A Successive Linear Programming Approach to IMRT Optimization Problem

Michael Merritt, Yin Zhang
Department of Computational and Applied Mathematics
Rice University

Helen Liu, Radhe Mohan
Department Of Radiation Physics
University of Texas M.D. Anderson Cancer Center

December 19, 2002

Abstract

We propose to solve the IMRT optimization problem through a successive linear programming approach. Taking advantage of the sensitivity information in linear programming and the re-optimization ability of simplex methods, the proposed approach provides an affordable methodology to efficiently solve problems with dose-volume constraints. Preliminary computational results indicate that, compared to the standard weighted least squares approach, the new approach leads to higher tumor dosage escalation and better conformation.

1 Introduction

For many years, radiation has been used as a treatment for cancer. Bombarding a malignant tumor with high-energy particles can destroy the cancerous cells or at least slow down their growth. Since tumors grow in the presence of healthy tissue and even near critical organs, it is usually impossible to irradiate the tumor without allowing some damage to the nearby critical organs. In fact, complications can occur when neighboring critical organs receive too much of this collateral radiation. Moderate damage to critical organs may be acceptable, however, if the effect can be accurately predicted, since cancerous cells do not have the ability to repair themselves as efficiently as normal cells. Therefore, an important research activity in radiotherapy is to develop methods of escalating dosage delivered to the tumor while carefully controlling the dosage deposited in neighboring critical organs and healthy tissues.

Modern external radiotherapy seeks to conform the "shape" of the delivered radiation to specific three-dimensional structures (tumors, critical organs, etc.) within the patient using a linear accelerator that emits beams of high-energy photons from predetermined angles around the patient. Conventional 3D conformal radiotherapy (3DCRT) uses beams of uniform intensity. Parameters such as beam intensity and aperture are carefully adjusted to give the best conformation and least chance of complications. In recent years, intensity-modulated radiotherapy (IMRT) has been gaining support as a more advantageous method of delivering radiation treatments. The IMRT approach breaks each beam up into hundreds of tiny constituent beamlets, or pencil-beams, each of which is assigned its own intensity. The resulting IMRT beam has better resolution than a uniform 3DCRT beam for conforming to the 3D shape of the patient's tumor and avoiding the critical organs. For instance, in IMRT, intensities for certain beamlets which are collinear with critical organs might be set to low or zero values, while other beamlets collinear with the tumor can be set to high values. In standard 3DCRT, however, the entire beam must be one uniform intensity, so maintaining the same safety tolerances for the critical organ would require delivering far less dosage to the tumor. Although IMRT has many advantages over the conventional 3DCRT approach, it also complicates treatment planning. Whereas 3DCRT requires the assignment of only a few intensities (one for each beam), IMRT requires thousands (one for each beamlet). The considerable task of planning the IMRT beamlet intensities such that a desirable treatment is achieved necessitates extensive use of optimization formulations and techniques.

The standard optimization formulation for the IMRT problem minimizes the average least squares deviation from the prescribed tumor dose, penalizing when critical organ dosages exceed some tolerance dose. Also, weight parameters are used in each term of the objective function to control the relative priorities of matching tumor dosage to the prescription versus protecting the critical organs. Although the weighted least squares approach is predominant in IMRT literature and practice, its solutions can be highly sensitive to the choice of weight parameters. In addition, these weight parameters do not directly represent intuitively meaningful quantities that can guide practitioners to their selections. Thus, current IMRT planning optimization requires significant trial and error, at the cost of complete re-optimization each time the weights are changed. Moreover, because this standard approach measures deviation from the prescribed and tolerance doses in the least squares sense, there is no explicit control over whether computed dosages are above or below these values. This is clearly undesirable as we would like to have strict control over this property for at least the dosage to the critical organs.

Indeed, many recent IMRT planning schemes seek to improve solutions, while maintaining precise control of the critical organ dosage, by making use of so-called "dose-volume" constraints. Once IMRT optimization has been performed, it is often discovered that the tolerance dose for the critical organ is too conservative to allow the delivery of a sufficiently large dosage to the tumor. Rather than increasing the tolerance dose for the critical organ and thereby risking damage to the entire organ, the physician are usually willing to accept damaging some healthy cells in a predetermined volume of the critical organ because the remaining cells could still perform the organ's normal function, and the critical organ can eventually repair such a limited damage. A typical dose-volume constraint given by a physician might specify that no more than 25% of the critical organ volume may receive dosage exceeding 10Gy. Dose-volume constrains are very intuitive for physicians to specify and they lead to a great deal of flexibility in tumor dosage escalation. How to impose such constraints, however, has been the subject of much research. The weighted least squares approach, for instance, attempts to impose dose-volume constraints through penalizing the objective function in certain ways [10].

In this paper, we introduce a successive linear programming approach to the IMRT optimization problem as an alternative to the weighted least squares formulation. The application of linear programming to radiotherapy is not new (for example, see [7, 5] and the references therein for some early works, and see [9, 2, 3] for some more recent works).

In our formulation, we maximize the minimum dosage delivered to the tumors and control maximum dosage delivered to the tumors and critical organs explicitly. Furthermore, we impose dose-volume control by means of successive relaxation of upper bounds on critical organ dosages until the specified maximum allowable volume has been relaxed. Preliminary numerical results are presented, using two dimensional phantom data and some standard test configurations, in which a standard weighted least squares formulation is compared to the proposed successive linear programming approach.

2 Preliminaries

Most formulations of the IMRT optimization problem are such that their solution requires using a relatively general nonlinear programming algorithm such as steepest-descent, Newton-like methods, or SQP [6]. In our case, however, the special structure in our formulation allows us to make efficient use of certain theoretical properties of the linear program. We now present some of the theory of linear programs necessary to understand our approach.

2.1 Duality Theory of Linear Programming

Given a linear program (LP) in standard form,

$$\begin{aligned} \max \quad & c^T x && \text{(1a)} \\ \text{s.t.} \quad & Ax \leq b && \text{(1b)} \\ & x \geq 0, && \text{(1c)} \end{aligned}$$

where $c \in \mathbb{R}^n, b \in \mathbb{R}^m$, and $A \in \mathbb{R}^{m\times n}$, and all inequalities are understood componentwise, then the associated dual problem is

$$\begin{aligned} \min \quad & b^T y && \text{(2a)} \\ \text{s.t.} \quad & A^T y \geq c && \text{(2b)} \\ & y \geq 0. && \text{(2c)} \end{aligned}$$

We call the original problem (1) the primal problem. Note the symmetry between the two problems. From primal to dual, we replace max with min, $\leq$ with $\geq$, A with A^T, and swap the roles of the primal objective coefficients c and the right-hand side vector b. Also, we have a new set of (dual) variables $y \in \mathbb{R}^m$ in place of the primal variables $x \in \mathbb{R}^n$. We note that taking the dual of the dual problem returns us to the original primal problem. Based simply on the inequality constraints in (1) and (2), one can easily show that weak duality holds for any feasible x and y, that is $c^T x \leq b^T y$, so the dual objective value always provides an upper bound on the primal objective value. In fact, the principle of strong duality states that if the primal problem (1) has an optimal solution $x^* \in \mathbb{R}^n$, then the dual problem (2) has an optimal solution $y^* \in \mathbb{R}^m$ such that $c^T x^* = b^T y^*$. This means that there is no gap at optimality between the primal and dual optimal objective values. Thus, an algorithm to solve an LP of the form (1) need only check that $x^* \in \mathbb{R}^n$ and $y^* \in \mathbb{R}^m$ satisfy (1b), (2b), and $c^T x^* = b^T y^*$ in order to conclude that x^* is an optimal solution of (1). Also, note that unlike nonlinear programs, all solutions to linear programs are *global* minimizers. Indeed, the simplex algorithm which we use to solve our linear program relies heavily on these concepts of duality to solve LP's efficiently. More on these concepts of duality and the simplex algorithm may be found in [1].

2.2 Sensitivity Analysis

Now, we examine the key technique we use in our method for improving on a solution. The LP in (1) is designed to force the objective function to take on the largest value possible without violating the constraints. But, if the objective represents some physical quantity we wish to maximize, then once the problem has been solved, we face the possibility that the objective is not "good enough" in whatever context makes sense. One obvious solution is to relax the constraints placed on the variable space in the hope that we might obtain a larger objective value with the new constraint set. Moreover, it would be nice if we had some guidance on which constraints to relax such that the changes we make will increase the objective function the most.

As it turns out, the m dual variables y in the problem (2) each correspond to exactly one of the m constraints (1b). In fact, the optimal dual variable y_i^* represents how sensitive the optimal objective value $c^T x^*$ is to changes in the constraint right-hand side b_i. That is, relaxing right-hand sides with corresponding sensitivities (optimal dual variables) that are high will yield greater increases in the optimal objective value than relaxing those with lower sensitivities. So, first we obtain a solution to the original problem and determine the largest dual variables $R = \{j \in S : y_j^* > \lambda\}$, where $\lambda > 0$ is some threshold value we choose and S is the set of all the constraints. Next, we relax the corresponding right-hand sides $b_i, i \in R$, by increasing the right-hand sides to $\hat{b}_i = b_i + \epsilon_i$ for $i \in R, \epsilon_i > 0$, while keeping $\hat{b}_i = b_i$ for $i \notin R$. Finally, we solve the primal problem (and corresponding dual) again with a new right-hand side vector $\hat{b}$ to obtain a larger objective value and repeat the process if the desired value has not been achieved.

3 Formulation

For simplicity, we only consider the case of one tumor, one critical organ, and one dose-volume constraint. In a straightforward manner, however, our approach can be extended to address multiple tumors, critical organs, and dose-volume constraints. Consequently, we will give conjecture on generalizing our approach in Section 7. Also, we will assume everywhere in this paper that beam angles are already fixed when IMRT planning ensues since optimizing with respect to these angles can be viewed as a distinct (discrete) optimization problem.

3.1 Formulation with Exact Dose-Volume Constraint

We assume that within the patient, a "region of interest" surrounding the tumor has been defined and discretized into m three-dimensional "voxels," some of which are located within the tumor, some within the critical organ, and some within neither. Let the voxels in the region of interest be indexed by the set $\{1, \ldots, m\}$. Now let $\mathcal{T}$ and $\mathcal{C}$ be subsets of this set which correspond to the patient's tumor and critical organ, respectively, and let $|\mathcal{C}|$ be the cardinality of the set $\mathcal{C}$. If $H \in \mathbb{R}^{m \times n}$ is the given (nonnegative) dose matrix and $w \in \mathbb{R}^n$ the vector of unknown beamlet intensities, or beamlet weights, then we have the following optimization problem:

$$\begin{aligned} \max \quad & \tau && \text{(3a)} \\ \text{s.t.} \quad & d = Hw, \qquad w \geq 0, && \text{(3b)} \\ & \tau \leq d_i \leq u_t, \quad i \in \mathcal{T}, && \text{(3c)} \\ & |\{j \in \mathcal{C} : d_j > u_c\}|/|\mathcal{C}| \leq \rho, && \text{(3d)} \end{aligned}$$

where $d_i, i = 1, \ldots, m$, is the dosage at the i-th voxel and u_t and u_c are upper bounds on dosage delivered to the tumor and critical organ, respectively. The beamlet weights $w_j, j = 1, \ldots, n$, and the tumor dose lower bound τ are actually the optimization variables and the dose matrix entries H_{ij} model physically how the radiation from the j-th beamlet contributes to the resultant dosage in the i-th voxel. Also, $\rho \in [0, 1]$ in the dose-volume constraint (3d) is the maximum ratio of voxels in the critical organ that are allowed to receive dosage greater than u_c. In fact, if $\rho = 0$, then we have the case that the entire critical organ is constrained to have dosage of at most u_c, i.e. no dose-volume constraint.

Clearly, the formulation (3) is a natural way to express the problem we want to solve and models the dose-volume constraint exactly. Notice, however, that although the dose-volume constraint (3d) allows a certain number of voxels within the critical organ to violate $d_i \le u_c, i \in \mathcal{C}$, it makes no assumption about *which* points may do so. Implicitly, every critical organ voxel in this formulation has the binary property that it either has no upper bound constraint on its dosage or its dosage is constrained such that $d_i \le u_c$. Furthermore, choosing different combinations of points to have $d_i \le u_c$ could effectively change the feasible set and thus possibly increase the optimal objective value. This implies the optimization problem (3) is in fact a mixed-integer program (MIP), i.e. some of the variables are continuous (beamlet weights, min tumor dosage) and others are discrete. For instance, we could impose the dose-volume constraint (3d) with the condition $(\sum_{i \in \mathcal{C}} \phi_i)/|\mathcal{C}| \le \rho$ where the discrete variables $\phi_i = 1$ if d_i could be greater than u_c and $\phi_i = 0$ if we explicitly constrain $d_i \le u_c$. In general, however, mixed-integer programs are computationally expensive to solve.

3.2 Successive Linear Programming Formulation

We would like to find a more computationally tractable way of imposing dose-volume constraints. Note that without the "exact dose-volume constraint" (3d), the problem would simply be an LP, which we know how to solve efficiently. If we only knew *a priori* which critical organ voxels are constrained to have $d_i \le u_c$ at optimality of the MIP, say $i \in S$, we could simply solve (3) as an LP with (3d) replaced by $d_i \le u_c, i \in S$. Although this is not possible, our motivation behind solving successive LP's (rather than one MIP) is the conjecture that iteratively selecting the combination of points that produce the largest increase in the objective locally results in a final combination that is close to optimal, i.e. yields a final LP solution that is close to the MIP solution.

We now make explicit the form of the LP we will solve successively and the scheme by which the points are chosen. We define the "k-th LP" to be

$$
\begin{aligned}
\max \quad & \tau && && \text{(4a)}\\
\text{s.t.} \quad & d = Hw, && w \ge 0, && \text{(4b)}\\
& \tau \le d_i \le u_t, && i \in \mathcal{T}, && \text{(4c)}\\
& d_i \le u_r, && i \in R_k, && \text{(4d)}\\
& d_i \le u_c, && i \in \mathcal{C} \backslash R_k, && \text{(4e)}
\end{aligned}
$$

for $k \ge 1$ where $R_k \subseteq \mathcal{C}$ is the set of indices corresponding to right-hand sides that have been relaxed after solving LP's $1, 2, \ldots, (k-1)$ and $u_r > u_c$ is some upper bound on those constraints (so that dosage at even the relaxed points does not get "too high"). From our discussion on sensitivity analysis, we define

$$R_k = \{j \in \mathcal{C} : y^*_{j,k-1} > \lambda\} \cup R_{k-1}$$

for $k \geq 2$ and $R_1 = \emptyset$, where $y^*_{j,k-1}$ is the j-th sensitivity in the $(k-1)$-th LP and $\lambda > 0$ is some constant parameter.

Thus, our general approach is the following. First, we solve an LP with constraints on all the critical organ points $i \in \mathcal{C}$ $(k = 1)$. Next, we utilize the sensitivity information in the optimal dual variables $y^*_{i,k-1}$ to choose which constraints to relax in the next iteration k. Then, we relax those corresponding right-hand sides and re-optimize. We repeat the last two steps until we have attained the prescribed minimum tumor dose or we determine that the next LP will violate the relation $|\{j \in \mathcal{C} : d_j > u_c\}|/|\mathcal{C}| \leq \rho$. Recalling our discussion of sensitivity analysis in Section 2, we observe that solving the k-th LP by relaxing the critical organ constraints with the largest sensitivities in the $(k-1)$-th LP yields an increase in the k-th optimal objective value, relative to the $(k-1)$-th.

Moreover, at each step, we "greedily" choose the constraints whose relaxation will produce the largest increase in τ. Consequently, since we simply choose the best relaxation locally at each iteration, it is not yet known how close the set of points iteratively chosen by this method is to the globally optimal combination of points determined by the solution of the MIP (3). Preliminary experiments suggest, however, that our method yields solutions that are indeed close to the exact MIP solutions.

4 Eliminating Hot Spots

We have given a formulation which allows us to explicitly control the dosage delivered to the tumor and critical organ, but we would now like to consider the remaining "normal tissue" surrounding these structures. Thus far, applying our method may result in a treatment plan which deposits high dosages in the tumor and low dosages in the critical organ. However, since we have placed no constraints on the normal tissue dosages in the region of interest, it is still possible (and in fact we have observed) that this strategy can result in the creation of "hot spots" in the normal tissue, or small regions that receive intense radiation due to the additive effect of overlapping beamlets. We would like to minimize this effect as much as is reasonable because high dosages even in the normal tissue can cause complications for the patient.

The obvious approach to dealing with this is to simply constrain every voxel within the region of interest: tumor, critical organ, and normal tissue. That is, the constraint matrix would have dimensions $(2|\mathcal{T}| + |\mathcal{C}| + |\mathcal{N}|) \times$ (# of beamlets), where $\mathcal{N} = \{1, \ldots, m\} \backslash \{\mathcal{C} \cup \mathcal{T}\}$ denotes the index set corresponding to what we will term the "normal tissue" voxels. So, for example, if we constrain the dosage of every 0.4 cm $\times$ 0.4 cm $\times$ 0.4 cm voxel in a 12 cm $\times$ 12 cm $\times$ 12 cm region of interest, then we would have at least 27,000 constraints in our optimization problem. Even if we have enough memory to store a 27,000 × (# beamlets) matrix, constraining every voxel in our problem slows the optimization to a crawl on most computers. However, we know that while $|\mathcal{N}|$ is typically very large relative to $|\mathcal{C}|$ and $|\mathcal{T}|$, only a few of those normal dosages are sufficiently "hot." So, the idea is to limit the size of our constraint matrix, thus speeding up computation and reducing memory requirements, by only including constraints for normal tissue voxels corresponding to hot spots.

Thus, we implement the following procedure to "cool" the hot spots that occur in solving the k-th LP. After solving the k-th LP, we compute the resulting dosages in the normal tissue voxels. Then, we determine which of the normal tissue voxels have dosages above some threshold u_h that is considered "hot" and add corresponding constraints to prevent the dosage at those points from exceeding u_h. Finally, we re-optimize (no relaxation) using this new constraint set and repeat the process until we have sufficiently cooled the normal tissue, i.e. all normal tissue dosages are at most u_h. Once we have finished cooling, we relax to get the $(k+1)$-th LP and begin again.

5 General Algorithm

Now that we have stated our main successive linear programming approach in Section 3 and a strategy for eliminating "hot spots" in Section 4, we assemble both components into a general algorithm. Of course, this pseudocode reflects only the significant tasks in the process, neglecting details of implementation.

5.1 Pseudocode

Let $\mathcal{C}, \mathcal{T}$, and $\mathcal{N}$ be index sets corresponding to the critical organ, tumor, and normal tissue voxels, respectively, such that $\mathcal{C} \cup \mathcal{T} \cup \mathcal{N} = \{1, \ldots, m\}$. We denote the prescribed tumor dose by d^T, the total number of voxels in the region of interest by m, and the total number of beamlets by n. Given these index sets, a nonnegative dose matrix $H \in \mathbb{R}^{m \times n}$, and suitable choices for the input parameters $d^T, u_t, u_c, \rho, u_h, \lambda, u_r$, we have the following algorithm for the solution of the IMRT optimization problem:

Algorithm:

1. **Set Up Initial LP in CPLEX**:

$$\begin{array}{rl} \max & \tau \\ \text{s.t.} & (Hw)_i \geq \tau,\ i \in \mathcal{T} \\ & (Hw)_i \leq u_t, i \in \mathcal{T} \\ & (Hw)_i \leq u_c, i \in \mathcal{C} \\ & \quad w \ \ \geq 0 \end{array}$$

2. **Iterate**:

$k = 1$

while $(\tau < d^T)$ **and** ($|\{j \in \mathcal{C} : d_j > u_c\}|/|\mathcal{C}| \leq \rho$)

  Call CPLEX Solver on Current LP

  Compute Dose $d = Hw$

  $\mathcal{H} := \{i \in \mathcal{N} : d_i > u_h\}$

  while $|\mathcal{H}| > 0$

    Add Constraints to Current LP: $d_i \leq u_h, i \in \mathcal{H}$

    Call CPLEX Solver on Current LP

    Compute Dose $d = Hw$

    $\mathcal{H} := \{i \in \mathcal{N} : d_i > u_h\}$

  end while

  $R_k := \{i \in \mathcal{C} : y^*_{i,k-1} > \lambda\}$

  Relax Current LP: $(rhs)_i \leftarrow u_r, i \in R_k$

  $k \leftarrow k + 1$

end while

5.2 CPLEX

The CPLEX software environment provides us with an efficient and robust implementation of the simplex algorithm for solving linear programs [1]. A program may interface with the CPLEX environment through a callable library of functions which we use to not only solve our linear programs, but also to facilitate data structure management. For instance, our algorithm sets up an initial LP in the CPLEX environment with input data and then simply updates that LP as it proceeds with CPLEX library functions that add constraints to the current set or modify the current right-hand side vector. Indeed, such functions might be very tedious for the programmer to implement and difficult for observers to understand without the level of abstraction that the CPLEX library provides.

In addition, CPLEX has the advantage that it is particularly efficient at re-optimizing an LP when small changes are made to the LP data (e.g. constraint right-hand sides). That is, if we compute the solution to an LP, relax some constraints, and optimize the relaxed LP, CPLEX can use the solution obtained previously to quickly compute the new solution, rather than optimizing the relaxed LP "from scratch." This makes CPLEX ideal as our solver since we are solving so many (slightly perturbed) linear programs.

6 Numerical Experiments

In this section, we illustrate that our new approach deserves further attention for its increased ability to conform radiation to the target volume and maintain precise dose-volume control, relative to the most widely-used formulation of the IMRT optimization problem, the weighted least squares approach.

6.1 Weighted Least Squares Formulation

Although the IMRT optimization problem can be formulated in many ways, the most popular formulation both in literature and practice is what we will call "weighted least squares." Essentially, it minimizes the average least squares deviation of dosages in the tumor voxels from the prescribed tumor dose. In addition, critical organ dosages may be encouraged to stay below a tolerance dose u_c by penalizing the objective by the least squares deviation from u_c, when those dosages exceed u_c. This "dose-based objective" [10] formulation would have a mathematical formulation similar to

$$\begin{aligned} \min \quad & \theta_t \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} (d_i - d^T)^2 + \theta_c \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} P(d_i - u_c)(d_i - u_c)^2 && \text{(5a)} \\ \text{s.t.} \quad & d = Hw && \text{(5b)} \\ & w \geq 0, && \text{(5c)} \end{aligned}$$

where the penalty (step) function P is designed to turn on the second term whenever the dosage d_i exceeds the critical organ tolerance dose u_c. That is,

$$P(d_i - u_c) = \begin{cases} 1, & \text{if } d_i > u_c \\ 0, & \text{otherwise.} \end{cases}$$

The penalty, or weight, parameters θ_t, θ_c are used to control the relative priorities of minimizing each term in the objective. For instance, a large value of θ_t (relative to θ_c) encourages tumor dosage to match the prescribed tumor dose d^T, but may neglect the suppression of critical organ dosage.

To compare results in this paper, we want a model that incorporates dose-volume constraints as well. Imposing such constraints in this least squares framework has, however, been the subject of much research. We choose to compare our results to a particular formulation whose scheme for imposing the dose-volume constraints can be found in Wu and Mohan [10]:

$$\begin{aligned} \min \quad & \theta_t \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} (d_i - d^T)^2 + \theta_c \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} P(d_i - u_c) P(d^R - d_i)(d_i - u_c)^2 & \text{(6a)} \\ \text{s.t.} \quad & d = Hw & \text{(6b)} \\ & w \geq 0. & \text{(6c)} \end{aligned}$$

Note the only difference between this dose-volume-based objective and the dose-based objective is the addition of the function $P(d^R - d_i)$ in the critical organ term which simply causes critical organ dosages between u_c and d^R to be penalized. We will now state the explanation as stated in [10] for how to choose d^R in order to satisfy the dose-volume constraint: no more than ρ volume of the critical organ may receive more than u_c dosage. Assume we solve the problem (5) without the second term, i.e. try to match the tumor dosage as close as possible to the prescription without considering the critical organ, obtaining a treatment plan with some dosage distribution. In order to impose the dose-volume constraint in the objective function, we seek another dosage d^R for which the volume of the critical organ with dosages greater than d^R is at most ρ in the current dosage distribution [10]. Then, if we solve the problem (6) with this value of d^R, we should obtain a treatment plan in which the dose-volume constraint is *nearly* satisfied. This scheme for implementing dose-volume constraints is clearly only a heuristic and experimentally we have found that many times it will not even enforce feasibility of the dose-volume constraint.

In fact, weighted least squares suffers from the drawback that penalizing dosages that cross some boundary dose does not necessarily prohibit them from crossing that boundary. Specifically, since this approach measures deviation from the prescribed and tolerance doses in the least squares sense, there is no explicit control in the formulation over whether computed dosages are above or below these limits. This is especially dangerous in the case of dose-volume constraints, which already involve some risk to the patient even if they are strictly satisfied. The other major disadvantage of this approach is that it is extremely sensitive to the choice of the weight parameters θ_t, θ_c and they must be chosen ahead of time. Furthermore, they represent no physical or intuitively meaningful quantities that could guide us in their selection. Even experienced practitioners in IMRT typically resort to trial and error to find parameters that produce acceptable treatment plans. Indeed, each time these weight parameters are varied by the slightest amount, complete re-optimization must be performed. This can be substantial as optimization for one set of parameters can take up to 2 hours in many clinical cases.

For our experiments, we solved the weighted least squares problem (6) using the scheme for selecting d^R from [10] and the method of steepest-descent (gradient method) with some judicious algorithmic choices for enforcing $w \geq 0$ that are similar to the gradient projection method [6]. In the interest of comparing our results to "good" weighted least squares plans, we developed the following scheme for approximating good weight parameters. We first reduce the number of parameters from two to one by using $\theta_t = 1$ and $\theta_c = \hat{\theta}$, then solve the problem for various values of $\hat{\theta}$ ranging from 0 to 100 in increments of 0.05. Finally, we search the results for the plan that (1) satisfies the dose-volume constraint and (2) has the greatest minimum tumor dosage, then we accept this $\hat{\theta}$ as the best reduced weight parameter. Essentially, the parameter $\hat{\theta}$ represents the priority of minimizing the critical organ term relative to minimizing the tumor term. Mathematically, the problems have the same solutions, but the values of the objectives will vary by a constant factor. However, we note that choices made in the optimization algorithm may result in local minima obtained by using the reduced parameter $\hat{\theta}$ that are not the same as if we had used two separate parameters with the ratio $\frac{\theta_c}{\theta_t} = \hat{\theta}$. In all our experiments, though, the 1-parameter solutions have not fallen significantly far from the corresponding 2-parameter solution. Thus, in lieu of trial and error for each configuration, we use this scheme to get the "best" weighted least squares parameters. Even though this is in no way an optimal strategy for choosing the optimal parameters, it is a systematic way of approximating optimal parameters which gives us some advantage over a random selection.

Incidentally, we have not mentioned any corresponding strategy for eliminating "hot spots" using the weighted least squares formulation. Although some ideas about how to do this may be gleaned from [10], we have not implemented a cooling procedure for the weighted least squares results. All our successive linear programming results do, however, include the cooling procedure described in Section 4. Clearly, any cooling procedure is a process which further constrains the optimization. Thus, it should be clearly understood that the weighted least squares results actually have that advantage over our successive linear programming results since without the cooling procedure, our method would be less constrained and the results would likely be even better.

6.2 Experimental Details

Our experiments are all performed on a 2D cross section of a simulated patient. We assume all the beam angles are fixed in advance to be equally-spaced. In general, better conformation can be achieved with more beams, but this increases the size of the optimization problem, resulting in slower planning times. The problems of how many beams to use and which angles to choose so that optimal plans result will be treated as a distinct optimization problem and will not be discussed in this paper. For all our results we will use 9 coplanar beams, as this gives very good conformation in general for any formulation. The parameters used in generating results from our formulation are as follows: $d^T = 80$ Gy, $u_t = 88$ Gy, $u_c = 25$ Gy, $\rho = 0.25, \lambda = 0.01$. Of course, the same values of $d^T = 80$ Gy and $u_c = 25$ Gy were used to generate the weighted least squares results. The choices for u_h and u_r are more complicated and will be discussed in Section 7. Finally, the best weighted least squares "reduced parameter" $\hat{\theta}$ is chosen as described above. But, weighted least squares results seem to be slightly better when two parameters θ_t, θ_c are chosen such that $\frac{\theta_c}{\theta_t} = \hat{\theta}$. So, once we get the "best" value of $\hat{\theta}$, we choose two relatively small parameters θ_t and θ_c with that ratio and perform the optimization once more to get the final weighted least squares result. Admittedly, this is a rather tedious heuristic for choosing the parameters, but we wanted to be as fair as possible to weighted least squares and besides, this should underscore how difficult it is to find good parameters in practice. Now, we need to describe the data we use for our experiments and the specific test cases we will simulate.

6.2.1 Description of Dose Contribution Data

The dose matrix H in (3b), which models how beamlet intensity maps to delivered dosage, happens to be very expensive to compute and requires vast memory to store. There is an entire body of theory on how to compute this matrix, but we will not focus on that in this paper. Our goal is simply to demonstrate the effectiveness of our method in a controlled environment where expensive dose calculations as well as inhomogeneity in the internal structures and complicated physical concerns will not be relevant factors. Toward this end, we have obtained data for this dose matrix directly from The University of Wisconsin-Madison Tomotherapy Research Group (`http://madrad.radiology.wisc.edu/penbeam/`). In their optimization studies, they have computed a set of pencil-beam (beamlet) data using the convolution/superposition method [4], to be used in investigating IMRT optimization formulations and algorithms. Their dose calculations are for a phantom consisting of two cylinders sharing a common central axis. The inner cylinder is of radius 10cm and contains "lung" tissue with a relative density of 0.3, while the outer cylinder is of radius 20cm and surrounds "water" with a relative density of 1.0. This allows us to test various tumor/critical organ configurations by defining regions within the "lung" corresponding to each. Consequently, our experiments always assume the tumor and critical organ have the same relative density (0.3 in this case). Also, this data is intended to be used in simulated regions of interest that are 2D slices of this cylindrical phantom, rather than the 3D ones considered in practice. So, here the region of interest used in computing the data consists of 2D voxels of size 0.2cm $\times$ 0.2cm, resulting in a discretized 2D region of interest that is 101 voxels $\times$ 101 voxels.

Moreover, the data represents the dose calculations for one beam of 99 beamlets, with each beamlet having its own 101×101 "dose contribution matrix," which gives the dosage contributed to every voxel in the region of interest per unit of beamlet intensity. Note that because there is circular symmetry about the central axis and homogeneity within the cylinders, we can simply rotate the given dose contribution matrix by any angle to obtain the dosage contributed to the region of interest by each beamlet in a beam emitted from that angle. Finally, given one dose contribution matrix corresponding to each beamlet, we construct the "dose matrix" H in (4b) in the following way. Each column of H corresponds to one of the dose contribution matrices, reshaped into a vector of length 101^2. Thus, the product $H_{ij}w_j$ gives the dosage contributed to the i-th voxel by the j-th beamlet and similarly $d_i = H_i w$ (H_i the i-th row of H) gives the total dosage contributed to the i-th voxel by all the beamlets. Further details about this data can be found in [8, 9].

6.2.2 Configurations

We investigate four simple tumor/critical organ configurations, some of which are too simple to be useful for anything other than testing purposes while others may actually be encountered in practice. The first configuration has an "L"-shaped tumor with a small critical organ between the arms of the "L." Note that intuitively one could plan an IMRT session even with only two beams quite easily by setting up fairly high-intensity beams to be collinear with the arms of the "L," thus depositing a large amount of radiation in the tumor and especially at the overlap, while completely avoiding hitting the critical organ directly. Of course, this would require making the beamlets so intense that the radiation bleeding over from the tumor into the critical organ would likely be unacceptable. Thus, a little more finesse and more beams are necessary for a good plan. Next, the "C" configuration has a "C"-shaped tumor with a critical organ placed between the arms. Again, naively we could simply use three beams to cover the area of the tumor, but we would encounter the same problem. This case is important because it shows up frequently in practice when a tumor is, for instance, wrapped around a patient's spinal column. Both of these examples share the property that one side of the tumor is open to the critical organ tissue, but both the "O" and "OA" configurations have tumors that are essentially rings surrounding a critical organ in the center. This implies there will likely be no way to sufficiently irradiate the tumor without the beams passing through the critical organ. Trading off the gains achieved by high dosage to the tumor and the risk of doing so to the critical organ make these cases particularly important because they are closer to the difficult choices that must be made in clinical cases. In addition, the "OA" case is distinct in that it also includes a second critical organ region (same dosage upper bound) directly above the tumor. Finally, we should mention that all these geometries have sharp corners due to the discretization, that is the "O" for instance, actually has a square tumor boundary with a square inner critical organ.

6.2.3 Visualization of Results

We will present plots of dosage distributions and dose-volume histograms for the configurations described above, obtained using both our successive linear programming approach and weighted least squares. By dosage distribution, we mean a representation showing the intensity of the absorbed radiation at all voxels in the region of interest. The scale is in units of Gray (Gy), the standard unit for measuring radiation absorption in IMRT. We will pay close attention to features such as how well the radiation conforms to the tumor and how much it avoids the critical organ. A dose-volume histogram (DVH) for a given dosage distribution relates dosage to volume for each structure. In each DVH, there are curves for the tumor, critical organ, and what we are calling "normal" tissue (the remaining region of interest). Each curve in the DVH is a plot of the volume (ratio of voxels) within that structure that has dosage greater than that corresponding dosage. For instance, we can interpret the critical organ DVH curve passing through the point $(10, 0.5)$ as meaning 50% of the critical organ voxels in this treatment plan have dosages of 10 Gy or more. These dose-volume histograms are familiar visualization tools for practitioners in the field of IMRT, so we include them for completeness, although we will not draw any conclusions from them.

6.3 Results

First, we consider the L-shaped tumor and note the WLS weight parameters we use are $\theta_t = 10, \theta_c = 5$. The dosage distribution plots in Figure 1 demonstrate that in this case, both the successive linear programming (SLP) approach and the weighted least squares (WLS) approach conform radiation to the tumor fairly well with SLP perhaps allowing slightly more radiation to bleed out. However, these plots are only helpful to the analysis qualitatively. For more quantitative information, Table 1 gives the minimum and maximum dosages planned for the tumor and critical organ as well as the percentage of voxels in the critical organ that satisfy $d_i > u_c = 25$ Gy, which should be no greater than $100\rho = 25\%$ if the dose-volume constraint has been satisfied. We can see from Table 1 that the minimum tumor dosage for SLP is much closer to the prescribed dose of $d^T = 80$ Gy than WLS, but we observe that SLP has violated the dose-volume constraint.

In our SLP implementation, it happens that we have yet to integrate into the relaxation procedure the capability to stop exactly at the ρ dose-volume threshold. Thus, some of our results slightly violate the dose-volume constraint. Future efforts will include a more sophisticated relaxation procedure, so that we can allow the DV % to increase gradually in a more controlled manner. In fact, since the SLP solution improves monotonically as we get closer to the ρ threshold, if we had the ability to stop the optimization just before it was violated, we would still get a minimum tumor dosage that was close to the prescribed dose of 80 Gy. Contrast this with WLS, which can *encourage* the dose-volume constraint to be satisfied, yet cannot explicitly control the DV % in any way as the optimization proceeds.

Next, we look at the C-shape simulation and use the WLS parameters $\theta_t = 10, \theta_c = 4.5$. In Figure 2, we observe better conformation using SLP than WLS and in Table 2, we can see SLP achieves the prescribed dose in the tumor while the dose-volume constraint is violated only slightly. By our reasoning above, we submit that with

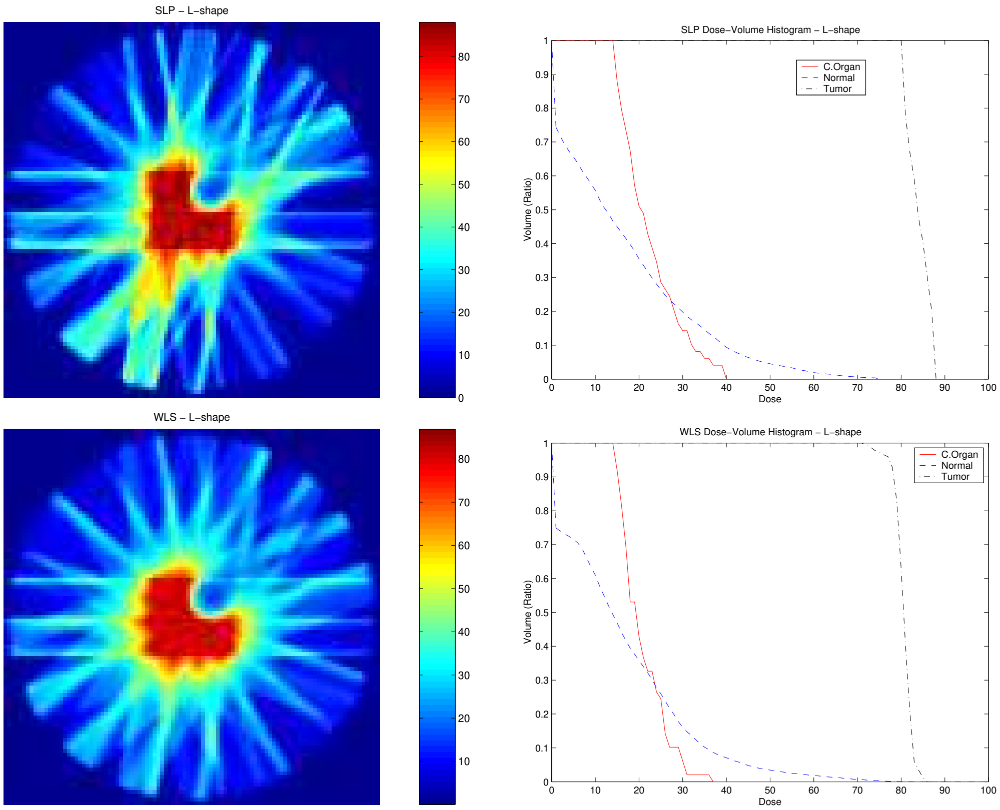

Figure 1: Plots for L-shape Configuration.

a more accurate relaxation procedure SLP could achieve the prescribed tumor dose while satisfying the dose-volume constraint. Notice, however, that WLS is relatively far from the prescribed dose while the DV % is only 13.33 %. This is bad, of course, because it means that WLS could effectively do much better if we only knew better weight parameters. That is, even with our (not necessarily optimal) strategy for picking weight parameters, we cannot be assured WLS will even get close to the dose-volume threshold ρ.

For the O-shape configuration, we use WLS weight parameters $\theta_t = 10, \theta_c = 6.5$. The dosage distributions in Figure 3 again suggest SLP does a better job of conforming the radiation to the target region. This "O" case is particularly interesting because

	Tumor		C.Organ		
	Min	Max	Min	Max	DV %
WLS	71.42	85.26	14.78	36.39	24.49
SLP	80.90	88.00	14.18	40.00	28.57

Table 1: Summary Data for L-shape Configuration

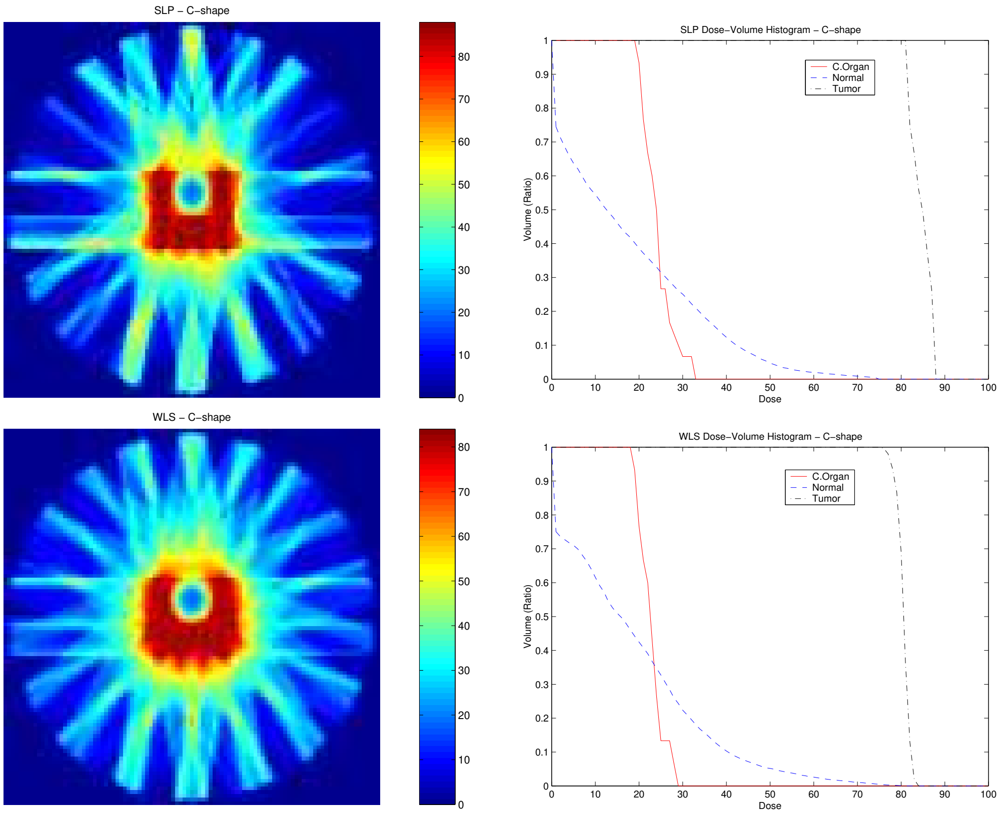

Figure 2: Plots for C-shape Configuration.

both SLP and WLS are similar in their treatment of the critical organ. That is, observe in Table 3 that both the critical organ min and max and the DV % are very close. Clearly, though, the minimum tumor dosage is significantly higher using SLP than WLS.

Finally, we have the OA-shape configuration for which we use WLS parameters $\theta_t = 1, \theta_c = 6.75$. In addition to the critical organ in the center of the "O," there is another critical organ region just above the upper boundary of the "O." We treat this simply as an extension of the first and proceed as if they were one critical organ. Although less definitive, we believe that the dosage distributions in Figure 4 bear out that SLP also conforms the radiation to the tumor shape better than WLS. The data

	Tumor		C.Organ		
	Min	Max	Min	Max	DV %
WLS	75.63	84.00	18.09	28.68	13.33
SLP	81.84	88.00	19.11	32.63	26.67

Table 2: Summary Data for C-shape Configuration

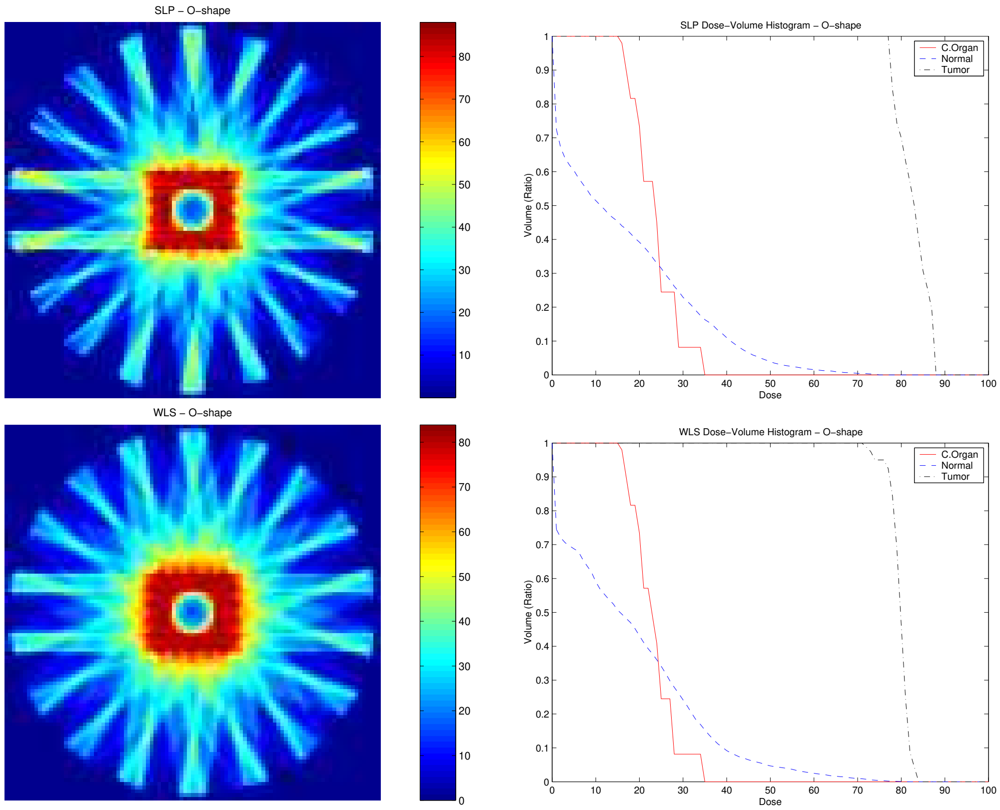

Figure 3: Plots for O-shape Configuration.

summarized in Table 4 show a dramatic difference in the minimum tumor dosage, with SLP coming much closer to the prescribed dose than WLS. Also, note our (admittedly simplistic) relaxation procedure prevents us from doing even better since with a more sophisticated relaxation procedure, we could improve our solution by increasing the DV % from 20.19 % to closer to the 25 % boundary.

	Tumor		C.Organ		
	Min	Max	Min	Max	DV %
WLS	71.32	83.88	15.70	34.21	24.49
SLP	77.99	88.00	15.66	34.93	24.49

Table 3: Summary Data for O-shape Configuration

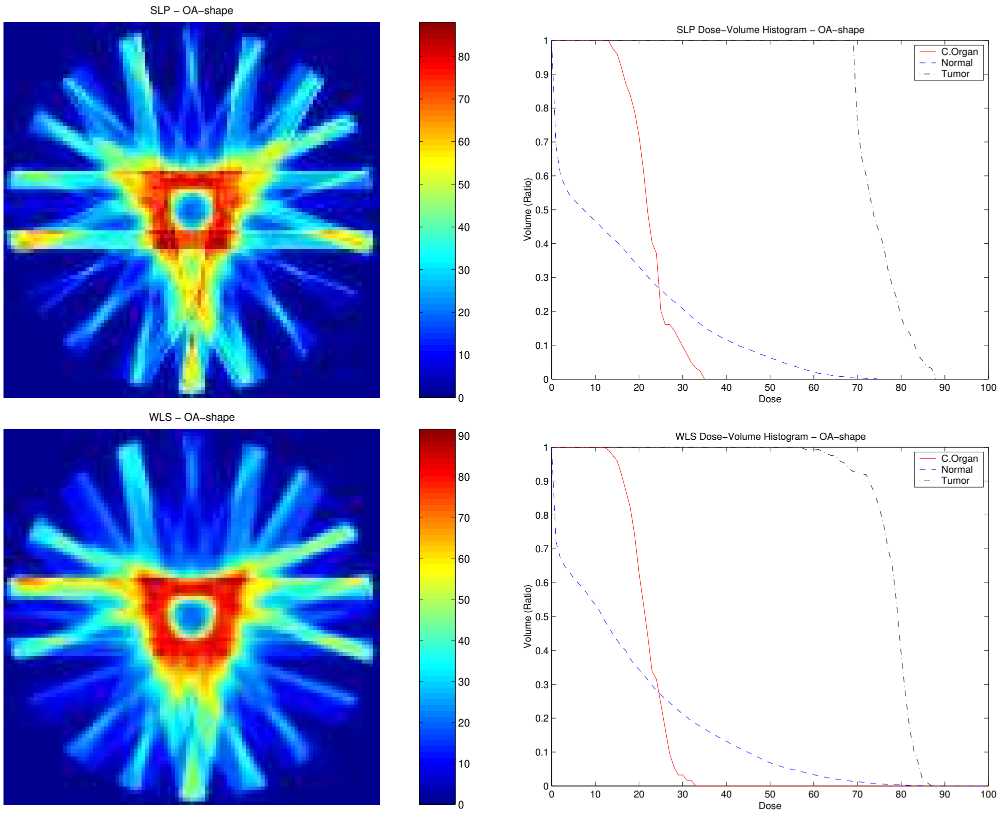

Figure 4: Plots for OA-shape Configuration.

7 Future Research

Because this paper represents the first effort to formulate and solve the IMRT beamlet weight optimization problem in this way, many opportunities for future research directions present themselves. We attempt to organize them here and suggest some possibilities.

The algorithm given in Section 5 requires several parameters including u_t and u_h. We have included u_t as an upper bound on the tumor dosage in the formulation (4). For our experiments, we chose $u_t = 1.1\ d^T$, or 10% more than the prescribed dose. Adjusting this value may have the effect of enforcing more or less homogeneity in

	Tumor		C.Organ		
	Min	Max	Min	Max	DV %
WLS	57.09	86.69	12.71	32.74	24.19
SLP	69.31	88.00	13.45	34.97	20.16

Table 4: Summary Data for OA-shape Configuration

the tumor dosage distribution. Also, to eliminate hot spots in the "normal" tissue, we proposed a threshold value u_h which specifies how large a dosage must be to be considered"hot" (See Section 4). This presentation was used for simplicity to explain the cooling procedure in general, but in practice it may be more efficient to use a much more adaptive procedure. In our numerical experiments, at each cooling iteration, we searched the normal dosages for those that exceeded 80% of the maximum normal dosage. These we considered as being "hot" for that iteration and we then constrained those corresponding voxels to have dosages of at most u_h, resolved the problem, and continued until all normal dosages are at most u_h. We believe this type of adaptive cooling procedure may reduce the number of cooling iterations required dramatically.

7.1 Dose Matrix

Since our approach relies heavily on computed dosages, providing an accurate method for calculating dose will undoubtedly be an important research direction. Any questions of efficiency will almost certainly involve considering how the dose matrix H is computed and how accurately it is computed. We have used a very simple dose matrix for our experiments which can be stored and rotated to obtain dose matrices for various angles around the phantom. Real people are, however, much more irregular and inhomogeneous which forces dose calculations to be even more expensive and makes the matrix H so large it typically cannot be stored ahead of time. To circumvent this, many commercial IMRT planning systems calculate dose "on the fly," requiring various levels of accuracy at different stages in the optimization to reduce computational cost. Although we have not considered how and when to compute *approximations* to H in this paper, having a solid understanding of how this is done and how to improve its interface with the optimization algorithm will undoubtedly be a key factor in increasing algorithm performance.

7.2 Relaxation Procedure

A key idea in our approach is improving the objective function by successively relaxing the constraint right-hand sides. However, our presentation of the relaxation procedure was oversimplified to convey the idea. In fact, we implemented a very different procedure to generate our results in Section 6. Once the most sensitive constraints were found by forming $R_k = \{i \in \mathcal{C} : y^*_{i,k-1} > \lambda\}$, we relaxed the right-hand sides with the update $b_i \leftarrow 2\ b_i, i \in R_k$. Simple alternatives include setting the upper bounds to some constant value u_r as laid out in Section 3 or using some other update such as the one suggested in Section 2: $b_i \leftarrow b_i + \epsilon_i, i \in R_k$, where the ϵ_i are chosen by a more sophisticated scheme. As seen in some of the results in Section 6, we have not yet built into our algorithm the capability to stop our relaxation before the ρ dose-volume constraint threshold is crossed. This will require considering how many right-hand sides $b_i, i \in R_k$ can be relaxed in order to achieve a dose-volume ratio of no more than ρ and possibly adaptive choices for λ. Since better relaxation procedures have the potential to decrease time to an acceptable plan and produce solutions which are closer to the optimal solution of the mixed-integer program we really want to solve, this is an extremely important direction for future research.

7.3 Extensions

In the formulation (4), we have assumed only one tumor and one critical organ with one dose-volume constraint for that critical organ. Here, we briefly present a possibility for a more general formulation. Assume we have N_t tumors with prescribed doses $d_1^p, \dots, d_{N_t}^p$ and N_c critical organs with corresponding tolerance doses of $u_1^c, \dots, u_{N_c}^c$. We then define the "k-th LP" to be

$$\begin{aligned}
\max \quad & \tau_1 + \cdots + \tau_{N_t} && && \text{(7a)}\\
\text{s.t.} \quad & d = Hw, && w \geq 0, && \text{(7b)}\\
& d_1^p \tau_1 \leq d_i \leq u_1^t, && i \in \mathcal{T}_1, && \text{(7c)}\\
& \vdots && && \text{(7d)}\\
& d_{N_t}^p \tau_{N_t} \leq d_i \leq u_{N_t}^t, && i \in \mathcal{T}_{N_t}, && \text{(7e)}\\
& d_i \leq u_1^r, && i \in R_1^k, && \text{(7f)}\\
& \vdots && && \text{(7g)}\\
& d_i \leq u_{N_c}^r, && i \in R_{N_c}^k, && \text{(7h)}\\
& d_i \leq u_1^c, && i \in \mathcal{C}_1 \backslash R_1^k, && \text{(7i)}\\
& \vdots && && \text{(7j)}\\
& d_i \leq u_{N_c}^c, && i \in \mathcal{C}_{N_c} \backslash R_{N_c}^k, && \text{(7k)}
\end{aligned}$$

where $\mathcal{T}_1, \dots, \mathcal{T}_{N_t}$ and $\mathcal{C}_1, \dots, \mathcal{C}_{N_c}$ are the index sets for the tumors and critical organs, respectively, $u_1^t, \dots, u_{N_t}^t$ are the upper bounds on tumor dosage, $u_1^r, \dots, u_{N_c}^r$ are the relaxed upper bounds for the critical organs, and $R_1^k, \dots, R_{N_c}^k$ are the corresponding relaxation sets for the critical organs. This formulation requires every tumor voxel's dosage to be at least some fraction of that tumor's prescribed dose, while maximizing the sum of those fractions over all the tumors. All the details generalize in a straightforward way.

In the formulation 7, we have assumed that each time we solve the "k-th LP," we relax constraint right-hand sides corresponding to every critical organ. In practice, it would seem more reasonable to relax according to priority. In this more general setting, we have implicitly assumed having at least one dose-volume constraint associated with each critical organ. However, the physician often specifies that protecting certain critical organs is more important than protecting others. So, we might think of relaxing constraint right-hand sides for the least important organ $\mathcal{C}_1$ first until the maximum allowable for that organ, say ρ_1, had been relaxed. Then, if we had not achieved the prescribed dosages in all tumors, we would proceed to the second least important organ $\mathcal{C}_2$ and continue relaxing. In this way, we could make good improvement to the minimum tumor dosages while gradually incurring risk of complications in a controlled, prioritized manner.

8 Conclusion

In this paper, we have proposed a new formulation of the IMRT planning optimization problem. Our successive linear programming (SLP) approach provides a precise control on dose-volume constraints and eliminates the weight parameter guesswork inherent in weighted least squares (WLS) approach. Preliminary results indicate that SLP conforms radiation to the tumor shape better than WLS and achieves a higher tumor-dosage escalation.

This present paper is a feasibility study for test of concept. Works are currently underway to implement the proposed approach in a realistic setting, and to compare it with the standard approach on clinical cases.

Acknowledgments

We would like to thank The University of Wisconsin-Madison Tomotherapy Research Group for making their pencil beam data sets available online that have facilitated our numerical simulation.

References

[1] V. Chvátal. *Linear Programming.* W.H. Freeman and Company, New York, 1983.

[2] Allen Holder. Designing radiotherapy plans with elastic constraints and interior point methods. Mathematics Technical Report 49, Trinity University, San Antonio, TX, 2000.

[3] Allen Holder. Radiotherapy treatment design and linear programming. Mathematics Technical Report 70, Trinity University, San Antonio, TX, 2002.

[4] T. Mackie, P. Reckwerdt, and N. Papanikolaou. 3D radiation treatment planning and conformal therapy, 1995.

[5] S. Morrill, R. Lane, J. Wong, and I. Rosen. Dose-volume considerations with linear programming optimization. *Medical Physics*, 18(6):1201, 1991.

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

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

[8] D. Shepard. *A Framework for the Optimization of Intensity Modulated Radiotherapy.* PhD thesis, University of Wisconsin-Madison, 1999.

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

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