# A Concept for the Optimization of Radiotherapy

Dissertation

zur Erlangung des Grades eines Doktors

der Naturwissenschaften

der Fakultät für Physik

der Eberhard-Karls-Universität zu Tübingen

vorgelegt von

Markus Lothar Alber

aus Hof/Saale

2000

Tag der mündlichen Prüfung: 11. Oktober 2000

| | |
|---|---|
| Dekan: | Prof. Dr. G. Wagner |
| 1. Berichterstatter: | Prof. Dr. F. Nüsslin |
| 2. Berichterstatter: | Prof. Dr. H. Müther |
| 3. Berichterstatter: | Prof. Dr. W. Schlegel |

# Contents

# Chapter 1

# Introduction

In the fight against cancer, radiotherapy is one of three weapons. It is applied together with surgery and chemotherapy, or as a single modality. With a significant fraction of cancer deaths associated with the failure to control the primary tumour, enhancing the effectiveness of radiotherapy is a worthy goal. A recent study on prostate cancer produced unambiguous evidence for the benefit of pushing the technical limits of conventional radiotherapy with photons to achieve a higher therapeutic dose. The concomitant development of ever faster computers, powerful imaging methods and sophisticated treatment units has provided the means to overcome many long-standing limitations of radiotherapy.

The key stimulus for physicists to instigate renewed activities in radiotherapy optimization was the concept of modulated as opposed to homogeneous radiation intensity (intensity modulated radiotherapy, IMRT) [1, 2]. As the limitations of conventional therapy techniques with multiple homogeneous photon beams are removed with this new technique, established methods of treatment planning become unavailing. Novel biological and physical models for optimizing radiotherapy have to be conceived to keep up with the pace of the technical and clinical development, which is in turn driven by novel software solutions. As a consequence of this mutual stimulation, treatment planning is about to undergo a metamorphosis to computer based treatment simulation embracing the fields of physics, biology and medical sciences.

The invention of IMRT followed an analogy to image reconstruction in x-ray tomography. Starting from a *definition of the desired dose distribution*, the modulated fluence distribution was obtained by a formal inversion of the calculus of tomographic image reconstruction. Whilst the latter is a well defined problem in that the solution (the density distribution of the image object) certainly exists, the solution of the former (the fluence distribution which creates a certain dose distribution) may very well not exist.

Over the years, the development moved ever further away from the inverse problem approach towards the concept of optimization. This evolution is expressed in efforts to formulate rules for deviations from the dose prescription in case it is unattainable. Simultaneously, IMRT was increasingly understood as a chance to escalate tumour doses beyond conventional limits. The majority of methods requires the prescription of a dose

for the target (tumour) volume and a tolerance dose for a number of normal tissue volumes (c.f. [3, 4, 5]). Very soon it became clear that at least some biological considerations had to be included into the definition of both dose prescription and rules for its violation [6, 7, 8, 9, 10, 11]. By today, prescriptions for partial volumes (dose-volume or dose-volume histogram constraints) are regarded as standard [12, 13, 14, 15, 16].

The common feature of all dose-based approaches is that the optimum solution is defined by the therapists with the specification of the prescription dose. However, the clinical guidelines which govern conventional techniques may not apply equally well to IMRT. The common understanding of an optimum dose distribution is shaped by the available means. The hugely greater flexibility of IMRT requires a new definition of optimality for radiotherapy. Undoubtedly, IMRT has the potential to improve on the cure rate of established dose prescriptions, yet there is also the danger of unwanted side effects. For a safe advancement of treatment, the modelling of normal tissue reactions to radiation in the optimization process is crucial.

While dose-based optimization constitutes an attempt to bridge the gap between the desired and the feasible dose distribution, the concept of *evidence-based biological optimization* as introduced here aims to incorporate the biological knowledge and clinical evidence of conventional radiotherapy to explore the potential of IMRT, yet stay on safe ground. Consequentially, normal tissues move into the focus of the optimization concept.

Earlier attempts at biological optimization placed less emphasis on clinical aspects and met with controversity (c.f. [6, 17, 18, 19, 20]) - the treatment objectives had been specified in a less stringent form and by unspecific models. Nevertheless, this development draws great inspiration from these earlier sources.

We understand evidence-based biological optimization as the inversion of the traditional planning scheme. Instead of conflicting prescriptions for therapeutic dose and dose tolerance, the *maximum possible therapeutic dose* obtains as the result of *rigorously limited normal tissue tolerance doses*. While the risk of side effects can be expressed in the language of traditional clinical experience, the full potential of dose escalated treatments can be explored without ad-hoc restrictions of the target prescription dose. This necessitates the development of biological models which impose the rules according to which the dose distribution is optimized. These rules are applied implicitly by therapists when prescribing the dose and dealing with the feasibility gap of the dose prescription.

From the point of view of biological optimization of radiotherapy, intensity modulation is a multiplication of the degrees of freedom of the problem rather than a new class of problem. With each degree of freedom comes a number of restrictions and problems of various nature which have to be taken into account to maintain maximum clinical utility of the algorithm. The algorithm presented here constitutes an advance in three key issues:

- **Evidence-based biological optimization:** For radiotherapy optimization, biological models have to be employed which pertain to two classes of effects: the response of tissues to the dose per treatment fraction and the response to inhomogeneous dose distributions. While tumour tissue response is fairly straightforward, the dose-response of normal tissues is very involved. The modelling discriminates the tissue specific variability of the relation between irradiated volume and dose tolerance.

- **Monte Carlo dose computation:** The modelling of radiation transport through complex geometries is a central problem of IMRT. Field geometries are much more irregular and smaller than in standard radiotherapy. IMRT has the potential to generate dose distributions with accuracies of about one millimeter in very inhomogeneous regions of the body like the head and neck or the thorax. With smaller field sizes, the modelling of scatter from the collimators or compensator filters becomes more important. These effects can be modelled precisely with Monte Carlo methods. The simulation of radiation transport with these methods imitates the physical processes at the price of significantly longer computation times. Nevertheless, Monte Carlo dose computation was included into the algorithm with clinically acceptable computation times.

- **Clinical utility factors:** A radiotherapy optimization algorithm can facilitate clinical routine in two ways: treatment planning can become more intuitive, faster and more standardized, and the dose application can become more practical, error tolerant and verifyable. The method of treatment prescription was designed to accomodate a data base for class solutions providing biological and clinical parameters. The algorithm delivers technically feasible fluence distributions. The complexity of the treatment is reduced to the least possible extent.

# Chapter 2

# Mathematical Modelling

For a non-linear optimization problem of the size of radiotherapy optimization, fortuitious circumstances must come to aid the modelling. In the present development, this is the linearity of the Boltzmann radiation transport equation which links incident energy fluence to dose in the patient. The entire formalism relies on an adaption of the concept of Green's functions to the problem in hand. These 'rays' lie right at the foundation of the formalism in the first section. The connection to the formalism of variational problems will be made there which delineates the further development.

The second section deals with the simplification of the biological dose response by virtue of a mean-field approximation which is expressed by the notion of objective measures. In combination with the ray formalism this yields the concept of ray derivatives. This latter concept is used to derive a powerful theorem in section three. This theorem identifies the global solution of the radiotherapy optimization problem with an equilibrium condition of the Lagrange density introduced in section two.

## 2.1 Definitions and Terminology

### 2.1.1 Fluence and Dose Spaces

The central quantity of radiotherapy certainly is the *energy dose* $D(x, y, z)$ at some point $P = (x, y, z) \in \mathbb{R}^3$ in the patient volume. We often use the term *dose distribution* if we want to highlight the character of dose as a three-dimensional non-negative scalar field

$$D : \mathbb{R}^3 \to \mathbb{R}^+, \quad (x, y, z) \to d \quad . \tag{2.1}$$

When we refer to *dose space* $\mathcal{D}$, we think of the space of all dose distributions $D$, which can be chosen as a subset of $\mathcal{L}_2(\mathbb{R}^3)$ since the support of $D$ is finite, yet dependent on the actual patient.

The dose is a result of the energy flux due to particles like photons, electrons, positrons, hadrons (or more exotic particles) through the patient. The incident *particle fluence* $\Phi(\nu, E, \phi, \theta, u, v)$ is defined on tangent planes to the unit sphere centered at the iso-centre

which is considered to enclose the entire support of $\mathcal{D}$. It is a function of particle type $\nu$ and energy $E$, the angles $\phi \in [0, 2\pi[$ and $\theta \in [0, \pi]$ and the position in the tangent plane $(u, v) \in \mathbb{R}^2$. With this definition, $(u, v)$ is the offset of a parallel line to the radius (central ray) impinging from solid angle $\Omega = (\phi, \theta)$. By using the term *fluence distribution* we highlight the character of fluence as a five-dimensional scalar field for each particle quality $\nu$

$$\Phi : \mathbb{R}^+ \times S_2 \times \mathbb{R}^2 \to [0, \Phi_{max}], \qquad (E, \phi, \theta, u, v) \to \Phi \quad . \tag{2.2}$$

In the following, we will also often refer to *fluence profiles* which are fluence distributions on a certain tangent plane and often with a certain energy spectrum and particle quality. With *fluence space* $\mathcal{F}$ we have in mind the space of all fluence distributions which can also be chosen as a subset of $\mathcal{L}_2(\mathbb{R}^+ \times S_2 \times \mathbb{R}^2)$.

### 2.1.2 The Absorption Operator

The link between fluence and dose space is mediated by the 'energy absorption per mass and fluence unit' operator $T$ which is the local energy dissipation of the Green's function of the Boltzmann transport equation for the given patient. We define

$$T : \mathcal{F} \to \mathcal{D}, \qquad \Phi \to T\Phi = D \quad . \tag{2.3}$$

As a consequence of the linearity of the Boltzmann equation, $T$ is a linear operator, and one can also assume that all dose distributions are continuous since they are a solution of the Boltzmann equation. The latter statement highlights the fundamental importance of the feasibility gap for the inverse problem: the dose cannot assume different values on adjacent points in the patient volume which would be necessary for undiscerning dose prescriptions, even if negative fluences were allowed.

The most limiting factor of the radiotherapy optimization problem in this framework becomes already apparent. Both fluence and dose space lack an inverse element with respect to addition, and hence cannot support groups. Even if one cannot establish a vector space structure of both spaces, it is worthwhile to introduce the notion of a basis in fluence space. Since the time-invariant Boltzmann equation is a linear homogeneous partial differential equation of first order, it can be inferred from the theorem of Picard-Lindelöf that the corresponding operator $T$ is injective, i.e. even if negative fluences are permitted, a dose distribution which is zero everywhere can only be generated by a zero fluence. It is, however, not surjective so that the image of $\mathcal{F}$ in $\mathcal{D}$ is sparse. While $T$ could be inverted in principle, the origin $T^{-1}D$ of any given dose distribution $D$ will almost certainly not lie in $\mathcal{F}$.

The true value of the injectiveness of $T$ lies in the fact, that any basis of $\mathcal{F}$ is also a basis of the image of $\mathcal{F}$ in $\mathcal{D}$ and is not redundant. While one can thus freely go from $\mathcal{F}$ to $\mathcal{D}$, the opposite direction is hampered by the incompleteness of fluence space with respect to $T^{-1}$. It is for this reason that the central position of dose is abandoned and the foundations of the development laid in fluence space.

### 2.1.3 The Ray Basis

The question of a suitable set of basis functions for fluence space is related to physical and practical issues. Commonly, basis functions are designed for some special purpose, like spherical harmonics. Often, basis functions are chosen to be orthogonal with respect to some metric. In this case there is no a priori metric in fluence space, but one could use a metric in dose space via $T$. However, as will be shown, this clinically relevant metric on dose space does not exist.

To the best of common knowledge, two dose distributions have to be considered biologically equivalent if they have equal dose-volume statistics for a homogeneous target volume. If a metric were to reflect this fact, an infinite number of different dose distributions would have no 'distance' from each other, in contradiction with the axiome of definiteness[1]. Although it would be possible to restrict the dose space to a set of dose distributions with unique dose-volume statistics, this is not feasible in practice and highly arbitrary. We conclude that it is not necessary to take into account orthogonality in the construction of a basis of fluence space. The downside lies in the fact that there is also no way to obtain the linear coefficients by orthogonal projections.

It is important to notice that this basis need not span the entire fluence space $\mathcal{F}$ as defined above. The space of all *practically achievable* fluences is subject to very stringent limitations, such as continuous and differentiable profiles with finite penumbra. It is sufficient to construct a basis which is complete for the applicable fluence space. Thereby, it is ensured that the resulting fluence distribution is not grossly compromised by applicability limitations.

Henceforth we consider as fluence space the subset $\mathcal{F}_\mathcal{B} \subset \mathcal{F}$ which is in the range of a given basis $\mathcal{B} \subset \mathcal{F}$. The basis $\mathcal{B}$ can be enlarged to access a greater subset of $\mathcal{F}$, e.g. for a refinement of field discretisation. The range of $\mathcal{B}$ needs clarification. Usually, in vector spaces all linear combinations of basis vectors lie in the vector space. In this case, only those linear combinations are permitted which yield an element of $\mathcal{F}$ [2]. Although in practical computations one can choose the basis such that all linear coefficients have to be non-negative, for theoretical considerations negative coefficients may be allowed. The details of this basis are given in chapter 5.

The concept of a finite basis of elementary fluence distributions is a cornerstone of the *ray formalism*. An element $\eta_i$ of the basis $\mathcal{B} = \{\eta_1, \ldots \eta_n\}$ will be termed *ray*, its dose distribution $T_i = T\eta_i$ *ray dose* (at unit weight). $\mathcal{B}$ is called the *ray basis*. The image of $\mathcal{F}_\mathcal{B}$ : $T\mathcal{F}_\mathcal{B} \subset \mathcal{D}$ is the subset of all practically feasible dose distributions. Since $T$ is an injection, the $\{T_i\}, i = 1, 2, \ldots$ form a *basis* of the accessible dose space $\mathcal{D}_\mathcal{B} \subset \mathcal{D}$. The linear coefficients $\phi_i$ of $\Phi = \sum_i \phi_i \eta_i$ are often referred to as *ray weights*. Again, the restriction applies that only those linear combinations are permitted which do not lead to negative fluences. Notice that the ray doses are the equivalent of Green's functions for

[1] Two elements $a, b$ of a set have $d(a, b) = 0$ if and only if $a = b$

[2] It is for this reason that the notation $\mathcal{F}_\mathcal{B} = \text{span}(\mathcal{B})$ is abandoned

this special function space. We can consider the rays as 'field modes' or 'single particle states' of radiotherapy optimization, their doses as their 'charge' by virtue of which they are coupled to each other and to an external potential which is introduced below.

### 2.1.4 The Variation Problem - Constrained Optimization

The fluence space $\mathcal{F}_\mathcal{B}$ was established as the parameter space from which the solution of the radiotherapy optimization problem will originate. The treatment philosophy requires that certain normal tissue dose limits not be exceeded while the probability of treatment failure is minimised. In the language of optimization this means that the optimum solution has to meet a set of *constraints* while an *objective function* is minimised. The following is concerned with the fundamental formulation of the problem; numerical methods to solve it are the subject of chapter 6.

Commonly, an objective function

$$F : \mathcal{F} \to \mathbb{R}_0^+ \tag{2.4}$$

is defined to model a problem such that it attains its global minimum at the optimum solution $\Phi^*$. For the setup here, $F$ is a functional rather than a function. Frequently it is required that $F$ is twice continuously differentiable with respect to its argument - in this case it is required that the first and second *variation* of $F$ with respect to $\Phi$, in our sense

$$\frac{\partial F(\Phi)}{\partial \eta} = \lim_{\epsilon \to 0} \frac{F(\Phi \pm \epsilon\, \eta) - F(\Phi)}{\pm \epsilon} \quad \text{for all test rays } \eta \in \mathcal{B}, \Phi \in \mathcal{F} \tag{2.5}$$

and

$$\frac{\partial^2 F(\Phi)}{\partial \eta^2} = \lim_{\epsilon \to 0} \frac{\partial F(\Phi \pm \epsilon\, \eta) - \partial F(\Phi)}{\pm \epsilon} \quad \text{for all test rays } \eta \in \mathcal{B}, \Phi \in \mathcal{F} \tag{2.6}$$

exists. In this case a necessary condition for optimality is that the first variation of $F$ vanishes and the second variation is positive definite for all test functions $\eta$. It is important to notice that the variation was defined here with respect to the ray basis $\mathcal{B}$ rather than fluence space $\mathcal{F}$. With this notion we emphasize the fact that the objective function may be defined for a much greater parameter space, but the variation (and subsequent optimality conditions) are restricted to the ray basis $\mathcal{B}$. If an arbitrary fluence distribution $\Phi$ meets the optimality conditions with respect to $\mathcal{B}$, it is not possible to improve it further within the parameter space $\mathcal{F}_\mathcal{B}$. It need not be that $\Phi \in \mathcal{F}_\mathcal{B}$. There may be an enlarged basis with respect to which the fluence distribution is not optimum. By using the basis property, the variation problem can be transformed into a vector function. If $\Phi = \sum_i \phi_i \eta_i$ is restricted to $\mathcal{F}_\mathcal{B}$, the objective functional becomes a function of the linear coefficients $\phi_i$ and the variation the gradient with respect to the vector $\Phi = (\phi_1, \phi_2, \ldots)$.

This dichotomy between fluence (and its corresponding dose distribution) and the fluence variations (as effected by the ray basis) is crucial for the further development. The underlying dose distribution and its related fluence variations are regarded as separate

entities, which is brought to bear on the optimization of beam angles, Monte Carlo dose computation or consideration of physical constraints.

The objective function as a measure of treatment success is clearly related to the dose distribution. In the following we use $F(D) = F(T\Phi)$ as a shorthand notation when we are not concerned with the aspect of radiation transport. In the next section a general concept for the formulation of $F(D)$ is given, and in chapter 3 the functional form is derived from biological principles.

The optimization of the fluence is subject to a number of *constraints*, some of which are also related to dose space when they correspond to normal tissue reactions. These constraints

$$G : \mathcal{F} \to \mathbb{R}_0^+ \tag{2.7}$$

are treated similarly to $F$, i.e. it is required that the first and second variation exists and the notation $G(D) = G(T\Phi)$ is used where appropriate. Other constraints act directly on fluence space to take into account the limitations of the treatment equipment. Those will be dealt with in chapter 4.

The optimization problem of radiotherapy then becomes

$$\begin{aligned} \text{minimise} \quad & F(\Phi) \\ \text{subject to} \quad & G_1(\Phi) \leq C_1, \ldots, G_N(\Phi) \leq C_N, \quad C_k \geq 0 \end{aligned} \tag{2.8}$$

where $N$ constraints are taken into account. The special nature of the problem allows some fundamental statements at this early point. Firstly, the objective is to maximise the dose to the tumour, which translates into the minimisation of the objective function, hence $F$ will be the only strictly decreasing function of fluence. Secondly, the normal tissue constraints which are associated with the $G_k$ will be strictly increasing functions of fluence. Therefore, these constraints cannot be mutually exclusive. For any reasonable set of constraints there will be a feasible solution of the problem, be it even zero fluence. In practice, the solution may be unsatisfactory because the constraints were too stringent. In general, not all constraints will be active, i.e. fluence limiting, at the optimum.

A non-linearly constrained optimization problem is commonly solved by transformation into an unconstrained subproblem for which a great number of algorithms is available. This is achieved by the method of (Lagrange-) multipliers which will be briefly motivated in the following. Let $L(\Phi)$ be the objective function of the unconstrained problem which has a solution $\Phi^*$ that solves problem eq.(2.8). Hence, $\frac{\partial L(\Phi^*)}{\partial \Phi} = 0$, yet not necessarily $\frac{\partial F(\Phi^*)}{\partial \Phi} = 0$. This is prohibited by any active constraint with $G_k(\Phi^*) = C_k$, i.e. even if $F$ could attain a smaller value elsewhere, the solution is bound to the manifold defined by this constraint. Any direction in which $\Phi^*$ could be changed (any feasible variation) must be orthogonal to the normal vector of this manifold, for any variation in the direction of the gradient would change the constraint. For all these feasible directions, it is required that the optimality condition hold, i.e. the variation of $F$ must be zero. In the remaining directions, the gradient of $F$ need not vanish. Hence, the gradient $\frac{\partial F(\Phi^*)}{\partial \Phi}$ is a linear combination of

the gradient vectors $\frac{\partial G_k(\Phi^*)}{\partial \Phi}$ with linear coefficients $\lambda_1 \geq 0, \ldots, \lambda_N \geq 0$, the Lagrange multipliers. If a constraint is not active, the corresponding multiplier is zero. The Lagrange function reads

$$L(\Phi, \lambda) = F(\Phi) + \sum_{k=1}^{N} \lambda_k G_k(\Phi). \tag{2.9}$$

If the gradients of the constraints are linearly independent then there exists a unique vector $\lambda^*$. The solution of the constrained problem is a pair $(\Phi^*, \lambda^*)$. For more details about Lagrange multiplier theory we refer to [21, 22]. The Lagrange multipliers may not be taken as 'penalty factors' or 'weights' because they are a mere mathematical construction. However, they convey relevant information about the rate of change of $F$ with respect to changes of $G = (G_1, \ldots, G_N)$, consider

$$\begin{aligned} \text{minimise} \quad & F(\Phi) \\ \text{subject to} \quad & G(\Phi) \leq u \end{aligned} \tag{2.10}$$

then

$$\nabla_u p(u) = -\lambda(u) \tag{2.11}$$

where $p(u)$ is the optimum objective parameterised by $u$, i.e.

$$p(u) = F(\Phi^*(u)) \tag{2.12}$$

(see [21], pp 277). This correlation provides the answer to an important question of treatment planning: how much effect in the target volume can be gained if the risk of side-effects is increased.

In practice, the unconstrained problem is solved for a sequence of vectors of Lagrange multipliers which converges to $\lambda^*$. The solution of the constrained problem is hampered by the difficulties associated with the determination of the multipliers. Fortunately, the radiotherapy optimization problem is well behaved in many aspects such as conditioning, convexity, and above all, degeneracy. Many dose distributions are equivalent with respect to the objective and constraint functions, so that a rather large set of solutions of eq.(2.8) exists within numerical uncertainty. Because of degeneracy, the solution of the problem is rather tolerant towards inexactf Lagrange multipliers.

## 2.2 Local Objective Measures

The inescapable complexity of biological modelling requires far-reaching approximations. With an eye to physical, dose-based optimization, biology is forced into a scheme which only clinical experience may eventually prove adequate.

By its physical nature, dose is a density function and supports, in mathematical terms, a measure on $\mathbb{R}^3$. A measure assigns a non-negative real number to any set of its support in a consistent 'linear' way. In that sense, a measure is an entirely local quantity in that

the measure of a set is the sum of its constituent subsets. If the *radiation effect* rather than the dose is regarded as a measure, one does of course make a far reaching assumption about the non-linearities in spatial interactions of complication mechanisms. This does not imply that long-range interactions cannot be dealt with, but there are certain difficulties associated with, for example, 'diffusion of radiation damage': tissue damage wrought on a confined volume may spread to adjacent tissues and thereby violate the linearity in volume of a measure. However, the use of such a concept for radiotherapy optimization offers great advantage: a local variety of 'optimality' may be found which reduces the complexity of the problem. This concept is invoked later in this and in the next section.

In the following, a 'linearisation in volume' of the hypothetic biological objective function $F$ is introduced by reducing it to an equivalent local biological measure[3]. The idea is to substitute $F$ with a measure which is equivalent with respect to local variations of the dose on a finite set containing $\Phi^*$. The approximation is similar to mean-field techniques in statistical mechanics where long-range interactions are combined to a background effect. This method works mainly because therapeutic dose distributions vary only slowly on mesoscopic length scales where short-range effects could lead to a breakdown of the approximation.

### 2.2.1 The Variation Density

We call $\mu$ a measure on $\mathbb{R}^n$ if $\mu$ assigns a non-negative number, possibly $\infty$, to each subset of $\mathbb{R}^n$ such that:

1.
$$\mu(\emptyset) = 0 \quad ; \tag{2.13}$$

2.
$$\mu(A) \leq \mu(B) \quad \text{if } A \subset B \quad ; \tag{2.14}$$

3. If $A_1, A_2, \ldots$ is a countable (or finite) sequence of sets then
$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mu(A_i) \tag{2.15}$$
with equality if the $A_i$ are disjoint sets.

A measure can be used to 'weigh' a volume if we envisage it as some kind of mass density. Likewise, if $\mu$ is the local radiation effect density, we may arrive at the total effect by integrating it over the whole tissue volume. So, in case $\mu(\vec{x})$ is a finite function on a closed set $A$, we find
$$\mu(A) = \int_A \mu(\vec{x}) dx^3 \quad , \tag{2.16}$$

[3]Although 'local measure' is a tautology, we use this term to stress the mathematical meaning in our use of the word.

where $\mu(A)$ is the 'effect' or 'damage' accumulated in the volume $A$.

It is this last property which becomes important in this formulation of the radiotherapy optimization problem. Since the optimality condition relies primarily on the first variation of the objective (or Lagrange) function with dose, for the purposes of optimization it is sufficient to approximate $F$ with a function with equivalent first variation. By virtue of the density nature of dose a local variation of the dose distribution defines a 'variation density', which is a valid approximation for small variations of the dose distribution. By integrating the variation density with respect to the local dose, an objective density can be derived which is equivalent to the original objective function with respect to local variations. A global variation of the dose distribution can be decomposed into locally confined variations in subvolumes and subsequently their effect summed up; by this way a first order approximation in dose mediates the measure quality of the objective function. This highlights the fact that biological modelling for optimization need not achieve as high a standard as for predicting treatment outcome.

First the notion of local variation of the objective function is introduced. Let $\{A_i\}$ be a decomposition of the support of $D$ into disjoint sets called 'test volumes' and let $F(D)$ be an objective function of the dose distribution. With $\chi(A_i)$ we denote the characteristic function of $A_i$, i.e. $\chi(\vec{x}) = 1$ if $\vec{x} \in A_i$, and else $\chi(\vec{x}) = 0$. If

$$\lim_{\epsilon \to 0} F(D + \epsilon\chi(A_i)) = \lim_{\epsilon \to 0} F(D - \epsilon\chi(A_i)) \tag{2.17}$$

holds for all $D$, we call $F$ continuous and if

$$\frac{\partial F(D)}{\partial D(A_i)} = \lim_{\epsilon \to 0} \frac{F(D \pm \epsilon\chi(A_i)) - F(D)}{\pm\epsilon \text{ vol}\,(A_i)} \tag{2.18}$$

exists for all $D$, we call $F$ locally differentiable[4]. In the following, $F$ is assumed to be continuous and differentiable. By letting $A_i \to \vec{x}$, we arrive at the *variation density* $\frac{\partial F}{\partial D(\vec{x})}$ if $F$ is sufficiently well behaved ($D$ is continuous and differentiable). The modulus of this quantity behaves like a measure.

The optimality of the dose distribution is identified with a vanishing first variation of the objective function. It can be seen, that this variation density rather than the objective function itself plays an important role in the solution of the problem. In principle, if a substitute objective function can be devised whose pointwise derivative with respect to dose equals the variation density, the solutions of both optimization problems will be identical. Thus, it is sufficient to define the objective by means of its variation density - that can be integrated pointwise with respect to local dose $D(\vec{x})$ to yield an *objective measure*. In the following we interprete

$$f(D(\vec{x}), \vec{x}) = \int \frac{\partial F(D)}{\partial D'(\vec{x})} dD'(\vec{x}) \tag{2.19}$$

[4]This definition is not intended to follow rigorous mathematical standards and should be seen in connection with the inevitable discretisation of the patient volume into voxels.

as an 'objective density', where we take care that $f(\vec{x}) \geq 0$ and consequently refer to

$$f(D(\vec{x}), A) = \int_A f(D(\vec{x}), \vec{x}) dx^3 \tag{2.20}$$

as objective (function) of the volume $A$, and to

$$\frac{\partial f(D(\vec{x}), A)}{\partial D} = \int_A \frac{\partial f(D(\vec{x}), \vec{x})}{\partial D(\vec{x})} dx^3 \tag{2.21}$$

$$= \int_A \frac{\partial F(D(\vec{x}))}{\partial D(\vec{x})} dx^3 \tag{2.22}$$

as objective variation of the volume $A$. The integrand of eq. (2.21) is the *variation density* of $F$.

Since the objective density originates in a variation of the objective functional, it need not be a function of the local dose only. In fact, it may have a complex dependency on the dose distribution and therefore change during the iterative approach to the optimum. This may potentially cause problems if the response exhibits great variations for small variations of the dose, not unlike a critical behaviour. There are two arguments in favour of this approximation. Firstly, the optimum solution will always correspond to low complication probabilities, so that the effect of non-linearities will also be small. Secondly, the dose distribution close to the optimum will never be very inhomogeneous, especially on mesoscopic length scales of the size of cell migration or diffusion of cytokines. Therefore, non-local interactions act in front of a rather homogeneous backdrop and can be incorporated into a mean-field approximation.

We will see in the third chapter, that several dose-response mechanisms can be very well described with a local model. For yet other complication mechanisms, it may suffice to take into account the global coupling by means of a single bias term in analogy with mean-field approximations in many particle physics. An instance where the variation density is not guaranteed to exist and thus also no objective density can be defined is for complication mechanisms with a strong dynamic component, for example the time-dependence of diffusion-like processes cannot be treated adiabatically (see chapter 3.2) or the damage propagates along complicated geometric structures like blood vessels. To the present day, no clear experimental evidence has been given that could support a model with these intricate features.

### 2.2.2 The Variation Problem for Measures

With the introduction of objective densities, the radiotherapy optimization problem eq. (2.8) now reads

$$\begin{aligned} \text{minimise} \quad & \int_{V_T} f(T(\vec{x})\Phi)\, dx^3 \\ \text{subject to} \quad & \int_{V_k} g_k(T(\vec{x})\Phi)\, dx^3 \leq C_k, k = 1, 2, \ldots, M \\ \text{and} \quad & G_k(\Phi) \leq C_k, k = M+1, \ldots, N \end{aligned} \tag{2.23}$$

with $D(\vec{x}) = T(\vec{x})\Phi$ and $C_k > 0$ being the *total effect* of constraint $k$. The volumes $V_k$ and $V_T$ (target volume) delineate the support of the corresponding objective density $g_k$ respectively $f$. Notice that not all $V_k$ need to be different; in this case more than one constraint is defined in a volume. In case $V_k = V_T$, a dose limiting constraint is defined in the target volume. The additional constraints $G_k(\Phi), k = M+1, \ldots, N$ correspond to physical restrictions of the fluence distribution.

For the corresponding Lagrange function we obtain

$$L(\Phi, \lambda) = \int_{V_T} f(T\Phi)\, dx^3 + \sum_{k=1}^{M} \frac{\lambda_k}{C_k} \int_{V_k} g_k(T\Phi)\, dx^3 + \sum_{k=M+1}^{N} \lambda_k G_k(\Phi) \tag{2.24}$$

where the biological constraints are normalised to unity. The physical constraints are dealt with in a variety of ways which will be the subject of chapter 4. Until then, they are stripped from the Lagrange function. The definition of the Lagrange density $l(\Phi)$ is straightforward

$$l(\vec{x}, \Phi, \lambda) = f(T(\vec{x})\Phi)\chi(V_T) + \sum_{k=1}^{M} \frac{\lambda_k}{C_k} g_k(T(\vec{x})\Phi)\chi(V_k). \tag{2.25}$$

In this formulation, a number of issues concerning the uniqueness of the solution and the existence of local minima can easily be addressed. One fundamental theorem of optimization theory states that a convex function assumes a unique minimum on a convex set. (For definitions of convexity of functions and sets see [21, 23]). Since convexity is preserved in addition, hence integrals, the Lagrange function is convex if all functions $f$ and $g_k$ are convex. For a twice differentiable function of one variable, convexity is tantamount to a positive definite second derivative, which can be easily checked. The objective density $f$ is always convex. The parameter space $\mathcal{F}_\mathcal{B}$ is also convex. As a consequence, together with the injectiveness of $T$, the radiotherapy optimization problem has a single (global) minium if all $g_k$ are convex, which is often the case as is shown in chapter 3. However, this does not mean that there exists a unique solution in practice. The reason is, that the integrals are rather insensitive to minute changes of the objective density, in particular when a slight increase in one region is compensated for by a slight decrease in another. The approximate solution as obtained from an algorithm will therefore be sensitive to pertubations depending on the degree of *degeneracy* of the Lagrange function.

### 2.2.3 The Ray Derivative

The variation of the Lagrange function with respect to a ray can be captured by a powerful intuitive picture if the measure properties are included. Starting with eq.(2.24) we find for the variation of $L$ with respect to the weight $\phi_\eta$ of ray $\eta$

$$\begin{aligned} \frac{\partial L(D)}{\partial \phi_\eta} &= \frac{\partial L(D)}{\partial D} \frac{\partial D}{\partial \phi_\eta} \\ &= \int \frac{\partial\, l(D(\vec{x}))}{\partial D} T(\vec{x})\, \eta\, dx^3. \end{aligned} \tag{2.26}$$

Again, it is important to notice that the ray $\eta$ need not contribute to the dose distribution $D$. We find that the variation of $L$ is affected only by those volumes which lie in the path of the ray; this is a consequence of the 'linearisation' induced by the introduction of a Lagrange density. Although this may seem obvious from the point of view of dose based optimization, for a biological objective function this originates only from the approximation leading to the measure quality of the objective.

Eq.(2.26) can be seen as a 'dose weighted mean effect' along the path of the ray. The ray 'picks up' negative and positive contributions to the variation density and sums them weighted with the dose at these points; only if this balance is even, the variation vanishes. The intuitive picture that beam directions which align to the greatest extension of a target volume are favourable, or that the beam should enter from the direction where the target volume is closest to the surface is formalised in this concept of a *ray derivative*. The potential of a ray to improve on a given dose distribution is a *linear* combination of the effect that has already been accumulated along its path and the dose deposited by this ray.

In case the dose distribution is comprised of the basis rays $\eta_i$, we introduce the shorthand notation

$$\left(\frac{\partial L(D)}{\partial \Phi}\right)_i = \int \frac{\partial\, l\,(D(\vec{x}))}{\partial D}\,\frac{\partial D(\vec{x})}{\partial \phi_i}\,dx^3 \tag{2.27}$$

$$= \int \frac{\partial\, l\,(D(\vec{x}))}{\partial D}\,T_i(\vec{x})\,dx^3. \tag{2.28}$$

This equation governs the practical computation of the optimum dose distribution. However, eq.(2.26) also is of practical importance, because the dose distribution need not be a linear combination of the rays. This earlier equation can be used to determine refinements of the ray basis, like a finer decomposition of the beams or in the form of additional beams. The latter leads to the question of beam angle optimization, and seen together with clinical issues, the necessary number of beams.

## 2.3 The Global Relaxation Theorem

The original setting of fluence space does not know finitely many beam directions. However, both in computations and most treatment applications the number of beams is limited to a number between 3 and 100. If this were taken into account for the optimization, it would become a discrete combinatorical optimization problem with a continuous subproblem. This would increase the complexity tremendously, not in the least because an abundance of local minima would emerge[5]. Also, the optimization of beam angles and fluence profiles cannot be separated, as we will see later. As a consequence, to evaluate each beam arrangement, a full optimization of the fluence profiles has to be run.

[5]This problem is caused mainly because the set of all $n$ beam arrangements is not convex: the superposition of two $n$ beam plans can be anything from an $n$ to a $2n$ beam plan.

These difficulties may make it appear poorly justified to expend so much effort on what may be such a small gain. Indeed, the high degeneracy of the radiotherapy optimization problem causes the attainable minimum objective to saturate for as few as 5 to as many as 15 beams in virtually all cases. The true problem lies in the fact that the saturation threshold is patient dependent and close to the threshold, beam angle optimization does yield a considerable gain in difficult cases located in the head or thorax. Especially in the head, non-coplanar beams can be indispensable with the result that a technique relying on coplanar beam angles will waste chances for tumour cure. In many cases the search for solutions with few beams is predominantly driven by the clinical requirement for short treatment times and may lose importance with improvements in technology.

A common misnomer lies in the notion that finding a suitable set of beam angles is an optimization problem[6]. In the framework of our ray formalism, beam angle optimization is tantamount to picking the 'best' set of basis rays $\mathcal{B}$ which are bundled to beams from certain angles. It is similar to an *approximation* problem: a truly optimum dose distribution of rays using the whole fluence space is to be approximated by few beams such that the loss expressed in the changee of the objective function is acceptable. Degeneracy is exploited here: it is not necessary to approximate the fluence distribution, which would be quite difficult, but the dose distribution which supports the only relevant measure of 'proximity' of dose distributions, the objective function. Normal tissue constraints enter the balance indirectly by causing a certain increase of the objective function by shielding since they are never violated by an optimum dose distribution. Once the basis is chosen, the radiotherapy optimization problem can be solved - we bear in mind that optimum in this context means optimum with respect to this particular basis.

Obviously, the global optimum dose distribution cannot be generated. Also, due to degeneracy, it is not indicated to accumulate too much information about it. The following propositions give a number of criteria which characterise the global optimum without computing it, and aid in the search for a suitable basis.

### 2.3.1 Optimality for a Reduced Fluence Space

The first proposition gives a necessary criterion for optimality of a dose distribution without the need to know the ray doses $T\eta$ so that this criterion can also be applied to dose distributions for which the fluence distribution is not known.

PROPOSITION I. If the dose distribution $D^*$ is a solution of eq.(2.23) with the Lagrange multipliers $\lambda^*$ then

$$\int \frac{\partial\, l(D^*(\vec{x}))}{\partial D}\, D^*(\vec{x})\, dx^3 = 0. \tag{2.29}$$

Proof: Consider $\frac{\partial L(\kappa D^*)}{\partial \kappa}$ which must vanish for $\kappa = 1$.

Of course, it will not be possible to obtain $\lambda^*$ without the knowledge of the basis which created $D^*$. The true value of this proposition lies in the fact that it can be used

[6]The more adequate term 'customisation' has been proposed.[24]

to verify whether two dose distributions obtained from different ray basis sets and the global solution are degenerate. If $L$ is convex, then for two dose distributions $D_1^*, D_2^*$ with equivalent Lagrange multipliers $\lambda^*$

$$\kappa L(D_1^*, \lambda^*) + (1-\kappa)L(D_2^*, \lambda^*) \geq L(\kappa D_1^* + (1-\kappa)D_2^*, \lambda^*). \tag{2.30}$$

If equivalence holds and the condition of proposition I is met, it can be supposed all convex combinations are degenerate to the global optimum and are solutions of eq.(2.23). If they are, either of the sets $\mathcal{B}_1, \mathcal{B}_2$ yields a suitable beam angle optimized basis. Otherwise, other basis sets need to be created. In that sense, proposition I is a 'termination criterion' for an iterative basis refinement.

### 2.3.2 Global Relaxation

The last proposition gave a necessary condition that a dose distribution is optimum with respect to its basis. To verify that a given dose distribution is equivalent to the global optimum there is no other way than to use rays from all possible angles as test functions. This is impractical for finely grained rays. However, it is possible to exploit a quality of the optimum dose distribution for a heuristic method which speeds up the selection of suitable beam angles.

Turning back to the Lagrange function, we recall that all normal tissue constraints are an increasing function of dose, whereas the objective function is the only decreasing function of dose. So in the ray derivatives, all positive contributions to the integral stem from constraints, whereas all negative contributions stem from the target volume. Since at the optimum every ray derivative is a balanced sum of normal tissue and target volume terms, the modulus of each contribution is equal. The fortuitious detail is that many rays overlap in any given point in the target volume so that the volume centered at this point contributes to the negative terms of many ray derivatives, and thus also to the positive terms. In a highly symmetric setting the normal tissue dose load as expressed by the positive contributions to all ray derivatives is equivalent for every ray. The optimum solution can thus be seen as an equilibrium state of all rays: the variations of the objective function (and the constraints) have relaxed to the common ground state; in that sense, every ray which contributes to the dose distribution has equal 'importance'.

We formalise this finding in the following proposition. With $[.]_+$ we denote the positive (belonging to normal tissue), with $[.]_-$ we denote the modulus of the negative (belonging to the target volume) part of the ray derivative. Let $\mathcal{B}_{S_2(\vec{x})}$ be a basis of conical rays which originate from every point of the unit sphere centered at $\vec{x} \in V_T$ with ray diameter $r$ in the plane containing $\vec{x}$ perpendicular to the ray.

PROPOSITION II. Let $D^* = T\Phi^*, \lambda^*$ be the solution of eq.(2.23) with respect to $\mathcal{F}$. Let $\eta_\Omega \in \mathcal{B}_{S_2(\vec{x})}$ be the ray impinging on $\vec{x}$ from solid angle $\Omega$. Let $\mathcal{B}'_{S_2(\vec{x})}$ be the set of all rays $\eta_\Omega \in \mathcal{B}_{S_2(\vec{x})}$ with non-zero fluence, i.e. for every point $P$ with $\eta(P) > 0$ it holds that

$\Phi^*(P) > 0$. Then for all rays $\eta_\Omega$

$$\min_{\eta_\Omega \in \mathcal{B}'_{S_2(\vec{x})}} \left\{ \int_{S_r(\vec{x})} \left| \frac{\partial f(D^*)}{\partial D} \right| T\eta_\Omega \, dx^3 \right\} \leq \left[ \frac{\partial L(D^*)}{\partial \phi_\Omega} \right]_+ \leq \max_{\eta_\Omega \in \mathcal{B}'_{S_2(\vec{x})}} \left[ \frac{\partial L(D^*)}{\partial \phi_\Omega} \right]_- \tag{2.31}$$

where $S_r(\vec{x})$ is the ball with radius $r$ centered at $\vec{x}$.

Proof: For every ray $\eta_\Omega \in \mathcal{B}'_{S_2(\vec{x})}$ we have

$$\left[ \frac{\partial L}{\partial \phi_\Omega} \right]_+ = \left[ \frac{\partial L}{\partial \phi_\Omega} \right]_- \tag{2.32}$$

since for these rays the optimality condition holds. The second inequality follows. Also, with

$$\int_{S_r(\vec{x})} \left| \frac{\partial f}{\partial D} \right| T\eta_\Omega \, dx^3 < \int_{\mathrm{supp}(T\eta_\Omega)} \left| \frac{\partial f}{\partial D} \right| T\eta_\Omega \, dx^3 < \left[ \frac{\partial L}{\partial \phi_\Omega} \right]_- \tag{2.33}$$

follows the first inequality since $S_r(\vec{x}) \subset \mathrm{supp}(T\eta_\Omega)$.

This proposition gives upper and lower boundaries for the ray derivative of the normal tissue constraints. It is the formal expression of the intuitive view that the damage to the normal tissue has to be spread evenly to obtain the best treatment plan. This view is implicit in many treatment techniques, most clearly in rotational irradiation[7]. From a different perspective, the proposition can be interpreted as 'Only if the tolerance of all normal tissues is exploited to the same extent, the effect to the target can be optimum'. At the global optimum, the fluence distribution is relaxed in the sense that all rays fulfill the 'equilibrium' conditions of proposition II, i.e. not only are all ray derivatives zero if the rays contribute to the solution, but also are all positive contributions to the ray derivatives within a given interval whose width depends on the patient geometry.

The definitive measure for the impact of a given ray $\eta$ on the normal tissue is the ray derivative $\left[ \frac{\partial L(D^*, \lambda^*)}{\partial \phi_\eta} \right]_+$ which can of course be computed for any dose distribution and with respect to any ray. If

$$\left[ \frac{\partial L(D^*, \lambda^*)}{\partial \phi_\eta} \right]_+ > \left[ \frac{\partial L(D^*, \lambda^*)}{\partial \phi_\eta} \right]_- \tag{2.34}$$

this ray does not contribute to the dose distribution $D^*$. If a relatively coarse ray is used as a test function, this may be taken as indicative of a 'bad' or 'good' angle of incidence. However, this should be taken with care: if the ray is decomposed into a number of smaller rays, the condition eq.(2.34) may only apply to a few, so that on total the given beam angle may well be included into the ray basis. An example: while an unmodulated field may exceed the tolerance of some high risk structure and should not be chosen, a modulated field may spare this structure completely. As a consequence, the best beam angles of a ray basis depend on the size of its constituent rays.

In general practice, proposition II will be applied to approximations of the optimum dose distribution and Lagrange multipliers. The import of proposition II were limited if the

[7] Whereas the rationale of rotational irradiation and the proposition agree for a convex Lagrange density, they do not for a non-convex setting where 'evenly spread damage' may not mean 'evenly spread dose'

Lagrange multipliers had to be known exactly. However, it will generally be sufficient for the selection of beam angles to work with an approximation of the multipliers. Unless the patient geometry is irregular to a degree that a satisfying therapy will barely be possible, the boundaries of proposition II will be fairly tight. If, however, the first inequality is violated, the dose distribution is not a good approximation to the optimum. The main purpose is to motivate a heuristic method for the selection of beam angles.

### 2.3.3 Annotations to Beam Angle Optimization

In the following we give a recipe how a suitable few-beams basis can be constructed when a good approximation to the optimum solution is known. This solution $D^*, \lambda^*$ may be obtained from a basis with significantly more beams than clinically practical; it is only important that this solution is close to the 'equilibrium' in the sense of the above proposition II.

As already pointed out, constructing such an 'optimized' basis is an approximation problem: how can the relaxation property be met approximately with the smallest number of fields. With $\mathcal{B}_\Omega$ we denote the set of all rays $\eta_i \in \mathcal{B}$ which belong to a field which impinges from a solid angle $\Omega$. With $\Omega \in \mathcal{F}_{\mathcal{B}\Omega}$ we denote the beam composed of these rays, which can itself be understood as a ray. It is essential to measure the relative 'importance' of a beam to the dose distribution, maybe by

$$\Delta L = L(D^* - \kappa T\Omega) - L(D^*). \tag{2.35}$$

where $0 < \kappa < 1$ is some weighting parameter. The problem with this approach is that there may emerge solutions with negative total fluence. A better way would be

$$\begin{aligned}
\Omega_{\text{opt}} &= \arg\max_{\Omega\in\mathcal{F}_{\mathcal{B}\Omega}} L(D^* - \kappa T\Omega) && (2.36)\\
&= \arg\max_{\Omega\in\mathcal{F}_{\mathcal{B}\Omega}} L\left((1-\kappa)D^* + \kappa \sum_{\Omega'\in\mathcal{F}_{\mathcal{B}\setminus\mathcal{B}_\Omega}} T\Omega'\right) && (2.37)\\
&= \arg\min_{\Omega\in\mathcal{F}_{\mathcal{B}\Omega}} L\left((1-\kappa)D^* + \kappa T\Omega\right) && (2.38)\\
&= \arg\min_{\Omega\in\mathcal{F}_{\mathcal{B}\Omega}} L((1-\kappa)D^*) + \kappa \int \frac{\partial\, l((1-\kappa)D^*)}{\partial D} T\Omega\, dx^3 && (2.39)\\
&= \arg\max_{\Omega\in\mathcal{F}_{\mathcal{B}\Omega}} \left[\frac{\partial L((1-\kappa)D^*)}{\partial \phi_\Omega}\right]_- && (2.40)
\end{aligned}$$

where we expand the Lagrange function to first order in the test fluence and use the fact that the ray derivative will be dominated by the negative (target volume) term. Of course, this chain of arguments is impossible to prove and constitutes an abuse of notation for the sake of clarity. The result implies the intuitive picture that the beam which causes the greatest variation of the objective function *on top of* some bias dose distribution should be included

into the basis. It is important to see the role of the bias dose distribution: it generates the proper background of normal tissue dose load which determines the proximity to the optimum by virtue of the relaxation property. The parameter $\kappa$ mediates a redistribution of normal tissue dose from many rays to few beams according to the effect in the target volume of these beams.

In practice, the problem is to choose suitable test rays $\Omega$ which in this picture would have to be intensity modulated beams. This can be accomplished by constructing the beam as a set of its constituent rays $\eta_i$ and determining the optimum fluence distribution $\psi_\Omega = (\phi_1, \phi_2, \ldots)$ in a run of the fluence optimization with the basis of this beam.

In the following, an iterative scheme for the selection of the optimum set of beam angles is devised.

1. Find $D^*, \lambda^*$ with respect to some basis $\mathcal{B}$. Set $\mathcal{B}^0 = \emptyset$ and $0 < \kappa^1 < 1$. Number of beams $k = 0$.

2. $k = k + 1$.

3. Find the best beam angle $\Omega^k$ by

$$\Omega^k = \arg\min_{\Omega \in S_2} \min_{\{\phi_i : \eta_i \in \mathcal{B}^{k-1} \cup \mathcal{B}_\Omega\}} L((1-\kappa^k)D^* + \sum_i T\eta_i\phi_i) \tag{2.41}$$

4. $\mathcal{B}^k = \mathcal{B}^{k-1} \cup \mathcal{B}^k_\Omega$

5. Find

$$\kappa^{k+1} = \arg\max_{\kappa \leq 1} \left\{ \min_{\{\phi_i : \eta_i \in \mathcal{B}^k\}} L\left( (1-\kappa)D^* + \sum_i T\eta_i\phi_i \right) \leq (1+\epsilon)L(D^*) \right\} \tag{2.42}$$

6. If $\kappa = 1$ exit. Else goto 2.

The maximum deviation of the few-field solution from the optimum is denoted by $\epsilon$.

The advantage of this algorithm over search schemes including simulated annealing is that the computation time goes with $N^2$ if $N$ is the number of beams as compared to $a^N$ for some $a$. Nevertheless, computation times will be too long for routine clinical use. The gain of beam angle optimization may often not justify the effort, especially if class solution based arrangements of beams are available. The benefit of beam angle optimization may eventually lie in generating these class solutions.

# Chapter 3

# Biological Objective Modelling

The traditional measures of treatment success and side effects in radiotherapy are *tumour control probability* (TCP) and *normal tissue complication probability* (NTCP). It is erroneous to believe that these quantities have to be employed for biological radiotherapy optimization - as has been sketched, an equivalent objective density is more expedient.

The mechanics of normal tissue response to therapeutic radiation can be intricate. As the process propagates in time and space, non-linear coupling and feedback loops may take effect. Not only is it very difficult to monitor microscopic changes *in vivo*, but there is also significant variability in individual response so that a comprehensive theory for a population of patients can only deliver estimated values. These difficulties have led to a tradition of phenomenological descriptions of dose response which tried to integrate the limited biological knowledge and the diffuse clinical experience with the aim to give predictive assays of treatment success [25, 26, 27, 28, 29, 30, 31]. Using these models would be rash: they had not been designed for radiotherapy optimization and are not ideally suited.

The complexity of the evolution in time and space necessitates far reaching approximations. For time evolution, it is almost always assumed that the dose-response can be modelled in the limit of infinite time and complete relaxation to the final state. Although this is a gross oversimplification of acute responding tissues, to our knowledge no suitable model has been devised to describe these. The propagation in space may often be entirely trivial, but in some cases it may be tied to complex geometric patterns or be affected by mesoscopic and macroscopic interactions. Due to the lack of biological data, any approach to model these non-local volume dependencies must be avid to follow clinical experience, often at the cost of poorly satisfying model assumptions.

Intensity modulated radiotherapy offers the possibility for radically altered treatment concepts. The present development was designed to exert control on side-effects. For this reason, the therapeutic dose to the target is limited by the tolerance of the normal tissue, which can be set to the equivalent of conventional treatments. Furthermore, the specification of treatment objectives should be intuitive and should allow the creation of a data base of class solutions to standardise treatments. This is achieved with the notion of

universal iso-effects which are used to prescribe the dose to normal tissues.

## 3.1 Complication Limited Tumour Control: Iso-Effects

A classical score function of radiotherapy is the probability of uncomplicated tumour control $P_+$

$$P_+ = \text{TCP} \times \prod_{i=1}^{M} (1 - \text{NTCP}_i) \tag{3.1}$$

for $M$ complications [32, 33, 34, 35, 36]. An attempt to maximise $P_+$ may sometimes lead to inacceptable complication probabilities, although this will occur rarely. However, the general lack of control over the result of the optimization is not in accordance with the common treatment philosophy. One can amend eq.(3.1) with a set of weight factors $\lambda_i, i = 1, 2, \ldots, M$, and taking logarithms yields

$$\log P = \log \text{TCP} + \sum_{i=1}^{N} \lambda_i \log(1 - \text{NTCP}_i) \tag{3.2}$$

from where one can immediately go to eq.(2.24) by multiplying with $-1$. With this heuristic argument the treatment objectives are established as log-probabilities. With hindsight, this choice is backed up by the resulting expediency of the TCP and certain NTCP objective densities.

The prescription of the treatment objectives is accomplished by defining the maximum permissible total effect for each complication, following eq.(2.20). This quantity is most conveniently expressed in terms of the *iso-effect* which relates the effect of the given dose distribution to standardised conditions. The meaning of the iso-effect depends on the complication in question; for some complications an iso-effective dose can be defined which is the homogeneous dose to some reference volume which causes the same effect as the given dose distribution. For other complications the iso-effect is the percentage of the maximum damage or the destroyed fraction of the organ volume. The iso-effect is the key to the biological modelling as presented here because it enables an intuitive and unambiguous translation of the treatment rationale to numerical quantities.

Some concepts pertinent to biological modelling are given in greater detail in Appendix A.

## 3.2 A Classification of Normal Tissue Constraints

Regarding the transition from complication log-probabilities to the objective measures used for optimization, the fundamental question is how long-range and non-local interactions can be taken into account. Only if a complication mechanism is entirely local and does not show any propagation of damage (or repair) at all, a measure which depends on the

local dose alone is a valid approximation. Only then, the functional form of the measure will not change during optimization. For all other complication mechanisms, the objective density would have to be updated with each change of the dose distribution.

At the other end of the spectrum are complications which are only set off if a certain threshold of some global quantity is exceeded. In this case one can assume that it is sufficient to approximate all non-local effects by their behaviour at the critical threshold - although they take effect for all levels of damage or dose, only close to the onset of the complication they will have a detectable impact. In this case the local measure is also independent of the actual dose distribution, but the local effect is always overestimated far away from the critical threshold[1]. The global coupling can be expressed by a factor which can be absorbed in the Lagrange multiplier.

These two extreme positions are currently available for biological optimization. As yet, clinical evidence is not clear as to how valid these approximations are. At present they seem to afford sufficient means to tailor the dose distributions according to clinical experience, but the fundamental lack of clinical data also prevents this approach from full-blown biological optimization.

The terminology for the two types of complication as outlined above stems from the tolerance of tissues towards partial damage.

An NTCP function $P(D)$ and its corresponding constraint function $G(D) = -\log(1 - P(D))$ is called *serial*, if for all volumes $A \subset \mathbb{R}^3$ and all $D \in \{D \in \mathcal{D} : D(\vec{x} \in A) = \infty\}$ , the condition

$$\lim_{\mathrm{vol}A \to 0} P(D) = 1 \tag{3.3}$$

$$\lim_{\mathrm{vol}A \to 0} G(D) = \infty \tag{3.4}$$

holds. All NTCP functions and constraint functions which do not conform to this definition are called *parallel*.

This definition tries to capture one of the most evasive concepts of radiotherapy, the biological volume effect. This concept describes the tissue specific increase in dose tolerance with a reduction of the irradiated volume. The volume effect has quite profound importance for radiotherapy optimization because it defines the shape of the optimum dose distribution in the normal tissues. The dose distributions are commonly characterized in the form of a (differential or cumulative) dose volume histogram (DVH). The width of the dose-volume distribution function in the differential histogram respectively the slope in the cumulative is determined by the volume effect: the greater the tolerance towards partial overdose, the wider/shallower can be the histogram; in the most extreme form of parallel complications the differential histogram will show two isolated peaks at low and high dose. The greatest advantage of biological optimization over dose based objectives is that each tissue can be equipped with its particular volume effect, whereas else all tissues are treated alike.

[1]That is, in the absence of negative feedback loops.

## 3.3 The Treatment of Time-Dependent Effects

The temporal evolution of radiation damage occurs on several time scales. While primary DNA repair mechanisms are activated within minutes, the repair can be completed within hours. Sublethal damage or incomplete repair may preserve DNA damage for days and weeks. The repopulation of tissues with stem cells and reconfiguration of the tissue matrix may take months and years.

For these reasons, the assessment of both treatment success and morbity can only be performed after the total dose was applied; hence, the total dose forms a natural measure of clinical evidence. Since the single dose per treatment fraction $D_{fx}$ may vary, it is necessary to relate total dose to a standard course of treatment, commonly in $D_{fx} = 2$ Gy fraction doses.

In this simplified picture, all processes on short time scales are treated as instantaneous and all changes to the fractionation scheme are judged by their long term effect. Acute reactions and complications which seem to imply critical behaviour (e.g. pneumonitis) are not amenable to this scheme directly, whereas the model applies very well to tumours. The time dependence enters the computations by virtue of the number of fractions $N_{fx}$ where a constant dose per fraction is assumed.

The standard time-independent model of cell kill assumes that a certain amount of dose kills (or sterilizes) a constant fraction of cells

$$p = \exp(-\alpha D_{fx}) \tag{3.5}$$

where $p$ is the probability of cell survival and $\alpha$ the cell sensitivity. This law does not take into account that cells may retain a higher sensitivity due to sublethal damage to their DNA after repair has completed, which leads to an expansion of $\alpha$ in $D_{fx}$

$$p = \exp(-\alpha D_{fx} - \beta D_{fx}^2) = \exp(-\alpha D_{fx}(1 + \frac{\beta}{\alpha} D_{fx})). \tag{3.6}$$

This is the standard 'Linear-Quadratic' model [37, 38, 39]. If the total dose $D$ is applied in $N_{fx}$ fractions, it can be transformed into the equivalent dose in 2 Gy fractions via

$$D_2 = D \frac{1 + \frac{\beta}{\alpha}\frac{D}{N_{fx}}}{1 + 2\,\mathrm{Gy}\frac{\beta}{\alpha}}. \tag{3.7}$$

The parameter $\beta$ may be significantly smaller for some tumours than for their surrounding normal tissues. If this is the case, a reduction of the fraction dose results in a better sparing of the normal tissues respectively in a higher tumour dose, see also figure 7.2.

The L-Q model does not take into account the effect that DNA repair mechanisms may need a threshold damage to be initiated. This so-called hypersensitivity to low doses per fraction is expressed by the induced repair model [40, 41]

$$p = \exp\left(-\alpha_R \left[1 + (\frac{\alpha_S}{\alpha_R} - 1)\exp\left(-\frac{D}{N_{fx} D_C}\right)\right] D - \beta \frac{D^2}{N_{fx}}\right) \tag{3.8}$$

with $\alpha_R$ being the sensitivity for large doses per fraction when repair is fully active, $\alpha_S$ being the increased sensitivity below the activation threshold and $D_C$ being the critical activation dose. This model predicts a significantly lower probability of cell survival for doses around 0.5 Gy if the parameter $\alpha_s$ is large. This effect is potentially of great importance for tissues with a low dose tolerance where the dose per fraction is always in the critical range. This effect has been found for certain tumours and less expressed for lung and kidney tissue in vivo [42, 43]. The modelling of normal tissue dose response is most notably affected by this effect for radiation pneumonitis where the optimum dose distribution may be highly dependent on the fraction number.

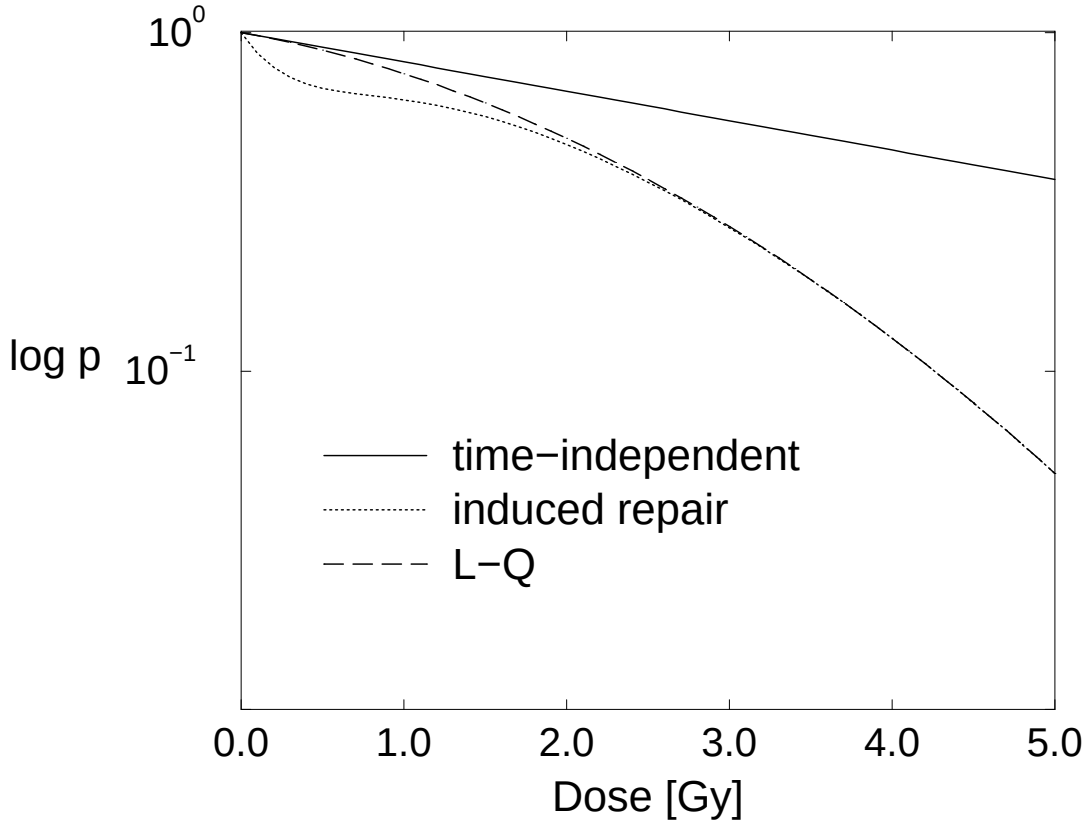


Figure 3.1: *The probability of cell survival as a function of single fraction dose for a fractionated treatment. The time-independent model does not take into account any 'latency' effects. The L-Q model accounts for incomplete repair of DNA damage which results in an apparently higher cell sensitivity for larger dose fractions. The induced repair model assumes that the DNA repair mechanisms are inactive if the damage stays below a certain threshold, which also results in an apparently higher cell sensitivity for very small dose fractions.*

Although treatment morbidity could be reduced by smaller doses per fraction, this should not lead to a longer course of treatment. The reason is that the number of tumour cells is replenished due to repopulation with the effect that towards the end of the treatment

a growing share of the daily dose is needed to decimate the clonogenes to the number of the previous day. Repopulation with a constant rate $\gamma$ leads to

$$p = \exp\left(\gamma N_{fx} - \alpha D(1 + \frac{\beta}{\alpha}\frac{D}{N_{fx}})\right) \tag{3.9}$$

in combination with the L-Q-model. More elaborate models with accelerated repopulation may be used [44, 45].

The adiabatic treatment of the time evolution allows to construct time-independent tumour and normal tissue dose response models in the following three sections. For want of relevant models for acute reactions, these cannot yet be included adequately into the optimization. For these complications, both the biological constraints and the generic DVH constraints can be utilised to express the clinical evidence.

## 3.4 Tumour Control

The failure of radiotherapy is the survival of a single clonogenic tumour cell. While this assumption is certainly on the safe side, modelling the expected value of surviving clonogenic cells for optimization purposes can be perilous because the spatial distributions of cell density and cell sensitivity have to be known. If some partial volume of the tumour is assumed to show a better response to radiotherapy than the residual, the optimum solution will increase the dose to the latter at the expense of the dose to the former, especially if dose-limiting constraints are present. Hence, an attempt to model the local probability of cell survival should excercise caution.

The efficacy of radiation depends on several factors which influence cell survival: the spatial distributions of cell sensitivity, oxygenation status, cell doubling times and cell density. All of these quantities are difficult to determine in practice, so that some assumptions have to be made to bias the result towards the safe side.

The objective measure can be found from the standard Poisson model of TCP [46, 47] to be

$$f(D(\vec{x})) = \rho(\vec{x}) \exp(-\alpha(\vec{x})D(\vec{x})) \tag{3.10}$$

where the spatial dependence of the *relative* cell density $\rho$ and the sensitivity $\alpha$ is indicated. Of course, the cell survival probability may be amended by the factors introduced in the previous section. Inter-patient heterogeneity has little effect on the objective measure although values for $\alpha$ may vary significantly. However, only relative values of the objective function are of concern. Quite to the contrary, the computation of absolute TCP values would depend very sensitively on inter-patient variability of $\alpha$.

The iso-effect is the homogeneous dose to the total volume which yields the same expected value of surviving clonogenes for an average sensitivity $\alpha'$

$$D_{\text{eff}} = \frac{-1}{\alpha'} \log\left(1/V \int_V f(D(\vec{x}))\, dx^3\right). \tag{3.11}$$

This quantity is of merely informative character and does not affect the optimization.

## 3.5 Serial Complications

For many serial complications, especially if they show a predominantly local dose response mechanism, a generic phenomenological model applies. This model is discussed in Appendix B. The objective density

$$g(D(\vec{x})) = \left(\frac{D(\vec{x})}{d_0}\right)^k \tag{3.12}$$

follows from this model with some reference dose $d_0$ and the volume effect parameter $k$. The iso-effective dose is the $k$-norm of the dose distribution

$$D_{\text{eff}} = d_0 \left(1/V \int_V g(D(\vec{x}))\, dx^3\right)^{1/k}. \tag{3.13}$$

The volume effect parameter $k$ determines the steepness of the dose response. The greater $k$, the less dose tolerance can be gained from a reduction of the irradiated volume, see figure 3.2.

This objective measure takes into account inter-patient variability directly by its functional form, see Appendix B. A special trait of this objective is that the volume effect does not depend on the iso-effect with the result that the typical shape of the DVH does not vary with the iso-effective dose. This facilitates an intuitive use of the constraint.

The model does not apply to inhomogeneous tissues, like the vascular structure or the heart. It may be used for complications involving damage to blood vessels, but in this case the whole volume would have to be assumed homogeneous. This leads to an overestimation of radiation effect for the embedding tissue matrix. In view of clinical safety these possible minute overall gains in dose tolerance should not be exploited anyhow.

## 3.6 Parallel Complications

Following the definition of parallel constraints, the dose to some partial volume of the organ in question can be arbitrarily high without causing this particular complication (other complications of the same organ may well be triggered!). If the complication is a genuine loss of function as for the parotids or the liver, this subvolume is termed 'functional reserve'. In other cases, like pneumonitis it appears more favourable to think of a 'critical damage' if loss of function is just a consequence of the complication, and if the mechanism involves the entire organ or body.

In any case, the requirement to spare a certain fraction of the organ volume leads to a non-convex constraint. In principle this may lead to a non-convex Lagrange function, but the seriousness of the problem depends crucially on the nature of the objective measure. Since the integral of the objective density over the sub-critical fraction of the volume has to be finite even for arbitrarily high doses, the objective measure has to be a finite function and thus be of sigmoidal shape. For mildly non-convex functions such as introduced here, it

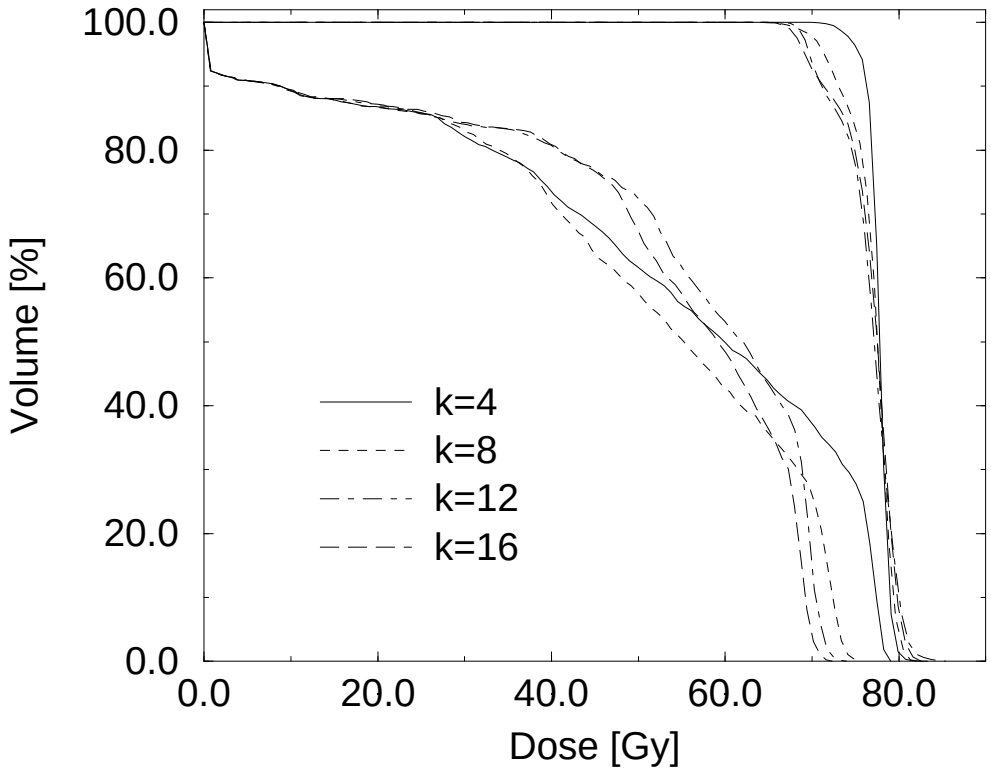


Figure 3.2: *A prostate case study with an overlap of rectum and planning target volume. The volume effect parameter was varied from* $k = 4, 8, 12, 16$*, i.e. from a rather pronounced tolerance against hot spots to a very distinct threshold behaviour. The iso-effect in all cases was* $D_{\text{eff}} = 65$ *Gy. Whereas a significant volume effect allows to treat the overlap region to the same dose as the remaining PTV, a strict volume threshold results in an underdosage in the PTV, yet better sparing of the rectum. The case shown here is a clinical example where trade-offs between PTV and rectum often have to be made because of the proximity of these volumes.*

may be conjectured that the Lagrange function is largely convex, and due to the degeneracy saddle points and local minima are virtually undetectable.

The functional form of these local dose-response mechanisms is extremely difficult to determine. The choice here is guided by mathematical expediency rather than biological insight

$$g(D(\vec{x})) = \left(1 + \left(\frac{d_0}{D(\vec{x})}\right)^k\right)^{-1} \tag{3.14}$$

where $k$ and $d_0$ are parameters which determine the shape of the sigmoidal *logit*-function. The sigmoidal shape of the local dose response can be supported in case of pneumonitis by experimental findings [48, 49], yet the logit function is no more than a template. In fact, for photon therapy the shape of the sigmoidal function is not of great importance. Recall that the ray derivative is the dose-weighted mean of the local response: the value of this

integral depends essentially on the fraction of the volume below and above the threshold dose $d_0$, where the response is close to 0 respectively close to 1.

The sigmoidal shape of the local dose response if sufficient to model a parallel complication mechanism with its corresponding volume effect. The heuristic argument may be linked to biological models by assuming that the complication mechanism is a critical process similar to a phase transition with the mean damage

$$\nu = \int_V g(D(\vec{x}))dx^3 \tag{3.15}$$

as critical parameter. Taking the analogy further, one obtains

$$\mathrm{NTCP}(D) \propto \left(\frac{\theta - \nu}{\theta}\right)^{-\kappa} \tag{3.16}$$

with some critical exponent $\kappa$ and some critical damage $\theta$. The global coupling which was assumed in the phase-transition model leads to a prefactor $(\theta - \nu(D))^{-1}$ by virtue of the transformation of the functional into an objective density eq.(2.19), see also Appendix A. This prefactor depends on the current dose distribution and has to be updated during optimization. It can be absorbed into the Lagrange multiplier and handled during the search for the proper multipliers. This is reflected in the update rule for Lagrange multipliers in chapter 6.2.

The iso-effect is the mean damage

$$\nu_{\mathrm{eff}} = 1/V \int_V g(D(\vec{x}))\, dx^3 \tag{3.17}$$

which is bounded by $0 \leq \nu_{\mathrm{eff}} < 1$. Although this could be inverted into a homogeneous total organ dose, we refrain from this step because there is no clinical experience for total irradiation of organs expressing a parallel complication mechanism.

A parallel constraint aims to redistribute the dose into low dose subvolumes and subvolumes above the threshold dose. As a consequence, the dose per fraction varies significantly across the organ volume. If this organ shows low-dose hypersensitivity, the two effects counteract. This is most pronounced for radiation pneumonitis. It is very unsatisfactory that most clinical findings do not take into account this (fairly recently discovered) effect with the consequence that to stay close to clinical experience, hypersensitivity should not be used for optimizations involving lung tissue for the time being.

## 3.7 Generic DVH Constraints

In case the treatment objectives cannot be formulated as a biological model or the dose-volume statistics as represented in the dose-volume histograms are to be manipulated directly, two generic dose-volume histogram constraints can be utilised. They are the equivalent of the serial and parallel constraints with the difference that they do not balance

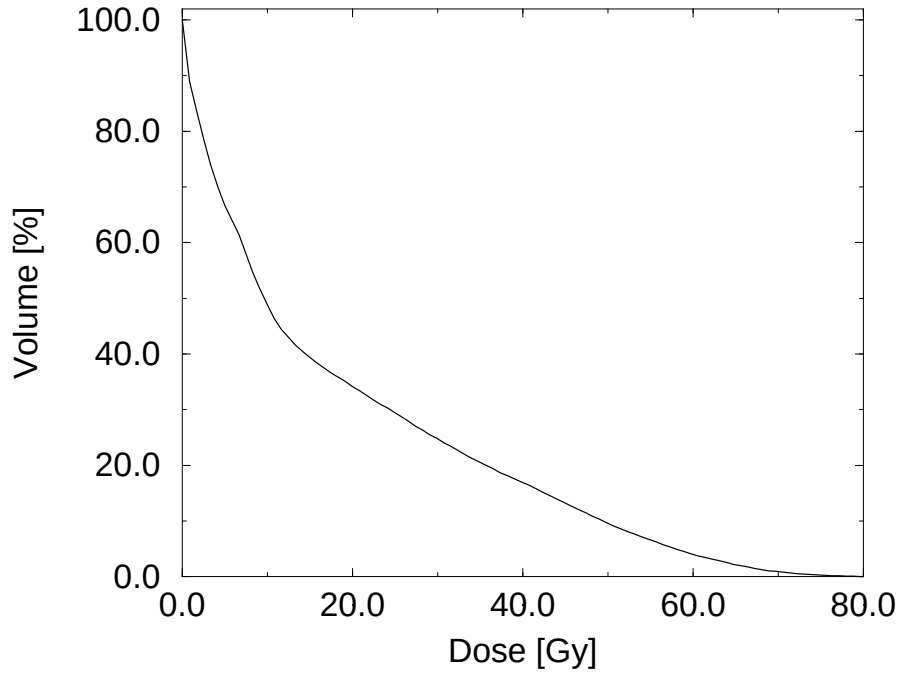


Figure 3.3: *A typical DVH for a parallel complication mechanism, in this case pneumonitis. The threshold dose for the sigmoidal local dose-response was set to 15 Gy, the iso-effect (the mean damage) was set to 30%. The DVH shows the characteristic kink at about these values; some fraction of the volume smaller than 30% can be exposed to a higher dose.*

high and low dose volumes according to some inherent dose-volume relation. In other words, the biological volume effect depends on the current threshold dose and iso-effect defined by the planner.

The equivalent of the serial model is the quadratic overdose penalty

$$g(D(\vec{x})) = \Theta(D(\vec{x}) - d_0)(D(\vec{x}) - d_0)^2 \tag{3.18}$$

with maximum dose threshold $d_0$. $\Theta(.)$ is the Heaviside function. The corresponding iso-effect is the rms-overdose

$$d_{\text{eff}} = d_0 + \sqrt{1/V \int_V g(D(x))\, dx^3}. \tag{3.19}$$

This constraint can be used if the maximum dose to the target volume should be limited.

The equivalent of the parallel model is the volume restricted overdose constraint. This constraint limits the volume to $V_0$ per cent which receives in excess of $d_0$ Gy. The objective measure is heuristically defined as

$$g(D(\vec{x})) = \begin{cases} 0 & : \quad D(\vec{x}) < d_0 - 1 \\ 1/2\,(D(\vec{x}) - d_0 + 1)^2 & : \quad d_0 - 1 \leq D(x) < d_0 \\ 1.5 - (D(\vec{x}) - d_0 + 1)^{-1} & : \quad d_0 \leq D(\vec{x}) \end{cases} \tag{3.20}$$

The iso-effect is the volume which receives more than $d_0$ Gy.

The characteristics of the physical constraints can be seen in figure 7.1.

# Chapter 4

# Physical Constraint Modelling

In the sense that planning is regarded as comprehensive treatment optimization, physical modelling can achieve more than ensure the practical feasibility of fluence profiles. The optimization touches the clinical fields of efficiency, safety and quality assurance together with the physical fields of dosimetry, device design and error tolerance. The radiotherapy optimization problem affords substantial latitude to shape the result in an expedient fashion.

Whilst a number of 'hard' limitations have to be dealt with rigorously by means of constraints or dedicated radiation transport models, neither the necessity nor the gain of other stipulations may be easily justified or evaluated. Nevertheless these 'soft' limitations appear desirable from a pragmatic point of view. Contrary to real constraints the soft constraints may be violated if they compromise the result unduly; therefore they are characterised by their effect on the optimum objective function.

## 4.1 Minimum Fluence Constraints

The restriction of fluence space to non-negative fluences was implicit in all prior considerations. The practical implementation of this constraint can be done by means of barrier functions or projection operators whereby the latter are not suitable for all optimization algorithms. The minimum fluence constraint translates into constraints of the ray weights of the current ray basis. Only in the special case that these fluence distributions do not 'overlap' and are positive semi-definite, these ray weights have to be non-negative. In all other cases, the ray weights may be negative as long as the non-negativity of the total fluence is ensured.

Depending on the treatment technique, the minimum fluence at a point within the field perimeter may be a function of maximum fluence across the field and collimator transmission or solely dependent on absolute parameters like dose rate and leaf velocity. The latter applies to the sliding window dynamic MLC technique [50, 51], whereas the former applies to the static MLC (step&shoot) technique [52, 53] with or without physical compensators or physical compensators alone. Also, it may be necessary to employ Megavoltage portal

imaging to control patient setup which increases the minimum fluence within the field outline further. At any rate, the minimum fluence is never exactly zero. This facilitates the use of barrier functions.

### 4.1.1 The Method of Barrier Functions

These functions are added to the objective function and are devised to increase with the violation of a constraint. In case the minimum violates the constraint, it is shifted towards the feasible set. These barrier functions are only exact if an associated multiplier is let to infinity which can spoil the condition of the problem; if a mild violation of the constraint is acceptable, it is preferrable to keep this multiplier as small as possible. Commonly, a mild violation of a positive minimum fluence constraint does not have any impact on the dose distribution.

The barrier function which showed the best performance was a fourth-order polynomial

$$G = \sum_{i=1}^{n} \Theta(\phi_{\min,i} - \phi_i)(\phi_i - \phi_{\min,i})^4 \tag{4.1}$$

where $\phi_{\min,i}$ is the minimum weight and $\Theta(.)$ the Heaviside step function. The corresponding multiplier is set to the sum of all normal tissue Lagrange multipliers. The minimum weight $\phi_{\min,i}$ may change during the optimization if it is a fraction of the maximum fluence of the corresponding field. If the basis consists of overlapping rays, the minimum weight may become negative. In this case, the barrier applies to the total fluence of all rays.

### 4.1.2 The Method of Projection Operators

Projection operators map a fluence distribution to the closest feasible fluence distribution; in this case any ray weight which is below the minimum weight is set to the minimum weight. Projection operators are only used as a fall-back because they disturb the convergence of the main optimization algorithm used here. In addition to the projection, the ray which violated the minimum weight constraint can be removed from the active set of optimization parameters and its weight value fixed to the minimum.

## 4.2 Profile Smoothing as a Soft Constraint

The optimum fluence profiles are smooth in the sense that they do not show erratic variations (noise) and a clear relation of fluence modulations to anatomical structures. Depending on the grain of the decomposition of the fields into rays the fluence distribution may nevertheless show significant jumps.

Due to the high degeneracy of the problem, a serious perturbation of the solution due to noise may occur if the convergence limits of the algorithm are not set very low. Although the problem could be solved by lower convergence limits, this method is inefficient because

the dose distribution will not improve despite the longer computation times. An additional soft constraint which acts as a 'noise filter' would be a preferable solution. Technically, such a constraint reduces the search space to the 'desirable' solutions. Appendix C is concerned with the definition of 'desirability' by virtue of a soft constraint which minimises the area of the surface given by the two-dimensional fluence profile.

Another source of erratic fluence jumps are discretisation and dose computation artefacts. It may be conjectured that these artefacts cannot be controlled by a soft constraint as used here. The performance of such a device may deteriorate if it is stretched to balance these artefacts since there is no *a priori* reason to assume that a fluence profile should be a minimum surface.

The smoothing constraint multiplier is by default set to about $10^{-4}$ of the objective function at the minimum. This value was found experimentally as the onset of dominance of the smoothing over the 'anatomical' fine structure of the profiles. At an acceptable convergence threshold the elements of the gradient of the Lagrange function are in the range of $10^{-3}...10^{-5}$ which means that the smoothing constraint and the biological constraints including the objective function are of equal magnitude at the optimum.

If the smoothing constraint multiplier is increased further, the optimum is more and more determined by the minimum surface condition. While this could be used to generate more clinically expedient fluence profiles, it appears arbitrary because it is not specific to treatment technique. However, the smoothed fluence profiles are almost always feasible for a dynamic MLC application together with a suitable minimum fluence constraint; to enforce feasibility, an additional projection operator would be necessary. Minimum surface smoothing should never exceed the role of a numerical tuning device.

# Chapter 5

# Radiation Transport Modelling

The conflicting requirements of dose computation for radiotherapy optimization are not equally well met by any single algorithm. Whilst phenomenological models can be very expeditious by specific approximations, the highest accuracy can only be achieved with Monte Carlo methods. The particular requirements of IMRT lend a new angle to the assessment of the merits of Monte Carlo methods. Small, irregular and MLC constrained fields make heuristic dose models difficult to develop. Also, adverse effects of electron scatter at low density interfaces in the patient can be compensated for by modulations of the primary fluence.

It was a design objective of the radiotherapy optimization algorithm to combine the benefit of Monte Carlo and phenomenological models. Under the constraints of clinical conditions a purely Monte Carlo based algorithm does not appear feasible at present. The hybrid method proposed here has an advantage of a few orders of magnitude in computation time over a pure Monte Carlo approach with no conceivable practical disadvantage. The technical considerations in the following section and appendix D apply equally to the hybrid and the pure Monte Carlo method.

## 5.1 Technical Aspects

Some practical limitations and numerical issues have such a fundamental impact on the performance of a dose computation model and its compatibility to the optimization engine that they need to be considered before the formulation of the method. They affect primarily the discretisation of the dose space $\mathcal{D}$ and the design of the ray basis $\mathcal{B}$.

The sampling of the dose distribution with a limited number of points is a very intricate problem. After all, it is the dose distribution of a narrow ray which has to be sampled for a large number of independent rays. For a ray diameter of a few millimeters this results in a sampling point density of more than 1 per cubic millimeter, which leads to a presently intractable problem of $10^7$ to $10^8$ sampling points. It becomes already clear that the dose distribution will always be undersampled within the limits of present day computer hardware.

It appears straightforward to reduce the number of sampling points and concentrate them in regions of interest or high dose gradients. A number of problems are associated with this approach. Firstly, each single ray has to be modelled in a sound way, which can not be achieved by taking samples only in regions of large variation of the beam dose, let alone the total dose distribution. With IMRT, it is rarely the case that the position of all in-field fluence gradients can be guessed from the start. Furthermore, the ray must be sampled by an equivalent density of points in each subvolume of the patient in order that all objectives are equally represented in the computation of the ray derivatives. This rules out inhomogenous *and* anisotropic (if couch angles are permitted) sampling in a large part of the patient volume. Secondly, the dose computation grid must be expedient for voxel-based Monte Carlo computation methods. This excludes hexagonal grids and narrows down the choice to cubic grids.

By its nature, dose is a density. Thus, each sampling point represents an integral of absorbed energy over some small volume around this point, divided by the size of this volume. This equals the way a voxel-based Monte Carlo algorithm computes the dose. Since the ray derivatives are weighted dose integrals, a volume based definition of the dose computation is only natural and ensures energy conservation.

At contrast, phenomenological models could be based on point doses which can cause serious definition problems in combination with the undersampling of dose space[1]. Dose computation is numerically more stable if it is based on an average over some volume. The larger this volume, the less prone to discretisation artefacts is the dose computation. Thus, to some extent it is possible to alleviate the problem of undersampling at the cost of some blurring of the gradients of the dose distribution. Notice that these sampling volumes need not be space filling; it is entirely sufficient if the average dose is computed in some small ball centered at the vertex of a cubic grid.

The stable computation of the ray derivatives is of paramount importance for the performance of the radiotherapy optimization algorithm (see appendix D). In essence, the dose computation has to ensure that all rays are represented in an equivalent fashion. The design of the ray basis can assist in the solution of this problem of phenomenological dose models. The ray basis was devised to construct the space of all practically achievable fluence distributions. This precludes the use of piecewise constant, discontinuous functions such as the characteristic functions on some regular decomposition of the cross-section of the beam. Of course, it is virtually impossible to define the shape of the proper penumbra of some small field for arbitrary positions in the field. However, the essential quality of a ray basis is not accuracy, but self-consistency and completeness.

One can construct a self-consistent ray basis by the notion that a certain number of rays have to add up to a homogeneous fluence. Let $h(x, a)$ be the one-dimensional fluence

[1]Consider a ray whose fluence rises from say 20% to 80% within 1 mm. If the dose computation were point based, the *ray derivatives* would be extremely sensitive to geometrical variations; in this case by up to a factor 4 for shifts within a range of 1 mm.

profile of a ray with width $a$. Then by

$$h(x,(2n+1)a) = \sum_{i=-n}^{n} h(x+ia,a) \tag{5.1}$$

the shape of the ray can be determined if for some $n = n_r$

$$\forall x \in [-a/2, a/2] : h(x,(2n_r+1)a) = 1 \tag{5.2}$$

where $n_r$ is the smallest such number. The number $n_r$ describes the number of rays necessary to reconstruct a locally homogeneous dose. It can be seen that $h(x,a) = 0$ for $|x| > n_r a$. The fluence distribution of a ray may then be considered as the product of two functions $h(x,y,a,b) = h(x,a)\,h(y,b)$ for some width $a,b$. If the rays are arranged on a rectangular grid with gridlines at multiples of $a,b$, the ray basis consists of functions $h(x-ia, y-jb, a, b)$ centred at each vertex $(i,j)$ of the grid. It is important to notice that the property of self-consistency translates to the ray dose by virtue of the linearity of the energy absorption operator $T$.

It is more expedient to create a self-consistent cross-profile of the ray *dose* by some function which complies to eq.(5.1) rather than computing the dose distribution for such a ray *fluence* profile. The hypothetical ray fluence is then the 'inverse' of the such defined ray dose; since the fluence distribution of a ray is never needed, it need not be determined. The ray dose exhibits rather shallow radial gradients because the effects of both photon penumbra and electron transport are incorporated into the model. The shallow gradients of these self-consistent ray doses also help to overcome the problem of the sensitivity of the dose sampling. The concept of self-consistent ray doses does not account in detail for secondary photon scatter or diffuse scatter sources in the linac head. These contributions to the total dose are considered implicitely in the ray dose distributions. Thus, in numerical practice a ray dose will *not* represent a physical dose distribution of a small fluence element, but can be considered as a constituent of a physical dose distribution of an extended intensity modulated field. This approximation is valid in view of the small influence a modulation of the primary photon fluence on the fluence of secondary photons, and the huge computational effort to compute this phantom scatter.

For the phenomenological dose model in the next section, these two countermeasures against undersampling of the dose distribution and discretisation artefacts are used: dose averaging over small balls centred at the vertices of a cubic grid, and the use of a basis of self consistent rays which incorporate some penumbra into their fluence- and dose profile.

## 5.2 Finite Size Cone Pencil Beams

The most widely used algorithms in clinical routine belong to a family of methods for dose computation which are based on a convolution of (primary or secondary) photon fluence and dose deposition kernels which incorporate electron and photon interactions. In one

way or other, most convolution algorithms [18, 54, 55, 56, 57, 58, 59, 60, 61, 62] resort to approximations to meet the clinical computation time requirements. Under partial consideration of heterogeneities of electron density a 'dose spread kernel' is convolved with the incident photon fluence. Several algorithms [18, 59, 60, 61] collapse the convolution along the ray axis and consequently use a decomposition of the dose into a direct product of a function which varies with depth and a kernel which depends on the distance to the ray axis alone. These algorithms are usually termed 'pencil beam' methods.

In essence, this model equates dose with the kinetic energy released by primary photon interactions to electrons per unit mass (KERMA) and is only strictly valid in case secondary electron equilibrium holds. The influence of heterogeneities on the track of the primary photons is taken into account by density corrections to the radiological depth of a point rather than its geometric depth.

In addition to the conceptual approximations, for a fast implementation allowances have to be made. Most pencil beam algorithms use Fourier transform for convolutions which necessitates coordinate transforms and precludes the use of space variant kernels. Pencil beam convolution methods reach a global accuracy of a few per cent, with local aberrations due to electron scatter of up to ten per cent. The following implementation of a pencil beam model shares the same deficiencies yet pursues a different arithmetic method.

The challange for the dose computation is constituted by the need to compute and store the dose distributions of several thousand rays separately, not in the bulk of a field. In the same way as the conventional pencil beam kernel can be understood as the ray dose of a parallel photon flux of infinitesimal width, the ray dose distribution can be understood as the kernel of a finite size pencil beam algorithm. The convolution becomes the sum over all constituent rays, weighted with their respective weight.

To compute the ray dose distribution, a local dose average of the self-consistent ray profiles over the sampling volume $S$ (a sphere centred at the origin) has to be performed

$$d(\vec{r}, Z) = \int_S h(x - r_x, y - r_y, a, b)\, t(Z - z)\, dx^3 \tag{5.3}$$

where $h$ is the dose cross-profile and $t$ the depth-dose function at a source distance Z. This integral again has the form of a convolution. However, if Fast Fourier transform were used, this could only be done on a temporary coordinate grid which would have to be much finer than the dose sampling grid. The dose computation is very time critical, since the convolution would have to be performed some $10^7$ times for the typical treatment case. For this reason, tabulated convolution values would provide a significant gain.

It is expedient to reduce the dimension of the convolution table to the least possible number. Consider a conical ray is traced through a cubic dose grid. Let the dose at each vertex be the average over a ball with diameter equal to the grid pitch. Since the ray is conical and the sampling volume spherical, the dose at each vertex of the dose grid *does not depend* on the orientation of the ray. Indeed, there are only two relevant variables: the distance of the point to the source $Z$ and the distance of the point to the ray axis $r = |\vec{r}|$.

The dependence on radiological depth can be treated by a separate factor by virtue of the common approximation of depth-independent scatter kernels.

The dose to a point $\vec{x}$ can then be computed by

$$D(\vec{x}) = \frac{1}{Z(\vec{x})^2}\, t(\mathrm{RD}(\vec{x}))\; d_{\mathrm{CT}}(r(\vec{x}), Z(\vec{x})) \tag{5.4}$$

where the first factor corresponds to the inverse-square law, $t$ is the depth-dose curve for the given photon energy at the radiological depth RD of $\vec{x}$, and $d_{\mathrm{CT}}$ is the convolution table. The convolutions of the ray with the sampling spheres with respect to $R$ can be precomputed for all $Z$. The three dimensional integral can be reduced analytically to an elliptic integral which can be tabulated.

The ray dose profile $h(Z)$ can be determined from measurements or Monte Carlo simulations. A function which fulfills the reconstruction condition eq.(5.1) can be fitted to the data. The reconstruction number $n_r$ will be influenced by practical considerations. It may appear advisable to truncate the ray dose distribution at some distance from the central axis to save computation time during the optimization. Notice that the ray derivative is an integral over the support of the ray dose; if secondary photons were to be taken into account in their entirety, this integral would stretch over the whole patient volume.

The fit function which was used by default with $n_r = 2$ was

$$R(r,a) = \begin{cases} 1 - 3k_3 - k_1(\exp(k_2(x/a-2)) + \\ \exp(k_2(x/a-4)) + \exp(-k_2(x/a+2))) & : \quad |x| \le a \\ k_1 \exp(-k_2 x/a) + k_3 & : \quad a < |x| \le 2a \\ 0 & : \quad |x| > 2a \end{cases} \tag{5.5}$$

with

$$k_1 = \frac{1}{2}\left(\exp(-k_2) + \exp(-3k_2) - 2\exp(-4k_2)\right)^{-1} \tag{5.6}$$

$$k_3 = -k_1 \exp(-4k_2). \tag{5.7}$$

The parameter $k_2$ is determined by a depth-dependent fit to a Monte Carlo computed ray (see figure 5.1).

The drawback of this method is that the conical rays cannot be arranged to be space filling. The best compromise can be found by using a hexagonal discretisation of the fluence profiles. With this arrangement, about 5 per cent of the beam cross-section are not covered by rays. Since the rays overlap, the fluence is not zero in these regions. However, because the dose grid is generally too coarse to resolve the irregularity of the fluence, this does not cause problems. In some sense, the undersampling of the dose distribution is exploited by the dose model to gain time. The computation of a ray dose takes about 10 ms on modern computer equipment for a dose grid of $(2\mathrm{mm})^3$.

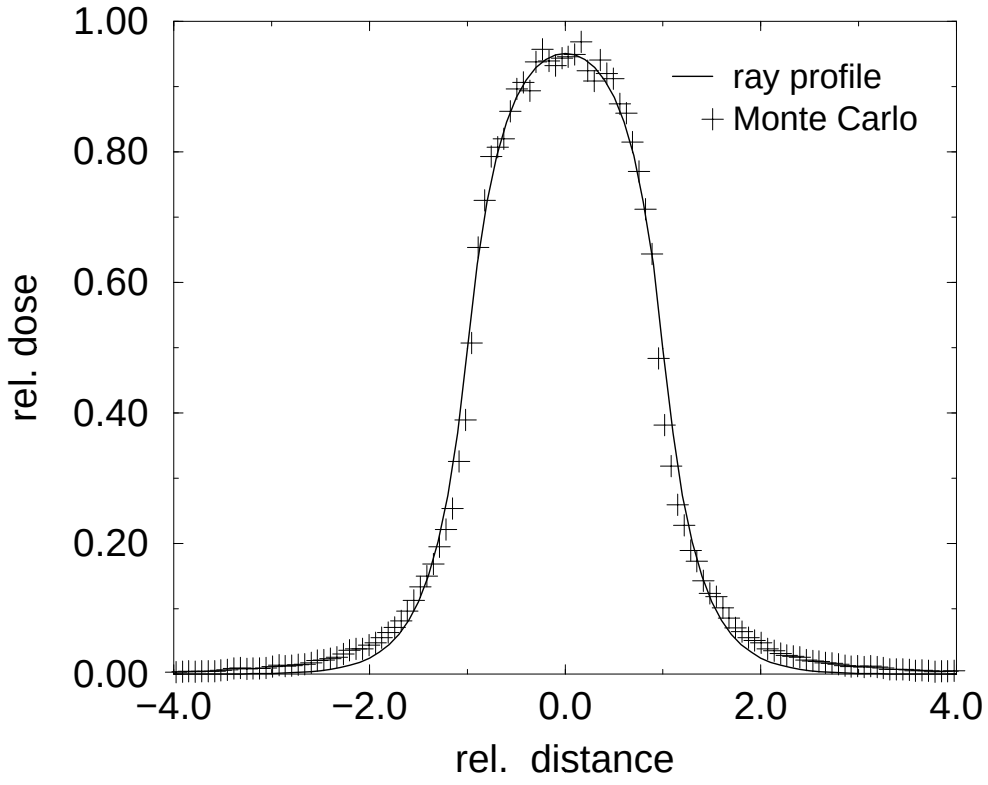


Figure 5.1: *A one parameter fit of the self consistent ray cross profile eq.(5.5) to a Monte Carlo computed ray dose for a conical ray (15 MV) of 5 mm diameter in 10 cm depth. An even better correspondence could be generated by using a second exponential for the tails [59].*

## 5.3 Intensity Modulation and Monte Carlo

The benefit of Monte Carlo dose computation lies in the fact that the entire path of the photons, from the source through the collimators into the patient, can be simulated; at the same time, a more homogeneous dose distribution can be delivered if electron scatter across low density surfaces is compensated for by primary fluence.

The rationale for separating the optimization process into a stage which allows interference by the therapist and a stage which runs the Monte Carlo dose engine is the high computational cost of solving a constrained problem as compared to an unconstrained problem. The meanderings of the algorithm during determination of the Lagrange multipliers $\lambda^*$ which solve the constrained problem would literally waste an enormous number of Monte Carlo particle histories. The first stage procures the proper set of Lagrange multipliers, so that during the second stage only minor adjustments of the multipliers have to be made.

In appendix D, the details of the IMMC technique are given. The method is essentially independent of the optimization algorithm and the Monte Carlo code. The implementa-

tion of the Monte Carlo code EGS4 follows [63]. The phenomenological algorithm used to compute the ray derivatives has to fulfill certain accuracy conditions to ensure that the descend property of the gradient optimization algorithm still holds. The very same conditions apply if a Monte Carlo algorithm were used to compute the ray doses. It must be noted that to obtain a statistical uncertainty in the order of the systematical error of the phenomenological dose model, a disproportionately high number of particles would have to be simulated; this for no obvious gain. Since the Monte Carlo dose computation has to include the collimators to bring to bear its full accuracy, the ray doses which were naturally computed without the collimators would not add up to the total dose distribution.

# Chapter 6

# The Optimization Engine

With the biological and physical modelling as described in the preceding chapters, radiotherapy optimization poses a large-scale optimization problem with nonlinear constraints, yet of a predominantly convex nature if the beam angles are fixed. Hence, if only fluence profiles for predefined beam directions are to be obtained, the most suitable optimization engine can be selected from a great variety of gradient based algorithms.

There is a number of independent classes of algorithms for the solution of such an optimization problem [21, 64]. Since execution time is of prime importance for the clinical application of an radiotherapy optimization algorithm, allowances have to be made where possible. The most elementary approach aims to determine the Lagrange multipliers and minimize the Lagrange function straight away. Although this method has significant methodological deficiencies, it is amenable to many accelerating heuristics. The more sophisticated SQP [21] (sequential quadratic programming) class algorithms await thorough testing in the context of this development and may need considerable fine-tuning despite their conceptual superiority to match the speed of the simple method of Lagrange multipliers [19].

The method of multipliers transforms the constrained optimization problem into a sequence of unconstrained sub-problems with varying Lagrange multipliers. The algorithm tries to construct the Lagrange multipliers $\lambda^*$ which solve the problem from the solutions of the trial sub-problems. The next section describes the algorithm which solves the unconstrained problem, section two is devoted to the heuristics which are used to estimate the Lagrange multiplier updates.

## 6.1 Solution of the Unconstrained Sub-Problem

A great number of algorithms for the solution of unconstrained optimization problems has been devised [21, 22, 23, 65]. The special requirements of the radiotherapy optimization problem narrow down the choice considerably. Due to the size and structure of the problem (number of parameters $\approx 10^3 \ldots 10^4$) the computation of the Hessian matrix of second derivatives is very time consuming; the memory requirement of some 64 MB does not

preclude its use on modern computers. Likewise, the computation of a gradient vector is much more expensive than an evaluation of the objective function. However, even the latter usually requires a complete recalculation of the current dose distribution.

These considerations influence the choice of algorithm although the total performance can only be assessed in numerical trials. The special nature of the radiotherapy optimization problem with its high degeneracy certainly affects algorithms which make use of the Hessian more than simple gradient techniques. To some extent the statement is valid that the algorithm predominantly has to separate out and solve the non-degenerate sub-problem to be most efficient.

A number of algorithms was tried: steepest descend algorithms with fixed and variable step size, quasi-Newton algorithms with and without line searches and a variety of conjugate gradient algorithms. In general, the Polak-Ribiere [65, 66] method of conjugate gradients was the most successful for reasonable starting points. Although other conjugate gradient methods were tested [67], none could gain a clear advantage over this method. The algorithm was implemented with a line search according to Brent [65, 68]. The line search does not use gradient information but multiple objective function calls instead. Since the ratio of computational cost of a gradient computation to an evaluation of the objective function is approximately 10:1, a slightly higher number of objective function calls is acceptable. Restarts in the direction of steepest descend were performed if the maximum change in a ray fluence exceeded a certain threshold, thereby indicating that the search directions were no longer aligned with the principal descent directions of the objective function. The termination criterion is met if the fractional change

$$\frac{L(\Phi^k) - L(\Phi^{k+1})}{L(\Phi^{k+1})} < \epsilon \tag{6.1}$$

falls below a threshold $\epsilon \approx 10^{-3}$. Most notably, the conjugate gradient method needed a number of iterations which was at most 5 per cent of the number of optimization parameters[1]. This indicates that the conjugate gradient algorithm separates the non-degenerate sub-problem very well.

For rather crude starting points, like homogeneous fluence distributions, the search direction updates become dominated by the directions of largest initial decrease. In this case, a steepest descent algorithm can descend faster than a conjugate gradient method. This is a consequence of the fact that the Hessian matrix varies strongly during the first iterations so that the conjugacy of search directions, which relies on an invariant Hessian, cannot unfold its potential. Once a satisfactory starting position for the conjugate gradient algorithm has obtained, the steepest descent method is terminated. This scheme reflects the varying ratio of efficiency to computational cost for gradient computations and objective function evaluations during the approach to the minimum.

[1]If the Hessian were not singular, the number of iterations would be at least equal to the number of optimization parameters.

The quasi-Newton method of [69] which was successfully applied to radiotherapy optimization did not perform well in this setting. The reason is that this problem has a highly singular Hessian close to the minimum and the Hessian has more off-diagonal entries. Approximating the matrix by its diagonal elements degrades the curvature information to an extend which disturbs rather than accelerates convergence. Other quasi-Newton schemes like BFGS or DFP [19, 22, 23] could potentially yield a small gain for the solution of the unconstrained problem. However, the greatest acceleration will result from replacing the method of multipliers with a superior handling of the biological constraints.

## 6.2 Determination of the Lagrange Multipliers

The essential drawback of the method of multipliers is the cumbersome search for the Lagrange multipliers $\lambda^*$ which solve the constrained problem. The overall convergence properties of the algorithm are determined by the convergence of a set of estimated Lagrange multipliers $\lambda^l \to \lambda^*, l = 1, 2, \ldots$ in an outer loop of the optimization which solves the unconstrained problem at each pass for the current Lagrange multipliers $\lambda^l$. The convergence of the Lagrange multipliers is frequently only linear, so that a significant amount of time is lost in this outer loop.

However, the special properties of the radiotherapy optimization problem allow to apply this technique despite its inefficiency. Firstly, due to the convexity and monotonicity properties of objective and constraint functions, the latter can be seen as their own barrier functions[2]. Secondly, due to the high degeneracy of the problem, the value of the objective function is very insensitive to the Lagrange multiplier estimates, although the corresponding fluence may be influenced comparatively strongly by variations of the multipliers. As a consequence, the method of Lagrange multipliers is capable of delivering a feasible and acceptable solution in a reasonable time, whereas the proper solution of the constrained problem is nigh impossible on clinical time scales with this method. An Augmented Lagrangian technique [21, 22, 70, 71] may provide an advantage over the simple Lagrange function, yet this remains to be tested.

The update of the current Lagrange multipliers employs a first-order rule which is applied at the termination of the inner loop, the solution of the unconstrained problem. The update rule derives from the sensitivity relation eq.2.11. The estimate of the new Lagrange multipliers is accurate to $\mathcal{O}(|\Phi^l - \Phi^*|^2)$ where $\Phi^l$ is the fluence which minimizes the Lagrange function with the multipliers $\lambda^l$. Let $F(d_f), G(d_g)$ be the objective and a given constraint function and let $d_{f,g}$ be the iso-effective homogeneous dose of $D$. Let $F'(d_f)$, $G'(d_g)$ be derivatives of $F, G$ with respect to the iso-effective dose $d_{f,g}$. At the

[2]This is in fact the point of view of dose-based optimization where quadratic penalty functions are employed to produce a slightly perturbed solution of a problem which is restricted by maximum dose constraints on normal tissues.

minimum

$$F'(d_f) \approx \lambda^l G'(d_g) \tag{6.2}$$

for the current multiplier $\lambda^l$ (which is only a number here). Similarly, for the constrained solution,

$$F'(d_f^*) \approx \lambda^* G'(d_g^*). \tag{6.3}$$

Since $F$ is an exponential function, one has $F'(d_f^*) = F'(d_f)\exp(-\alpha(d_f^* - d_f))$ . By $G(d_g^*) = 1$ and dividing eq.(6.2) and (6.3) one obtains

$$\frac{\lambda^*}{\lambda^l} = \exp(-\alpha(d_f^* - d_f))\frac{G'(d_g)}{G'(G^{-1}(1))} \tag{6.4}$$

where $d_f^* - d_f$ can be set to 0, or be replaced with $d_g^* - d_g$. This equation defines the update $\lambda^{l+1} = \lambda^*$ for all constraints for which an iso-effective dose can be defined.

For parallel complications, where some global interaction has to be taken into account, the update is governed by the rule

$$\frac{\lambda^*}{\lambda^l} = \exp(-\alpha(d_f^* - d_f))G(D)\frac{\Theta - 1}{\Theta - G(D)} \tag{6.5}$$

in case the global coupling is assumed to show a phase-transition like behaviour at the mean damage $\theta/\nu^* = \Theta$, where $\nu^*$ is the (prescribed) upper bound to the mean damage $\nu$. Generally, all coupling mechanisms may use the rule

$$\frac{\lambda^*}{\lambda^l} = \exp(-\alpha(d_f^* - d_f))\frac{\nu}{\nu^*} = \exp(-\alpha(d_f^* - d_f))G(D). \tag{6.6}$$

Since all normal-tissue constraints are inequality constraints, the Lagrange multipliers are always non-negative. If a constraint does not belong to the active set, its multiplier is 0. In practice, the constraints are removed from the active set if their multiplier is below some threshold after the multiplier updates. The multiplier updates are sensitive to the termination criterion of the inner optimization loop which naturally delivers only an approximation to the solution $D^l$ for a given set of multipliers $\lambda^l$. If the updates do not appear to be stable, the convergence threshold of the inner loop is forced down to provide better estimates. The algorithm terminates if the result of an inner loop is feasible with respect to the constraints and the multiplier updates are within certain bounds $[1-\epsilon, 1+\epsilon]$. The penalties for the barrier functions and soft fluence constraints are not subject to these update rules; their handling is described in chapter 4.

# Chapter 7

# Applications

The current state of biological and clinical knowledge makes it impossible to provide sufficient models or data for a full-blown biological optimization which does not require the definition of treatment objectives by the therapist. Although there is still some uncertainty as to the applicability of the concepts of time/fractionation and volume effects to all tissues, it can be shown that these effects do have an impact on the optimum dose distribution. The method of evidence based biological optimization offers a radically different design of clinical studies, with an intuitive definition of treatment objectives in terms of limited morbidity. If advantage is taken of the individual patient's potential for dose escalation, clinical radiotherapy gains an entirely new quality. Also, the physical modelling can be shown to improve on the clinical quality of treatment plans. Of the following sections each highlights a particular aspect of the optimization model with a clinical example in hand.

## 7.1 Physical or Evidence-Based Biological Optimization?

Within the confinements of standard conformal radiotherapy, the prescription of the target dose and the limits of normal tissue tolerance could be outlined by a few points in the DVHs of the respective volumes. At best, the degrees of freedom of the treatment technique afforded a feasible solution to the treatment prescription. The introduction of IMRT removed these limitations and led to a multiplication of the possibilities to shape the dose distribution. Whereas for standard techniques a few limiting points in the DVH were sufficient to define the treatment objectives, for IMRT these few points leave substantial uncertainty about the shape of the dose distribution. It is an inevitable consequence of the potential of IMRT, that the rules which lead to the prescription of the treatment doses have to be incorporated into the formulation of the objectives of the algorithm. Only then the arbitrary definition of DVH limiting points can be abandoned.

In essence, the definition of the treatment objectives has to offer sufficient means to specify the quality of the dose distribution. This can be done by physical or biological

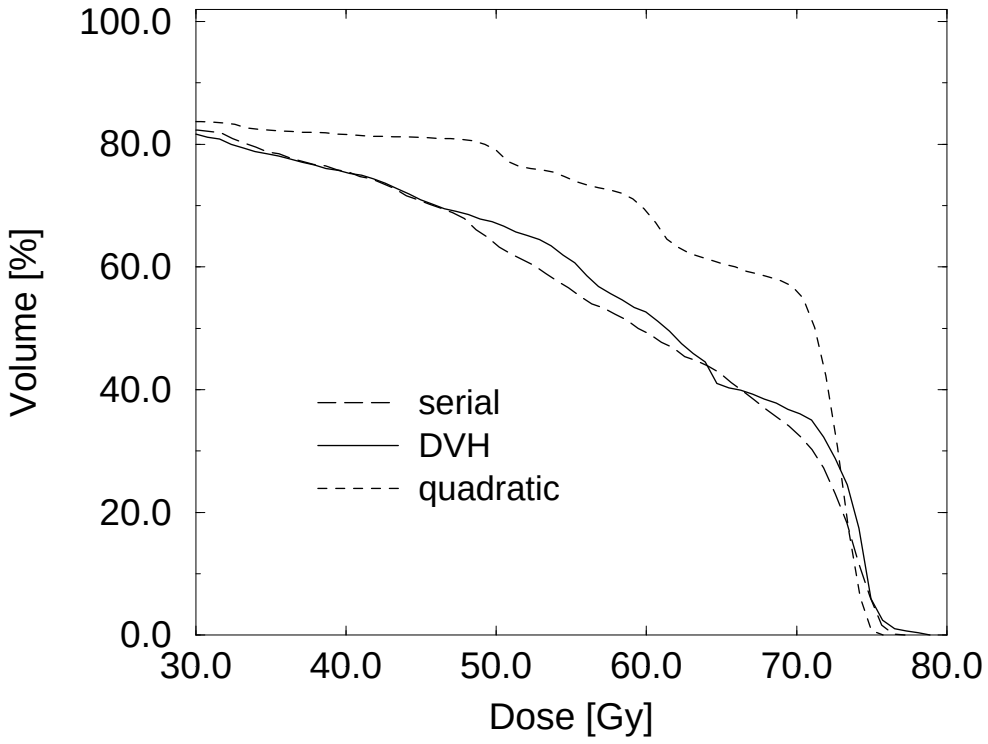


Figure 7.1: *A comparison of three normal tissue constraints for the rectum of a prostate case, as in figure 3.2. The dose-volume constraints result in a kink in the DVH at 65 Gy, 40% of the volume and a maximum dose of about 80 Gy. The quadratic overdose penalty does not take effect for doses below its threshold of 72 Gy so that the dose prescription is considerably exceeded. Both generic DVH constraints do not model the volume effect in a consistent way. The biological serial constraint with a volume effect parameter* $k = 8$ *delivers a DVH which is in accordance with the clinical experience that went into the definition of the DVH limiting points without further stipulations on the dose distribution.*

indices, yet in either case *implicit* assumptions about the dose-response of each tissue are made. In case a dose-based penalty function is used which is common to all organs at risk, *all organs are assumed to have the same dose-response mechanism.* The definition of DVH limiting points is highly arbitrary and rarely sufficient. As an example, figure 7.1 shows a comparison of dose-based and biological objective functions. This figure should be seen in combination with figure 3.2, page 27, which was generated from the same clinical prostate case. The latter figure demonstrates how the concept of dose-volume effect which is implicit in many clinical decisions influences the shape of the DVH. In the figure, one DVH is constrained by three limiting points (65 Gy, 40%), (72 Gy, 30%) and (75 Gy, 5%). It can be deduced from the stair stepped appearance of this DVH that these three points are not sufficient to define the treatment objectives. At the same time, the prescription of a volume-effect and a tolerance dose for the biological constraint leads to a DVH which approximately fulfills the DVH limits. Obviously, the quadratic overdose penalty is not

capable of modelling the volume effect of the rectum since the DVH corresponds to a much higher tolerance towards high doses.

The example was chosen because the volume effect of the rectum is frequently exploited in dose escalated prostate treatment to increase the dose to a small target volume. Most treatment protocols permit some higher dose to some smaller volume than would be tolerated for the whole rectum. These considerations become a part of the automated treatment planning process represented by any IMRT algorithm. The concept of evidence based biological optimization allows to express clinical experience in the volume effect parameter and the iso-effect prescription in a reproducible manner.

As IMRT allows higher target doses and prescriptions are changed towards dose escalated treatments, the issue of dose fractionation becomes important. If the course of treatment is prolonged to accomodate a greater number of standard dose fractions, repopulation of the tumour with clonogenes can offset the effect of a higher total dose. At the same time, as a consequence of better sparing, the dose per fraction to the normal tissues might even decrease. With dose escalation, there is a need to reconsider fractionation schemes. For this reason, fractionation effects need to be taken into account by the optimization algorithm. These effects are more important for normal tissues since a wide range of dose per fraction sizes is covered. Figure 7.2 shows a comparison of two fractionation schemes for an extensive lung tumour with an escalated dose of 70 Gy to a boost volume. The normal fractionation scheme was 2 Gy per fraction, the other an 'accelerated hyperfractionated scheme' of 70 fractions of 1 Gy administered twice a day. It can be seen that due to the great difference in $(\beta/\alpha)_{\text{tumour}} = 10$ Gy to $(\beta/\alpha)_{\text{lung}} = 2.5$ Gy, the hyperfractionated scheme maintains a more homogeneous boost dose - the better tolerance of lung of this scheme is exploited to redistribute the dose to the target volume more homogeneously. As a consequence, fractionation effects have an impact on the optimum physical dose which goes beyond a mere rescaling.

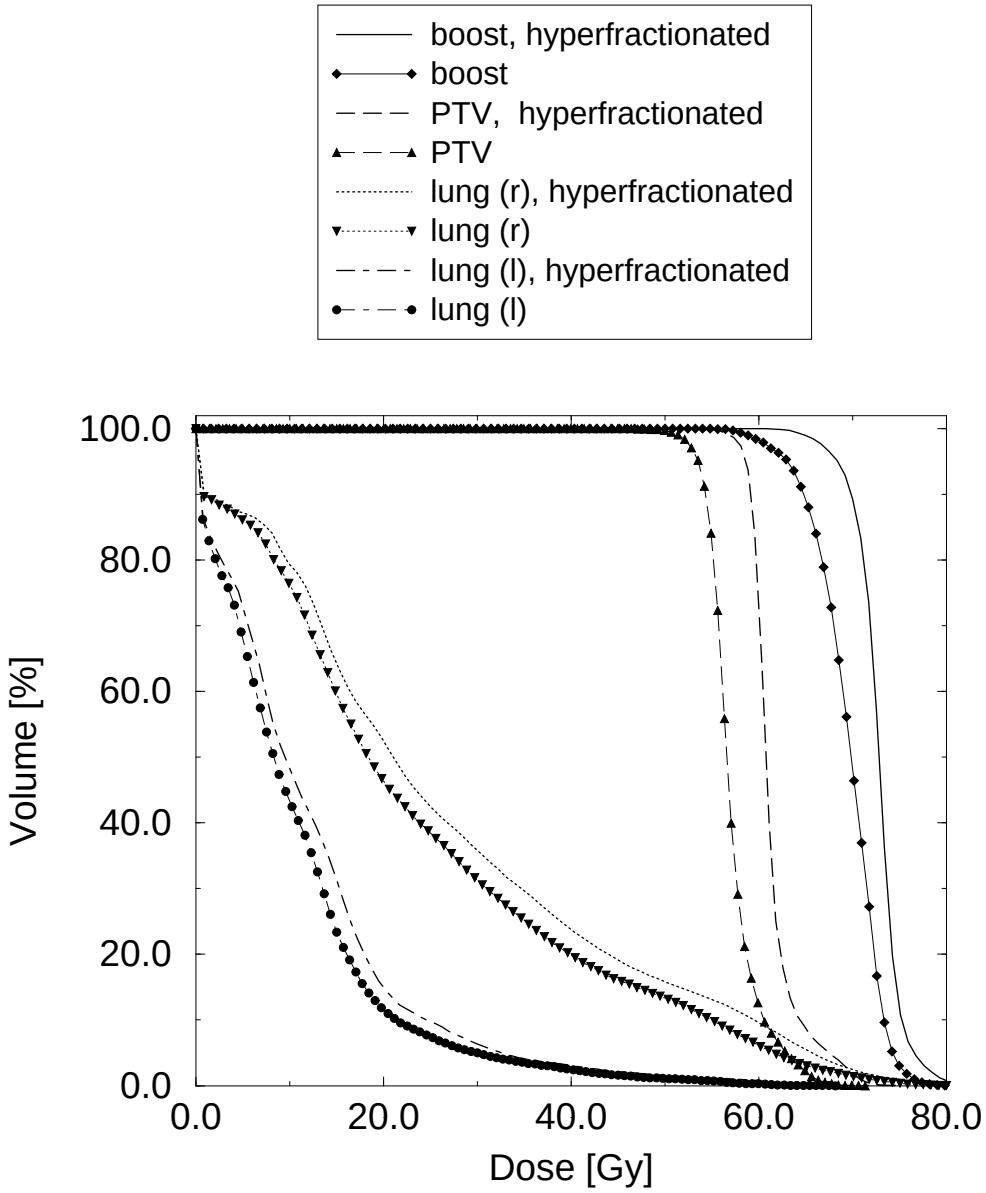


Figure 7.2: *A lung case with a prescription of 70 Gy to the boost volume and 55 Gy to the PTV. The normal fractionation scheme was 35 fractions of 2 Gy, whereas the accelerated hyperfractionated scheme was 1 Gy twice daily. The shift in the curves which show physical dose is caused by the smaller sensitivity of tissues for smaller dose fractions. The biologically equivalent DVHs for the lungs would coincide. However, the dose distribution in the boost volume is more homogeneous for the hyperfractionated scheme which shows that the algorithm uses the greater tolerance of the lungs to re-distribute the dose. The biologically equivalent dose to the boost volume was equivalent in both cases.*

## 7.2 The Clinical Benefit of Monte Carlo Optimization

Arguably, Monte Carlo dose computation has the potential to be more accurate than any phenomenological model. It has been pointed out that Monte Carlo verification computations provide a clinical benefit[72, 73, 74]. At contrast with standard 3D conformal planning, IMRT can gain from the better accuracy of the Monte Carlo computed dose distribution by fine tuning the fluence distributions. A Monte Carlo verification computation provides an estimate of the adverse effects of scatter, yet optimization results on the basis of Monte Carlo possess a different quality: they demonstrate what can be done clinically to counteract the physical effects which would go unnoticed with phenomenological models.

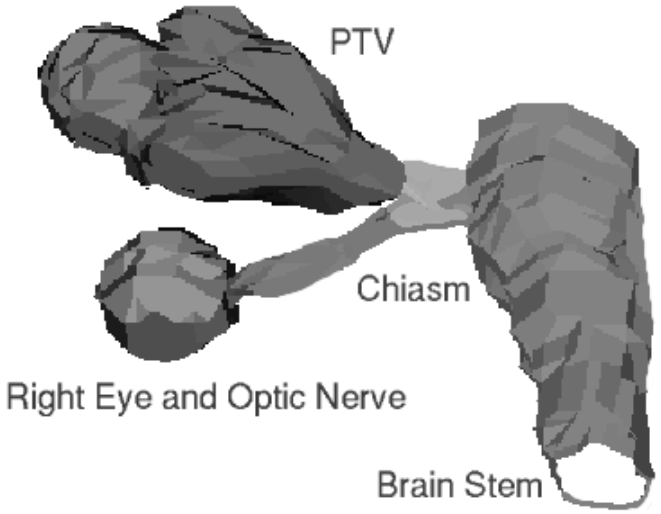


Figure 7.3: *A beam's eye view corresponding to the fluence distribution of figure 7.4. The field is tangential to the sphenoidal sinuses at the lower part of the PTV.*

The example is a schwannoma of the optical nerve (figure 7.3, 7.5) which stands for a class of paranasal target volumes with a genuine tumour-air interface. The complex geometry of the skull with bone/soft tissue/air interfaces leads to an overestimation of dose in the target, or underestimation of scatter in critical structures by phenomenological algorithms. The vicinity of organs with no significant volume effects prompts sharp dose gradients, and together with a clear tumour outline (in this case) and tight margins of the PTV, any reduction of the dose to the PTV is clinically significant.

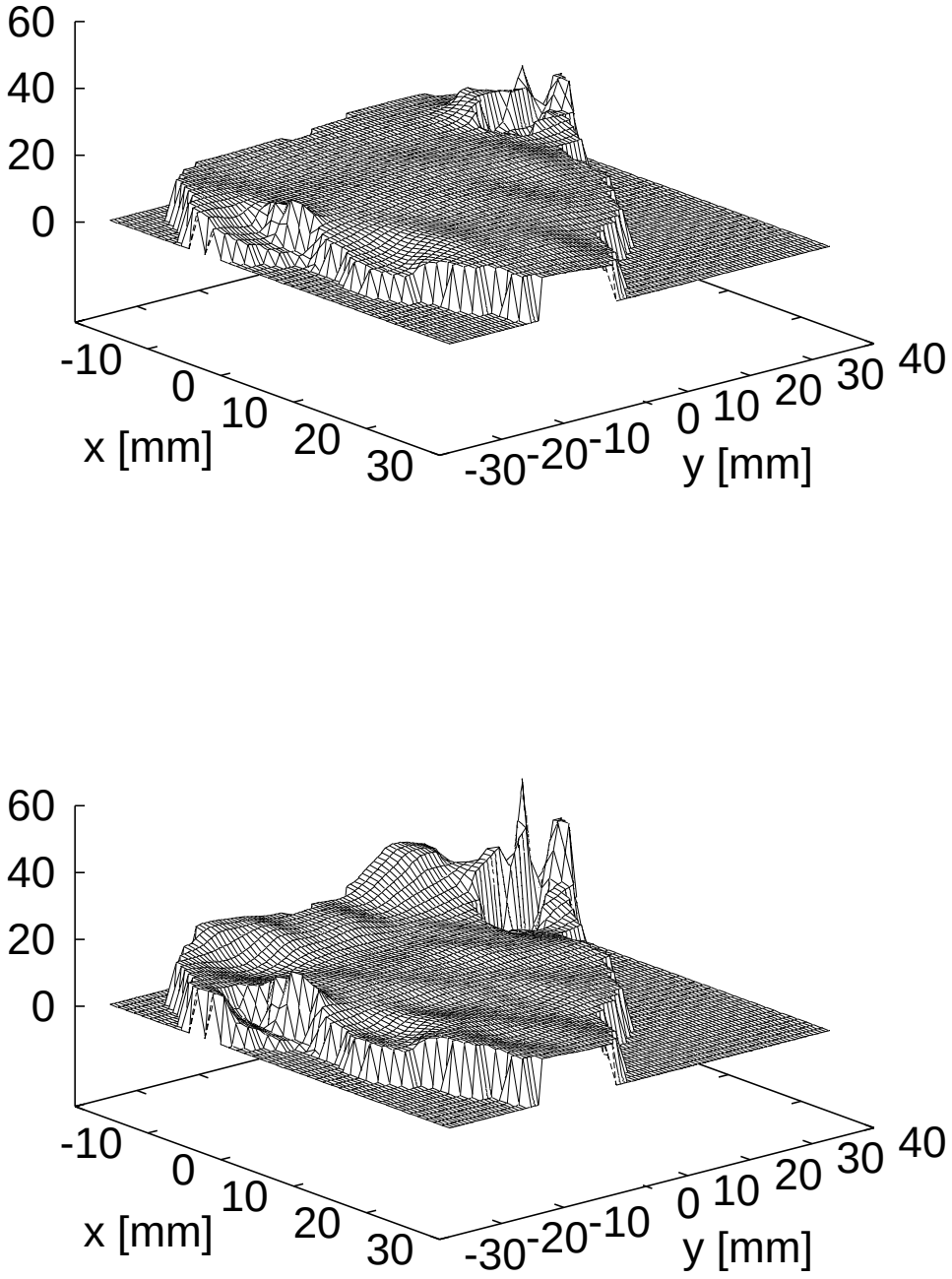


Figure 7.4: *The fluence distribution as obtained from the pencil-beam optimization (above) and the Monte Carlo optimization (below). The Monte Carlo profile shows compensation for the electron scatter in the third quadrant and the fourth quadrant of the field corresponding to the position of the sphenoidal and frontal sinuses, see figure 7.3.*

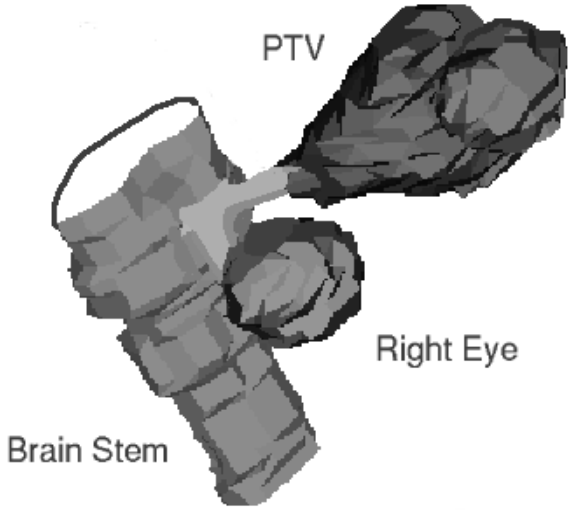


Figure 7.5: *The beam's eye view corresponding to the fluence distribution of figure 7.6. The beam passes through the nose which causes a dose re-buildup region in the PTV.*

The arrangement of four beams followed clinical practice $((-58°, -41°)$, $(90°, -57°)$, $(102°, 39°)$, $(-34°, 38°))$, the fluence matrices had a resolution of $1.25 \times 1.6\text{mm}^2$ which is feasible with micro-multileaf collimators. The photon energy was 6 MV, about 60 million histories were simulated. Compared to the pencil beam optimization results, the Monte Carlo profiles show clear evidence of the compensation of lateral electron scatter by increasing the primary fluence along tangents to the low density surfaces (see figure 7.4, 7.6).

| VOI | $C_{\text{iso}}$ [Gy] | k | PB [Gy] | VERI [Gy] | IM/MC [Gy] |
|---|---|---|---|---|---|
| PTV | 70 | | 72.7 | 62.8 | 70.9 |
| Chiasm | 35 | 10 | 34.9 | 34.1 | 35.3 |
| Nerve (r) | 35 | 10 | 2.7 | 5.8 | 7.4 |
| Eye (r) | 15 | 6 | 4.1 | 5.4 | 8.4 |
| Eye (l) | 40 | 6 | 39.8 | 38.8 | 40.8 |
| Brainstem | 30 | 8 | 3.7 | 2.6 | 4.1 |
| Brain | 36 | 8 | 36.4 | 35.3 | 36.4 |

Table 7.1: *The prescribed iso-effective doses $C_{\text{iso}}$ and the volume effect parameters k for the volumes of interest (VOIs) involved in the planning study. The resulting iso-effective doses for the pencil beam optimization (PB), the Monte Carlo verification (VERI) and the Monte Carlo optimization (IM/MC) show that essentially the dose in the PTV was overestimated. A large dose reduction in the PTV even in small volumes leads to a considerable reduction of iso-effective tumour dose.*

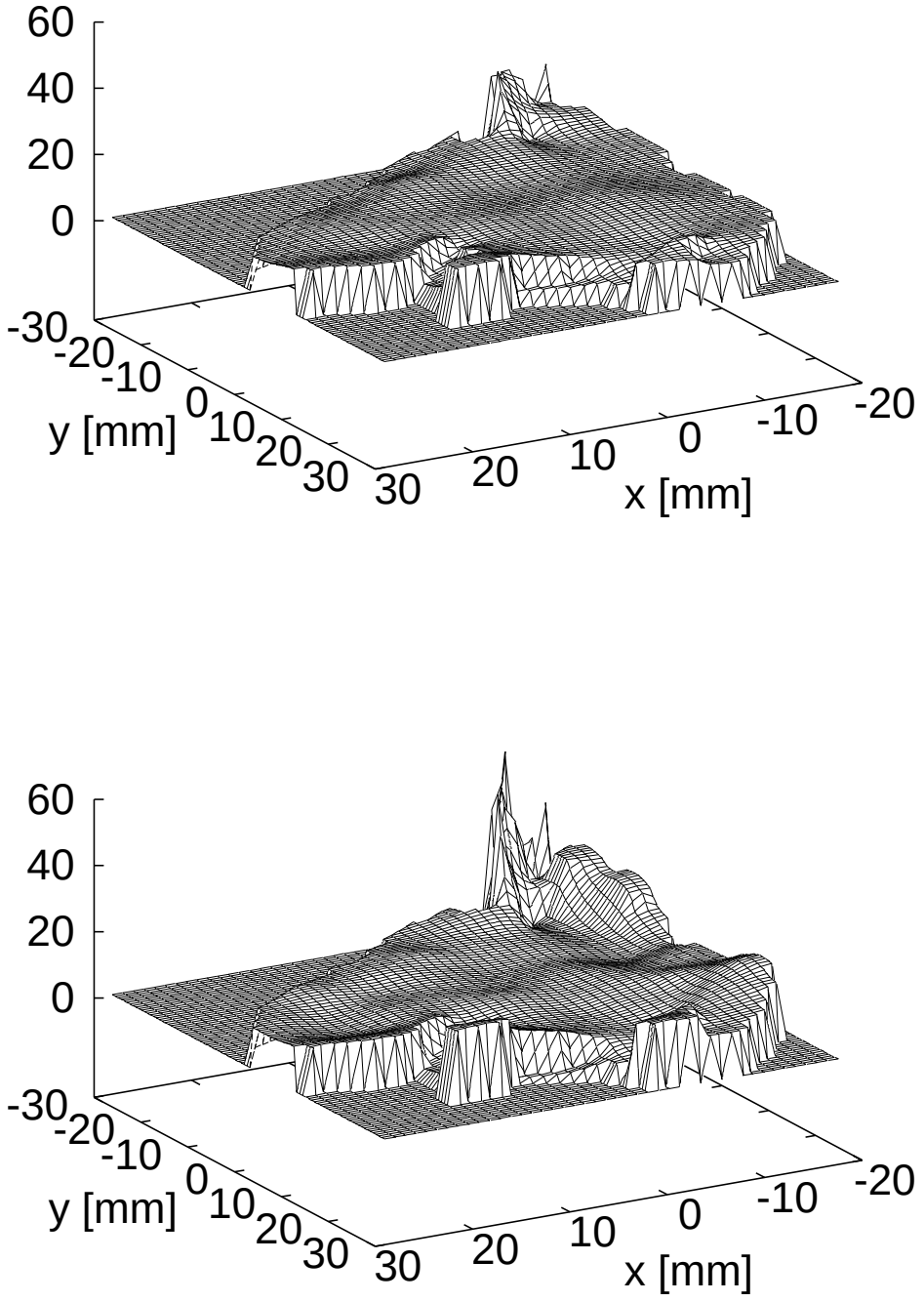


Figure 7.6: *The fluence distribution as obtained from the pencil-beam optimization (above) and the Monte Carlo optimization (below). The Monte Carlo profile shows compensation for the electron scatter in the fourth quadrant of the field corresponding to the position of the sphenoidal and frontal sinuses, see figure 7.4.*

Table 7.1 gives the optimization results for the pencil beam optimization, the Monte Carlo verification of the pencil beam dose, and the Monte Carlo optimization. The small field sizes and low energy of 6 MV produce dose re-buildup and lateral scatter regions with extensions of about 1 cm. Since the diameter of the PTV is in the range of 4 cm, a significant volume is affected by these effects. Consequently, the reduction in iso-effective dose to the PTV is in the range of 10 Gy. This dose cannot be fully restored by the Monte Carlo optimization since in the presence of active, dose limiting constraints the dose can only be redistributed, yet usually not increased. However, the benefit is significant and restores the tumour dose to 71 Gy. Since the beam arrangement largely avoided the organs at risk, only the chiasm, the affected eye and the brain were dose limiting together with a dose homogeneity constraint on the PTV.

The optimization did not take into account the influence of scatter from the leaves of the micro-MLC. It can be expected that this will yield a further advantage for the Monte Carlo optimization.

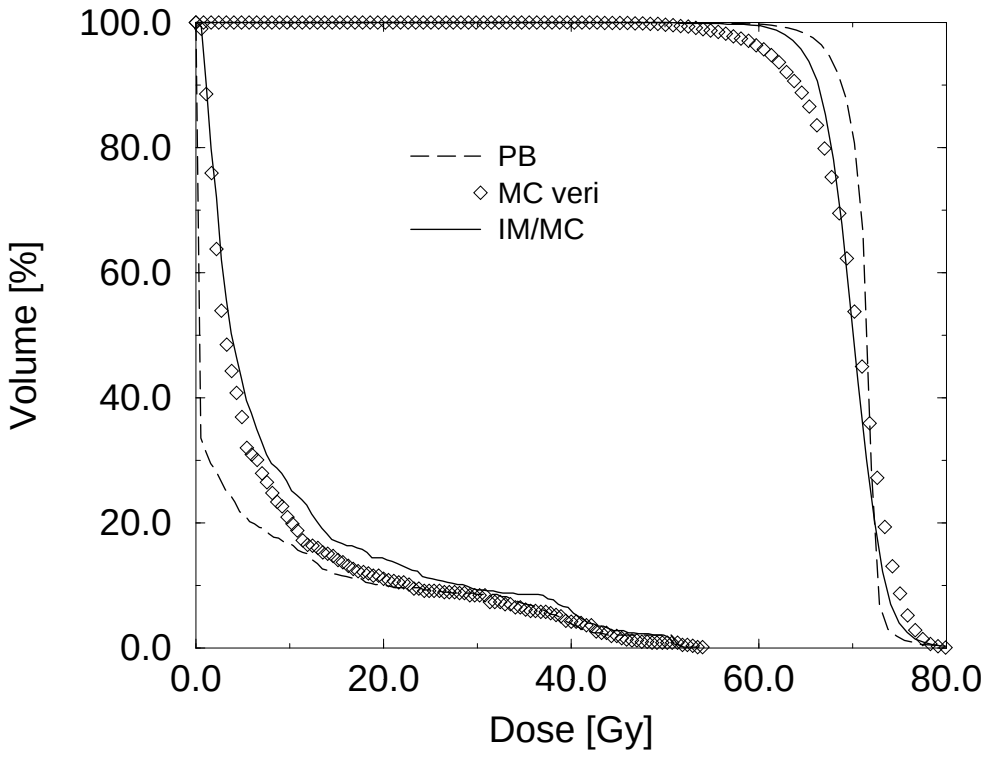


Figure 7.7: *A comprison of the dose distributions of PTV and chiasm for the pencil beam optimization (PB), the Monte Carlo verification (MC veri) and the Monte Carlo optimization (IM/MC). The PTV verification shows a pronounced underdosage to 10% of the volume, and some underdosage to about 50% of the volume. The IM/MC optimization restores the dose to the most exposed volumes, however, since the fluence can only be redistributed in the presence of active constraints, it cannot fully restore the pencil beam dose. The dose to the chiasm was unchanged safe the underestimation of scatter by the pencil beam algorithm.*

## 7.3 Enhanced Clinical Utility of Fluence Profiles

The clinical application of IMRT has to meet high quality standards. Regardless of the details of the implementation of intensity modulation at the treatment machines, smooth fluence profiles offer significant clinical advantages for an error tolerant application of radiation. For the static techniques, where the leaves of the MLC move to the configuration of the next field segment while the radiation is interrupted, smooth profiles lead to fewer field segments, a better efficiency and shorter treatment times.

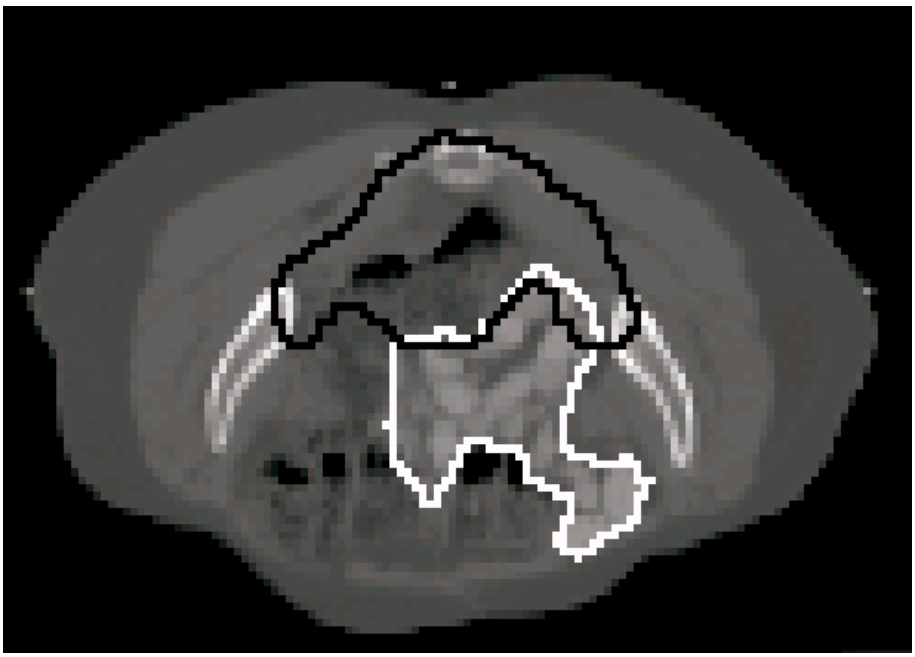

Figure 7.8: *A CT slice image of the colon-rectum example case. The PTV (black) encompasses the local lymph nodes and overlaps with the small bowel (white). The multiply concave shape of the PTV and the close proximity to the organ at risk necessitate heavily modulated fields.*

The static MLC technique is implemented at the William Beaumont Hospital, Detroit, with the help of a software module which translates the output of a radiotherapy optimization algorithm into a piecewise constant fluence profile suitable for static MLC application. In March 2000 the WBH embarked on IMRT of colon-rectal cancer which generally necessitates rather complex fluence distributions. The smoothness of the fluence profiles as produced by the algorithm described here facilitated a treatment within clinically acceptable time [75]. This patient was probably also the first treated with a biologically optimized plan.

Figure 7.8 shows a slice image of a target volume of a similar case where the prime organ at risk (small bowel) and the PTV are closely entwined. This case was part of a study which preceded the first treatment. The goal of the optimized treatment was both to reduce the volume of the small bowel receiving a high dose and intermediate dose in order

to avoid the complication of acute diarrhea. The PTV consists of the tumour bed and the locally involved pelvic lymph nodes. The standard treatment involves five coplanar fields with 72 degrees spacing.

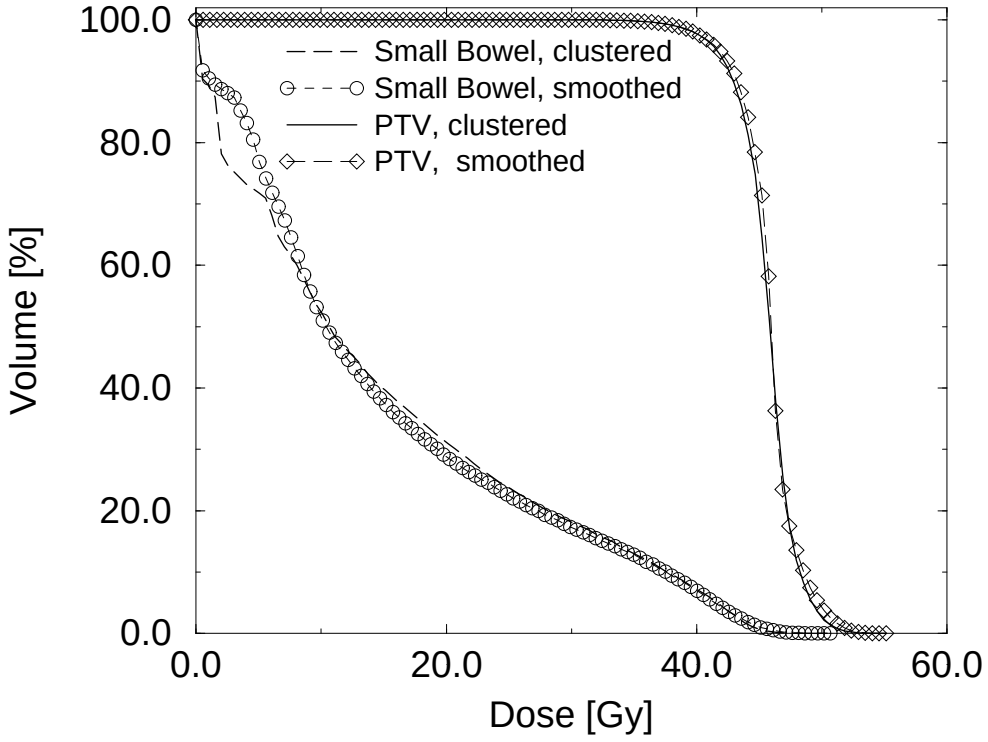


Figure 7.9: *A comparison of DVHs for the minimum surface smoothing and the clustering of rays into static MLC field segments. While the PTV receives an equivalent dose, the DVH of small bowel exhibits the stair stepped appearance which is typical for discrete homogeneous fields at low fluences. A translation of the smoothed fluence profiles to static MLC field segments would not have preserved the high level of dose homogeneity in the PTV.*

In figure 7.10 and 7.11 the fluence distributions of two fields are shown. The resolution of the fluence matrices corresponds to a clinical MLC of 1 cm leaf width in the isocentre and a discretisation of 2.1 mm in leaf direction. While the output of the smooth fluence profiles greatly diminishes the deterioration of plan quality due to the translation of the fluence profile into static MLC field segments, this step is still problematic. A full solution of the MLC problem can only be found if the creation of MLC field segments is fully included into the optimization. The results compare favourably with the translator output: the reduction in tumour dose is barely noticable, the number of field segments is about 20% lower. The method is the subject of future work [76].

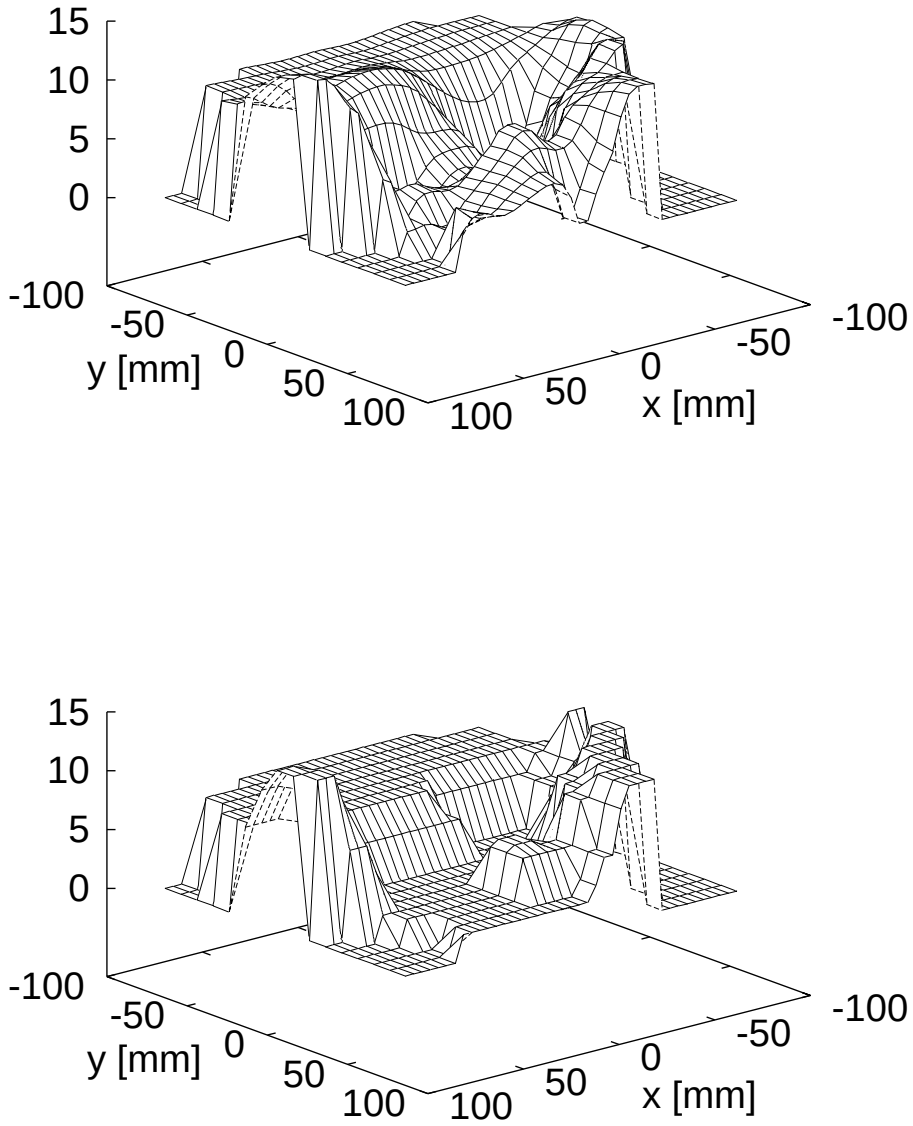


Figure 7.10: *The fluence profile of the posterior-anterior field with the minimal surface smoothing constraint (above) and with direct optimization of a piecewise constant fluence suitable for static MLC application (below). The difference from a simple translation into a piecewise constant fluence can be seen in the third quadrant of the field, where a small field segment was created in the optimization which is not present in the upper profile.*

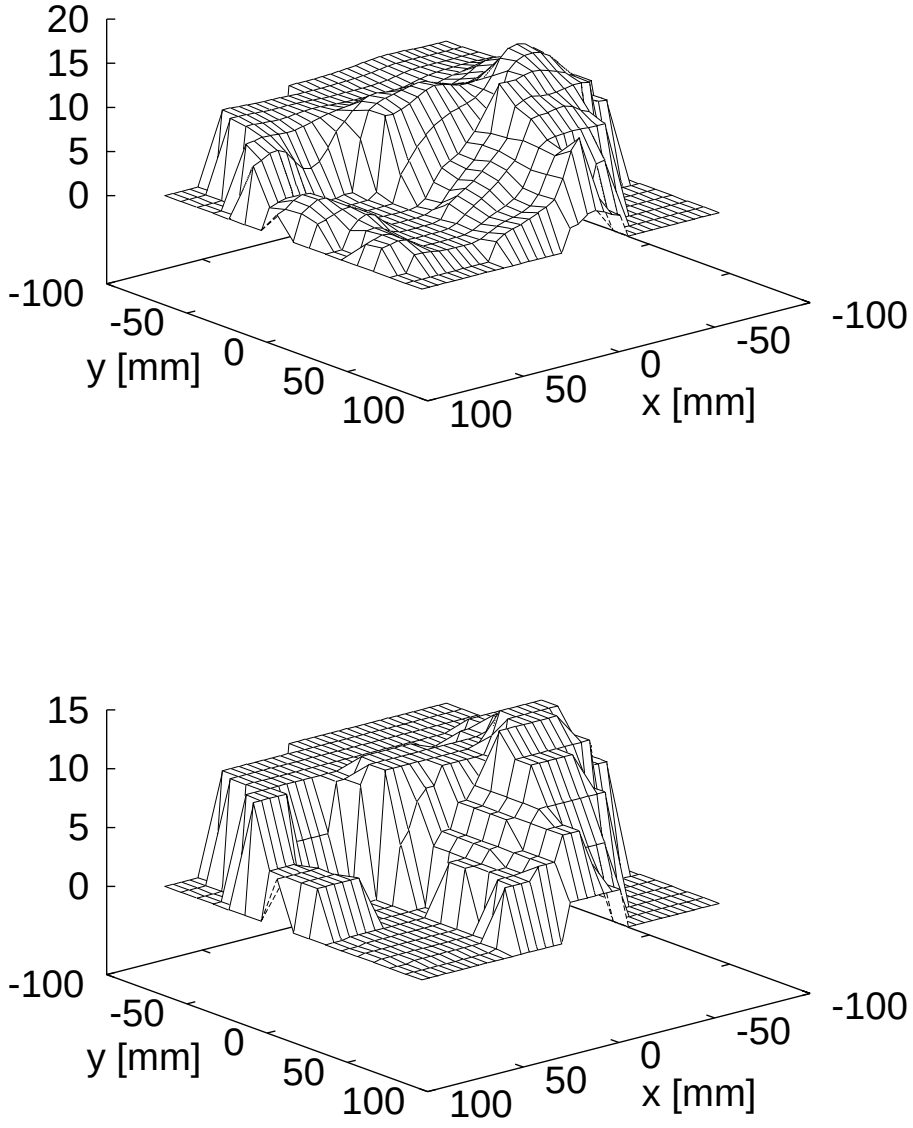


Figure 7.11: *The fluence profile of the anterior oblique field, again with minimal surface smoothing (above) and direct optimization of the static MLC field segments (below). In this instance, the lower field is bimodal, which ensures a rapid application.*

# Chapter 8

# Conclusion

The development set out here formulates the task of planning intensity modulated radiotherapy as a comprehensive optimization problem, including biological treatment objectives, measures for enhanced clinical utility and Monte Carlo radiation transport codes.

In its most general form, radiotherapy optimization is a variation problem. The treatment outcome as a functional of the fluence distribution is subject to a number of restrictions. To make the problem numerically tractable, a ray formalism is introduced which is a generalization of the method of Green's functions. A ray is the constituent entity of a practically feasible fluence distribution. It is shown that the functionals of dose which correspond to biological objectives can be expressed in the form of a 'radiation effect density' by virtue of a mean field approximation. This approximation can be motivated by a separation of biological interactions according to length scales; microscopic and macroscopic interactions can be treated explicitely whereas mesoscopic interactions can be taken into account by the mean-field approximation for photon therapy. A classification of normal tissue dose response mechanisms along these lines is given and specific effect functions are devised.

A dose computation algorithm is described which was specifically designed for the use in an optimization algorithm. The dose distribution of some $10^3$ to $10^4$ rays is precomputed and stored. The algorithm can be combined with Monte Carlo dose computation to form a very time efficient hybrid technique. A proof of convergence of the method is given. The clinical benefit of Monte Carlo dose computation is significant for tumours located close to low density interfaces. The modelling of the fluence distributions also incorporates a method which generates smooth fluence profiles. In an additional step, the complicated restrictions of multi-leaf collimators are included into the optimization. The full realization of these constraints as well as the treatment of the beam shaping elements with Monte Carlo methods during optimization are the subject of future work.

The treatment of physical and technical aspects of IMRT in the present optimization algorithm has the potential to establish IMRT as an alternative to conventional therapy for a large group of patients. The clinical introduction of the technique has been protracted by dosimetry problems and treatment time limitations which were entirely due to the neclect of

application constraints in early implementations of IMRT planning algorithms. The careful design of the dose computation algorithm and the introduction of a smoothing operation reduced the treatment complexity significantly and led to a considerable improvement of treatment quality. The full handling of application constraints will eventually allow to reduce treatment complexity even further and could help to establish 'IMRT for everybody'.

The concept of evidence based biological optimization constitutes a novel approach to treatment planning. It was conceived to deal with the hugely increased freedom to shape the dose distribution afforded by IMRT. The incentive of dose escalation to the tumour is to increase chances for cure by making use of this freedom. However, this can only be done if normal tissue effects are kept at the levels of conventional radiotherapy. Thus, the success of IMRT depends crucially on the capability of planning to express normal tissue reactions in a way which makes the advanced treatment comparable to the established. Evidence based biological optimization provides the means to achieve this. The complicated interplay of time/fractionation and volume effects of normal tissues makes dose based treatment planning intractable for radically new treatment concepts. In the quest for providing the best treatment, the next logical step is the individualized description of target volumes and prescription of target doses and dose fractions [77, 78, 79]. IMRT and evidence based biological optimization are no more, yet no less than the prerequisites to successfully meet the challenge of creating image guided, adaptive radiotherapy.

# Bibliography


[1] A. Brahme, J. E. Roos, and I. Lax. Solution of an integral equation encountered in rotation therapy. *Phys. Med. Biol.*, 27:1221–1229, 1982.

[2] A. Brahme. Optimisation of stationary and moving beam radiation therapy techniques. *Radiother. Oncol.*, 12:129–140, 1988.

[3] S. Webb. *The Physics of Three-dimensional Radiation Therapy: Conformal Radiotherapy, Radiosurgery and Treatment Planning.* Medical Science Series. IOP Publishing, Bristol, 1993.

[4] A. Brahme. Treatment optimization using physical and radiobiological objective functions. In A. R. Smith, editor, *Radiation Therapy Physics*, pages 209–246. Springer, Berlin, 1995.

[5] S .Webb. *The Physics of Conformal Radiotherapy: Advances in Technology.* Medical Science Series. IOP Publishing, Bristol, 1997.

[6] C. Raphael. Mathematical modelling of objectives in radiation therapy treatment planning. *Phys. Med. Biol.*, 37:1297–1311, 1992.

[7] R. Mohan, X. Wang, A. Jackson, T. Bortfeld, A. L. Boyer, G .J. Kutcher, S. A. Leibel, Z. Fuks, and C. C. Ling. The potential and limitations of the inverse radiotherapy technique. *Radiother. Oncol.*, 32:232–248, 1994.

[8] M. Goitein and A. Niemierko. Intensity modulated therapy and inhomogeneous dose to the tumor: A note of caution. *Int. J. Rad. Oncol. Biol. Phys.*, 36:519–522, 1996.

[9] T. Bortfeld, W. Schlegel, C. Dykstra, S. Levegrün, and K. Preiser. Physical vs. biological objectives for treatment plan optimization. *Radiother. Oncol.*, 40(2):185, 1996. letter, comment.

[10] R. Mohan and X.-H. Wang. Response to Bortfeld et al. re Physical vs biological objectives for treatment plan optimization. *Radiother. Oncol.*, 40(2):186–187, 1996. letter, comment.

[11] A. Brahme and B. K. Lind. The importance of biological modeling in intensity modulated radiotherapy optimization. In *Proceedings of the XII. International Conference on the use of Computers in Radiotherapy*, pages 5–8, 1997.

[12] M. Langer and J. Leong. Optimization of beam weights under dose-volume restrictions. *Int. J. Rad. Oncol. Biol. Phys.*, 13:1255–1260, 1987.

[13] X. Wang, S. Spirou, T. LoSasso, C. S. Chui, and R. Mohan. Dosimetric verification of an intensity modulated treatment. *Med. Phys.*, 23:317–327, 1996.

[14] T. Bortfeld, J. Stein, and K. Preiser. Clinically relevant intensity modulation optimization using physical criteria. In *Proceedings of the XII. International Conference on the use of Computers in Radiotherapy*, pages 1–4, 1997.

[15] S. V. Spirou and C.-S. Chui. A gradient inverse planning algorithm with dose-volume constraints. *Med. Phys.*, 25:321–333, 1998.

[16] P. S. Cho, S. Lee, R. J. Marks II, S. Oh, S. G. Sutlief, and M .H. Phillips. Optimization of intensity modulated beams with volume constraints using two methods: Cost function minimization and projections onto convex sets. *Med. Phys.*, 25:435–443, 1998.

[17] P. Källman, B. K. Lind, and A. Brahme. An algorithm for maximizing the probability of complication free tumor control in radiation therapy. *Phys. Med. Biol.*, 37:871–890, 1992.

[18] A. Gustafsson, B. K. Lind, and A. Brahme. A generalized pencil beam algorithm for optimization of radiation therapy. *Med. Phys.*, 21:343–356, 1994.

[19] Anders Gustafsson. *Development of a versatile algorithm for optimization of radiation therapy.* PhD thesis, University of Stockholm, 1996.

[20] S. Söderström and A. Brahme. Which is the most suitable number of photon beam portals in coplanar radiation therapy. *Int. J. Rad. Oncol. Biol. Phys.*, 33:1701–1709, 1995.

[21] D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, MA, 1995.

[22] P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization.* Academic Press, London, 1981.

[23] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty. *Nonlinear Programming.* Wiley, New York, 1983.

[24] C. G. Rowbottom, M. Oldham, and S. Webb. Constrained customization of non-coplanar beam orientations in radiotherapy of brain tumours. *Phys. Med. Biol.*, 44:383–399, 1999.

[25] J .T. Lyman. Complication probabilities as assessed from dose-volume histograms. *Radiat. Res.*, 104:S13–S19, 1985.

[26] J .T. Lyman and A. B. Wolbarst. Optimization of radiation therapy III: a method of assessing complication probabilities from dose-volume histograms. *Int. J. Rad. Oncol. Biol. Phys.*, 13:103–109, 1987.

[27] J .T. Lyman and A. B. Wolbarst. Optimization of radiation therapy IV: A dose-volume histogram reduction algorithm. *Int. J. Rad. Oncol. Biol. Phys.*, 17:433–436, 1989.

[28] G. J. Kutcher and C. Burman. Calculation of complication probability factors for non-uniform normal tissue irradiation: the effective volume method. *Int. J. Rad. Oncol. Biol. Phys.*, 16:1623–1630, 1989.

[29] G. J. Kutcher, C. Burman, L. Brewster, M. Goitein, and R. Mohan. Histogram reduction method for calculating complications probabilities for three-dimensional treatment planning evaluations. *Int. J. Rad. Oncol. Biol. Phys.*, 21:137–146, 1991.

[30] C. Burman, G. J. Kutcher, B. Emami, and M. Goitein. Fitting of normal tissue tolerance data to an analytic function. *Int. J. Rad. Oncol. Biol. Phys.*, 21:123–135, 1991.

[31] B. Emami, J. Lyman, A. Brown, L. Coia, M. Goitein, J. E. Munzenrieder, B.Shank, L. J. Solin, and M. Wesson. Tolerance of normal tissue to therapeutic irradiation. *Int. J. Rad. Oncol. Biol. Phys.*, 21:109–122, 1991.

[32] L. Cohen. The statistical prognosis in radiotherapy. *Am. J. Roentgenol.*, 84:741–753, 1960.

[33] D. H. Moore and M. L. Mendelsohn. Optimal treatment levels in cancer therapy. *Cancer*, 30:95–106, 1972.

[34] T. E. Schultheiss, C. G. Orton, and R. A. Peck. Models in radiotherapy: volume effects. *Med. Phys.*, 10:410–415, 1983.

[35] A. B. Wolbarst. Optimization of radiation therapy. II. the critical-voxel model. *Int. J. Rad. Oncol. Biol. Phys.*, 10:741–745, 1984.

[36] J. R. Andrew. Benefit, risk and optimization by ROC analysis in cancer radiotherapy. *Int. J. Radiat. Oncol. Biol. Phys.*, 11:1557–1562, 1985.

[37] A. M. Kellerer and H. H. Rossi. RBE and the primary mechanism of radiation action. *Radiat. Res.*, 47:15–34, 1971.

[38] K. H. Chadwick and H. P. Leenhouts. A molecular theory of cell survival. *Phys. Med. Biol.*, 18:78–87, 1973.

[39] B. G. Douglas and J. F. Fowler. Fractionation schedules and a quadratic dose-effect relationship. *Br. J. Radiol.*, 48:502–504, 1975.

[40] M. C. Joiner and H. Johns. Renal damage in the mouse: The response to very small doses per fraction. *Radiat. Res.*, 114:385–398, 1988.

[41] A. Dasu and J. Denekamp. Superfractionation as a potential hypoxic cell radiosensitizer: prediction of an optimum dose per fraction. *Int. J. Rad. Oncol. Biol. Phys.*, 43:1083–1094, 1999.

[42] E. P. Malaise, P. Lambin, and M. C. Joiner. Radiosensitivity of human cell lines to small doses. Are there some clinical implications? *Radiat. Res.*, 138:S25–27, 1994.

[43] M. C. Joiner, Lambin P., and E. P. Malaise. Hypersensitivity to very-low single radiation doses: Its relationship to the adaptive response and induced radioresistance. *Mutat. Res.*, 358:171–183, 1996.

[44] H. R. Withers, J. M. G. Taylor, and B. Maciejewski. The hazard of accelerated tumor clonogen repopulation during radiotherapy. *Acta Oncolog*, 27:131–146, 1988.

[45] R. Mohan, Q. Wu, M. Manning, and R. Schmidt-Ullrich. Radiobiological considerations in the design of fractionation strategies for intensity-modulated radiation therapy of head and neck cancers. *Int. J. Radiat. Oncol. Biol. Phys.*, 46:619–630, 2000.

[46] A. E. Nahum and D. M. Tait. Maximising tumour control by customized dose prescription for pelvic tumours. In A. Breit, editor, *Advanced Radiation Therapy: Tumour Response Monitoring and Treatment Planning*, pages 425–431. Springer, Heidelberg, 1992.

[47] S. Webb and A. E. Nahum. A model for calculating tumour control probability in radiotherapy including the effects of inhomogenoeus distributions of dose and clonogenic cell density. *Phys. Med. Biol.*, 38:653–666, 1993.

[48] L. J. Boersma, E. M. F. Damen, R .W. de Boer, S. H. Muller, C. M. Roos, R. A. Valdes Olmos, N. van Zandwijk, and J. V. Lebesque. Dose-effect relations for local functional and structural changes of the lung after irradiation for malignant lymphoma. *Radiother. Oncol.*, 32:201–209, 1994.

[49] L. J. Boersma, E. M. F. Damen, R .W. de Boer, S. H. Muller, C. M. Roos, R. A. Valdes Olmos, N. van Zandwijk, and J. V. Lebesque. Estimation of overall pulmonary function after irradiation using dose-effect relations for local functional injury. *Radiother. Oncol.*, 36:15–23, 1995.

[50] D. J. Convery and M. E. Rosenbloom. The generation of intensity-modulated fields for conformal radiotherapy by dynamic collimation. *Phys. Med. Biol.*, 37:1359–1374, 1992.

[51] C. S. Chui, T. LoSasso, and S. Spirou. Dose calculation for photon beams with intensity modulation generated by dynamic jaw or multileaf collimations. *Med. Phys.*, 21:1237–1244, 1994.

[52] T. R. Bortfeld, D. L. Kahler, T. J. Waldron, and A. L. Boyer. X-ray field compensation with multileaf collimators. *Int. J. Radiat. Oncol. Biol. Phys.*, 28:723–730, 1994.

[53] A. L. Boyer. Use of MLC for intensity modulation. *Med. Phys.*, 21:1007, 1994.

[54] T. R. Mackie, J. W. Scrimger, and J. J. Battista. A convolution method for calculating dose for 15-MV X-rays. *Med. Phys.*, 12:188–196, 1985.

[55] R. Mohan, C. Chui, and L. Lidofsky. Differential pencil beam dose computation model for photons. *Med. Phys*, 13:64–73, 1986.

[56] R. Mohan and C. Chui. Use of Fast Fourier transforms in calculating dose distributions for irregularly shaped fields for three-dimensional treatment planning. *Med. Phys.*, 14:70–77, 1987.

[57] A. Ahnesjö. Collapsed cone convolution of radiant energy for photon dose calculation in heterogenenous media. *Med. Phys.*, 16:577–592, 1989.

[58] A. L. Boyer, Y. Zhu, L. Wang, and P. Francois. Fast Fourier transform convolutions of x-ray isodose distributions in homogenous media. *Med. Phys.*, 16:248–253, 1989.

[59] A. Ahnesjö, M. Saxner, and A. Trepp. A pencil beam model for photon dose calculation. *Med. Phys.*, 19:263–273, 1992.

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

[61] P. Storchi and E. Woudstra. Calculation of the absorbed dose distribution due to irregularly shaped photon beams using pencil beam kernels derived from basic beam data. *Phys. Med. Biol.*, 41:637–656, 1996.

[62] A. Ahnesjö and M. M. Aspradakis. Dose calculations for external photon beams in radiotherapy. *Phys. Med. Biol.*, 44:R99–R156, 1999.

[63] W. U. Laub. *Monte-Carlo Simulationen zur Validierung von Dosisberechnungsalgorithmen für Photonenstrahlung in der 3D-Bestrahlungsplanung.* PhD thesis, Eberhard-Karls-Universität Tübingen, Tübingen, 1998.

[64] D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods.* Academic Pres, New York, 1982.

[65] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C.* Cambridge University Press, 1994.

[66] E. Polak. *Computational Methods in Optimization.* Academic Press, New York, 1971.

[67] D. Touati-Ahmed and C. Storey. Efficient hybrid conjugate gradient techniques. *J. Opt. Th. Appl.*, 64:379–397, 1990.

[68] R. P. Brent. *Algorithms for minimization without derivatives.* Prentice-Hall, Englewood Cliffs, NJ, 1973.

[69] T. Bortfeld, J. Bürkelbach, R Boesecke, and W. Schlegel. Three-dimensional solution of the inverse problem in conformation radiotherapy. In Breit, editor, *Advanced Radiation Therapy: Tumor Response Monitoring and Treatment Planning*, pages 503–508. Springer, Berlin, 1992.

[70] J. Nocedal and S. J. Wright. *Numerical Optimization.* Springer, New York, 1999.

[71] A. R. Conn, N. I. M. Gould, and P. L. Toint. *LANCELOT A Fortran package for large-scale nonlinear optimization (Release A).* Springer, Berlin, 1992.

[72] A. E. Nahum. Conformal therapy needs Monte Carlo dose computation. *Proc. Challenges in Conformal Radiotherapy (ESTRO)*, pages 1–11, 1997.

[73] R. Mohan. Why Monte Carlo. In *Proc. Int. Conf. on the Use of Computers in Radiation Therapy*, pages 16–18, Madison, Wi, 1997. Medical Physics Publishing.

[74] P. Keall, J. Siebers, and R. Mohan. The impact of Monte Carlo dose calculations on treatment outcomes. In W. Schlegel and T. Bortfeld, editors, *Proceedings of the XIII International Conference on the Use of Computers in Radiation Therapy*, pages 425–427, 2000.

[75] W. U. Laub, D. Yan, M. Sharpe, J. Nuyttens, J. Robertson, and J. Wong. A comparison of IMRT planning systems in the treatment of colon-rectal cancer. In W. Schlegel and T. Bortfeld, editors, *Proceedings of the XIII International Conference on the Use of Computers in Radiation Therapy*, pages 529–531, Heidelberg, 2000. Springer.

[76] M. Alber and F. Nüsslin. IMRT optimization under constraints for static and dynamic MLC delivery. *submitted to Phys. Med. Biol.*, 2001.

[77] D. Yan, J. Wong, F. Vicini, J. Michalski, C. Pan, A. Frazier, E. Horwitz, and A. Martinez. Adaptive modification of treatment planning to minimize the deleterious effects of treatment setup errors. *Int. J. Radiat. Oncol. Biol. Phys.*, 38:197–206, 1997.

[78] D. Yan, D. A. Jaffray, and J. W. Wong. A model to accumulate the fractionated dose in a deforming organ. *Int. J. Radiat. Oncol. Biol. Phys.*, 44:665–675, 1999.

[79] D. Yan. Treatment strategies for daily image feedback adaptive radiotherapy. In W. Schlegel and T. Bortfeld, editors, *Proceedings of the XIII International Conference on the Use of Computers in Radiation Therapy*, pages 518–520, Heidelberg, 2000. Springer.

# Acknowledgments

This work thrived on friendship, idealism, trust and - scepticism. I received an abundance of each of them - with varying composition - from all the people to whom I wish to express my most sincere gratitude.

To my academic teacher Prof. Fridtjof Nüsslin who lent a sense of perspective to all my wanderings.
To my friend Wolfram Laub who so often was the completion - and so often the contradiction.
To Dr. Th. W. Kaulich for warmhearted advice and support.
To Mattias Birkner whose support I will never be able to reciprocate.
To Prof. Gunther Christ for listening and calm pragmatism.
To Dr. Mathias Fippel for his stoic qualities in the face of my more head-strong phases.
To Michael Reinert, Bernd Weigel, Andre Mondry and Markus Buchgeister who earned my greatest respect when I joined them in routine treatment planning. To Andre and Markus in particular for all the forbearance towards my computer-fuelled bad tempers. To Anne Bakai for her forbearance towards all my tempers. To Christoph Benk for unequally shared six-packs and equally shared sorrows. To all other colleagues at Tübingen for making those three years a good time.
To Prof. Rembert Reemtsen at the University of Cottbus for endless patience and tireless struggle for mathematic rigour and against methodologic handwaving.
To Di Yan, John Wong, Mike Sharpe, Dave Jaffray, Mark Oldham and John Robertson at William Beaumont Hospital, Detroit for a lession in New World idealism and for letting Hyperion shine.
To Werner De Gersem at the University of Gent for an overdose of enthusiasm.
To Philip Mayles, Robert Price, Geoff Lawrence, John Fenwick, Tanja Wolff and all the others at Clatterbridge Centre for Oncology, Liverpool who made my stay there such a gainful time.
To Dr. Thomas Bortfeld at the DKFZ and Dr. Jörg Stein and Dr. Carsten Schulze at MRC Systems for inspiring discussions and the invaluable cooperation on the basis of KonRad.
To the Deutsche Krebshilfe for providing the funding for a project which certainly began on a different track from where it ended.

# Appendix A

# An objective function for radiation treatment optimization based on local biological measure

*published in* Physics in Medicine and Biology **44** p.479-493 (1999).

# Appendix B

# A representation of a NTCP function for local complication mechanisms

*published in* Physics in Medicine and Biology **46** p.439-447 (2001).

# Appendix C

# Intensity modulated photon beams subject to a minimal surface smoothing constraint

*published in* Physics in Medicine and Biology **45** p.N49-N52 (2000).

# Appendix D

# Monte Carlo Dose calculation for IMRT optimization

*published in* Physics in Medicine and Biology **45** p.1741-1754 (2000).