Chapter 4

A TUTORIAL ON RADIATION ONCOLOGY AND OPTIMIZATION

Allen Holder
Trinity University
Department of Mathematics
aholder@trinity.edu

Bill Salter
University of Texas Health Science Center, San Antonio
Associate Director of Medical Physics
Cancer Therapy & Research Center
bsalter@saci.org

Abstract Designing radiotherapy treatments is a complicated and important task that affects patient care, and modern delivery systems enable a physician more flexibility than can be considered. Consequently, treatment design is increasingly automated by techniques of optimization, and many of the advances in the design process are accomplished by a collaboration among medical physicists, radiation oncologists, and experts in optimization. This tutorial is meant to aid those with a background in optimization in learning about treatment design. Besides discussing several optimization models, we include a clinical perspective so that readers understand the clinical issues that are often ignored in the optimization literature. Moreover, we discuss many new challenges so that new researchers can quickly begin to work on meaningful problems.

Keywords: Optimization, Radiation Oncology, Medical Physics, Operations Research

4.1 Introduction

The interaction between medical physics and operations research (OR) is an important and burgeoning area of interdisciplinary work. The first optimization model used to aid the design of radiotherapy treatments was a linear model in 1968 [1], and since this time medical physicists have recognized that optimiza-

tion techniques can support their goal of improving patient care. However, OR experts were not widely aware of these problems until the middle 1990s, and the last decade has witnessed a substantial amount of work focused on medical physics. In fact, three of the four papers receiving the Pierskalla prize from 2000 to 2003 address OR applications in medical physics [14, 25, 54].

The field of medical physics encompasses the areas of Imaging, Health Physics, and Radiation Oncology. These overlapping specialties typically combine when a patient is treated. For example, images of cancer patients are used to design radiotherapy treatments, and these treatments are monitored to guarantee safety protocols. While optimization techniques are useful in all of these areas, the bulk of the research is in the area of Radiation Oncology, and this is our focus as well.

Specifically, we study the design and delivery of radiotherapy treatments. Radiotherapy is the treatment of cancerous tissues with external beams of radiation, and the goal of the design process is to find a treatment that destroys the cancer but at the same time spares surrounding organs. Radiotherapy is based on the fact that unlike healthy tissue, cancerous cells are incapable of repairing themselves if they are damaged by radiation. So, the idea of treatment is to deliver enough radiation to kill cancerous tissues but not enough to hinder the survival of healthy cells.

Treatment design was, and to a large degree still is, accomplished through a trial-and-error process that is guided by a physician. However, the current technological capabilities of a clinic make it possible to deliver complicated treatments, and to take advantage of modern capabilities, it is necessary to automate the design process. From a clinical perspective, the hope is to improve treatments through OR techniques. The difficulty is that there are numerous ways to improve a treatment, such as delivering more radiation to the tumor, delivering less radiation to sensitive organs, or shortening treatment time. Each of these improvements leads to a different optimization problem, and current models typically address one of these aspects. However, each decision in the design process affects the others, and the ultimate goal is to optimize the entire process. This is a monumental task, one that is beyond the scope of current optimization models and numerical techniques. Part of the problem is that different treatment goals require different areas of expertise. To approach the problem in its entirety requires a knowledge of modeling, solving, and analyzing both deterministic and stochastic linear, nonlinear, integer, and global optimization problems. The good news for OR experts is that no matter what niche one studies, there are related, important problems. Indeed, the field of radiation oncology is a rich source of new OR problems that can parlay new academic insights into improved patient care.

Our goals for this tutorial are threefold. First, we discuss the clinical aspects of treatment design, as it is paramount to understand how clinics assess treat-

ments. It is easy for OR experts to build and solve models that are perceived to be clinically relevant, but as every OR expert knows, there are typically many attempts before a useful model is built. The clinical discussions in this tutorial will help new researchers avoid traditional academic pitfalls. Second, we discuss the array of optimization models and relate them to clinical techniques. This will help OR experts identify where their strengths are of greatest value. Third, the bibliography at the end of this tutorial highlights some of the latest work in the optimization and medical literature. These citations will quickly allow new researchers to become acquainted with the area.

4.2 Clinical Practice

As with most OR applications, knowledge about the restrictions of the other discipline are paramount to success. This means that OR experts need to become familiar with clinical practice, and while treatment facilities share many characteristics, they vary widely in their treatment capabilities. This is because there are differences in available technology, with treatment machines, software, and imaging capabilities varying from clinic to clinic. A clinic's staff is trained on the clinic's equipment and rarely has the chance to experiment with alternate technology. There are many reasons for this: treatment machines and software are extremely expensive (a typical linear accelerator costs more than $1,000,000), time restrictions hinder exploration, etc.... A dialog with a clinic is invaluable, and we urge interested readers to contact a local clinic.

We begin by presenting a brief overview of radiation therapy (RT) concepts, with the hope of familiarizing the reader with some of the terminology used in the field, and then describe a "typical" treatment scenario, beginning with patient imaging and culminating with delivery of treatment.

4.2.1 Radiation Therapy Concepts and Terminology

Radiation therapy (RT) is the treatment of cancer and other diseases with ionizing radiation; ionizing radiation that is sufficiently energetic to dislodge electrons from their orbits and send them penetrating through tissue depositing their energy. The energy deposited per unit mass of tissue is referred to as *Absorbed Dose* and is the source of the biological response exhibited by irradiated tissues, be that lethal damage to a cancerous tumor or unwanted side effects of a healthy tissue or organ. Units of absorbed dose are typically expressed as Gy (pronounced Gray) or centiGray (cGy). One Gy is equal to one Joule (J) of energy deposited in one kilogram (kg) of matter.

Cancer is, in simple terms, the conversion of a healthy functioning cell into one that constantly divides, thus reproducing itself far beyond the normal needs of the body. Whereas most healthy cells divide and grow until they encounter another tissue or organ, thus respecting the boundaries of other tissues, cancer-

ous cells continue to grow into and over other tissue boundaries. The use of radiation to "treat" cancer can adopt one of two general approaches.

One delivery approach is used when healthy and cancerous cells are believed to co-mingle, making it impossible to target the cancerous cells without also treating the healthy cells. The approach adopted in such situations is called *fractionation*, which means to deliver a large total dose to a region containing the cancerous cells in smaller, daily fractions. A total dose of 60 Gy, for example, might be delivered in 2 Gy daily fractions over 30 treatment days. Two Gy represents a daily dose of radiation that is typically tolerated by healthy cells but not by tumor cells. The difference between the tolerable dose of tumor and healthy cells is often referred to as a *therapeutic advantage*, and radiotherapy exploits the fact that tumor cells are so focused on reproducing that they lack a well-functioning repair mechanism possessed by healthy cells. By breaking the total dose into smaller pieces, damage is done to tumor cells each day (which they do not repair) and the damage that is done to the healthy cells is tolerated, and in fact, repaired over the 24 hours before the next daily dose. The approach can be thought of as bathing the region in a dose that tumor cells will not likely survive but that healthy cells can tolerate.

The second philosophy that might be adopted for radiation treatment dosage is that of Radio*Surgery*. Radiosurgical approaches are used when it is believed that the cancer is in the form of a solid tumor which can be treated as a distinct target, without the presence of healthy, co-mingling cells. In such approaches it is believed that by destroying all cells within a physician-defined target area, the tumor can be eliminated and the patient will benefit. The treatment approach utilized is that of delivering one fraction of dose (i.e. a single treatment) which is extremely large compared to fractionated approaches. Typical radiosurgical treatment doses might be 15 to 20 Gy in a single fraction. Such doses are so large that all cells which might be present within the region treated to this dose will be destroyed. The treatment approach derives its name from the fact that such methods are considered to be the radiation equivalent to surgery, in that the targeted region is completely destroyed, or ablated, as if the region had been surgically removed.

The physical delivery of RT treatment can be broadly sub-categorized into two general approaches: *brachytherapy* and *external beam* radiation therapy (EBRT), each of which can be effectively used in the treatment of cancer. Brachytherapy, which could be referred to as *internal* radiation therapy, involves a minimally invasive surgical procedure wherein tiny radioactive "seeds" are deposited, or implanted, in the tumor. The optimal arrangement of such seeds, and the small, roughly spherical distribution of dose which surrounds them, has been the topic of much optimization related research. *External* beam radiation therapy involves the delivery of radiation to the tumor, or target, from a source of radiation located outside of the patient; thus the external compo-

Figure 4.1. A Linear Accelerator

Figure 4.2. A Linear accelerator rotating through various angles. Note that the treatment couch is rotated.

nent of the name. The radiation is typically delivered by a device known as a *linear accelerator*, or linac. Such a device is shown in Figures 4.1 and 4.2. The device is capable of rotating about a single axis of rotation so that beams may be delivered from essentially 360 degrees about the patient. Additionally, the treatment couch, on which the patient lies, can also be rotated through, typically, 180 degrees. The combination of gantry and couch rotation can facilitate the delivery of radiation beams from almost any feasible angle. The point defined by the physical intersection of the axis of rotation of the linac gantry with the central axis of the beam which emerges from the "head" of the linac is referred to as *isocenter*. Isocenter is, essentially, a geometric reference point associated with the beam of radiation, which is strategically placed inside of the patient to cause the tumor to be intersected by the treatment beam.

External beam radiation therapy can be loosely subdivided into the general categories of *conventional radiation therapy* and, more recently, *conformal radiation therapy* techniques. Generally speaking, conventional RT differs from conformal RT in two regards; complexity and intent. The goal of conformal techniques is to achieve a high degree of conformity of the delivered distribution of dose to the shape of the target. This means that if the target surface is convex in shape at some location, then the delivered dose distribution will also be convex at that same location. Such distributions of dose are typically represented in graphical form by what are referred to as *isodose distributions*. Much like the isobar lines on a weather map, such representations depict iso-levels of absorbed dose, wherein all tissue enclosed by a particular isodose level is understood to see that dose, or higher. An isodose line is defined as a percentage of the target dose, and an isodose volume is that amount of anatomy receiving at least that much radiation dose. Figure 4.3 depicts a conformal isodose

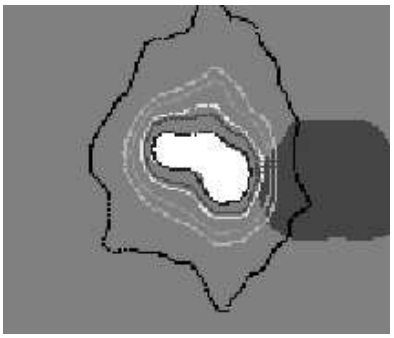

Figure 4.3. Conformal dose distribution. The target is shaded white and the brain stem dark grey. Isodose lines shown are 100%, 90%, 70%, 50%, 30% and 20%.

distribution used for treatment of a tumor. The high dose region is represented by the 60 Gy line(dark line), which can be seen to follow the shape of the convex shaped tumor nicely. The outer most curve is the 20 percent isodose curve, and the tissue inside of this curve receives at least 20 percent of the tumorcidal dose. By conforming the high dose level to the tumor, nearby healthy tissues are spared from the high dose levels. The ability to deliver a conformal distribution of dose to a tumor does not come without a price, and the price is complexity. Interestingly, the physical ability to deliver such convex-shaped distributions of dose has only recently been made possible by the advent of Intensity Modulating Technology, which will be discussed in a later section.

In conventional external beam radiation therapy, radiation dose is delivered to a target by the aiming of high-energy beams of radiation at the target from an origin point outside of the patient. In a manner similar to the way one might shine a diverging flashlight beam at an object to illuminate it, beams of radiation which are capable of penetrating human tissue are shined at the targeted tumor. Typically, such beams are made large enough to irradiate the entire target from each particular delivery angle that a beam might be delivered from. This is in contrast to IMRT approaches, which will be discussed in a later section, wherein each beam may treat only a small portion of the target. A fairly standard conventional delivery scheme is a so-called 2 field parallel-opposed arrangement (Figure 4.4). The figure depicts the treatment of a lesion of the liver created by use of an anterior to posterior-AP (i.e. from patient front to patient back) and posterior to anterior field-PA (i.e. from patient back to patient front). The isodose lines are depicted on computed tomography (CT) images of the patient's internal anatomy. The intersection of two different divergent fields delivered from two opposing angles results in a roughly rectangular shaped region of high dose (depicted by the resulting isodose lines for this plane). Note that the resulting high dose region encompasses almost the entire front to back

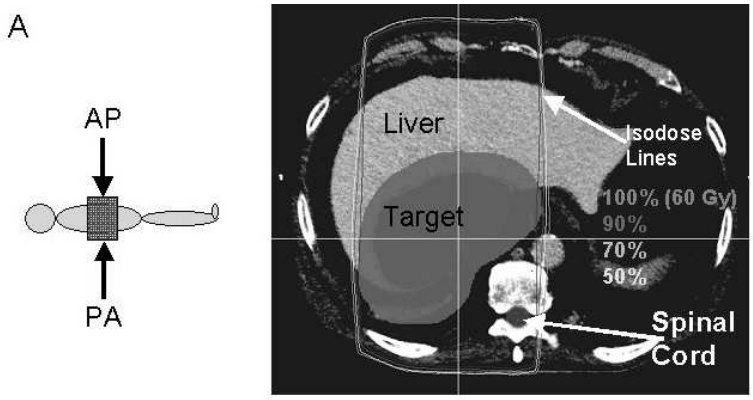

Figure 4.4. Two field, parallel opposed treatment of liver lesion.

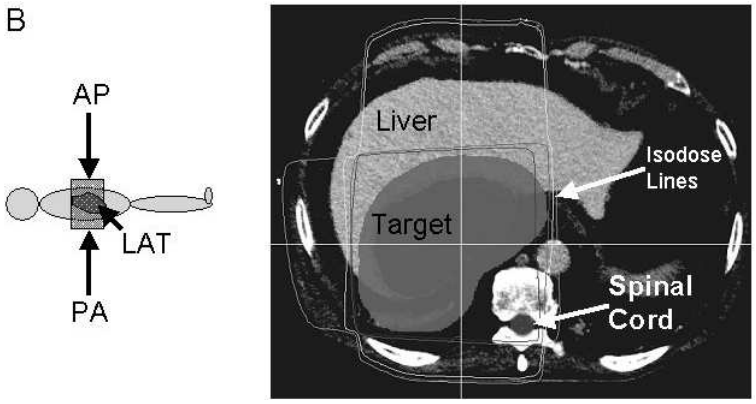

Figure 4.5. Three field treatment of liver lesion.

dimension of the patient, and that this region includes the spinal cord critical structure. The addition of a third field, which is perpendicular to the opposing fields, results in a box or square shaped distribution of dose, as seen in Figure 4.5. Note that the high dose region has been significantly reduced in size, but still includes the spinal cord. For either of these treatments to be viable, the dose prescribed by the physician to the high dose region would have to be maintained below the tolerance dose for the spinal cord (typically 44 Gy in 2 Gy fractions, to keep the probability of paralysis acceptably low) or a higher probability of paralysis would have to be accepted as a risk necessary to the survival of the patient. Such conventional approaches, which typically use 2-4

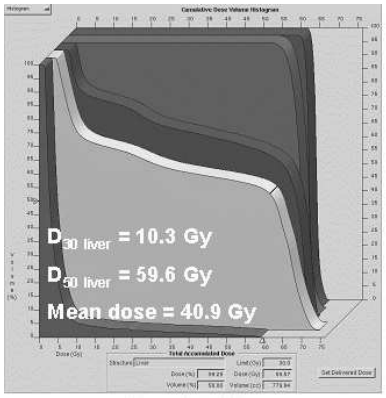

Figure 4.6. CDVH of two field treatment depicted in Figure 4.4

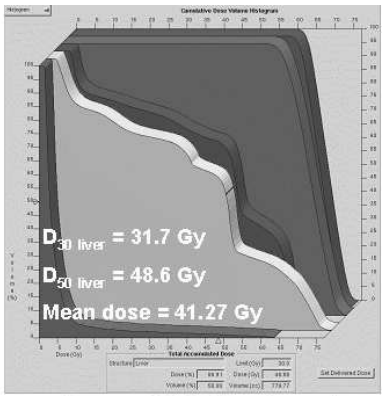

Figure 4.7. CDVH of three field treatment depicted in Figure 4.5

intersecting beams of radiation to treat a tumor, have been the cornerstone of radiation therapy delivery for years. By using customized beam blocking devices called "blocks" the shape of each beam can be matched to the shape of the projection of the target from each individual gantry angle, thus causing the total delivered dose distribution to match the shape of the target more closely.

The quality of a treatment delivery approach is characterized by several methods. Figures 4.6 and 4.7 show what is usually referred to as a "dose volume histogram" (DVH). More accurately, it is a cumulative DVH (CDVH). The curves describes the volume of tissue for a particular structure that is receiving a certain dose, or higher, and as such represents a plot of percentage of a particular structure versus Dose. The two CDVH's shown in Figures 4.6 and 4.7 are for the two conventional treatments shown in Figure 4.4 and 4.5, respectively. Five structures are represented in the figures from back to front, the Planning Target Volume (PTV) - a representation of the tumor that has been enlarged to account for targeting errors, such as patient motion; Clinical Target Volume (CTV) - The targeted tumor volume as defined by the physician on the 3-dimensional imaging set; The spinal Cord; the healthy, or non-targeted, Liver; all non-specific Healthy Tissue not specified as a critical structure. An ideal tumor CDVH would be a step function, with 100% of the target receiving exactly the prescribed dose (i.e. the 100% of prescribed level). Both treatments (i.e. Two Field and Three Field) produce near-step-function-like tumor DVH's. An ideal healthy tissue or critical structure DVH would be similar to that shown in Figures 4.6 and 4.7 for the Healthy Tissue, with 100% of the volume of the structure seeing 0% of the prescribed dose. The three field treatment in Figure 4.7 delivers less dose to the liver (second curve from front) and

spinal cord (third curve from front) in that the CDVH's for these structures are pushed to the left, towards lower delivered doses. With regard to volumetric sparing of the liver and spinal cord, the three field treatment can be seen to represent a superior treatment. Dose volume histograms capture the volumetric information that is difficult to ascertain from the isodose distributions, but they do not provide information about the location of high or low dose regions. Both the isodose lines and the DVH information are needed to adequately judge the quality of a treatment plan.

Thus far, the general concept of cancer and its treatment by delivery of tumorcidal doses of radiation have been outlined. The concepts underlying the various delivery strategies which have historically been employed were summarized, and general terminology has been presented. What has not yet been discussed is the method by which a treatment "plan" is developed. The treatment plan is the strategy by which beams of radiation will be delivered, with the intent of killing the tumor and sparing from collateral damage the surrounding healthy tissues. It is, quite literally, a plan of attack on the tumor. The process by which a particular patient is taken from initial imaging visit, through the treatment planning phase and, ultimately, to treatment delivery will now be outlined.

4.2.2 The Clinical Process

A patient is often diagnosed with cancer following the observation of symptoms related to the disease. The patient is then typically referred for imaging studies and/or biopsy of a suspected lesion. The imaging may include CT scans, magnetic resonance imaging (MRI) or positron emission tomography (PET). Each imaging modality provides different information about the patient, from bony anatomy and tissue density information provided by the CT scan, to excellent soft tissue information from the MRI, to functional information on metabolic activity of the tumor from the PET scan. Each of these sets of three dimensional imaging information may be used by the physician both for determining what treatment approach is best for the patient, and what tissues should be identified for treatment and/or sparing. If external beam radiotherapy is selected as the treatment option of choice, the patient will be directed to a radiation therapy clinic where they will ultimately receive radiation treatment(s) for a period of time ranging from a single day, to several weeks.

Before treatment planning begins, a 3-dimensional representation of the internal anatomy of the patient must be obtained. For treatment planning purposes such images are typically created by CT scan of the patient, because of CT's accurate rendering of the attenuation coefficients of each voxel of the patient, as will be discussed in the section on Dose Calculation. The 3-dimensional CT representation of the patient is built by a series of 2-dimensional

images (or slices), and the process of acquiring the images is often referred to as the Simulation phase. Patient alignment and immobilization is critical to this phase. The treatment that will ultimately be delivered will be based on these images, and if the patient's position and orientation at the time of treatment do not agree with this "treatment planning position", then the treatment will not be delivered as planned. In order to ensure that the patient's position can be reproduced at treatment time, an immobilization device may be constructed. Such devices may be as invasive as placing screws into the skull of the patient's head to ensure precise delivery of a radiosurgical treatment to the brain, to as simple as placing a rubber band around the feet of the patient to help them hold still for treatment of a lesion of the prostate. Negative molds of the patient's posterior can be made in the form of a cradle to assist in immobilization, and pediatric patient's may need to be sedated for treatment. In all cases, alignment marks are placed on the patient to facilitate alignment to the linac beam via lasers in the treatment vault.

Once the images and re-positioning device(s) are constructed, the treatment plan must be devised. Treatment plans are designed by a medical physicist, or a dosimetrist working under the direction of a medical physicist, all according to the prescription of a radiation oncologist. The planning process depends heavily on the treatment machine and software, and without discussing the nuances of different facilities, we explain the important distinction between forward and inverse planning. During treatment, a patient is exposed to the beams of radiation created by a high-energy radioactive source, and these beams deposit their energy as they travel through the anatomy (see Subsection 4.3). Treatment design is the process of selecting how these beams will pass through the patient so that maximum damage accumulates in the target and minimal damage in healthy tissues. Forward treatment design means that a physicist or dosimetrist manually selects beam angles and fluences (the **amount** of radiation delivered by a beam, controlled by the amount of time that a beam is "turned on"), and calculates how radiation dose accumulates in the anatomy as a result of these choices. If the beams and exposure times result in an unacceptable dose distribution, different beams and fluences are selected. The process repeats until a satisfactory treatment is found.

The success of the trial-and-error technique of forward planning depends on the difficulty of the treatment and the expertise of the planner. Modern technology is capable of delivering complicated treatments, and optimally designing a treatment that considers the numerous options is beyond the scope of human ability. As its name suggests, inverse planning reverses the forward paradigm. Instead of selecting beams and fluences, the idea is to prescribe absorbed dose in the anatomy, and then algorithmically find a collection of beams and fluences that satisfy the anatomical restrictions. This means that inverse planning

relies on optimization software, and the models that make this possible are the primary focus of this work.

Commercial software products blend forward and inverse planning, with most packages requiring the user to select the beam directions but not the fluences. The anatomical restrictions are defined on the patient images by delineating the target volume and any surrounding sensitive regions. A target dose is prescribed and bounds on the sensitive tissues are defined as percentages of this dose. For example, the tumor in Figure 4.4 is embedded in healthy surrounding liver, and located near the spinal cord. After manually identifying the tumor, the healthy liver, and the spinal cord on each 2-dimensional image, the dosimetrist enters a physician prescribed target dose, and then bounds how much radiation is delivered to the remaining structures as a percentage of the target dose. The dosimetrist continues by selecting a collection of beam angles and then uses inverse planning software to determine optimal beam fluences. The optimization problems are nontrivial, and modern computing power can calculate optimal fluence maps in about 20 minutes. We mention that commercial software varies substantially, with some using linear and quadratic models and others using complex, global optimization models solved by simulated annealing. Input parameters to the optimization software are often adjusted several times before developing a satisfactory treatment plan. Once an acceptable treatment plan has been devised, treatment of the patient, according to the radiation oncologist's dose and fractionation directive can begin.

In the following sections we investigate the underpinnings of the physics describing how radiation deposits energy in tissue, as well as many of the optimization models suggested in the literature. This discussion requires a more detailed description of a clinic's technology, and different clinical applications are explained as needed. We want to again stress that a continued dialog with a treatment facility is needed for OR techniques to impact clinical practice. In the author's experience, medical physicists are very receptive to collaboration. The OR & Oncology web site (`http://www.trinity.edu/aholder/HealthApp/oncology/`) lists several interested researchers, and we encourage interested readers to contact people on this list.

4.3 Dose Calculations

Treatment design hinges on the fact that we can accurately model how beams of high-energy radiation interact with the human anatomy. While an entire tutorial could be written on this topic alone, our objective is to provide the basics of how these models work. An academic dose model does not need to precisely replicate clinical dose calculations but does need to approximate how radiation is deposited into the anatomy. We develop a simple, 2-dimensional,

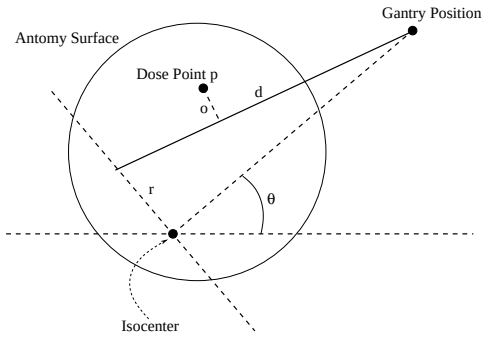

Figure 4.8. The geometry involved in calculating the contribution of sub-beam (θ, r) to the Dose Point.

continuous dose model and its discrete counterpart. The 3-dimensional model is a natural extension but is more complicated to describe.

Consider the diagram in Figure 4.8. The isocenter is in the lower part of the diagram, and the gantry is rotated to angle θ. Patients are often shielded from parts of the beam by devices such as a *multileaf collimator,* which are discussed in detail in Section 4.4. The sub-beam considered in Figure 4.8 is (θ, r), and we calculate this sub-beam's contribution to the dose point p. A simple but effective model uses the depth of the dose point along sub-beam (θ, r), labeled d, and the distance from the dose point to the sub-beam, denoted o (o is used because this is often referred to as the 'off axis' distance). The radiation being delivered along sub-beam (θ, r) attenuates and scatters as it travels through the anatomy. Attenuation means that photons of the beam are removed by scattering and absorption interactions as depth increases. So, if the dose point was directly in the path of sub-beam (θ, r), it would receive more radiation the closer it is to the gantry. While the dose point is not directly in the path of sub-beam (θ, r), it still receives radiation from this sub-beam because of scatter. A common model assumes that the percentage of deposited dose falls exponentially as d and o increase. So, if $g(\theta, r)$ is the amount of energy being delivered along sub-beam (θ, r) (or equivalently, the amount of time this sub-beam is not blocked), the dose point receives

$$g(\theta, r)e^{\eta o}e^{\mu d}$$

units of radiation from sub-beam (θ, r), where μ and η are parameters decided by the beam's energy. If $L_\theta = \{r : (\theta, r)$ is a sub-beam of angle $\theta\}$, we have that the total (or integral) amount of radiation delivered to the dose point from all gantry positions is

$$D_p = \int_L g(\theta, r)e^{\eta o}e^{\mu d}d\theta. \tag{3.1}$$

Calculating the amount of radiation deposited into the anatomy is a *forward* problem, meaning that the amount of radiation leaving the gantry is known and the radiation deposited into the patient is calculated. An *inverse* problem is one in which we know the radiation levels in the anatomy and then find a way to control the beams at the gantry to achieve these levels. Treatment design problems are inverse problems, as our goal is to specify the distribution of dose being delivered and then calculate a 'best' way to satisfy these limits. As an example, if the dose point p' is inside a tumor, we may desire that $D_{p'}$ be at least 60Gy. Similarly, if the dose point p'' was in a nearby, sensitive organ, we may want $D_{p''}$ to be no greater than 20Gy. So, our goal is to calculate $g(\theta, r)$ for each sub-beam so that

$$D_{p'} = \int_L g(\theta, r) e^{\eta o} e^{\mu d} d\theta \geq 60, \tag{3.2}$$

$$D_{p''} = \int_L g(\theta, r) e^{\eta o} e^{\mu d} d\theta \leq 20, \text{ and} \tag{3.3}$$

$$g(\theta, r) \geq 0 \text{ for all } (\theta, r). \tag{3.4}$$

From these constraints it is obvious that we need to invert the integral transformation that calculates dose, and while there are several numerical techniques to do so, such techniques do not guarantee the non-negativity of g. Moreover, the system may be inconsistent, which means the physician's restrictions are not possible. However, the typical case is that there are many choices of $g(\theta, r)$ that satisfy the physician's requirements, and in such a situation, the optimization question is which collection of $g(\theta, r)$'s is best?

The discrete approximation to (3.1) depends on a finite collection of angles and sub-beams. Instead of the continuous variables θ and r, we assume that there are q gantry positions, indexed by a, and that each of the gantry positions is comprised of τ sub-beams, indexed by s. The amount of radiation to deliver along sub-beam (a, s), which is equivalent to deciding how long to leave this sub-beam unblocked, is denoted by $x_{(a,s)}$. For the dose point p, we let $a_{(p,a,s)}$ be $e^{\eta o} e^{\mu d}$. The discrete counterpart of (3.1) is

$$\sum_{(a,s)} a_{(p,a,s)} x_{(a,s)} \approx D_p = \int_L g(\theta, r) e^{\eta o} e^{\mu d} d\theta.$$

We construct the *dose matrix*, A, from the collection of $a_{(p,a,s)}$'s by indexing the rows and columns of A by p and (a, s), respectively.

The dose matrix A adequately models how radiation is deposited into the anatomy as the gantry rotates around a single isocenter, which can be located at any position within the patient. Moreover, modern linear accelerators are capable of producing beams with different energies, and these energies correspond to different values of μ and η. So, for each isocenter i and beam energy

e, we construct the dose matrix $A_{(i,e)}$. The entire dose matrix is then

$$\left[A_{(1,1)}|A_{(1,2)}|\cdots|A_{(1,E)}|A_{(2,1)}|\cdots|A_{(2,E)}|\cdots|A_{(I,1)}|\cdots|A_{(I,E)}\right],$$

where there are I different isocenters and E different energies. The index on x is adjusted accordingly to (i,e,a,s) so that $x_{(i,e,a,s)}$ is the radiation leaving the gantry along sub-beam (a,s) while the gantry is rotating around isocenter i and the linear accelerator is producing energy e. Many of the examples in this chapter use a single isocenter, and all use a single energy, but the reader should be aware that clinical applications are complicated by the possibility of having multiple isocenters and energies.

The cumulative dose at point p is the p^{th} component of the vector Ax, denoted by $(Ax)_p$. We now see that the discrete approximations to (3.2) - (3.4) are

$$D_{p'} \approx (Ax)_{p'} \geq 60, \quad D_{p''} \approx (Ax)_{p''} \leq 20 \quad \text{and} \quad x \geq 0.$$

As before, there may not be an x that satisfies the system. In this case, we know that the physician's bounds are not possible with the discretization described by A. However, there may be a different collection of angles, sub-beams, and isocenters, and hence a different dose matrix, that allows the physician's bounds to be satisfied. Selecting the initial discretization is an important and challenging problem that we address in Section 4.4.

The vector x is called a *treatment plan* (or more succinctly a plan) because it indicates how radiation leaves the gantry as it rotates around the patient. The linear transformation $x \mapsto Ax$ takes the radiation at the gantry and deposits it into the anatomy. Both the continuous model and the discrete model are linear —i.e. the continuous model is linear in g and the discrete model is linear in x. The linearity is not just an approximation, as experiments have shown that the dose received in the anatomy scales linearly with the time a sub-beam is left unblocked. So, linearity is not just a modeling assumption but is instead natural and appropriate.

The treatment area and geometry are different from patient to patient, and the clinical dose calculations are patient specific. Also, depending on the region being treated, we may modify the attenuation to reflect different tissue densities, with the modified distances being called the *effective* depth and off-axis distance. As an example, if the sub-beam (a,s) is passing through bone, the effective depth is increased so that the attenuation (exponential decay) of the beam is greater as it travels through the bone. Similarly, if the sub-beam is passing through air, the effective depth is shortened so that less attenuation occurs.

We reiterate that there are numerous models of widely varying complexity that calculate how radiation is deposited into the anatomy. Our goal here was to introduce the basic concepts of a realistic model. Again, it is important

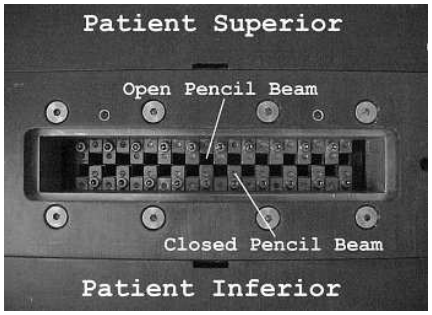

Figure 4.9. A tomotherapy multileaf collimator. The leaves are either open or closed.

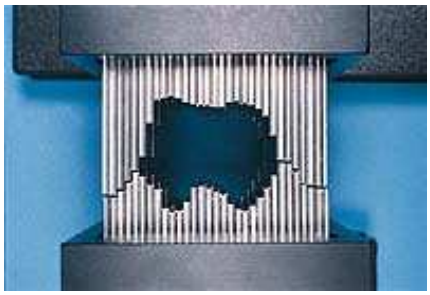

Figure 4.10. A multileaf collimator for static gantry IMRT.

to remember that for academic purposes, the dose calculations need only be reasonably close to those used in a clinic.

4.4 Intensity Modulated Radiotherapy (IMRT)

A recent and important development in the field of RT is that of Intensity Modulated Radiotherapy (IMRT). Regarded by many in the field as a quantum leap forward in treatment delivery capability, IMRT allows for the creation of dose distributions that were previously not possible. As a result, IMRT has allowed for the treatment of patients that previously had no viable treatment options.

The distinguishing feature of IMRT is that the normally large, rectangular beam of radiation produced by a linear accelerator is shaped by a multileaf collimator into smaller so-called pencil beams of radiation, each of which can be varied, or modulated, in intensity (or fluence). Figures 4.9 and 4.10 show images of two multileaf collimators used for delivery of IMRT treatments. The leaves in Figure 4.9 are pneumatically controlled by individual air valves that cause the leaves to open or close in about 30 to 40 milliseconds. By varying the amount of time that a given leaf is opened from a particular gantry angle the intensity, or fluence, of the corresponding pencil beam is varied, or modulated. This collimator is used in *tomotherapy*, which treats the 3-dimensional problem as a series of 2-dimensional sub-problems. In tomotherapy a treatment is delivered as a summation of individually delivered "slices" of dose, each of which is optimized to the specific patient anatomy that is unique to the treatment slice. Tomotherapy treatments are delivered by rapidly opening and closing the leaves as the gantry swings continuously about the patient.

The collimator in Figure 4.10 is used for static gantry IMRT. This is a process where the gantry moves to several static locations, and at each position the

patient is repeatedly exposed to radiation using different leaf configurations. Adjusting the leaves allows for the modulation of the fluence that is delivered along each of the many sub-beams. This allows the treatment of different parts of the tumor with different amounts of radiation from a single angle. Similar to tomotherapy, the idea is to accumulate damage from many angles so that the target is suitably irradiated.

4.4.1 Clinically Relevant IMRT Treatments

For an optimized IMRT treatment to be clinical useful, the problem must be modeled assuming clinically reasonable values for the relevant input variables.The clinical restrictions of IMRT depend on the type of delivery used. Tomotherapy has fewer restrictions with regard to gantry angles, in that any and all of the possible pencil beams may be utilized for treatment delivery. The linac gantry performs a continuous arc about the patient regardless of whether or not pencil beams from each gantry angle are utilized by the optimized delivery scheme. This is in contrast to the static gantry model where clinical time limitations make it impractical to deliver treatments comprised of, typically, more than 7-9 gantry angles. This means that the optimization process must necessarily select the optimal set of 7 to 9 gantry angles of approach from which to deliver pencil beams, from the much larger set of *possible* gantry angles of delivery, which leads to mixed integer problems. For either delivery approach, the gantry angles considered must, of course, be limited to those angles that do not lead to collisions of the gantry and treatment couch or patient. Clinical optimization software for static gantry approaches typically requires that the user pre-select the static gantry angles to be used. Such software provides visualization tools that help the user intelligently select gantry angles that can be visually recognized to provide unobstructed angles. This technique serves to reduce the complexity of the problem to manageable levels but does not, of course, guarantee a truly optimal solution. The continuous gantry movement of a tomotherapy treatment is approximated by modeling the variation of leaf positions every 5^o, and the large number of potential angles coupled with a typical fluence variation of 0 to 100% in steps of 10% causes tomotherapy to possess an extremely large solution space.

4.4.2 Optimization Models

Before we begin describing the array of optimization models that are used to design treatments, we point out that several reviews are already in the literature. Shepard, Ferris, Olivera, and Mackie have a particularly good article in *SIAM Review* [57]. Other OR reviews include the exposition by Bartolozzi, et. al. in the *European Journal of Operations Research* [2] and the introductory material by Holder in the *Handbook of Operations Research/Management Sci-*

ence Applications in Health Care [24]. In the medical physics literature, Rosen has a nice review in *Medical Physics* [55]. We also mention two web resources: the *OR & Oncology Web Site* at `www.trinity.edu/aholder/HealthApp/oncology/` and *Pub Med* at `www.ncbi.nlm.nih.gov/`. The medical literature can be overwhelming, with a recent search at Pub Med on "optimization" and "oncology" returning 652 articles.

We begin our review of optimization models by studying linear programs. This is appropriate because dose deposition is linear and because linear programming is common to all OR experts. Also, many of the models in the literature are linear [1, 22, 24, 33, 36]. Let A be the dose deposition matrix described in Section 4.3, and partition the rows of A so that

$$A = \begin{bmatrix} A_T \\ A_C \\ A_N \end{bmatrix} \begin{array}{l} \leftarrow \text{ Target Volume} \\ \leftarrow \text{ Critical Structures} \\ \leftarrow \text{ Unrestricted, Normal Tissue,} \end{array}$$

where A_T is $m_T \times n$, A_C is $m_C \times n$, and A_N is $m_N \times n$. The sets T, C, and N partition the dose points in the anatomy, with T containing the dose points in the target volume, C containing the dose points in the critical structures, and N contains the remaining dose points. We point out that A is typically large. For example, if we have a 512×512 patient image with each pixel having its own dose point, then A has $262,144$ rows. Moreover, A has $360,000$ columns if we design a treatment using 4 energies, 5 isocenters, 360 angles per isocenter, and 50 sub-beams per angle. So, for a single image we would need to apriori make 9.44×10^{10} dose calculations. Since there are usually several images involved, it is easy to see that generating the data for a model instance is time consuming. Romeijn, Ahuja, Dempsey and Kumar [54] have developed a column generation technique to address this computational issue.

The information provided by a physician to build a model is called a *prescription*. This clinical information varies from clinic to clinic depending on the design software. A prescription is initially the triple (TG, CUB, NUB), where TG is a m_T vector containing the goal dose for the target volume, CUB is a m_C vector listing the upper bounds on the critical structures, and NUB is a m_N vector indicating the highest amount of radiation that is allowed in the remaining anatomy. In many clinical settings, NUB is not decided before the treatment is designed. However, clinics do not routinely allow any part of the anatomy to receive doses above 10% of the target dose, and one can assume that $NUB = 1.1 \times TG$.

The simplest linear models are feasibility problems [5, 48]. In these models the goal is to satisfy

$$A_T x \geq TG, \quad A_C x \leq CUB, \quad A_N x \leq NUB, \text{ and } \quad x \geq 0.$$

The consistency of this system is not guaranteed because physicians are often overly demanding, and many authors have complained that infeasibility is a shortcoming of linearly constrained models [22, 33, 44, 55]. In fact, the argument that feasibility alone correctly addresses treatment design is that the region defined by these constraints is relatively small, and hence, optimizing over this region does not provide significant improvements in treatment quality.

If a treatment plan that satisfies the prescription exists, the natural question is which plan is best. The immediate, but naive, ideas are to maximize the tumor dose or minimize the critical structure dose. Allowing e to be the vector of ones, where length is decided by the context of its use, these models are variants of

$$\max\{e^T A_T x : A_T x \geq TG,\ A_C x \leq CUB,\ A_N x \leq NUB,\ x \geq 0\}, \quad (4.1)$$

$$\min\{e^T A_C x : A_T x \geq TG,\ A_C x \leq CUB,\ A_N x \leq NUB,\ x \geq 0\}, \quad (4.2)$$

$$\begin{aligned}\max\{z : A_T x \geq TG + ze,\ A_C x \leq CUB,\\ A_N x \leq NUB,\ x \geq 0,\ z \geq 0\}, \text{ or}\end{aligned} \quad (4.3)$$

$$\begin{aligned}\min\{z : A_T x \geq TG,\ A_C x \leq CUB - ze,\\ A_N x \leq NUB,\ x \geq 0,\ z \geq 0\}.\end{aligned} \quad (4.4)$$

Models (4.1) and (4.2) maximize and minimize the cumulative dose to the tumor and critical structures, respectively. Model (4.3) maximizes the minimum dose received by the target volume and (4.4) minimizes the maximum dose received by a critical structure.

The linear models in (4.1) - (4.4) are inadequate for several reasons. As already mentioned, if the feasibility region is empty, most solvers terminate by indicating that infeasibility has been detected. While there is a substantial literature on analyzing infeasibility (see for example [7–9, 20, 21]), discovering the source of infeasibility is an advanced skill, one that we can not expect physicians to acquire. Model (4.1) further suffers from the fact that it is often unbounded. This follows because it is possible to have sub-beams that intersect the tumor but that do not deliver numerically significant amounts of radiation to the critical structures. In this situation, it is obvious that we can make the cumulative dose to the tumor as large as possible. Lastly, these linear models have the unintended consequence of achieving the physician's bounds. For example, as model (4.3) increases the dose to the target volume, it is also increasing the dose to the critical structures. So, an optimal solution is likely to achieve the upper bounds placed on the critical structures, which is not desired. We also point out that because simplex based optimizers terminate with an ex-

treme point solution, we are **guaranteed** that several of the inequalities hold with equality when the algorithm terminates [22]. So, the choice of algorithm plays a role as well, a topic that we address later in this section.

An improved linear objective was suggested by Morrill [45]. This objective maximizes the difference between the dose delivered to the tumor and the dose received by the critical structures. For example, consider the following models,

$$\begin{aligned}\max\{e^T A_T x - e^T A_C x : A_T x \geq TG,\\ A_C x \leq CUB, A_N x \leq NUB, x \geq 0\}\end{aligned} \quad \text{and} \tag{4.5}$$

$$\begin{aligned}\max\{z - q : A_T x \geq TG + ze, A_C x \leq CUB - qe,\\ A_N x \leq NUB, x \geq 0, z \geq 0, q \geq 0\}.\end{aligned} \tag{4.6}$$

These models attempt to overcome the difficulty of attaining the prescribed limits on the target volume and the critical structures. However, model (4.5) is often unbounded for the same reason that model (4.1) is. Also, both of these models are infeasible if the physician's goals are overly restrictive.

Many of the limitations of models (4.1) - (4.6) are addressed by parameterizing the constraints. This is similar to goal programming, where we think of the prescription as a goal instead of an absolute bound. Constraints that use parameters to adjust bounds are called *elastic*, and Holder [25] used these constraints to build a linear model that overcame the previous criticisms. Before presenting this model, we discuss another pitfall that new researchers often fall into. The target volume is not exclusively comprised of tumorous cells, but rather normal and cancerous cells are interspersed throughout the region. Recall that external beam radiotherapy is successful because cancerous cells are slightly more susceptible to radiation damage than are normal tissues. The goal is to deliver enough dose to the target volume so that the cancerous cells die but not enough to kill the healthy cells. So, one of the goals of treatment planning is to find a plan that delivers a uniform dose to the tumor. The model suggested in [25] uses a uniformity index, ρ, and sets the tumor lower bound to be $TLB = TG - \rho e$ and the tumor upper bound to be $TUB = TG + \rho e$ (typical values of ρ in the literature range from 0.02 to 0.15). Of course, there is no reason why the upper and lower bounds on the target volume need to be a fixed percentage of TG, and we extended a prescription to be the 4-tuple (TUB, TLB, CUB, NUB), where TUB and TLB are arbitrary positive vectors such that $TUB \leq TLB$. Consider the model below.

$$\begin{aligned}\min\{\omega \cdot l^T \alpha + u_C^T \beta + u_N^T \gamma : TLB - L\alpha \leq A_T x \leq TUB,\\ A_C x \leq CUB + U_C \beta, A_N x \leq NUB + U_N \gamma, -CUB \geq U_C \beta,\\ 0 \leq U_N \gamma, 0 \leq x\}\end{aligned} \tag{4.7}$$

In this model, the matrices L, U_C, and U_N are assumed to be non-negative, semimonotone matrices with no row sum being zero. The term La measures the target volume's under dose, and the properties of L ensure that the target volume receives the minimum dose if and only if α is zero. Similarly, $U_C\beta$ and $U_N\gamma$ measure the amount the non-cancerous tissues are over their prescribed bounds. The difference between β and γ is that they have different lower bounds. If $U_B\beta$ attains its lower bound of $-CUB$, we have found a treatment plan that delivers no radiation to the critical structures. The lower bound on $U_N\gamma$ is 0, which indicates that we are willing to accept any plan where the dose to the non-critical tissue is below its prescribed limit.

The objective function in (4.7) penalizes adverse deviations and rewards desirable deviations. The term $l^T\alpha$ penalizes under dosing the target volume and $u_N^T\gamma$ penalizes overdosing the normal tissue. The role of $u_C^T\beta$ is twofold. If β is positive, it penalizes overdosing the critical structures, and if β is negative, it rewards under dosing the critical structures. The parameter ω weights the importance placed on attaining tumor uniformity.

One may ask why model 4.7 is stated in such general terms of measure and penalty. The reason is that there are two standard ways to measure and penalize discrepancies. If we want the sum of the discrepancies to be the penalty, then we let l, u_C, and u_N be vectors of ones and L, U_C, and U_N be the identity matrices. Alternatively, if we want to penalize the largest deviation, we let l, u_c, and u_N each be the scalar 1 and L, U_C and U_N be vectors of ones. So, this one model allows deviations to be measured and penalized in many ways but has a single mathematical analysis that applies to all of these situations.

The model in (4.7) has two important theoretical advantages to the previous models. The first result states that the elastic constraints of the model guarantee that both the primal and dual problems are feasible.

THEOREM 4.1 (HOLDER [25]) *The linear model in 4.7 and its dual are strictly feasible, meaning that each of the constraints can simultaneously hold without equality.*

The conclusion of Theorem 4.1 is not surprising from the primal perspective, but the dual statement requires all of the assumptions placed on l, u_C, u_N, L, U_C and U_N. The feasibility guaranteed by this result is important for two reasons. First, if the physician's goals are not possible, this model minimally adjusts the prescription to attain feasibility. Hence, this model returns a treatment plan that matches the physician's goals as closely as possible even if the original desires were not achievable. Second, Theorem 4.1 assures us that interior-point algorithms can be used, and we later discuss why these techniques are preferred over simplex based approaches.

The second theoretical guarantee about model (4.7) is that it provides an analysis certificate. Notice that the objective function is a weighted sum of the

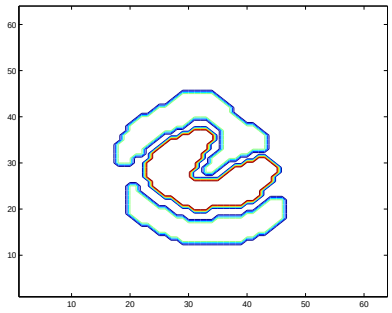

Figure 4.11. A tumor surrounded by two critical structures. The desired tumor dose is $80\text{Gy}\pm3\%$, and the critical structures are to receive less than 40Gy.

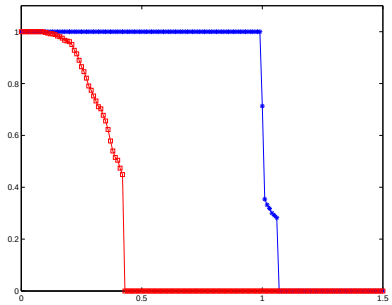

Figure 4.12. The dose-volume histogram indicates that 100% of the tumor receives its goal dose and that about 60% of the critical structures is below its bound of 40Gy.

competing goals of delivering a large amount of radiation to the target volume and a small amount of radiation to the remaining anatomy. The next result shows that the penalty assigned to under dosing the target volume is uniformly bounded by the inverse of ω.

THEOREM 4.2 (HOLDER [25]) *Allowing* $(x^*(\omega), \alpha^*(\omega), \beta^*(\omega), \gamma^*(\omega))$ *to be an optimal solution for a particular* ω*, we have that* $l^T\alpha^*(\omega) = O(1/\omega)$.

A consequence of Theorem 4.2 is that there is a positive scalar κ such that for any positive ω, we have that $l^T\alpha^*(\omega) \leq \kappa/\omega$. This is significant because we can apriori calculate an upper bound on κ that depends on the dose matrix A. If κ' is this upper bound, we have that $l^T\alpha^*(\omega) \leq \kappa/\omega \leq \kappa'/\omega$. So, we can make $l^T\alpha^*(\omega)$ as small as we want by selecting a sufficiently large ω. If we use this ω and $l^T\alpha^*(\omega)$ is larger that κ'/ω, then we know with certainty that we can not achieve the desired tumor uniformity. Moreover, we know that if $l^T\alpha^*(\omega)$ is less than κ'/ω and the remaining terms of the objective function are positive, then we can attain the tumor uniformity only at the expense of the critical structures. So, the importance of Theorem 4.2 is that it provides a guaranteed analysis.

Consider the geometry in Figure 4.11, where a tumor is surrounded by two critical structures. The goal dose for the tumor is $80\text{Gy}\pm3\%$, and the upper bound on the critical structures is 40Gy. Figure 4.12 is a dose-volume histogram for the treatment designed by Model (4.7), and from this figure we see that 100% of the tumor receives it's goal dose. Moreover, we see that about 60% of the critical structure is below its upper bound of 40Gy.

Outside of linear models, the most prevalent models are quadratic [36, 43, 61]. A popular quadratic model is

$$\min\{\|A_T x - TG\|_2 : A_C x \leq CUB, A_N x \leq NUB, x \geq 0\}. \tag{4.8}$$

This model attempts to exactly attain the goal dose over the target volume while satisfying the non-cancerous constraints. This is an attractive model because the objective function is convex, and hence, local search methods like gradient descent and Newton's method work well. However, the non-elastic, linear constraints may be inconsistent, and this model suffers from the same infeasibility complaints of previous linear models. Some medical papers have suggested that we instead solve

$$\begin{aligned}\min\{\|A_T x - TG\|_2 + \|A_C x - CUB\|_2 + \\ \|A_N x - NUB\|_2 : x \geq 0\}.\end{aligned} \tag{4.9}$$

While this model is never infeasible, it is inappropriate for several reasons. Most importantly, this model attempts to **attain** the bounds placed on the non-cancerous tissue, something that is clearly not desirable. Second, this model could easily provide a treatment plan that under doses the target volume and over doses the critical structures, even when there are plans that sufficiently irradiate the tumor and under irradiate the critical structures. A more appropriate version of (4.9) is

$$\min\{\|A_T x - TG\|_2 + \|A_C x\|_2 + \|A_N x\|_2 : x \geq 0\}, \tag{4.10}$$

but again, without constraints on the non-cancerous tissues, there is no guarantee that the prescription is (optimally) satisfied.

The only real difference between the quadratic and linear models is the manner in which deviations from the prescription are measured. Since there is no clinically relevant reason to believe that one measure is more appropriate than another, the choice is a personal preference. In fact, all of the models discussed so far have a linear and a quadratic counterpart. For example, the quadratic manifestation of (4.7) is

$$\begin{aligned}\min\{\omega \cdot \|l^T \alpha\|_2 + \|u_C^T \beta\|_2 + \|u_N^T \gamma\|_2 : TLB - L\alpha \leq A_T x \leq TUB, \\ A_C x \leq CUB + U_C \beta, A_N x \leq NUB + U_N \gamma, -CUB \geq U_C \beta, \\ 0 \leq U_N \gamma, 0 \leq x\}\end{aligned} \tag{4.11}$$

and the linear counterparts of (4.10) are

$$\min\{\|A_T x - TG\|_1 + \|A_C x\|_1 + \|A_N x\|_1 : x \geq 0\}, \text{ and} \tag{4.12}$$

$$\min\{\|A_T x - TG\|_\infty + \|A_C x\|_\infty + \|A_N x\|_\infty : x \geq 0\}. \tag{4.13}$$

We point out that Theorems 4.1 and 4.2 apply to model (4.11), and in fact, these results hold for any of the p-norms.

Each of the above linear and quadratic models attempts to 'optimally' satisfy the prescription, but the previous prescriptions of (TG, CUB, NUB) and (TLB, TUB, CUB, NUB) do not adequately address the physician's goals. The use of dose-volume histograms to judge treatments enables physicians to express their goals in terms of tissue percentages that are allowed to receive specified doses. For example, we could say that we want less than 80% of the lung to receive more than 60% of the target dose, and further, that less than 20% of the lung receives more than 75% of the target dose.

Constraints that model the physician's goals in terms of percent tissue receiving a fixed dose are called *dose-volume constraints*. These restrictions are biologically natural because different organs react to radiation differently. For example, the liver and lung are modular, and these organs are capable of functioning with substantial portions of their tissue destroyed. Other organs, like the spinal cord and bowel, lose functionality as soon as a relatively small region is destroyed. Organs are often classified as *rope* or *chain* organs [19, 53, 63, 64], with the difference being that rope organs remain functional even with large amounts of inactive tissue and that chain organs fail if a small region is rendered useless. Rope organs typically fail if the entire organ receives a relatively low, uniform dose, and the radiation passing through these organs should be accumulated over a contiguous portion of the tissue. Alternatively, chain organs are usually capable of handling larger, uniform doses over the entire organ, and it is desirable to disperse the radiation over the entire region. So, there are biological differences between organs that need to be considered. Dose-volume constraints capture a physician's goals for these organs.

We need to alter the definition of a prescription to incorporate dose-volume constraints. First, we partition C into $C^1, C^2, \ldots, C^K$, where C^k contains the dose points within the k^{th} critical structure. We know have that

$$A_C = \begin{bmatrix} A_{C^1} \\ A_{C^2} \\ \vdots \\ A_{C^K} \end{bmatrix} \begin{array}{l} \leftarrow \text{ Critical Structure 1} \\ \leftarrow \text{ Critical Structure 2} \\ \\ \leftarrow \text{ Critical Structure K.} \end{array} \tag{4.14}$$

The vector of upper bounds, CUB, no longer has the same meaning since we instead want to calculate the volume of tissue that is above the physician defined thresholds. For each k, let $T^{k_1}, T^{k_2}, \ldots, T^{k_{\Lambda_k}}$ be the thresholds for critical structure k. We let $\alpha_p^{k_\lambda}$ be a binary variable that indicates whether or not dose point p, which is in critical structure k, is below or above threshold T^{k_λ}. The percentage of critical structure k that is desired to be under threshold T^{k_λ} is $1 - \rho^{k_\lambda}$, or equivalently, ρ^{k_λ} is the percent of critical structure k that is allowed to violate threshold T^{k_λ}. Allowing M to be an upper bound on the

amount of radiation deposited in the anatomy, we have that any x satisfying the following constraints also satisfies the physician's dose-volume and tumor uniformity goals,

$$\left.\begin{array}{rcccl} TLB & \leq & A_T x & \leq & TUB \\ & & A_{C^k} x & \leq & T^{k_\lambda} e + \alpha^{k_\lambda} M, \quad \text{for each } k_\lambda \\ & & e^T \alpha^{k_\lambda} & \leq & \rho^{k_\lambda} |C^k|, \quad \text{for each } k_\lambda \\ & & A_N x & \leq & NUB \\ & & x & \geq & 0 \\ & & \alpha_p^{k_\lambda} & \in & \{0,1\} \qquad\qquad p \in C^k. \end{array}\right\} \tag{4.15}$$

The binary dose-volume constraints on the critical structures have replaced the previous linear constraints. In a similar fashion, we can add binary variables β_p, for $p \in T$, to measure the amount of target volume that is under dosed. If γ is the percentage of tumor that is allowed to be under its prescribed lower bound, we change the first set of inequalities in (4.15) to obtain,

$$\left.\begin{array}{rcll} A_T x & \leq & TUB, & \\ A_T x & \geq & TLB - \text{diag}(TLB)\beta, & \\ e^T \beta & \leq & \gamma |T|, & \\ A_T x & \leq & TUB, & \\ A_{C^k} x & \leq & T^{k_\lambda} e + \alpha^{k_\lambda} M, & \text{for each } k_\lambda, \\ e^T \alpha^{k_\lambda} & \leq & \rho^{k_\lambda} |C^k|, & \text{for each } k_\lambda, \\ A_N x & \leq & NUB, & \\ x & \geq & 0, & \\ \alpha_p^{k_\lambda} & \in & \{0,1\}, & p \in C^k, \\ \beta_p & \in & \{0,1\}, & p \in T. \end{array}\right\} \tag{4.16}$$

Of course we could add several threshold levels for the target volume, but the constraints in (4.16) describe how a physician prescribes dose in common commercial systems. Notice that a prescription now takes the form

$$\begin{aligned} &(TLB, TUB, NUB, T^{1_1}, \ldots, T^{1_{\Lambda_1}}, T^{2_1}, \ldots, T^{2_{\Lambda_2}}, \ldots, \\ &\quad T^{K_1}, \ldots, T^{2_{\Lambda_K}}, \gamma, \rho^{1_1}, \ldots, \rho^{1_{\Lambda_1}}, \rho^{2_1}, \ldots, \rho^{2_{\Lambda_2}}, \ldots, \rho^{K_1}, \ldots, \rho^{K_{\Lambda_K}}). \end{aligned}$$

For convenience, we let $\mathcal{P}$ be the collection of

$$u = (x, \alpha^{1_1}, \ldots, \alpha^{1_{\Lambda_1}}, \alpha^{2_1}, \ldots, \alpha^{2_{\Lambda_2}}, \ldots, \alpha^{K_1}, \ldots, \alpha^{K_{\Lambda_K}}, \beta)$$

that satisfy the constraints in (4.16).

From an optimization perspective, the difficulty of the problem has significantly increased from the earlier linear and quadratic models. Common objective functions are those that improve the under and over dosing. A linear

objective is

$$\min \left\{ w^1 \cdot e^T \beta + \sum_{k_\lambda} w^{k_\lambda} \cdot e^T \alpha^{k_\lambda} : u \in \mathcal{P} \right\}, \tag{4.17}$$

where the w's weight the importance of the respective under and over dosing. For more information on similar models, we point to Lee [42, 29–32].

A different modeling approach is to take a biological perspective [44, 53]. The concept behind these models is to use biological probabilities to find desirable treatments. In [53], Raphael presents a stochastic model that maximizes the probability of a successful treatment. The assumption is that tumorous cells are uniformly distributed throughout the target volume and that cells are randomly killed as they are irradiated. Allowing d_p to be the dose delivered at point p, we let $S(d_p)$ be the probability that any particular cell survives in the region represented by p under dose d_p. So, if there are $C(p)$ cancerous cells near dose point p, the expected number of survivors in this region under dose d_p is $\bar{N}(p) = C(p)S(d_p)$. If TV is the set of dose points within the target volume, then the expected number of surviving cancer cells is $\sum_{p \in TV} C(p)S(d_p)$. The actual number of survivors is the sum of many independent Bernoulli trials, whose distribution is assumed to be Poisson. This means that the probability of tumor control —i.e. when the expected number of survivors is zero, is

$$e^{-\sum_{p \in TV} C(p)S(d_p)}.$$

We want to maximize this probability, and the corresponding optimization problem is

$$\begin{aligned} \max \Big\{ & e^{-\sum_{p \in TV} C(p)S(d(p))} : \\ & A_T x = d,\ A_C x \leq CUB,\ A_N x \leq NUB,\ x \geq 0 \Big\}. \end{aligned} \tag{4.18}$$

This model has the favorable quality that it attempts to measure the overriding goal of treatment, that of killing the cancerous cells. However, this model simply introduces an exponential measure that increases the dose to the target volume as much as possible. As such, this model is similar to models (4.1) and (4.3), and it suffers from the same inadequacies.

Morrill develops another biologically based model in [44], where the goal is to maximize the probability of a complication free treatment. The idea behind this model is that different organs react to radiation differently, and that there are probabilistic ways to measure whether or not an organ will remain *complication free* [28, 39–41, 46]. To represent this model, we divide the rows of A_C as in (4.14). If $f(d^k)$ is the probability of critical structure k remaining complication free, where d^k is a vector of dose values in critical structure k,

the optimization model is

$$\max \left\{ \prod_{k=1}^{K} f(d^k) : TLB \leq A_T x \leq TUB, \right.$$
$$\left. A_{C_k} x = d^k,\ k = 1, 2, \ldots, K,\ A_N x \leq NUB \right\}. \tag{4.19}$$

As one can see, the scope of designing radiotherapy treatments intersects many areas of optimization. In addition to the models just discussed, several others have been suggested, and in particular, we refer to [14, 15, 33, 44, 57] for further discussions of nonlinear, nonquadratic models. All of the models in this subsection measure an aspect of treatment design, but they each fall short of mimicking the design process faced by a physician. Hence, it is crucial to continue the investigation of new models. While it may be impossible to include all of the patient specific information, the goal is to consistently improve the models so that they become flexible enough to work in a variety of situations.

4.4.3 New Directions

The models presented in Subsection 4.4.2 are concerned with the difficult task of optimally satisfying a physician's goals. In this section, we address some related treatment design questions that are beginning to benefit from optimization. The models in Subsection 4.4.2 have made significant inroads into the design of radiotherapy treatments, and the popular commercial systems use variants of these models in their design process. While this is a success for the field of OR, this is not the end of the story, and there are many, many clinically related problems that can benefit from optimization. This subsection focuses on the design questions that need to be made before a treatment is developed.

Several questions need to be answered before building any of the models in Subsection 4.4.2. These include deciding: 1) the distribution of the dose-points, 2) the number and location of the isocenters, and 3) the number and location of the beams. Each of these decisions is currently made by trial-and-error, and once these decisions are made, the previous models optimize the treatment. However, the fact that a model's representation depends on these decisions means that a treatment's quality depends on a physician's experience. Replacing the trial-and-error process with an optimization technique that provides consistently favorable answers to these questions is an important and relatively untapped area of research.

To formally study how these three questions affect a treatment plan, we let Opt be a function with the following arguments: $\mathbb{B}$ is a collection of isocenters and beams, $\mathbb{D}$ is a vector of dose points, and $\mathbb{P}$ is a prescription. Opt$(\mathbb{B}, \mathbb{D}, \mathbb{P})$ returns the optimal value of the optimization routine, denoted by optval, and

an optimal treatment, x. The argument $\mathbb{B}$ has the following form,

$$\mathbb{B} = \begin{pmatrix} ((x_1, y_1, z_2), (\theta_{(1,1)}, \psi_{(1,1)}, \rho_{(1,1)})), \ldots, (\theta_{(1,B_1)}, \psi_{(1,B_1)}, \rho_{(1,B_1)}) \\ ((x_2, y_2, z_2), (\theta_{(2,2)}, \psi_{(2,2)}, \rho_{(2,2)})), \ldots, (\theta_{(2,B_2)}, \psi_{(2,B_2)}, \rho_{(2,B_2)}) \\ \vdots \\ ((x_I, y_I, z_I), (\theta_{(I,I)}, \psi_{(I,I)}, \rho_{(I,I)})), \ldots, (\theta_{(I,B_I)}, \psi_{(I,B_I)}, \rho_{(I,B_I)}) \end{pmatrix},$$

where (x_i, y_i, z_i) is an isocenter and $\{(\theta_{(i,j)}, \psi_{(i,j)}, \rho_{(i,j)}) : j = 1, 2, \ldots, B_i\}$ is the collection of spherical coordinates for the beams around isocenter i. Assuming that there are m dose points in the anatomy (so the dose matrix has m rows), the vector of dose points looks like

$$\mathbb{D} = ((x_1, y_1, z_1), (x_2, y_2, z_2), \ldots, (x_m, y_m, z_m)) .$$

The form of the prescription $\mathbb{P}$ depends on the optimization problem. The functional dependence a treatment has on the isocenters, the beam positions, the dose points, and the prescription is represented by $\text{Opt}(\mathbb{B}, \mathbb{D}, \mathbb{P}) = (\text{optval}, x)$. The form of Opt is defined by the optimization model, and most of the research has been directed toward having a useful representation of Opt. However, the optimization models and subsequent treatments depend on $\mathbb{B}$, $\mathbb{D}$, and $\mathbb{P}$, and this dependence is not clearly understood.

The question of deciding how the dose points are distributed has received some attention (see for example [45]) and is often discussed as authors describe their implementation. However, a model that 'optimally' positions dose points within an anatomy has not been considered, and deciding what optimal means is open for debate. Most researchers use either a simple grid or the more complicated techniques of skeletization or octree, but none of these processes are supported by rigorous mathematics.

The question of deciding the number and position of the isocenter(s) is the least investigated of the three problems. Most treatment systems place the isocenter at the center-of-mass of a user defined volume, typically the target volume. This placement is intuitive, but there is no reason to believe that this is the 'best' isocenter. In fact, the clinics with which the authors are familiar have developed techniques for special geometries that place the isocenter at different positions. In addition to the location question, there has been no mathematical work on deciding the number of isocenters. Investigating these questions promises to be fruitful research.

The question of pre-selecting a candidate set of beams has witnessed some work [4, 12, 17, 51, 52, 56, 62]. The breadth of the research exhibits that the problem is complicated enough so that there is no clearly defined manner to address the problem. Indeed, the first author of this tutorial spent several years working on this problem to no avail. Much of the current research is structural, meaning that the set of beams is constructed by adding beams under

a decision rule. The other work selects a candidate set by solving a large, mixed-integer problem. For the sake of brevity, we omit a detailed discussion of these techniques and instead investigate a promising new process.

We suggest that rather than constructing a candidate set of beams, we instead begin with an unusually large collection of beams and then prune them to the desired number. The premise of the idea is that a treatment based on many beams indicates which beams are most useful. Ehrgott [12] uses this idea to select a candidate set of beams from a larger collection by solving a large, mixed-integer problem. Our technique is different and is based on the data compression technique of vector quantization.

A *quantizer* is a mapping that has a continuous, random variable as its argument and maps into a discrete set, called the code book. Each quantizer is the composition of an *encoder* and a *decoder*. If u is a random variable with possible values in V, an encoder takes the form $f : V \to \{1, 2, \ldots, n\}$, and a decoder looks like $g : \{1, 2, \ldots, n\} \to V$. The quantizer defined by f and g is $Q(u) = g(f(u))$. The encoder maps the realizations of u into the index set $\{1, 2, \ldots, n\}$ and partitions V into n sets. The decoder completes the quantization by assigning the possible realizations of u to a discrete subset of V, and the elements in this subset are called codewords. As an example, let u be uniformly distributed on $[0, 1]$. The process of rounding is a quantizer, and in this case we have that

$$f : [0,1] \to \{1,2\} : u \mapsto \begin{cases} 1, & 0 \le u < 0.5 \\ 2, & 0.5 \le u \le 1 \end{cases}$$

$$g : \{1,2\} \to [0,1] : u \mapsto \begin{cases} 0, & u = 1 \\ 1, & u = 2. \end{cases}$$

In this example, the interval $[0, 1]$ is partitioned by $\{[0, 0.5), [0.5, 1]\}$, with the first interval mapping to the codeword 0 and the second interval mapping to the codeword 1.

In the previous example, the interval $[0, 1]$ is quantized to the discrete set $\{0, 1\}$. The quantization error for any realization of u is $d(u, Q(u))$, where d is a metric on V (the most common error is $\|u - Q(u)\|_2$). The quantizer's distortion is the average error,

$$D_Q = E\, d(u, Q(u)) = \int_V P(u) \cdot d(u, Q(u)) du,$$

where $E\, d(u, Q(u))$ is the expected value of d and $P(u)$ is the probability distribution of u. A quantizer is *uniform* if the partitioning sets have the same measure and is *regular* if it satisfies the nearest neighbor condition —i.e.

$$d(u, Q(u)) = \min\{d(u, y) : y \text{ is a codeword}\}.$$

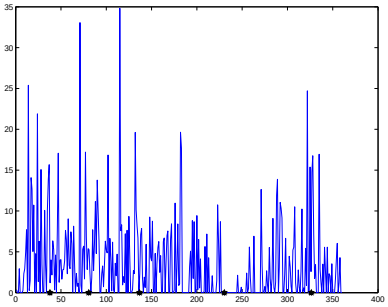

Figure 4.13. The dose profile of a treatment with one isocenter and 360 beams.

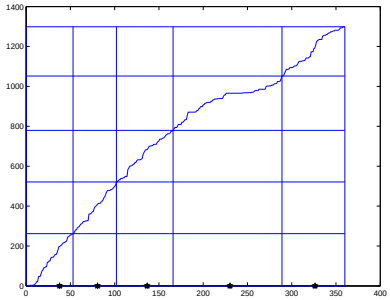

Figure 4.14. The cumulative dose distribution.

The quantizer design problem is to build a quantizer that minimizes the distortion, and the following necessary conditions [18] guide the design process:

- For any codebook, the partition must satisfy the nearest neighbor condition, or equivalently, the quantizer must be regular.
- For any Partition, the codevectors must be the centers of mass of the probability density function.

Since these are only necessary conditions, a quantizer satisfying these conditions may not minimize distortion. However, there are cases where these are necessary and sufficient, such as when the logarithm of the probability density function is convex [18].

We address the 2-dimensional beam selection problem by designing a quantizer from $[0, 2\pi)$ into $[0, 2\pi)$. The probability density function is patient specific and is calculated by approximating the continuous planning problem. For each isocenter, assume that there is a large number of beams, something on the order of one every degree. Solve $\text{Opt}(\mathbb{B}, \mathbb{D}, \mathbb{P}) = (\text{optval}, x)$, and from the treatment x calculate the amount of radiation delivered along each beam —i.e. aggregate each beam's sub-beams to attain the total radiation for the beam. As an example, the dose profile for the problem in Figure 4.11 is in Figure 4.13, where Opt is defined by model (4.7), there is a single isocenter in the middle of the 64×64 image, there are 360 beams, and dose points are centered within each pixel. The normalized dose profile is the probability density function, and the idea is that this function estimates the likelihood of using an angle. This assumption is reasonable because beams that deliver large amounts of radiation to the tumor are often the ones that intersect the tumor but not the critical structures.

Figure 4.14 is the cumulative dose distribution, and this function is used to define the encoder. Let h be the function that calculates the cumulative dose distribution from a treatment plan x, and let the range of h be the interval $[0, \gamma]$. So, $h_{(\mathsf{Opt}(\mathbb{B},\mathbb{D},\mathbb{P}))}(u)$ is a bijective mapping from $[0, 2\pi)$ onto $[0, \gamma)$ that depends on the isocenters, the beams, the dose points, the prescription, and the optimization model. If the physician desires an N beam plan, the encoder is defined by

$$f : [0, 2\pi) \to \{1, 2, \ldots, N\} :$$
$$h_{(\mathsf{Opt}(\mathbb{B},\mathbb{D},\mathbb{P}))}(u) \to i,\ u \in [(i-1)\gamma/N, i\gamma/N).$$

The encoder partitions the interval $[0, 2\pi)$ into code regions, and because h is monotonic, we are guaranteed that the quantizer is regular. For the example in Figure 4.14, the partition of $[0, 2\pi)$ for a 5 beam treatment is depicted along the horizontal axis (which is in degrees instead of radians).

The decoder assigns a codeword to each partition, and from the necessary conditions of optimality we have that the codewords of an optimal quantizer must be the centers of mass of the normalized dose profile. These codewords are highlighted on the horizontal axes of Figures 4.13 and 4.14 and are beams 32, 78, 140, 232, and 327. The quantizer for this example has the final form,

$$[0, 52) \mapsto 32,\ \ [52, 101) \mapsto 78,$$
$$[101, 164) \mapsto 140,\ \ [164, 290) \mapsto 232,\ \ [290, 360) \mapsto 327.$$

Dose-volume histograms of two pruned plans are in Figures 4.15 and 4.16. These images indicate that the critical structures fare better as more angles are used (compare to Figure 4.12).

Our initial investigations into selecting beams with vector quantization are promising, but there are many questions. The partition depends on the cumulative dose distribution, and this function depends on where accumulation begins. We currently start accumulating dose at angle 0, but this choice is arbitrary and not substantiated. A more serious challenge is to define a clinically relevant error measure that allows us to analyze distortion. We point out that any meaningful error measure relies on Opt and moreover that the dose matrices are different for the quantized and unquantized beams. Lastly, for this technique to have clinical meaning, we need to extend it to 3-dimensions. Any advancement in these areas will lead to immediate improvement in patient care.

We conclude this section by mentioning that the quality of a treatment not only depends on the model, which we have discussed in detail, but also on the algorithm, which we have ignored. As an example, many of the nonlinear models, including the least squares problems, are often solved by simulated annealing. Because this is a stochastic algorithm, it is possible to design different treatments with the same model. An interesting numerical paper for the

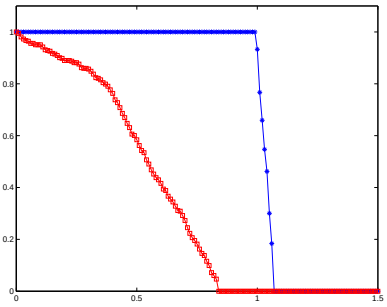

Figure 4.15. The dose-volume histogram for a treatment pruned from 360 to 5 angles for the problem in Figure 4.11.

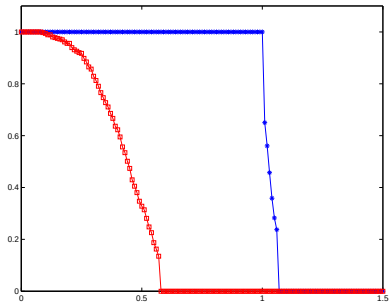

Figure 4.16. The dose-volume histogram for a treatment pruned from 360 to 9 angles for the problem in Figure 4.11.

nonlinear models would be to solve the same model with several algorithms to find if some of them naturally design better treatments. The linear models are likely to have multiple optimal solutions, and in this case, the solutions from a simplex algorithm and an interior-point algorithm are different. Both solutions have favorable and unfavorable characteristics. The basic solutions have the favorable property that the number of sub-beams is restricted by the number of constraints, and if constraints are aggregated, we can control the number of sub-beams [36]. However, the simplex solutions have the unfavorable quality that they guarantee that some of the prescribed bounds are attained [22]. The interior-point solutions have the reverse qualities, as they favorably ensure the prescribed bounds are strictly satisfied and they unfavorably use as many sub-beams as possible [23–25]. The fact that interior-point algorithms inherently produce treatments with many beams makes them well suited to tomotherapy.

4.5 The Gamma Knife

The Gamma Knife treatment system was specifically designed to treat brain, or intracranial, lesions. The first Gamma Knife was built in 1968 and is *radiosurgical* in its intent (the term radiosurgery was first used in [35]). The difference between radiosurgical and radiotherapy approaches has been previously described. The high dose delivered during a radiosurgery makes accuracy in both treatment planning and delivery crucial to a treatment's success.

The Gamma Knife uses 201 radioactive cobalt-60 sources to generate its treatment pencil beams instead of a linear accelerator. These sources are spherically distributed around the patient, and their width is controlled by a series of collimators. These collimators are different than those used in IMRT, with the

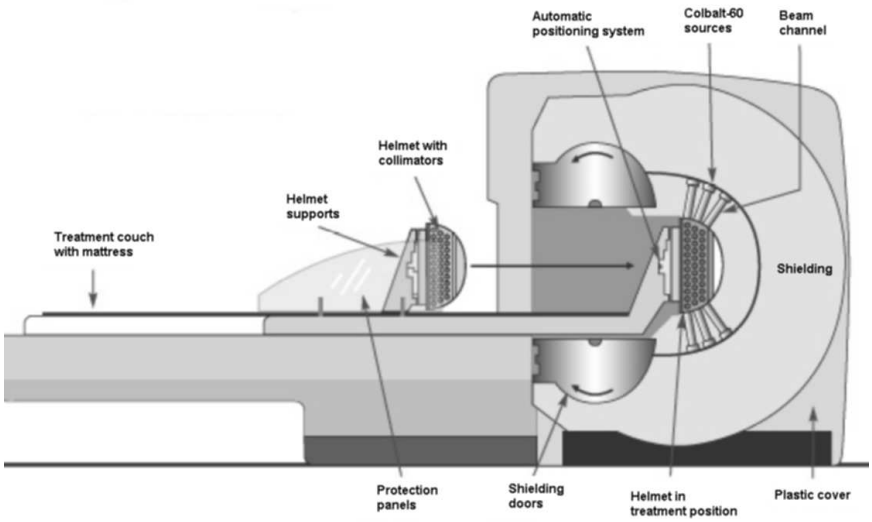

Figure 4.17. A Gamma Knife treatment machine.

Gamma Knife collimators being located in a helmet that fits the patient's head. Each helmet consists of 201 cylindrical holes of either 4, 8, 14, or 18mm. The 201 radiation beams thus produced intersect at a common focal point and form a spherically shaped high dose region whose diameter is roughly equal to the collimator size. These spheres are called *shots* and have the favorable property that radiation dose outside these regions falls off very quickly (i.e. a high dose gradient). It is this fact that makes the Gamma Knife well suited to deliver radiosurgeries, in that very high doses of radiation may be delivered to a target which is immediately adjacent to a critical structure, with relatively little dose delivered to the structure.

4.5.1 Clinically Relevant Gamma Knife Treatments

The primary clinical restriction on Gamma Knife treatments is that the number of shots must be controlled. Between each shot the patient is removed from the treatment area, re-aligned, possibly re-fitted with a different collimator helmet, and returned to the treatment area. This is a time consuming process, and most treatment facilities attempt to treat a patient in under 10 to 15 shots. We mention that it is possible to 'plug' some of the 201 collimator holes, which can produce an ellipsoidally shaped distribution of dose. While this is clinically possible, this is rarely undertaken because of time restrictions and the possibility of errors related to the manual process. In this tutorial we do not

consider plugged collimators, and we therefore assume that the dose is delivered in spherical packets.

4.5.2 Optimization Models

From a modeling perspective, the Gamma Knife's sub-beams are different than the sub-beams of IMRT. The difference is that in IMRT the amount of radiation delivered along each sub-beam is controlled by a multileaf collimator, but in the Gamma Knife each sub-beam delivers the same amount of radiation. So, Gamma Knife treatments do not depend on the same decision variables as IMRT, and consequently, the structure of the dose matrix for the Gamma Knife is different. The basic dose model discussed in Section 4.3 is still appropriate, but we need to alter the indices of $a_{(p,a,s,i,e)}$, which recall is the rate at which radiation accumulates at dose point p from sub-beam (a, s) when the gantry is focused on the i^{th} isocenter and energy e is used. The Gamma Knife delivers dose in spherical packets called *shots*, which are defined by their centers and radii. A shot's center is the point at which the sources are focused and is the same as the isocenter. The radius of a shot is controlled by the collimators that are placed on each source. As mentioned in the previous subsection, the same collimator size is used for every source per shot, and hence, a shot is defined by its isocenter i and its collimator c. Moreover, unlike the linear accelerators used in IMRT, the cobalt sources of the Gamma Knife produce a single energy, and hence, there is no functional dependence on the energy e. We alter the indices of $a_{(p,a,s,i,e)}$ by letting $a_{(p,c,i)}$ be the rate at which radiation accumulates at dose point p from shot (c, i). These values form the dose matrix A, where the rows are indexed by p and the columns by (c, i).

Since the Gamma Knife delivers dose in spherical shots, the geometry of treatment design is different than that of IMRT. The basic premise of irradiating cancerous tissue without harming surrounding structures remains, but instead of placing beams of radiation so that they avoid critical areas, we rather attempt to cover (or fill) the target volume with spheres. Since the cumulative dose is additive, regions where shots overlap are significantly over irradiated. These hot spots do not necessarily degrade the treatment because of its radiosurgical intention. However, it is generally believed that the best treatments are those that sufficiently irradiate the target and at the same time reduce the number and size of hot spots. This means that favorable Gamma Knife treatments fill the target with spheres of radiation so that 1) shots intersections are small and 2) shots do not intersect non-target tissue. Outside the fact that designing Gamma Knife treatments is clinically important, the problem is mathematically interesting because of its relationship to the sphere packing problem. While there is a wealth of mathematical literature on sphere packing, this connection has not been exploited, and this promises to be a fruitful research direction.

The problem of designing Gamma Knife treatments received significant exposure when it was one of the modeling problems for the 2003 COMAP competition in mathematical modeling [10], and there are many optimization models that aid in the design of treatments [15, 34, 38, 58–60, 65, 66]. We focus on the recent models by Ferris, Lim, and Shepard [14] (winner of the 2002 Pierskalla award) and Cheek, Holder, Fuss, and Salter [6]. Both of these models use dose-volume constraints and segment the anatomy into target and non-target, making A_N vacuous. The model proposed in [14] is

$$\begin{aligned}\min\{e^T u_T : d_T = A_T x, d_C = A_C x, \theta \le u_T + d_T,\\ 0 \le x \le sM, \rho(e^T d_T + e^T d_C) \le e^T d_T, e^T s \le n,\\ s_i \in \{0,1\}, 0 \le u_T, 0 \le u_C\}.\end{aligned} \tag{5.1}$$

The dose to the target and non-target tissues is contained in the vectors d_T and d_C, and u_T measures how much the target volume is under the goal dose θ. The objective is to minimize the total amount the target volume is under irradiated. The binary variables s_i indicate whether or not a shot is used or not, and the constraint $e^T s \le n$ limits the treatment to n shots (M is an arbitrarily large number). The parameter ρ is a measure of desired conformality, and the constraint $\rho(e^T d_T + e^T d_C) \le e^T d_T$ ensures that the target dose is at least ρ of the total dose. If ρ is 1, then we are attempting to design a treatment in which the entire dose is within the target.

Model (5.1) is a binary, linear optimization problem. The authors of [14] recognize that the size of the problem makes it impossible for modern optimization routines to solve the problem to optimality (small Gamma Knife treatments often require more than 500 Gigabytes of data storage). The authors of [14] replace the binary variables with a $\tan^{-1}$ constraint that transforms the problem into a continuous, nonlinear program. Specifically, they replace the constraints

$$\left.\begin{array}{c} 0 \le x \le sM, \\ e^T s \le n, \\ s \in \{0,1\} \end{array}\right\} \quad \text{with} \quad \left\{ \sum_{(c,i)} \tan^{-1}(\alpha x_{(c,i)}) \le n, \right.$$

where larger α values more accurately resemble the binary constraints. Together with other reductions and assumptions, this permits the authors to use CONOPT [11] to design clinically acceptable treatments.

Model (5.1) is similar to the IMRT models that use dose-volume constraints because it's objective function measures dose and ρ describes the volume of non-target tissue that we are allowed to irradiate. While this model successfully designed clinically relevant treatments, physicians often judge Gamma Knife treatments with a conformality index. These indices are scoring functions that quantify a treatment's quality by measuring how closely the irradiated tissue

resembles the target volume. So, in addition to the dose-volume histograms and the 2-dimensional isodose lines, Gamma Knife treatments are often judged by a single number. Collapsing large amounts of information into a single score is not always appropriate, but the radiosurgical intent of a Gamma Knife treatment lends itself well to such a measure —i.e. the primary goal is to destroy the target with an extremely high level of radiation and essentially deliver no radiation to the remaining anatomy. This means that conforming the high-dose region to the target is crucial, and hence, judging treatments on their conformity is appropriate.

Several conformality indices are suggested in the literature (see [37] for a review). Let D be the suggested target dose (meaning the physician desires $A_T x \geq De$) and define

$$\begin{aligned} TV &= \{p : \text{ dose point } p \text{ is in the target volume } \} \text{ and} \\ IV_T &= \{p : \text{ the dose at point } p \text{ is at least } T \cdot D\}, \end{aligned}$$

where T is between 0 and 1. If we assume that each dose point represents a volume V, the target volume is $V \cdot |TV|$ and the T^{th} isodose line encloses a volume of $V \cdot |IV_T|$. The standard indices are expressed in terms of the %100 isodose line and are

$$\begin{aligned} PITV &= |IV_1|/|TV|, \\ CI &= |TV \cap IV_1|/|IV_1|, \text{ and} \\ IPCI &= (|TV \cap IV_1|/|TV|) \cdot (|TV \cap IV_1|/|IV_1|). \end{aligned}$$

The last index is called Ian Paddick's conformality index [47] and is the product of the *over treatment ratio* and the *under treatment ratio*. These are defined for any isodose value by

$$OTR_T = |TV \cap IV_T|/|IV_T| \quad \text{and} \quad UTR_T = |TV \cap IV_T|/|TV|.$$

The over treatment ratio is at most 1 if the target volume contains the T^{th} isodose volume. Otherwise, OTR_T is between 0 and 1, and $1 - OTR_T$ is the volume of non-target tissue receiving a dose of at least $T \cdot D$. Similarly, the under treatment ratio is 1 if the target volume is contained in the T^{th} isodose line, and $1 - UTR_T$ is the percentage of target volume receiving less than $T \cdot D$Gy. The over and under treatment ratios are 1 only if the T^{th} isodose volume matches the target volume, and the conformality objective is to design plans that have $OTR_T = UTR_T = 1$. For any T, the Ian Paddick conformality index is $IPCI_T = UTR_T \cdot OTR_T$.

The authors of [6] suggest a model whose objective function is based on Ian Paddick's conformality index. Assume that there are I isodose lines and that

Θ_i is the i^{th} column of the matrix Θ. The model in [6] is

$$\min \Bigg\{ \sum_{i=1}^{I} w_i(1 - e_{TV}^T \Theta_i / e^T \Theta_i) + u_i(1 - (V/K) e_{TV}^T \Theta_i) : \\ Ax = d,\ \text{diag}(d) ee^T \leq ee^T HD + M\Theta,\ 0 \leq x \leq M\beta,\ e^T\beta \leq L, \\ \beta_i \in \{0,1\},\ \Theta_{(p,i)} \in \{0,1\} \Bigg\}. \tag{5.2}$$

The parameters V and K are the voxel and target volumes, and e_{TV} is the binary vector with ones where an index corresponds to a targeted dose point. The number of shots is measured by the binary vector β and is restricted by L (M is an arbitrarily large value). The vector d is the delivered dose, and $\text{diag}(d)$ is the diagonal matrix formed by d. The diagonal matrix H contains the isodose values that we are using, and $ee^T HD$ is a matrix with each column being $T_i De$. The matrix constraint $\text{diag}(d) ee^T \leq ee^T H + M\Theta$ guarantees that if the dose at point p is above the isodose value $T_i D$, then $\Theta_{(p,i)}$ is 1. From this we see that $OTR_{T_i} = e_{TV}^T \Theta_i / e^T \Theta_i$ and that $UTR_{T_i} = (V/K) e_{TV}^T \Theta_i$. The weights w_i and u_i express the importance of having the T_i^{th} isodose line conform to the target volume. We point out that the objective function of Model (5.2) is not *IPCI* but is rather a weighted sum of the over and under treatment ratios.

Neither model (5.1) or (5.2) penalizes over irradiating portions of the target, and controlling hot spots complicates the problem. This follows because measuring hot spots is often accomplished by adding a variable for each dose point that increases as the delivered dose grows beyond an acceptable amount. The problem is not with the fact that there are an increased number of variables, but rather that physicians are not concerned with high doses over small regions. A more appropriate technique is to partition the dose points into subsets, say H_r, and then aggregate dose over these regions to control hot spots. If a hot spot is defined by the average dose of a region exceeding τ, then adding the constraints,

$$\sum_{p \in H_r} d_p \leq |H_r|(\tau + q) \quad \text{and} \quad q \geq 0$$

to model (5.1) or (5.2) enables us to calculate the largest hot spot. Such a tactic was used in [6] for model (5.2), where each H_r contained 4 dose points in a contiguous, rectangular pattern. The objective function was altered to

$$\sum_{i=1}^{I} \left(w_i(1 - e_{TV}^T \Theta_i / e^T \Theta_i) + u_i(1 - (V/K) e_{TV}^T \Theta_i) \right) + 0.5q.$$

Model (5.2) is easily transformed into a binary, quadratic problem, but again, it's size makes standard optimization routines impractical. As an al-

ternative, fast simulated annealing is used in [6], where the research goal was to explore how treatment quality depends on the number of shots —i.e. how the standard indices depend on L. Treatments designed with this model are shown in Figure 4.18, and the CI and $IPCI$ values for different choices of L are in Table 4.1. Figure 4.19 shows how the dose-volume histograms improve as more shots are allowed.

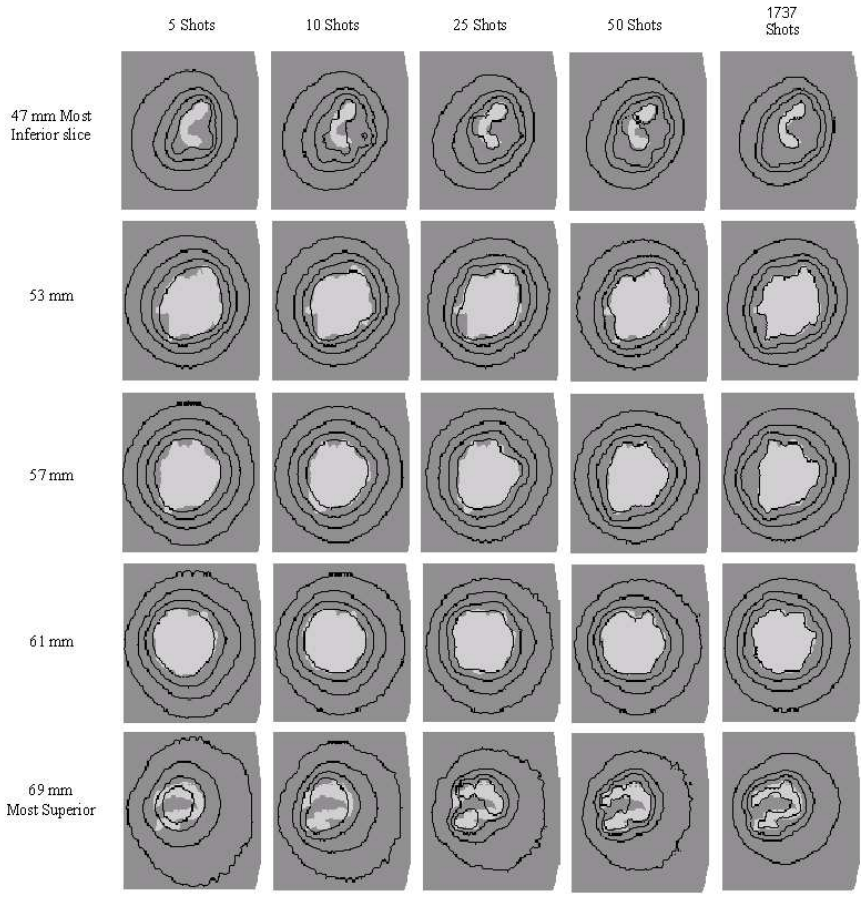

Figure 4.18. Isodose curves from treatments designed with Model (5.2). The value of L is listed across the top of each treatment, and the millimeter value on the left indicates the depth of the image.

4.6 Treatment Delivery

The previous sections focused on treatment design, and while these problems are interesting and important, the optimization community is now poised

	5 Shots	10 Shots	25 Shots	50 Shots	Unlimited	Ideal
PITV	0.934	0.996	0.992	0.990	0.999	1
CI	0.846	0.897	0.925	0.954	0.997	1
IPCI	0.767	0.808	0.863	0.919	0.995	1

Table 4.1. How the $PITV$, CI and $IPCI$ indices react as the number of possible shots increases.

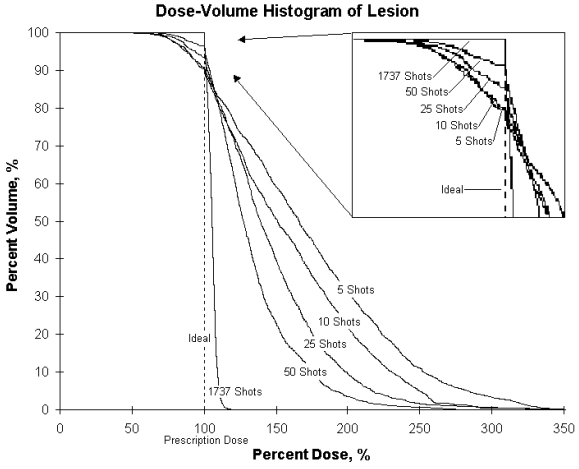

Figure 4.19. The dose-volume histograms for treatment plans with differing numbers of shots.

to significantly improve patient care with respect to the design process. So, even though it is important to continue the study of treatment design, there are related clinical questions where beginning researchers can make substantial contributions. In this section we focus on two treatment delivery questions that are beginning to receive attention.

As mentioned earlier, the difference between radiotherapy and radiosurgery is that radiotherapy is delivered in fractional units over several days. Current practice is to divide the total dose into N equal parts and deliver the overall treatment in uniform, daily treatments. The value of N is based on studies that indicate how healthy tissue regenerates after being irradiated, and the overall treatment is fractionated to make sure that healthy tissue survives. Dividing the total dose was particularly important when technology was not capable of conforming the high-dose region to the target, as this meant that surrounding tissues were being irradiated along with the tumor. However, modern technology permits us to magnify the difference between the dose delivered to the target and the dose delivered to surrounding tissues. The support for a uniform

division does not make sense with our improved technology, and Ferris and Voelker [16] have investigated different approaches.

Suppose we want to divide a treatment into N smaller treatments. If d^k is the cumulative dose after k treatments, the problem is to decide how much dose to deliver in subsequent periods. This leads to a discrete-time dynamic system, and if we let u^k be the dose added in period k and w^k be the random error in delivering u^k, then the system is

$$d_{k+1} = d_k + u_k(1 + w_k).$$

The random error is real because the planned dose often deviates from the delivered dose since patient alignment varies from day-to-day. The decision variables u_k must be nonnegative since it is impossible to remove dose after it is delivered. The optimization model used in [16] is

$$\begin{aligned} \min\{&E(\|w^T(d_N - D)\|_1) : \\ &d_{k+1} = d_k + u_k(1 + w_k), u_k \geq 0, w_k \in W\}, \end{aligned} \tag{6.1}$$

where D is the total dose to deliver, W is the range of the random variable w, and E is the expected value. This model can be approached from many perspectives, and the authors of [16] consider stochastic linear programming, dynamic programming, and neuro-dynamic programming. They suggest that a neuro-dynamic approach is appropriate and experiment with a 1-dimensional problem. Even at this low dimensionality the problem is challenging. They conclude that undertaking such calculations to guide clinical practice is not realistic, but they do use their 1-dimensional model to suggest 'rules-of-thumb.'

Model (6.1) requires a fixed number of divisions, and hence, this problem only address the uniformity of current delivery practices. An interesting question that is not addressed is to decide the number of treatments. If we can solve this problem independent of deciding how the dose is to be delivered, then we can calculate N before solving model (6.1). However, we suggest that it is best to simultaneously make these decisions.

Another delivery question that is currently receiving attention is that of leaf sequencing [3, 13, 26, 27, 49, 50]. This is an important problem, as complicated treatments are possible if we can more efficiently deliver dose. An average treatment lasts from 15 to 30 minutes, and if the leaves of the collimator are adjusted so that the desired dose profile is achieved quickly, then more beams are possible. This translates directly to better patient care because treatment quality improves as the number of beams increases (the same is true for the Gamma Knife as demonstrated in Section 4.5.2). We review the model in [3], which is representative, and encourage interested readers to see the other works and their bibliographies.

Suppose we have solved (in 3-dimensions) Opt$(\mathbb{B}, \mathbb{D}, \mathbb{P})$ and that an optimal treatment shows that a patient should be irradiated with the following exposure

pattern,

$$\mathcal{I} = \begin{bmatrix} 0 & 0 & 2 & 2 & 2 & 0 \\ 0 & 1 & 1 & 3 & 1 & 0 \\ 0 & 0 & 2 & 2 & 1 & 0 \\ 1 & 2 & 2 & 2 & 1 & 0 \\ 0 & 1 & 2 & 3 & 2 & 1 \\ 0 & 1 & 2 & 2 & 2 & 2 \end{bmatrix}. \tag{6.2}$$

The dose profile $\mathcal{I}$ contains our desired exposure times. Each element of $\mathcal{I}$ represents a rectangular region of the Beam's Eye View —i.e. the view of the patient as one was looks through the gantry. The collimator for this example has 12 leaves (modern collimators have many more), one on the right and left of each row. These leaves can move across the row to shield the patient.

The optimization model in [3] minimizes exposure time by controlling the leaf positions. The treatment process is assumed to follow the pattern: the leaves are positioned, the patient is exposed, the leaves are re-positioned, the patient is exposed, etc..., with the process terminating when the dose profile is attained. For each row i and time step t we let

$$l_{ijt} = \begin{cases} 1, & \text{if the left leaf in row } i \text{ is positioned in column } j \text{ at time } t \\ 0, & \text{otherwise} \end{cases}$$

$$r_{ijt} = \begin{cases} 1, & \text{if the right leaf in row } i \text{ is positioned in column } j \text{ at time } t. \\ 0, & \text{otherwise.} \end{cases}$$

The nonlinear, binary model studied is

$$\min \Bigg\{ \sum_t \alpha_t : \sum_j l_{ijt} = 1, \forall i, t;\ \sum_j r_{ijt} = 1, \forall i, t,$$
$$y_{ijt} = \sum_{k=0}^{j-1} l_{ikt} - \sum_{k=1}^{j} r_{ikt}, \forall t;\ \sum_{k=0}^{j} l_{ikt} \geq \sum_{k=1}^{j} r_{ikt}, \forall t,$$
$$\sum_t \alpha_t y_{ijt} = \mathcal{I}_{ij}, \forall i, j;\ l_{ijt}, r_{ijt}, y_{ijt} \in \{0, 1\}, \forall i, j, t;$$
$$\alpha_t \geq 0, \forall t \Bigg\}. \tag{6.3}$$

Model (6.3) is interpreted as finding a *shape matrix* at each time t. A shape matrix is a binary matrix such that the 1s in every row are contiguous (a row may be void of 1s). Each 1 indicates an unblocked region of the beam, and each shape matrix represents a positioning of the leaves. The y variables in

model (6.3) form an optimal collection of shape matrices. For example,

$$\begin{aligned}\mathcal{I} = \begin{bmatrix} 2 & 3 \\ 4 & 2 \end{bmatrix} &= 2\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + 4\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} + 3\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \\ &= 2\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + 1\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} + 1\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}.\end{aligned}$$

The first shape matrix in the first decomposition has $y_{111} = y_{221} = 1$ and $y_{121} = y_{211} = 0$. The total exposure time for the first decomposition is $2+4+3 = 9$ and for the second decomposition the exposure time is $2 + 1 + 1 = 4$. So, the second leaf sequence is preferred.

The authors of [3] show that model (6.3) can be re-stated as a network flow problem, and they further develop a polynomial time algorithm to solve the problem. This provides the following theorem (see [13] for related results).

THEOREM 4.3 (BOLAND, HAMACHER, AND LENZEN [3]) *Model (6.3) is solvable in polynomial time.*

We close this section by suggesting a delivery problem that is not addressed in the literature. As Figure 4.19 shows, Gamma Knife treatments improve as the number of shots increases. We anticipate that new technology will permit automated patient movement, which will allow the delivery of treatments with numerous shots. How to move a patient so that shots are delivered as efficiently as possible is related to the traveling salesperson problem, and investigations into this relationship are promising. In the distant future, we anticipate that patients will movement continuously within the treatment machine. This means shots will move continuously through the patient, and finding an optimal path is a control theory problem.

4.7 Conclusion

The goal of this tutorial was to familiarize interested researchers with the exciting work in radiation oncology, and the authors hope that readers have found inspiration and direction from this tutorial. We welcome inquiry and will be happy to answer questions. We conclude with a call to the OR community to vigorously investigate how optimization can aid medical procedures. The management side of health care has long benefited from optimization techniques, but the clinical counterpart has enjoyed much less attention. The focus of this work has been radiation oncology, but there are many procedures where standard optimization routines and sound modeling can make substantial improvements in patient care. This research is mathematically aesthetic, challenging, and intrinsically worthwhile because it aids mankind.

Acknowledgments

The authors thank Roberto Hasfura for his careful editing and support.

References

[1] G. K. Bahr, J. G. Kereiakes, H. Horwitz, R. Finney, J. Galvin, and K. Goode. The method of linear programming applied to radiation treatment planning. *Radiology*, 91:686–693, 1968.

[2] F. Bartolozzi et al. Operational research techniques in medical treatment and diagnosis. a review. *European Journal of Operations Research*, 121(3):435–466, 2000.

[3] N. Boland, H. Hamacher, and F. Lenzen. Minimizing beam-on time in cancer radiation treatment using multileaf collimators. Technical Report Report Wirtschaftsmathematik, University Kaiserslautern, Mathematics, 2002.

[4] T. Bortfeld and W. Schlegel. Optimization of beam orientations in radiation therapy: Some theoretical considerations. *Physics in Medicine and Biology*, 38:291–304, 1993.

[5] Y. Censor, M. Altschuler, and W. Powlis. A computational solution of the inverse problem in radiation-therapy treatment planning. *Applied Mathematics and Computation*, 25:57–87, 188.

[6] D. Cheek, A. Holder, M. Fuss, and B. Salter. The relationship between the number of shots and the quality of gamma knife radiosurgeries. Technical Report 84, Trinity University Mathematics, San Antonio, TX, 2004.

[7] J. Chinneck. An effective polynomial-time heuristic for the minimum-cardinality iis set-covering problem. *Annals of Mathematics and Artificial Intelligence*, 17:127–144, 1995.

[8] J. Chinneck. Finding a useful subset of constraints for analysis in an infeasible linear program. *INFORMS Journal on Computing*, 9(2):164–174, 1997.

[9] J. Chinneck and H. Greenberg. Intelligent mathematical programming software: Past, present, and future. *Canadian Operations Research Society Bulletin*, 33(2):14–28, 1999.

[10] Consortium for Mathematics and Its Applications (COMAP), `www.comap.com`. *Gamma Knife Treatment Planning, Problem B*.

[11] A. Drud. CONOPT: A GRG code for large sparse dynamic nonlinear optimization problems. *Mathematical Programming*, 31:153–191, 1985.

[12] M. Ehrgott and J. Johnston. Optimisation of beam direction in intensity modulated radiation therapy planning. *OR Spectrum*, 25(2):251–264, 2003.

[13] K. Engel. A new algorithm for optimal multileaf collimator field segmentation. Technical report, Operations Research & Radiation Oncology Web Site, w.trinity.edu/aholder/HealthApp/oncology/, 2003.

[14] M. Ferris, J. Lim, and D. Shepard. An optimization approach for the radiosurgery treatment planning. *SIAM Journal on Optimization*, 13(3):921–937, 2003.

[15] M. Ferris, J. Lim, and D. Shepard. Radiosurgery optimization via nonlinear programming. *Annals of Operations Research*, 119:247–260, 2003.

[16] M. Ferris and M. Voelker. Neuro-dynamic programming for radiation treatment planning. Technical Report Numerical Analysis Group Research Report NA-02/06, Oxford University Computing Laboratory, 2002.

[17] S. Gaede, E. Wong, and H. Rasmussen. An algorithm for systematic selection of beam directions for imrt. *Medical Physics*, 31(2):376–388, 2004.

[18] A. Gersho and M. Gray. *Vector Quantization and Signal Processing*. Kluwer Academic Publishers, Boston, MA, 1992.

[19] M. Goitein and A Niemierko. Biologically based models for scoring treatment plans. Scandanavian Symposium on Future Directions of Computer-Aided Radiotherapy, 1988.

[20] H. Greenberg. *A Computer-Assisted Analysis System for Mathematical Programming Models and Solutions: A User's Guide for ANALYZE*. Kluwer Academic Publishers, Boston, MA, 1993.

[21] H. Greenberg. Consistency, redundancy and implied equalities in linear systems. *Annals of Mathematics and Artificial Intelligence*, 17:37–83, 1996.

[22] L. Hodes. Semiautomatic optimization of external beam radiation treatment planning. *Radiology*, 110:191–196, 1974.

[23] A. Holder. Partitioning Multiple Objective Solutions with Applications in Radiotherapy Design. Technical Report 54, Trinity University Mathematics, 2001.

[24] A. Holder. Radiotherapy treatment design and linear programming. Technical Report 70, Trinity University Mathematics, San Antonio, TX, 2002. to appear in the Handbook of Operations Research/Management Science Applications in Health Care.

[25] A. Holder. Designing radiotherapy plans with elastic constraints and interior point methods. *Health Care and Management Science*, 6(1):5–16, 2003.

[26] T. Kalinowski. An algorithm for optimal collimator field segmentation with interleaf collision constraint 2. Technical report, Operations Research & Radiation Oncology Web Site, w.trinity.edu/aholder/HealthApp/oncology/, 2003.

[27] T. Kalinowski. An algorithm for optimal multileaf collimator field segmentation with interleaf collision constraint. Technical report, Operations Research & Radiation Oncology Web Site, w.trinity.edu/aholder/HealthApp/oncology/, 2003.

[28] G. Kutcher and C. Burman. Calculation of complication probability factors for non-uniform normal tissue irradiation. *International Journal of Radiation Oncology Biology and Physics*, 16:1623–30, 1989.

[29] M. Langer, R. Brown, M. Urie, J. Leong, M. Stracher, and J. Shapiro. Large scale optimization of beam weights under dose-volume restrictions. *International Journal of Radiation Oncology, Biology, Physics*, 18:887–893, 1990.

[30] E. Lee, T Fox, and I Crocker. *Integer Programming Applied to Intensity-Modulated Radiation Treatment Planning*. To appear in Annals of Operations Research, Optimization in Medicine.

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

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

[33] J. Legras, B. Legras, and J. Lambert. Software for linear and non-linear optimization in external radiotherapy. *Computer Programs in Biomedicine*, 15:233–242, 1982.

[34] G. Leichtman, A. Aita, and H. Goldman. Automated gamma knife dose planning using polygon clipping and adaptive simulated annealing. *Medical Physics*, 27(1):154–162, 2000.

[35] L. Leksell. The stereotactic method and radiosurgery of the brain. *Acta Chirurgica Scandinavica*, 102:316 – 319, 1951.

[36] W. Lodwick, S. McCourt, F. Newman, and S. Humphries. Optimization methods for radiation therapy plans. In C. Borgers and F. Natterer, editors, *IMA Series in Applied Mathematics - Computational, Radiology and Imaging: Therapy and Diagnosis*. Springer-Verlag, 1998.

[37] N. Lomax and S. Scheib. Quantifying the degree of conformality in radiosurgery treatment planning. *International Journal of Radiation Oncology, Biology, and Physics*, 55(5):1409–1419, 2003.

[38] L. Luo, H. Shu, W. Yu, Y. Yan, X. Bao, and Y. Fu. Optimizing computerized treatment planning for the gamma knife by source culling. *International Journal of Radiation Oncology Biology and Physics*, 45(5):1339–1346, 1999.

[39] J. Lyman. Complication probability as assessed from dose-volume histrograms. *Radiation Research*, 104:S 13–19, 1985.

[40] J. Lyman and A. Wolbarst. Optimization of radiation therapy iii: A method of assessing complication probabilities from dose-volume histograms. *International Journal of Radiation Oncology Biology and Phsyics*, 13:103'109, 1987.

[41] J. Lyman and A. Wolbarst. Optimization of radiation therapy iv: A dose-volume histogram reduction algorithm. *International Journal of Radiation Oncology Biology and Phsyics*, 17:433–436, 1989.

[42] J. Leong M. Langer. Optimization of beam weights under dose-volume restrictions. *International Journal of Radiation Oncology, Biology, Physics*, 13:1255–1260, 1987.

[43] S. McDonald and P. Rubin. Optimization of external beam radiation therapy. *International Journal of Radiation Oncology, Biology, Physics*, 2:307–317, 1977.

[44] S. Morrill, R. Lane, G. Jacobson, and I. Rosen. Treatment planning optimization using constrained simulated annealing. *Physics in Medicine & Biology*, 36(10):1341–1361, 1991.

[45] S. Morrill, I. Rosen, R. Lane, and J. Belli. The influence of dose constraint point placement on optimized radiation therapy treatment planning. *International Journal of Radiation Oncology, Biology, Physics*, 19:129–141, 1990.

[46] A. Niemierko and M. Goitein. Calculation of normal tissue complication probability and dose-volume histogram reduction schemes for tissues with critical element architecture. *Radiation Oncology*, 20:20, 1991.

[47] I. Paddick. A simple scoring ratio to index the conformality of radiosurgical treatment plans. *Journal of Neurosurgery*, 93(3):219–222, 2000.

[48] W. Powlis, M. Altschuler, Y. Censor, and E. Buhle. Semi-automatic radiotherapy treatment planning with a mathematical model to satisfy treatment goals. *International Journal of Radiation Oncology, Biology, Physics*, 16:271–276, 1989.

[49] F. Preciado-Walters, M. Langer, R. Rardin, and V. Thai. Column generation for imrt cancer therapy optimization with implementable segments. Technical report, Purdue University, 2004.

[50] F. Preciado-Walters, M. Langer, R. Rardin, and V. Thai. A coupled column generation, mixed-integer approach to optimal planning of intensity modulated radiation therapy for cancer. Technical report, Purdue University, 2004. to appear in Mathematical Programming.

[51] A. Pugachev, A. Boyer, and L. Xing. Beam orientation optimization in intensity-modulated radiation treatment planning. *Medical Physics*, 27(6):1238–1245, 2000.

[52] A. Pugachev and L. Xing. Computer-assisted selection of coplaner beam orientations in intensity-modulated radiation therapy. *Physics in Medicine and Biology*, 46:2467–2476, 2001.

[53] C. Raphael. Mathematical modeling of objectives in radiation therapy treatment planning. *Physics in Medicine & Biology*, 37(6):1293–1311, 1992.

[54] H. Romeijn, R. Ahuja, J.F. Dempsey, and A. Kumar. A column generation approach to radiation therapy treatment planning using aperture modulation. Technical Report Research Report 2003-13, Department of Industrial and Systems Engineering, University of Florida, 2003.

[55] I. Rosen, R. Lane, S. Morrill, and J. Belli. Treatment plan optimization using linear programming. *Medical Physics*, 18(2):141–152, 1991.

[56] S. Sɵderstrɵm and A. Brahme. Selection of suitable beam orientations in radiation therapy using entropy and fourier transform measures. *Physics in Medicine and Biology*, 37(4):911–924, 1992.

[57] D. Shepard, M. Ferris, G. Olivera, and T. Mackie. Optimizing the delivery of radiation therapy to cancer patients. *SIAM Review*, 41(4):721–744, 1999.

[58] D. Shepard, M. Ferris, R. Ove, and L. Ma. Inverse treatment planning for gamma knife radiosurgery. *Medical Physics*, 27(12):2748–2756, 2000.

[59] H. Shu, Y. Yan, X. Bao, Y. Fu, and L. Luo. Treatment planning optimization by quasi-newton and simulated annealing methods for gamma unit treatment system. *Physics in Medicine and Biology*, 43(10):2795–2805, 1998.

[60] H. Shu, Y. Yan, L. Luo, and X. Bao. Three-dimensional optimization of treatment planning for gamma unit treatment system. *Medical Physics*, 25(12):2352–2357, 1998.

[61] G. Starkshcall. A constrained least-squares optimization method for external beam radiation therapy treatment planning. *Medical Physics*, 11(5):659–665, 1984.

[62] J. Stein et al. Number and orientations of beams in intensity-modulated radiation treatments. *Medical Physics*, 24(2):149–160, 1997.

[63] H. Withers, J. Taylor, and B. Maciejewski. Treatment volume and tissue tolerance. *International Journal of Radiation Oncology, Biology, Physics*, 14:751–759, 1987.

[64] A. Wolbarst. Optimization of radiation therapy II: The critical-voxel model. *International Journal of Radiation Oncology, Biology, Physics*, 10:741–745, 1984.

[65] Q. Wu and J. Bourland. Morphology-guided radiosurgery treatment planning and optimization for multiple isocenters. *Medical Physics*, 26(10):2151–2160, 1999.

[66] P. Zhang, D. Dean, A. Metzger, and C. Sibata. Optimization o gamma knife treatment planning via guided evolutionary simulated annealing. *Medical Physics*, 28(8):1746–1752, 2001.