This is an author produced version of an article that appears in:

**MEDICAL PHYSICS**

The internet address for this paper is:

https://publications.icr.ac.uk/2639/

Copyright information:

http://www.aip.org/pubservs/web_posting_guidelines.html

**Published text:**


Institute of Cancer Research Repository
https://publications.icr.ac.uk

Please direct all emails to:
publications@icr.ac.uk
Constrained segment shapes in direct-aperture optimization for step-and-shoot IMRT

James L. Bedford and Steve Webb
Joint Department of Physics, The Institute of Cancer Research and Royal Marsden NHS Foundation Trust, Downs Road, Sutton, Surrey SM2 5PT, United Kingdom

(Received 3 June 2005; revised 20 October 2005; accepted for publication 8 December 2005; published 17 March 2006)

Previous studies have shown that, by optimizing segment shapes and weights directly, without explicitly optimizing fluence profiles, effective IMRT plans can be generated with fewer segments. This study proposes a method of direct-aperture optimization with aperture shape constraints, which is designed to provide segmental IMRT plans using a minimum of simple, regular segments. The method uses a cubic function to create smoothly curving multileaf collimator shapes. Constraints on segment dimension and equivalent square are applied, and each segment can be constrained to lie within the previous one, for easy generation of fluence profiles with a single maximum. To simply optimize the segment shapes and reject any shapes which violate the constraints is too inefficient, so an innovative method of feedback optimization is used to ensure in advance that viable aperture shapes are generated. The algorithm is demonstrated using a simple cylindrical phantom consisting of a hemi-annular planning target volume and a central cylindrical organ-at-risk. A simple IMRT rectum case is presented, where segments are used to replace a wedge. More complex cases of prostate and seminal vesicles and prostate and pelvic nodes are also shown. The algorithm produces effective plans in each case with three to five segments per beam. For the simple plans, the constraint that each segment should be contained within the previous one adds additional simplicity to the plan, for a small reduction in plan quality. This study confirms that direct-aperture optimization gives efficient solutions to the segmental IMRT inverse problem and provides a method for generating simple apertures. By using such a method, the workload of IMRT verification may be reduced and simplified, as verification of fluence profiles from individual beams may be eliminated. © 2006 American Association of Physicists in Medicine. [DOI: 10.1118/1.2163832]

Key words: radiotherapy, IMRT, segmental, step-and-shoot, optimization, treatment plan

I. INTRODUCTION

The traditional approach to inverse planning for intensity-modulated radiation therapy (IMRT) has been to produce a map of desired fluence for each beam in the treatment plan, and then to determine in a separate step, how to deliver this fluence. For step-and-shoot or segmental IMRT, this second step involves determining the shapes and intensities of a series of beamlets defined by a multileaf collimator, such that the individual fluences collectively sum to yield the overall desired fluence profile. This method is fast and flexible, but the practical implementation suffers from a major limitation, namely that the optimization phase takes no account of the delivery constraints. Although the optimization may be approximately guided by a knowledge of what is practically achievable, it is inherently unintelligent as to the precise delivery constraints. Thus, it may produce a solution which subsequently requires very many irregularly shaped segments to realize, when, due to the degeneracy of the inverse-planning problem, it could actually have selected fluence profiles which could be very simply delivered. The excessive complexity in this case leads to difficulties in treatment-plan verification and ultimately to uncertainty in the quality of the treatment. In particular, there may be many intricate match lines between the segments, leading to lines of incorrect fluence in the profiles of individual beams. These effects may even propagate into the composite dose distribution resulting from all beams. A further difficulty with the two-step approach is that the initial optimization assumes a continuous distribution of fluence, whereas the segmental plan is composed of a series of step functions. The conversion from one to the other inevitably introduces a loss of plan quality, which is difficult to recover. Several authors have realized these limitations and proposed methods to overcome them. For example, Cho and Marks adopt an iterative scheme which carries out a continuous optimization but periodically examines the effectiveness of delivery of the evolving treatment plan. Siebers et al. take a similar approach.

A further step in progress is made when it is realized that, because the final segmental treatment plan consists of a series of apertures which are closely related to the anatomy of the patient and the relative positions of the anatomical structures, then it is possible to generate apertures based directly on the anatomy. These apertures are then adjusted in shape and intensity to provide the desired treatment plan. These ideas are generalized in the elegant concept of direct-aperture optimization. In this method, a single optimization is used, the variables being the positions of the multileaf collimator (MLC) leaves and the intensities of the
segments. Because the delivery constraints are taken into account at all stages of the optimization, the method promises to provide treatment plans which are as regular as possible. Consequently, verification of fluence profiles from individual beams may be eliminated, thereby reducing the workload of verification. Also, by definition, all of the physics of head-scatter, leaf transmission, and other effects can be directly incorporated. There is no further work to be done on leaf sequencing. A similar method of aperture shape optimization for nonmodulated conformal radiotherapy has been described by van Dalen et al. 14

Initially, up to 25 segments per beam were thought necessary for effective IMRT, but the studies based on anatomical apertures 11,15 and other studies 16 now suggest that rather fewer segments may be required for most problems. Direct aperture optimization appears to provide solutions with the most efficient use of the segments specified. 12,13 Furthermore, because the fluences are optimized as well as the aperture shapes, fewer apertures may be needed to realize a given number of intensity levels in the intensity-modulated beams. The general rule is that for \( N \) segments, between \( N \) and \( 2^{N-1} \) intensity levels may be obtained, depending upon the intensity pattern. 12

Taking into account these considerations, direct-aperture optimization appears to be the method of choice for providing simple, practical segmental IMRT with rational, easy-to-verify aperture shapes. The work described in this paper extends this topic by applying aperture-shape constraints to provide the simplest possible segments with fewest matches. The resulting algorithm is also explored with a view to determining the limit of what can be achieved in IMRT with a few segments.

II. METHODS
A. Software framework

For the work described in this paper, direct-aperture optimization has been implemented into the AUTOPLAN inverse planning program developed at this Center. The details of this program have been published elsewhere. 17,18 The software is written in Java (Sun Microsystems, Santa Clara, CA) and is designed to be interfaced to a commercial treatment-planning system. In the normal mode of operation, structure delineation takes place on the commercial treatment-planning system. The computed tomography (CT) scan and structure set are then read by AUTOPLAN and optimization is carried out using a convolution dose calculation including full heterogeneity correction. 17 The final plan is then returned to the commercial treatment-planning system for final calculation and evaluation. This allows the benefit of a fully commissioned treatment-planning system to be exploited. However, in order to present the details of the optimization as clearly as possible, the results shown in this paper are as they appear within AUTOPLAN.

B. Objectives and constraints

To initiate the optimization, the user supplies a series of objectives and constraints [Fig. 1(b)]. The objectives are to minimize or maximize various dose, dose-volume, or biological treatment-plan statistics, with an importance factor between 1 and 1000. The constraints are to ensure that various dose, dose-volume, or biological statistics are maintained above or below a specified tolerance. This is similar to the scheme described previously, 18 except that multiple objectives are now allowed. This provides greater flexibility in balancing dose between different anatomical structures. It also allows a structure to be included as both an objective and constraint. This is useful, for example for the spinal cord, where the main constraint is for the maximum dose to be 45 Gy, but additionally, as an objective with a much lower importance, the maximum spinal cord dose should be minimized beyond this to as low a dose as possible.

From these objectives and constraints, the objective function is constructed. In algorithmic form, this is:

\[
\text{If the constraints are satisfied:} \\
\text{Minimize the weighted sum of the objectives;} \\
\text{Else:} \\
\text{Add a constant to the objective function to signify a poor solution, and minimize the weighted sum of the amounts by which the constrained statistics exceed their respective tolerances. In addition, minimize the weighted sum of the objectives with reduced importance factors.}
\]

Mathematically, the goal is to minimize an objective function \( f(d(x)) \), where \( d(x) \) is a distribution of dose among the voxels of the patient model, and \( x \) is a vector of variables to be selected, such as beam shapes and weights. If the treatment plan statistics are represented as \( f_1, f_2, f_3, \ldots \), in the form of \( M \) objectives and \( N \) constraints, the objective function is as follows:

\[
f(d(x)) = \sum_{m=1}^{M} \beta_m f_m(d) \quad \text{if} \quad \sum_{n=1}^{N} H(f_n - c_n) = 0,
\]

\[
= A + \sum_{n=1}^{N} H(f_n - c_n)(f_n - c_n) + B \sum_{m=1}^{M} \beta_m f_m(d) \quad \text{otherwise},
\]

where the \( \beta_m \)'s are importance factors, the \( c_n \)'s are limits on the \( N \) constraints, \( A \) is a large constant, and \( B \) is a small importance factor. This is illustrated in Fig. 2.

The basic form of the objective function follows that described previously, 18 in which the objective is minimized unless the constraints are not met, when the objective function is forced artificially high to signify a poor solution. However, the previous formulation suffers from several drawbacks.
FIG. 1. AUTOPLAN user interface showing the various screens guiding the user through the inverse planning process. (a) Welcome screen; (b) clinical constraints screen, where objectives and constraints are entered; (c) physical constraints screen, where number of segments and beam angles are entered; (d) settings screen, where segment constraints such as minimum dimension and requirement for leaves to close in are entered; (e) beam’s eye view screen; (f) dose-volume histogram screen; and (g) status panel, showing the progress of optimization.
First, if the constraints are not met, the single high value returned by the objective function offers the optimization engine no incentive to move downhill toward the region where the constraints are satisfied. Second, if the constraints cannot be met due to the inherent physical properties of the radiation transport, the simple formulation described previously does not attempt to optimize the objectives as it simply returns a high value.

In the present formulation, the objective function minimizes the objectives if the constraints are met and a constant is added to the objective function if the constraints are not met, to signify a poor solution. However, the weighted sum of the amount by which the constraints exceed their respective tolerances ensures that the objective function reduces as the treatment plan comes closer to satisfying the constraints. Furthermore, by including a weighted sum of the objectives with a relatively low weight when the constraints are not satisfied ensures that even when it is impossible to satisfy the constraints, the optimization will at least provide the best it can achieve with the objectives.

C. Treatment plans

For each field in the treatment plan, the user selects a machine and the desired number of segments for that field [Fig. 1(c)]. The number of segments can be varied from beam to beam, and there can be a mixture of conformal and segmental beams if required, thereby allowing considerable flexibility. The nominal penumbra margin is also selected. This is the maximum margin allowed around the edge of the planning target volume (PTV), within which all segments must lie. The structure designated as the PTV can be chosen from all of the available contoured structures. For each beam, the gantry, couch, and collimator angles are specified.

The segment apertures are specified and optimized by means of the two collimators in the one direction and the two MLC leaf banks in the orthogonal direction. The backup collimators behind the MLC leaf banks are aligned with the rearmost MLC leaf and play no part in the optimization. For each segment, the two collimators orthogonal to the MLC leaf banks are set first, followed by the MLC leaf banks.

A key feature of the algorithm is the use of a polynomial curve to specify the positions of the leaves in each bank of the MLC (Fig. 3). For one bank of leaves, the endmost exposed leaves and two intermediate leaves are set explicitly. The remaining leaves on that bank are positioned by fitting a cubic function to these preset leaves. Four field widths, adjacent to the four explicitly set leaves, are then used to fit a second cubic function to characterize the field width. The leaves in the second bank are then positioned according to the positions of the leaves in the first bank and the cubic field width. The reason for using field width to generate the leaves in the second bank, rather than setting four leaves and fitting the rest of the leaves to these, is that the optimization can suggest field widths relatively independently of the positions of leaves in the first bank. If the second leaf bank were to be set directly, the leaves may collide with those in the first bank, but by using field width, the leaves in the second bank can be maintained a suitable distance from the leaves in the first bank.

The advantages of this method are (1) that the number of variables involved in the optimization of each leaf bank is just four, regardless of the field size, and (2) the field shapes are smoothed and regularized by this process. Each segment is therefore characterized by 10 variables—two jaw positions with which the length of the field is delineated, four MLC leaf positions on the first leaf bank, and four field widths from which the width of the field is defined. This approach can be considered analogous to the use of basis functions for fluence optimization.

It may be possible to use other polynomial functions besides a cubic function. The higher the order of the polynomial, the greater the flexibility of the segment shape. However, too much flexibility prevents the production of simple,
regular apertures. A cubic function provides sufficient flexibility in shape while retaining an overall regular form. Note that in smaller apertures, the four MLC leaves chosen for optimization are closer together, so a more intricate aperture shape may be created. Also note that abrupt changes in contour shape may be handled by positioning the collimators appropriately, without necessarily requiring an abrupt change in leaf settings.

Each segment is constrained by four main constraints [Fig. 1(d)]. The first of these is the minimum dimension of the field. This is a user-defined distance which gives the minimum width or length of the segment. Its significance in the optimization is that it prevents excessively short or narrow segments from being generated. The second constraint is the equivalent square. This is necessary to prevent segments with small areas. The minimum dimension precludes many of such segments, but for example, a narrow diagonal segment has large length and width but only a small area. The additional constraint is therefore necessary.

The remaining two constraints can be selected as required and apply to the sequence of segments in a beam. The first is a specification that the first segment in the beam must be a normal conformal aperture. This is to provide confidence that part of the treatment plan is geometrically conformal, and it ensures that during patient setup, the light field for the first segment is larger than the light fields for the other segments. Thus, if there is any potential collision with the couch bars, this can be seen easily by selecting the first segment. This option also allows for ease of portal imaging. The second constraint (if applied) is that each segment must lie within the previous. This minimizes any matchplanes within the beam and consequently eliminates any tongue-and-groove artifact. It also provides an ideal basis for generating simple IMRT plans where there is only one maximum in the fluence profile for each beam, for example where the intensity profiles are used to simulate wedges. In this case, the $2^{N-1}$ rule breaks down, and the number of segments required is equal to the number of conceptual intensity levels into which each beam is divided.

Besides these constraints, the normal collimator and leaf constraints for the linear accelerator are taken account of. For the Elekta Precise linear accelerator and MLCi (Elekta Oncology Systems, Crawley, UK) used in this work, these are that the MLC leaves and backup collimators can travel beyond the central axis by no more than 125 mm, opposing leaves can move no nearer to each other than 10 mm, and the collimators orthogonal to the MLC leaf banks cannot travel beyond the central axis.

Segment weights can be selected to be adjusted in fine intervals, coarse intervals, or not to be optimized at all. If fine weights are selected, the algorithm assigns to each segment a relative applied weight of between $0.1/N$ and $1/N$, in steps of $0.1/N$, where $N$ is the number of segments per beam. If coarse weights are selected, the algorithm increments the beam weights in steps of $0.25/N$, from $0.25/N$ to $1/N$. If unit weights are selected, each beam contributes an equal number of monitor units to the treatment plan, but the segments within each beam are weighted such that the first segment has a relative weight of $1/N$, reducing linearly down to the last segment, which has a weight of $0.1/N$. This is to ensure that realistic beam modulations can be produced with fixed segment weights.

### D. Patient model and dose calculation

The treatment plans are constructed in conjunction with a model of the patient. The patient model is a three-dimensional grid of voxels which cover a computed tomography (CT) scan of the patient. The CT scan and outline set are resampled onto a regular grid for this purpose. The resolution of this grid is $10 \, \text{mm} \times 10 \, \text{mm} \times 10 \, \text{mm}$ for the first 75% of the optimization, and $5 \, \text{mm} \times 5 \, \text{mm} \times 5 \, \text{mm}$ thereafter. This allows the approximate shape of the segments to be built up in a relatively short time, and then the fine detail of the segments to be adjusted with a more accurate dose calculation.

Each voxel contains all essential information relating to a particular region of the patient. It includes a position coordinate, the width, height and length of the voxel, the CT number of that voxel, the absorbed dose due to each segment of a treatment plan, and the total dose. Storage of the dose due to each segment allows for the total dose to be simply summed if only the beam weights of the plan change. The voxel also contains a list of the volumes of interest of which that voxel forms part, for calculation of dose-volume histograms (DVHs) and dose statistics. For performance purposes, the volumes of interest are also stored separately as lists of voxel indices.

Information regarding ray tracing is stored within each field object as a list of coordinates traversed by each ray, and the equivalent path length to those coordinates. Storage of these variables allows for their re-use, which facilitates fast calculation.

For each treatment plan, dose is calculated using a fast convolution algorithm with full heterogeneity correction. This scheme uses a restricted scatter kernel to provide a full three-dimensional scatter convolution without the time requirement of a traditional convolution dose calculation algorithm. The collimators and MLC leaves have distinct transmissions. These are typically 1% for the collimators and 3% for the MLC leaves. Head scatter is implicitly taken into account by the off-axis dose profiles and central axis output factors used by the dose calculation. The calculation has been shown to provide accuracy in isocenter dose, relative to a commercial treatment planning system, of better than 5%, and root-mean-square accuracy in dose-volume histograms of better than 5%. There is therefore very little difference between the dose distribution produced by AUTOPLAN and that produced by a commercial treatment planning system after a final dose calculation. The calculation time is around 2 s per beam. This, therefore, allows the evaluation of a wide range of treatment plans in a feasible length of time. After calculation of the dose distribution using the arbitrary relative weights, the dose distribution is normalized to the mean dose in the PTV.
E. Optimization engine

The optimization engine is in principle a random search over the aperture shape variables. Aperture shapes which violate the constraints are rejected. However, implemented literally, this algorithm is prohibitively inefficient. This is because most randomly generated apertures violate the constraints, and the result is that a large proportion of the optimization time is spent generating and rejecting unsatisfactory apertures. This is particularly so when the option is selected that each aperture should lie within the aperture for the previous segment. In this case, the previous segment effectively forms the constraint. Accordingly, a method called feedback optimization has been developed. This is illustrated in Fig. 4.

The optimization begins with all apertures conformally shaped to the beam’s eye view of the PTV. This is a reasonable starting point, but also ensures that the optimization begins with apertures which satisfy the constraints. The PlanOptimiser module handles the optimization. To optimize the segment shapes, it calls a FeedbackOptimiser module, which is a generic module for feedback optimization. The feedback optimizer cycles through the variables to be optimized one by one, beginning with the first segment of the first beam, making a random perturbation to each variable in turn. The perturbation is constrained by the constraints on the segment apertures. The magnitude of the perturbation is selected from a uniform distribution whose width initially spans the range of allowed values for the variable, but which narrows as the optimization progresses. In other words, at the beginning of the optimization, a particular MLC leaf may be moved to any other allowed position, whereas at the end of the optimization, the leaf may only move a short distance from its current position. If the variable is already close to a constraint boundary, the uniform distribution is skewed to encourage selection of a new value within the allowed range.

There is also a possibility that the new value will be the same as the current value, particularly as the optimization progresses and the distribution narrows. As this is inefficient, three attempts are made to find a new parameter value which differs from the present one.

Once a new parameter value has been selected by the feedback optimizer, it is implemented into the plan by the plan optimizer. It is at this point that the feedback optimization method differs from a classical iterative scheme. Once the new parameter value has been implemented, the boundary for the next variable is updated. For example, if the variable being implemented is a collimator position, once it has been set in the plan, this immediately imposes a constraint that the opposite collimator cannot proceed closer to it than the minimum segment dimension. In addition, the opposite collimator cannot proceed over the central axis. The update() method therefore updates the boundary of the opposite collimator accordingly. This may result in a situation where the opposite collimator is now outside of its boundary, so the feedback optimizer realigns the position of the opposite collimator as necessary. This process of feedback and realignment takes place for all of the variables in the plan every time one of the variables is changed. A similar process also holds for the opposite leaves of the MLC. This process is illustrated in Fig. 5.

The above process ensures that each variable is selected within the constraints, thereby maximizing efficiency. However, there are still circumstances in which the constraints are not satisfied. For example, when the leaf positions are all set and the intermediate leaves are set using the cubic function,
it may be found that a leaf protrudes closer to the adjacent leaf in the opposite leaf bank than the 10 mm constraint that the linear accelerator imposes. In this case, each leaf is addressed in turn, and retracted until it satisfies the constraint. Another example is the minimum equivalent-square constraint. It is almost impossible to generate segments which satisfy this constraint a priori. Thus, any segment which violates the constraint is repeatedly expanded to satisfy the constraint. Finally, if any constraint is not met after these adjustments, the new treatment plan, resulting from the implementation of the new parameter value, is rejected. However, this only occurs in a very small proportion of iterations.

After an updated variable has been implemented into the treatment plan and the boundaries updated, the feedback optimizer requests that the new plan is evaluated (Fig. 5). The plan optimizer recomputes any segment which has changed. (Note that although only one variable has actually been changed deliberately by the feedback optimizer, the process of realignment may cause a number of segments to change.) The total dose due to all of the segments is then computed and the objective function evaluated. At every tenth iteration, corresponding to the renewal of all the variables for one segment, segment weights are also optimized. This takes place using a traditional fast simulated annealing optimizer. The very first time that the segment weight optimization is used, the segment weights are initialized to a range of values, the first segment having a relative weight of $1/N$, reducing linearly to the last segment, which has a weight of $0.1/N$, where $N$ is the number of segments per beam. When segment weight optimization is commenced subsequently, the previous optimal set of segment weights is used as a starting point. At each iteration of the optimization, all of the segment weights are perturbed randomly according to a uniform distribution. As with the feedback optimizer, the width of the distribution is reduced as the optimization proceeds, and the distribution is skewed to encourage selection of valid segment weights. As the segment weight optimization is nested within the aperture-shape-optimization scheme, it is not necessary to recalculate the dose distribution for each new combination of beam weights: the individual segment dose distributions must only be re-summed. The overall time taken for the complete optimization of segment shapes and weights is around 1 to 2 h for five beams and three to five segments per beam on a 1.28 GHz SunBlade 2500 workstation (Sun Microsystems, Santa Clara, CA).

F. Phantom case

The performance of the algorithm was evaluated first for the case of a cylindrical phantom. The phantom was 250 mm in diameter and 300 mm long. The phantom contained a PTV and an organ at risk (OR), which were both 100 mm long, situated centrally along the length of the phantom. The PTV was hemi-annular and had an outer diameter of 150 mm and an inner diameter of 70 mm. Meanwhile, the central OR had a diameter of 50 mm. Five equally spaced 6 MV beams were applied to an isocenter situated at the center of the PTV. The objective was to minimize the root-mean-square deviation from 50 Gy in the PTV, such that the mean dose to the OR was less than 25 Gy. The choice of mean dose was somewhat arbitrary. For some organs at risk (e.g., rectum), the volume irradiated to a high dose is of interest, for other organs at risk (e.g., spinal cord), the maximum dose is of interest, and for yet others (e.g., parotid glands) the mean dose is of interest. Given this range of options, the mean dose was chosen as it was closely related to the integral dose commonly used in objective functions for IMRT. This therefore allowed the present study to be related to previous work. Either one, three, five, or seven segments per beam were specified to the optimization algorithm and the first segment should be conformal. The plan with one segment per beam represented an unmodulated conformal plan. Minimum dimension was 20 mm and minimum equivalent square was 40 mm (i.e., the side of the equivalent square field was a minimum of 40 mm). There was no requirement that each segment should lie within the previous one. A plan was also created with nine segments per beam, no constraint that the first segment should be conformal, minimum dimension 10 mm, minimum equivalent square 20 mm, and no requirement that each segment should lie within the previous one. This plan represented an unconstrained IMRT plan. In addition, in order to evaluate the effect of the requirement that each segment should lie within the previous one, the plan with five segments per beam was repeated with the constraint applied.

G. Clinical case: Rectum

The next demonstration was for the case of a prone rectal treatment. At this Center, these patients are treated using an anterior open field and two lateral wedged fields. Since the wedge is oriented perpendicular to the MLC, this means that on the lateral fields, the MLC leaves approach from superiorly and inferiorly and are therefore unable to spare the bladder and/or bowel which projects into a concavity on the anterior surface of the PTV. In such cases, the collimators are therefore rotated on the lateral fields, so that the MLC orientation is appropriate, and the wedge is removed. Three to four segments are used to create a wedge effect instead.

The direct-aperture-optimization algorithm was demonstrated for one of these cases. The anterior field was an open 6 MV field and the two lateral fields were 10 MV fields with four segments each. The objective was to minimize the root-mean-square variation of dose around 45 Gy within the PTV. The first segment was required to be conformal, and the leaves were required to close in, i.e., each segment was required to lie within the previous one. Minimum dimension was 20 mm and minimum equivalent square was 40 mm.

H. Clinical case: Prostate and seminal vesicles

The prostate is a treatment site which may benefit significantly from the simplified IMRT method described in this paper. This is because the volume is fairly regular and not grossly concave. The reduction in the number of intricate match lines facilitated by the approach given in this paper...
may therefore allow the volume to be treated simply and effectively. It may also allow centers which verify the fluence profile of each beam separately to reduce the extent of their verification. Large numbers of patients are typically treated with prostate IMRT, so this kind of reduction in verification would have a large impact on workload.

For the case planned in this study, the patient was supine and the prostate and seminal vesicles and the prostate alone were both outlined. PTV1 then consisted of the entire prostate and seminal vesicles plus a 10 mm margin in all directions. PTV2 consisted of the prostate alone plus a 10 mm margin in all directions except posteriorly, where there was a 5 mm margin. PTV3 consisted of the prostate alone with a 5 mm margin, but with no margin posteriorly. The prescription was 74 Gy to PTV3 and the goal was to cover PTV3 with the 70 Gy isodose (i.e., 95% of the prescription dose), PTV2 with the 67 Gy isodose and PTV1 with the 56 Gy isodose. This was in accord with the conventional arm of the CHHIP trial (Conventional versus Hypofractionated High-dose IMRT for Prostate cancer).23,24

Five 10 MV fields were used, at gantry angles 0°, 45°, 100°, 260°, and 315°. The objectives and constraints are shown in Table I. Note that although the objectives specified to the inverse-planning algorithm were approximately comparable to the actual clinical objectives, they were not always identical, as some flexibility was used in order to control the shape of the DVHs. Three segments per beam were used, the first segment must be conformal, and the leaves were required to close in. Minimum segment dimension was 20 mm and minimum equivalent square was 30 mm. The isocenter was positioned at the geometric center of PTV3. For comparison purposes, a plan similar to that used routinely for this type of treatment at this Center was also created. This plan consisted of three fields with gantry angles 0°, 90°, and 270°. Each field consisted of two segments, the first conforming to the beam’s eye view of PTV1 and the second to the beam’s eye view of PTV3, with a fixed penumbra margin of 6 mm. For the larger three segments, collimator angles were 0°, 270°, and 90°, respectively, and wedges were used on the lateral fields. For the smaller three segments, all collimator angles were 0° and no wedges were used. For consistency, AUTOPLAN was used for this plan also, but only the segment weights and wedge angles were optimized. Finally, an IMRT plan was created with nine segments per beam, no constraint

### Table I. Clinical objectives and constraints used for optimization of the prostate and seminal vesicles case.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Objective</th>
<th>Optimization Objective</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTV1</td>
<td>$V_{56%} &gt; 99%$</td>
<td>Maximize PTV1 minimum dose with weight 9</td>
</tr>
<tr>
<td>PTV2</td>
<td>$V_{67%} &gt; 99%$</td>
<td>Maximize PTV2 minimum dose with weight 60</td>
</tr>
<tr>
<td>PTV3</td>
<td>$V_{70%} &gt; 99%$</td>
<td>Minimize PTV3 RMS dose around 74 Gy with weight 60</td>
</tr>
<tr>
<td>Rectum</td>
<td>$V_{50%} &lt; 60%$</td>
<td>Minimize rectum volume irradiated to 50 Gy with weight 1</td>
</tr>
<tr>
<td>Bladder</td>
<td>$V_{60%} &lt; 50%$</td>
<td>Minimize rectum volume irradiated to 60 Gy with weight 2</td>
</tr>
<tr>
<td>Femoral heads</td>
<td>$V_{65%} &lt; 30%$</td>
<td>Minimize rectum volume irradiated to 65 Gy with weight 2</td>
</tr>
<tr>
<td></td>
<td>$V_{50%} &lt; 50%$</td>
<td>Minimize bladder volume irradiated to 60 Gy with weight 2</td>
</tr>
<tr>
<td></td>
<td>$V_{60%} &lt; 25%$</td>
<td>Minimize bladder volume irradiated to 65 Gy with weight 1</td>
</tr>
<tr>
<td></td>
<td>$V_{52%} &lt; 10%$</td>
<td>Femoral head volume irradiated to 52 Gy &lt; 10% (constraint)</td>
</tr>
</tbody>
</table>

I. Clinical case: Prostate and pelvic nodes

This Center is currently carrying out a randomized clinical trial of radiotherapy for prostate and pelvic nodes. The trial aims to simultaneously deliver 70 Gy to the prostate and 50-60 Gy to the pelvic nodes, in 35 fractions. The use of the present algorithm is demonstrated for one of these patients.

The primary CTV (CTV1) consisted of the prostate and base of seminal vesicles and any further involved region of the seminal vesicles. The remaining seminal vesicles and the pelvic lymph nodes up to the sacral promontory formed the secondary CTV (CTV2). The primary PTV (PTV1) consisted of CTV1 plus a 10 mm margin in all directions except posteriorly, where the margin was 5 mm. The secondary PTV (PTV2) consisted of CTV2 plus a 5 mm margin. The patient was treated in the supine position.

Five 6 MV fields were used, at gantry angles 180°, 260°, 340°, 20°, and 100°. The clinical objectives are shown in Table II. For the bowel, the objectives were originally expressed in cubic centimeters so as to be applicable to all patients, but for consistency, these absolute volumes were converted to percentages based upon the total volume of bowel in this particular patient. The objectives given for bowel were designed to give no more than grade 1 toxicity. Four segments per beam were used, the first segment must be conformal, and there was no requirement for the leaves to close in. Minimum segment dimension was 30 mm and minimum equivalent square was 40 mm. The isocenter was positioned at the geometric center of PTV1. Since it was known that the left anterior oblique field should ideally treat the left nodes and the right anterior oblique field should treat the right nodes,15 the left anterior oblique field was confined to the envelope of PTV1 and the left side of PTV2 and the right anterior oblique field was confined to PTV1 and the right side of PTV2. The plan was calculated for 50 Gy and the more challenging case of 60 Gy to PTV2. The prescribed dose to PTV1 was 70 Gy in both cases.
Table II. Clinical objectives and constraints used for optimization of the prostate and pelvic nodes case.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Objective</th>
<th>Optimization Objective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate PTV (PTV1)</td>
<td>( V_{95%} &gt; 99% )</td>
<td>Minimize PTV1 RMS dose around 70 Gy with weight 800</td>
</tr>
<tr>
<td></td>
<td>( V_{95%} &gt; 95% )</td>
<td></td>
</tr>
<tr>
<td>Nodal PTV (PTV2)</td>
<td>( V_{99%} &gt; 99% )</td>
<td>Maximize PTV2 (excluding PTV1) RMS dose around 50 Gy with weight 100</td>
</tr>
<tr>
<td></td>
<td>( V_{95%} &gt; 95% )</td>
<td>Minimize PTV2 volume irradiated to 45 Gy with weight 100</td>
</tr>
<tr>
<td>Bowel</td>
<td>( V_{50 \text{ Gy}} &lt; 3% )</td>
<td>Minimize PTV2 volume irradiated to 55 Gy with weight 30</td>
</tr>
<tr>
<td></td>
<td>( V_{50 \text{ Gy}} &lt; 2% )</td>
<td></td>
</tr>
<tr>
<td>Bladder</td>
<td>( V_{50 \text{ Gy}} &lt; 50% )</td>
<td>Minimize bladder volume irradiated to 55 Gy with weight 10</td>
</tr>
<tr>
<td></td>
<td>( V_{50 \text{ Gy}} &lt; 25% )</td>
<td></td>
</tr>
<tr>
<td>Rectum</td>
<td>( V_{50 \text{ Gy}} &lt; 30% )</td>
<td>Minimize rectum volume irradiated to 50 Gy with weight 50</td>
</tr>
<tr>
<td></td>
<td>( V_{70 \text{ Gy}} &lt; 15% )</td>
<td>Minimize rectum volume irradiated to 65 Gy with weight 50</td>
</tr>
</tbody>
</table>

### III. RESULTS

#### A. Phantom case

The progress of optimization is shown for the phantom plan with three segments per beam in Fig. 6. As the process begins with a conformal plan, the constraint that the mean OR dose should be less than 25 Gy is not initially met, so the objective function is artificially elevated to signify this. As the solution reduces the mean OR dose, the objective function gradually falls. As soon as the solution satisfies the OR constraint, the objective function falls to a lower level. Thereafter, the objective function values fall into one of two levels, one signifying that the OR constraint has been met, the other signifying that it has not. As the optimization probes around the constraint boundary, either of these objective function outcomes may result, giving the two-level appearance of the objective function plot. After 3000 iterations, the voxel size of the patient is reduced and with the more accurate dose calculation, it is found that the OR constraint is not met, so the optimum value of the objective function rises. However, the plan, based upon the coarse patient grid, is close to optimum, so with very few adjustments, the constraints are again satisfied. The optimum objective function therefore falls again. The plateau region at the end of the optimization indicates that an optimum solution has been found. To verify this, the optimization has also been run for ten times more iterations, and the solution has not significantly improved.

Segment shapes for one of the beams are shown in Fig. 7 for the case of three segments per beam. It can be seen that the segment shapes are simple and regular, focusing on the part of the PTV which can be irradiated without irradiating the OR. Consequently, they provide a dose distribution which covers the PTV while sparing the OR, even though there are only a few segments per beam. The apertures are not equal in width all along the length of the phantom, even though the phantom is uniform. This does not appear to compromise the performance of the algorithm, and is thought to result from degeneracy between the apertures of the different beams, i.e., a change in aperture width for one beam is compensated by a corresponding change in one or more other beams. For the PTV, the root-mean-square dose variation is 12.0% for one segment per beam, 9.7% for three segments per beam, 6.6% for five segments per beam, and 6.4% for seven segments per beam, with mean OR dose limited to 25 Gy. The dose-volume histograms for PTV and OR are shown in Fig. 8(a) for one, three, and five segments per beam. Increasing the number of segments gradually improves the PTV dose homogeneity for equivalent OR mean dose, as observed by Cotrutz and Xing. No further improvement is seen beyond five segments per beam. In Fig. 8(b), the DVHs for five segments per beam are compared with those for nine unconstrained segments per beam. The PTV coverage for nine unconstrained segments per beam is marginally better than for five constrained segments per beam, with a root-mean-square variation of 5.4%. This difference represents the effect of the constraints on the IMRT plans. Also shown in Fig. 8(c) is a comparison of DVHs for five segments per beam, with and without the use of the constraint that the leaves must close in. It can be seen that the constraint reduces the efficacy of the treatment slightly, the root-mean-square PTV dose variation increasing to 9.9%. This is to be expected, as this case is expected to produce

![Graph](image)
fluence profiles with more than one maximum. For example, the anterior field might direct fluence either side of the OR.

B. Clinical case: Rectum

Typical segment shapes and dose volume histograms are shown for the rectum case in Fig. 9. Again it can be seen that the method provides effective coverage of the PTV, using segments that can be delivered with confidence. Note that although the bladder is included in Fig. 9 for reference, this does not form part of the optimization, as the conformal shaping of the lateral beams inherently spares the bladder. The final segment [segment 4, Fig. 9(d)] is an example of a diagonal field where the minimum equivalent square is needed. The minimum dimension has been satisfied, but the actual exposed area is narrow. The minimum equivalent square prevents the exposed area from becoming unreasonable. The long narrow segment is at about the limit of what can be treated with confidence.

C. Clinical case: Prostate and seminal vesicles

The shape of the segments for the left anterior oblique field are shown in Fig. 10. The requirement that the leaves should close in provides a natural and efficient means of effecting the coning down of the segments in accord with the progressively tightening PTV margins. The resulting dose-volume histograms are also shown in Fig. 11. Compared to the conformal plan [Fig. 11(a)], the IMRT plan provides equivalent PTV coverage. The rectal dose is similar for the two plans, but the bladder dose is significantly lower for the IMRT plan. The femoral head dose is lower for the IMRT plan, but this is likely to be due to the use of five fields for
the IMRT plan as opposed to three for the standard conformal plan. Compared to the unconstrained IMRT plan with nine segments per beam [Fig. 11(b)], the constrained IMRT plan has slightly improved PTV coverage and consequently slightly higher bladder and rectal dose. This difference is primarily due to the difficulty of producing exactly equivalent plans. The unconstrained IMRT plan is not significantly worse in quality, showing that the benefits of constrained segment shaping can be achieved without detriment to the clinical benefit of the plan.

D. Clinical case: Prostate and pelvic nodes

The shapes of the segments for the right anterior oblique field with 50 Gy prescription to PTV2 are shown in Fig. 12. Note that the segments never expose PTV2 entirely as they are constrained to treat the right side of PTV2 only. The segments focus on the part of PTV2 not overlapping bowel or bladder. A boost is given by this field to PTV1 so as to provide the additional dose required by this PTV. Transaxial, sagittal, and coronal dose distributions, recalculated in Pinnacle\textsuperscript{3} (Philips Radiation Oncology Systems, Milpitas, CA), are shown in Fig. 13, where the higher dose to PTV1 and the concavity between the pelvic nodes can be clearly
seen. Dose-volume histograms for both 50 and 60 Gy prescriptions are shown in Fig. 14. For the 50 Gy prescription, all of the objectives have been reached. For the 60 Gy prescription, the objectives for the PTVs, rectum, and femoral heads have been met. However, the volumes of bladder and bowel irradiated to 50–60 Gy are slightly high (bladder $V_{50\text{ Gy}}$ 56% vs 50%, bladder $V_{60\text{ Gy}}$ 33% vs 25%, bowel $V_{55\text{ Gy}}$ 23% vs 8%, bowel $V_{60\text{ Gy}}$ 5% vs 2%). These irradiated volumes can only be reduced at the expense of PTV coverage.

IV. DISCUSSION

All of the presented cases demonstrate that direct-aperture optimization with constrained segment shapes provides effective IMRT dose distributions with apertures which are as regular in shape as possible. The effectiveness of the algorithm is indicated by the almost equivalent performance to the unconstrained plans but with fewer, more regular segments. The value of using simple segments is that the planning system can reliably calculate the dose distribution. Verification checks should then agree more closely with the predicted dose. The overall result is that confidence in the ability of the treatment-planning system to accurately calculate the resulting dose can be quickly built up. Effective, reliable IMRT may therefore be delivered with the least requirement for elaborate verification. In particular, the verification of the fluence profiles of individual beams may be
reduced. The extensive verification which accompanies IMRT treatments\textsuperscript{29–31} can therefore be minimized as soon as possible.\textsuperscript{32}

The range of cases presented shows that this method is ideally suited to both simple IMRT techniques consisting of a few segments, perhaps simulating a wedge, and to more complex techniques involving multiple, highly concave, PTVs. Whatever the complexity of the case, the same principle of achieving the most effective solution with the simplest set of apertures is applicable. The disadvantage of the method presented is that it is much slower than traditional fluence optimizations, which are usually accomplished by means of a fast gradient-descent method. However, when time has been allowed for segmentation following fluence optimization, which sometimes requires several runs to achieve the desired outcome, the present method is not so disadvantageous.

There may also be a tradeoff when using few segments for IMRT delivery. Although the direct-aperture-optimization method described in this paper is able to provide the most effective treatment plan with the allocated segments and within the delivery constraints, it may be possible to over-constrain the algorithm so that the optimum dose distribution cannot be found. The goal is to use the fewest possible segments, with the simplest possible aperture shapes, but without unduly limiting the quality of the treatment plan. This tradeoff is not yet fully understood, and further work is planned to explore the minimum number of segments and the lower limit on the segment size constraints that can be used to obtain a clinically effective treatment plan while minimizing the number of match lines.

By minimizing the number of segments in the IMRT plan, artifacts such as tongue-and-groove effect are minimized. In

Fig. 14. Results for the prostate and pelvic nodes case: dose-volume histograms for (a) 50 Gy prescription to PTV2, and (b) 60 Gy prescription to PTV2. RF: right femoral head; LF: left femoral head.
particular, tongue-and-groove effect is eliminated by the option to require that each segment lies within the previous segment. Even if this is not selected, the apertures are rarely abutting and therefore the effect is minimal. For this reason, the tongue-and-groove design of the MLC leaves has not been modeled in this study. However, should it be required in future, the appropriate approach is to model the tongue-and-groove behavior of the MLC and then leave the optimization to eliminate the underdosage naturally according to its effect on the overall dose distribution.

The optimization of segment shapes is a subset of the global IMRT optimization problem, involving in its completeness, the optimization of radiation modality, beam energy, and beam orientations. Accordingly, the algorithm is expected to be further developed into IMRT beam orientation optimization. The segment shape optimization, together with its nested module for segment weight optimization, have been designed to be conveniently nested within higher level modules for optimization of gantry, couch, and collimator angle. Although the algorithm will have to be reduced so as to provide sufficient speed for this application, it will have the advantage of providing realistic segmentation at each stage of the optimization. Thus, the final outcome of orientation optimization will be reliable because the solution for each specific orientation examined will also be realistic.

V. CONCLUSION

A method of direct-aperture optimization using constrained segment shapes has been presented and the performance of the algorithm demonstrated for simple and complex phantom and clinical cases. The algorithm provides apertures which are as regular as possible, due to the use of a cubic shaping function and constraints on minimum aperture size and equivalent square. There is also an option to specify that each segment should lie within the previous one. A method of feedback optimization is described. This method increases the efficiency of the search for appropriate segments, particularly when each leaf lies within the previous one. The demonstrations show that the algorithm provides simple, effective segmental IMRT solutions.

ACKNOWLEDGMENTS

This work has been supported by Cancer Research UK (Grant Reference SP2312/0201), Philips Radiation Oncology Systems and the National Cancer Research Institute, to whom we are grateful. We also thank A. P. Warrington for his supervision of this work.

48Electronic mail: james.bedford@icr.ac.uk
50S. Webb, Contemporary IMRT: Developing Physics and Clinical Applications (IOP, Bristol, 2005).