Mathematical Optimization For The Inverse Problem Of Intensity-Modulated Radiation Therapy

Yair Censor, D.Sc.
Department of Mathematics, University of Haifa
Haifa, Israel

Introduction

We consider intensity-modulated radiation therapy (IMRT) where beams of penetrating radiation are directed at the lesion (tumor) from external sources. Based on understanding of the physics and biology of the situation, there are two principal aspects of radiation teletherapy that call for mathematical modeling. The first is the calculation of the *radiation dose,* which is a measure of the actual energy absorbed per unit mass everywhere in the irradiated tissue. In dose calculation the relevant physical, geometric and biological characteristics of the irradiated object and the relevant information about the radiation source (geometry, physical nature, intensity, etc.) serve as input data. The result of the calculation is a *dose function* (distribution) whose values are the dose absorbed as a function of location inside the irradiated body. This dose calculation is the *forward problem* of IMRT.

The second aspect is, mathematically speaking, the *inverse problem* of the first. In addition to the availability of the physical and biological parameters of the irradiated object we assume here that the relevant information about the capabilities and specifications of the available *treatment machine* (i.e., radiation source) is given. Based on medical diagnosis, knowledge, and experience, the physician prescribes a *desired* dose function to the case. The output of a solution method for the inverse problem should be a *radiation intensity function,* whose values are the radiation intensities at the sources as a function of source location, that would result in a dose function that is identical to the desired one. To be of practical value, this resulting radiation intensity function must be implementable, in a clinically acceptable form, on the available treatment machine.

Historically, working in two dimensions (2-D) where only a single plane through the center of the target is considered, the treatment planning was done (and is still frequently done) in a trial-and-error fashion. A machine setup that gives rise to a certain external radiation intensity field (function) is picked up and then, using a forward-problem-solver software package, the resulting dose function is determined. If the discrepancy between this dose function and the prescribed dose function is unacceptable, then some changes are made to the external radiation intensity field (i.e., the machine setup parameters) and the process is repeated until the physician and dosimetrist are satisfied with the resulting dose function. Only then is actual patient treatment performed.

Such 2D-RTTP (radiation therapy treatment planning) has achieved success due to accumulated experience and also because of the ever-increasing quality, sophistication, and speed of forward-problem-solvers. However, automated solution of the inverse problem of IMRT should be useful in handling difficult planning cases, particularly in three dimensions (3-D) (see figure 1). There, it would be much more difficult to reach an acceptable plan by trial-and-error because of the multitude of potential directions from which the 3-D object can be irradiated. Nonetheless, even a 2-D discussion, as given here, is enough to expose the nature of the dilemmas that we consider in the sequel.

In the next section we present the continuous forward and inverse problems and then we give their discretizations. The feasibility approach is formulated and optimization formulations are given in later sections. Finally, we very briefly discuss some of the methods and techniques that have been applied to the inverse problem of IMRT, namely, global optimization (including simulated annealing), multi-objective optimization, linear and mixed integer programming, and projection methods (including Cimmino’s algorithm). This paper is written as a tutorial and there is neither an intention to present a full survey of optimization methods in RTTP, nor an attempt to properly cover the literature. We also admit a slight bias in space allocation below towards projection methods, which are our own main field of research.

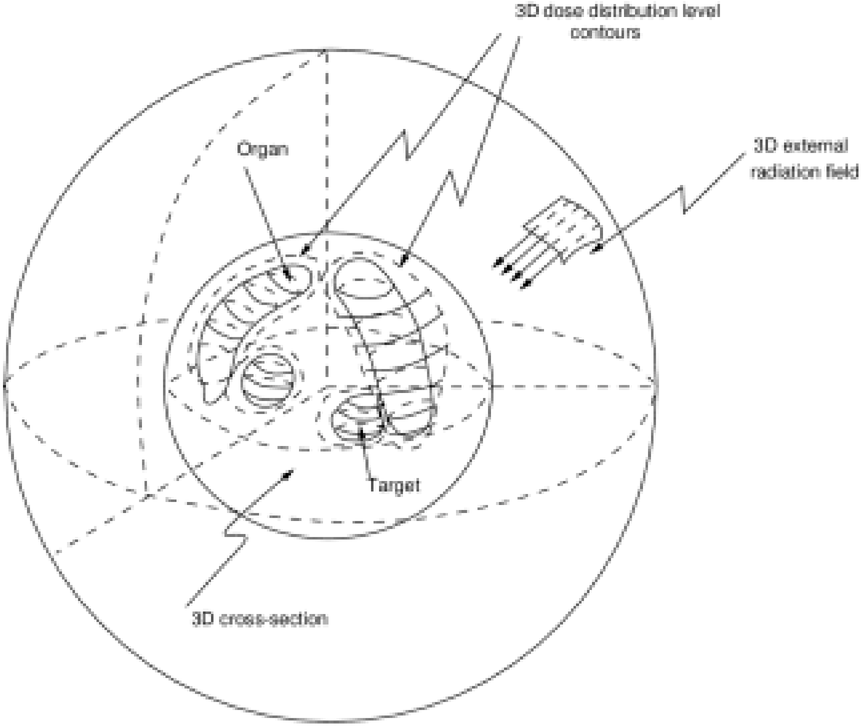

3D dose distribution level contours
3D external radiation field
Organ
Target
3D cross-section

Figure 1. A 3-D cross-section, external radiation field, and dose distribution for 3-D IMRT planning.

Problem Definition And The Continuous Model

Let $D(r, \theta)$ be a real-valued nonnegative function, of the polar coordinates r and θ, whose value is the dose absorbed at a point in the patient's planar cross-section Ω coincident with the plane of the machine's gantry motion. This is the *dose function*, or dose distribution. A *ray* is a directed line along which radiated energy travels away from the *source* (the *teletherapy source*). Rays are parametrized by variables u and w in some well-defined way and the real-valued nonnegative function $\rho(u, w)$ represents the *radiation intensity* along the ray (u, w) due to a point source on the gantry circle, located at (u, w).

Problem 1. ***The continuous forward problem of IMRT***. *Assume that the cross-section Ω of the patient and its radiation absorption characteristics are known. Given an external radiation intensity function $\rho(u, w)$, for $0 \leq u < 2\pi$ and $-W \leq w \leq W$, find the dose function $D(r, \theta)$, for all $(r, \theta) \in \Omega$, from the formula*

$$D(r,\theta) = \mathfrak{D}\left[\rho(u,w)\right](r,\theta) \tag{1}$$

where $\mathfrak{D}$ is the dose operator which relates the dose function to the radiation intensity function.

In other words, the forward problem amounts to the calculation of the total dose absorbed at each point of a patient's cross-section when all parameters of all radiation beams are specified and the description of the patient's cross-section is known. The difficulties associated with the forward problem stem from the fact that to this date there exists no closed-form analytic representation of the dose operator $\mathfrak{D}$ that will enable us to use equation (1) for the calculation of $D(r, \theta)$. Although the interaction between radiation and tissue is measured and understood at the atomic level, the situation is so complex that, to solve the forward problem in practice, a state-of-the-art computer program, which represents a *computational approximation* of the operator $\mathfrak{D}$ and which enables reasonably good dose calculations, must be used.

By stating that "there exists no closed-form analytic representation of the dose operator $\mathfrak{D}$" we mean that only if drastically simplifying assumptions are made about the physics of the model as well as of the particulars of the desired dose distribution, then it is sometimes possible to express the dose operator in a closed-form analytic formula. This has been done first by Brahme, Roos, and Lax (1982) and extended by Cormack and co-workers; consult the review paper of Cormack and Quinto (1990) for further references. See also Brahme's review (Brahme 1995) and Goitein's editorial (Goitein 1990). In current practice of IMRT, when dose calculations are performed to verify the dose that will result from a proposed treatment plan, the goal is to obtain results that are as accurate as possible. To achieve this, various empirical data, which are often condensed in look-up tables, are incorporated into the forward calculation. Thus, the true forward calculation, or true dose operator, is not represented by a closed-form analytic relation between the radiation intensity function $\rho(u, w)$ and the dose function $D(r, \theta)$, but by a software package that calculates $D(r, \theta)$ from $\rho(u, w)$. We choose to adhere to the software representation of $\mathfrak{D}$ rather than to compromise by allowing simplifying assumptions that might lead to a closed-form analytic mathematical formula at the expense of the physical and biological reality of the forward calculation.

Problem 2. ***The continuous inverse problem of IMRT***. *Assume that the cross-section Ω of the patient and its radiation absorption characteristics are known. Given a prescribed dose function $D(r, \theta)$, find a radiation intensity function $\rho(u, w)$ such that equation* (1) *holds, or, equivalently,*

$$\rho(u,w) = \mathfrak{D}^{-1}\left[D(r,\theta)\right], \tag{2}$$

where $\mathfrak{D}^{-1}$ is the ***inverse operator*** *of $\mathfrak{D}$.*

Solving problem 2 gives an external configuration and relative intensities of radiation sources (i.e., the radiation field) that will deliver the prescribed radiation dose distribution (or some acceptable approximation thereof). This inversion problem needs to be solved, in a computationally tractable way, although no closed-form analytic mathematical representation is available for the dose operator $\mathfrak{D}$. Without such a mathematical representation of $\mathfrak{D}$ it is impossible to employ mathematical methods for analytic inversion to find the inverse operator $\mathfrak{D}^{-1}$. This is why full discretization of the problem has to be adopted, as we did in Altschuler and Censor (1984) and Censor, Altschuler, and Powlis (1988a).

The dose at (r, θ) is the sum of the dose contributions from the sources at all the different gantry angles. Thus,

$$D(r,\theta) = \sum_{s=1}^{S} y_s D_s(r,\theta) \tag{3}$$

where, for each $s = 1, 2, \ldots, S$, the value $D_s(r, \theta)$ is the dose deposited at point (r, θ) by a beam of unit intensity from the sth source, and y_s is the time the sth beam is kept on. It will be assumed here that the dose $D_s(r, \theta)$ can be calculated accurately once the beam parameters and patient's cross-section information are specified. That is, we assume that we can solve the forward problem and calculate $D(r, \theta)$ accurately from (3). This assumption is confirmed by innumerable direct measurements in water and tissue-equivalent phantoms. Whereas a dose distribution that solves the forward problem is always obtained for a specified external radiation intensity field, the inverse problem may have no solution at all, since some prescribed dose functions may be unobtainable from any radiation field.

Discretization Of The Problem

In the approach presented here, we adhere to the computational approximation of the dose operator $\mathfrak{D}$. Full discretization of the problem at the outset is used to circumvent the difficulties associated with the analytic inversion of $\mathfrak{D}$. We also neglect in the present description the effects of scattered radiation. The patient's cross-section Ω is discretized into a grid of points represented by the pairs$\{(r_j, \theta_j) \mid j = 1, 2, \ldots, J\}$. Define $\mathfrak{D}_j[\rho]$ by

$$\mathfrak{D}_j[\rho] := [\mathfrak{D}\rho](r_j,\theta_j) \tag{4}$$

and call $\mathfrak{D}_j$ a *dose functional,* for every $j = 1, 2, \ldots, J$. Acting on a radiation intensity function $\rho(u, w)$, the functional $\mathfrak{D}_j$ provides $\mathfrak{D}_j[\rho]$, which is the dose absorbed at the jth grid point of the patient's cross-section Ω due to the radiation intensity field ρ. To continue the discretization process of the problem it is assumed that a set of *I basis* *radiation intensity fields* is fixed and that their nonnegative linear combinations can give adequate approximations to any radiation intensity field we wish to specify. This is done by discretizing the region $0 \leq u < 2\pi$, $-W \leq w \leq W$ in the (u, w)-plane into a grid of points given by $\{(u_i, w_i) \mid i = 1, 2, \ldots, I\}$. A radiation intensity function

$$\sigma_i(u,w) := \begin{cases} 1, & \text{if } (u,w) = (u_i, w_i) \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

is a *unit intensity ray* (or *beamlet*) and serves as a member of the set of basis radiation intensity fields for $i = 1, 2, \ldots, I$. In this fully discretized model, a desired radiation intensity function ρ that solves the inverse problem is always approximated by

$$\hat{\rho}(u,w) = \sum_{i=1}^{I} x_i \sigma_i(u,w) \tag{6}$$

where x_i is the intensity of the ith ray and it is required to be nonnegative, i.e., $x_i \geq 0$ for all $i = 1, 2, \ldots, I$. Once the grid points are fixed, any radiation intensity function $\hat{\rho}$, that can be represented as a nonnegative linear combination of the rays, is uniquely determined by the intensity coefficients x_i. The latter form the components of the vector $x = (x_i)_{i=1}^{I} \in R^I$, in the I-dimensional Euclidean space, referred to as the *radiation intensity vector*.

Further, assume that the dose functionals $\mathfrak{D}_j$ are linear and continuous. This assumption cannot be mathematically verified due to the absence of an analytic representation of either $\mathfrak{D}$ or $\mathfrak{D}_j$, but it is a reasonable assumption based on the empirical knowledge of $\mathfrak{D}_j$. Using linearity and continuity of all $\mathfrak{D}_j$'s, we can write

$$\mathfrak{D}_j[\rho] \cong \mathfrak{D}_j(\hat{\rho}) = \sum_{i=1}^{I} x_i \mathfrak{D}_j[\sigma_i] \tag{7}$$

For $j = 1, 2, \ldots, J$, and $i = 1, 2, \ldots, I$, denote by

$$a_{ij} := \mathfrak{D}_j[\sigma_i] \tag{8}$$

the dose deposited at the jth grid point (r_j, θ_j), in the patient's cross-section Ω, due to a unit intensity ray $\sigma_i(u, w)$, and define vectors $a^j = (a_{ij})_{i=1}^{I} \in R^I$, for $j = 1, 2, \ldots, J$. Then the right-hand side of (7) becomes equal to the inner product $\langle a^j, x \rangle = \sum_{i=1}^{I} a_{ij} x_i$ in R^I. The desired dose functional is also discretized by defining

$$b_j := D(r_j, \theta_j), \text{ for all } j = 1, 2, \ldots, J. \tag{9}$$

Problem 3. ***The fully discretized inverse problem of IMRT***. *Let a_{ij} be as in* (*8*) *and let b_j be the desired doses as in* (*9*), *for j* = 1, 2, . . . , *J, and i* = 1, 2, . . . , *I. Find a radiation intensity vector $x^* \in R^I$ such that*

$$\langle a^j, x^* \rangle = b_j,\ for\ \mathrm{j} = 1, 2, \ldots, J, \tag{10}$$

$$and\ x_i^* \geq 0,\ for\ i = 1, 2, \ldots, I. \tag{11}$$

Defining the $J \times I$ matrix A as the matrix whose transpose A^T has a^j in its *j*th column, and the *J*th dimensional vector $b = \left(b_j\right)_{j=1}^J$, the system (10)–(11) can be rewritten as

$$Ax^* = b\ and\ x^* \geq 0. \tag{12}$$

This fully discretized model calls for the quantities a_{ij} which can be precalculated with any state-of-the-art forward-problem-solver. Numerous iterative techniques are available for the solution of (12), some of which are discussed in the sequel. The tendency to make the discretization finer results in very large values of I and J. If the available treatment machine cannot deliver such finely discretized radiation intensity fields, by shooting energy along rays, we need an additional computational step after a solution vector x^* (or an approximation thereof) of the system (12) has been obtained. This is a “consolidation” step in which a clinically acceptable machine setup, usually with few (up to 5 to 6) beam positions, is derived from the fully discretized solution vector x^* by using the individual ray intensities to rank the prominence of beams; see, e.g., Censor, Altschuler, and Powlis (1988a). Modern computer-controlled *multileaf collimator* (MLC) technology, capable of generating arbitrary intensity modulation, fills in the gap that existed between the fully discretized beamlet-based solution of the inverse problem and the delivery capabilities; see, e.g., Cho and Marks (2000) and references therein. To sum up, the fully discretized model is not difficulties-free, but it offers a route of circumventing the inversion problem of the computational dose operator $\mathfrak{D}$ without compromising on any of the heuristics and empiricism involved in advanced dose calculations. Brahme (1995) reaches also a conclusion in favor of full discretization and says: “...In either case it is very useful to transform the relevant integral equation into an algebraic form by discretizing the transport quantities along the coordinates of the free variables.”

The Feasibility Approach

The feasibility formulation relaxes the equality in (1). Let $\overline{D} = \overline{D}(r,\theta)$ and $\underline{D} = \underline{D}(r, \theta)$ be two dose functions whose values represent upper and lower bounds, respectively, on the permitted and required dose inside the patient's cross-section.

Problem 4. ***The feasibility formulation for the continuous inverse problem of IMRT***. *Assume that the cross-section* Ω *of the patient and its radiation absorption characteristics are known. Given prescribed dose functions* $\overline{D}(r,\theta)$, *and* $\underline{D}(r, \theta)$, *find a radiation intensity function* $\rho(u, w)$ *such that*

$$\underline{D}(r,\theta) \leq \mathfrak{D}\left[\rho(u,w)\right](r,\theta) \leq \overline{D}(r,\theta), \textit{ for all } (r,\theta) \in \Omega, \tag{13}$$

where $\mathfrak{D}$ *is the dose operator*.

A radiation therapist defines $\overline{D}(r,\theta)$ and $\underline{D}(r, \theta)$ for each given case and will accept as a solution to the IMRT inverse problem any radiation intensity function $\rho(u, w)$ that satisfies (13). In target regions (tumors) the lower bound $\underline{D}$ is usually the important factor because the dose there should exceed that given value. In critical organs and other healthy tissues $\underline{D}(r, \theta) = 0$, and $\overline{D}(r,\theta)$ is the dose that cannot be exceeded. Any solution $\rho(u, w)$ that fulfills (13), for given $\overline{D}$ and $\underline{D}$, is a *feasible solution* to the IMRT continuous inverse problem 4. In order to discretize (13) we must specify the dose functions $\overline{D}$ and $\underline{D}$ at the grid points by giving, for all $j = 1, 2, \ldots, J$,

$$\overline{D}\left(r_j,\theta_j\right) = \overline{D}_j \text{ and } \underline{D}\left(r_j,\theta_j\right) = \underline{D}_j \tag{14}$$

thus, converting (13) into a finite system of *interval inequalities*

$$\underline{D}_j \leq \mathfrak{D}_j\left[\rho\right] \leq \overline{D}_j, \quad j = 1,2,\ldots,J. \tag{15}$$

Denoting hereafter by $\overline{D}$ ($\underline{D}$) the J-dimensional column vector whose jth component is $\overline{D}_j$($\underline{D}_j$), the inverse problem of IMRT can be restated as follows:

Problem 5. ***The feasibility formulation for the fully discretized inverse problem of IMRT***. *Assume that the cross-section* Ω *of the patient and its radiation absorption characteristics are known. Given vectors* $\overline{D} = (\overline{D}_j)$ *and* $\underline{D} = (\underline{D}_j)$ *of permitted and required doses, respectively, at J grid points in the patient's cross-section* Ω, *find a radiation intensity vector* $x \in R^I$ *such that*

$$\underline{D}_j \leq \sum_{i=1}^{I} x_i a_{ij} \leq \overline{D}_j. \quad j = 1,2,\ldots,J, \tag{16}$$

$$x_i \geq 0, \quad i = 1, 2, \ldots, I. \tag{17}$$

where a_{ij} are as in (*8*).

Let the set of pixels in the discretized patient's cross-section be denoted by $N = \{1, 2, \ldots, J\}$, so that the jth pixel is identified with the jth grid point at (r_j, θ_j). Organs within the patient's cross-section are then defined as subsets of N. The subsets $B_l \subset$ N, where $l = 1, 2, \ldots, L$, denote L *critical organs* that have to be spared from excessive radiation. Let the values b_l denote the corresponding upper bounds on the dose permitted in each critical organ. The subsets $T_q \subset N$, where $q = 1, 2, \ldots, Q$, denote Q *target regions*. Let the values t_q denote the corresponding prescribed lower bounds for the absorbed dose in each target organ. All the B_l and T_q are pairwise disjoint. The set of pixels inside the patient's cross-section that are not in any B_l or T_q are called the *complement*, denoted as the subset $C \subset N$, and c is the upper bound for the permitted dose there. It is assumed that the definition of all subsets B_l and T_q and C and the prescription of all b_l, t_q, and c are given by the radiotherapist as input data for the treatment planning process. Problem (16)–(17) then becomes the following system of linear inequalities.

$$\sum_{i=1}^{I} a_{ij} x_i \leq b_l, \quad \text{for all } j \in B_l, l = 1, 2, \ldots, L, \tag{18}$$

$$t_q \leq \sum_{i=1}^{I} a_{ij} x_i, \quad \text{for all } j \in T_q, q = 1, 2, \ldots, Q, \tag{19}$$

$$\sum_{i=1}^{I} a_{ij} x_i \leq c, \quad \text{for all } j \in C \tag{20}$$

$$x_i \geq 0, \quad \text{for all } i = 1, 2, \ldots, I. \tag{21}$$

With b_l, t_q, and c given and the a_{ij}'s pre-calculated from (8) with a forward problem solver, the mathematical question represented by the basic model (18)–(21) is to find a nonnegative solution vector $x^* = (x_i^*)$ for a system of linear inequalities. This fully discretized feasibility inverse problem appeared in Altschuler and Censor (1984) and Censor, Altschuler, and Powlis (1988a).

Optimization Formulations

We use the term optimization as it is used in the field of *mathematical optimization*, namely to designate a situation where an *objective function* (also called: cost function or merit function) has to be optimized (i.e., minimized or maximized). This explains why the feasibility approach, discussed above, is not called optimization, although in the field of IMRT the term *optimization* is frequently used in a more general manner to refer to the process of treatment planning where the treatment has to be "optimized" even if the underlying mathematical model is a feasibility model where no objective function appears. When it comes to discussing an optimization approach to IMRT we must distinguish between two different kinds of optimization problems depending on the space in which they are formulated. One possibility is to define an objective function f: $R^I \rightarrow R$ over the space of radiation intensity vectors x and use either the system (12) or the constraints (18)–(21) as the feasible set (i.e., the constraints set). For example, choosing $f(x) = (1/2) \| x \|^2$ (where $\| \cdot \|$ stands for the Euclidean norm) and solving a minimization problem

$$\min\left\{(1/2) \| x \|^2 \mid (18) - (21) \text{ hold}\right\}, \tag{22}$$

leads to a minimum-norm solution vector x^*; i.e., a feasible vector closest to the origin so that the total radiation intensity is smallest possible in the Euclidean norm sense. This was recently studied via a special-purpose iterative minimization method in Xiao et al. (2003a).

Regardless of the specific choice of f, in this approach the *interval-constrained optimization* problem

$$\min\left\{f(x) \mid \alpha \leq Ax \leq \beta,\ \ x \geq 0\right\} \tag{23}$$

where $\alpha \leq Ax \leq \beta$ represents the system (18)–(20), with appropriately defined $\alpha, \beta \in R^I$, is still aiming at a solution of the fully discretized formulation of the inverse problem. A solution vector x^* will represent a radiation field that will deliver a dose which is both feasible (i.e., adheres to the upper and lower dose bounds imposed by the physician) and is optimal in the sense that it minimizes the objective function f. This approach of optimization in the space of radiation intensity vectors is called *radiation intensity optimization*.

The second possibility for introducing an optimization problem in IMRT is to use (12) or (18)–(21) as constraints but choose an objective function $g : R^J \rightarrow R$ defined over the space of dose vectors. Such objective functions may be either *biological*, or *physical*. Biological objective functions represent knowledge (statistical or other) about various biological mechanisms that affect our ability to control the disease. An example is the conditional probability of having tumor control without severe injury, denoted in the literature by P_+. Physical objective functions aggregate physical features which are important for tumor control and prevention of normal tissue complications, such as dose variance over target volume or peak dose to organs at risk. A thorough discussion of biological and physical objective functions can be found in Brahme (1995), see also Alber and Nüsslin (1999). Let us call this kind of optimization, over the space of dose vectors, *dose optimization*.

Early work on dose optimization was not geared towards solving an optimization problem but rather towards *comparing rival plans.* In this mode, several treatment plans were compared, based on their score with respect to some pre-determined quality index. The treatment plans were all fixed prior to the comparison and, therefore, the selection of the plan of choice depended largely on the choice of the quality index. Various quality indices were proposed and advocated on different grounds; see, e.g., Wolbarst et al. (1980), Dritschilo et al. (1978) and Kartha et al. (1982). In general, the dose optimization approach leads to a problem of the form

$$\min \{g(y) \mid \alpha \leq y \leq \beta\}, \tag{24}$$

where $g : R^J \to R$ assigns real values to dose vectors $y = \left(y_j\right)_{j=1}^{J} \in R^J$ whose jth component y_j is dose at pixel j. The question of feasibility versus optimization is not crucial if only radiation intensity optimization is considered because both the feasibility formulation and the optimization formulation [regardless of the particular choice of the objective function $f(x)$] occur in the same space (of radiation intensity vectors) and, thus, aim at a solution of the discretized inverse problem. Therefore, the difference between these two formulations is, from the mathematical point of view, only technical. Raphael (1992) studied the inverse problem of RTTP as constrained optimization in the L^2 Hilbert space. Recently, Cho et al. (1997) reported on the advantage of the feasibility approach over a global optimization model solved by simulated annealing; see also Cho et al. (1998). In case when the composite function $g(Ax)$ is simple enough the approach of (24) can still be efficiently used for solving directly the discretized inverse problem in its full generality. Otherwise, the inversion problem has to be abandoned and the optimization can be performed with respect to only few parameters of the external radiation field. See, for example, Gustafsson (1996) and Gustafsson, Lind, and Brahme (1994). This is done while other important parameters are left out of the optimization problem and must be given as input to the process; see also the discussion in Censor and Zenios (1997, section 11.7). The question whether to use biological or physical objective functions in the space of dose vectors (and thereby possibly compromise on the full generality of the inverse problem) remains unsettled.

Mathematical Optimization Techniques

A variety of mathematical optimization techniques have been applied to the inverse problem of IMRT. Additional methods and approaches are being applied and tested as the collaboration between researchers in this field with experts in mathematical optimization and operations research surges in recent years. This trend is evident from the growing number of special issues devoted to the interface between optimization theory and radiation therapy, e.g., Lee and Sofer (2003), Holder and Newman (2003), and Ferris and Zhang (2003), and the recent dedicated site on the Internet (Holder 2003). In this section we briefly review the following methods and approaches that have been applied to solving the inverse problem of IMRT, with emphasis on more recent publications in each category: (1) Simulated annealing and global optimization, (2) Multi-objective optimization, (3) Linear optimization and mixed integer programming (MIP), and (4) Cimmino's algorithm and other projection methods.

Other optimization models and methods, not mentioned here, were also used in RTTP in recent years; see Shepard et al. (2000), Xing and Chen (1996), Bortfeld et al. (1990) and the gradient and gradient-like methods of Spirou and Chui (1998) and others.

Simulated Annealing And Global Optimization

The NEOS (Network Enabled Optimization System) Guide Optimization Tree (at: http://www-fp.mcs.anl.gov/otc/Guide/OptWeb/index.html) uses, in its *Introduction to Global Optimization*, the following definition: "Global optimization is the task of finding the absolutely best set of admissible conditions to achieve your objective, formulated in mathematical terms. It is the hardest part of a subject called nonlinear programming (NLP)." It goes on to supply references and links to the field that are most useful to anyone who wishes to learn about it. A general mathematical optimization problem has the form

$$\min\{f(x) \mid x \in Q\}, \tag{25}$$

where

$$Q = \{x \in R^{l} \mid x \in \Gamma, g_j(x) \le 0, (j = 1, 2, \ldots, J), h_m(x) = 0, (m = 1, 2, \ldots, M)\} \tag{26}$$

is the feasible set of the problem, represented by a set-constraint Γ and equality $h_m(x) = 0$ and inequality $g_j(x) \le 0$ constraints. A point $x^* \in Q$ is a global optimal solution (global minimizer) of (25) if

$$f(x^*) \le f(x), \text{ for all } x \in Q. \tag{27}$$

A point $\tilde{x} \in Q$ is a *local optimal solution* (local minimizer) of (25) if there exists a neighborhood $U \subset R^I$ of $\tilde{x}$ so that

$$f(\tilde{x}) \leq f(x), \quad \text{for all } x \in Q \cap U. \tag{28}$$

The problem (25) is *multi-extremal* if it has multiple local minimizers with different objective function values. The occurrence of multiple extrema makes problem solving in nonlinear optimization a hard task. Without supplying global information, which is usually unavailable, the search for a global optimizer is not simple. There are stochastic methods and deterministic methods for global optimization, but one can also classify different methods based on their underlying philosophy, as Rinnooy Kan and Timmer (1989) do, as follows. (a) *Partition and search*: the feasible set *Q* is partitioned into successively smaller subregions among which the global minimum is sought. (b) *Approximation and search*: the objective function *f* is replaced by an increasingly better approximation that is easier from a computational point of view. (c) *Global decrease*: in this class of methods the aim is for permanent improvement in the values of *f*, culminating in arrival at the global minimum. (d) *Improvement of local minima*: exploiting the availability of an efficient local search routine, these methods seek to generate a sequence of local minima of decreasing function values. (e) *Enumeration of local minima*: here one strives to reach a complete enumeration of all local minima or, at least, of a promising subset of them.

Simulated annealing (SA) is a global optimization method. Its fundamental idea appears in Metropolis et al. (1953) and was applied to optimization problems by Kirkpatrick, Gelatt, and Vecchi (1983). The underlying principle of SA is to simulate the cooling process of material in a heat bath and it uses this simulation to systematically search for feasible points in a way that makes the generated sequence converge to a global minimum. Webb first introduced the SA algorithm into the field of RTTP (Webb 1989); see also his book (Webb 2001). A concise description of Webb's application of SA appears in his book (Webb 1993, subsection 2.5.4).

Global optimization can be used also for the IMRT inverse problem when the trajectories of the leaves of the MLC are integrated into the model. This has been recently done by Trevo et al. (2003) who arrived at a very high-dimensional constrained nonlinear global optimization problem and solved it by a, commercially available, software package called LGO (Lipschitz (continuous) Global Optimizer).

Multi-Objective Optimization

Multi-objective (also called *multicriteria*) *optimization* handles problems in which more then one objective function is defined over the feasible set. The standard form of such a problem is

$$\min\{F(x) \mid x \in Q\}, \tag{29}$$

where $F(x)$ is a vector of (objective) functions, i.e., for $S \geq 2$,

$$(F(x))^T = (f_1(x), f_2(x), \dots, f_S(x)), \tag{30}$$

and for each $s = 1, 2, \dots, S$, the function $f_S(x)$ maps $R^I \to R$. We denote by $Q \subset R^I$ the feasible set that may be defined by equality constraints, inequality constraints, set constraints, or any mixture of them, as in (26). An optimal point for problem (29) is a point that is feasible ($x \in Q$) and minimizes $F(x)$. But what does it mean to minimize a vector of functions? Moving from one point x^1 in R^I to another point x^2 may cause some function values to decrease while others increase; so how should one decide if a move is acceptable? The situation thus differs from the case of *scalar optimization,* when $S = 1$, because the vector of functions $F(x)$ induces on the feasible set Q a *partial order* and not a *linear order,* i.e., not every two points are ordered by their F values. Therefore, the solution (or solutions) to problem (29) depends *a priori* on which *solution concept* is adopted for solving the problem; see, e.g., Censor (1977). One way to handle this is by employing *scalarization*. This refers to the conversion of the multi-objective problem into a family of scalar optimization problems. This family has the form

$$\min\left\{\sum_{s=1}^{S} \gamma_s f_s(x) \mid x \in Q\right\}, \tag{31}$$

where $\gamma = (\gamma_s)_{s=1}^{S} \in R^S$ is a parameter vector whose components γ_s are the weights of relative importance which combine all scalar functions $f_s(x)$ into the linear combination. The difficulty here is that one usually does not know how to choose appropriately a vector γ by which a specific scalar optimization problem of the form (31) will be picked out of the family of all possible such problems. Obviously, this choice strongly affects the final outcome.

An alternative approach to multi-objective optimization is to preserve the multi-objective nature of the problem and use a solution concept that does not involve scalarization. A frequently used such concept is the *Pareto optimality,* also termed *Pareto efficiency*.

Definition 6. *A point $x^* \in R^I$ is called* ***Pareto optimal*** (***efficient***) *for problem* (*29*) if *$x^* \in Q$ and there is no other $x \neq x^*$ such that $x \in Q$, for which $f_s(x) \leq f_s(x^*)$ for all $s =$* $1, 2, \dots,$ *S, with a strict inequality for at least one s*, $1 \leq s \leq S$.

This means that x^* is Pareto efficient if it is impossible to decrease the value of any individual scalar objective function from its value at x^* without increasing at least one other scalar objective function. See, e.g., Ehrgott's recent book (Ehrgott 2000). In a recent paper, Hamacher and Küfer (2002) propose and investigate a *linear multicriteria programmin*g (LMP) problem for the inverse problem in RTTP. The concept of multicriteria optimization without prior scalarization is indeed tempting to work with.

It has been utilized in a variety of real-world problems in other technological and scientific fields. In operations research and the management sciences, multiple criteria decision making has developed into a solid body of literature in the past 30 years. The International Society on Multiple Criteria Decision Making (MCDM) offers relevant software solutions (at: http://www.mit.jyu.fi/MCDM/soft.html). Further scientific research and evaluation are needed to gauge the usefulness of this methodology in the field of RTTP. Küfer et al. (2003) develop this approach further and find that they must use adaptive reduction by appropriate approximation schemes to cope with the large scale nature of the LMP problem.

Linear Optimization And Mixed Integer Programming (MIP)

Linear optimization (traditionally called *linear programming*) is the field of study of optimization problems in which all constraints as well as the objective function are linear. The literature of this field is vast and the leading algorithms for solving such problems are the famous SIMPLEX method and primal-dual interior point methods. We direct the reader to Shepard et al. (1999, subsection 4.1) and to Holder (2003) for recent works that describe this approach and supply many useful references. An early application of linear optimization to radiotherapy treatment planning is Bahr et al. (1968) where the approach is used to optimize the treatment plan with respect to just a few setup parameters which are kept free after the plan has been obtained by the trial and error methods of those days. Rosen et al. (1991) critically compares linear optimization approaches, as used until 1990, with simulated annealing and with projection methods for the feasibility approach.

There are several ways to apply linear optimization to the IMRT inversion problem, depending mainly on the choice of the objective function. For example, if we choose to minimize the total dose to all pixels in the patient's cross-section while obeying the upper and lower bounds on organs we may consider the linear optimization problem

$$\min\left\{\sum_{j=1}^{J}\sum_{i=1}^{I} a_{ij}x_i \;\middle|\; (18)-(21) \text{ hold}\right\}. \tag{32}$$

Alternatively, one can use an organ-weighted total dose objective function of the form

$$\sum_{l=1}^{L}\beta_l \sum_{j\in B_l}\sum_{i=1}^{I} a_{ij}x_i + \sum_{q=1}^{Q}\theta_q \sum_{j\in T_q}\sum_{i=1}^{I} a_{ij}x_i + \gamma\sum_{j\in C}\sum_{i=1}^{I} a_{ij}x_i, \tag{33}$$

and minimize it over (18)–(21) after choosing user-specified weights of importance

$$\{\beta_l\}_{l=1}^{L}, \{\theta_q\}_{q=1}^{Q}$$

and γ. See, e.g., Shepard et al. (1999, subsection 4.1) for other formulations.

Mixed integer programming (MIP) occurs when some variables in an optimization problem are restricted to take only integer values. In the case of linear optimization, the MIP problem has the general form

$$\min\left\{\langle c, x\rangle \mid Ax = b,\ l \le x \le u, \quad \text{and some or all } x_i \text{ are integers}\right\}, \tag{34}$$

where $c \in R^I$, the matrix A, and the vectors $b \in R^J$, $l, u, \in R^I$, are given, see, e.g., Bixby et al. (2000). In RTTP the need for this kind of optimization arises in a natural way when *dose-volume constraints* (also called *partial volume constraints*) are considered. Such constraints appear when the oncologist is willing to sacrifice a portion of a region at risk in order to improve the probability of curing the disease. In such a case, in addition to the upper and lower bounds on required and permitted radiation doses, as formulated in (18)–(21), he might state that "up to φ% of all pixels inside a certain organ B_l in (18) might be allowed to exceed b_l by ψ%", without specifying *a priori* which of the pixels in B_l will actually use this relaxed upper bound. The MIP formulation reached at in this way can be found in Langer et al. (1996) and also in Shepard et al. (1999, p. 737). Other applications of MIP in this field include Lee, Fox, and Crocker (2000) who used it for radiosurgery treatment planning, Boland, Hamacher, and Lenzen (2002) who employed a nonlinear MIP formulation to incorporate MLC settings within the treatment planning, and Bednarz et al. (2002) who compared MIP performance with that of Cimmino's algorithm. Ferris, Meyer, and D'Souza (2002) give details of the mathematical formulations and algorithmic approaches as well as pointers to supporting literature for MIP-based approaches to problems of RTTP. As Ferris, Meyer, and D'Souza correctly notice, the main difficulty associated with the MIP approach is that these formulations can become quickly impractical due to large numbers of voxels in the region of interest (i.e., the number J, above). These difficulties have then to be attacked by approximate techniques.

Cimmino's Algorithm And Other Projection Methods

The *convex feasibility problem* is to find a (i.e., any) point in the nonempty intersection $C := \cap_{j=1}^{J} C_j \neq \emptyset$ of a family of closed convex subsets $C_j \subseteq R^I$, $1 \le j \le J$, of the I-dimensional Euclidean space. It is a fundamental problem in many areas of mathematics and the physical sciences, see, e.g., Combettes (1993, 1996) and references therein. It has been used to model significant real-world problems such as image reconstruction from projections [see, e.g., Herman (1980)] and crystallography [see Marks, Sinkler, and Landree (1999)] and has been used under additional names such as *set theoretic estimation* or the *feasible set approach*. A common approach to such problems is to use projection algorithms; see, e.g., Bauschke and Borwein (1996). *Projection algorithms* employ projections onto convex sets in various ways. They may use different kinds of projections and, sometimes, even use different projections within the same algorithm. They serve to solve a variety of problems, which are either of the feasibility or the optimization types. They have different algorithmic structures, of which some are particularly suitable for parallel computing, and they demonstrate nice convergence properties and/or good initial behavior patterns. This class of algorithms has witnessed great progress in recent years and its member algorithms have been applied with success to fully discretized models of problems in image reconstruction and image processing; see, e.g., Stark and Yang (1998), Censor and Zenios (1997).

Projection algorithms often employ *orthogonal projections* (i.e., nearest point mappings) onto the individual sets C_j. The orthogonal projection $P_\Omega(z)$ of a point $z \in R^I$ onto a closed convex set $\Omega \subseteq R^I$ is defined by

$$P_\Omega(z) := \arg\min\left\{ \| z - x \| \;\middle|\; x \in \Omega \right\}. \tag{35}$$

Frequently a *relaxation parameter* is introduced so that

$$P_{\Omega,\lambda}(z) := (1 - \lambda)z + \lambda P_\Omega(z) \tag{36}$$

is the relaxed projection of z onto Ω with relaxation λ. Another problem that is related to the convex feasibility problem is the *best approximation problem* of finding the projection of a given point $y \in R^I$ onto the nonempty intersection C of a family of closed convex subsets $C_j \subseteq R^I$, $1 \leq j \leq J$; see, e.g., Deutsch's recent book (Deutsch 2001). In both problems the convex sets $\left\{C_j\right\}_{j=1}^{J}$ represent mathematical constraints obtained from the modeling of the real-world problem, e.g., in IMRT, each constraint of (18)–(20) can be used to define a halfspace C_j. In the convex feasibility approach any point in the intersection is an acceptable solution to the real-world problem whereas the best approximation formulation is usually appropriate if some point $y \in R^I$ is given and one wishes to find the point in the intersection of the convex sets which is closest to the point y. Iterative projection algorithms for finding a projection of a point onto the intersection of sets are more complicated than algorithms for finding just any feasible point in the intersection. This is so because they must have, in their iterative steps, some built-in "memory" mechanism to remember the original point whose projection is sought after. The sequential or parallel algorithms of Dykstra [see, e.g., Bregman, Censor, and Reich (1999)], Haugazeau [see, e.g., Bauschke and Combettes (2001)], Bauschke (1996), and others and their modifications employ different such memory mechanisms.

Projection algorithmic schemes for the convex feasibility problem or for the best approximation problem are, in general, either *sequential* or *simultaneous* or *block-iterative* (see, e.g., Censor and Zenios (1997) for a classification of projection algorithms into such classes, and the review paper of Bauschke and Borwein (1996) for a variety of specific algorithms of these kinds). In what follows we explain and demonstrate these structures along with the recently proposed *string-averaging* structure. The philosophy behind these algorithms is that it is easier to calculate projections onto the individual sets C_j than onto the whole intersection of sets. Thus, these algorithms call for projections onto individual sets as they proceed sequentially, simultaneously, or in the block-iterative or the string-averaging algorithmic modes.

The String-Averaging Algorithmic Structure

The *string-averaging* algorithmic scheme was proposed by Censor, Elfving, and Herman (2001). For $t = 1, 2, \ldots, M$, let the *string* J_t be an ordered subset of $\{1, 2, \ldots, J\}$ of the form

$$J_t = \left(j_1^t, j_2^t, \ldots, j_{J(t)}^t\right). \tag{37}$$

with $J(t)$ denoting the number of elements in J_t. Suppose that there is a set $Q \subseteq R^I$ such that there are operators $R_1, R_2, \ldots, R_J$ mapping Q into Q and an operator R which maps $Q^M = Q \times Q \times \cdot\cdot\cdot \times Q$ (M times) into Q. Initializing the algorithm at an arbitrary $x^0 \in Q$, the iterative step of the string-averaging algorithmic scheme is as follows. Given the current iterate x^k, calculate, for all $t = 1, 2, \ldots, M$,

$$T_t\left(x^k\right) = R_{j_{J(t)}^t} \ldots R_{j_2^t} R_{j_1^t}\left(x^k\right) \tag{38}$$

and then calculate

$$x^{k+1} = R\left(T_1\left(x^k\right), T_2\left(x^k\right), \ldots, T_M\left(x^k\right)\right). \tag{39}$$

For every $t = 1, 2, \ldots, M$, this algorithmic scheme applies to x^k successively the operators whose indices belong to the tth string. This can be done in parallel for all strings and then the operator R maps all end-points onto the next iterate x^{k+1}. This is indeed an algorithm provided that the operators $\left\{R_j\right\}_{j=1}^{J}$ and R all have algorithmic implementations. In this framework we get a sequential algorithm by the choice $M = 1$ and $J_1 = (1, 2, \ldots, J)$. The well-known "Projections Onto Convex Sets" (POCS) algorithm for the convex feasibility problem is such a *sequential* projection algorithm; see Bregman (1965), Gubin, Polyak, and Raik (1967), Youla (1987), and the review papers by Combettes (1993, 1996). Starting from an arbitrary initial point $x^0 \in R^I$, the POCS algorithm's iterative step is

$$x^{k+1} = x^k + \lambda_k\left(P_{C_{j(k)}}\left(x^k\right) - x^k\right), \tag{40}$$

where $\{\lambda_k\}_{k\geq 0}$ are relaxation parameters and $\{j(k)\}_{k\geq 0}$ is a *control sequence*, $1 \leq j(k) \leq m$, for all $k \geq 0$, which determines the individual set $C_{j(k)}$ onto which the current iterate x^k is projected. A commonly used control is the *cyclic control* in which $j(k) = k \bmod J + 1$, but other controls are also available (Censor and Zenios 1997). This algorithm was used in RTTP in Censor, Altschuler, and Powlis (1988b) and by Cho and Marks and co-workers in Cho et al. (1997, 1998) and Cho and Marks (2000). The celebrated ART (Algebraic Reconstruction Technique) of Gordon, Bender, and Herman (1970) [see also Herman (1980)], is equivalent to the application of POCS to a system of linear equations.

A *simultaneous* algorithm is obtained by the choice $M = J$ and $J_t = (t)$, $t = 1, 2, \ldots, M$, and Cimmino's projections method is indeed such an algorithm. Using relaxation parameters $\{\lambda_k\}_{k\geq 0}$ and *weights of importance* $\left\{w_j\right\}_{j=1}^{J}$, such that $w_j > 0$ and $\sum_{j=1}^{J} w_j = 1$, the iterative step of Cimmino's algorithm for the derivation of the next iterate x^{k+1} from the current one x^k is

$$x^{k+1} = x^k + \lambda_k \left(\sum_{j=1}^{J} w_j P_{C_j}\left(x^k\right) - x^k \right). \tag{41}$$

For halfspaces as constraints sets, i.e.,

$$C_j = \left\{ x \in R^I \;\middle|\; \left\langle a^j, x \right\rangle \leq d_j \right\}, \quad \text{for all } j = 1, 2, \ldots, J, \tag{42}$$

the formula becomes:

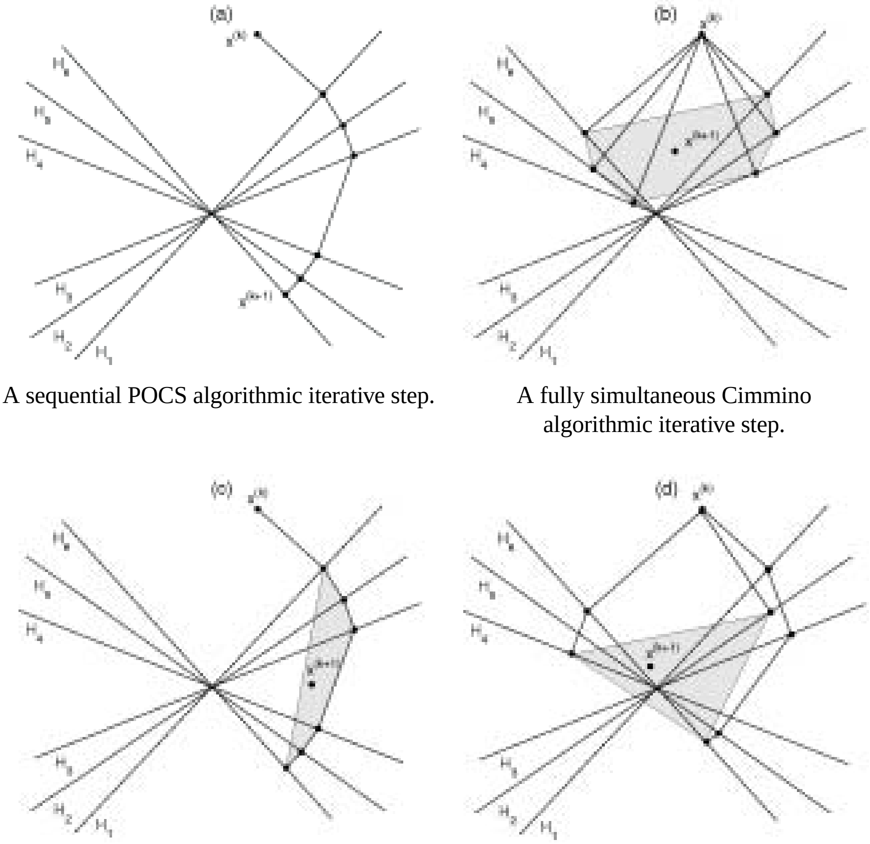
$$x^{k+1} = x^k + \lambda_k \sum_{j=1}^{n} w_j c_j\left(x^k\right) a^j, \tag{43}$$

where

$$c_j\left(x^k\right) = \min\left(0, \frac{d_j - \left\langle a^j, x^k \right\rangle}{\| a^j \|^2} \right). \tag{44}$$

String-averaging schemes offer a variety of options for steering the iterates towards a solution of the convex feasibility problem. It is an *inherently parallel* scheme in that its mathematical formulation is parallel (like the fully simultaneous method mentioned above). We use this term to contrast such algorithms with others that are sequential in their mathematical formulation but can, sometimes, be implemented in a parallel fashion based on appropriate model decomposition (i.e., depending on the structure of the underlying problem). Being inherently parallel, this algorithmic scheme enables flexibility in the actual manner of implementation on a parallel machine. At the extremes of the "spectrum" of possible specific algorithms, derivable from the string-averaging algorithmic scheme, are the generically sequential method, which uses one set at a time, and the fully simultaneous algorithm, which employs all sets at each iteration.

The *block-iterative projections* (BIP) scheme of Aharoni and Censor (1989) also has the sequential and the fully simultaneous methods as its extremes in terms of block structures, but the string-averaging algorithmic structure gives users further options to design new inherently parallel computational schemes. The behavior of the string-averaging algorithmic scheme in the inconsistent case when the intersection $C = \cap_{j=1}^{J} C_j$ is empty is not known at this time. For results on the behavior of the fully simultaneous algorithm with orthogonal projections in the inconsistent case see, e.g., Combettes (1994). We demonstrate some of the algorithmic possibilities offered by the general string-averaging method in figure 2. The constraints sets of the convex feasibility problem are assumed in this special case to be the six hyperplanes H_1, H_2, H_3, H_4, H_5, and H_6.



A sequential POCS algorithmic iterative step.

A fully simultaneous Cimmino algorithmic iterative step.

Averaging sequential strings of consecutive projections.

An example for a string-averaging algorithmic iterative step.

Figure 2. Some algorithmic possibilities offered by the string-averaging method.

The fully simultaneous Cimmino algorithm, applied to halfspaces, defined by the inequalities of the system (18)–(20), was first used in RTTP by Censor, Altschuler, and Powlis (1988a) [see also Powlis et al. (1989)]. There are several advantages of the simultaneous projections Cimmino algorithm over the sequential projections POCS algorithm for the linear feasibility problem arising from the fully discretized model of IMRT. When initialized at zero intensities it generates an approximate LIF (least-intensity feasible) solution; see Xiao et al. (2003a). It converges globally to a feasible solution, if such a solution exists, or to a minimal value of a proximity function in the inconsistent case; see Combettes (1994) and Byrne and Censor (2001). It is an inherently parallel iterative algorithm, thus, implementable on parallel computing equipment regardless of problem structure. It can be accelerated by using strong over-relaxation, see Höffner et al. (1996), or by using it with oblique projections (i.e., the CAV algorithm; see, e.g., [Xiao et al. (2203b), appendix] and references therein). It is a special case of more general algorithmic schemes, such as BIP and the string-averaging scheme, which allow processing of various sets of constraints instead of a single (POCS) or all (Cimmino) constraints, in each iterative step. It generates smoother intensity patterns; see Xiao et al. (2003b).

Acknowledgments

The author gratefully acknowledges useful discussions with Martin Altschuler, Greg Bednarz, Christopher Börgers, Yan Chen, James Galvin, Chris Hauser, Gabor Herman, Darek Michalski, William Powlis and Ying Xiao. The support and hospitality and the inspiring research atmosphere at the Medical Physics Division (Prof. James Galvin, Director), Department of Radiation Oncology, Thomas Jefferson University, Philadelphia, PA, where some of the work reported here was done, are gratefully acknowledged. This work was partially supported by grant 592/00 of the Israeli Science Foundation, founded by The Israel Academy of Sciences and Humanities. Some material was adopted from Censor (1999) which appeared in the book: *Computational Radiology and Imaging: Therapy and Diagnosis*, C. Börgers and F. Natterer (eds.), The IMA Volumes in Mathematics and its Applications, Vol. 110, New York, NY: Springer-Verlag, copyright Springer-Verlag, with permission.

References

Aharoni, R., and Y. Censor. (1989). “Block-iterative projection methods for parallel computation of solutions to convex feasibility problems.” *Linear Algebra and Its Applications* 120:165–175.

Alber, M., and F. Nüsslin. (1999). “An objective function for radiation treatment optimization based on local biological measures.” *Phys. Med. Biol.* 44:479–493.

Altschuler, M. D., and Y. Censor. “Feasibility Solutions in Radiation Therapy Treatment Planning” in *Proceedings of the Eighth International Conference on the Use of Computers in Radiation Therapy*. J.R. Cunningham, D. Ragan, and J. Van Dyk, (eds.), Silver Spring, MD, USA: IEEE Computer Society Press, pp. 220–224, 1984.

Bahr, G. K., J. G. Kereiakes, H. Horowitz, R. Finney, J. Galvin, and K. Goode. (1968). “The method of linear programming applied to radiation treatment planning.” *Radiol*. 91:686–693.

Bauschke, H. H. (1996). “The approximation of fixed points of compositions of nonexpansive mappings in Hilbert space.” *J. Math. Analys. Appl.* 202:150–159.

Bauschke, H. H., and J. M. Borwein. (1996). “On projection algorithms for solving convex feasibility problems.” *SIAM Review* 38:367–426.

Bauschke, H. H., and P. L. Combettes. (2001). “A weak-to-strong convergence principle for Fejér-monotone methods in Hilbert spaces.” *Math. Oper. Res*. 26:248–264.

Bednarz, G., D. Michalski, C. Hauser, M. S. Huq, Y. Xiao, P. R. Anne, and J. M. Galvin. (2002). “The use of mixed-integer programming for inverse treatment planning with pre-defined field segments.” *Phys. Med. Biol*. 47:1–11.

Bixby, R. E., M. Fenelon, Z. Gu, E. Rothberg, and R. Wunderling. “MIP: Theory and Practice—Closing the Gap” in *System Modelling and Optimization: Methods, Theory and Applications*. M. J. D. Powell and S. Scholtes (eds.). Boston, MA: Kluwer Academic Publishers, 2000.

Boland, N., H. W. Hamacher, and F. Lenzen. “Minimizing beam on time in cancer radiation treatment using multileaf collimators.” Technical report in Wirtschaftsmathematik Nr. 78/2002. Department of Mathematics, University of Kaiserslautern, Germany. 2002.

Bortfeld, T., J. Bürkelbach, R. Boesecke, and W. Schlegel. (1990). “Methods of image reconstruction from projections applied to conformation radiotherapy.” *Phys. Med. Biol*. 35:1423–1434.

Brahme, A. “Treatment Optimization Using Physical and Radiological Objective Functions” in *Medical Radiology: Radiation Therapy Physics*. A.R. Smith (eds.), Berlin: Springer-Verlag, pp. 209–246, 1995.

Brahme, A., J.-E. Roos, and I. Lax. (1982). “Solution of an integral equation encountered in rotation therapy.” *Phys. Med. Biol*. 27:1221–1229.

Bregman, L. M. (1965). “The method of successive projections for finding a common point of convex sets.” *Soviet Mathematics Doklady* 6:688–692.

Bregman, L. M., Y. Censor, and S. Reich. (1999). “Dykstra’s algorithm as the nonlinear extension of Bregman’s optimization method.” *J. Convex Analys*. 6:319–333.

Byrne, C., and Y. Censor. (2001). “Proximity function minimization using multiple Bregman projections, with applications to split feasibility and Kullback-Leibler distance minimization.” *Ann. Oper. Res*. 105:77–98.

Censor, Y. (1977). “Pareto optimality in multiobjective problems.” *Appl. Math. Optim*. 4:41–59.

Censor, Y. “Mathematical Aspects of Radiation Therapy Treatment Planning: Continuous Inversion Versus Full Discretization and Optimization Versus Feasibility” in *Computational Radiology and Imaging: Therapy and Diagnosis*, C. Börgers and F. Natterer (eds.). The IMA Volumes in Mathematics and its Applications, Vol. 110, New York: Springer-Verlag, pp. 101–112, 1999.

Censor, Y., and S. A. Zenios. *Parallel Optimization: Theory, Algorithms, and Applications*. New York: Oxford University Press, 1997.

Censor, Y., M. D. Altschuler, and W. D. Powlis. (1988a). “On the use of Cimmino’s simultaneous projections method for computing a solution of the inverse problem in radiation therapy treatment planning.” *Inverse Problems* 4:607–623.

Censor, Y., M. D. Altschuler, and W. D. Powlis. (1988b). “A computational solution of the inverse problem in radiation therapy treatment planning.” *Appl. Math. Comput*. 25:57–87.

Censor, Y., T. Elfving, and G. T. Herman. “Averaging Strings of Sequential Iterations for Convex Feasibility Problems” in *Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications*. D. Butnariu, Y. Censor, and S. Reich (eds.). Amsterdam: Elsevier Science Publishers, 2001, pp. 101–114, 2001.

Cho, P. S., and R. J. Marks II. (2000). “Hardware-sensitive optimization for intensity modulated radiotherapy.” *Phys. Med. Biol*. 45:429–440.

Cho, P. S., S. Lee, R. J. Marks II, J. A. Redstone, and S. Oh. “Comparison of Algorithms for Intensity Modulated Beam Optimization: Projections onto Convex Sets and Simulated Annealing” in *Proceedings of the XII International Conference on the Use of Computers in Radiation Therapy*. May 27–30, 1997, Salt Lake City, Utah. D. D. Leavitt, and G. Starkschall (eds.). Madison, WI: Medical Physics Publishing, pp. 310–312, 1997.

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

Combettes, P. L. (1993). “The foundations of set-theoretic estimation.” *Proc. IEEE* 81:182–208.

Combettes, P. L. (1994). “Inconsistent signal feasibility problems: Least squares solutions in a product space.” *IEEE Trans. Sig. Proc*. 42:2955–2966.

Combettes, P. L. (1996). “The convex feasibility problem in image recovery.” *Adv. Imag. Electron Phys*. 95:155–270.

Cormack, A. M., and E. T. Quinto. (1990). “The mathematics and physics of radiation dose planning using X-rays.” *Contemp. Math*. 113:41–55.

Deutsch, F. *Best Approximation in Inner Product Spaces*. New York: Springer-Verlag, 2001.

Dritschilo, A., J. T. Chaffey, W. D. Bloomer, and A. Mark. (1978). “The complication probability factor: A method for selection of radiation treatment plans.” *Br. J. Radiol*. 51:370–374.

Ehrgott, M. *Multicriteria Optimization*. Lecture Notes in Economics and Mathematical Systems. Vol. 491. Berlin: Springer-Verlag, 2000.

Ferris, M. C., and Y. Zhang (eds.). Special Issue on Mathematical Programming in Biology and Medicine. *Mathematical Programming Series B,* 2003. To appear.

Ferris, M. C., R. R. Meyer, and W. D’Souza. (2002). “Radiation treatment planning: Mixed integer programming formulations and approaches.” Optimization Technical Report 02-08, Computer Science Department, University of Wisconsin, Madison, WI, October 2002.

Goitein, M. (1990). “The inverse problem.” *Int. J. Radiat. Oncol. Biol. Phys*. 18:489–491.

Gordon, R., R. Bender, and G. T. Herman. (1970). “Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography.” *J. Theoret. Biol*. 29:471–481.

Gubin, L., B. Polyak, and E. Raik. (1967). “The method of projections for finding the common point of convex sets.” *USSR Computational Mathematics and Mathematical Physics* 7:1–24.

Gustafsson, A. (1996). Development of a Versatile Algorithm for Optimization Of Radiation Therapy. Ph.D. Thesis. Department of Medical Radiation Physics. University of Stockholm, Sweden.

Gustafsson, A., B. K. Lind, and A. Brahme. (1994). “A generalized pencil beam algorithm for optimization of radiation therapy.” *Med. Phys*. 21:343–356.

Hamacher, H. W., and K.-H. Küfer. (2002). “Inverse radiation therapy planning—a multiple objective optimization approach.” *Discrete Appl. Math*. 118:145–161.

Herman, G. T. *Image Reconstruction From Projections: The Fundamentals of Computerized Tomography*. New York: Academic Press, 1980.

Höffner, J., P. Decker, E. L. Schmidt, W. Herbig, J. Rittler, and P. Weiss. (1996). "Development of a fast optimization preview in radiation treatment planning." Strahlentherap. Onkol. 172:384–394.

Holder, A. (2003). "Radiotherapy Treatment Design and Linear Programming" in *The Handbook of Operations Research/Management Science Applications in Health Care*. M. Brandeau, F. Sainfort, M. Brandeau, and W. Pierskalla (eds.). New York: Kluwer Academic Press, 2003. In press.

Holder, A. (Internet site founder and maintainer). "Operations Research & Radiation Oncology." http://www.trinity.edu/aholder/HealthApp/oncology/.

Holder, A., and F. Newman (eds.). Special Issue on Radiation Oncology. Optimization and Engineering. In preparation. http://www.trinity.edu/aholder/research/CallForPapers.html.

Kartha, P. K. I., A. Pagnamenta, A. Chung-Bin, and F. R. Hendrickson. (1982). "Optimization of radiation treatment planning for the isocentric technique by use of a quality index." *Appl. Radiol*. 11:101–110.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. (1983). "Optimization by simulated annealing." *Science* 220:671–680.

Küfer, K.-H., A. Scherrer, M. Monz, F. Alonso, H. Trinkaus, T. Bortfeld, and C. Thieke. (2003). "Intensity-modulated radiotherapy—A large scale multi-criteria programming problem." *OR Spectrum*. In press.

Langer, M., S. Morrill, R. Brown, O. Lee, and R. Lane. (1996). "A comparison of mixed integer programming and fast simulated annealing for optimizing beam weights in radiation therapy." *Med. Phys*. 23:957–964.

Lee, E. K., and A. Sofer (eds.). (2003). Special Issue on Optimization in Medicine. *Ann. Oper. Res.* 119.

Lee, E. K., T. Fox, and I. Crocker. (2000). "Optimization of radiosurgery treatment planning via mixed integer programming." *Med. Phys*. 27:995–1004.

Marks, L. D., W. Sinkler, and E. Landree. (1999). "A feasible set approach to the crystallographic phase problem." *Acta Crystallograph*. A55:601–612.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. R. Teller. (1953). "Equation of state calculation by fast computing machines." *J. Chem. Phys*. 21:1087–1091.

Powlis, W. D., M. D. Altschuler, Y. Censor, and E. L. Buhle Jr. (1989). "Semi-automatic radiotherapy treatment planning with a mathematical model to satisfy treatment goals." *Int. J. Radiat. Oncol. Biol. Phys*. 16:271–276.

Raphael, C. (1992). "Radiation therapy treatment planning: An L^2 approach." *Appl. Math. Computat*. 52:251–277.

Rinnooy Kan, A. H. G., and G. T. Timmer. (1989). "Global Optimization" in *Optimization. Handbooks in Operations Research and Management Science, Vol. 1*. G. L. Nemhauser, A. H. G. Rinnooy Kan, and M. J. Todd (eds.). Amsterdam: Elsevier Science Publishers B.V., pp. 631–662, 1989.

Rosen, I. I., R. G. Lane, S. M. Morrill, and J. A. Belli. (1991). "Treatment plan optimization using linear programming." *Med. Phys.* 18:141–152.

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

Shepard, D. M., G. H. Olivera, P. J. Reckwerdt, and T. R. Mackie. (2000). "Iterative approaches to dose optimization in tomotherapy." *Phys. Med. Biol*. 45:69–90.

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

Stark, H., and Y. Yang. *Vector Space Projections: A Numerical Approach to Signal and Image Processing, Neural Nets, and Optics*. New York: John Wiley & Sons, 1998.

Trevo, J., P. Kolmonen, T. Lyyra-Laitinen. J. D. Pinter, and T. Lahtinen. (2003). "An optimization-based approach to the multiple static delivery technique in radiation therapy." *Ann. Oper. Res*. 119. In press.

Webb, S. (1989). "Optimisation of conformal radiotherapy dose distribution by simulated annealing." *Phys. Med. Biol*. 34:1349–1370.

Webb, S. *The Physics of Three-Dimensional Radiation Therapy*. Bristol, UK: Institute of Physics Publishing (IOP), 1993. Reprinted with corrections 2001.

Webb, S. *Intensity-Modulated Radiation Therapy*. Bristol, UK: Institute of Physics Publishing (IOP), 2001.

Wolbarst, A. B., E. S. Sternick, B. H. Curran, and A. Dritschilo. (1980). "Optimized treatment planning using the complication probability factor (CPF)." *Int. J. Radiat. Oncol. Biol. Phys.* 6:723–728.

Xiao, Y., Y. Censor, D. Michalski, and J. M. Galvin. (2003a). "The least intensity feasible solution for aperture-based inverse planning in radiation therapy." *Ann. Oper. Res.* 119:183–203.

Xiao, Y., Y. Censor, D. Michalski, and J. M. Galvin. (2003b). "Inherent smoothness of intensity patterns for intensity modulated radiation therapy generated by a simultaneous projection algorithm." Technical Report, May 19, 2002. Revised: January 21, 2003.

Xing, L., and G. T. Y. Chen. (1996). "Iterative methods for inverse treatment planning." *Phys. Med. Biol.* 41:2107–2123.

Youla, D.C. "Mathematical Theory of Image Restoration by the Method of Convex Projections" in *Image Recovery: Theory and Applications*. H. Stark, (ed.). Orlando, Florida: Academic Press, pp. 29–77, 1987.

Intensity modulated radiation therapy (IMRT) is an advanced radiation therapy treat-ment method, for which equipment and treatment planning algorithms have been contin-. 2. uously developed for more than a decade. With the help of computer-controlled linear accelerator, the IMRT machine is capable of delivering a high-precision dose distribution that conforms to the three-dimensional shape of the tumor, and creates sharp dose gradi-ents (measured by how quickly the dose changes between two adjacent areas with dierent dose levels) to eectively avoid the tissues surrounding the tumor (see Intens... In order to facilitate the leaf sequencing process in intensity modulated radiation therapy (IMRT), and design of a practical leaf sequencing algorithm, it is an important issue to smooth the planned fluence maps. The objective is to achieve both high-efficiency and high-precision dose delivering by considering characteristics of leaf sequencing process. The key factor which affects total number of monitor units for the leaf sequencing optimization process is the max flow value of the digraph which formulated from the fluence maps. Therefore, we believe that one strategy for compromising dose