KTH Engineering Sciences

Utilizing Problem Structure in Optimization of Radiation Therapy

FREDRIK CARLSSON

Doctoral Thesis
Stockholm, Sweden 2008

TRITA-MAT 08/OS/03
ISSN 1401-2294
ISRN KTH/OPT SYST/DA-08/03-SE
ISBN 978-91-7178-920-4

Optimization and Systems Theory
Department of Mathematics
Royal Institute of Technology
SE-100 44 Stockholm, SWEDEN

Akademisk avhandling som med tillstånd av Kungl Tekniska Högskolan framlägges till offentlig granskning för avläggande av teknologie doktorsexamen fredagen den 25 april 2008 klockan 10.00 i sal F3, Lindstedtsvägen 26, Kungl Tekniska Högskolan, Stockholm.

Print: Universitetsservice US AB

Till min familj

Abstract

In this thesis, optimization approaches for intensity-modulated radiation therapy are developed and evaluated with focus on numerical efficiency and treatment delivery aspects. The first two papers deal with strategies for solving fluence map optimization problems efficiently while avoiding solutions with jagged fluence profiles. The last two papers concern optimization of step-and-shoot parameters with emphasis on generating treatment plans that can be delivered efficiently and accurately.

In the first paper, the problem dimension of a fluence map optimization problem is reduced through a spectral decomposition of the Hessian of the objective function. The weights of the eigenvectors corresponding to the p largest eigenvalues are introduced as optimization variables, and the impact on the solution of varying p is studied. Including only a few eigenvector weights results in faster initial decrease of the objective value, but with an inferior solution, compared to optimization of the bixel weights. An approach combining eigenvector weights and bixel weights produces improved solutions, but at the expense of the pre-computational time for the spectral decomposition.

So-called iterative regularization is performed on fluence map optimization problems in the second paper. The idea is to find regular solutions by utilizing an optimization method that is able to find near-optimal solutions with non-jagged fluence profiles in few iterations. The suitability of a quasi-Newton sequential quadratic programming method is demonstrated by comparing the treatment quality of deliverable step-and-shoot plans, generated through leaf sequencing with a fixed number of segments, for different number of bixel-weight iterations. A conclusion is that over-optimization of the fluence map optimization problem prior to leaf sequencing should be avoided.

An approach for dynamically generating multileaf collimator segments using a column generation approach combined with optimization of segment shapes and weights is presented in the third paper. Numerical results demonstrate that the adjustment of leaf positions improves the plan quality and that satisfactory treatment plans are found with few segments. The method provides a tool for exploring the trade-off between plan quality and treatment complexity by generating a sequence of deliverable plans of increasing quality.

The final paper is devoted to understanding the ability of the column generation approach in the third paper to find near-optimal solutions with very few columns compared to the problem dimension. The impact of different restrictions on the generated columns is studied, both in terms of numerical behaviour and convergence properties. A bound on the two-norm of the columns results in the conjugate-gradient method. Numerical results indicate that the appealing properties of the conjugate-gradient method on ill-conditioned problems are inherited in the column generation approach of the third paper.

Key words: Optimization, intensity-modulated radiation therapy, conjugate-gradient method, step-and-shoot delivery, column generation, quasi-Newton method, regularization, sequential quadratic programming.

Preface

After three months of backpacking in Asia to celebrate my graduation, followed by a lovely Christmas in Halmstad with my family, I was not sure of what to do next. I applied for a few jobs in different fields, cities and countries, while taking a course in economics at Handels in Göteborg. One morning in January I spotted a big flashy colour advert in Ny Teknik, a paper that my flatmate David so kindly provided. A company called RaySearch Laboratories was looking for an industrial PhD student in optimization of radiation therapy, in a joint project with KTH. Why not, I thought? I sent off my application and within a few weeks I was invited to Stockholm by Henrik Rehbinder, director of research at RaySearch. I took the train to Stockholm, and looking back, I am so glad I did.

It has been a great journey ever since and I have truly enjoyed working with one foot in academia and one foot in industry. In particular, it has been inspiring to work so close to the clinical applications, with real patient data and software used at the clinics. It has been fascinating to realize how small the gap is between applied mathematics and clinical implementation in the field of radiation therapy. The research of this thesis has been co-funded by the Swedish Research Council (VR) and RaySearch Laboratories. I am grateful to both for making this project possible.

There are many people who have supported me in my work with this thesis. First of all I owe many thanks to my advisor Anders Forsgren, it has been a privilege working with you. Your positive approach and mathematical expertise have been of great value, both for the project and for me personally. Thanks also to Krister Svanberg for being an academic member of the reference group of this project.

Many thanks to Johan Löf, the founder and CEO of RaySearch. This project would not have existed had it not been for you. The great impression I got of you at the job interview was an important reason for me taking on this project. Five years later, this strong impression still stands, both on a personal and a professional level. Another key player that has been of great support is Henrik Rehbinder. Your interest and active participation in my project combined with your thorough knowledge in mathematics and radiation therapy has meant a great deal to my research. Also, I am indebted to Göran Sporre for always having time for discussions and for providing precious feedback on my work.

It has been a true pleasure to spend time with all of you at the Division of

Optimization and Systems Theory, KTH. I want especially to thank my roommate Tove Gustavi for being in such a positive spirit, I have really enjoyed your company. Also, I am grateful to my “brother-in-arm” in optimization, Mats Werme, for our collaboration in numerous courses, for taking interest in my work and for always making me laugh.

When I started we were 18 people working at RaySearch, now we have grown to almost the triple. It has been a great experience to work in such an emerging company with so many skilled, friendly and down-to-earth people. I would like to thank all of you for making RaySearch a special place to work. In particular, I would like to thank Kjell Eriksson, Anna Lundin and Catharina Sundgren for your support with C++ and Björn Hårdemark for always providing a valuable answer, no matter what question I ask.

I am also grateful for the encouragement and interesting discussions with other researchers in the field of optimization and radiation therapy.

I would like to thank all my other friends and colleagues for brightening my days, you mean very much to me. Thanks to my lovely family for your support, encouragement and kindness. Finally, thanks Lotta for being such a wonderful person and for making me look forward to every day.

Stockholm, March 2008

Fredrik Carlsson

Contents

Introduction

Radiation therapy, the use of ionizing radiation to treat cancer disease, is one of the three most common types of cancer treatments. The other two are surgery and chemotherapy. Of the approximately 1.4 million people diagnosed with cancer in USA 2006[1], 68% received some form of radiation therapy [37]. This thesis deals with optimization approaches for an advanced and increasingly used form of radiation therapy called *intensity-modulated radiation therapy* (IMRT).

A challenge in IMRT is to design treatment plans that can be delivered efficiently and accurately while fulfilling the designated treatment goals. The aim of the research described in this thesis is to develop and evaluate optimization approaches that solve IMRT optimization problems efficiently while finding solutions that are advantageous from a clinical perspective. To develop such approaches, the problem structure of the IMRT optimization problems must be understood and utilized.

The content of this thesis is divided into an introduction and four appended papers. The introduction gives the basics of radiation therapy and introduces fundamental concepts of optimization theory. The latter part of the introduction deals with optimization of IMRT treatment plans, with particular emphasis on mathematical structure and treatment delivery requirements. In the final part of the introduction, the main contributions of this thesis are discussed and a summary of the appended papers is given.

1 Radiation therapy

Radiation therapy, or radiotherapy, may be used either as a stand-alone treatment or in conjunction with other forms of treatment such as surgery or chemotherapy. Radiotherapy is used both as a curative treatment with the aim of curing the cancer, and as a palliative treatment to control symptoms and improve quality of life if a cancer is too advanced to cure. Radiotherapy is a common treatment for many different cancer types, such as cancer in the prostate, head-and-neck region, breast, lung, brain and skin [37].

Radiotherapy treatments can be classified as either *external (beam)* or *internal*, referring to the location of the source of radiation relative the patient. External beam radiotherapy is by far the most commonly used. For example, 90% of all radiotherapy treatments in USA 2006 were external beam treatments according to [37].

---

[1] http://www.cancer.gov

The extensive development of software and hardware during the last decades for imaging and external beam radiotherapy has paved the way for the IMRT technique. IMRT belongs to the class of external photon beam radiotherapy, which uses megavoltage X-rays as the treatment modality. IMRT can be seen as a generalization of *three-dimensional conformal radiation therapy* (3DCRT); see Section 1.2. 3DCRT is, in turn, an enhancement of *conventional radiation therapy*, where the treatments are based primarily on two-dimensional X-ray images. For a discussion on the advances in external photon beam radiotherapy, see [19]. External beam radiotherapy also includes other treatment modalities such as protons, neutrons and light ions. Other radiotherapy treatment techniques include brachytherapy (internal) and stereotactic radiosurgery (external). From here on, radiotherapy refers to external photon beam radiation therapy.

1.1 Radiobiology

Radiotherapy strives for destroying as many (all if curative treatment) cancer cells as possible, while limiting damage to the healthy tissue. This is accomplished by directing high energy photons to the *target volume* with high precision. The photons interact with the tissue in the patient through elastic and inelastic collisions. Electrons and free radicals that are released from these collisions scatter through the tissue and eventually collide with the DNA molecules of the cells. These collisions break the DNA molecule by ionizing atoms in the molecule. A small fraction of the damages are non-repairable, which results in that the cells eventually die. Healthy cells have a better ability to recover from sublethal damages than cancer cells. The radiotherapy treatment is therefore divided into *fractions* that are given daily over a specific time period, typically five days a week for six to eight weeks. The healthy cells can then recover and repopulate between each treatment delivery at a faster rate than the cancer cells. For a thorough description of radiobiology, see, e.g., [71].

1.2 Hardware

A *linear accelerator* (*linac*) generates the megavoltage photon *fluences* used in radiotherapy. The linac accelerates electrons in a strong electric field onto a brehmsstrahlung target made of high density material, where collisions result in scattering of high-energy photons. This target is referred to as the primary photon source. A portion of the photons are collected and pass a *flattening filter* before leaving the linac through the *gantry head* (see Figure 1, number 2). The gantry (Figure 1, number 1) rotates around the patient in order to deliver the photon fields, or *beams*, from different directions. The gantry rotation is centered at the *isocenter point*, and the patient is normally positioned such that the isocenter point lies in the target volume.

The amount of radiation absorbed by the tissue is called *dose* and has the unit Gray (abbreviated as Gy), with $1\,\mathrm{Gy} = 1\,\mathrm{J/kg}$. The output of a linac is measured

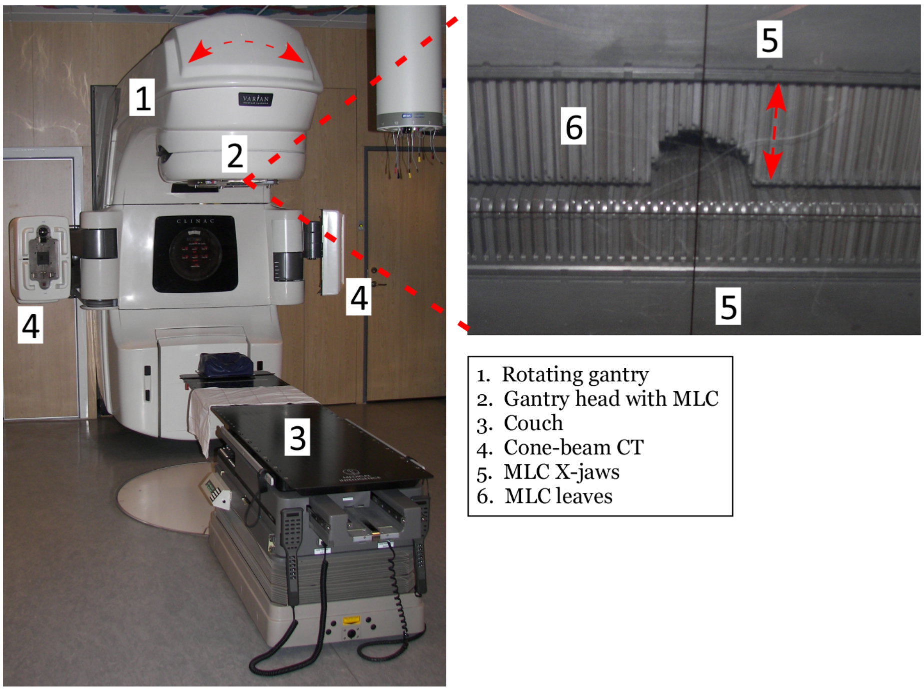

Figure 1: A treatment room with a linear accelerator (a Varian Clinac at the Karolinska University Hospital, Stockholm), equipped with an MLC and a cone-beam CT system.

in *monitor units* (MUs), and 1 MU is defined as the fluence of a square field that results in a dose of 1 cGy at a specific depth in a water phantom. The *dose-rate* of a linac is measured in MU/min. Typically, a dose-rate between 100-600 MU/min is used and the fraction dose to the target volume is in the order of 2 Gy. This results in a beam-on-time between 20 and 120 seconds for a fraction. With smaller field sizes, which are typical for IMRT treatments, the beam-on-time is longer.

In 3DCRT, each beam is shaped to match the projection of the target volume onto the *fluence plane* of the beam; see Figure 2. The evolvement of three-dimensional imaging technology providing accurate information of the tumor geometry and location has made it possible to compute these projections accurately. One or two pairwise opposed movable metal blocks called *jaws* (Figure 1, number 5) are positioned in the gantry head to block parts of the beam not intersecting with the target volume, resulting in a rectangular beam shape. A *multileaf collimator* (MLC) can be used to improve the matching further. The MLC is mounted in the gantry head and consists of several pairwise opposed tungsten leaves (Figure 1,

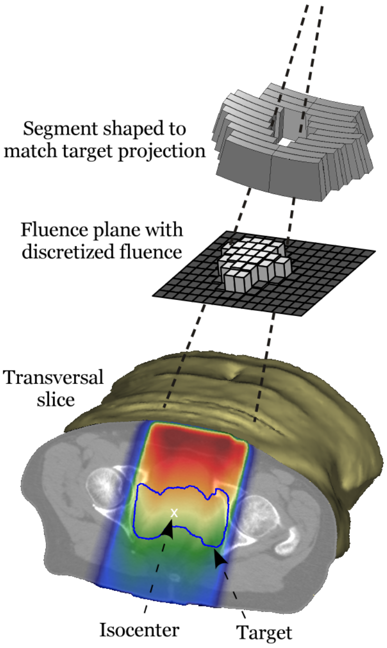
Segment shaped to match target projection
Fluence plane with discretized fluence
Transversal slice
Isocenter
Target

Figure 2: A schematic illustration of one projection-based segment for a prostate case and the resulting dose distribution visualized in a transversal slice. The blue contour outlines the target volume in the slice. Red regions have high dose and blue regions have low dose.

number 6). The leaves can be independently positioned with high accuracy to fine-tune the shape of the beam. A configuration of the jaws and the leaves of the MLC is called a *segment*, or *aperture*. Due to mechanical limitations of the MLC, not every combination of leaf positions can be realized. These limitations differ between MLCs from different manufacturers. An extensive description of some MLCs used clinically is given in [30]. Figure 2 shows a schematic illustration of one projection-based segment for a prostate case, the resulting fluence and the dose distribution in the patient (red regions have high dose and blue regions have low dose). Note that the dose is higher close to the intersection of the beam with the patient than further away. This is a characteristic of photon fields, which implies that more than one beam is necessary for generating conform dose distributions to target volumes.

1.3 Beam model and dose calculation

An accurate computation of the dose delivered to the patient requires an accurate estimate of the photon energy fluence distribution incident on the patient. The calculation of the incident fluence is based on a beam model. It models scattering effects in the gantry head and the impact of the jaws and the MLC on the fluence. Often, beam models split the fluence into primary and secondary photons emerging from the source and the flattening filter, respectively. The models also account for effects induced by the jaws and the MLC such as leakage and scatter around the edges of the leaves. It is hard to accurately model the transmitted fluence of small and/or irregular segments since the scattering and leakage effects are considerable [22, 66]. Additional requirements on minimum area and regularity of segments may therefore be imposed by the clinicians to reduce this source of error.

A crucial component of the software for radiotherapy is the dose engine. It computes a dose distribution d in the patient volume V given the incident fluence τ and the patient geometry G describing the patient surface and the tissue density. In all four papers, a pencil beam algorithm [34] is used as the dose engine during optimization since it is very fast. In paper C, a collapsed cone algorithm is used to compute a more accurate final dose distribution. The increased accuracy is a result of a more precise handling of how heterogeneities in the patient, i.e., varying tissue density, affect the dose deposition [2]. For a description of the collapsed cone algorithm, see [1].

The pencil beam algorithm is based on a *pencil beam kernel* which is precalculated for a homogeneous medium (water) using a Monte Carlo particle transport method. The pencil beam kernel is then applied to the treatment and patient geometries, which results in *beamlets* $p(r, \rho, G(r))$, describing the energy deposition per unit mass at a point r in the patient volume due to fluence incident on a point ρ on a fluence plane. Assuming radially symmetric beamlets, a parametrization of p in cylindrical coordinates (ρ, z) can be made such that $p = p(r, \rho, G(r)) = p(\rho - \rho_0, z(r, \rho, G(r)))$, where ρ is a coordinate in the fluence geometry, ρ_0 lies on the fluence plane on the line between the source and r, and z is the depth; see Figure 3 for an illustration of the geometry. The influence of tissue heterogeneity may be corrected for when computing z [39]. The total dose $d(r)$ in a point $r \in V$ is given by the convolution integral

$$d(r) \quad = \quad \iint\limits_{S} p(\rho - \rho_0, z(r, \rho, G(r)))\, \tau(\rho)\, \mathrm{d}\rho, \tag{1}$$

where $\tau(\rho)$ is the incident fluence at ρ and S is the union of the fluence planes, or cross-sections, of all beams. In (1), mono-energetic photons are assumed. A more general formulation includes an integration over photon energies with the beamlets being functions of the photon energy.

In practice, V is discretized into m cubic *voxels* and S is discretized into n rectangular *bixels* which results in that (1) can be written as

$$d = P\tau, \tag{2}$$

where d is the m-dimensional dose distribution vector, P is the $m \times n$ dose matrix, and τ is the n-dimensional fluence vector, or bixel vector. Typical sizes are $4 \times 4 \times 4$

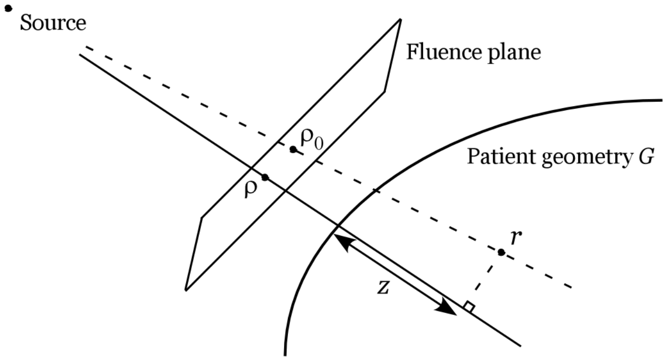

Figure 3: Geometry for calculating the dose in r due to fluence incident on a point ρ on the fluence plane of a beam.

mm^3 for the voxels and 5×5 mm^2 for the bixels. This results in n being in the order of 10^3 and m being in the order of 10^5. The speed of the dose computation (2) can be increased by approximating the pencil beam kernel by a decomposition [15]. A similar method to this has been used in papers B and C.

1.4 Imaging techniques and patient geometry

The quality of radiotherapy treatments relies heavily on the accuracy of the geometrical data of the patient provided by digital images. The technology for generating high-resolution and high-contrast images of the patient in three-dimensions (3D) has evolved rapidly over the last two decades. This has radically improved the conditions for accurate delineation of tumor regions and normal structures which is an important part of the radiotherapy treatment planning process.

Imaging techniques for radiotherapy planning include computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), and single photon emission computed tomography (SPECT). They all belong to the class of tomographic techniques, where 2D projection data from multiple directions is gathered and fed into a tomographic reconstruction software algorithm to yield a 3D dataset of the patient. The dataset may be viewed in 2D slices orthogonal to different axis of the body. By far the most common imaging technique for radiotherapy planning is CT, where X-rays are used to acquire data about tissue density. CT images are necessary for dose computation since they hold density information of the patient. High-contrast resolution images are obtained in regions with varying densities, e.g., the surroundings of a bone structure, while the soft tissue contrast is lower. However, the delineation of organs is often performed on CT images also in soft tissue regions. MRI visualizes the structure and function of the body by creating a strong magnetic field which, combined with radio waves, results in that hydrogen atoms emit a weak radio signal that is detected. PET visualizes functional processes in the body by detecting pairs of gamma rays generated when positrons emitted from a radio-isotope are annihilated by electrons. SPECT uses a gamma camera to acquire multiple 2D images from different directions. Many modern scanners combine CT with either MRI or PET to yield images that combine the high-contrast resolution of CT with the functional imaging capabilities of MRI and PET. For more details on medical imaging, see, e.g., [72].

Once the images are acquired, so-called *regions of interest* (ROIs) of the patient are specified. The ROIs represent regions of the patient of specific interest for the treatment, such as the tumor region(s) and healthy organs. ROIs representing healthy organs are called *organs-at-risk* (OARs). The ROIs are often delineated manually slice by slice, which may be time-consuming. Image segmentation software can speed-up the delineation process by automating part of it.

The definition of target volume is commonly separated into different ROIs [60]. The *gross target volume* (GTV) is defined as the gross extent of the malignant growth as determined by images or palpation. The *clinical target volume* (CTV) is specified as an expansion of the GTV to account for spread of microscopic malignant disease that cannot be seen in the images. The task of specifying a correct CTV region is, of course, very complicated and based on clinical experience. Consequently, delineation of the CTV is one of the more prominent sources of error in radiotherapy planning [75].

1.5 Geometrical uncertainties

The regions of interest delineated on the acquired *planning images* represent the patient at the time of scanning. Between and during fractions, the actual shape and position of the organs that the ROIs represent change due to factors such as rectal filling and breathing motion. Other factors include patient setup errors, tumor shrinkage and weight loss. Since IMRT plans typically have dose distributions with high dose to the tumor and a sharp dose fall off outside the tumor, it is important to handle these geometrical uncertainties.

A common practice for reducing setup errors is to position the patient using laser alignment or to perform a couch correction, where real-time images are compared to the planning images and the couch is moved to compensate for deviations. Such real-time images may be acquired by a portal imaging device or a cone-beam CT; see Figure 1, number 4. The motion during the treatment delivery can be reduced for head-and-neck cases by immobilization techniques such as fixation masks and biteplates. For cancers located in regions affected by the breathing cycle, gating techniques can be used to avoid irradiating the patient when the displacements of ROIs are intolerable.

The precautions described above can reduce the geometrical uncertainties, but cannot remove them entirely. There are also other sources of errors present in the treatment such as delineation errors and inaccuracies in the treatment delivery. To compensate for these, a margin is applied to the CTV to generate a *planning target volume* (PTV) [60]. The size of this margin is, of course, very important; if the margin is too large, a large portion of healthy cells receives high dose, and if the margin is too small, there is a risk that cancerous cells do not receive high doses in some fractions and survive the treatment.

It is possible to compensate for delivery errors in previous fractions by adaptively replanning between fractions [11, 47, 59]. This requires that images of the patient are acquired during treatment. For a review of motion effects and compensation approaches in radiotherapy, see [78].

1.6 Treatment planning and IMRT

The goal of radiotherapy treatment planning is to design a treatment plan that handles the conflict of delivering high dose to the target volume while avoiding excessive dose to OARs in the best possible way. The field of radiotherapy treatment planning can be divided into *forward treatment planning* and *inverse treatment planning*. Figure 4[2] illustrates the conceptual differences between these approaches.

Forward treatment planning is essentially a trial-and-error procedure. Given the patient geometry data, the planner defines a set of beams, their angles and possibly attenuates some beams by using wedge-shaped metal blocks. The dose is then computed and if the planner is not satisfied with the dose distribution, the setup is altered and a new dose is computed. The procedure continues until the planner is satisfied with the plan. Forward treatment planning is the original procedure for generating treatment plans and it is commonly used for conventional treatment planning and 3DCRT treatment planning. In inverse planning, computer algo-

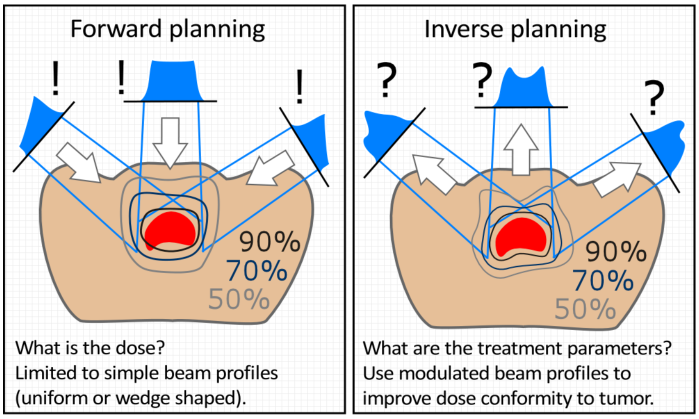
Forward planning
!
!
!
90%
70%
50%
What is the dose?
Limited to simple beam profiles
(uniform or wedge shaped).
Inverse planning
?
?
?
90%
70%
50%
What are the treatment parameters?
Use modulated beam profiles to
improve dose conformity to tumor.

Figure 4: A comparison of forward planning and inverse planning.

---

[2] The figure is based on an illustration by Anders Brahme.

rithms are used to convert treatment goals usually formulated in the dose domain into treatment parameters associated with the delivery system. Inverse treatment planning problems are formulated as optimization problems which are solved iteratively. Inverse treatment planning is always used for IMRT, but may also be used for other forms of radiotherapy.

Radiotherapy treatment plans are often evaluated by studying dose distributions on 2D slices and so-called *dose-volume histograms* (DVHs), where the dose distribution of a ROI is displayed as a curve. The interpretation of a point (x, v) on a DVH curve is that v percent of the ROI receives a dose of at least x Gy. The DVH holds no spatial information of the dose distribution but is nevertheless useful since many treatment protocols are based on dose-to-volume requirements. Other measures of plan quality include radiobiological functions such as *tumor control probability* (TCP) and *normal tissue complication probability* (NTCP), which are based on biological models of the response of the cells to dose; see, e.g., [46].

The concept of IMRT was first introduced in [18], where it was shown that non-uniform fluences improve the dose conformity to a nonconvex target. Instead of introducing yet another definition of IMRT, the one presented in [13] is quoted: "*IMRT is a radiation treatment technique with multiple beams in which at least some of the beams are intensity-modulated and intentionally deliver a non-uniform intensity to the target. The desired dose distribution in the target is achieved after superimposing such beams from different directions. The additional degrees of freedom are utilized to achieve a better target dose conformity and/or better sparing of critical structures*". Comprehensive introductions to IMRT can be found in [3, 13, 77].

The benefits of IMRT are most prominent for cases with the tumor located close to healthy organs, such as cancer in the head-and-neck region and prostate cancer. For cases with simpler geometry, 3DCRT or conventional radiotherapy is often used. However, the rate of clinical acceptance for IMRT has increased significantly the last few years [49]. The potential for dose escalation to the target volume and enhanced normal tissue sparing are the two main reasons for this increase [49].

Figure 5 illustrates an MLC-based IMRT plan for a head-and-neck case with nine angularly equidistant beams. The dose distribution presented is the total dose delivered over all fractions. The top images show a transversal slice (left) and a sagittal slice (right) of the CT images of the patient together with the delineated ROIs. The plan has two CTV regions; one called CTV 72 for escalating dose to the GTV and one called CTV 49.5 for directing radiation to a larger volume where microscopic malignant disease may be present. Here, the PTV is given by a five millimeter expansion of the CTV 49.5 region. The image in the middle of the figure illustrates the treatment geometry and the beam profiles, while the DVH curves for some of the ROIs of the plan are shown at the bottom.

By viewing the CT slices and the DVH curves, it is clear that the high dose region is concentrated to the CTV 72 region while the maximum dose to the cord is low. Note that one parotid gland is sacrificed while the other is spared, this strategy allows for higher conformity to the CTV 72 region. A uniform dose to a

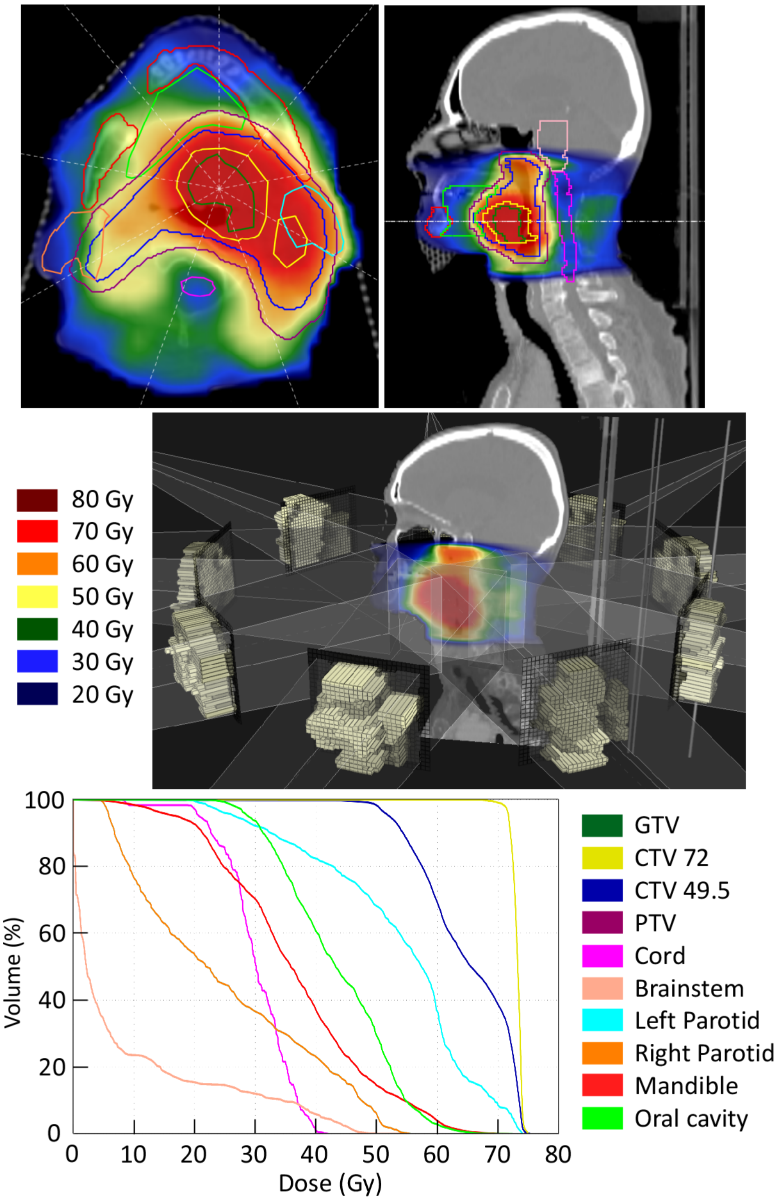

Figure 5: A head-and-neck IMRT plan, illustrating the ROIs and the dose distribution in a transversal slice (top left) and a sagittal slice (top right). The treatment setup and the beam profiles are shown in the middle and the DVH curves for some of the ROIs are shown at the bottom.

ROI corresponds to the DVH curve being a vertical line, which is almost the case for the CTV 72 region here. The DVH curves corresponding to the OARs should, ideally, be as far to the left as possible. Overlap between OARs and target ROIs and the scatter of dose in the patient will, however, imply that some parts of the OARs will receive high dose. There is obviously a conflict in delivering high dose to the target volume while avoiding excessive dose to OARs.

1.7 IMRT delivery techniques

Modulating the beam profiles to deliver IMRT treatment plans can be done in various ways. This thesis considers the commonly used step-and-shoot IMRT delivery technique, where a fixed set of beams are defined and the fluence of each beam is modulated by superimposing the fluence of a few MLC segments. Typically, three to nine beams are used. Since the radiation is off when the leaves and jaws move to form the next segment, a step-and-shoot plan with many segments may lead to a long delivery time. This issue is addressed in paper C, and to some extent in papers B and D. It is also preferable to use large and regular segments since such plans have a low number of MUs and can be delivered accurately. This issue is also addressed in paper C.

An alternative to step-and-shoot delivery is dynamic MLC (DMLC) delivery, where the fluence modulation is achieved by moving the MLC leaves while the radiation is on. An increasingly popular technique for performing IMRT is tomotherapy, where the treatment is delivered with a narrow slit beam. The patient is moved through a rotating gantry and irradiated continuously. Other techniques for delivering modulated fluence include inserting a metal compensator in the beam, a computer-controlled scanned beam and a linac mounted on a highly manoeuvrable robotic arm. A relatively new approach to IMRT delivery with the potential for shorter delivery times is referred to as volumetric arc therapy (VMAT) or intensity-modulated arc therapy (IMAT). The technique is a generalization of DMLC in that the gantry is rotating continuously while the beam is on and the leaves move. For more thorough descriptions of IMRT delivery techniques, see, e.g., [28, 38, 56].

2 Optimization concepts

In optimization, also referred to as mathematical programming, the goal is to determine the values of a set of *variables* such that the *objective function* is minimized (or maximized) while satisfying predefined restrictions. This is done by formulating and solving an *optimization problem.* For real-life applications, the formulation of the optimization problem is based on a model of the underlying problem. The model aims at describing the problem as accurately as possible, while allowing for a formulation that is suitable for optimization solvers. For instance, if the underlying problem is infinite-dimensional, a discretization is often necessary to make the problem practically solvable. All optimization problems formulated in this thesis are finite-dimensional, i.e., the variable set is represented by a finite vector. Further, all problems are formulated as minimization problems. Maximization problems are equivalent to minimization problems with the sign of the objective function reversed.

The restrictions on the variables form a *feasible region* $\mathcal{F}$ with elements denoted by *feasible solutions*. The objective function F quantifies the quality of every feasible solution by associating a real value to it, i.e., $F : \mathcal{F} \rightarrow I\!R$. The *optimal solution*, or *minimizer*, is given by the feasible solution with the lowest objective value. The optimization problem of minimizing F over $\mathcal{F}$ is written

$$\begin{array}{ll} \underset{x}{\text{minimize}} & F(x) \\ \text{subject to} & x \in \mathcal{F}, \end{array} \tag{3}$$

where the variables are denoted by x. From here on, this section deals with *continuous optimization*, where the variables are allowed to assume real values, as opposed to *discrete optimization*, where the variables are restricted to assume integer values. The feasible region is assumed to be a subset of $I\!R^n$, the n-dimensional Euclidean space.

A point $x^* \in \mathcal{F}$ is a *global minimizer* to (3) if $F(x^*) \leq F(x)$ for all $x \in \mathcal{F}$. A point $x^* \in \mathcal{F}$ is a *local minimizer* to (3) if there exists an $\epsilon > 0$ such that $F(x^*) \leq F(x)$ for all $x \in \mathcal{F}$ that satisfy $\|x - x^*\| < \epsilon$, that is, there is no point in the neighbourhood of x^* with a lower objective value. A global minimizer is also a local minimizer, but the converse is not true in general, as is discussed in the next section.

2.1 Convexity

The concept of convexity is central in optimization and much research has been devoted to the field of *convex optimization*; see, e.g., [61].

A set $\mathcal{F}$ is said to be a *convex set* if $\alpha x + (1 - \alpha)y \in \mathcal{F}$ for all $x, y \in \mathcal{F}$ and $0 < \alpha < 1$. In other words, for every pair of points in $\mathcal{F}$, the entire straight line segment that joins them lies in $\mathcal{F}$. If F is defined on a convex set $\mathcal{F}$, then F is said to be a *convex function* on $\mathcal{F}$ if

$$F(\alpha x + (1 - \alpha)y) \leq \alpha F(x) + (1 - \alpha)F(y), \tag{4}$$

for all $x, y \in \mathcal{F}$ and $0 < \alpha < 1$. If F is a convex function, then $-F$ is a *concave* function. Affine functions fulfill (4) with equality and are thus both convex and concave. Finally, if F fulfills (4) with "$\leq$" replaced by "$<$" for all $x, y \in \mathcal{F}$ such that $x \neq y$ and for all $0 < \alpha < 1$, then F is *strictly convex*.

An optimization problem with a convex objective function and a convex feasible region is a *convex optimization problem*. For convex optimization problems, local minimizers are global minimizers. This is a great advantage from a practical viewpoint; many efficient optimization methods are designed to find local minimizers.

If F, in addition, is strictly convex, then the optimal solution is unique. A nonconvex optimization problem is any problem where either the feasible region or the objective function is nonconvex.

2.2 Nonlinear programming

This section considers general nonlinear optimization problems and their so-called first order necessary optimality conditions. These conditions are the foundation of the optimization methods used in this thesis.

The feasible region is commonly defined by a set of *constraint functions* $C_i(x)$, $i = 1, \ldots, m$. More specifically, the feasible region consists of points satisfying the *inequality constraints* $C_i(x) \geq 0$ for $i \in \mathcal{I}$ and the *equality constraints* $C_i(x) = 0$ for $i \in \mathcal{E}$, where $\mathcal{I}$ and $\mathcal{E}$ partition the set $\{1, \ldots, m\}$. The nonlinear programming problem is given by

$$(NLP) \qquad \begin{array}{ll} \underset{x}{\text{minimize}} & F(x) \\ \text{subject to} & C_i(x) = 0, \quad i \in \mathcal{E}, \\ & C_i(x) \geq 0, \quad i \in \mathcal{I}. \end{array} \tag{5}$$

The feasible region of (5) is convex if $C_i(x), i \in \mathcal{I}$, are concave functions on $I\!R^n$ and $C_i(x), i \in \mathcal{E}$, are affine functions on $I\!R^n$. If, in addition, F is convex, then (5) is a convex problem. A constraint $C_i(x) \geq 0$ is *active* at x if $C_i(x) = 0$, and consequently, all equality constraints are active in the feasible region. Throughout Section 2, it is assumed that F and C_i, $i = 1, \ldots, m$, are twice continuously differentiable. The gradient of the objective function at a point x is denoted by $\nabla F(x)$, and the Jacobian of the constraints $C(x)$ is denoted by $J(x)$. The Jacobian is an $m \times n$ matrix, with the ith row given by $\nabla C_i(x)^T$.

Let x^* be a local minimizer to (5) and assume that the gradients of the active constraints at x^* are linearly independent. Then, there exists a vector λ^* such that the first order necessary conditions hold, i.e., such that:

$$\begin{aligned} \nabla F(x^*) &= J(x^*)^T \lambda^*, & & \text{(6a)} \\ C_i(x^*) &= 0, & i \in \mathcal{E}, & & \text{(6b)} \\ C_i(x^*) &\geq 0, & i \in \mathcal{I}, & & \text{(6c)} \\ \lambda_i^* &\geq 0, & i \in \mathcal{I}, & & \text{(6d)} \\ C_i(x^*)\lambda_i^* &= 0, & i \in \mathcal{I}, & & \text{(6e)} \end{aligned}$$

where λ_i^* is the so-called Lagrange multiplier associated with the ith constraint. The conditions (6) are often referred to as the *Karush-Kuhn-Tucker (KKT) conditions* [40, 41]. For convex problems, these conditions are sufficient for determining (global) optimality. This is generally not true for nonconvex problems. The assumption of linearly independent gradients of the active constraints can be weakened, see [8] for a discussion.

A special case of (5) is the quadratic programming (QP) problem, which is given by

$$(QP) \qquad \begin{array}{ll} \underset{x}{\text{minimize}} & \frac{1}{2}x^T H x + c^T x \\ \text{subject to} & Ax = b, \\ & x \geq 0, \end{array} \tag{7}$$

where H is an $n \times n$ symmetric matrix and A is an $m \times n$ matrix. The constraints of (7) are called *linear constraints* and *bound constraints*, respectively.

Let x be a local minimizer to (7). Applying the KKT conditions, there exist vectors $s \in I\!R^n$ and $\lambda \in I\!R^m$ such that

$$Hx + c = A^T\lambda + s, \tag{8a}$$

$$Ax = b, \tag{8b}$$

$$x_j s_j = 0, \qquad j = 1, \ldots, n, \tag{8c}$$

$$x, s \geq 0. \tag{8d}$$

If H is positive semi-definite, then (7) is convex and the conditions of (8) are sufficient for global optimality.

2.3 Solvers

The optimization methods utilized in this work are designed for finding a KKT point, i.e., a feasible point that satisfies the KKT conditions, at least in some approximate sense. Given a starting point x_0, the methods proceed by generating a sequence of iterates $\{x_k\}_{k \geq 0}$ until a termination criterion is fulfilled. In each iteration k, the algorithms compute a *search direction* p_k and the new point is given by $x_{k+1} = x_k + \alpha_k p_k$, where α_k is the *step length*. The step length is (ideally) given as the solution of

$$\underset{\alpha > 0}{\text{minimize}} \quad F(x_k + \alpha p_k), \tag{9}$$

but it is often impractical to solve (9) exactly. Instead, an approximate step length is computed by evaluating $F(x_k + \alpha p_k)$ for a few different values of α. For more details on methods for computing step lengths, see, e.g., [54].

In the following discussion, it is assumed that the objective function is strictly convex. First, unconstrained problems where the feasible region equals $I\!R^n$ are considered. For these problems, minimizing $F(x)$ is equivalent to finding a point x^* such that $\nabla F(x^*) = 0$. Three related search direction strategies for this problem class are described, starting with *Newton's method*.

The search direction of Newton's method is given by the step to the minimizer of a local second-order approximation of F about x_k. The quadratic model, denoted by q_k, is given by

$$q_k(x_k + p) = F(x_k) + \nabla F(x_k)^T p + \tfrac{1}{2} p^T H(x_k) p, \tag{10}$$

where $H(x)$ denotes the Hessian $\nabla^2 F$ at a point x. The *Newton direction* p_k is given by the unique minimizer to $\nabla_p q_k = 0$, which results in the *Newton equations*

$$H(x_k)p_k = -\nabla F(x_k). \tag{11}$$

The more accurately the quadratic model approximates $F(x_k+p)$, the more reliable search direction p_k is obtained. Newton's method has a fast rate of local convergence, but requires explicit computation of the Hessian in every iteration. This is a major drawback for large problems, which the following two algorithms avoid by not requiring computation of the Hessian.

The search directions of *quasi-Newton* methods are given by (11), with the true Hessian $H(x_k)$ replaced by a symmetric approximation B_k. This approximation is updated in every iteration to satisfy $B_{k+1}(x_{k+1} - x_k) = \nabla F(x_k + 1) - \nabla F(x_k)$. For practical reasons, an update strategy that preserves positive definiteness while being of low rank may be preferable. An important class of update strategies fulfilling these requirements is the *Broyden class*. For an overview of quasi-Newton methods, see, e.g., [54].

The nonlinear *conjugate-gradient* method computes search directions through $p_k = -\nabla F(x_k) + \beta_k p_{k-1}$, where β_k is a scalar. In this thesis (papers B and D), the conjugate gradient method is solely used for solving QP problems. Then, the search directions are conjugate, i.e., $p_k^T H p_l = 0$ if $k \neq l$, and the method converges in at most n iterations in exact arithmetic. The quasi-Newton methods of the Broyden class generate identical iterates to the conjugate-gradient method for QP problems, given that (9) is solved exactly in every iteration [53].

The final part of this section is devoted to *sequential quadratic programming* (SQP) methods for solving general nonlinear problems of the form (5). An SQP method proceeds by solving a sequence of QP subproblems. In the kth iteration, the search direction p_k is computed by solving the QP problem (with $x = x_k$ fixed),

$$\begin{array}{lll} \underset{p}{\text{minimize}} & \frac{1}{2}p^T\nabla^2_{xx}L(x,\lambda)p + \nabla F(x)^T p & \\ \text{subject to} & \nabla C_i(x)^T p = -C_i(x), & i \in \mathcal{E}, \\ & \nabla C_i(x)^T p \geq -C_i(x), & i \in \mathcal{I}, \end{array} \tag{12}$$

where $L(x,\lambda)$ is the Lagrangian function $L(x,\lambda) = F(x) - \lambda^T C(x)$ and $\nabla^2_{xx}L(x,\lambda)$ is positive definite (otherwise it is replaced by a positive definite approximation). The Hessian of the Lagrangian may be approximated by a quasi-Newton approximation B_k. The Lagrange multipliers λ are updated in every iteration. Note that (12) has linear constraints, which means that a strategy for generating feasible solutions must be adopted. The original constraints $C(x)$ may be violated in the new point $x_{k+1} = x_k + \alpha_k p_k$ since p_k is feasible with respect to the linearizations of the original constraints in (12). However, the SQP method will asymptotically converge to a feasible solution to the original problem (5). A comprehensive presentation of the SQP method can be found in [32]. In papers A and B, the SQP solver NPSOL[3] [33] is used, while an SQP solver developed at RaySearch is used in paper C.

---

[3] NPSOL is a registered trademark of Stanford University.

3 Optimization of IMRT treatment plans

The first optimization approaches to the inverse problem of IMRT, which occurred twenty years ago, were often inspired by optimization methods used for image reconstruction problems. The early publications include [16], where an inverse back projection is performed to find the optimal shapes of the incident beam profiles, and [76], where an approach using a global optimization method called simulated annealing is presented. Optimization approaches to IMRT using local methods more related to the approaches used in this thesis, and in many modern treatment planning software packages, were first introduced in [14, 45]. Other IMRT optimization strategies include integer optimization approaches, see, e.g, [43], and approaches focusing on robustness with respect to treatment uncertainties [21, 55]. A presentation of several optimization approaches to IMRT is given in [68].

3.1 Mathematical formulation

The inverse problem of finding the fluence τ that generates the prescribed dose distribution $\hat{d}$ is equivalent to finding the solution to the Fredholm integral equation of the first kind,

$$\hat{d}(r) \quad = \quad \iint_S p(\rho - \rho_0, z(r, \rho, G(r)))\, \tau(\rho)\, \mathrm{d}\rho, \tag{13}$$

with notation following (1). This is an ill-posed problem since it in general has no solution, even if negative fluence is allowed. The integration with the beamlet p has a smoothing effect on τ in the sense that high-frequency components in τ are smoothed out. The reverse process, i.e., computing τ from $\hat{d}$, therefore tends to amplify high-frequency components in $\hat{d}$. Such components are typical for IMRT problems since the dose prescriptions to the target volume and the OARs are conflicting, which results in a discontinuous $\hat{d}$.

To solve the inverse problem numerically, (13) is discretized, which results in the problem of finding the non-negative n-dimensional fluence τ such that the difference between $\hat{d}$ and $P\tau$ is minimized. Here, $\hat{d}$ is the m-dimensional prescription vector and P is the $m \times n$ dose matrix introduced in Section 1.3. In practice, $m \gg n$ and P has full column rank. Measuring the difference between $\hat{d}$ and $P\tau$ by the two-norm results in the QP

$$\begin{array}{ll} \underset{\tau}{\text{minimize}} & \|\hat{d} - P\tau\|_2^2 \\ \text{subject to} & \tau \geq 0, \end{array} \tag{14}$$

which is convex since the Hessian $H = P^T P$ is positive definite. The ill-posedness of (13) is inherited in (14) in that H is ill-conditioned with many eigenvalues close to zero [4]. This ill-conditioning results in that many different fluence vectors produce similar dose distributions, and thus similar objective values. The unique optimal solution of (14) is typically very jagged due to the high frequencies associated with eigenvectors corresponding to small eigenvalues; see the left part of Figure B.1. Jagged fluence profiles should be avoided in radiotherapy since they result in an increased number of MUs and may affect the accuracy of MLC-based deliveries [52]. One approach for avoiding this problem is to apply regularization techniques to obtain solutions with less jagged fluence profiles; see Section 3.3. Another approach is to optimize directly on the treatment parameters rather than the fluence. This approach is described in Section 3.4.

3.2 Optimization functions

In practice, it is not viable to specify the m-dimensional prescription vector $\hat{d}$ used in (14). Instead, the treatment goals of an IMRT plan are described by optimization functions F_k, $k = 1, \ldots, K$. Each function maps the dose distribution of one ROI to a single number, which serves as a measure of the quality of the dose distribution of the ROI. One ROI can have many associated functions since it may be hard to capture the treatment goals of a ROI by a single function. The optimization functions can be partitioned into physical functions and biological functions. Physical functions are based on direct measures in the dose domain, e.g., the maximum dose should not exceed a certain dose level in a ROI, while biological functions are based on radiobiological models that predict the clinical outcome of the dose distribution, see, e.g., [17].

This thesis concerns optimization functions that are commonly used in the clinics. They all belong to the class of physical functions and are based on quadratic penalties from some prescribed dose level. The *uniform dose*, *max dose*, and *min dose* functions are given by

$$F^k(d) = \frac{1}{2} \sum_{i \in V} f(d_i, \hat{d}^k) \Delta v_i \left(\frac{d_i - \hat{d}^k}{\hat{d}^k} \right)^2, \tag{15}$$

where $f(d_i, \hat{d}^k) = \max(d_i - \hat{d}^k, 0)$ for the max dose function, $f(d_i, \hat{d}^k) = \max(\hat{d}^k - d_i, 0)$ for the min dose function and $f(d_i, \hat{d}^k) = 1$ for the uniform dose function, V specifies the voxels included in the ROI, Δv_i is the relative volume of voxel i, d_i is the dose in voxel i, and $\hat{d}^k$ is the function specific prescribed dose level. The max dose function is typically used for OARs, since only voxels with dose exceeding the prescribed dose level are penalized. Conversely, the min dose function is only used for target ROIs.

The *max dose-volume* and *min dose-volume* functions are based on fulfilling DVH requirements for the ROI, e.g., no more than x percent of the ROI should receive a dose that exceed y Gy. Dose volume functions are given by (15) with the modification that V depends on a specified volume level and the dose distribution. This is illustrated in Figure 6, where a max dose-volume function is applied to an OAR and a min dose-volume function is applied to a PTV. The crosses in the

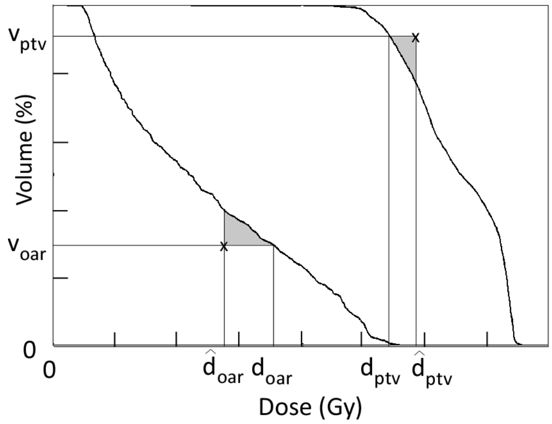

Figure 6: An illustration of a max dose-volume function to an OAR and a min dose-volume function to a PTV. The crosses specify the prescribed DVH requirements and the grey areas point out the violations of these.

figure specify the prescribed DVH requirements and the grey areas point out the violations of these. The prescribed dose levels are denoted by $\hat{d}_{oar}$ and $\hat{d}_{ptv}$, and the specified volume levels are denoted by v_{oar} and v_{ptv} for the OAR and the PTV, respectively. The voxels of the OAR included in (15) are the ones with dose between $\hat{d}_{oar}$ and d_{oar}. For the PTV, the voxels included in (15) are the ones with dose between d_{ptv} and $\hat{d}_{ptv}$. The dose-volume functions are nonconvex and not continuously differentiable [27, 64]. However, in practice, the impact of the local minimas induced by this nonconvexity on the outcome is clinically insignificant [81]. An approach similar to the one presented in [80] has been used in this thesis for handling of the dose-volume functions, where the set of voxels included in (15) is updated in every iteration.

Finally, the *max mean-dose* and *min mean-dose* functions, which are used in paper C, are given by (with the same notation as above)

$$F^k(d) = \frac{1}{2} f(\bar{d}, \hat{d}^k) \left(\frac{\bar{d} - \hat{d}^k}{\hat{d}^k} \right)^2 , \tag{16}$$

where $\bar{d} = \sum_{i \in V} \Delta v_i d_i$ is the mean dose of the ROI.

3.3 Fluence map optimization

The original IMRT optimization problem is the *fluence map optimization problem*

$$\begin{array}{ll} \underset{\tau \in \mathbb{R}^n}{\text{minimize}} & F(d(\tau)) \\ \text{subject to} & \tau \geq 0, \end{array} \tag{17}$$

where τ denotes the variables of the discretized fluence of all beams and $d(\tau) = P\tau$. This is also referred to as the *bixel-weight optimization problem.* The objective function F is composed of the optimization functions F^k, $k = 1, \ldots, K$, described in the previous section. Throughout this thesis, F is given by a weighted sum of the optimization functions, with weights reflecting the relative importance of the treatment goals. No nonlinear constraints are used. The weights often need to be adjusted a few times in order to find a solution of (17) where the trade-off between high dose to target ROIs and sparing of OARs is well-balanced. An interesting alternative to the weighted sum approach in IMRT is multi-objective optimization [51], where the K optimization functions form a K-dimensional objective function. The compromises between conflicting treatment goals can then be explored in a more intuitive manner, see, e.g., [26, 36, 42, 63] and references therein.

The structure of (17) is very similar to the structure of (14) with F as above. This means that (17) is an ill-conditioned problem, typically with a jagged optimal solution. To generate solutions of practical interest, optimization approaches applied to (17) must incorporate some regularization or smoothing strategy. Three popular regularizing strategies for ill-conditioned problems are: (i) *Tikhonov's method*, (ii) *truncated SVD*, and (iii) *iterative methods*; see [35]. The equivalence of these techniques under certain conditions is discussed in the same paper.

Tikhonov's method works by adding a stabilizing function to the objective function [74]. This method is used in [24], and it is demonstrated that adding a quadratic term based on the gradient of F to the objective function of (17) results in less jagged solutions. A method based on so-called L-curve analysis to select an appropriate weight of this term is described in [23].

Truncated SVD is based on the singular value decomposition (SVD), which, for an $m \times n$ matrix M of full rank and with $m \geq n$, is given by

$$M = \sum_{i=1}^{n} \sigma_i u_i v_i^T, \tag{18}$$

where u_i and v_i, $i = 1, \ldots, n$, are the singular vectors and the singular values σ_i are ordered so that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$. In the truncated SVD method, the right hand side of (18) is truncated to remove the terms associated with small singular values. In paper A, a variant of this regularization strategy is applied to (17) by performing an SVD of a matrix M such that $H = M^T M$, where H is the Hessian of the objective function of (17). This produces singular vectors, or eigenvectors, v_i, $i = 1, \ldots, n$, of H. An optimization problem of reduced dimension is obtained by using ξ as variables, where $\tau = V\xi$ and V consists of eigenvectors corresponding to large eigenvalues.

Iterative methods refer to using optimization methods that initially proceed in directions corresponding to the dominant singular values, e.g., a conjugate-gradient method. This approach is explored in paper B, where regular solutions to (17) are obtained by applying a quasi-Newton SQP method to solve (17). The optimization is terminated before the method proceeds in directions corresponding to small singular values.

Several other approaches for obtaining smooth solutions to (17) have been proposed. In [83], an algorithm which inherently finds smooth solutions is used. Functions different from the Tikhonov function are incorporated into the objective function in [5, 48, 70], while upper bounds on τ are added in [25]. In [70, 79], high-frequency components of the fluence are removed between iterations.

Since τ is not a treatment parameter, the solution of (17) is not deliverable. With an MLC-based delivery system, a so-called leaf sequencing step is required, where the fluence profiles are converted into feasible MLC segments such that the deliverable fluence resembles the solution of (17). For step-and-shoot IMRT plans, the leaf sequencing approaches aim at minimizing either the number of MUs or the number of segments, while resembling the original fluence to some accuracy. The latter problem is in fact NP complete [7]. There is a vast literature on leaf sequencing methods; see, e.g., [20,44,82] and references therein. There is a potential risk for plan quality degradation if the objective function of (17) is not incorporated into the process of generating segments. Some approaches address this issue by alternating between solving (17) and performing leaf sequencing; see, e.g., [6,65,69].

3.4 Step-and-shoot parameter optimization

By formulating IMRT optimization problems with the treatment parameters as optimization variables, the generated solutions correspond to deliverable treatment plans and no post-processing such as leaf sequencing is needed. Also, the ill-conditioning of the dose matrix is no longer an issue. However, this formulation is generally nonconvex even if F is convex in dose. Further, a beam model must be incorporated into the optimization problem with this formulation.

The available degrees of freedom in step-and-shoot delivery, and thus possible optimization variables, include: gantry angles, collimator (MLC) angles, couch angles, leaf positions and segment weights. At a higher level, one may also include fractionation schedule and photon energy as variables in the problem. However, to formulate a tractable optimization problem, one has to limit the choice of variables. Fixing all parameters listed above except for the leaf positions and the segment weights results in the *direct step-and-shoot optimization problem*

$$\begin{array}{ll} \underset{x,w}{\text{minimize}} & F(d(\tau(x,w))) \\ \text{subject to} & A^{(s)}x^{(s)} \geq b^{(s)}, \quad s = 1,\dots,S, \\ & w \geq w_0, \end{array} \tag{19}$$

where $x^{(s)}$ denotes the leaf position variables for segment s, $w \in I\!R^S$ denotes the segment weight variables for all S segments, $d(\tau) = P\tau$, $\tau = (\tau_1^T \ \dots \ \tau_B^T)^T$, and

$$\tau_b(x, w) = \sum_{s \in S_b} w_s \tau(x^{(s)}), \quad b = 1, \dots, B, \tag{20}$$

where B is the number of beams in the plan, S_b specifies the segments of beam b and $\tau(x^{(s)})$ is the transmitted fluence distribution of segment s. The bounds on w are included to ensure that all segments fulfill their lower monitor unit limit in order to avoid segments with very short beam-on-time. The linear constraints represent MLC requirements such as interdigitation, minimum gaps and minimum segment areas; see Figure 7 for an illustration.

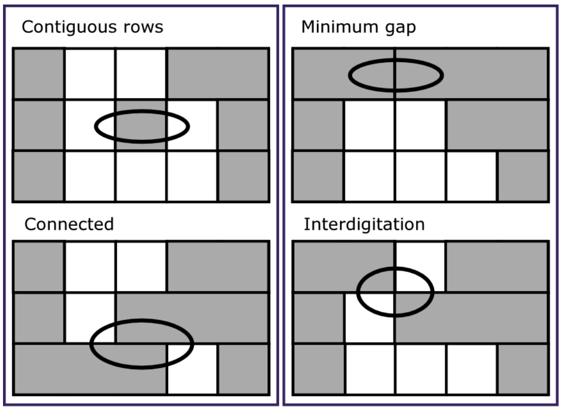

Figure 7: An illustration of four common requirements on MLCs, highlighted with ovals. The contiguous rows requirement must always be fulfilled, while the other three may or may not need to be fulfilled depending on MLC type. Grey areas correspond to leaves and white areas correspond to openings.

The computation of the fluence distribution $\tau(x^{(s)})$ is based on integration of the intensity distributions of the primary source and the flattening filter. Assuming Gaussian intensity distributions of both these results in a fluence distribution described by a combination of error functions [29]. Figure 8 illustrates the transmitted fluence distribution in one dimension with a leaf pair intersecting the beam. Clearly, $\tau(x^{(s)})$ is a nonconvex function.

Many optimization approaches to (19) start with a set of predefined segments specifying the number of segments, their distribution over the beams and the set of leaves included in the optimization problem. The variable sets w_s and $x^{(s)}$, $s = 1, \dots, S$, of (19) are thus fixed throughout the optimization. In many of these approaches, the initial segments are based on projections of ROIs onto the fluence planes of the beams. Another strategy for generating initial segments is to

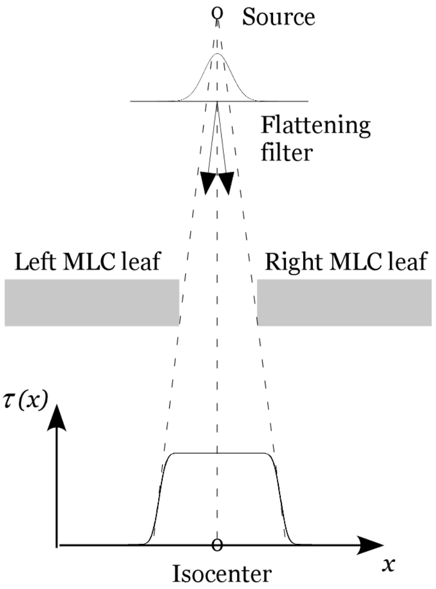

Figure 8: An illustration of the transmitted fluence distribution with one intersecting leaf pair. The computation of $\tau(x)$ is based on integration of two Gaussian intensity distributions, originating from the source and the flattening filter.

solve the fluence map optimization problem (17) approximately and then perform leaf sequencing. Since (19) is a nonconvex optimization problem, one must either utilize global optimization methods or rely on the initial set of segments and their distribution being sufficiently good to reach high-quality solutions. Approaches using global stochastic optimization methods for solving (19) are found in [10,67,73], while two approaches more based on heuristics are discussed in [9, 31].

An approach for solving (19) that dynamically alters the variable set was introduced in [62]. This approach has two main advantages compared to the previously mentioned approaches: (i) The nonconvexity induced by the leaf position variables can be removed and (ii) the set of segments is not fixed. This gives an opportunity to study the impact of adding segments on the plan quality and, more generally, the relation between plan quality and delivery time. Papers C and D are both inspired by this approach, which uses an optimization method called column generation. Other approaches using column generation for generating deliverable step-and-shoot plans are presented in [50, 58].

The idea of column generation applied to (19) is to start with few or no segments and then only generate segments that have potential to improve the objective function value. Consider a pool of segments where all feasible segment shapes for all beams are included such that the leaves are aligned with the bixel grid and the leaf positions are fixed to the bixel boundaries. The working set $\mathcal{W}$ specifies the segments generated, or picked from this pool, during the optimization process. The column generation approach proceeds by alternating between solving a *master problem* and a *subproblem*, where the master problem is given by

$$(MASTER) \qquad \begin{array}{lll} \underset{w}{\text{minimize}} & F(d(\tau(x,w))) & \\ \text{subject to} & w_i \geq w_0 & i \in \mathcal{W}, \\ & w_i = 0 & i \notin \mathcal{W}, \\ & x \text{ fixed}, & \end{array} \tag{21}$$

which is a convex problem if F is convex in d (since d is linear in w). The role of (21) is to optimize the segment weights of the segments included in the working set. Problem (21) may be viewed as a restricted version of (19) with $S = |\mathcal{W}|$ and with the leaf positions fixed.

Since the leaf positions are fixed to the bixel boundaries, one may view each segment as a set of exposed bixels. For each beam, the solution of the subproblem corresponds to the most promising segment, in terms of the gradient of F with respect to the exposed bixels, not yet included in $\mathcal{W}$. The subproblem for beam b is given by

$$(SUB) \qquad \begin{array}{ll} \underset{z}{\text{minimize}} & \left(\frac{\partial F}{\partial \tau_b}\right)^T z \\ \text{subject to} & z \in \mathcal{Z}, \\ & z \in \{0,1\}, \end{array} \tag{22}$$

where $\mathcal{Z}$ is the set of bixel regions corresponding to feasible segments with respect to the MLC used. Such a bixel region is represented by a binary vector, where zero components correspond to bixels covered by a leaf while the components with value one correspond to exposed bixels. The solutions of (22) are easily transformed into MLC segments by placing the leaves such that all zeros in the solution vector are covered. After solving (22) for each beam, some or all of the corresponding segments are included in $\mathcal{W}$ and (21) is solved again, using the previous solution as starting point. The solution process proceeds until the user is satisfied with the plan quality or until no solutions to (22) can be found such that the optimal value of (22) is negative.

The strategy for solving (22) depends on the MLC requirements. If the requirements are separable in leaf pairs, i.e., if the only requirement is to have contiguous rows (see Figure 7), an algorithm presented in [62] solves (22) efficiently. If the MLC does not support interdigitation or requires a connected opening, the subproblem cannot be separated in bixel rows (leaf pairs). However, the subproblem of each beam can be formulated as a shortest-path problem that incorporates all of the requirements illustrated in Figure 7 [12, 62]. The shortest-path problem is to find a path between a certain pair of nodes in a graph such that the sum of the weights of its constituent arcs is minimized; see, e.g., [57] for an introduction. For each beam, a layered graph is constructed where each layer corresponds to a bixel row. Each node represents a leaf pair configuration and the weights of all arcs incident on a node is given by the sum of the components of the gradient of F with respect to the exposed bixels for that leaf pair configuration. In paper C, the problem is modified by scaling the weight of each arc with a factor based on the relative overlap of the exposed bixels for the two nodes of the arc. By doing this, arcs that may lead to jagged segment shapes can be avoided.

A drawback of the column generation approach compared to the approaches solving (19) directly is that the leaf positions are fixed to the bixel boundaries. This can be overcome by combining the column generation approach with direct step-and-shoot optimization to fine-tune the leaf positions after the solution of every master problem. This is the idea of paper C; see Figure C.1 for an illustration of the solution process. In that paper, both (19) and (21) are solved with a quasi-Newton SQP method developed at RaySearch.

4 Main contributions

Although iterative regularization is widely known in the field of inverse problems, it was first introduced in the context of IMRT optimization in paper B. The results of that paper clearly demonstrate the suitability of a quasi-Newton SQP method for performing iterative regularization. The paper also provides an explanation of the efficiency of this optimization method on IMRT problems, which is based on the numerical behaviour of the conjugate-gradient method on ill-conditioned problems. Since this optimization method is widely used clinically in a very similar way to the setup in paper B, an important message of the paper is to avoid over-optimizing the treatment plans prior to leaf sequencing.

Generating high-quality step-and-shoot treatment plans with few and regular segments is a challenge, and the number of required segments varies from case to case depending on the patient geometry and the choice of optimization functions. The approach in paper C provides a method that gives support in exploring the trade-off between plan quality and treatment complexity. The novelty of the method is the combination of the flexibility of dynamically altering the set of segments with the ability to fine-tune the segment shapes. The method generates a sequence of deliverable plans while being capable of finding satisfactory treatment plans with few segments. Column generation approaches to step-and-shoot IMRT tend to find near-optimal solutions with very few segments compared to the problem dimension. This behaviour is, to some extent, explained in paper D by interpreting the conjugate-gradient method as a special case of a column generation method.

5 Summary of the appended papers

Of the four papers included in this thesis, the first two papers focus on methods for solving the fluence map optimization problem efficiently while avoiding jagged solutions. The last two papers deal with a column generation approach for generating segments dynamically when optimizing step-and-shoot parameters.

Paper A: Using eigenstructure of the Hessian to reduce the dimension of the intensity modulated radiation therapy optimization problem

Paper A is co-authored with Anders Forsgren, Henrik Rehbinder and Kjell Eriksson, and has been published in *Annals of Operations Research*, Vol. 148, pp. 81-94, 2006.

The effect of reducing the dimension of the fluence map optimization problem through a spectral decomposition of an approximation of the Hessian of the objective function is studied in this paper. An optimization problem with lower dimension is formulated by introducing eigenvector weights as optimization variables, where only eigenvectors corresponding to large eigenvalues are included. The approach is evaluated on a prostate case by applying a quasi-Newton SQP method to a suite of problems, where the number of included eigenvectors is varied.

Optimization of a few eigenvector weights results in a faster initial decrease of the objective value, but with an inferior solution after 25 iterations, compared to optimization of bixel weights. By combining eigenvector weights and bixel weights as variables, a lower objective value is obtained after 25 iterations. However, this advantage comes at the expense of the pre-computational time for the spectral decomposition.

Paper B: Iterative regularization in intensity-modulated radiation therapy optimization

Paper B is co-authored with Anders Forsgren, and has been published in *Medical Physics*, Vol. 33(1), pp. 225-234, 2006.

The suitability of using a quasi-Newton SQP method for performing iterative regularization of fluence map optimization problems is demonstrated in this paper. This is done by comparing the treatment quality of deliverable step-and-shoot plans, generated through leaf sequencing with a fixed number of segments, for different number of bixel-weight iterations.

Numerical results for ten IMRT problems show that the SQP method with diagonal initial Hessian estimate fulfills the requirements for performing iterative regularization; it initially proceeds in directions corresponding to smooth fluence profiles and finds high-quality solutions in few iterations. The deliverable plans obtained after 35 iterations of fluence map optimization and a leaf sequencing step outperform the deliverable plans obtained if 100 bixel-weight iterations are performed instead. It is concluded that performing too many bixel-weight iterations deteriorates the quality of the deliverable plan.

Paper C: Combining segment generation with direct step-and-shoot optimization in intensity-modulated radiation therapy

Paper C has been submitted to *Medical Physics*. Part of the material has been published in the *ICCR 2007 Conference Proceedings*.

In this paper, a method that combines generation of new segments with the optimization of segment shapes and weights is presented. The method may be viewed either as (i) a generalization of direct step-and-shoot optimization methods by dynamically altering the set of optimization variables or as (ii) an extension of a column generation approach to step-and-shoot problems by fine-tuning the leaf positions of the generated segments.

The method is evaluated on a test suite consisting of ten cases and it is found that the adjustment of leaf positions improves the plan quality. The improvement in plan quality when adding segments is larger for plans with few segments. Eventually, adding more segments contributes very little to the plan quality. The method provides a tool for controlling the number of segments and, indirectly, the delivery time. The generated sequence of deliverable plans can thus support the planner in finding a sound trade-off between plan quality and treatment complexity.

Paper D: A conjugate-gradient based approach for approximate solutions of quadratic programs

Paper D is co-authored with Anders Forsgren, and has been submitted to *Annals of Operations Research*.

An attempt to explain the promising numerical results obtained with the column generation approaches of paper C and [50, 62] on step-and-shoot IMRT problems is carried out in this paper. The impact of different restrictions on the generated columns of a column generation method is studied, both in terms of numerical behaviour and convergence properties. It is noted that a bound on the two-norm of the columns results in that the column generation method is equivalent to the conjugate-gradient method. The column generation approach for IMRT is obtained by employing a restriction based on the infinity-norm and non-negativity.

The column generation method has weak worst-case convergence properties if restricted to generating feasible step-and-shoot plans. However, the numerical results of three IMRT QPs indicate that the appealing properties of the conjugate-gradient method on ill-conditioned problems are inherited in the column generation approach for IMRT; as observed in paper C, near-optimal solutions are found in very few iterations compared to the problem dimension.

6 References

[1] A. Ahnesjö. Collapsed cone convolution of radiant energy for photon dose calculation in heterogeneous media. *Medical Physics*, 16(4):577–592, 1989.

[2] A. Ahnesjö and M. M. Aspradakis. Dose calculations for external photon beams in radiotherapy. *Physics in Medicine and Biology*, 44(11):R99–R155, 1999.

[3] A. Ahnesjö, B. Hårdemark, U. Isacsson, and A. Montelius. The IMRT information process—mastering the degrees of freedom in external beam therapy. *Physics in Medicine and Biology*, 51(13):R381–R402, 2006.

[4] M. Alber, G. Meedt, F. Nüsslin, and R. Reemtsen. On the degeneracy of the IMRT optimization problem. *Medical Physics*, 29(11):2584–2589, 2002.

[5] M. Alber and F. Nüsslin. Intensity modulated photon beams subject to a minimal surface smoothing constraint. *Physics in Medicine and Biology*, 45(5):N49–N52, 2000.

[6] M. Alber and F. Nüsslin. Optimization of intensity modulated radiotherapy under constraints for static and dynamic MLC delivery. *Physics in Medicine and Biology*, 46(12):3229–3239, 2001.

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

[8] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty. *Nonlinear Programming: Theory and Algorithms.* John Wiley and Sons, New York, second edition, 1993. ISBN 0-471-55793-5.

[9] J. L. Bedford and S. Webb. Constrained segment shapes in direct-aperture optimization for step-and-shoot IMRT. *Medical Physics*, 33(4):944–958, 2006.

[10] A. M. Bergman, K. Bush, M.-P. Milette, I. A. Popescu, K. Otto, and C. Duzenli. Direct aperture optimization for IMRT using Monte Carlo generated beamlets. *Medical Physics*, 33(10):3666–3679, 2006.

[11] M. Birkner, D. Yan, M. Alber, J. Liang, and F. Nüsslin. Adapting inverse planning to patient and organ geometrical variation: algorithm and implementation. *Medical Physics*, 30(10):2822–2831, 2003.

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

[13] T. Bortfeld. IMRT: a review and preview. *Physics in Medicine and Biology*, 51(13):R363–R379, 2006.

[14] T. Bortfeld, J. Bürkelbach, R. Boesecke, and W. Schlegel. Methods of image reconstruction from projections applied to conformation radiotherapy. *Physics in Medicine and Biology*, 35(10):1423–1434, 1990.

[15] T. Bortfeld, W. Schlegel, and B. Rhein. Decomposition of pencil beam kernels for fast dose calculations in three-dimensional treatment planning. *Medical Physics*, 20(2):311–318, 1993.

[16] A. Brahme. Optimization of stationary and moving beam radiation therapy. *Radiotherapy Oncology*, 12:129–140, 1988.

[17] A. Brahme. Optimized radiation therapy based on radiobiological objectives. *Seminars in radiation oncology*, 9(1):35–47, 1999.

[18] A. Brahme, J. E. Roos, and I. Lax. Solution of an integral equation encountered in rotation therapy. *Physics in Medicine and Biology*, 27(10):1221–1229, 1982.

[19] M. K. Bucci, A. Bevan, and M. Roach. Advances in Radiation Therapy: Conventional to 3D, to IMRT, to 4D, and Beyond. *CA Cancer J Clin*, 55(2):117–134, 2005.

[20] D. Cao, M. A. Earl, S. Luan, and D. M. Shepard. Continuous intensity map optimization (CIMO): A novel approach to leaf sequencing in step and shoot IMRT. *Medical Physics*, 33(4):859–867, 2006.

[21] T. C. Y. Chan, T. Bortfeld, and J. N. Tsitsiklis. A robust approach to IMRT optimization. *Physics in Medicine and Biology*, 51(10):2567–2583, 2006.

[22] C. W. Cheng, I. J. Das, and M. S. Huq. Lateral loss and dose discrepancies of multileaf collimator segments in intensity modulated radiation therapy. *Medical Physics*, 30(11):2959–2968, 2003.

[23] A. V. Chvetsov. L-curve analysis of radiotherapy optimization problems. *Medical Physics*, 32(8):2598–2605, 2005.

[24] A. V. Chvetsov, D. Calvetti, J. W. Sohn, and T. J. Kinsella. Regularization of inverse planning for intensity-modulated radiotherapy. *Medical Physics*, 32(2):501–514, 2005.

[25] M. M. Coselmon, J. M. Moran, J. D. Radawski, and B. A. Fraass. Improving IMRT delivery efficiency using intensity limits during inverse planning. *Medical Physics*, 32(5):1234–1245, 2005.

[26] D. L. Craft, T. F. Halabi, H. A. Shih, and T. R. Bortfeld. Approximating convex pareto surfaces in multiobjective radiotherapy planning. *Medical Physics*, 33(9):3399–3407, 2006.

[27] J. O. Deasy. Multiple local minima in radiotherapy optimization problems with dose-volume constraints. *Medical Physics*, 24(7):1157–1161, 1997.

[28] J. Fenwick, S. Riley, and A. Scott. *Radiation Oncology Advances*, chapter Advances in Intensity-Modulated Radiotherapy Delivery. Springer US, 2008.

[29] M. Fippel, F. Haryanto, O. Dohm, F. Nüsslin, and S. Kriesen. A virtual photon energy fluence model for Monte Carlo dose calculation. *Medical Physics*, 30(3):301–311, 2003.

[30] J. Galvin. The multileaf collimator: a complete guide. In *Proc. AAPM Annual Meeting*, 1999.

[31] W. D. Gersem, F. Claus, C. D. Wagter, B. V. Duyse, and W. D. Neve. Leaf position optimization for step-and-shoot IMRT. *International Journal of Radiation Oncology, Biology, Physics*, 51(5):1371–1388, 2001.

[32] P. E. Gill, W. Murray, and M. A. Saunders. SNOPT: An SQP algorithm for large-scale constrained optimization. Numerical Analysis Report 97-2, Department of Mathematics, University of California, San Diego, La Jolla, CA, 1997.

[33] P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright. User's guide for NPSOL (version 4.0): A Fortran package for nonlinear programming. Stanford University SOL 86-2, 1986.

[34] A. Gustafsson, B. K. Lind, and A. Brahme. A generalized pencil beam algorithm for optimization of radiation therapy. *Medical Physics*, 21(3):343–356, 1994.

[35] P. C. Hansen. Numerical tools for analysis and solution of Fredholm integral equations of the first kind. *Inverse Problems*, 8(6):849–872, 1992.

[36] A. L. Hoffmann, A. Y. D. Siem, D. den Hertog, J. H. A. M. Kaanders, and H. Huizenga. Derivative-free generation and interpolation of convex pareto optimal IMRT plans. *Physics in Medicine and Biology*, 51(24):6349–6369, 2006.

[37] IMV. *2006 Radiation Oncology Market Summary Report*. IMV Medical Information Division Inc., www.imvinfo.com, 2007.

[38] Intensity Modulated Radiation Therapy Collaborative Working Group. Intensity-modulated radiotherapy: Current status and issues of interest. *International Journal of Radiation Oncology, Biology, Physics*, 51(4):880–914, 2001.

[39] U. Jelen and M. Alber. A finite size pencil beam algorithm for IMRT dose optimization: density corrections. *Physics in Medicine and Biology*, 52:617–633(17), 2007.

[40] W. Karush. Minima of functions of several variables with inequalities as side constraints. Master's thesis, Department of Mathematics, University of Chicago, 1939.

[41] H. W. Kuhn and A. W. Tucker. Nonlinear programming. In J. Neyman, editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492, Berkeley, 1951. University of California Press.

[42] K.-H. Küfer, A. Scherrer, M. Monz, F. Alonso, H. Trinkaus, T. Bortfeld, and C. Thieke. Intensity-modulated radiotherapy - a large scale multi-criteria programming problem -. *OR Spectrum*, 25:223–249, 2003.

[43] E. Lee, T. Fox, and I. Crocker. Integer programming applied to intensity-modulated radiation therapy treatment planning. *Annals of Operations Research*, 119:165–181(17), March 2003.

[44] F. Lenzen. An integer programming approach to the multileaf collimator problem. Master's thesis, University of Kaiserslautern, 2000.

[45] B. K. Lind. Properties of an algorithm for solving the inverse problem in radiation therapy. *Inverse Problems*, 6(3):415–426, 1990.

[46] J. Löf. *Development of a general framework for optimization of radiation therapy.* PhD thesis, 2000, Stockholm University.

[47] J. Löf, B. K. Lind, and A. Brahme. An adaptive control algorithm for optimization of intensity modulated radiotherapy considering uncertainties in beam profiles, patient set-up and internal organ motion. *Physics in Medicine and Biology*, 43:1605–1628, 1998.

[48] M. M. Matuszak, E. W. Larsen, and B. A. Fraass. Reduction of IMRT beam complexity through the use of beam modulation penalties in the objective function. *Medical Physics*, 34(2):507–520, 2007.

[49] L. K. Mell, A. K. Mehrotra, and A. J. Mundt. Intensity-modulated radiation therapy use in the U.S., 2004. *Cancer*, 104(6):1296–1303, 2005.

[50] C. Men, H. E. Romeijn, Z. C. Taskin, and J. F. Dempsey. An exact approach to direct aperture optimization in IMRT treatment planning. *Physics in Medicine and Biology*, 52(24):7333–7352, 2007.

[51] K. Miettinen. *Nonlinear Multiobjective Optimization*. Springer, 1999.

[52] R. Mohan, M. Arnfield, S. Tong, Q. Wu, and J. Siebers. The impact of fluctuations in intensity patterns on the number of monitor units and the quality and accuracy of intensity modulated radiotherapy. *Medical Physics*, 27(6):1226–1237, 2000.

[53] L. Nazareth. A relationship between the BFGS and conjugate gradient algorithms and its implications for new algorithms. *SIAM Journal on Numerical Analysis*, 16(5):794–800, 1979.

[54] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, 1999. ISBN 0-387-98793-2.

[55] A. Olafsson and S. J. Wright. Efficient schemes for robust IMRT treatment planning. *Physics in Medicine and Biology*, 51(21):5621–5642, 2006.

[56] K. Otto. Volumetric modulated arc therapy: IMRT in a single gantry arc. *Medical Physics*, 35(1):310–317, 2008.

[57] C. Papadimitriou and K. Steiglitz. *Combinatorial Optimization*. Dover, New York, 1998. ISBN 0-486-40258-4.

[58] F. Preciado-Walters, M. P. Langer, R. L. Rardin, and V. Thai. Column generation for IMRT cancer therapy optimization with implementable segments. *Annals of Operations Research*, 148:65–79, 2006.

[59] H. Rehbinder, C. Forsgren, and J. Löf. Adaptive radiation therapy for compensation of errors in patient setup and treatment delivery. *Medical Physics*, 31(12):3363–3371, 2004.

[60] D. J. Reviewer. ICRU report 50 - Prescribing, recording and reporting photon beam therapy. *Medical Physics*, 21(6):833–834, 1994.

[61] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, New Jersey, 1970. ISBN 0-691-08069-0.

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

[63] H. E. Romeijn, J. F. Dempsey, and J. G. Li. A unifying framework for multi-criteria fluence map optimization models. *Physics in Medicine and Biology*, 49(10):1991–2013, 2004.

[64] C. G. Rowbottom and S. Webb. Configuration space analysis of common cost functions in radiotherapy beam-weight optimization algorithms. *Physics in Medicine and Biology*, 47(1):65–77, 2002.

[65] J. Seco, P. M. Evans, and S. Webb. An optimization algorithm that incorporates IMRT delivery constraints. *Physics in Medicine and Biology*, 47(6):899–915, 2002.

[66] M. B. Sharpe, B. M. Miller, D. Yan, and J. W. Wong. Monitor unit settings for intensity modulated beams delivered using a step-and-shoot approach. *Medical Physics*, 27(12):2719–2725, 2000.

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

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

[69] J. V. Siebers, M. Lauterbach, P. J. Keall, and R. Mohan. Incorporating multileaf collimator leaf sequencing into iterative IMRT optimization. *Medical Physics*, 29(6):952–959, 2002.

[70] S. V. Spirou, N. Fournier-Bidoz, J. Yang, C. Chui, and C. C. Ling. Smoothing intensity-modulated beam profiles to improve the efficiency of delivery. *Medical Physics*, 28:2105–2112, October 2001.

[71] G. G. Steel. *Basic Clinical Radiobiology.* Hodder Arnold, third edition, 2002. ISBN 0-340-80783-0.

[72] P. Suetens. *Fundamentals of Medical Imaging.* Cambridge University Press, 2002.

[73] J. Tervo, P. Kolmonen, T. Lyyra-Laitinen, J. Pintér, and T. Lahtinen. An optimization-based approach to the multiple static delivery technique in radiation therapy. *Annals of Operations Research*, 119:205–227, 2003.

[74] A. N. Tikhonov. Regularization of incorrectly posed problems. *Soviet Mathematics, Translation of Doklady Akademii Nauk SSSR (American Mathematical Society, Providence)*, 4:1624–1627, 1963.

[75] M. van Herk. Errors and margins in radiotherapy. *Seminars in radiation oncology*, 14(1):52–64, 2004.

[76] S. Webb. Optimization by simulated annealing of three-dimensional conformal treatment planning for radiation fields defined by a multileaf collimator. *Physics in Medicine and Biology*, 36:1201–1226, 1991.

[77] S. Webb. The physical basis of IMRT and inverse planning. *The British Journal of Radiology*, 76(910):678–689, 2003.

[78] S. Webb. Motion effects in (intensity modulated) radiation therapy: a review. *Physics in Medicine and Biology*, 51(13):R403–R425, 2006.

[79] S. Webb, D. J. Convery, and P. M. Evans. Inverse planning with constraints to generate smoothed intensity-modulated beams. *Physics in Medicine and Biology*, 43(10):2785–2794, 1998.

[80] Q. Wu and R. Mohan. Algorithms and functionality of an intensity modulated radiotherapy optimization system. *Medical Physics*, 27(4):701–711, 2000.

[81] Q. Wu and R. Mohan. Multiple local minima in IMRT optimization based on dose-volume criteria. *Medical Physics*, 29(7):1514–1527, 2002.

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

[83] Y. Xiao, D. Michalski, Y. Censor, and J. M. Galvin. Inherent smoothness of intensity patterns for intensity modulated radiation therapy generated by simultaneous projection algorithms. *Physics in Medicine and Biology*, 49(14):3227–3245, 2004.