gaps still present in the current solutions (23.34% and 22.15%, respectively). The differing allocation effects would likely be reduced if the solver were given more time to find a plan that is closer to optimality.



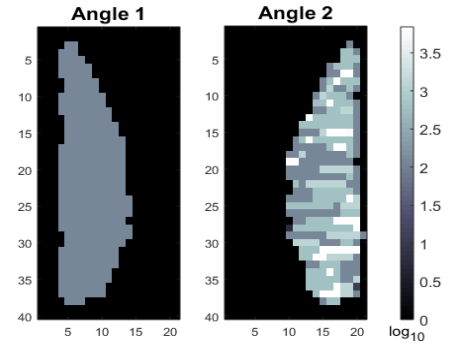
(a) FM of the DAO (no continuity constraints)



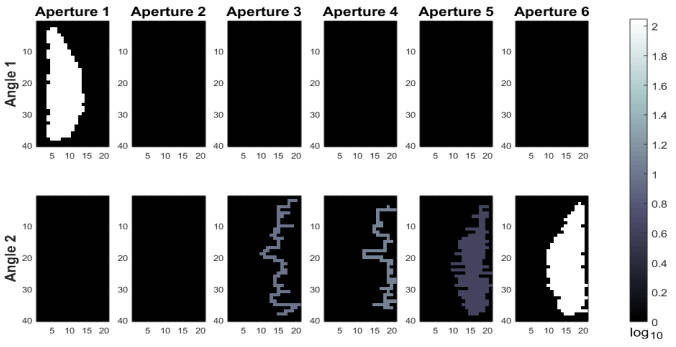
(b) Combined FM of the DAO



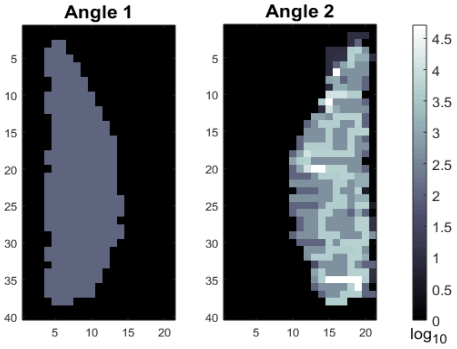
(c) FM of the RDAO (no continuity constraints)



(d) Combined FM of the RDAO



(e) FM of the RDAO (with continuity constraints)



(f) Combined FM of the RDAO

Figure 4.4: FM of the **M-DAO** and **M-RDAO** models applied to the clinical problem. They are each combined into a single fluence map, for reference. Note: the robust models were plotted on a log base 10 scale, to enhance visualization.

4.3.3 DVH Results

In Figure 4.5, we compare the DVHs of the **M-FMO** and **M-RFMO** models when a non-nominal breathing motion, or corner point of the uncertainty set is applied. We note that while the DVH of the **M-FMO** model loses quality towards the edge of the DVH, the **M-RFMO** model continues to deliver a high-quality dose throughout, and the full 100% by volume region of the CTV recieves 100% of the dose.

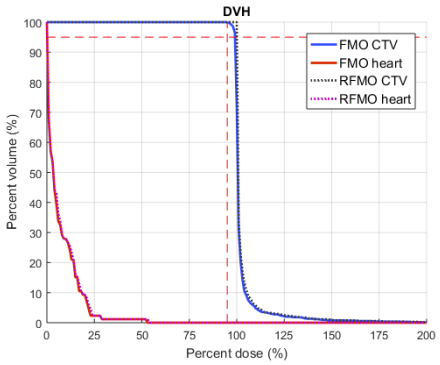


Figure 4.5: DVH of the **M-FMO** v.s. **M-RFMO** model with non-nominal breathing.

While it is not a totally fair comparison, as our best **M-RDAO** model was much further from optimality than the **M-DAO** model, we may also compare the DVH across the two models. When the nominal breathing pattern is realized, as in Figure 4.6 , it is primarily just evident that the CTV is receiving excessive dose in the **M-RDAO** model, due to the optimality gap, with no immediately obvious gain.

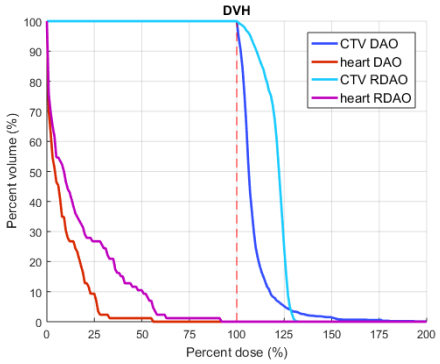


Figure 4.6: DVH of the **M-DAO** v.s. **M-RDAO** model with nominal breathing.

The benefits of the **M-RDAO** model are seen, however, when non-nominal breathing patterns are realized. Figure 4.7 shows that the CTV will be underdosed when plans are generated using the **M-DAO** model at a non-nominal realization, while the **M-RDAO** model consistently ensures consistent coverage.

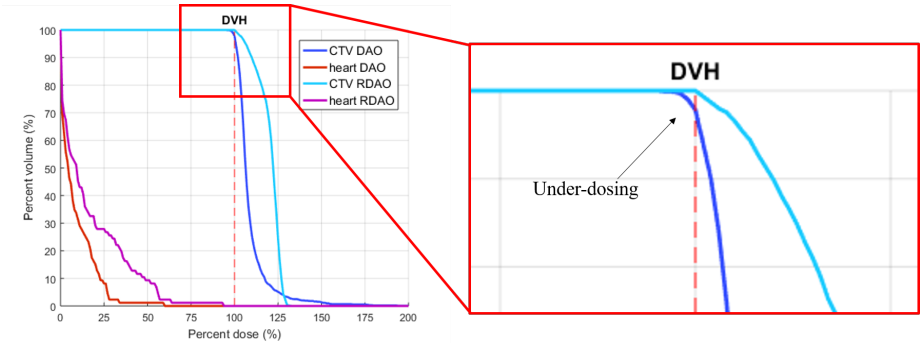


Figure 4.7: DVH of the **M-DAO** v.s. **M-RDAO** model with non-nominal breathing.

4.4 Computational Results

All clinical data was run with objective function weighting values, c_s, set to $c_T = 0.4$ and $c_H = 0.6$. These values were empirically found to produce a good balance between our objectives of conformity and heart sparing. The following section demonstrates the results in terms of the improvement methods outlined in Chapter 3, including symmetry breaking 4.4.1, sampling 4.4.2 and the warm-start method 4.4.3.

4.4.1 Angle Symmetry Elimination

Angle symmetry was tested out empirically in the models. It was found to perform poorly in the FMO models, increasing run-time for the algorithm. It was however very beneficial for the DAO based models. Due to time constraints and empirical advantage, symmetry breaking constraints were used for all tests in which the uniformity constraint's M value was not heavily restricted (this is a more manual strategy for reducing the decision space).

4.4.2 Sampling Techniques

We began by running the continuous **M-FMO** and **M-RFMO** models both quickly and locally with our given objective function. To obtain more near-optimal solutions, we also sampled more aggressively, increasingly the sample rate 3 times by a factor of 4; a choice that does result in a loss of treatment quality, but allows us to better demonstrate our methodology. The resultant values are of interest as they provide the lower-bound fluence map solutions for their mixed integer DAO counterparts. The results of these runs are summarized in Table 4.1.

Model	**Sample Rate**	PreProc.	Time	z^*	Avg T	Avg H	Max b
FMO	$\times 1$	0.11	24.96	19.0765	43.42	2.847	5.549e+04
FMO	$\times 4$	0.03	0.41	17.872	42.56	1.413	1.541e+05
FMO	$\times 16$	0.00	0.04	16.9693	42.4	0.01554	2.669e+05
FMO	$\times 64$	0.00	0.01	16.96	42.4	0	1.235e+07
RFMO	$\times 1$	0.51	537.33	19.2029	43.63	2.917	6.098e+04
RFMO	$\times 4$	0.10	36.54	18.0295	42.8	1.513	1.415e+05
RFMO	$\times 16$	0.03	0.57	17.0959	42.56	0.1208	2.09e+05
RFMO	$\times 64$	0.01	0.08	16.96	42.4	0	4.876e+05

Table 4.1: Continuous **M-FMO** and **M-RFMO** model results run on clinical datasets. Note: times are reported in seconds. The preprocessing time (PreProc) is separated from the optimization run-time (Time). Objective function value (z^*), average dose to target (Avg T), heart (Avg H), and max planned beamlet intensity (Max b) are reported for each plan.

As expected, the **M-RFMO** takes longer to run than the **M-FMO** model, and under nominal assumptions, it delivers more dose to the heart and the CTV. The most aggressive sampling method ($\times 64$) results in a perfect objective function, with 0 Gy going to the heart and 42.4 Gy going to all included CTV voxels. This means that 16.96 is the lowest possible objective value given our selected objective weightings. It also means that $\times 64$ sampling is a bit aggressive for this dataset. In reality, it is only hitting a small subset of the voxels, as are all the sampled datasets to a degree, an effect that can be seen in the DVH in Figure 4.8. For that reason, the higher sampling rates were used, not to suggest clinically acceptable plans, but to demonstrate and compare model performance.

The maximum beamlet intensity is also reported in Table 4.1, to demonstrate that if left unchecked in an **M-FMO** model, it does grow quite large, which would increase treatment time. An entire treatment, including motion should be between 15-30 minutes.

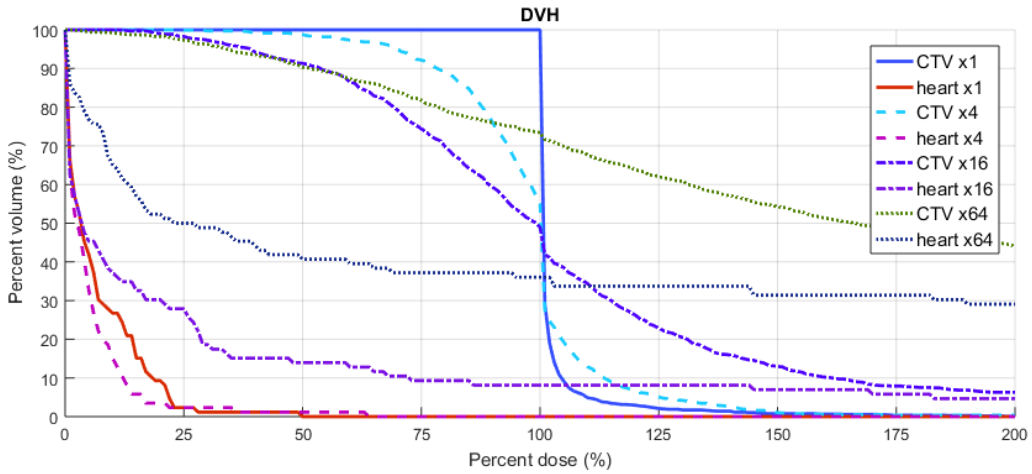


Figure 4.8: Depiction of the quality loss in the nominal FMO model, assuming a nominal distribution, with more aggressive sampling. The figure is plotted on the clinically sampled grid, so the clinically sampled dataset performs nearly perfectly. The remaining down-sampled plans get worse as the sampling becomes more aggressive.

To give some context, the dose delivered by most machines is 300 to a maximum of 2400 MU/minute (Kalantzis et al., 2012). Monitor units (MU) are calibrated to 1 cGy, which is .01 Gy, so we can deliver an upper limit of 24 Gy/minute. We need to give the patient 42.4 Gy of treatment over a set number of fractions. Treatments are divided into fraction of 1.5 Gy at the low end (and up to 3 Gy at the higher end), so it should be safe to say that this plan should require a maximum of 30 fractions (Baskar et al., 2012). In order to deliver the lowest intensity in the above table, which is 55,490 Gy, we would require 55,490/30 = 1849.66 Gy to be delivered per fraction. At a maximum of 24 Gy/minute, this would require roughly 77 minutes of beam on time per session. This is an hour and 17 minutes, in addition to setup time and beam reorientation. If we use a more realistic dose-rate estimate of 600 MU/minute, each treatment would require the beam on time to be longer than 5 hours. This is clearly unacceptable, and provides further incentive to generate more uniform plans, rather than trying to decompose a fluence map as discussed in section 1.1.2.

In an effort to strike a balance between up-keeping plan quality and aperture flexibility, we chose to run our test cases with six apertures (total) for the **M-DAO** and **M-RDAO** models. According to Jiang et al. (2005), allocating less than 3 apertures per angle has been shown to negatively impact plan quality, whereas 3-7 apertures per angle was shown to be ideal for avoiding diminishing returns. We also did not want to over-allocate angles,

however, as choosing more, lower dose angles has been shown to increase the impact of miss-calibration and leakages by Sudahar et al. (2012). Six apertures was also found to be the median number of apertures required in the clinical breast cancer study presented in Vicini et al. (2002).

Running the large-scale **M-DAO** and **M-RDAO** models proved to be extremely time consuming, taking 2 weeks run-time only to remain beyond a 50% optimality gap. In some cases the problem also ran out of memory. Table 4.2 shows direct aperture results along with the closest optimality gap we have for them within a two-week run time limit.

Model	**Sample**	Total time	$z^*_{(R)FMO}$	z_{Best}	Max b	BOT	Gap(%)
DAO	×1	2 Wks	19.0765	20.6143	100	573.1	7.46
DAO	×4	2 Wks	17.872	19.1858	100	589.9	6.85
DAO	×16	Memory	16.9693	17.8005	100	598.9	4.67
DAO	×64	28766.6	16.96	16.9794	100	595.9	0.11
RDAO	×1	2 Wks	19.2029	24.6657	111.39	245.7	22.15
RDAO	×4	2 Wks	18.0295	21.8619	117.16	223.4	17.53
RDAO	×16	2 Wks	17.0959	19.0686	60.46	250.8	10.34
RDAO	×64	2 Wks	16.96	17.1472	95.34	235.6406	1.09

Table 4.2: Six segment **M-DAO** and **M-RDAO** model results run on clinical datasets. Note: times reported in seconds, unless indicated otherwise. Here, z_{Best} is the objective we achieved that is closest to the optimal before the algorithm timed out or finished. BOT stands for beam-on-time, a clinical metric. Gap is the gap between z_{Best} and $z^*_{(R)FMO}$, a lower bound on the best possible objective function value for the MIP problem.

4.4.3 Warm Start Algorithm

Since the models can easily run for 2 weeks each, sometimes without making any noticeable progress, particularly for the more complex **M-RDAO** models, we turned to our proposed improvement methods. It is possible to get fairly good problem approximations in very little time with the warm start method, described in detail in Section 3.3.

To motivate the need for a warm start, Table 4.3 shows the exact same **M-RDAO** model set-up being run on a warm started problem, versus without the start. In two weeks time, the two larger problems stayed beyond a 50% optimality gap. This is noteworthy, seeing as we know based on our warm-start methods that there are far better plans available.

Model	**Sample**	Time	$z^*_{(R)FMO}$	z_{Warm}	z	$\text{Gap}_{Warm}(\%)$	$\text{Gap}(\%)$
RDAO	$\times 1$	2 Wks	19.2029	**24.6657**	43.8719	**22.15**	56.23
RDAO	$\times 4$	2 Wks	18.0295	**21.8619**	44.0599	**17.53**	59.08
RDAO	$\times 16$	2 Wks	17.0959	**19.0686**	19.8256	**10.34**	13.77
RDAO	$\times 64$	2 Wks	16.96	**17.1472**	17.1783	**1.09**	1.27

Table 4.3: Non-warm start 2 week objective values (z), v.s. the warm-started ones (z_{Warm}) reported above for the **M-RDAO** model reported in Table 4.2. Better solutions are bolded.

A complete set of warm start solutions are shown in Table 4.1. Note, the gap-filling heuristic component of the warm start completes in fractions of a second, and the initial optimization for the lower bound is shown in Table 4.1, so the only times reported in this table are for the warm start min-max phase of the optimization. It should also be noted that the warm start reported objectives in Table 4.4 is the calculated objective of the **M-DAO** and **M-RDAO** models with the warm start + heuristic $w_{b,a}$ values. This value may be further adjusted downwards by CPLEX, since the heuristic does not reduce the CTV dose, even if the heuristically-added new beamlets now lead to overdosing the CTV. CPLEX will use these values to find a basis, then adjust them until we are at a corner point in the solution space.

A sample output of the warm start algorithm is depicted in Appendix D. These outputs show a low-fidelity version of the FM and DVH plots which are used to assess and build on the given plan.

The continuous model ran quickly enough that the non-robust constraint generation proved to be unnecessary. Since MIP problems do not respond well to add-hock row generation algorithms, it was not actually used after implementation. In the robust case, however, constraint generation proved very useful for running both **M-RFMO** and the robust warm start model for the larger size models.

The impact of the constraint generation of the regular **M-RFMO** model is shown in Table 4.5. As expected, for smaller problem sizes, the constraint generation does not provide much benefit, as you have to solve essentially the same problem several times. At larger sizes, however, the constraint generation starts really paying off. The effect of the column generation start was larger in the warm start models, to the extent that none of the warm starts were run with the regular robust without column generation as it took on the order of 10 times longer to run.

Model	**Sample**	AC?	Time	$z^*_{(R)FMO}$	z^{ws}	z_{DAOadj}	Max b	Gap(%)	Iter
wDAO	×1	No	48.77	19.0765	27	-	134.5	29.36	-
wDAO	×4	No	9.02	17.872	22.1	21.7696	40.14	19.14	-
wDAO	×16	No	1.65	16.9693	20.54	20.3110	45.08	17.36	-
wDAO	×64	No	0.39	16.96	17.71	17.1475	67.73	4.221	-
wDAO	×1	Yes	177.17	19.0765	26.19	25.2713	47.11	27.16	-
wDAO	×4	Yes	9.02	17.872	22.1	21.7696	40.14	19.14	-
wDAO	×16	Yes	1.65	16.9693	20.54	20.3110	45.08	17.36	-
wDAO	×64	Yes	0.38	16.96	17.85	17.5028	67.73	5.001	-
wRDAO	×1	No	94.51	19.2029	27.44	-	135.7	30.02	5
wRDAO	×4	No	22.96	18.0295	22.64	-	96.37	20.37	6
wRDAO	×16	No	2.8	17.0959	21.32	-	123.4	19.82	6
wRDAO	×64	No	0.48	16.96	17.86	17.7350	90.66	4.834	3
wRDAO	×1	Yes	699.96	19.2029	27.37	25.1077	129.7	42.55	9
wRDAO	×4	Yes	226.02	18.0295	23.41	117.16	120.5	22.98	5
wRDAO	×16	Yes	64.08	17.0959	20.83	20.4170	64.02	17.93	5
wRDAO	×64	Yes	6.6	16.96	17.91	17.6558	82.07	5.59	3

Table 4.4: Warm start plans run on clinical datasets. Note: times are reported in seconds. The prescribed dose is 42.4 Gray. $e_1 = 0.005$, $e_2 = 1.1$. In cases where the MIP model was run following the warm start, the adjusted objective function value (z_{DAOadj}) is listed. Here, AC stands for aperture continuity, which indicates whether or not the continuous aperture constraints have been enforced.

Model	**Sample**	PreProc.	Time (s)	Obj	Iter	Avg T	Avg H	Max b
RcgFMO	×1	0.10	**85.96**	19.2029	8	43.63	2.917	6.097e+04
RcgFMO	×4	0.02	**13.95**	18.0295	9	42.8	1.513	1.418e+05
RcgFMO	×16	0.01	0.93	17.0959	12	42.56	0.1208	2.09e+05
RcgFMO	×64	0.01	0.11	16.96	8	42.4	4.2e-05	6.114e+05
RFMO	×1	0.51	537.33	19.2029	-	43.63	2.917	6.098e+04
RFMO	×4	0.10	36.54	18.0295	-	42.8	1.513	1.415e+05
RFMO	×16	0.03	**0.57**	17.0959	-	42.56	0.1208	2.09e+05
RFMO	×64	0.01	**0.08**	16.96	-	42.4	0	4.876e+05

Table 4.5: The **M-RFMO** model was run using the general robust versus the constraint generation algorithm. The quicker run times of the two are bolded. Iter stands for the number of iterations or optimizations run for the constraint generation problems.

Chapter 5

Conclusions

Providing optimized treatment plans for breast cancer patients is vital for obtaining high post-treatment quality of life. In this thesis, practical solutions are proposed for real-world inhibitors of integrating advanced mathematical modelling into modern intensity modulated radiation therapy (IMRT) devices. These inhibitors include target region motion uncertainty, as well as mechanical delivery constraints.

The proposed methodology centres around combining previously proposed robust fluence map optimization (RFMO) frameworks, which can be used to account for breathing motion during left-sided breast cancer treatment, with novel direct aperture optimization (DAO) techniques, which incorporate deliverability requirements, into a single, holistic mixed integer programming (MIP) model. This unified robust direct aperture optimization (RDAO) model provides a way to immunize against uncertainty without worrying about downstream losses of quality, which arise in the conventional approaches that leave deliverability to post-processing heuristics.

This thesis outlines contributions to three aspects of the RDAO problem. The first contribution is in providing the unifying framework which aligns the goals of robust and DAO models, as discussed above. The second contribution is with regards to providing fast, high-quality heuristic estimations for this difficult-to-solve MIP problem. Then, as a final contribution, methods for simplifying the decision space are provided and tested against the initial model.

The proposed RDAO framework primarily hinges on the partitioning of RFMO decision variables into multiple, separately-deliverable beam apertures. After this substitution has been made, the two models can be integrated into one. Proposed constraints for enforcing DAO requirements, including uniformity constraints, angle selection, island removal and

optionally, vertical beam continuity are added on top of existing robust constraints. This combination of constraints forms a large MIP problem, which is solvable on a small-scale 2D phantom model, but would often run for weeks without proper convergence on larger clinical datasets.

The slow convergence time of the large-scale model, motivated the secondary contribution that is a heuristic approach to estimating a DAO solution. This mini-max framework was inspired by the uniformity constraints in the DAO model, since mini-max constraints have a tendency to flatten the decision space, in a manner that is analogous to the constrained uniformity. This framework provides a rapid, DAO-like solution, but must be post-processed heuristically to provide a true DAO or RDAO plan. This heuristic makes the problem highly nonlinear and difficult to tune.

To get more from the heuristic solution, it was incorporated into the third contribution, which is decision space simplification. Rather than evaluating the heuristic approach on its own, it was applied as a warm start to the existing DAO and RDAO models. The warm start cut down the decision space, often saving days or weeks of time. Furthermore, symmetry breaking constraints were introduced for improving MIP convergence time and constraint generation methods were introduced for speeding up the continuous parts of the problem.

Future opportunities for research lie in applying decompositions techniques to this large-scale model. Column generation approaches have been popular in the literature; however, the binary structure of the subproblem does lend itself to other methods of separation, which require fewer additional variables, such as logic based benders approaches.

Some lower hanging fruit in terms of future research might come from examining the impact of the weighting variable M, within the uniformity constraint, as the idea of incorporating a "very large number" into the optimization, naturally leads to the question "how large is large?", and thus far, that question remains unanswered. More effective sampling methodologies such as k-mean clustering or adjusting the aggressiveness of sampling based on organ location could be another simple way to increase algorithm effectiveness in the future. Parallelizing the code is another option for speeding up the MIP results, although many clinics do not have the infrastructure to implement that sort of approach, making it less practical.

Finally, analyzing the effectiveness of the heuristic approach on larger-scale problems such as volumetric modulated radiation therapy (VMAT) could be an interesting avenue for future research. The beam-filling algorithm could be improved with some fairly straight forward tweaks, with minor time penalties, such as iteratively evaluating connection options before selecting one, and the development of stricter weighting calibration rules.

References

[1] R. K. Ahuja and H. W. Hamacher. A network flow algorithm to minimize beam-on time for unconstrained multileaf collimator problems in cancer radiation therapy. *Networks: An International Journal*, 45(1):36–41, 2005.

[2] E. Ahunbay and X. A. Li. Investigation of the reliability, accuracy, and efficiency of gated IMRT delivery with a commercial linear accelerator. *Medical Physics*, 34(7): 2928–2938, 2007.

[3] E. E. Ahunbay, G.-P. Chen, S. Thatcher, P. A. Jursinic, J. White, K. Albano, and X. A. Li. Direct aperture optimization–based intensity-modulated radiotherapy for whole breast irradiation. *International Journal of Radiation Oncology, Biology, Physics*, 67 (4):1248–1258, 2007.

[4] American Cancer Society. External beam radiation therapy. http://www.princetonneurologicalsurgery.com/radiosurgery-institute/treatment-tools/elekta-synergy-s/, 2018.

[5] D. Baatar, H. W. Hamacher, M. Ehrgott, and G. J. Woeginger. Decomposition of integer matrices and multileaf collimator sequencing. *Discrete Applied Mathematics*, 152(1-3):6–34, 2005.

[6] R. Baskar, K. A. Lee, R. Yeo, and K.-W. Yeoh. Cancer and radiation therapy: current advances and future directions. *International Journal of Medical Sciences*, 9(3):193, 2012.

[7] J. Bernier, E. J. Hall, and A. Giaccia. Radiation oncology: a century of achievements. *Nature Reviews Cancer*, 4(9):737, 2004.

[8] D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.

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

[10] T. Bortfeld. IMRT: a review and preview. *Physics in Medicine & Biology*, 51(13): R363, 2006.

[11] T. Bortfeld and W. Schlegel. Optimization of beam orientations in radiation therapy: some theoretical considerations. *Physics in Medicine & Biology*, 38(2):291, 1993.

[12] T. Bortfeld, T. C. Chan, A. Trofimov, and J. N. Tsitsiklis. Robust management of motion uncertainty in intensity-modulated radiation therapy. *Operations Research*, 56 (6):1461–1473, 2008.

[13] M. Broderick, M. Leech, and M. Coffey. Direct aperture optimization as a means of reducing the complexity of intensity modulated radiation therapy plans. *Radiation Oncology*, 4(1):8, 2009.

[14] Canadian Cancer Statistics Advisory Committee. *Canadian Cancer Statistics*. Canadian Cancer Society, 2018.

[15] T. C. Chan, T. Bortfeld, and J. N. Tsitsiklis. A robust approach to IMRT optimization. *Physics in Medicine & Biology*, 51(10):2567, 2006.

[16] T. C. Chan, H. Mahmoudzadeh, and T. G. Purdie. A robust-CVaR optimization approach with application to breast cancer therapy. *European Journal of Operational Research*, 238(3):876–885, 2014.

[17] M. Chu, Y. Zinchenko, S. G. Henderson, and M. B. Sharpe. Robust optimization for intensity modulated radiation therapy treatment planning under uncertainty. *Physics in Medicine & Biology*, 50(23):5463, 2005.

[18] L. Conroy, S. Quirk, and W. L. Smith. Realistic respiratory motion margins for external beam partial breast irradiation. *Medical Physics*, 42(9):5404–5409, 2015.

[19] S. C. Darby, D. J. Cutter, M. Boerma, L. S. Constine, L. F. Fajardo, K. Kodama, K. Mabuchi, L. B. Marks, F. A. Mettler, L. J. Pierce, et al. Radiation-related heart disease: current knowledge and future prospects. *International Journal of Radiation Oncology, Biology, Physics*, 76(3):656, 2010.

[20] A. Gladwish, M. Oliver, J. Craig, J. Chen, G. Bauman, B. Fisher, and E. Wong. Segmentation and leaf sequencing for intensity modulated arc therapy. *Medical Physics*, 34(5):1779–1788, 2007.

[21] M. J. Hooning, A. Botma, B. M. Aleman, M. H. Baaijens, H. Bartelink, J. G. Klijn, C. W. Taylor, and F. E. Van Leeuwen. Long-term risk of cardiovascular disease in 10-year survivors of breast cancer. *Journal of the National Cancer Institute*, 99(5): 365–375, 2007.

[22] International Agency for Research on Cancer et al. GLOBOCAN 2012: Estimated cancer incidence, mortality and prevalence worldwide in 2012, 2012.

[23] Z. Jiang, M. Earl, G. Zhang, C. Yu, and D. Shepard. An examination of the number of required apertures for step-and-shoot IMRT. *Physics in Medicine & Biology*, 50 (23):5653, 2005.

[24] G. Kalantzis, J. Qian, B. Han, and G. Luxton. Fidelity of dose delivery at high dose rate of volumetric modulated arc therapy in a truebeam linac with flattening filter free beams. *Journal of Medical Physics/Association of Medical Physicists of India*, 37 (4):193, 2012.

[25] S. Kamath, S. Sahni, J. Li, J. Palta, and S. Ranka. Leaf sequencing algorithms for segmented multileaf collimation. *Physics in Medicine & Biology*, 48(3):307, 2003.

[26] P. J. Keall, G. S. Mageras, J. M. Balter, R. S. Emery, K. M. Forster, S. B. Jiang, J. M. Kapatoes, D. A. Low, M. J. Murphy, B. R. Murray, et al. The management of respiratory motion in radiation oncology report of AAPM task group 76 a. *Medical Physics*, 33(10):3874–3900, 2006.

[27] L. L. Kestin, M. B. Sharpe, R. C. Frazier, F. A. Vicini, D. Yan, R. C. Matter, A. A. Martinez, and J. W. Wong. Intensity modulation to improve dose uniformity with tangential breast radiotherapy: initial clinical experience. *International Journal of Radiation Oncology, Biology, Physics*, 48(5):1559–1568, 2000.

[28] M. Langer, V. Thai, and L. Papiez. Improved leaf sequencing reduces segments or monitor units needed to deliver IMRT using multileaf collimators. *Medical Physics*, 28(12):2450–2458, 2001.

[29] Y. Li, J. Yao, and D. Yao. Genetic algorithm based deliverable segments optimization for static intensity-modulated radiotherapy. *Physics in Medicine & Biology*, 48(20): 3353, 2003.

[30] H. Mahmoudzadeh, T. Chan, and T. Purdie. Th-a-116-05: A robust direct aperture optimization approach for left-sided breast IMRT. *Medical Physics*, 40(6):529–529,

2013. ISSN 2473-4209. doi: 10.1118/1.4815734. URL http://dx.doi.org/10.1118/1.4815734.

[31] H. Mahmoudzadeh, J. Lee, T. C. Chan, and T. G. Purdie. Robust optimization methods for cardiac sparing in tangential breast IMRT. *Medical Physics*, 42(5):2212–2222, 2015.

[32] H. Mahmoudzadeh, T. G. Purdie, and T. C. Chan. Constraint generation methods for robust optimization in radiation therapy. *Operations Research for Health Care*, 8: 85–90, 2016.

[33] M. Mahnam, 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*, 62(14):5589, 2017.

[34] C. Men, 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*, 52(24):7333, 2007.

[35] M.-P. Milette. *Direct optimization of 3D dose distributions using collimator rotation*. PhD thesis, University of British Columbia, 2008.

[36] K. D. Miller, R. L. Siegel, C. C. Lin, A. B. Mariotto, J. L. Kramer, J. H. Rowland, K. D. Stein, R. Alteri, and A. Jemal. Cancer treatment and survivorship statistics, 2016. *CA: A Cancer Journal for Clinicians*, 66(4):271–289, 2016.

[37] P. L. Nguyen and A. L. Zietman. Proton-beam vs intensity-modulated radiation therapy: which is best for treating prostate cancer? *Oncology*, 22(7):748, 2008.

[38] A. Olafsson and S. J. Wright. Efficient schemes for robust IMRT treatment planning. *Physics in Medicine & Biology*, 51(21):5621, 2006.

[39] Princeton Neurological Surgery. Elekta Synergy® S. https://www.cancer.org/treatment/treatments-and-side-effects/treatment-types/radiation/external-beam-radiation-therapy.html. Accessed: 2018-08-17.

[40] T. G. Purdie, R. E. Dinniwell, D. Letourneau, C. Hill, and M. B. Sharpe. Automated planning of tangential breast intensity-modulated radiotherapy using heuristic optimization. *International Journal of Radiation Oncology, Biology, Physics*, 81(2): 575–583, 2011.

[41] S. Quirk, L. Conroy, and W. L. Smith. When is respiratory management necessary for partial breast intensity modulated radiotherapy: A respiratory amplitude escalation treatment planning study. *Radiotherapy and Oncology*, 112(3):402–406, 2014.

[42] H. Romeijn and J. Dempsey. Intensity modulated radiation therapy treatment plan optimization. *TOP*, 16(2):215–243, 2008.

[43] H. E. Romeijn, 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*, 15(3):838–862, 2005.

[44] E. Salari and J. Unkelbach. A column-generation-based method for multi-criteria direct aperture optimization. *Physics in Medicine & Biology*, 58(3):621, 2013.

[45] E. Salari, C. Men, and H. E. Romeijn. Accounting for the tongue-and-groove effect using a robust direct aperture optimization approach. *Medical Physics*, 38(3):1266–1279, 2011.

[46] D. Shepard, M. Earl, X. Li, S. Naqvi, and C. Yu. Direct aperture optimization: A turnkey solution for step-and-shoot IMRT. *Medical Physics*, 29(6):1007–1018, 2002.

[47] R. A. C. Siochi. Minimizing static intensity modulation delivery time using an intensity solid paradigm. *International Journal of Radiation Oncology, Biology, Physics*, 43(3): 671–680, 1999.

[48] K. E. Sixel, M. C. Aznar, and Y. C. Ung. Deep inspiration breath hold to reduce irradiated heart volume in breast cancer patients. *International Journal of Radiation Oncology, Biology, Physics*, 49(1):199–204, 2001.

[49] H. Sudahar, P. Kurup, V. Murali, J. Velmurugan, et al. Dose linearity and monitor unit stability of a G4 type cyberknife robotic stereotactic radiosurgery system. *Journal of Medical Physics*, 37(1):4, 2012.

[50] A. A. S. Sultan. *Optimization of Beam Orientation in Intensity Modulated Radiation Therapy Planning*. PhD thesis, Technische Universität Kaiserslautern, 2006.

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

[52] C. Vachani. All about proton therapy. *Oncolink*, 2018.

[53] F. A. Vicini, M. Sharpe, L. Kestin, A. Martinez, C. K. Mitchell, M. F. Wallace, R. Matter, and J. Wong. Optimizing breast cancer treatment efficacy with intensity-modulated radiotherapy. *International Journal of Radiation Oncology, Biology, Physics*, 54(5):1336–1344, 2002.

[54] W. Wang, T. G. Purdie, M. Rahman, A. Marshall, F.-F. Liu, and A. Fyles. Rapid automated treatment planning process to select breast cancer patients for active breathing control to achieve cardiac dose reduction. *International Journal of Radiation Oncology, Biology, Physics*, 82(1):386–393, 2012.

[55] P. Xia and L. J. Verhey. Multileaf collimator leaf sequencing algorithm for intensity modulated beams with multiple static segments. *Medical Physics*, 25(8):1424–1434, 1998.

[56] G. Zhang, Z. Jiang, D. Shepard, B. Zhang, and C. Yu. Direct aperture optimization of breast IMRT and the dosimetric impact of respiration motion. *Physics in Medicine & Biology*, 51(20):N357, 2006.

APPENDICES

Appendix A

Step-by-Step Construction of the Robust Counterpart

In Bertsimas et al. (2011), it is shown that a polyhedral uncertainty set can be converted to a linear dual, by first breaking down the problem into a problem and subproblem, and then taking a dual of the subproblem an placing it back into the main problem.

A.1 The primal subproblem

This process starts with the following constraint, with a polyhedral uncertainty set $\mathcal{P}$:

$$\sum_{b\in\mathcal{B}}\sum_{i\in\mathcal{I}}\tilde{p}_i D_{b,v,i} w_b \geq \quad L_v \quad \forall v \in \mathcal{V}_T,\ \forall \tilde{\mathbf{p}} \in \mathcal{P}.$$

Since there are an infinite number of sets of $\tilde{\mathbf{p}}$ in $\mathcal{P}$, this formula represents an infinite number of constraints. But in reality, the only concern is that the *lowest possible* value of the left hand side is still higher than the minimum prescribed dose to each voxel, for a given intensity set w_b. That way, the left hand side of the constraint can be reformulated as a minimization problem with variable $\hat{p}_i$, as follows:

$$\min_{\{\hat{\mathbf{p}}\in\mathcal{P}\}} \sum_{i\in\mathcal{I}}\sum_{b\in\mathcal{B}}\hat{p}_i D^t_{b,v,i} w_b \geq \quad L_v \quad \forall v \in \mathcal{V}_T. \tag{A.1}$$

Since each voxel's worst case is independent, the inner optimization with variable $\hat{p}_i$ may

be inspected independently with separate constraints for each voxel, v. This does holds because the problem will still be bounded by the same worst-case $\hat{p}_i$, regardless of whether the problem is separated or kept together. The worst-case $\hat{p}_i$ is that which causes the maximum violation, so if at optimality an individual constraint could cause a larger violation than 0, that would be the current worst-case $\hat{p}_i$, and the solution would not be at optimality, contradicting the initial condition.

For the subproblem derivation, w_b can be assumed to be held constant, an assumption that can be dropped by the end of the derivation. The separated derivation is as follows, first we rewrite the equation A.1 as,

$$\begin{aligned} \min_{\hat{p}} \quad & \sum_{i\in\mathcal{I}}\sum_{b\in\mathcal{B}} \hat{p}_i D^t_{b,v,i} w_b \\ \text{s.t.} \quad & \sum_{i\in\mathcal{I}} \hat{p}_i = 1, \\ & (p_i - \underline{p}_i) \le \hat{p}_i \le (p_i + \bar{p}_i) \quad \forall i \in \mathcal{I}. \end{aligned} \tag{A.2}$$

For convenience, we may assign $M_i = \sum_{b\in\mathcal{B}} D^t_{b,v,i} w_b$:

$$\begin{aligned} \min_{\hat{p}} \quad & \sum_{i\in\mathcal{I}} \hat{p}_i M_i \\ \text{s.t.} \quad & \sum_{i\in\mathcal{I}} \hat{p}_i = 1, \\ & (p_i - \underline{p}_i) \le \hat{p}_i \le (p_i + \bar{p}_i) \quad \forall i \in \mathcal{I}. \end{aligned} \tag{A.3}$$

Expanded:

$$\begin{aligned} \min_{\hat{p}} \quad & \hat{p}_1 M_1 + \hat{p}_2 M_2 + \cdots + \hat{p}_I M_I \\ \text{s.t.} \quad & \hat{p}_1 \quad + \hat{p}_2 \quad + \cdots + \hat{p}_I \quad = 1, \\ & \hat{p}_i \le p_i + \bar{p}_i \quad \forall i \in \mathcal{I}, \\ & \hat{p}_i \ge p_i - \underline{p}_i \quad \forall i \in \mathcal{I}. \end{aligned} \tag{A.4}$$

A.2 The dual subproblem

Taking the dual of this problem yields the following:

$$
\begin{aligned}
\max \quad & y_0 + \sum_{i\in\mathcal{I}} (p_i + \bar{p}_i) y_i + \sum_{i\in\mathcal{I}} (p_i - \underline{p}_i) y_{I+1+i} & & \\
\text{s.t.} \quad & y_0 + y_i + y_{|\mathcal{I}|+i} &= M_i & \forall i \in \mathcal{I}, \\
& y_0 & \text{URS}, & \\
& y_i &\leq 0 & \forall i \in \mathcal{I}, \\
& y_{|\mathcal{I}|+i} &\geq 0 & \forall i \in \mathcal{I}.
\end{aligned}
\tag{A.5}
$$

Flipping the sign on $y \leq 0$ variables and replacing M_i back with the original constraints we get:

$$
\begin{aligned}
\max \quad & y_0 - \sum_{i\in\mathcal{I}} (p_i + \bar{p}_i) y_i + \sum_{i\in\mathcal{I}} (p_i - \underline{p}_i) y_{I+1+i} & & \\
\text{s.t.} \quad & y_0 - y_i + y_{|\mathcal{I}|+i} - \sum_{b\in\mathcal{B}} D^t_{b,v,i} w_b &= 0 & \forall i \in \mathcal{I}, \\
& y_0 & \text{URS}, & \\
& y_i, y_{|\mathcal{I}|+i} &\geq 0 & \forall i \in \mathcal{I}.
\end{aligned}
\tag{A.6}
$$

We can further simplify the equations using the equality:

$$
y_{|\mathcal{I}|+i} = \sum_{b\in\mathcal{B}} D^t_{b,v,i} w_b - y_0 + y_i \quad \forall i \in \mathcal{I}, \tag{A.7}
$$

(note: (15) use the equivalent of $y_i = y_0 + y_{|\mathcal{I}|+i} - \sum_{b\in\mathcal{B}} D^t_{b,v,i} w_b \; \forall i \in \mathcal{I}$ for their derivation, to a similar effect).

The number of variables are then reduced by subbing the equality into the lower bound constraint as follows:

$$
\sum_{b\in\mathcal{B}} D^t_{b,v,i} w_b - y_0 + y_i \geq 0 \quad \forall i \in \mathcal{I}. \tag{A.8}
$$

As well as into the objective function:

$$
y_0 - \sum_{i\in\mathcal{I}} (p_i + \bar{p}_i) y_i + \sum_{i\in\mathcal{I}} (p_i - \underline{p}_i) (\sum_{b\in\mathcal{B}} D^t_{b,v,i} w_b - y_0 + y_i). \tag{A.9}
$$

Which can be rewritten as:

$$y_0 + \sum_{i \in \mathcal{I}} -(p_i + \bar{p}_i)y_i + (p_i - \underline{p}_i)(\sum_{b \in \mathcal{B}} D^t_{b,v,i} w_b - y_0 + y_i). \tag{A.10}$$

Which can be rearranged:

$$y_0 + \sum_{i \in \mathcal{I}} -(p_i + \bar{p}_i)y_i + (p_i - \underline{p}_i)y_i + (p_i - \underline{p}_i)(\sum_{b \in \mathcal{B}} D^t_{b,v,i} w_b - y_0). \tag{A.11}$$

Then simplified:

$$y_0 + \sum_{i \in \mathcal{I}} \Big[- (\underline{p}_i + \bar{p}_i)y_i + (p_i - \underline{p}_i)(\sum_{b \in \mathcal{B}} D^t_{b,v,i} w_b - y_0) \Big]. \tag{A.12}$$

Further, factor out the extra y_0, since sum p_i is just 1 and y_0 is constant:

$$y_0 - \sum_{i \in \mathcal{I}} (p_i - \underline{p}_i)y_0 + \sum_{i \in \mathcal{I}} \Big[- (\underline{p}_i + \bar{p}_i)y_i + (p_i - \underline{p}_i) \sum_{b \in \mathcal{B}} D^t_{b,v,i} w_b \Big],$$

$$\sum_{i \in \mathcal{I}} \Big[\underline{p}_i y_0 - (\underline{p}_i + \bar{p}_i)y_i + (p_i - \underline{p}_i) \sum_{b \in \mathcal{B}} D^t_{b,v,i} w_b \Big]. \tag{A.13}$$

The final set of robust constraints which replace constraints (2.2) in the nominal model, look as follows:

$$\begin{aligned}
&\sum_{i \in \mathcal{I}} \Big[\underline{p}_i y_{0,v} - (\underline{p}_i + \bar{p}_i)y_{i,v} + (p_i - \underline{p}_i) \sum_{b \in \mathcal{B}} D_{b,v,i} w_b \Big] \geq L_v && \forall v \in \mathcal{V}_T, \\
&\sum_{b \in \mathcal{B}} D_{b,v,i} w_b - y_{0,v} + y_{i,v} \geq 0 && \forall i \in \mathcal{I}, \forall v \in \mathcal{V}_T, \\
&y_{0,v} \text{ URS} && \forall v \in \mathcal{V}_T, \\
&y_{i,v} \geq 0 && \forall i \in \mathcal{I}, \forall v \in \mathcal{V}_T.
\end{aligned}$$

Appendix B

Island Restriction Constraints Relaxation Example

In each iteration of branch and bound, a relaxed linear program (LP) version of the MIP is solved. Both sets of island constraints produce the same feasible set of x, within that relaxation, however, the version in Boland et al. (2004) leads to infeasibility when the l or r values are rounded, or branched on, making it harder for the solver to find the associated feasible solutions.

We demonstrate this problem on a set of feasible relaxations for a row of a $|\mathcal{K}| = 9$ column MLC setup.

	Proposed Island Removal									Boland et al. (2004) Island Removal										
k	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	10
L	.9	.9	.9	1	1	1	1	1	1	.9	0	0	.1	0	0	0	0	0	0	-
R	1	.9	.9	.9	.8	.7	.6	.5	.4	-	0	.1	0	0	.1	.1	.1	.1	.1	.4
x	.9	.8	.8	.9	.8	.7	.6	.5	.4	-	.9	.8	.8	.9	.8	.7	.6	.5	.4	-

If we round both solution spaces, (we assume .5 is rounded up) we get 2 very different solutions.

	Proposed Island Removal									Boland et al. (2004) Island Removal										
k	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	10
L	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	-
R	1	1	1	1	1	1	1	1	0	-	0	0	0	0	0	0	0	0	0	0
x	1	1	1	1	1	1	1	1	0	-	1	1	1	1	1	1	1	1	0	-

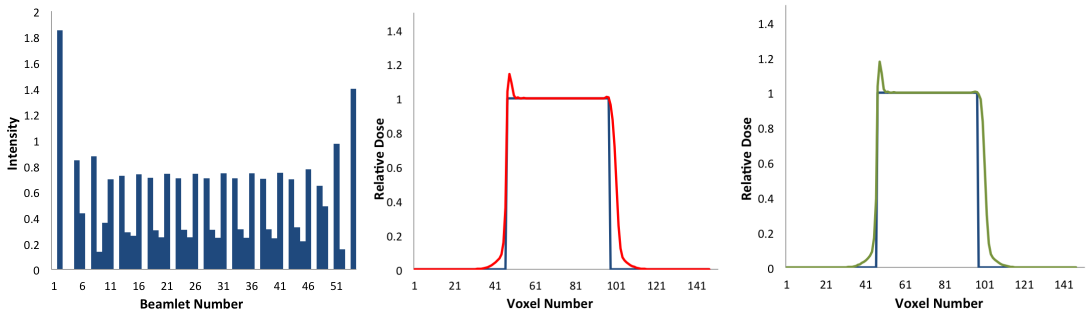
Note, that in the first island removal plan, solutions are internally consistent and propagation of constraint changes leads to feasible systems. In the case of the Boland et al. (2004) island constraints, the right hand constraints now violate the initial problem, as they no longer sum to 1, so the solver will have to find a new solution with which to branch on.

Appendix C

Further 1D Results

C.1 RFMO 1D Results

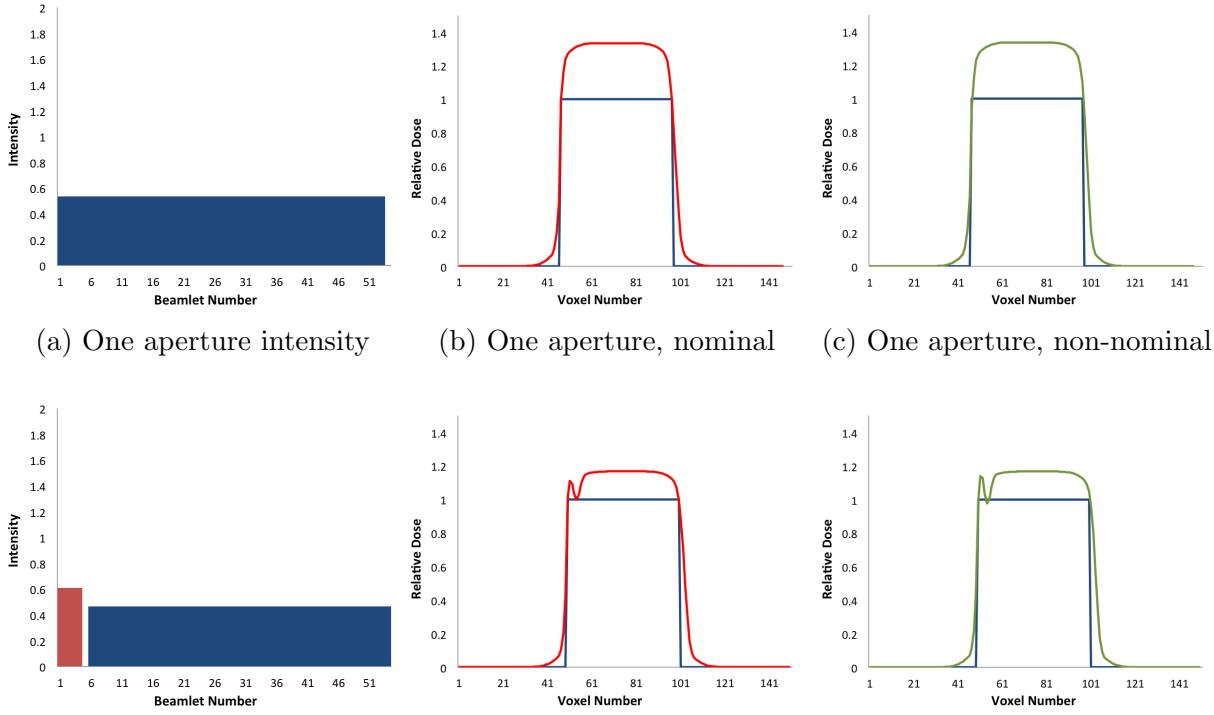
The results from the **M-FMO** model, as expected did not behave well under uncertainty. To address this behaviour, the **M-RFMO** model addresses the uncertainty, as shown in Figure C.1c. While the beamlet intensities in Figure C.1a and the nominal dose distribution in Figure C.1b looks similar to the **M-FMO** model results previously shown in Figure 4.2c, the non-nominal results in Figure C.1c set them apart, as there is no under-dosing of the clinical target volume (CTV), even in the worst case realization of uncertainty.



(a) Beamlet intensities (b) Robust nominal (c) Robust non-nominal

Figure C.1: RFMO model run on the 1D phantom. Figure (a) is intensities. Figures (b), (c) both show the prescribed dose in blue. The nominal realization is depicted in (b) in red, whereas the non-nominal realization is depicted in green in (c).

C.2 DAO 1D Results

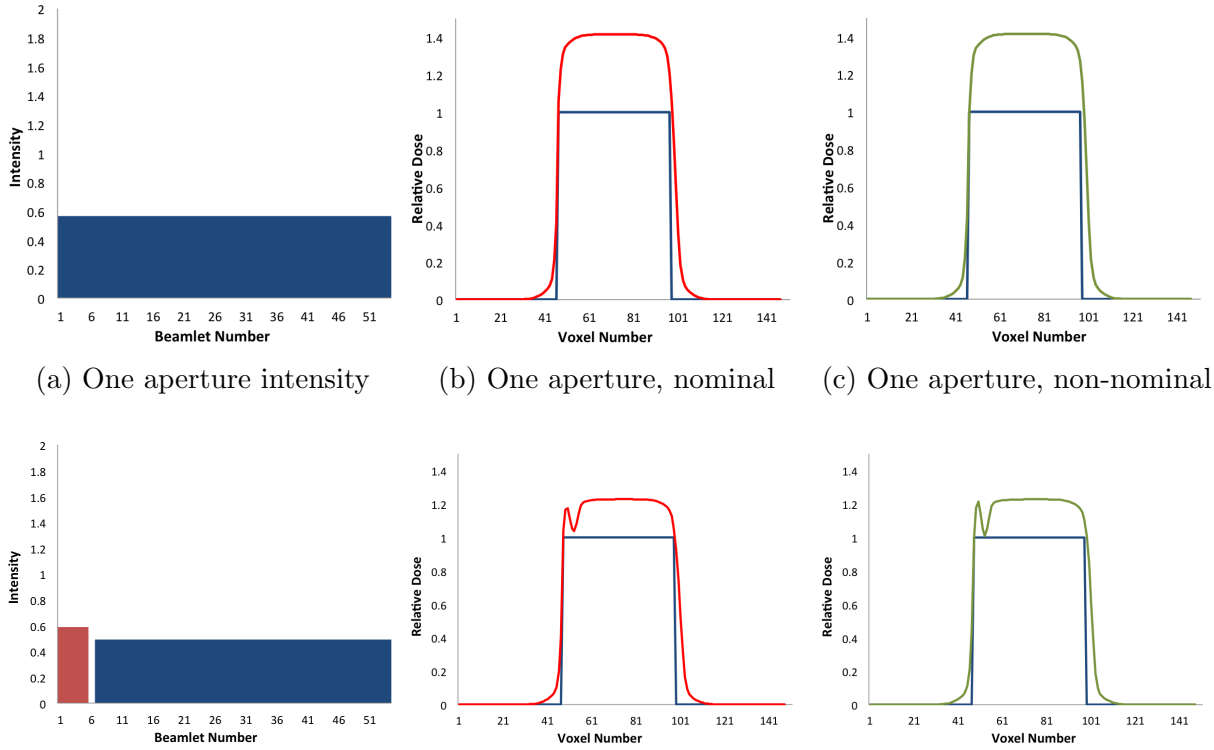


(a) One aperture intensity (b) One aperture, nominal (c) One aperture, non-nominal

(d) Two aperture intensities (e) Two apertures, nominal (f) Two apertures, non-nominal

Figure C.2: DAO models with 1 and 2 apertures run on the 1D phantom. Parts (a) and (d) show segment intensities. Figures (b), (c), (e) and (f) show the prescribed dose in blue, along with its realized dose in red (for nominal) and green (for non-nominal).

C.3 RDAO 1D Results



(a) One aperture intensity (b) One aperture, nominal (c) One aperture, non-nominal

(d) Two aperture intensities (e) Two apertures, nominal (f) Two apertures, non-nominal

Figure C.3: RDAO models with 1 and 2 apertures run on the 1D phantom. Parts (a) and (d) show segment intensities. Figures (b), (c), (e) and (f) show the prescribed dose in blue, along with its realized dose in red (for nominal) and green (for non-nominal).

Appendix D

Warm Start Pre-Post Heuristic Aperture

A single aperture example of a warm-start as depicted in the low fidelity C++ output. The aperture changes places in the final plan, based on its total value, since the algorithm is compatible with the no symmetry constraints, and the apertures must be distributed, as only one aperture per angle is feasible in the final plan. Within the apertures the symbol "`M`" represent a beamlet at the highest intensity in that particular beam, while "`x`" is any lower value.

The heuristic is demonstrated first on a case with free vertical placement (Section D.1), then on a case with no vertical gaps (Section D.2).

D.1 Free Vertical Aperture

```
[Warm start, pre-heuristic]

 Aperture 3, Angle 1 (Max 50.36)
----------------------------------
 1:
 2:     MM
 3:     MMM
 4:     MMM
 5:      M
 6:
 7:
 8:
 9:     MMM
10:     MMM
11:      M M
12:       MMM
13:       MM
14:      MM
15:      MM MM
16:         MMM
17:          M
18:    MM
19:    MMMMM
20:      MMMM
21:      MMMM   M
22:       MM   MMM
23:            MM
24:            Mx
25:           MM
26:           MM
27:         MMx
28:     MM  MM
29:    MMM
30:     MM   M
31:         MMM
32:         MMM
33:       MM M
34:      MMMM
35:      MMM
36:      MM
37:
38:
39:
40:
```

```
[Warm-start, post-heuristic]

 Aperture 1, Angle 1 (Max 50.36)
----------------------------------
 1:
 2:     MM
 3:     MMM
 4:     MMM
 5:      M
 6:
 7:
 8:
 9:     MMM
10:     MMM
11:      MMM
12:       MMM
13:       MM
14:      MM
15:      MMMMM
16:         MMM
17:          M
18:    MM
19:    MMMMM
20:      MMMM
21:      MMMMMMMM
22:       MMMMMMMM
23:            MM
24:            MM
25:           MM
26:           MM
27:         MMM
28:     MMMMMM
29:    MMM
30:     MMMMMM
31:         MMM
32:         MMM
33:       MMMM
34:      MMMM
35:      MMM
36:      MM
37:
38:
39:
40:
```

D.2 Restricted Vertical Aperture

```
[Warm start, pre-heuristic]

Segment 1, Angle 2 (Max 82.07)
---------------------------------
 1:
 2:
 3:
 4:                  MM
 5:                 MMM
 6:                  Mx
 7:
 8:                  M
 9:                 MMM
10:                 MMM
11:                M
12:               MMM
13:                MM
14:                MMM
15:             MMM MM
16:             MMM
17:              M    Mx
18:                   MM
19:                 x MM
20:                MMM
21:           M    MMM
22:          MMx   MMM
23:           MM
24:           MM
25:           MM
26:           MMM
27:             MMM
28:             MMM
29:              M  MMMM
30:              Mx MMMM
31:             MMM  MM
32:             MMM
33:              M M
34:               MMM
35:               MMM
36:                 MM
37:                 MM
38:                 MM
39:
40:
```

```
[Warm-start, post-heuristic]

  Segment 6, Angle 2 (Max 82.07)
---------------------------------
 1:
 2:
 3:
 4:                  MM
 5:                 MMM
 6:                  MM
 7:                  M
 8:                  M
 9:                 MMM
10:                MMMM
11:                M
12:               MMM
13:                MM
14:                MMM
15:             MMMMMM
16:             MMM
17:              MMMMMMM
18:                   MM
19:                 MMMM
20:                MMM
21:           MMMMMMMM
22:          MMMMMMMMM
23:           MM
24:           MM
25:           MM
26:           MMM
27:             MMM
28:             MMM
29:              MMMMMMM
30:              MMMMMMM
31:             MMMMMMM
32:             MMM
33:              MMM
34:               MMM
35:               MMM
36:                 MM
37:                 MM
38:                 MM
39:
40:
```