

Instituto de Engenharia de Sistemas e Computadores de Coimbra
Institute of Systems Engineering and Computers
INESC - Coimbra

Humberto Rocha Joana Matos Dias
Brígida da Costa Ferreira Maria do Carmo Lopes

**From fluence map optimization to fluence map delivery:
the role of combinatorial optimization**

No. 5

2011

ISSN: 1645-2631

Instituto de Engenharia de Sistemas e Computadores de Coimbra
INESC - Coimbra
Rua Antero de Quental, 199; 3000-033 Coimbra; Portugal
www.inescc.pt



From fluence map optimization to fluence map delivery: the role of combinatorial optimization

H. Rocha ^{*} J. M. Dias ^{*,†} B.C. Ferreira ^{§,‡} M.C. Lopes [§]

May 5, 2011

Abstract

The intensity modulated radiation therapy (IMRT) treatment planning problem is usually divided in three smaller problems that are solved sequentially: geometry problem, intensity problem, and realization problem. That division has the consequence of causing a plan quality deterioration arising from the transition between the intensity problem and the realization problem. Typically, on the beamlet-based approach, after the optimal beamlet intensities are determined, they are discretized over a range of values using a distance criteria (rounding). However, that decision criteria is not appropriate and we present empirical evidence that this can lead to a significant deterioration of the treatment plan quality regardless of the model used to tackle the intensity problem. We propose a combinatorial optimization approach and a probabilistic binary tabu search method to enable an improved transition from optimized to delivery fluence maps in IMRT by minimizing the deterioration of the treatment plan quality and improving organ sparing at the same time. Two head & neck clinical examples were used to test the ability of the proposed formulation and resolution method to obtain improved plans compared to the usual rounding procedure. The results obtained present a clear improvement of the treatment plan quality both in terms of target coverage and also in terms of parotid sparing.

Key words. IMRT; fluence map optimization; combinatorial optimization.

^{*}*INESC-Coimbra, Coimbra, Portugal.*

[†]*Faculdade de Economia, Universidade de Coimbra, Coimbra, Portugal.*

[‡]*I3N, Departamento de Física, Universidade de Aveiro, Aveiro, Portugal.*

[§]*Serviço de Física Médica, IPOC-FG, EPE, Coimbra, Portugal.*

1 Introduction

The goal of radiation therapy is to deliver a dose of radiation to the cancerous region to sterilize the tumor minimizing the damages on the surrounding healthy organs and tissues. Radiation therapy is based on the fact that cancerous cells are focused on fast reproduction and are not as able to repair themselves when damaged by radiation as healthy cells. Therefore, the goal of the treatment is to deliver enough radiation to kill the cancerous cells but not so much that jeopardizes the ability of healthy cells to survive.

In the inverse planning of the radiation treatment plan, for a prescribed treatment plan, a correspondent set of parameters (beams and fluences) is algorithmically computed in order to fulfil the prescribed doses and restrictions. Inverse treatment planning allows the modeling of highly complex treatment planning problems and optimization has a fundamental role in the success of this procedure. An important type of inverse treatment planning is intensity modulated radiation therapy (IMRT) where the radiation beam is modulated by a multileaf collimator as illustrated in Figure 1(a). Multileaf collimators (MLC) enable the transformation of the beam into a grid of smaller beamlets of independent intensities (see Figure 1(b)). Despite the illustration of Figure 1(b), beamlets do not exist physically. Their existence is generated by the movement of the leaves of the MLC in Figure 1(a) that block part of the beam during portions of the delivery time. The MLC has movable leaves on both sides that can be positioned at any beamlet grid boundary. MLC can operate in two distinct ways: dynamic collimation or multiple static collimation. In the first case, the leaves move continuously during irradiation. In the second case, the “step and shoot mode”, the leaves are set to open a desired aperture during each segment of the delivery and radiation is on for a specific fluence time or intensity. This procedure generates a discrete set (the set of chosen beam angles) of intensity maps like in Figure 1(b). Here, one will consider multiple static collimation.

A common way to solve the inverse planning in IMRT optimization problems is to use a beamlet-based approach. For optimization purposes, each structure’s volume is discretized in voxels (volume elements). Each voxel in a structure is identified by a three dimensional coordinate (x, y, z) . Let us assume that there are $m \times n$ beamlets identified by the index pair (p, q) . The weight (intensity) of the beamlet (p, q) delivered over an angle θ is defined by $w(\theta, p, q)$. Using the superposition principle, the total dose, $D(x, y, z)$, that a voxel

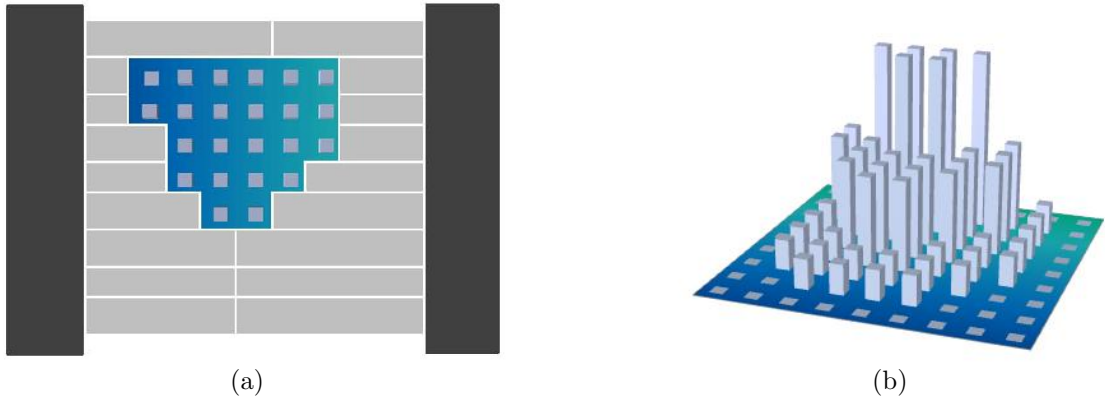


Figure 1: Illustration of a multileaf collimator (with 9 pairs of leaves) – 1(a) and illustration of a beamlet intensity map (9×9) – 1(b).

(x, y, z) receives is

$$D(x, y, z) = \sum_{(\theta, p, q)} w(\theta, p, q) \cdot d_{(\theta, p, q)}(x, y, z), \quad (1)$$

where $d_{(\theta, p, q)}(x, y, z)$ is the dose delivered to voxel (x, y, z) by beamlet (p, q) from angle θ . This beamlet-based approach leads to a large-scale programming problem with thousands of variables (beamlets) and hundreds of thousands of (dose-volume) constraints. The quality of the plan, depending on the programming models and the solution methods, determines the clinical treatment effect. Due to the complexity of the whole optimization problem, the treatment planning is usually divided into three smaller problems which can be solved separately: geometry problem, intensity problem, and realization problem.

The geometry problem consists in finding the minimum number of beams and corresponding directions that satisfy the treatment goals using optimization algorithms [6, 8, 18]. In clinical practice, most of the times, the number of beams is assumed to be defined a priori by the treatment planner and the beam directions are still manually selected by the treatment planner that relies mostly on his experience. After deciding what beam angles should be used, an optimal plan is obtained by solving the intensity (or fluence map) problem - the problem of determining the optimal beamlet weights for the fixed beam angles. Therefore, the linkage between the geometry problem and the intensity problem is straightforward, since the angle values given by the geometry problem are an input to the intensity problem (regardless of whether they are computed or manually selected by the treatment planner).

Many mathematical optimization models and algorithms have been proposed for the intensity problem, including linear models (e.g. [21, 22]), mixed integer linear models (e.g. [13, 17]), nonlinear models (e.g. [2, 25]), and multiobjective models (e.g. [23, 26]). The outcome of the intensity problem is a set of optimized “continuous” fluence maps (one for each beam angle). However, in order to deliver the optimized fluence in the “step and shoot mode”, the beamlet intensities must be discretized by levels of intensity. Typically, the beamlet intensities are discretized over a range of values (0 to 5, e.g.) and the resulting fluence maps are the input of the realization problem. This discretization process is the linkage between intensity and realization problem and is the core subject of this research report.

After an acceptable set of intensity maps is produced and discretized, one must find a suitable way for delivery (realization problem) using one of the many existing techniques ([1, 3, 21, 24]) to construct the apertures and intensities that approximately match the intensity maps previously determined. However, reproducing the optimized discrete intensity maps efficiently, i.e., minimizing the radiation exposure time, is a challenging optimization problem. Moreover, we need to find an efficient way for MLC devices to produce the exact same optimized discrete intensity profiles. Due to leaf collision issues, leaf perturbation of adjacent beamlet intensities, tongue-and-groove constraints, etc., the intensity maps actually delivered may be different from the optimized discrete ones. Those problems have been tackled ([11], e.g.) and are still a prosperous field of research.

However, one of the main reasons for the deterioration of the treatment plan quality, is related to the simplistic linkage between the intensity problem and the realization problem, rather than only caused by the difficulties of segmentation during the realization problem related with the complexity of the fluence map or physical restrictions of the MLC. This subject is poorly documented in the literature (see [19, 20], e.g.) and the general idea transmitted is that deterioration is mainly caused by segmentation issues.

Some approaches choose to optimize directly apertures and fluences instead of solving the realization problem after the intensity problem. By solving both problems at once, one has the guarantee that solutions are always feasible for delivery without any post-processing. The column generation approach [17, 21] belongs to this type of approaches since the generated columns can be restricted to correspond to feasible hardware

settings. An advantage of the column generation approach to the other IMRT optimization approaches is that the complexity of the treatment plan can be controlled through the number of columns. This approach allows us to investigate the non-trivial trade-off between plan quality and treatment complexity. Another approach that belong to this group of approaches is the direct aperture optimization (DAO) scheme ([24]). DAO and other variations, namely direct parameter machine optimization (DMPO), are integrated in the latest generation of planning systems along with the beamlet-based approaches. However, no approach have proved to be better than the other. Therefore, it continues to be of the utmost interest to improve the transition from the intensity problem to the realization problem for the beamlet-based approach.

The objective of this research report is twofold. First, to present empirical evidence of the plan's quality deterioration caused by the usual transition from optimized to delivery fluence maps in IMRT treatment planning regardless of using linear or nonlinear models to model the intensity map optimization problem. Second, to propose a combinatorial optimization approach that enables an efficient transition from optimized to delivery fluence maps in IMRT treatment planning minimizing the plan deterioration while improving organ sparing. Two clinical examples of head & neck cancer cases are used to present numerical evidence of the resulting deterioration of plan quality using the usual rounding procedure and to highlight the advantages of the proposed combinatorial optimization.

The research report is organized as follows. In the next section we describe and use two clinical examples of head & neck cancer cases to show numerical evidence of the deterioration of the plan quality resulting from the transition of the intensity problem to the realization problem using the usual rounding procedure. Both linear and nonlinear models are formulated and used to highlight that the deterioration occurs regardless of the model used to address the intensity problem. In section 3, we propose a combinatorial optimization approach and a probabilistic binary tabu search method to enable an improved transition from optimized to delivery fluence maps in IMRT by minimizing the deterioration of the treatment plan quality and improving organ sparing at the same time. In the 4th section, numerical results of the proposed formulation and resolution method are presented for the two head & neck clinical examples. In the last section we have the concluding remarks.

2 Illustration of the plan’s quality deterioration using linear and nonlinear models

The process of converting an optimal fluence map into a set of MLC segments is called segmentation. Segmentation needs to receive as input integer matrices, that are obtained by the discretization of each beamlet intensity over a range of values. The outcome of the intensity problem is a set of optimal fluence maps (one for each fixed beam) that can be represented by real matrices of $m \times n$ beamlet weights (intensity assigned to each beamlet), i.e., there are m leaf pairs and for each leaf there are $n+1$ possible positions. These matrices, solutions of the intensity problem, cannot be directly implemented, because of hardware constraints. The matrices have to be transformed to accommodate hardware settings, with a resulting degradation of the plan quality. This discretization is one of the main causes for deterioration of plan quality.

It is important to remark that this rounding procedure is the usual procedure of many treatment planning systems such as KonRad ([12], section 7.3.4). In “step & shoot” (static) delivery mode, the user can specify the number of intensity levels. The calculated continuous fluence distributions for each beam direction are converted into equidistant intensity levels. In Figure 2 we have the optimal fluence for a beam obtained by KonRad and the correspondent deliverable fluence after rounding using 7 levels.

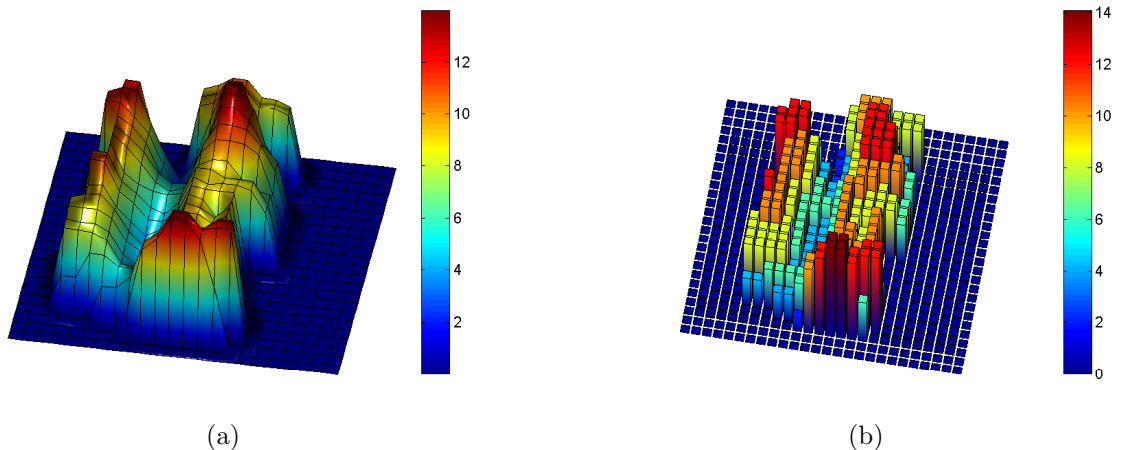


Figure 2: Ideal theoretic fluence – 2(a) and deliverable fluence – 2(b) for a beam using KonRad.

An illustration of the usual way this transition is done, and the consequent treatment plan deterioration, is presented using two head & neck clinical examples.

2.1 Head & neck clinical examples

Two clinical examples of retrospective treated cases of head and neck tumors at the Portuguese Institute of Oncology of Coimbra are used to verify the deterioration caused by the rounding of the optimal fluence maps. The patients' CT sets and delineated structures were exported via Dicom RT to a freeware computational environment for radiotherapy research. In general, the head & neck region is a complex area to treat with radiotherapy due to the large number of sensitive organs in this region (e.g. eyes, mandible, larynx, oral cavity, etc.). For simplicity, in this study, the OARs used for treatment optimization were limited to the spinal cord, the brainstem and the parotid glands. The tumor to be treated plus some safety margins is called planning target volume (PTV). For the head & neck cases in study it was separated in two parts: PTV left and PTV right (see Figure 3). The prescribed doses for all the structures considered in the optimization are presented in Table 1.

Structure	Mean Dose	Max Dose	Prescribed Dose
Spinal cord	–	45 Gy	–
Brainstem	–	54 Gy	–
Left parotid	26 Gy	–	–
Right parotid	26 Gy	–	–
PTV left	–	–	59.4 Gy
PTV right	–	–	50.4 Gy
Body	–	70 Gy	–

Table 1: Prescribed doses for all the structures considered for IMRT optimization.

In order to facilitate convenient access, visualization and analysis of patient treatment planning data, the computational tools developed within Matlab [16] and CERR [7] (computational environment for radiotherapy research) were used as the main software platform to embody our optimization research and provide the necessary dosimetry data to perform

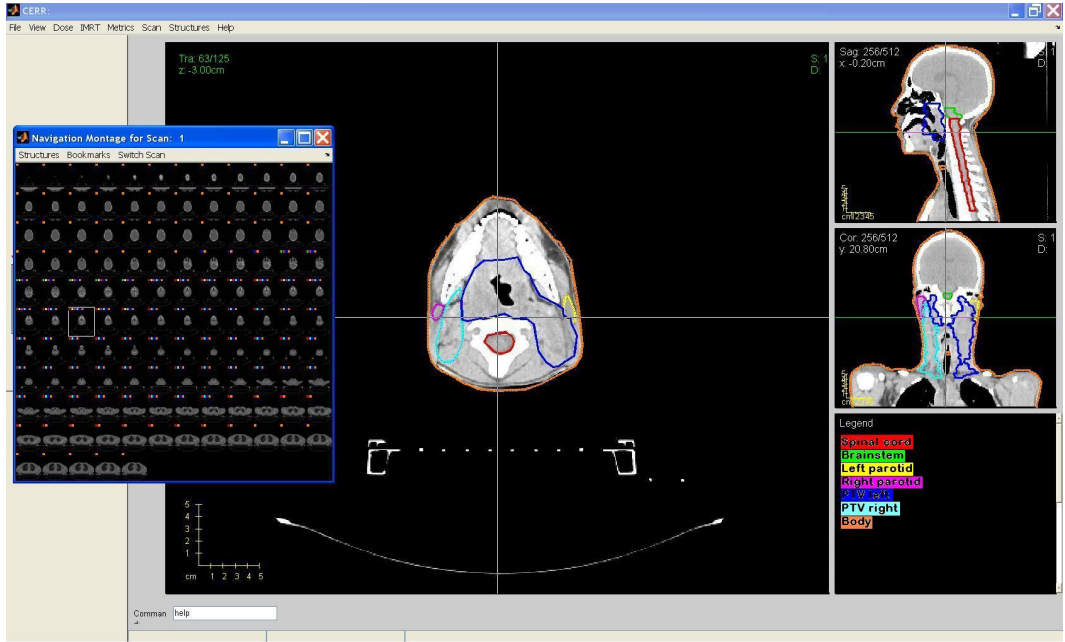


Figure 3: Structures considered in the IMRT optimization visualized in CERR.

optimization in IMRT.

2.2 Fluence map optimization models

The outcome of the fluence map optimization is a set of optimal fluence maps (one for each fixed beam) that can be represented by real matrices of $m \times n$ beamlet weights (intensity assigned to each beamlet), i.e., there are m leaf pairs and for each leaf there are $n + 1$ possible positions. In order to convert an optimal fluence map into a set of MLC segments we need to transform the optimal real matrices into integer matrices, that are obtained by the discretization of each beamlet intensity over a range of values. We might think that, in general, the degradation of the treatment quality will be smaller when converting smoother fluence maps. Since typically nonlinear models produce smoother fluence maps compared to linear models, two different models were used to perform IMRT optimization: a linear model and a convex penalty function nonlinear model.

The first attempts to tackle the intensity problem used linear models, due to the fact that dose deposition is linear, they are easy to implement and are broadly used. Most of the formulations of the linear models belong to a class of constrained optimization models such that an objective function is optimized while meeting these dose requirements (see

[14], e.g.). We choose to minimize the mean dose of the organs at risk and normal tissue (body) while imposing hard constraints for the maximum organs at risk doses and for the minimum and maximum targets dose to ensure that enough radiation is delivered to fulfill the dose prescription. The mean dose of a given structure is the sum of the dose for all voxels of the structure divided by the number of voxels of the structure. The linear model used for this study was

$$\begin{aligned}
\min_w \quad & \frac{1}{\text{card}(OAR)} \cdot \sum_{(x,y,z) \in OAR} D_{OAR}(x,y,z) + \frac{1}{\text{card}(NT)} \cdot \sum_{(x,y,z) \in NT} D_{NT}(x,y,z) \\
s.t. \quad & D(x,y,z) = \sum_{(\theta,p,q)} w(\theta,p,q) \cdot d_{(\theta,p,q)}(x,y,z), \forall (x,y,z) \in PTV \cup OAR \cup NT \\
& LB_{PTV} \leq D_{PTV}(x,y,z) \leq UB_{PTV}, \forall (x,y,z) \in PTV, \\
& D_{OAR}(x,y,z) \leq UB_{OAR}, \forall (x,y,z) \in OAR, \\
& D_{NT}(x,y,z) \leq UB_{NT}, \forall (x,y,z) \in NT, \\
& 0 \leq w(\theta,p,q) \leq M, \quad \forall \theta, p = 1, \dots, m, \\
& \quad \quad \quad q = 1, \dots, n,
\end{aligned}$$

where LB_{PTV} and UB_{PTV} are the lower and upper bounds for the planning target volume (PTV) dose, D_{PTV} , UB_{OAR} are the upper bounds for the organs at risk (OAR) doses, D_{OAR} , UB_{NT} is the upper bound for the normal tissue (NT) dose, D_{NT} , and M is the upper bound for all the beamlet weights w .

Similar linear models were used before [6]. Other used models include convex penalty function structure-based [15] or voxel-based [2] approaches. Here, we will use as well, a convex penalty function voxel-based nonlinear model similar to the one used in Aleman et al. [2]. In this model, each voxel is penalized according to the square difference of the amount of dose received by the voxel and the amount of dose desired/allowed for the voxel. The total dose received by the voxel (x,y,z) is $D(x,y,z) = \sum_{(\theta,p,q)} w(\theta,p,q) \cdot d_{(\theta,p,q)}(x,y,z)$, as seen in Eq. 1. Typically, a dose matrix D is constructed from the collection of all beamlet weights, by indexing the rows of D to each voxel (x,y,z) and the columns to each beamlet, i.e., the number of rows of matrix D equals the number of voxels (V) and the number of columns equals the number of beamlets (N) from all angles considered. Therefore, using matrix format, we can say that the total dose received by the voxel i is given by $\sum_{j=1}^N D_{ij} w_j$. This formulation yields a quadratic programming problem with only linear non-negativity

constraints on the fluence values:

$$\min_w \sum_{i=1}^V F_i \left(\sum_{j=1}^N D_{ij} w_j \right) \tag{2}$$

$$s.t. \quad w_j \geq 0, \quad j = 1, \dots, N.$$

Similarly to Romeijn et al. [22] we use asymmetric quadratic penalty functions:

$$F_i \left(\sum_{j=1}^N D_{ij} w_j \right) = \frac{1}{v_S} \left[\underline{\lambda}_i \left(T_i - \sum_{j=1}^N D_{ij} w_j \right)_+^2 + \bar{\lambda}_i \left(\sum_{j=1}^N D_{ij} w_j - T_i \right)_+^2 \right],$$

where T_i is the desired dose for voxel i , $\underline{\lambda}_i$ and $\bar{\lambda}_i$ are the penalty weights of underdose and overdose of voxel i , and $(\cdot)_+ = \max\{0, \cdot\}$. Although this formulation allows unique weights for each voxel, similarly to the implementation in Aleman et. al. [2], weights are assigned by structure only so that every voxel in a given structure has the weight assigned to that structure divided by the number of voxels of the structure (v_S). This nonlinear formulation imply that a very small amount of underdose or overdose may be accepted in clinical decision making, but larger deviations from the desired/allowed doses are decreasingly tolerated.

2.3 Illustration of the plan’s quality deterioration using head & neck clinical examples

Our tests were performed on a 2.66Ghz Intel Core Duo PC with 3 GB RAM. We used CERR 3.2.2 version and MATLAB 7.4.0 (R2007a). The dose was computed using CERR’s pencil beam algorithm (QIB) with seven equispaced beams in a coplanar arrangement, with angles 0, 51, 103, 154, 206, 257 and 309, and with 0 collimator angle. This equispaced beam angle configuration is the typically used in clinical practice. To address the linear problem we used one of the most effective commercial tools to solve large scale linear programs – Cplex[5]. We used a barrier algorithm (*baropt* solver of Cplex 10.0) to tackle our linear problem. To address the convex nonlinear formulation we used a trust-region-reflective algorithm (*fmincon*) of MATLAB 7.4.0 (R2007a) Optimization Toolbox.

In order to acknowledge the degree of plan quality deterioration, results obtained for the optimal fluence maps were compared with the fluence maps obtained after rounding

Level	level intensity	beamlet intensity range
0	0.0000	[0.0000 ; 1.8000)
1	3.6000	[1.8000 ; 5.4000)
2	7.2000	[5.4000 ; 9.0000)
3	10.800	[9.0000 ; 12.600)
4	14.400	[12.600 ; 16.200)
5	18.000	[16.200 ; 18.000]

Table 2: Beamlet distribution to correspondent intensity level for 5 levels.

optimal intensities using 5 intensity levels. In Table 2 we have the beamlet intensity range for each intensity level. By increasing the number of levels, the beamlet intensity range will decrease, which can lead to a decrease on the deterioration of the results. However, by increasing the number of levels, the segmentation problem will be more complex resulting in a less efficient delivery. In the best case scenario, there are no differences between the optimal intensities and the rounded intensities. However, for the worst case scenario, the difference between the optimal and the rounded intensity for each beamlet is 1.8.

In Figure 4, we have the resulting beamlet intensities for the two head & neck cancer cases considered using both the linear and the nonlinear formulations. By simple inspection, we can verify that the nonlinear formulation, and the corresponding solution method, originate smoother solutions compared to the solutions obtained by the linear programming approach. Nonetheless, the deterioration in the quality of the treatment plan when the usual rounding procedure is used is similar for both formulations and solutions.

The quality of the results can be perceived considering a variety of metrics and can change from patient to patient. Typically, results are judged by their cumulative dose-volume histogram (DVH). An ideal DVH for the tumor would present 100% volume for all dose values ranging from zero to the prescribed dose value and then drop immediately to zero, indicating that the whole target volume is treated exactly as prescribed. Ideally, the curves for the organs at risk would instead drop immediately to zero, meaning that no volume receives radiation.

In Figure 5, DVH curves are presented for optimal fluences obtained by the linear model and for the rounded optimal intensities when using 5 intensity levels. DVH curves for OARs

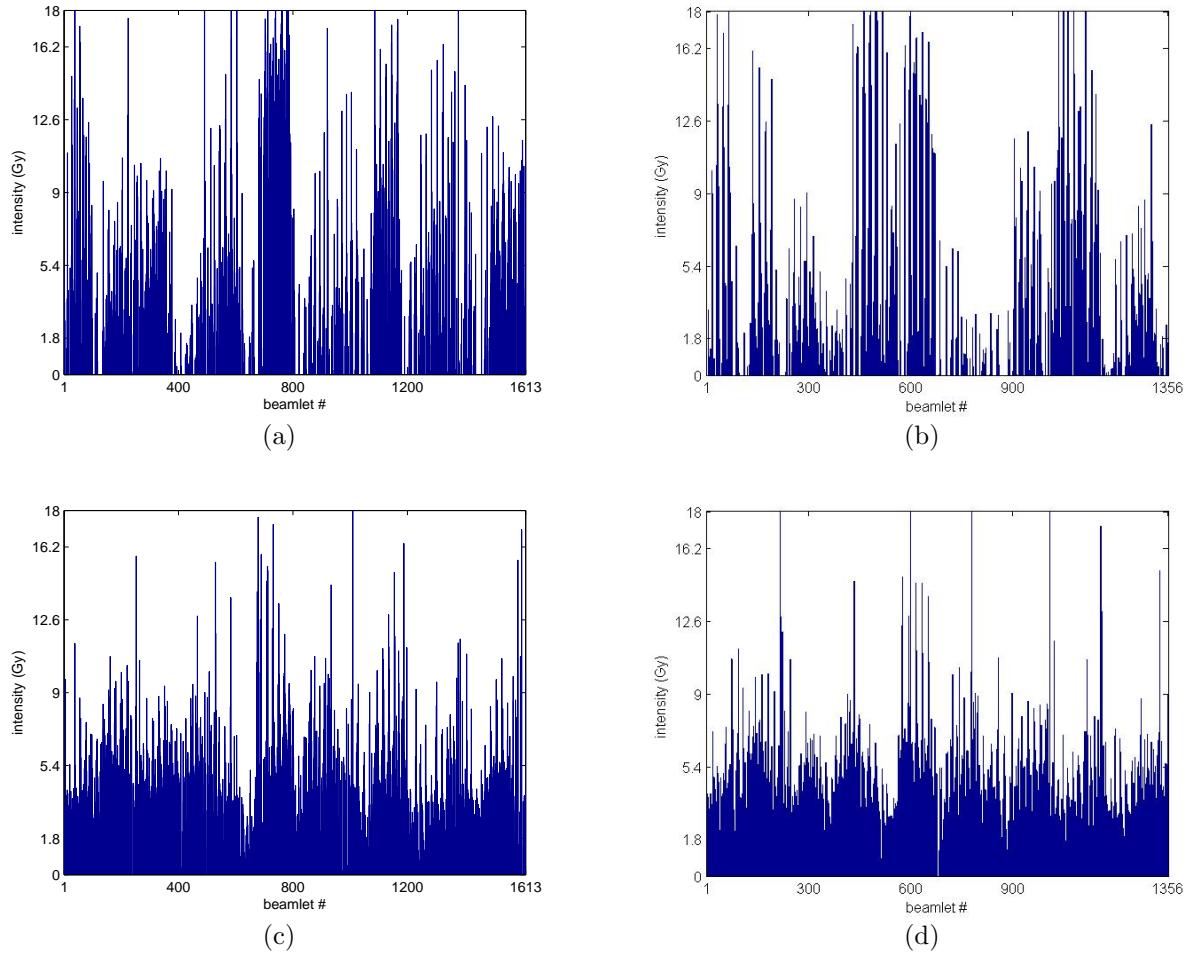


Figure 4: Beamlet intensities obtained by the linear formulation for the first head & neck case – 4(a) and for the second head & neck case – 4(b) and beamlet intensities obtained by the nonlinear formulation for the first head & neck case – 4(c) and for the second head & neck case – 4(d).

only suffer residual changes with the rounding procedure. However, by simple inspection of Figure 5, we can observe the deterioration of the PTVs dose-volume curves, consequence of the transition of the optimal fluence maps to the rounded ones, in both head & neck cases considered. Another metric usually used considers the prescribed dose that 95% of the volume of the PTV receives (D_{95}). Typically, 95% of the prescribed dose is required. D_{95} is represented in Figure 5 with an asterisk and we can observe that the rounded fluences fail to meet that quality criteria. Note that no segmentation was done and the observed deterioration is exclusively caused by the rounding of the optimal fluence maps. The fact

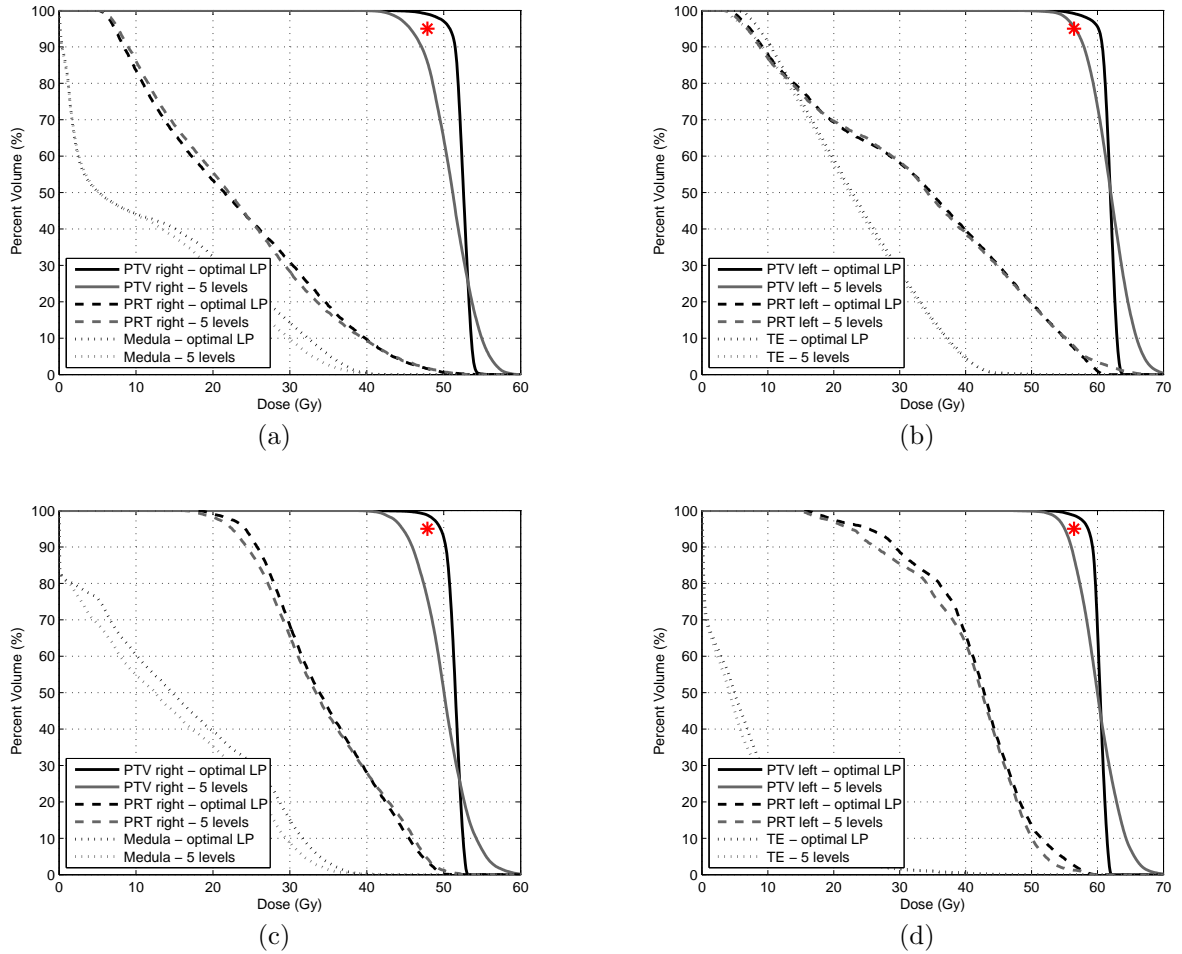


Figure 5: Cumulative dose volume histograms comparing the optimal beamlets obtained by the linear model (optimal LP) and the rounded optimal beamlets using 5 levels (5 levels) for right PTV, right parotid (PRT) and medula – 5(a) and for left PTV, left PRT and brainstem (TE) – 5(b) of the first head & neck case and for right PTV, right PRT and medula – 5(c) and for left PTV, left PRT and TE – 5(d) of the second head & neck case.

that the deterioration of the results affect mostly the PTVs is expected since only a small amount of beamlets reach the OARs while each beamlet was chosen so that it can irradiate the PTVs.

In Figure 6, DVH curves are presented for optimal fluences obtained by the nonlinear model and for the rounded optimal intensities when using 5 intensity levels. The results are very similar to the ones presented for the linear model. Therefore, this presents empirical evidence that, regardless the model used to tackle the intensity problem, the transition from

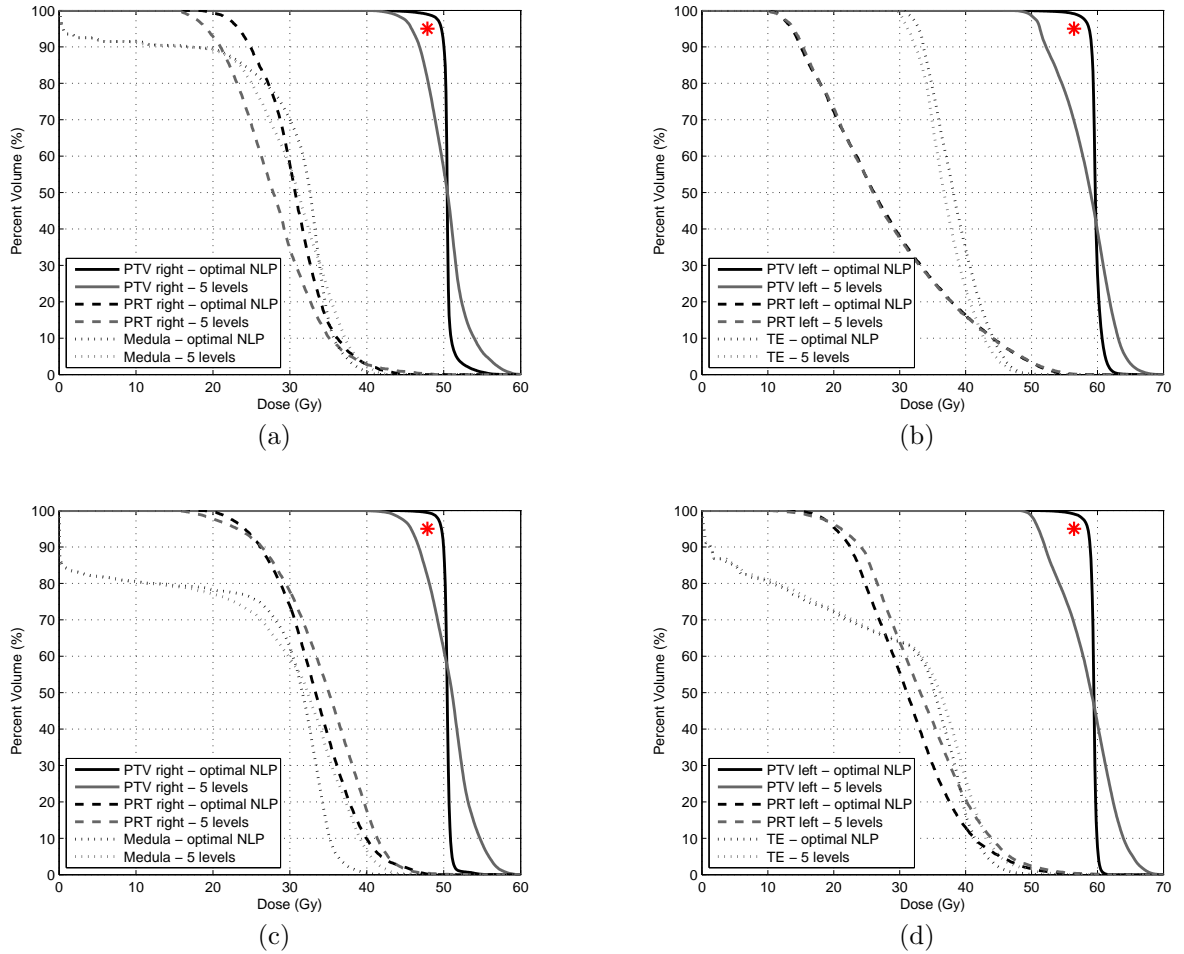


Figure 6: Cumulative dose volume histograms comparing the optimal beamlets obtained by the linear model (optimal LP) and the rounded optimal beamlets using 5 levels (5 levels) for right PTV, right parotid (PRT) and medula – 6(a) and for left PTV, left PRT and brainstem (TE) – 6(b) of the first head & neck case and for right PTV, right PRT and medula – 6(c) and for left PTV, left PRT and TE – 6(d) of the second head & neck case.

optimized to delivery fluence maps in IMRT treatment planning with the usual rounding procedure leads to a deterioration of the treatment plan quality.

3 Combinatorial optimization for improving the transition from optimized to delivery fluence maps

The transition from the optimal fluences to the discretized matrices that are used as inputs for the segmentation problem is almost never mentioned in the literature, and typically consists in a simple rounding of the optimal fluence (e.g., see [12]). Here, we will address it directly by using a combinatorial optimization approach.

3.1 Model formulations

After obtaining an optimal fluence map, and defining the number of levels of intensity to consider, we need to decide to which level of intensity each beamlet should be assigned to. The typical approach is to decide based on smaller distance and assign the level intensity closer to the optimal fluence (rounding). However, that decision criteria can put two beamlets with very close optimal intensities in distinct levels of intensity. Moreover, in such a complex large-scale optimization process, with such interdependence between beamlet intensity values, increasing or reducing the intensity of a beamlet should not be based on distance to closest intensity level. An alternative decision criteria is to decide between the two adjacent levels of the optimal beamlet intensity, based on a dose-volume response, rather than on a distance criteria.

If we use the convex nonlinear formulation (2) to model the fluence map optimization problem, we can decide based on the same criteria to which neighbor intensity level a beamlet intensity should be assigned to. Let x^{opt} denote the vector of the optimal beamlet intensities obtained in the end of the intensity problem. Let x^{round} denote the vector of the usual rounded intensities and let x^{trunc} denote the vector of the truncated intensities, i.e., the vector of the intensities corresponding to the smaller intensity value of the two neighbor level intensities. A straightforward combinatorial optimization formulation of deciding, based on the same criteria of the convex nonlinear formulation (2), to which neighbor intensity level a beamlet intensity should be assigned to, is the following:

$$\min \sum_{i=1}^V F_i \left(\sum_{j=1}^N D_{ij} \cdot (x^{trunc} + x \cdot L) \right)$$

s.t. x binary,

where L is the amplitude of an intensity level. The difference of intensity levels between x^{round} and x^{trunc} is a binary vector, where each 1 represents a choice of an upper level of intensity, and each 0 represents a choice of a lower intensity level. If we consider such binary vector as our initial point, our combinatorial optimization problem aim to improve the usual rounded fluences. Although the effectiveness of this formulation in improving the usual rounding procedure, the best scenario we can expect is the improvement of the dose coverage for the PTVs compared to the rounded solution. However, with such approach, it is not possible to expect improvements on organ sparing by solving the same problem using a combinatorial approach instead of a more effective continuous one.

If a different criteria is used to decide to which neighbor intensity level a beamlet intensity should be assigned to, we might aim to improve the rounded results both in terms of tumor dose requirements and also in terms of sparing the organs that fail to meet the dose limits prescribed. Usually, the algorithms used by commercial software for the resolution of the intensity problem include DVH penalty terms in the objective function. The most important DVH requirement consists in obtaining 95% of the prescribed dose for 95% of the target volume (D_{95}). Additionally, only a small percentage of the target volume should receive more than 110% of the prescribed dose (V_{110}). We adopted similar criteria to measure the distances between the DVH curves for x^{opt} and DVH curves for $x^{trunc} + x \cdot L$ for targets. For OARs, the criteria used are simply the mean or maximum dose within required limits. The combinatorial optimization problem of deciding, based on a dose-volume criteria, to which neighbor intensity level a beamlet intensity should be assigned to, can be stated as a binary optimization problem of the form:

$$\min f(x)$$

s.t. x binary,

where $f(x)$ is the following penalty function

$$\begin{aligned}
f(x) = & \lambda_1 [D_{95}^{PTV}(x^{opt}) - D_{95}^{PTV}(x^{trunc} + x \cdot L)]_+ + \\
& \lambda_2 [V_{110}^{PTV}(x^{trunc} + x \cdot L) - V_{110}^{PTV}(x^{opt})]_+ + \\
& \lambda_3 [\text{mean}_{OAR}(x^{trunc} + x \cdot L) - T(OAR)]_+ + \\
& \lambda_4 [\text{max}_{OAR}(x^{trunc} + x \cdot L) - T(OAR)]_+, \tag{3}
\end{aligned}$$

where $T(OAR)$ are the mean or maximum doses prescribed for a given OAR. Note that, since $[\cdot]_+ = \max\{0, \cdot\}$, if the mean or maximum dose of a given OAR is within the prescribed limits, that term in the objective function is null. However, they have the important role of preventing a given organ of violating its prescribed limits. The term becomes “active” if during the optimization process those limits are exceeded. For the two head & neck cancer cases considered, only the PTVs and the parotids have an active role in the combinatorial optimization process. The λ parameters are tuned in a trial and error process. The importance of each term of (3) as well as the difficulty of obtaining the desired improvements give a good insight to tune those penalty parameters. In Figure 7 we have an illustration of the purpose of each penalty term of the penalty function (3). The aim of the first term, illustrated by arrow 1, is to improve the D_{95} metric. This is our priority since the main goal is to irradiate the tumor volume. The goal of the second term, illustrated by arrow 2, is to improve the V_{110} metric. The goal of the third term, illustrated by arrow 3, is to improve the mean dose of a OAR while the objective of the fourth term, illustrated by arrow 4, is to improve the maximum dose of a OAR.

This formulation originates a large combinatorial optimization problem. For the head and neck problems introduced in the previous section, the number of beamlets is 1613 for the first case and 1356 for the second case, which means we have $2^{1613} = 3.6423 \text{ E } 485$ and $2^{1356} = 1.5727 \text{ E } 408$ possibilities to consider, respectively. The magnitude of those numbers implies that both an exhaustive approach and an exact approach (branch and bound) are inviable.

3.2 Combinatorial optimization algorithm

There exists a number of heuristics to address successfully this problem. We implemented versions of binary genetic algorithms (using the Genetic Algorithm Optimization

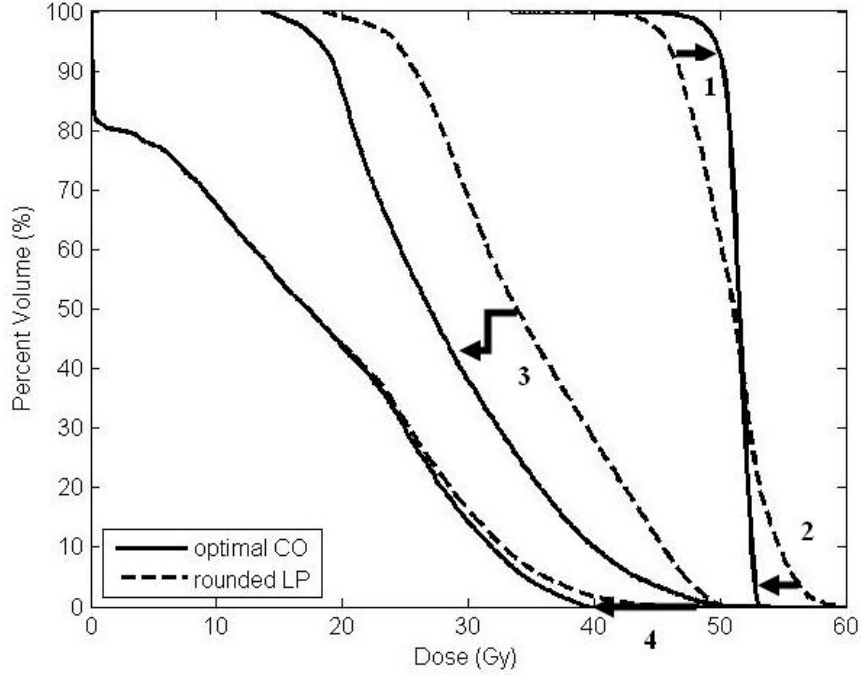


Figure 7: Illustration of the purpose of each penalty term of penalty function (3).

Toolbox of MATLAB) and versions of binary particle swarm algorithms (using the Particle Swarm Optimization Toolbox of MATLAB). However, the most successful approach was consistently obtained by a tabu search algorithm.

Tabu search is a meta-heuristic, proposed in the 80's by Glover [9], used successfully in many science fields, particularly in combinatorial optimization problems. Tabu search is an iterative neighborhood search procedure that moves towards a local optimal solution in the neighborhood of the current iterate.

The core of tabu search is the construction and update of a tabu list. This is its distinguishing feature when compared with other neighborhood or local search algorithms. The tabu list is a short-term memory containing the points that have been tested in recent past iterations. It avoids repetition or cycling, by keeping track of recent iterates information. Other types of memory lists can be used depending on the characteristics of the problem at hand. Here, we will use a simple tabu search implementation, because it contains only the main ingredient of the tabu search – a short-term memory tabu list without aspiration criteria. For more details concerning tabu search see, e.g., Glover and Laguna [10].

Typically, in the local or neighborhood search, the points to examine are selected ran-

domly from a candidate list, subset of the points of the current iterate’s neighborhood. The use of random sampling may be contrasted with the use of probabilities in a variant called probabilistic tabu search. Recently, probabilistic tabu search methods have been found effective in solving 0–1 mixed integer programming problems [4].

For our binary problem, given a current iterate (current best point), the candidate points are obtained by changing one or more than one coordinate from 0 to 1 or from 1 to 0. In practice, the candidate points are obtained from the current best point by switching the intensity level of a beamlet, or more than one beamlet intensity level at a time. Instead of picking the beamlets that are to be changed randomly, we decided to do that choice based on probabilities. A beamlet will have a low probability of being selected if its optimal fluence value is very close to an intensity level while a beamlet whose optimal fluence value is roughly in the middle of two intensity levels will have a higher probability of being selected. Those probabilities remain constant during all the optimization process. Additionally, for the OARs that fail to meet the prescribed dose limits, the probability of a beamlet that irradiates those OARs and whose value of the binary vector x is 1, i.e., is on the upper level of intensity, is kept at a high value. On the other hand, beamlets that irradiate OARs that fail to meet the prescribed dose limits and have a 0 entry in x are kept at a low probability of being selected.

If the neighborhood search procedure used within the tabu search scheme only progresses when a decrease of the objective function is obtained, the neighborhood method is called *descent method*. We used a descent method in our implementation. In Algorithm 3.1 we describe the probabilistic tabu search framework we implemented:

Algorithm 3.1 Binary Probabilistic Tabu Search

1. Choose a starting binary vector x_0 . Determine the probability of choosing each coordinate. Initialize the tabu list. Set $k = 0$.
2. With the goal of decreasing $f(x_k)$, try to obtain x_{k+1}^{trial} in the neighborhood of x_k , $\mathcal{N}(x_k)$, by evaluating f at a finite number of points, picked probabilistically, that are not in the tabu list.
3. If $x_{k+1}^{trial} \in \mathcal{N}(x_k)$ is found satisfying $f(x_{k+1}^{trial}) < f(x_k)$, then set $x_{k+1} = x_{k+1}^{trial}$ (iteration is declared successful) else $x_{k+1} = x_k$ (iteration is declared unsuccessful).

4. Set $k = k + 1$. Update the tabu list.
5. If a stopping criteria is met then stop. Else go to (2).

For our binary optimization problem, the tabu list is simply a boolean vector indicating which coordinates of the current iterate have been switched (tested). After each successful iteration, the tabu list is re-initialized and the last successful switched coordinate(s) is (are) introduced in the tabu list.

3.3 Numerical results for the head & neck clinical examples

The choice of the starting binary vector x_0 in Algorithm 3.1 is vital for the success of the method. Typically, for general problems, multi-start procedures are used in tabu search schemes. Here, the choice of the starting point is natural and has a physical meaning. Since $f(x)$ can be perceived as a penalty function of the distances between the DVH curves for $x^{trunc} + x \cdot L$ and the DVH curves for x^{opt} , if we consider $x_0 = \frac{1}{L} \cdot (x^{round} - x^{trunc})$ our starting point is, in practice, x^{round} , the rounded fluences we aim to improve. By considering the previous initial starting binary vector, our objective function will start by comparing the dose-volume of the rounded solution with the dose-volume of the optimal solution. Since we use a descent method, the physical meaning of each successful iteration is an improvement of the rounded solutions.

An alternative starting point, derived from the previous one, can be obtained by setting to 0 all entries of the binary vector x that correspond to beamlets that irradiate the OARs that fail to meet the dose limits prescribed. By starting the optimization with such initial point, we start with the best possible organ sparing solution and the combinatorial optimization process will try to improve the PTV dose coverage. In our numerical tests, this approach is not as good as the previous one since it became harder to fulfil the primary objective of proper irradiation of the tumor. Nevertheless, this binary vector is important since it give us guidance on the best possible result we can expect for the OARs that we aim to spare in a more effective way.

Since the solutions obtained by linear models and nonlinear models produces similar decrease in the quality of the treatment plans, we decided to use the fluence maps that present worst organ sparing. The only organ that fail to meet the dose limits prescribed

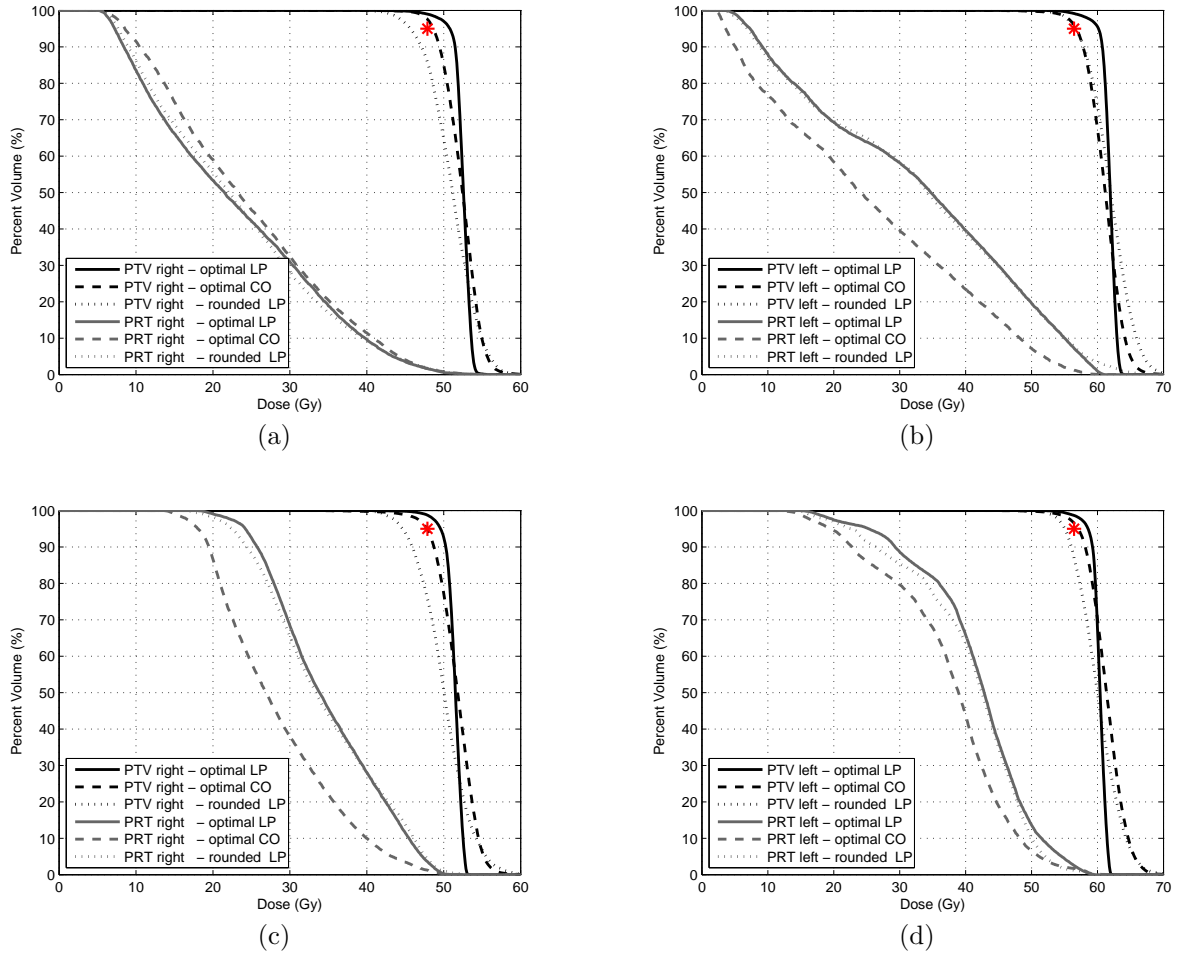


Figure 8: Cumulative dose volume histograms comparing the optimal beamlets obtained by the linear model (optimal LP), the rounded optimal beamlets using 5 levels (rounded LP), and the beamlets solution of the combinatorial optimization problem (optimal CO), for PTV right – 8(a) and PTV left – 8(b) of the first head & neck case and for PTV right – 8(c) and PTV left – 8(d) of the second head & neck case.

are the parotids, and the worst results are presented by the linear model. Therefore, we will try to improve the rounded solutions obtained by the linear model using our combinatorial approach. In Figure 8, DVH curves for PTV's and parotids are presented for optimal fluences obtained by the linear model, for the rounded optimal fluences using 5 levels of intensity, and the fluences obtained by the resolution of the combinatorial optimization problem using the binary probabilistic tabu search method. Looking at the DVH curves for PTV's we can see the benefit of the combinatorial optimization approach on the im-

Case	Target coverage	opt. LP	opt. CO	round LP
1	PTV left at 95 % volume	58.73 Gy	56.78 Gy	56.58 Gy
	PTV left % > 93% of Rx (%)	99.06	98.01	97.73
	PTV left % > 110% of Rx (%)	0.00	3.19	13.09
	PTV right at 95 % volume	49.78 Gy	48.58 Gy	45.57 Gy
	PTV right % > 93% of Rx (%)	98.94	98.96	90.65
	PTV right % > 110% of Rx (%)	0.02	5.53	6.75
2	PTV left at 95 % volume	58.53 Gy	57.03 Gy	54.98 Gy
	PTV left % > 93% of Rx (%)	99.23	98.13	93.61
	PTV left % > 110% of Rx (%)	0.00	5.56	5.43
	PTV right at 95 % volume	49.53 Gy	47.78 Gy	44.58 Gy
	PTV right % > 93% of Rx (%)	99.33	97.14	83.19
	PTV right % > 110% of Rx (%)	0.00	3.43	4.87

Table 3: Target coverage obtained by treatment plans.

provement of the rounded solution. That improvement is particularly notorious in Figures 8(a), 8(c) and 8(d), since the DVH curves of the rounded LP fluences failed to meet the criteria of having 95% of the volume of the PTV receiving 95% of the prescribed dose. Not only the DVH curves of the optimal CO fluences for the PTVs meet that criteria (DVH curve is over the asterisk) but in some cases they are almost as good as the DVH curves for the optimal LP fluences. The improvements obtained by the combinatorial approach for target coverage can be also perceived by the results presented in Table 3.

For the cases where parotids have higher mean dose than required, Figures 8(b), 8(c) and 8(d), we can verify that the combinatorial approach produced improvements on parotid sparing. The numbers of all organ sparing results are presented in Table 4. Only parotids fail to meet the dose limits prescribed. However, the combinatorial approach obtained parotid sparing within limits for both parotids in the first head & neck case and significant improvements for the second head & neck case, namely for the right parotid, obtaining almost the desired mean dose. Here, we have to highlight that, using the alternative binary vector mentioned before, the parotid sparing is almost identical to the one obtained after

Case	OAR	Mean Dose (Gy)			Max Dose (Gy)		
		opt. LP	opt. CO	round LP	opt. LP	opt. CO	round LP
1	Spinal cord	–	–	–	42.97	44.18	43.17
	Brainstem	–	–	–	50.74	52.52	51.25
	Left parotid	32.75	25.46	32.73	–	–	–
	Right parotid	22.88	24.44	23.01	–	–	–
2	Spinal cord	–	–	–	42.56	44.14	44.55
	Brainstem	–	–	–	50.04	48.24	49.65
	Left parotid	41.64	37.35	40.53	–	–	–
	Right parotid	34.75	28.30	34.37	–	–	–

Table 4: OARs sparing obtained by treatment plans.

the combinatorial approach, which means that the combinatorial optimization approach produced almost “the best” solution we could have expected.

4 Conclusion Remarks

A common way to solve the inverse planning in IMRT optimization problems is to use a beamlet-based approach. This approach leads to a large-scale programming problem, with thousands of variables and hundreds of thousands of constraints, and as a consequence, typically, the treatment planning is divided into three smaller problems which can be solved separately: geometry problem, intensity problem, and realization problem. That division has the consequence of causing a plan quality deterioration arising from the transition between the intensity problem and the realization problem. Typically, on the beamlet-based approach, after the optimal beamlet intensities are determined, they are discretized over a range of values using a distance criteria (rounding). However, that decision criteria is not appropriate and can lead to severe plan quality deterioration.

The deterioration of the treatment plan affects mostly the target coverage. That is natural since all beamlets are meant to irradiate the tumor volume and only few also irradiate the different OARs in the neighborhood. Therefore, our main focus is to improve

the target coverage of the rounded fluences and obtain a coverage similar to the one achieved by the continuous optimal fluences. We propose an alternative decision criteria based on a dose-volume response instead of the usual rounding procedure. That criteria has physical meaning and originates a combinatorial optimization problem of deciding, based on a dose-volume criteria, to which intensity level a beamlet intensity should be assigned to. Since the optimal continuous fluences can fail to properly spare some OARs, while improving the target coverage we also attempt to improve the sparing of those OARS.

A binary probabilistic tabu search method was proposed to solve the combinatorial optimization problem. Two head and neck clinical examples were used to test the ability of the proposed formulation and resolution method to obtain improved plans compared to the usual rounding procedure. The results obtained did improve the rounded solutions, with a clear increase of the plan quality both in terms of target coverage and also in parotid sparing. Although these results were obtained for particular clinical examples we believe that the transition using this combinatorial approach can always improve the usual transition between the intensity problem and the realization problem, regardless the model used to solve the intensity problem and the clinical case at hand.

References

- [1] M. Alber, G. Meedt, F. Nusslin, and R. Reemtsen, *On the degeneracy of the IMRT optimization problem*, Med. Phys. 29 (2002), pp. 2584–2589.
- [2] D. M. Aleman, D. Glaser, H. E. Romeijn, and J. F. Dempsey, *Interior point algorithms: guaranteed optimality for fluence map optimization in IMRT*, Phys. Med. Biol. 55 (2010), pp. 5467–5482.
- [3] G. Bednarz, D. Michalski, C. Houser, M.S. Huq, Y. Xiao, P. R. Anne, and J. M. Galvin, *The use of mixed-integer programming for inverse treatment planning with pre-defined segments*, Phys. Med. Biol. (2002), pp. 2235–2245.
- [4] A. Colorni, M. Dorigo, F. Maffioli, V. Maniezzo, G. Righini, and M. Trubian, *Heuristics from nature for hard combinatorial optimization problems*, Int. Trans. Op. Res. 3 (1996), pp. 1–21.

- [5] CPLEX, ILOG CPLEX, <http://www.ilog.com/products/cplex>.
- [6] D. Craft, *Local beam angle optimization with linear programming and gradient search*, Phys. Med. Biol. 52 (2007), pp. 127–135.
- [7] J. O. Deasy, A. I. Blanco, and V. H. Clark, *CERR: A Computational Environment for Radiotherapy Research*, Med. Phys. 30 (2003), pp. 979–985.
- [8] M. Ehrgott, A. Holder, and J. Reese, *Beam Selection in Radiotherapy Design*, Linear Algebra and its Applications 428 (2008), pp. 1272–1312.
- [9] F. Glover, *Future paths for integer programming and links to artificial intelligence*, Comp. Oper. Res. 13 (1986), pp. 533–549.
- [10] F. Glover and M. Laguna, *Tabu Search*, Kluwer Academic, Hingham, MA, 1997.
- [11] T. Kalinowski, *A duality based algorithm for multileaf collimator field segmentation with interleaf collision constraint*, Discrete Appl. Math. 152 (2005), pp. 52–88.
- [12] KonRad V2.1 User Manual, MRC Systems GmbH, Heidelberg, Germany, 2002.
- [13] E. K. Lee, T. Fox, and I. Crocker, *Integer programming applied to intensity-modulated radiation therapy treatment planning*, Ann. Oper. Res. 119 (2003), pp. 165–181.
- [14] G. J. Lim, M. C. Ferris, S. J. Wright, D. M. Shepard, and M. A. Earl, *An optimization framework for conformal radiation treatment planning*, INFORMS J. Comput. 19 (2007), pp. 366–380.
- [15] G. J. Lim, J. Choi, and R. Mohan, *Iterative solution methods for beam angle and fluence map optimization in intensity modulated radiation therapy planning*, OR Spectrum 30 (2008), pp. 289–309.
- [16] MATLAB, The MathWorks Inc., <http://www.mathworks.com>.
- [17] F. Preciado-Walters, M. P. Langer, R. L. Rardin, and V. Thai, *Column generation for IMRT cancer therapy optimization with implementable segments*, Ann. Oper. Res. 148 (2006), pp. 65–79.

- [18] H. Rocha, J. M. Dias, B. C. Ferreira, and M. C. Lopes, *Direct search applied to beam angle optimization in radiotherapy design*, Inesc Research Report 06/2010, ISSN: 1645–2631. Available at www.inescc.pt/documentos/6_2010.PDF.
- [19] H. Rocha, J. M. Dias, B. C. Ferreira, and M. C. Lopes, *Towards efficient transition from optimized to delivery fluence maps in inverse planning of radiotherapy desing*, Inesc Research Report 07/2010, ISSN: 1645–2631. Available at www.inescc.pt/documentos/7_2010.PDF.
- [20] H. Rocha, J. M. Dias, B. C. Ferreira, and M. C. Lopes, *On the transition from fluence map optimization to fluence map delivery in intensity modulated radiation therapy treatment planning*, Proc. of the VII ALIO-EURO – Workshop on Applied Combinatorial Optimization, Porto, Portugal, May 46, 2011, pp. 171–174.
- [21] H. E. Romeijn, R. K. Ahuja, J. F. Dempsey, and A. Kumar, *A column generation approach to radiation therapy treatment planning using aperture modulation*, SIAM J. Optim. 15 (2005), pp. 838–862.
- [22] H. E. Romeijn, R. K. Ahuja, J. F. Dempsey, A. Kumar, and J. Li, *A novel linear programming approach to fluence map optimization for intensity modulated radiation therapy treatment planing*, Phys. Med. Biol. 48 (2003), pp. 3521–3542.
- [23] H. E. Romeijn, J. F. Dempsey, and J. Li, *A unifying framework for multi-criteria fluence map optimization models*, Phys. Med. Biol. 49 (2004), pp. 1991–2013.
- [24] D. M. Shepard, M. A. Earl, X. A. Li, S. Naqvi, and C. Yu, *Direct aperture optimization: a turnkey solution for the step-and-shoot IMRT*, Med. Phys. 29 (2002), pp. 1007–1018.
- [25] S. Spirou and C. -S. Chui, *A gradient inverse planning algoritm with dose-volume constraints*, Med. Phys. 25 (1998), pp. 321–333.
- [26] Thieke C., Kufer K. H., Monz M., Scherrer A., Alonso F., Oelfke U., Huber P. E., Debus J. and Bortfeld T., *A new concept for interactive radiotherapy planning with multicriteria optimization: first clinical evaluation*, Radiother. Oncol. 85 (2007), pp. 292–298.