On Linear Infeasibility Arising in Intensity-Modulated Radiation Therapy Inverse Planning

Yair Censor[1], Adi Ben-Israel[2], Ying Xiao[3]
and James M. Galvin[3]

[1]Department of Mathematics, University of Haifa,
Mt. Carmel, Haifa 31905, Israel.
(yair@math.haifa.ac.il)

[2]RUTCOR-Rutgers Center for Operations Research,
Rutgers, The State University of New Jersey,
Piscataway, NJ 08854, USA.
(adi.benisrael@gmail.com)

[3]Medical Physics Division, Radiation Oncology Department,
Thomas Jefferson University Hospital,
Philadelphia, PA 19107, USA.
({ying.xiao, james.galvin}@mail.tju.edu).

November 28, 2006. Revised: November 1, 2007.

Abstract

Intensity–modulated radiation therapy (IMRT) gives rise to systems of linear inequalities, representing the effects of radiation on the irradiated body. These systems are often infeasible, in which case one settles for an approximate solution, such as an $\{\alpha, \beta\}$–relaxation, meaning that no more than α percent of the inequalities are violated by no more than β percent. For real-world IMRT problems, there is a

feasible $\{\alpha, \beta\}$–relaxation for sufficiently large $\alpha, \beta > 0$, however large values of these parameters may be unacceptable medically.

The $\{\alpha, \beta\}$–relaxation problem is combinatorial, and for given values of the parameters can be solved exactly by Mixed Integer Programming (MIP), but this may be impractical because of problem size, and the need for repeated solutions as the treatment progresses.

As a practical alternative to the MIP approach we present a heuristic non–combinatorial method for finding an approximate relaxation. The method solves a Linear Program (LP) for each pair of values of the parameters $\{\alpha, \beta\}$ and progresses through successively increasing values until an acceptable solution is found, or is determined non–existent. The method is fast and reliable, since it consists of solving a sequence of LP's.

1 Introduction

The fully-discretized feasibility model of the inverse problem of intensity-modulated radiation therapy (IMRT) gives rise to a system of linear inequalities that describes the effects of radiation on the irradiated body and the treatment prescription, see Censor, Altschuler and Powlis [8, 9], see also Censor [7]. As an illustration, consider a simple representative system

$$A_1\mathbf{x} \leq \mathbf{u}^1, \tag{1}$$

$$A_2\mathbf{x} \geq \boldsymbol{\ell}^2, \tag{2}$$

$$\mathbf{x} \geq \mathbf{0}, \tag{3}$$

where the nonnegative nonzero matrices $A_1 \in R^{n_1 \times m}$ and $A_2 \in R^{n_2 \times m}$ and vectors $\mathbf{u}^1 \in R^{n_1}$ and $\boldsymbol{\ell}^2 \in R^{n_2}$ are given. The inequalities (1) represent healthy tissues where radiation is undesirable (hence the upper bound $\mathbf{u}^1$), and the inequalities (2) represent malignant tissues that must receive a minimal amount of radiation dose (expressed by the lower bound $\boldsymbol{\ell}^2$). The IMRT application is described in more detail in the Section 3.

Often the system (1)–(3) is infeasible, in which case the inequalities may be relaxed to obtain a feasible solution. *Relaxation* means here changing the right-hand side of an inequality so as to allow more feasible solutions, by raising the upper bounds in the inequalities of type (1), and/or lowering the lower bounds in the inequalities of type (2). There is a sizable literature on inconsistent subsystems of linear systems, see, e.g., Chinneck [12, 13] for a

thorough and up to date coverage of this field. However, it appears, from [13, **?**, **?**], that our proposed model and method has not been discussed before. Although the question: "what is the smallest adjustment to the constraints in the model that will render it feasible?" certainly has.

In Section 2 we describe and justify our approach. In Section 3 we discuss the application to IMRT and describe the linear model for the inverse problem of IMRT and its potential infeasibility. Our successive $\{\alpha, \beta\}$–relaxation method is given in Section 4 which includes also a brief discussion of other methods. Experimental results are presented in Section 5 and some concluding comments are given in Section 6.

2 A relaxation to achieve feasibility

If the system (1)–(3) is infeasible, we may relax the inequalities in a controlled manner until a feasible system is obtained. For simplicity we consider here the case where only the inequalities (1) are relaxed. In the IMRT application various subsets of the inequalities may be relaxed and several different relaxation levels may be applied to a single subset of inequalities. We define an $\{\alpha, \beta\}$-*relaxation* as follows.

Definition 2.1 *β–**relaxation** and $\{\alpha, \beta\}$–**relaxation.***

*(i) Given $\mathbf{a} \in R^m$, $b \in R$, and $\beta \geq 0$, a β–**relaxation** of the inequality,*

$$\langle \mathbf{a}, \mathbf{x} \rangle \leq b, \tag{4}$$

is the inequality

$$\langle \mathbf{a}, \mathbf{x} \rangle \leq (1 + \beta)\, b\,. \tag{5}$$

*(ii) Given $0 \leq \alpha \leq 1$ and $\beta \geq 0$, an $\{\alpha, \beta\}$–**relaxation** of the system (1) is any system obtained from (1) in which at most a fraction α of the constraints (1) undergoes a β–relaxation.*

The parameter β is a measure of the relaxation (violation) of the original inequality constraint. This relaxation can be modelled by a linear program (LP) that attempts to minimize the violations. Let the rows of the $n_1 \times m$ matrix A_1 be denoted by $\mathbf{a}^j$. The β–relaxations of the inequalities (1) can be written, for $j = 1, 2, \ldots, n_1$, as follows,

$$\left\langle \mathbf{a}^j, \mathbf{x} \right\rangle \leq t_j\, u_j^1, \tag{6}$$

$$1 \leq t_j \leq (1 + \beta), \tag{7}$$

where u_j^1 are the components of the upper bound $\mathbf{u}^1$. We relax the inequalities (7) further,

$$0 \leq t_j \leq (1+\beta), \tag{8}$$

noting that smaller values of t_j are preferred, since they represent less radiation on healthy tissues. We collect the inequalities (6), (8) in the system,

$$A_1\mathbf{x} \leq U_1\mathbf{t}, \quad \mathbf{0} \leq \mathbf{t} \leq (1+\beta)\mathbf{1}, \tag{9}$$

where U_1 is the diagonal matrix with the components of $\mathbf{u}^1$ in its diagonal, $\mathbf{t} = (t_j)$ and $\mathbf{1}$ is the vector of ones. Since β–relaxation applies to at most a fraction α of the n_1 constraints of (1), it follows that the components t_j of $\mathbf{t}$ must satisfy,

$$\sum_{j=1}^{n_1} t_j \leq \alpha\, n_1\, (1+\beta) + (1-\alpha)\, n_1 = n_1(1+\alpha\beta), \tag{10}$$

a consistency condition called here the **check inequality**.

We propose the following LP formulation to approximate the solution of the system (1)–(3),

$$\begin{aligned}
\min \quad & \langle \mathbf{1}, \mathbf{t} \rangle && \text{(LP}(\alpha,\beta)\text{)} \\
\text{such that} \quad & A_1\mathbf{x} \leq U_1\,\mathbf{t}, && \text{(LP.a)} \\
& A_2\mathbf{x} \geq \boldsymbol{\ell}^2, && \text{(LP.b)} \\
& \mathbf{t} \leq (1+\beta)\mathbf{1}, && \text{(LP.c)} \\
& \langle \mathbf{1}, \mathbf{t} \rangle \leq n_1(1+\alpha\beta), && \text{(LP.d)} \\
& \mathbf{x}, \mathbf{t} \geq 0, && \text{(LP.e)}
\end{aligned}$$

where (LP.d) is the check inequality (10). In applications to IMRT the parameters α and β in (LP(α,β)) are bounded above,

$$0 \leq \alpha \leq \alpha_{\max}\ ,\ 0 \leq \beta \leq \beta_{\max}, \tag{11}$$

where $\alpha_{\max}$ and $\beta_{\max}$ are the upper bounds acceptable to the user.

For IMRT problems (1)–(3) arising in practice, the inequalities (2) or (LP.b), representing the minimum dosage on the target, are assumed consistent. We also assume that the upper bound $\mathbf{u}^1$ is a positive vector. It is therefore obvious that there exist values $0 \leq \alpha \leq 1$ and $\beta > 0$ such that the problem (LP(α,β) is feasible.

However these values may violate the upper bounds (11), and therefore be unacceptable.

It should be noted that (LP(α, β)) is not a precise model of $\{\alpha, \beta\}$–relaxation, since the inequalities (8) are weaker than (7), and on the other hand, the violations may be spread over more than a fraction α of the constraints in (1). The check inequality (LP.d) is thus a relaxation of the combinatorial condition of Definition 2.1(ii). However, if the most relaxed problem (LP($\alpha_{\max}, \beta_{\max}$)) is infeasible, it follows that (LP(α, β)) is infeasible for all α, β satisfying (11). We use this LP formulation as a tool for the relaxation process and not as a model of the IMRT inverse problem. The latter is done by various workers, see, e.g., Holder [18], Romeijn *et al.* [27].

3 The linear model for the inverse problem in intensity-modulated radiation therapy

Intensity-modulated radiation therapy (IMRT) has been implemented widely since the time it was first introduced. The radiation therapy community has gathered considerable experience in taking advantage of the currently-available software and hardware tools to devise and deliver good IMRT plans that better treat targets while sparing critical structures, see, e.g., Palta and Mackie [24], Webb [30] and references therein. A critical input needed for the generation of good IMRT plans is the definition of reasonable treatment planning goals. A set of goals that are self-contradicting or too stringent may steer the planning in an undesirable direction. Based on the relaxation methodology presented in Section 2, we study here an approach that can take a set of goals that could be infeasible and gradually relax them selectively until a feasible solution is reached. This IMRT planning tool has the potential for significantly shortening the traditional iterative trial and error planning process used for this treatment modality. This is because it avoids bringing the responsible physician back to re-adjust constraints when a plan is not acceptable. Instead, the physician can choose from a number of solutions with different constraints that have been relaxed in a methodologically-organized way.

We consider a prototypical system of linear inequalities that arises in the fully-discretized approach to the inverse problem of IMRT. Historically, we proposed and studied this fully-discretized model of the inverse problem

prior to the time when IMRT was identified as a complete process, see Censor, Altschuler and Powlis [9, 8], or Censor and Zenios [11, Chapter 11]. For recent review papers see, e.g., Reemtsen and Alber [26], Shepard, Ferris, Olivera and Mackie [28], Censor [7], Galvin *et al.* [17], Bortefld [5] and references therein. The problem is formulated as follows. A system of linear inequalities of the form

$$0 \leq \sum_{v=1}^{m} a_v^j x_v \leq u_j, \quad \text{for all} \quad j \in B_s, \ \text{for all} \quad s = 1, 2, \ldots, S, \tag{12}$$

$$0 \leq l_j \leq \sum_{v=1}^{m} a_v^j x_v, \quad \text{for all} \quad j \in L_q, \quad \text{for all} \quad q = 1, 2, \ldots, Q, \tag{13}$$

$$0 \leq x_v, \quad \text{for all} \quad v = 1, 2, \ldots, m, \tag{14}$$

must be solved where the m-dimensional vector $x = (x_v)_{v=1}^m$ is the vector of unknowns. These are the *beamlets intensities* (or beam segments weights) in the inverse planning problem of IMRT and the index v counts all the beamlets in some agreed manner. The $\{B_s\}_{s=1}^S$ are specified subsets of inequalities that have some common properties. In IMRT, these represent *Organs at Risk* (OARs) whose total absorbed radiation dose should stay below the specified upper bounds, denoted by u_j. The individual inequalities for the OARs are indexed by j. The $\{L_q\}_{q=1}^Q$ are additional specified subsets of inequalities. In IMRT they represent *Planning Target Volumes* (PTVs) whose total absorbed radiation dose should stay above the specified lower bounds, denoted by l_j.

The real numbers $a_v^j \geq 0$ are modelling-related quantities that are predetermined. In IMRT they represent the dose deposited in voxel (i.e., volume-element) j due to unit radiation intensity coming from beamlet v. They are known in IMRT ahead of time by performing appropriate *"forward" dosimetry calculations*. We denote the vectors $a^j = (a_v^j)_{v=1}^m$, for all $j \in (\cup_{s=1}^S B_s) \cup (\cup_{q=1}^Q L_q)$, where the total number of elements in all the sets is identified with the total number n of inequalities in the generic system (12)–(13).

The original formulation of the inverse problem of IMRT in its fully-discretized "feasibility" (as opposed to "optimization") approach is to find an intensities vector $x = (x_v)_{v=1}^m$ such that the system (12)–(14) will be satisfied. Using special hardware called a multileaf collimator (MLC) it is possible to modulate the intensity of a radiation beam according to the individual beamlet intensities given by the vector x, see, e.g., Cho and Marks II [14] for

a recent application in the field of IMRT. However, in cases when a feasible solution to that system cannot be found it is possible to relax the system in a certain controlled manner so that a feasible solution to the relaxed system is found. Since relaxing the system means, inevitably, violating the original system to some extent, the method proposed here finds such a solution by relaxing the original system by as little as possible.

This objective is achieved by imposing on the system (12)–(14) an additional condition that will dynamically control such violation/relaxation of the original system. This condition can be formulated for more then one subset of the inequalities which represent either an OAR or a PTV and several such conditions may be applied with different levels to one organ. However, to keep the notation and presentation simple, we assume for now, without loss of generality, that a single condition is formulated for only one OAR. Thus, the additional condition, called a *dose-volume constraint* (DVC) in IMRT, is formulated as follows for an OAR.

Condition 3.1 *For one subset of the system of inequalities (12)–(14), say* B_1*, to be explicit, allow up to* $\alpha_{\max}\%$ *of the total number of inequalities in* B_1 *to have their right-hand side values* $\{u_j \mid j \in B_1\}$ *increase by up to* $\beta_{\max}\%$*.*

The permissions on both the total number of inequalities in B_1 that may be violated and the increase of their right-hand side values use the words "up to". It is desirable to rely on these permissions as little as possible. Therefore, we should not aim initially at the maximal $\alpha_{\max}\%$ and $\beta_{\max}\%$ values. Condition 3.1 is not common in the literature on feasibility problems outside IMRT, see, e.g., the recent special issue on algorithms and computational methods in feasibility and infeasibility, Chinneck [12]. It does not specify a priori which inequalities in B_1 should or could be violated, but instead leaves this as another degree of freedom for the algorithm that solves the problem. For a PTV organ a DVC would be formulated as follows.

Condition 3.2 *For one subset of the system of inequalities (12)–(14), say* L_1*, to be explicit, allow up to* $\alpha_{\max}\%$ *of the total number of inequalities in* L_1 *to have their left-hand side values* $\{l_j \mid j \in L_1\}$ *decrease by up to* $\beta_{\max}\%$*.*

It is possible to handle infeasible systems and, in particular, the combination of (12)–(14) and Conditions 3.1 or 3.2 using mixed-integer programming (MIP) or other methods, as discussed briefly at the end of the next section.

The advantages of the approach described here are that (i) it does not resort to MIP which generally suffers from difficulties in handling high-dimensional problems, (ii) the linear nature of the model is preserved, in contrast with other approaches that resort to nonlinear formulations, (iii) the model permits dynamic incrementation of the violations permitted by Conditions 3.1 or 3.2 so that we can search for a solution with small violations and then, gradually, increase the permissible violations, up to $\alpha_{\max}$ and $\beta_{\max}$, (iv) the intensities vector that is obtained from our algorithm is checked after each incrementation of the permissible violation and additional increase of the violation is employed only if necessary, and (v) the method is very fast and can, therefore, be re-applied in an adaptive planning environment.

4 The successive $\{\alpha, \beta\}$-relaxation method

In order to solve the combination of (12)–(14) and Condition 3.1, as described above, we identify (12)–(14) with the system (1)–(3) and apply the methodology of Section 2 to obtain the following LP problem.

$$\min \sum_{j \in B_1} t_j, \tag{15}$$

$$\text{such that } \ 0 \leq \sum_{v=1}^{m} a_v^j x_v \leq t_j u_j, \quad \text{for all } \ j \in B_1, \tag{16}$$

$$0 \leq \sum_{v=1}^{m} a_v^j x_v \leq u_j, \quad \text{for all } \ j \in B_s, \quad \text{for all } \ s = 2, 3, \ldots, S, \tag{17}$$

$$0 \leq l_j \leq \sum_{v=1}^{m} a_v^j x_v, \quad \text{for all } \ j \in L_q, \quad \text{for all } \ q = 1, 2, \ldots, Q, \tag{18}$$

$$0 \leq x_v, \quad \text{for all } \ v = 1, 2, \ldots, m, \tag{19}$$

$$0 \leq t_j \leq 1 + \beta, \quad \text{for all } \ j \in B_1, \tag{20}$$

$$\sum_{j \in B_1} t_j \leq n_1(1 + \alpha\beta). \tag{21}$$

Each inequality j of the OAR B_1 (the subset of inequalities to which Condition 3.1 applies) is assigned its, real, not necessarily integer, t_j that controls the amount by which the right-hand side of an inequality of B_1, associated with a voxel j, will go above its original prescribed upper bound

u_j. In (20) we confine all t_js (for all inequalities (voxels) $j \in B_1$) to lie between 0 and $1 + \beta$. This β is a user-chosen parameter that can be set by the user to any value $0 \leq \beta \leq \beta_{\max}$. Equations (17), (18) and (19) are the same as in the original problem (12)–(14) because for all remaining OARs and all PTVs no changes are permitted. In (17) the indexing starts from $s = 2$ since B_1 is treated separately. The last constraint (21) is the check inequality (10). The number n_1 is the total number of voxels (inequalities) in the OAR (subset) B_1, thus, $\sum_{j \in B_1} = \sum_{j=1}^{n_1}$. The α is a user-chosen parameter that can be set to any value $0 \leq \alpha \leq \alpha_{\max}$.

The next comments elaborate on this LP and its application to the combination of (12)–(14) and Condition 3.1. As mentioned before, the check constraint (21) does not guarantee that the algorithm will find a relaxed solution, even if such a solution exists. This is because a solution that violates *all* inequalities of B_1 (or just more than $\alpha\%$ of them) might be found which violates all inequalities (or just more than $\alpha\%$ of them) by a very small amount, so small that the total still fulfills (21) although violations occur in more than $\alpha\%$ of the inequalities. Thereby, it will violate the "up to $\alpha_{\max}\%$" permission of Condition 3.1. This is why the formulation presented here does not stop at a relaxed feasibility problem consisting of (16)–(21) alone but we use an LP that strives to minimize $\sum_{j \in B_1} t_j$ over all inequalities of B_1. LP solutions are "extreme points" and, therefore, it is expected that many of the t_js will attain their upper bound value $1 + \beta$ in (21), in which case the solution may come close to satisfying the condition on α. However, even the LP cannot guarantee that a solution that violates *all* of (or just more than $\alpha\%$ of) the inequalities of B_1 by a very small amount will not be obtained. This is because "minimize $\sum_{j \in B_1} t_j$" is still of a global nature over all voxels in B_1.

This situation has not been remedied at this time. However, the experimental work performed here indicates that, when solving the LP, this kind of violation occurs very rarely. Second and more importantly, even if the LP finds such a solution it will not pass the "substitution" step where there is a check of each solution for compliance to the combination of (12)–(14) and Condition 3.1.

The *successive* $\{\alpha, \beta\}$*-relaxation method* works by dynamically incrementing the values of α and β and successively applying the LP. All of $[0, \beta_{\max}] \times [0, \alpha_{\max}]$ is discretely searched bottom-up. There is an outer-loop constructed around the LP-solver (which is the `linprog` function of MATLAB applied to the IMRT situation, see details in Section 5) that allows a

selection of the values of α and β that the LP-solver will run with. The execution is started by putting $\alpha = 0$ and $\beta = 0$, i.e., trying to solve the original system without any violation of the inequalities of the OAR (i.e., no dose-volume constraints at all). At the end of each run of the LP-solver the $(x_1, x_2, \ldots, x_m)$ part of the solution vector of the LP-solver is substituted back into the original system (12)–(14) to verify that it solves the original system without violating Condition 3.1. This guarantees that no inequality in the OAR B_1 overflows the $(1 + \beta_{\max})$ upper bound on the right-hand side, and that no more then $\alpha_{\max}\%$ of the inequalities of B_1 are satisfied with $u_j(1 + \beta_{\max})$, and that all other constraints are fulfilled. The desired final solution is obtained when the current run of the LP-solver satisfies (12)–(14) with Condition 3.1. The outer loop automatically increments the values of α and β by $\Delta\alpha$ and/or $\Delta\beta$, re-runs the LP-solver, and then re-checks the LP solutions repeatedly until an acceptable solution is obtained. If α and β have reached their maximal values without reaching a solution then it is necessary to go back to the radiation oncologist and report infeasibility with Condition 3.1. The radiation oncologist has then to decide whether to raise the values of $\alpha_{\max}$, $\beta_{\max}$, or both, or to modify his prescription in some other way. If a satisfactory solution has been found then it is presented to the "owner" of the original problem (the radiation oncologist) for final approval.

The above is summarized in the following algorithm statement; it is assumed that an LP-solver is available before hand.

Algorithm 4.1 ***The successive*** $\{\alpha, \beta\}$***-relaxation method.***

1. Initialization: *(i) read the data of the original system (12)–(14), i.e., all* a_v^j, u_j, l_j, *for all subsets of inequalities, (ii) obtain from the problem "owner" his prescription for the values of* $\alpha_{\max}$ *and* $\beta_{\max}$, *(iii) choose values for* $\Delta\alpha$ *and* $\Delta\beta$, *(iv) specify the subset* B_1 *to which Condition 3.1 is applied, (v) define* $\alpha_0 = \beta_0 = 0$.

2. The first step: *set* $\alpha \leftarrow \alpha_0$ *and* $\beta \leftarrow \beta_0$ *and apply the LP-solver on the problem (15)–(21).*

2.1 *If the LP-solver fails, i.e., reports "no solution" to the LP, go to step* ***3.***

2.2 *If the LP-solver returns a solution to the LP then take the LP solution vector* x *as a solution to the original problem (12)–(14) and exit the program.*

3. The k***-th Iterative Step:*** *Given the values* α_k *and* β_k *from the end of the previous iterative step, do the following:*

3.1 Solving the LP: *set* $\alpha \leftarrow \alpha_k$ *and* $\beta \leftarrow \beta_k$ *and apply the LP-solver on the problem (15)–(21).*

3.1.1 *If the LP-solver fails, i.e., reports "no solution" to the LP, go to step* ***3.2.***

3.1.2 *If the LP-solver returns a solution to the LP then take the LP solution vector* x *and substitute it into the original system (12)–(14) to check if* x *solves the combination of (12)–(14) and Condition 3.1. If this is the case then save the vector* x *as the desired solution and exit the program, otherwise, delete it and go to step* ***3.2.***

3.2 Incrementing α_k ***and*** β_k***:*** *Define* $\rho = \beta_k + \Delta\beta$.

3.2.2 *If* $\rho \leq \beta_{\max}$ *then define* $\beta_{k+1} = \rho$ *and* $\alpha_{k+1} = \alpha_k$ *and go to step* ***3.***

3.2.3 *If* $\rho > \beta_{\max}$ *then reset* $\beta_k = \beta_0$ *and define* $\sigma = \alpha_k + \Delta\alpha$.

3.2.3.1 *If* $\sigma \leq \alpha_{\max}$ *then define* $\alpha_{k+1} = \sigma$ *and* $\beta_{k+1} = \beta_k$ *and go to step* ***3.***

3.2.3.2 *If* $\sigma > \alpha_{\max}$ *then exit the program and report that no solution has been found.*

It is instructive to briefly mention other existing approaches: MIP and proximity function minimization. The first treatment of dose-volume constraints in IMRT was Bortfeld's conference report [4] followed by Spirou and Chui's journal publication [29], see [5, p. R368]. There are other approaches to handle the infeasibility of the system (12)–(14), see, e.g., Michalski *et al.* [23]. The mixed integer programming (MIP) method and the proximity function minimization approach are of special interest. Both methods are fundamentally different than the one presented here. In MIP one considers an optimization formulation over the constraint set defined by (12)–(14). With some exogenous, user-chosen, linear objective function $f : R^m \rightarrow R$. For IMRT, this MIP formulation can be found in Langer *et al.* [19] and also in Shepard *et al.* [28, p. 737]. Other applications of MIP in this field include Lee, Fox and Crocker, [20] and [21], who used it for radiosurgery treatment planning, Boland, Hamacher and Lenzen [3] who employed a nonlinear MIP formulation to incorporate MLC settings within the treatment planning, and Bednarz *et al.* [2] who compared MIP performance with that of Cimmino's algorithm, see also [31]. Ferris, Meyer and D'Souza [16] give details of the mathematical formulations and algorithmic approaches as well as pointers to supporting literature for MIP-based approaches to problems of radiation therapy. As Ferris, Meyer and D'Souza correctly notice, the main difficulty associated with the MIP approach is that it can become quickly impractical

due to large numbers of voxels in the region of interest (i.e., a large number of inequalities). These difficulties have then to be attacked by approximation techniques. See also Preciado-Walters *et al.* [25].

A completely different approach to infeasibility of a system such as (12)–(14) is to minimize a *proximity function* that measures in some manner the infeasibility of the system. A common choice is a (weighted) proximity function of the form

$$p(\mathbf{x}) := (1/2) \sum_{j=1}^{n} w_j \parallel P_j(\mathbf{x}) - \mathbf{x} \parallel^2 . \tag{22}$$

Here $P_j(\mathbf{x})$ is the orthogonal (nearest point) projection of the point $\mathbf{x}$ onto the j-th half-space determined by any of the inequalities in (12)–(13), where the distance is measured by the Euclidean norm $\parallel \mathbf{x} \parallel^2 = \langle \mathbf{x}, \mathbf{x} \rangle$. The real numbers $\{w_j\}_{j=1}^n$ are user-chosen positive weights such that $\sum_{j=1}^n w_j = 1$. Algorithms for unconstrained minimization of such proximity functions exist within the class of projection methods, see, e.g., Bauschke and Borwein [1], Xiao *et al.* [32]. Such algorithms are either simultaneous projection methods, see, e.g., Byrne and Censor [6] or steered sequential projection methods, see, e.g., Censor, De Pierro and Zaknoon [10]. Cimminio's simultaneous projection method performs proximity function minimization, see, e.g., Xiao *et al.* [31]. The proximity function is always nonnegative and if a sequence of iterates $\{\mathbf{x}^k\}_{k=0}^{\infty}$, generated by some algorithm, converges to a point $\mathbf{x}^*$ at which $p(\mathbf{x}^*) > 0$ then the underlying problem is infeasible and the size of $p(\mathbf{x}^*)$ reflects the degree of infeasibility. An algorithm that will generate such an $\mathbf{x}^*$, or an approximation of it, will simply give the problem-solver a solution that is "best" in the sense that it "minimally" violates the constraints but it does not give the problem-solver a tool to control the overall violation like a method that incorporates Condition 3.1.

5 Experimentation

Next we supply details on the implementation of Algorithm 4.1 and report our experiments. All computations were carried out within MATLAB6.5 [22] using the routine `linprog` as the LP-solver for Algorithm 4.1. The option for equality constraints $A_{eq}x = b_{eq}$ in `linprog` must be disabled by setting $A_{eq} = [\ ]$ and $b_{eq} = [\ ]$ and our LP problem (15)–(21) was appropriately transformed into the format of MATLAB's `linprog`.

5.1 A randomly-generated test problem

For a non-clinical example we use the system (1)–(3) that represents an example with a single OAR and a single PTV. We define the dimensions m (number of beamlets), n_1 (number of inequalities, i.e., voxels, in the OAR) and n_2 (number of inequalities, i.e., voxels, in the PTV (2)). On truly randomly-generated problems one cannot intelligently make any statement about when or why the algorithm reaches or does not reach an acceptable solution. The reason for this is that the creation of a truly randomly-generated matrix A and a truly randomly-generated right-hand sides vector may result in a very "large" infeasibility and it is possible that the $\beta_{\max}$ needed to reach feasibility is huge (thousands or millions %). Therefore, to test the method and algorithm described here we construct test-problems in which there is a priori control on the infeasibility. This is done by first randomly generating a matrix A and a "solution" $\mathbf{x}$, then calculating $A\mathbf{x} = \mathbf{d}$ and, finally, changing a certain number of the components of $\mathbf{d}$ by a certain percentage and taking the resulting vector as the right-hand side for the test problem.

Randomly-generated data for the m-dimensional vectors $\mathbf{a}^j$, for all $j = 1, 2, \ldots, n_1$, and for the vectors $\mathbf{a}^j$, for all $j = n_1+1, n_1+2, \ldots, n_1+n_2$, is used. We specified a range within which the random numbers are generated. All a_v^j are generated to be nonnegative. Then we generate randomly a nonnegative "solution" $\mathbf{x}$. Denoting $A_1\mathbf{x} = \mathbf{d}^1$, where A_1 is as in (1), a (randomly chosen) $\widehat{\alpha}\%$ of the components of $\mathbf{d}^1$ is decreased by a (randomly chosen) $\widehat{\beta}\%$ and the resulting vector is defined as the vector $\mathbf{u}^1 = (u_j^1)_{j=1}^{n_1}$ of upper bounds on the OAR inequalities. Denoting $A_2\mathbf{x} = \mathbf{d}^2$, where A_2 is as in (2), we let $\mathbf{d}^2 = \boldsymbol{\ell}^2$, the vector of lower bounds on the PTV inequalities. The final step is to let the user define the $\alpha_{\max}$ and $\beta_{\max}$ up to which he is willing to "sacrifice" voxels of the OAR. The user also defines parameters $\Delta\alpha$ and $\Delta\beta$ which indicate by how much the program should increment the parameters α and/or β in each iterative step of Algorithm 4.1 until an acceptable solution is reached or until both $\alpha_{\max}$ and $\beta_{\max}$ are reached.

When iterations of Algorithm 4.1 stop, due to the finiteness of the discretized grid of αs and βs, either no solution, or exactly one solution that solves (12)–(14) and Condition 3.1 can be found. If this were a real IMRT case then in the first case it would be up to the original problem "owner" to decide whether to raise $\alpha_{\max}$, or $\beta_{\max}$, or both or modify the prescription in any other way.

This methodology has been experimentally tested by running many ex-

amples of various sizes. All experiments reached acceptable relaxed solutions for the test problems randomly-generated in the controlled manner described above. An example is presented below. The size of this example is $n_1 = 600$, $n_2 = 450$ and $m = 80$. In the test problem generation we used $\widehat{\alpha} = 0.2$ and $\widehat{\beta} = 0.3$ to destroy in a controlled manner the feasibility of the generated problem. In the solution phase we let $\alpha_{\max} = \beta_{\max} = 0.5$ and employ incremental steps of $\Delta\alpha = \Delta\beta = 0.1$. The generated test problem turned out to be indeed infeasible, namely, no solution could be found for it with $\alpha = \beta = 0$. Then the program repeatedly applied the LP-solver while gradually incrementing the αs and βs. The first acceptable solution (i.e., a solution of the LP problem (15)–(21) whose $\mathbf{x}$ part solved (12)–(14) and Condition 3.1) was encountered for the pair $(\alpha, \beta) = (0.3, 0.4)$.

However, out of curiosity, we did not stop there and rather let the algorithm work on until the whole grid of (α, β)-pairs was exhausted. Naturally, with increased values of (α, β) more acceptable solutions were discovered before $(\alpha_{\max}, \beta_{\max})$ was reached. In the table below we list all pairs of αs and βs for which an acceptable solution was found. The third column in the table gives the values of the proximity function (22) that we calculated for each solution.

Solution #	α	β	$p(x)$
1	0.3	0.4	125.92
2	0.3	0.5	122.3
3	0.4	0.3	143.67
4	0.4	0.4	125.92
5	0.4	0.5	122.3
6	0.5	0.3	143.67
7	0.5	0.4	125.92
8	0.5	0.5	122.3

Mixed integer programming (MIP) was applied in studies of similar size in [2]. The time reported there was between 20–120 minutes. The time required for this case here is around 5 minutes for all the iterations. Similar computers were used for both experiments.

5.2 A clinical IMRT example

Next we present a clinical IMRT example. Since most institutions where IMRT is implemented offer treatment of prostate cancer, it is of general interest to pick an example from this particular disease site. Prostate cases are among the most common IMRT treatments, which is one of the reasons that they are chosen. We encounter infeasibility in clinical planning of these cases around 5-10% of the time. The physicians have to reconsider higher probabilities of either bladder or rectum toxicities for these infeasible cases. We use 18 MV beams for all our prostate IMRT planning and treatment. Some parallel opposed beams are also used routinely in our clinic for IMRT planning to obtain optimal target coverage and critical structure sparing. The geometrical center of the prostate PTV was chosen as the center of the IMRT radiation beams. The beam angles selected for the inverse planning system are 0^o, 55^o, 90^o, 145^o, 180^o, 215^o, 270^o and 305^o. The aperture-based inverse planning (ABIP) (see, e.g., Xiao *et al.* [31]) method was applied. The aperture definition was carried out with the commercial CMS FOCUS treatment planning system [15]. For the prostate cases, a 5 mm margin surrounding the *Clinical Target Volume* (CTV) was used to define the PTV. An additional 8 mm margin was added to accommodate the beam penumbra.

The total number of voxels for PTV (target volume), bladder and rectum are 685, 862 and 381, respectively. Using Matlab version 7 on a PC of processor 1.1GHz running Windows XP with 1.24 GB of RAM. The running time is around 5 minutes for all the iterations.

The apertures were selected according to the methodology of [31] and they include: fields that conform to the combined outline of all targets projected back to the radiation point source for all orientations of the treatment unit; fields that conform to the projection of the boost volume for all orientations; field segments that conform to the target but fully shield the critical structures; extra segments to adjust for the dose inhomogeneity that results from shielding critical structures that do not run along the whole length of the target. The number of apertures depended on the geometry and the topology of a particular site as well as on the complexity of the prescription, e.g., the number of boost regions. For treatment plans of prostate cancer, with bladder and rectum as the critical organs that have to be avoided, the total number of segments is usually in the range of 50–60. The dose lower and upper bounds for the prostate case are given in the next table. The lower bound is the minimum dose that has to be deposited in the organ and

the upper bound is the maximum dose that the organ can tolerate. For the target volume, upper bounds on the dose are also imposed to achieve acceptable dose homogeneity. The dose values in the table are in cGy units.

	lower bound	upper bound
PTV	7560	9000
Rectum	0	6500
Bladder	0	6500

Standard dose-volume constraints (DVCs) commonly applicable to this clinical case are specified in the next table. No more than 5% of the target volume is allowed to receive dose that is by at most 5% less than the lower bound goal dose. For critical organs no more then 20% of the OAR is allowed to receive dose that is by at most 20% more than the upper bound permitted dose. The dose calculations are performed with the calculation engine within the system. Dose matrices to voxels due to each of the beams are then extracted. The voxel sizes are $3 \times 3 \times 3$ mm^3. Only those voxels that intercept target volumes are included in the dose matrices extracted.

	upper/lower bound	DVC: vol below	DVC: vol above
PTV	7560	$\leq 5\%$	—
Rectum	6500	—	$\leq 20\%$
Bladder	6500	—	$\leq 20\%$

When applying our method and program to this prostate case we used the rectum as the single OAR for which we exercise a DVC. We set $\alpha_{\max} = 100\%$ and $\beta_{\max} = 100\%$ for this organ with the intent to be able to report on the smallest (α, β) pair which turns the original feasibility problem into a feasible one, even if it will not satisfy the physician's prescription of $\alpha = \beta = 20\%$. The PTV and the other OAR were not permitted any DVC. We then ran the program first with $\alpha = \beta = 0$ and no feasible solution could be found. Then we gradually incremented the values of α and β according to Algorithm

4.1 with $\Delta\alpha = 0.1$ and $\Delta\beta = 0.1$. Feasible solutions were obtained for all pairs (α, β) with $\alpha \geq 0.2$ and $\beta \geq 0.2$. A plot of a cumulative dose-volume frequency distribution, commonly known as a dose-volume histogram (DVH), graphically summarizes the radiation dose distribution within a volume of interest of a patient which would result from a proposed radiation treatment plan. DVHs are used as tools for comparing rival treatment plans for a specific patient by presenting the distribution of dose in the target volume and in volumes of adjacent normal organs or tissues. Figures 1, 2 and 3 show the DVH graphs for PTV, CTV coverage, rectum and bladder doses to percent volumes for different alpha and beta pairs. Of the two parameters, β seems to affect the DVH outcome more. Figures 2 and 3 show the results from (α, β) pairs of (0.3, 0.2) and (0.3, 0.3), respectively. The receding of the rectum DVH line to lower doses is evident with smaller β values, which is quite reasonable considering the fact that this is the relaxation upper dose limit we impose. The α value of 0.3 is higher than expected, especially when studying the DVH results for the rectum where less than 20% of the volume is getting 6500 cGy or less. However, from the model (15)–(21), the combination of (α, β) may be the dominating factor. When more organs are implemented in the study, as we plan to do next, the trade-off performance of various (α, β) pairs will be even more interesting.

6 Conclusion

The introduction of IMRT has in many ways revolutionized Radiation Oncology. Using this treatment modality, it is now possible to generate dose distributions that conform to treatment volumes that partially wrap around a critical structure or even surround it completely. However, a new problem has surfaced in that obtaining a good IMRT plan is highly dependent on the abilities of the clinician and dosimetrist to specify reasonable dose constraints. It is currently necessary for the individual operating the treatment planning system to manually modify input parameters and generate successive plans as a technique of working to an acceptable final result that is superior to other treatment plans obtained during the process. If the dose constraints used as a starting point for the process of generating alternative plans are too relaxed, better dose distributions might go undetected. On the other hand, if very strict constraints are used and convergence to an accept-

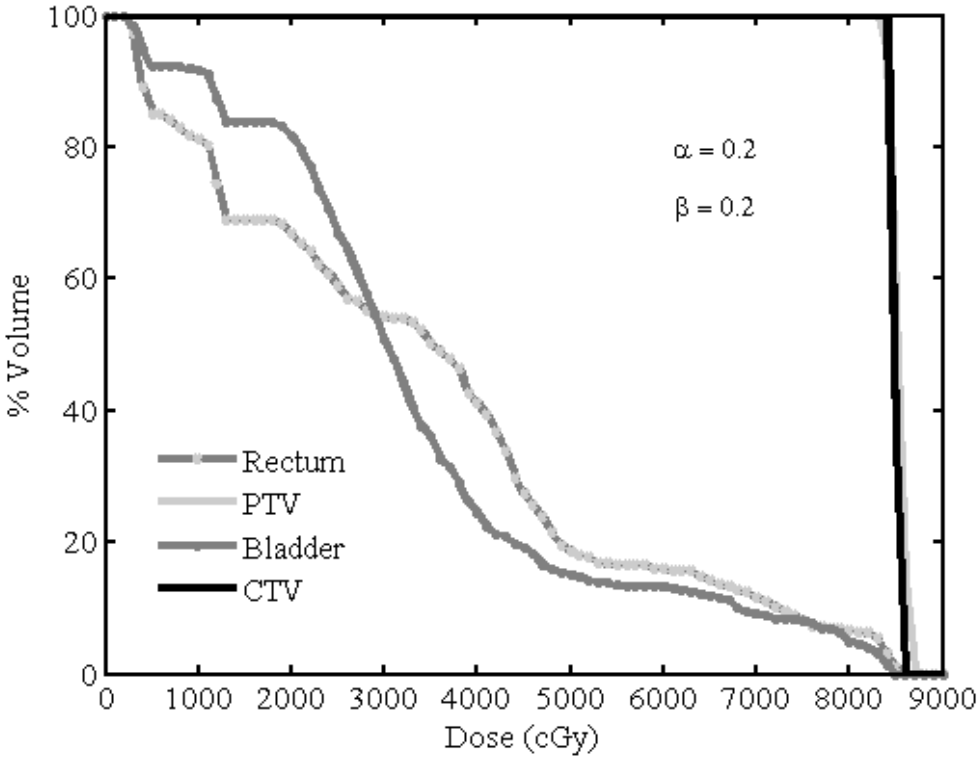

Figure 1: DVH plots for the pair $(\alpha, \beta) = (0.2, 0.2)$.

able plan does not occur, it is hard to know how to relax them in a stepwise fashion or when to stop trying to find a better result.

This paper discusses a new approach that allows the dose constraints to be varied (relaxed) in an organized way within the inverse planning process so that a feasible solution can be found for an otherwise infeasible problem. By automating the step of sorting through different combinations of dose constraints, it is possible to find a treatment plan or series of plans with target and critical structure DVHs that are at least near the original dose constraints specified by the attending physician responsible for a particular patient's care. This approach can be an advantage for busy clinicians that might otherwise be challenged by the prospect of remembering how alternate plans compare when they are generated over periods of time that could involve a number

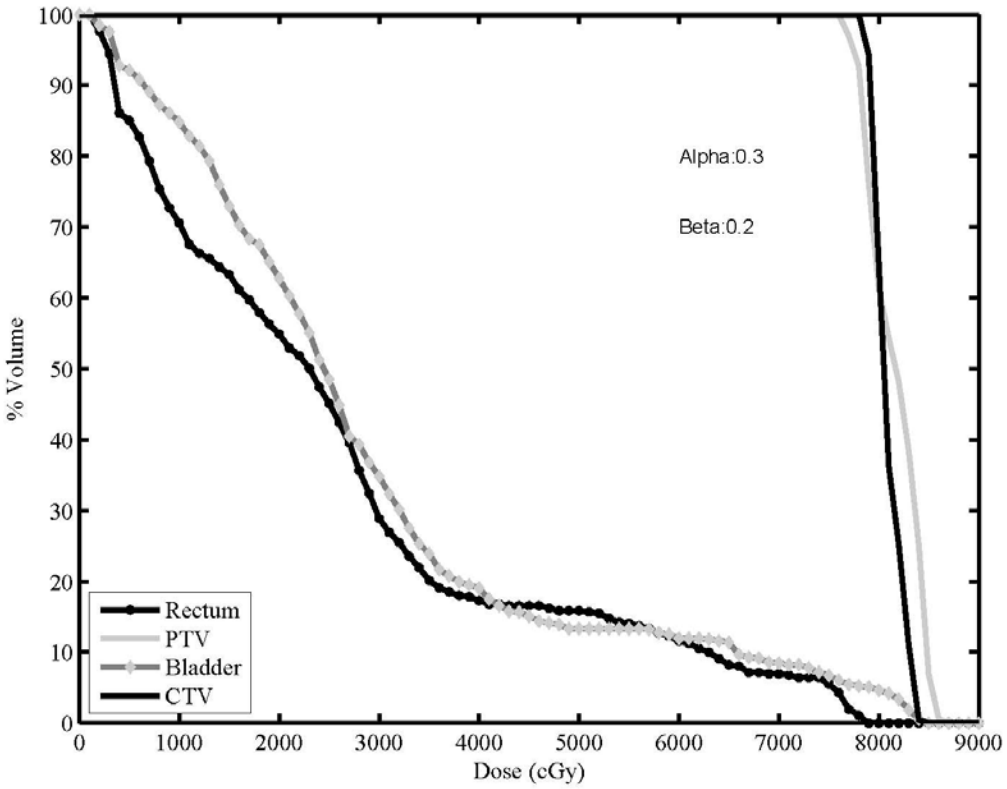

Figure 2: DVH plots for the pair $(\alpha, \beta) = (0.3, 0.2)$.

of days. The approach described can produce a series of plans with relaxed constraints when the original dose limits are not met, and the responsible physician can review and select a result from among those plans presented.

Acknowledgments. We thank two anonymous reviewers for their constructive reports which helped us to improve the presentation. We thank Arik F. Hatwell for his skillful programming and computational work. We thank Thomas Bortfeld for many fruitful discussions and Wei Chen for his comments on an earlier draft. The work of Y. Censor was supported by grant No. 2003275 of the United States-Israel Binational Science Foundation (BSF), by a National Institutes of Health (NIH) grant No. HL70472 and by grant No. 522/04 of the Israel Science Foundation (ISF) at the Center

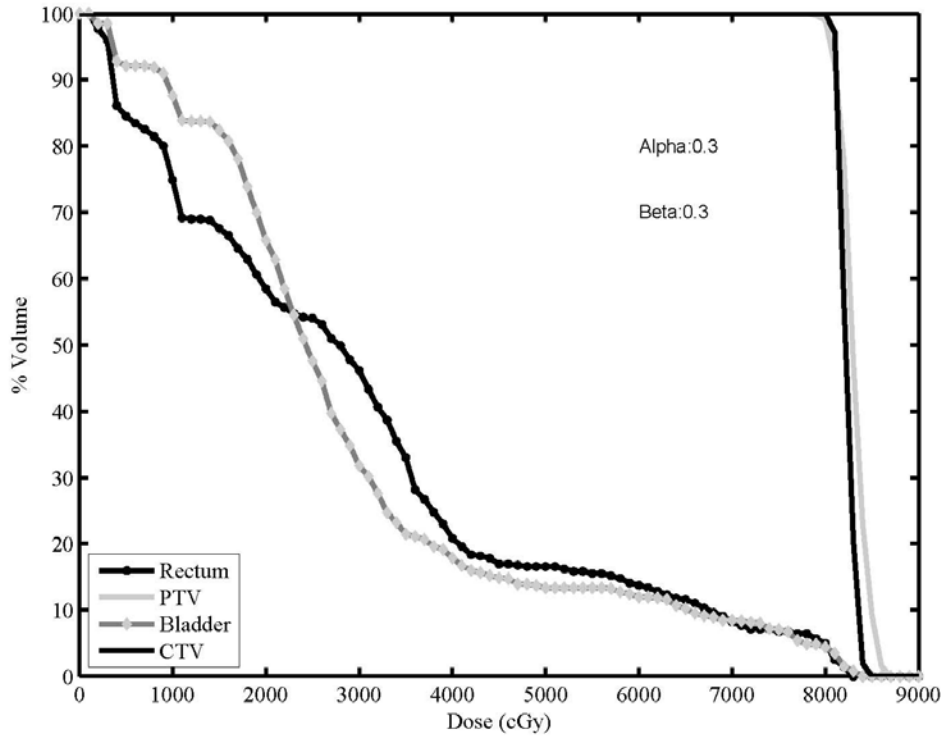

Figure 3: DVH plots for the pair $(\alpha, \beta) = (0.3, 0.3)$.

for Computational Mathematics and Scientific Computation (CCMSC) in the University of Haifa. Preliminary results of this study were presented at "The Interdisciplinary Experts' Workshop on Intensity-Modulated Radiation Therapy (IMRT), Medical Imaging, and Optimization Theory", June 6–11, 2004, that took place at the University of Haifa, Haifa, Israel. We thank Eva Lee, Ron Rardin and Mark Langer for their useful comments at that workshop.

References

[1] H.H. Bauschke and J.M. Borwein, On projection algorithms for solving convex feasibility problems, *SIAM Review* **38**, pp. 367–426, (1996).

[2] G. Bednarz, D. Michalski, C. Houser, M.S. Huq, Y. Xiao, P.R. Anne and J.M. Galvin, The use of mixed-integer programming for inverse treatment planning with pre-defined field segments, *Physics in Medicine and Biology*, Vol. **47**, pp. 1–11, (2002).

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

[4] T. Bortfeld, J. Stein and K. Preiser, Clinically relevant intensity modulation optimization using physical criteria, in: D.D. Leavitt and G. Starkschall (Editors), *The XIIth International Conference on the Use of Computers in Radiation Therapy (ICCR),* Salt Lake City, Utah, USA, May 27–30, 1997, pp. 1–4.

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

[6] C. Byrne and Y. Censor, Proximity function minimization using multiple Bregman projections, with applications to split feasibility and Kullback-Leibler distance minimization, *Annals of Operations Research*, Vol. **105**, pp. 77–98, (2001).

[7] Y. Censor, Mathematical optimization for the inverse problem of intensity-modulated radiation therapy, in: J.R. Palta and T.R. Mackie (Editors), *Intensity-Modulated Radiation Therapy: The State of The Art*, American Association of Physicists in Medicine, Medical Physics Monograph No. 29, Medical Physics Publishing, Madison, Wisconsin, USA, 2003, pp. 25–49.

[8] Y. Censor, M.D. Altschuler and W.D. Powlis, A computational solution of the inverse problem in radiation therapy treatment planning, *Applied Mathematics and Computation*, Vol. **25**, pp. 57–87, (1988).

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

[10] Y. Censor, A.R. De Pierro and M. Zaknoon, Steered sequential projections for the inconsistent convex feasibility problem, *Nonlinear Analysis:*

Theory, Methods & Applications (Series A: Theory and Methods), Vol. **59**, pp. 385–405, (2004).

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

[12] J.W. Chinneck (Guest Editor), Special Issue on "Algorithms and Computational Methods in Feasibility and Infeasibility", *Computers and Operations Research*, Volume **35**, in press, (2008).

[13] J.W. Chinneck, *Feasibility and Infeasibility in Optimization: Algorithms and Computational Methods*, International Series in Operations Research and Management Sciences, Vol. **118**, Springer-Verlag, 2007.

[14] P.S. Cho and R.J. Marks II, Hardware-sensitive optimization for intensity modulated radiotherapy, *Physics in Medicine and Biology*, Vol. **45**, pp. 429–440, (2000).

[15] Computerized Medical Systems, Inc., *Radiation Treatment Planning Software: FOCUS*, St. Louis, MO, USA 2001.

[16] M.C. Ferris, R.R. Meyer, and W. D'Souza, Radiation treatment planning: Mixed integer programming formulations and approaches, in: G. Appa, L. Pitsoulis and H.P. Williams (Editors), *Handbook on Modelling for Discrete Optimization,* Springer-Verlag, New York, NY, USA, 2006, pp. 317–340.

[17] J.M. Galvin, G. Ezzel, A. Eisbrauch, C. Yu, B. Butler, Y. Xiao, I. Rosen, J. Rosenman. M. Sharpe, L. Xing, P. Xia, T. Lomax, D.A. Low and J. Palta, Implementing IMRT in clinical practice: A joint document of the American Society for Therapeutic Radiology and Oncology and the American Association of Physicists in Medicine, *International Journal Radiation Oncology, Biology, Physics*, Vol. **58**, pp. 1616–1634, (2004).

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

[19] M. Langer, S. Morrill, R. Brown, O. Lee, and R. Lane, A comparison of mixed integer programming and fast simulated annealing for optimizing

beam weights in radiation therapy, *Medical Physics,* Vol. **23**, pp. 957–964, (1996).

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

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

[22] MATLAB6.5, The MathWorks, Inc., http://www.mathworks.com/.

[23] D. Michalski, Y. Xiao, Y. Censor and J.M. Galvin, The dose-volume constraint satisfaction problem for inverse treatment planning with field segments, *Physics in Medicine and Biology*, Vol. **49**, pp. 601–616, (2004).

[24] J.R. Palta and T.R. Mackie (Editors), *Intensity-Modulated Radiation Therapy: The State of The Art,* American Association of Physicists in Medicine, Medical Physics Monograph No. 29, Medical Physics Publishing, Madison, WI, USA, 2003.

[25] F. Preciado-Walters, R. Rardin, M. Langer and V. Thai, A coupled column generation, mixed integer approach to optimal planning intensity modulated radiation therapy for cancer, *Mathematical Programming, Series B*, Vol. **101**, pp. 319–338, (2004).

[26] R. Reemtsen and M. Alber, Continuous optimization of beamlet intensities for photon and proton radiotherapy, Technical Report, March 2006, available at: http://www.math.tu-cottbus.de/INSTITUT/lsing1/publications/reemtsen_e.html. To appear in: P.M. Pardalos and E. Romeijn (Editors), *Handbook of Optimization in Medicine*, Springer-Verlag.

[27] H.E. Romeijn, R.K. Ahuja, J.F. Dempsey, A. Kumar and J.G. Li, A novel linear programming approach to fluence map optimization for intensity modulated radiation therapy treatment planning, *Physics in Medicine and Biology*, Vol. **48**, pp. 3521–3542, (2003).

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

[29] S.V. Spirou and C.S. Chui, A gradient inverse planning algorithm with dose-volume constraints, *Medical Physics,* Vol. **25**, pp. 321–333, (1998).

[30] S. Webb, *Contemporary IMRT: Developing Physics and Clinical Implementation*, Institute of Physics (IOP) Publishing, Bristol, UK, 2005.

[31] Y. Xiao, Y. Censor, D. Michalski and J.M. Galvin, The least-intensity feasible solution for aperture-based inverse planning in radiation therapy, *Annals of Operations Research*, Vol. **119**, pp. 183–203, (2003).

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