Sparse Solution for the Intensity-Modulated Radiotherapy Problem Using Conic Programming

Avishai Adler and Michael Zibulevsky

Department of Computer Science
Technion - Israel Institute of Technology
E-mail: adlerav@tx.technion.ac.il, mzib@cs.technion.ac.il

December 16, 2009

Abstract
We address the problem of planning an effective IMRT (Intensity Modulated Radiotherapy) treatment for cancer. IMRT is a modern method of radiotherapy. It irradiates cancerous tissues by administering lethal dose of radiation to those cells, while sparing the healthy tissues. This method’s greatest advantage is the large degree of freedom which it allows the treatment planning, but it is also its’ shortcoming. Achieving a good treatment plan is a long and sophisticated process which rarely results in the optimal solution. In this report we will demonstrate an approach for solving the inverse-IMRT problem automatically. We treat IMRT as a constrained convex optimization problem, where the physical limitations are defined by the constraints, so any solution must be feasible. We will show that minimizing an $l^1$ norm based cost-function can produce a sparse solution. This solution usually produces radiation planning which utilizes small number of the radiation sources. This can give the radiotherapist a tool to treat the patient more quickly and efficiently. Using similar concept can also optimize the settings of the treatment machine, thus making the treatment safer and faster.

keywords: Radiotherapy, IMRT, conic programming, $l^1$ norm, sparse solution, inverse problem

1 Introduction
Radiotherapy is a medical method for treating cancerous tissues with external beams of radiation. The goal of the designed treatment is to destroy the tumor by administering lethal radiation dosage, while sparing the healthy surrounding tissues. The treatment is based on the fact that unlike healthy tissue, the cancerous cells are incapable of repairing themselves after exposure to radiation. The planning method should balance between the two opposite goals of the
treatment in order to achieve high recovery ratio, as described in [12]. IMRT (intensity-modulated radiation therapy) is a rather new method of radiotherapy. The large radiation beams are modulates from many tiny "beamlets" or "pencil-beams", each one has it’s own intensity. This procedure allows much greater degree of control but requires a much more sophisticated treatment planning to achieve the best treatment [6] [4].

In the pursuit of a good treatment plan, understanding the properties of the treatment machine itself is important. We assume that a device such as the multi-leaf collimators (MLC) is in use, as in [12]. This device can rotate over the patient’s body and emit radiation fields from many directions. The shape and intensity of each beam can be controlled in this device by shifting rulers to cover parts of the field, thus creating a rectangular beam in a precise dimension. These machines are very accurate and allow us to treat each beam as if it is assembled of many beamlet (pencil-beams or beam-pixels a.k.a. 'bixels').

In the past the procedure of planning the treatment was done by a try-and-error procedure. The proper dose for each tissue is determined by a physician (defining the goals of the treatment), and then a skilled radiotherapy technician (dosimetrist) is planning the configuration of the machine that will be used in the treatment. This solution can then be checked by strong forward-problem-solvers which were developed for this purpose. If the discrepancy between the planned dosage and the dosage prescribed by the physician is unacceptable, the dosimetrist changes some of the parameters and tries again. This procedure relies on accumulated experience of the staff that plans the treatment. Although this procedure seems inefficient, in many of the standard situations it works well [4]. Moreover, if the beams gantry angles (directions) are pre-defined, the amount of dosage administered from each direction and each beamlet can be obtained automatically as done in [3] [13]. However, it is clear that this planning procedure is time consuming and probably won’t reach the optimal solution, which may result in damage to the healthy tissues and an inefficient treatment (both time and energy considered).

A modern approach is the inverse-planning. Obtaining the goals of the treatment, defined by the physician, is just a feasibility problem and because of the high degree of freedom, provided by the IMRT method, there is usually more than a single solution to the problem. That is why another aspect of the problem can be considered, so the solution can achieve optimality in some manner while sustaining the conditions of the treatments.

Different goals produces different optimization problems. Many of the works done in this field suggests very elaborate solutions. Most of them try to use a better form of optimization to solve the inverse-IMRT problem. Some approaches use linear programming (LP) [13] or quadratic programming to define the settings of the beamlets [3], but the directions of the beams have to be given. Many of the solutions uses mixed-integer linear programming (MILP) or nonlinear global optimization problems to solve this problem [6] and include all the possible directions in the system - but the complexity of these solutions is very high. Others suggests the use of heuristics, but in that way there is no guarantee for optimality.
In this work we suggest a second order conic problem formulation (SOCP) [2] that can achieve optimality in the sense that the beams’ directions are sparse. The formulation of the problem minimizes the number of radiation sources (or directions) which are used to administer the required dose. Furthermore, we will show that in the case of MLC there is an easy way to determine the settings of the machine at each direction so that it requires minimum changes in the setting of the rulers. The goal of these methods is to apply the right amount of dosage while operating the machine as efficiently as possible, thus making the treatment safer and faster.

2 Mathematical Model

The mathematical model of the problem tries to imitate the physical condition in the patient’s body. Many such models were suggested in previous articles e.g. [3],[12], [6]. In this work we decided to use a linear approximation to the problem in order to formulate the optimization problem as a second order conic programming, as described here.

**Problem formulation**  Lets assume there are $p$ radiation fields, that is the directions of the beams. Each beam is divided into $n_i$, $i = 1, 2, ..., p$ beamlets, so that the total number of beamlets in the optimization problem is $N = \sum_{i=1}^{p} n_i$. Let's denote the intensity of each beamlet by $x_i$, and $x^j$ is a vector of the intensities of the beamlets that constructs the $j$-th radiation field. We also denote $x = [x^1 \ T \ x^2 \ T \ ... \ x^p \ T] \ T = [x_1 \ x_2 \ ... \ x_N] \ T$ as the vector of all the intensities.

The space is divided into $M$ small cubes named 'voxels'. The dose at each voxel is a function of the intensities of the beamlets. $D(x)$ is the dose function. Each voxel contains a known tissue, and there are prescribed lower and upper bounds for the dosage at the $j$-th voxel marked as $\Delta_j$ and $\bar{\Delta}_j$ respectively. The constraints of the system are:

$$\Delta \leq D(x) \leq \bar{\Delta} \quad (1)$$

$$x \geq 0 \quad (2)$$

Physical experiments indicate that it is reasonable to assume that the dose function $D(x)$ is continuous and linear. Under this assumption we can denote $d_{ji}$ as the dose induced by the $i$-th beamlets at the $j$-th voxel using an intensity unit. This value can be calculated using forward-problem solver. The dose function can be written using the $M \times N$ matrix $D$: $D_{ji} = d_{ji}$ as:

$$D(x) = Dx$$

and the constraint (1) can be expressed as:

$$\Delta \leq Dx \leq \bar{\Delta} \quad (3)$$
The matrix $D$ is very sparse since not many of the beams intersects, so it can be saved explicitly.

These constraints embed the physical model. Any dosage produced by a feasible solution will fall within the prescribed boundaries. Algorithms for constrained optimizations usually indicate whether there is a feasible solution for the system. If there is none, the dosage boundaries need to be altered.

**Cost function**  The cost-function defines a measure for the optimality of the solution. In the literature there is extended use of $l^p$ norm as the cost function. The definition of $l^p$ norm is:

$$||x||_p = \left( \sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} \quad (1 \leq p \leq \infty)$$

The most common is the $l^2$ norm, which solves a constrained least-squares problem. The problem definition in this case is:

$$\min_x \left\{ x^T x \right\} \text{ subject to } \left\{ \Delta \leq Dx \leq \Delta; \ x \geq 0 \right\}$$

However this is a naive approach, since there is no physical reason to use this cost-function. The most distinguished disadvantage of this norm can be seen in the case where there are two close beams. The solution will distribute the amount of intensity between them instead of administering the dosage from only one beam.

The use of $l^1$ norm minimizes the total sum of the intensities used to administer the treatment, thus minimizing the total dose of radiation that is absorbed by the patient during the treatment.

In the treatment plan there is also great importance to reducing the number of radiation-fields. In a normal treatment there are usually 3-6 direction in use and no more than 12. The cost function of the optimization problem should be chosen in a manner that reduces the number of directions (radiation sources) in use. The cost function that does is $l^0$ norm on the ‘energy’ of each beam. $l^0$ norm is the cardinality norm. In vectors notation this norm indicates whether there is at least one non-zero element in a vector. The problem formulation is:

$$(P0) : \min_x \sum_i ||x_i||_0$$

subject to:

$$\left\{ \begin{array}{l} \Delta \leq Dx \leq \Delta \\ x \geq 0 \end{array} \right.$$
However, \(l^0\) norm is not a convex function and (P0) is a hard combinatoric optimization problem (MILP). To reduce the complexity of the problem a convex relaxation of \(l^0\) norm is introduced:

\[
\min_x \sum_i \|x^i\|_2 \text{ subject to (6)} \tag{7}
\]

which is a sum of non-squared \(l^2\) norm. This formulation usually generates a sparse solution (e.g. [7], [10]) in the sense that most vectors \(x^i\) are zero vectors. Our problem has a resemblance to signal reconstruction problem in [8], only there is no error in our case and we need only to locate the signals sources (which are the beams) in space. An alternative formulation of (7) is:

\[
(P1) : \min_{x, \xi} \|\xi\|_1 \tag{8}
\]

subject to:

\[
\begin{cases}
\Delta - Dx \leq \Xi \\
x \geq 0 \\
\xi_i \geq \|x^i\|_2 & i = 1, 2, \ldots, p
\end{cases} \tag{9}
\]

Where \(\xi = [\xi_1 \xi_2 \ldots \xi_p]^T\). Here it is clear that the linear relaxation of the \(l_0\) norm is a \(l_1\) norm. This change of the cost function often leads to results with similar sparsity (e.g. [5], [8], [7], [11], [10]).

Since \(\xi \geq 0\) the cost function is simply the sum of \(\xi\). The problem can be expressed equivalently as:

\[
(P2) : \min_{x, \xi} 1^T \xi \text{ subject to (9)} \tag{10}
\]

This formulation is actually a second-order conic programming (SOCP), which is a convex problem. The general formulation of this family of problems is:

\[
\min_x c^T x \text{ subject to } \|A_i x + b_i\|_2 \leq c_i^T x + d_i, \quad i = 1, 2, \ldots m \tag{11}
\]

SOCP is a generalization of linear-programming (LP), and like LP there are very efficient large-scale algorithms to solve it. This research field was initialized in [9]. More information about this formulation can be found in [2] and [1, lecture 3].

**MLC setting** The formulation explained above can produce the dosage that should be administered from each beamlet by specifying the intensities vectors \(x^i\). A basis-pursuit method can be used to find the best setting of the MLC, in the meaning that it will require the minimum number of changes in the settings of the occluding rulers.

In a fixed position the MLC can produce different rectangular beams by occluding part of the radiation field. The simplest way to achieve the required radiation is by using a different rectangle for each beamlet. This combination
might not be optimal and may consume more time than the minimum required. Since there are many settings possible to the MLC, there can be more than a single combination of shapes and exposure periods which can obtain the intensity function described by $x^i$. In order to choose the best setting, the different shapes of the MLC can be treated as an over-complete basis and the exposure time at each shape as it’s coefficient.

Denote $\Psi^i$ as the matrix of the over-complete system of functions possible for the $i$-th direction of the beam. Each column in $\Psi^i$ is a particular function of the MLC radiation field. The size of this matrix is $|x^i| \times L_i$ where $L_i > |x^i|$ and the column-space spans $x^i$. The corresponding coefficients vector is the $L_i \times 1$ vector $c^i$ and we shall impose the constraints $x^i = \Psi^i c^i$. At this point we suggest two cost-functions, each for a different goal. The use of $l^1$ norm on $c^i$ minimizes the total time of exposure, since there is a linear relation between time and radiation’s intensities. In that case the formulation is:

$$\min_{c^i} 1^T c^i$$

subject to:

$$\left\{ \begin{array}{l} \Psi^i c^i = x^i \\ c^i \geq 0 \end{array} \right.$$  \hspace{1cm} (12)

Another goal is to minimize the number of changes of the MLC settings. That can be done by minimizing the number of different shapes in use. The cost-function should be $l^0$ norm, but as we mentioned above, this function can be replaced by the $l^1$ norm. This solution is sparse in the sense that the coefficients of the radiation functions are sparse (e.g. [8], [7], [11], [10]). This solution can be more sparse by the use of a weighting vector $W \geq 0$. Each value in $W$ corresponds to a basis in $\Psi^i$ and has an inverse relation to the area defined by the basis. The more area the support of the basis function covers the more likely
it will be part of the selected configuration:

$$\min \, W^T c \text{ subject to (13)}$$  

(14)

If $W$ is selected as described, this solution results in the same exposure time as the previous one, but has a sparser configuration.

**General formulation**  The following formulations contain the MLC settings coefficients. The variables in $x^i$ can be replaced by expressions that depend on the MLC’s coefficients. The first formulation assumes that the directions of the beams are already known and the goal is to minimize the total time of exposure.

The group of active angles is $\{\mathcal{R}_i\}_1^k$. Denote $c_{\mathcal{R}} = [c_{\mathcal{R}1}^T, c_{\mathcal{R}2}^T, \ldots, c_{\mathcal{R}k}^T]^T$ the coefficients vector of the selected beams, $\Psi_{\mathcal{R}} = [\Psi_{\mathcal{R}1}, \Psi_{\mathcal{R}2}, \ldots, \Psi_{\mathcal{R}k}]$ as the basis matrix of all the selected beams and $D_{\mathcal{R}}$ is the dose matrix $D$ without the columns that corresponds with the beams that are not active. The formulation is:

$$(P3): \min \, W^T c_{\mathcal{R}}$$  

subject to:

$$\begin{align*}
\Delta &\leq D_{\mathcal{R}} \Psi_{\mathcal{R}} c_{\mathcal{R}} \\
c_{\mathcal{R}} &\geq 0
\end{align*}$$

(15)

This is a LP problem which can be solved with reasonable effort because the beams’ directions are predefined.

A more general formulation does not assume which beam is active. In this formulation the goal is to facilitate fewer beams as possible, but it is also important to reduce the total exposure time. The parameter $\mu > 0$ defines the importance of minimizing the total exposure time over the sparseness of the beams in use. The general formulation is:

$$(P4): \min_{\xi,c} 1^T \xi + \mu W^T c$$

subject to:

$$\begin{align*}
\Delta &\leq D \Psi c \\
c &\geq 0 \\
\xi_i &\geq \|\Psi^i c^i\|_2
\end{align*}$$

(18)

where $\xi = [\xi_1, \xi_2, \ldots, \xi_p]^T$, $c = [c^1^T, c^2^T, \ldots, c^p^T]^T$ and $\Psi = [\Psi^1, \Psi^2, \ldots, \Psi^p]$. This is a very large-scale SOCP. We leave the numerical investigation of (P3) and (P4) to future work.

3 Simulation Results

The simulated problem is the 2D case where the MLC can be rotated over a fixed plane around the patient’s body. The simulation implemented (P2) (10) and also demonstrate the advantage over $l^2$ norm (4). The dose function is
simulated by the Radon transform. After the beams are determined using (P2), the MLC setting are found. The MLC settings are determined using the basis-pursuit method, were the basis functions are simulated by a sliding window. Coefficients are determined using $l^1$ norm (12, 13), weighted $l^1$ norm (14) and $l^2$ norm (used as a cost function in the basis-pursuit).

The bottle-neck of the suggested algorithm is the time needed to solve the large optimization problem. Although there are efficient algorithms for SOCP, it is recommended to eliminate unneeded constraint and variables. The simplest way to eliminate unneeded variables is to identify the beamlets which do not point at the tumor(s). This is done by identifying the rows that corresponds to the tumor’s section in the dose matrix $D$ and finding which variables do not have any positive coefficients in those rows. Eliminating redundant constraints can be done in many ways. An important observation is that most of the values in the vector $\Delta$ equals 0 since only in the tumor’s section the dose is required to be positive. Since the intensities $x \geq 0$ and the dose matrix $D$ has only non-negative values the condition $Dx \geq 0$ is always fulfilled, so whenever $\Delta_i = 0$ this constraint can be eliminated from the system.

Reducing the complexity of the optimization problem by elimination of variables can be used in the process of determining the MLC settings (12-14). The active beamlets can be identified as the beamlets which point at the tumor’s cells. In the optimization process only basis functions whose support is smaller or equal to the support of the active beamlets and overlap it might have coefficients which are greater than zero. All of the other coefficients can be ignored.

In the following results the 2D cross-section is simulated by a $100 \times 100$ pixels images where the gray-levels represents the dosage boundaries and there are 180 optional beams’ directions. In the following example set 2 is more complex scenario than set 1, considering both shape and dosage demands. The MLC is simulated by two rulers which can occlude a portion of the radiation field, resulting in a sliding rectangular beam with variable size, as in figure 1, where each function is defined by two properties - length and position.

---

**Figure 2:** Minimum and maximum dosage boundaries for set 1
Figure 3: Results using $l^2$ norm for set 1

Figure 4: Results using $l^1$ norm for set 1

Figure 5: Minimum and maximum dosage boundaries for set 2
Figure 6: Results using $l^2$ norm for set 2

Figure 7: Results using $l^1$ norm for set 2
In set 1 the solution using $l^2$ norm (10) required all the beams’ directions (180 directions), while the SOCP formulation (10) produced a solution which utilized only 7 directions. In set 2 the $l^2$ norm distributed the intensities between all the beams’ directions as before, while the SOCP formulation required only 11 directions.

Pursuing the best MLC configuration in both cases was done over the (P2) solution (known directions and intensities of the beams). The MLC’s coefficients results for set 1 are in table 1 and for set 2 in table 2. The $l^1$ norm (12, 13) and the weighted $l^1$ norm (14) always achieve a total sum of coefficients which is smaller then the sum when using $l^2$ norm. This means that the total time of exposure is smaller. However, the number of coefficients in the weighted $l^1$ norm case is smaller than the simple $l^1$ norm case. This means that there is a smaller number of settings in the treatment plan, which translates to less time spent in adjusting the machine. The results indicate that the suggested weighted $l^1$ norm reaches the minimum exposure time, while reducing the time needed to calibrate the MLC.

<table>
<thead>
<tr>
<th></th>
<th>Number of Coef.</th>
<th>Sum of Coef.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$l^2$ norm</td>
<td>102</td>
<td>18.53</td>
</tr>
<tr>
<td>$l^1$ norm</td>
<td>97</td>
<td>16.33</td>
</tr>
<tr>
<td>weighted $l^1$ norm</td>
<td>47</td>
<td>16.33</td>
</tr>
</tbody>
</table>

Table 1: Total sum and number of MLC’s coefficients after basis-pursuit on set 1

<table>
<thead>
<tr>
<th></th>
<th>Number of Coef.</th>
<th>Sum of Coef.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$l^2$ norm</td>
<td>172</td>
<td>22.42</td>
</tr>
<tr>
<td>$l^1$ norm</td>
<td>172</td>
<td>20.41</td>
</tr>
<tr>
<td>weighted $l^1$ norm</td>
<td>86</td>
<td>20.41</td>
</tr>
</tbody>
</table>

Table 2: Total sum and number of MLC’s coefficients after basis-pursuit on set 2

The simulation was developed in MATLAB using the package YALMIP to define the problem and the package SeDuMi as the solver. The program was tested on a Pentium 4 3.20Ghz with 512Mbytes of RAM. The time to solve (P2) on a 100 × 100 pixels image is about 5 seconds, and for a 200 × 200 pixels image the time to solve is about 25 seconds. This shows that even a large system can be solved in a reasonable time.

4 Conclusions

The study of IMRT holds great potential for better treatment of cancer. The main problem is to plan a treatment that will destroy the tumor(s) with minimum damage to the other organs. In this report we introduce two simple
Figure 8: MLC’s coefficients distribution for set 1. The x-direction indicates the length of the basis and the y-direction indicates its position.

Figure 9: MLC’s coefficients distribution for set 2. The x-direction indicates the size of the basis and the y-direction indicates its position.
methods which can improve the solution of the inverse-IMRT problem by defining the problem as a continuous and convex optimization problem. The first method uses a SOCP formulation to determine the beams’ directions and intensities, while using an energy-like constraints on each beam. The second method is used afterward to optimize the setting of the treatment machine.

There are some immediate developments that are implied from this report:

- Validation of the results on real data.
- Applying more efficient large-scale algorithms for this problem.
- Integrating the basis pursuit into the formulation of the IMRT problem (P3 and P4).
- Using elastic constraints, as done in [12]. This procedure adds variables to the constraints. Those variables allow the solution to administer overdose and under-dose. Taking these variables into consideration in the cost-function may produce new solutions which may be more sparse than what was possible before at the cost of small deviations from the dose boundaries.

5 Acknowledgments

This research was partly supported by the European Communitys FP7-FET program, SMALL project, under grant agreement no. 225913. The authors would like to thank Yair Censor for introducing this exciting research area.

References


