USING EIGENSTRUCTURE OF THE HESSIAN TO REDUCE THE DIMENSION OF THE INTENSITY MODULATED RADIATION THERAPY OPTIMIZATION PROBLEM

Fredrik CARLSSON[*†‡], Anders FORSGREN[†], Henrik REHBINDER[‡] and Kjell ERIKSSON[‡]

Technical Report TRITA-MAT-2004-OS1
Department of Mathematics
Royal Institute of Technology
June 2004

Abstract

Optimization is of vital importance when performing intensity modulated radiation therapy to treat cancer tumors. The optimization problem is typically large-scale with nonlinear objective function and bound constraints on the variables. Our optimization framework is an existing treatment planning system based on a quasi-Newton sequential quadratic programming solver. This study investigates the effect on the optimal solution, and hence treatment outcome, when solving an approximate optimization problem of lower dimension. The reduction of dimension is based on a spectral decomposition of an approximation to the Hessian. The Hessian has been observed to be degenerate in the sense that many eigenvalues are much smaller than the largest ones. This observation motivates an introduction of eigenvector weights as optimization parameters, and considering an approximate problem related to the large eigenvalues only.

The eigenvector weight optimizations performed on a prostate patient case show that a reduction in dimension results in faster initial decline in the objective function, but the approximate model is in general unable to give an entirely satisfactory final solution. Another approach, which combined eigenvector weights and bixel weights as variables is also investigated. Our results indicate that lower objective values than with the conventional starting guess are obtained. However, this advantage is at the expense of the pre-computational time for the spectral decomposition.

Key words. IMRT, optimization, sequential quadratic programming, quasi-Newton method

1. Introduction

The goal of external-beam *radiation therapy* (RT) is to obtain an acceptable balance between tumor control and complications to the normal tissue surrounding the tumor. In traditional conformal RT, often referred to as 3D-CRT, the clinician determines the incident beam angles and shapes the beams with, for example, wedges in order to find an acceptable treatment. This task is often time-consuming and can be rather difficult, since the possibilities of shaping

[*] +46 8 54506150 (fcar@math.kth.se).

[†] Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology (KTH), SE-100 44 Stockholm, Sweden, http://www.math.kth.se/optsyst/. Research supported by the Swedish Research Council (VR).

[‡] RaySearch Laboratories, Sveavägen 25, SE-111 34 Stockholm, Sweden, http://www.raysearchlabs.com.

the beams are restricted. The complexity of the task also increases for patients where healthy organs lie close to and surround the tumor.

These issues motivated the introduction of so called inverse treatment planning, where the clinician specifies certain characteristics of the desired dose distribution by introducing objective functions for each *region of interest* (ROI). The ROIs are certain regions in the patient that are of specific interest to the treatment, e.g. tumor and critical organs. These objective functions typically penalize low and non-conformal dose in the tumor and high dose in the *organs at risks* (OARs). The objectives can be either physical or biological. The former describes how well the dose distribution corresponds to a prescribed distribution while the latter measures clinical outcome based on dose-response models and estimates of biological parameters. The composite objective function is then minimized by an optimization algorithm that calculates the optimal fluences profiles of the beams. These fluence profiles generate a dose distribution as close to the prescribed one as possible. By using a computer controlled *multileaf collimator* (MLC), the optimal beam profiles can be delivered with good precision. The MLC consists of narrow tungsten blocks (leafs), which can be positioned to shape the transmitted fluence. The desired fluence profile is built by adding up the transmitted fluence from several setups of the leaves. The MLC delivery of optimized fluence profiles is called *intensity modulated radiation therapy* (IMRT), which significantly improves the dose distribution and facilitates the planning process for the clinician.

In general, the IMRT problem has a non-convex nature, leading to multiple local minimas [3, 10]. Despite this, the most common optimization techniques when solving the IMRT problem with fixed beam angles are gradient based methods. They are fast compared to stochastic methods such as simulated annealing and seem to deliver satisfactory treatment plans although not ensuring global optimality. Moreover, it has been suggested that the clinical difference between the treatment corresponding to a local minima and the treatment corresponding to the global minima is negligible [11]. A gradient based method therefore seems natural when considering the IMRT problem with fixed beam angles and we will use this kind of solver in this study.

One important issue with IMRT is the computational time required to solve the optimization problem. The IMRT problem often has several thousands of variables that are subject to bound constraints, leading to computationally heavy calculations.

This paper is focused on solving an approximate problem of lower dimension. Such an approach was introduced in [8]. The motivation for reducing the problem dimension is that the problem has been observed to be *degenerate* in the sense that the Hessian of the objective function has a large number of small eigenvalues and rather few large eigenvalues [2].

We obtain such a dimension reduction by extracting vital information of the Hessian from a spectral decomposition. The Hessian at the optimum is, naturally, unknown before starting the optimization. We therefore suggest a scheme for approximating the Hessian, and thereby its eigenvectors, on beforehand, and then approximately solve the IMRT problem with the eigenvector weights as variables.

2. Problem formulation

Radiation therapy results in a *particle fluence* incident on the patient. In this study we will focus on the most widely used particle type in RT, photons. Other suitable particle types for RT include electrons, neutrons and light ions.

The clinically relevant part of the patient and the beams are discretized in our model.

The body is discretized into a number of volume elements, denoted by *voxels*. Let m be the number of voxels, and let d_i, $i = 1, \dots, m$, denote the dose in voxel i. Normally m is very large, often in the order of hundreds of thousands. The dose deposited in the voxels for a certain particle fluence is given as the weighted sum of so called *pencil beam kernels* [1]. A pencil beam kernel describes how the energy distribution from a small area element of a beam is spread inside the patient due to interactions between the incident particles and the tissue. The shape of the pencil beam kernel, calculated with Monte Carlo simulations, depend on particle type, energy of the particles and the electron density of the tissue.

The beam element related to a specific kernel is denoted a *bixel* and the union of the beam elements from all beams is referred to as *bixels*. The *pencil beam matrix*, denoted by P, describes the beam and tissue interaction by relating the bixels to the voxels. One column in the P matrix represents a kernel from a bixel and the rows represent the voxels. The weights of the bixels are the optimization variables in the problem we consider and they are denoted Ξ_j, $j = 1, \dots, n$. The dose distribution $d(\Xi)$ is related to the P matrix and the bixel weights Ξ through the linear relation

$$d(\Xi) = P\Xi. \tag{2.1}$$

In this study we consider a typical IMRT problem with physical objective functions only. As mentioned in the introduction, the underlying idea with physical objectives is to give a uniform dose of adequate dose level to the tumor and low dose to the healthy tissues. Denoting such an objective function by $F(d)$ we obtain the optimization problem

$$\begin{array}{ll} \underset{d \in I\!R^m,\ \Xi \in I\!R^n}{\text{minimize}} & F(d) \\ \text{subject to} & d = P\Xi, \\ & \Xi \geq 0. \end{array} \tag{2.2}$$

In addition to the constraints given in (2.2) there may be other constraints present, such as prescribing that the average dose in a region of the patient must not exceed a certain level. However, these constraints are normally rather few, and no such constraints are included in the problems studied in this paper. We may eliminate d from (2.2) by (2.1) and write

$$\begin{array}{ll} \underset{\Xi \in I\!R^n}{\text{minimize}} & F(P\Xi) \\ \text{subject to} & \Xi \geq 0. \end{array} \tag{2.3}$$

The objective function F in (2.2) and (2.3) may be composed by a number of different objective functions. In our setting, F is a weighted sum of K objectives F^k, $k = 1, \dots, K$, where objective k is assigned weight w^k, $k = 1, \dots, K$. The weights are set by the clinician before starting the optimization. The idea is to set high weight on the tumor-related objectives and on the objectives tied to the most critical healthy organs, while low weight should be given to the objectives of the least important risk organs. The process of picking adequate values of these weights is a trial-and-error process, where the weight factors have no direct clinical meaning. If the optimal dose distribution is not good enough, the clinician has to change the weight factors and start the optimization again. An alternative and more flexible approach of ranking the significance of the objectives has been proposed in [6].

However, in this project the conventional weight factor approach will be used. $F(d)$ is then given by

$$F(d) = \sum_{k=1}^{N} w^k F^k(d). \tag{2.4}$$

The objectives $F^k(d)$ measure penalties for a given dose distribution d. In this paper, three types of physical least-squares type objectives are used, (i) *uniform dose*; (ii) *max(min) dose*; and (iii) *max(min) dose volume histogram (DVH).* They all penalize deviation from desired dose levels quadratically. Uniform dose penalizes all voxels in the addressed ROI from the desired level, while max(min) dose penalizes only the voxels exceeding (falling short of) the prescribed dose level d^k. For a certain ROI r and objective k, $F^k(d)$ for max(min) dose can be described by

$$F^k(d) = \frac{1}{2} \sum_{i \in V^r} f(d_i, d^k) \left(\frac{d_i - d^k}{d^k} \right)^2 \Delta v_i^k, \tag{2.5}$$

where $f(d_i, d^k) = H(d_i - d^k)$ for the max dose function, the opposite for the min dose function and $f(d_i, d^k) = 1$ for the uniform dose function, $H(\cdot)$ is the Heaviside function, V^r denotes the voxels included in ROI r, d_i is the dose in voxel i and Δv_i^k is the relative volume of voxel i.

A max(min) DVH objective function acts like a max(min) dose function, with an addition of a volume constraint. The objective is introduced from the requirement that only a fraction η of the voxels of a ROI should fulfill the dose prescription d^k. If we introduce $V_\eta^r(d)$ as

$$V_\eta^r(d) = \{i : \text{voxel } i \in \{(1-\eta) \text{ voxels in ROI } r \text{ with lowest (highest) dose}\}\}, \tag{2.6}$$

we may write

$$F^k(d) = \frac{1}{2} \sum_{i \in V_\eta^r(d)} f(d_i, d^k) \left(\frac{d_i - d^k}{d^k} \right)^2 \Delta v_i^k, \tag{2.7}$$

with the same notation as in (2.5). Note that the index set $V_\eta^r(d)$ depends on the dose distribution. This means that the function $F^k(d)$ of (2.7) is not continuously differentiable. This discontinuity may be eliminated by introducing binary variables [9]. Since we are interested in using gradient based techniques and want to avoid unnecessary complexity, we model this function by a local penalty function instead. This means that the set $V_\eta^r(d)$ is updated in each iteration, based on the current dose d.

3. Description of the patient case

The patient studied in this project is a patient with a prostate tumor, irradiated by seven beams and a total of 1633 bixels. Since the prostate is situated between the bladder and rectum, which are sensitive to radiation, this case requires very precise dose delivery. In addition to the tumor itself and the important risk organs, a *planning target volume* (PTV) is included in the problem formulation. The PTV consists of the tumor volume extended by a small margin surrounding the tumor. The reason for introducing the PTV is that the exact location of the tumor is uncertain due to set-up errors and motion of the organs. The margin added to the tumor is large enough to assure that the tumor stays inside the PTV throughout the treatment.

The objective functions and their weights for all ROIs considered are listed in Table 1. Note that the property of uniform dose distribution in the tumor is emphasized by putting high weights on these objectives. It is also clear that the prostate and the PTV are the important ROIs, the main objective is to remove the tumor, not to spare the risk organs. As seen on the weight factors, rectum is considered to be slightly more important than the bladder since the complications are more severe in the former. A representative slice of the patient, outlining the ROIs included in Table 1, is shown in Figure 1.

ROI	Prostate	PTV	Rectum	Bladder	Fem Heads
Function Type	Uniform	Min DVH	Max DVH	Max DVH	Max Dose
Weight	40	30	5	10	1
Dose Level (Gy)	80	74	40	40	40
Fraction η	-	0.95	0.45	0.55	-
Function Type		Min Dose	Max DVH	Max DVH	Max DVH
Weight		30	10	5	1
Dose Level (Gy)		71	60	65	28
Fraction η		-	0.20	0.25	0.45
Function Type		Uniform	Max DVH	Max DVH	
Weight		20	10	7	
Dose Level (Gy)		80	75	75	
Fraction η		-	0.05	0.10	

Table 1: Specification of the objective functions for each ROI.

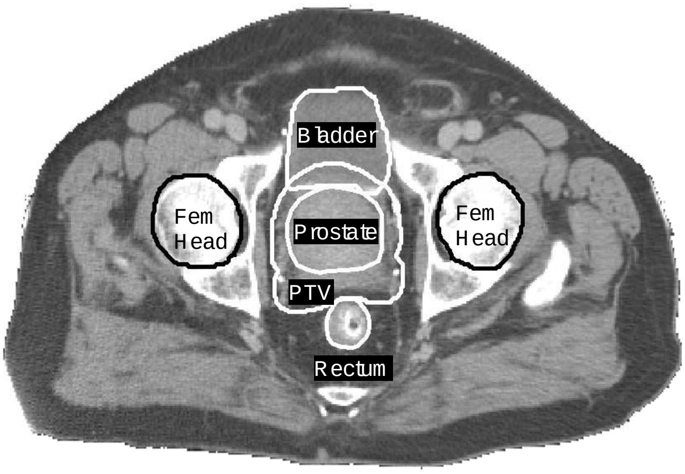

Figure 1: Slice of the patient showing the contours of the ROIs.

4. Optimization framework

A quasi-Newton sequential quadratic programming method is used to solve the optimization problems in this study. More specifically, we use ORBIT[1] [7], which adds optimization functionality to the treatment planning system Pinnacle^3®[2]. ORBIT is coupled to the quasi-Newton sequential quadratic programming solver NPSOL®[3] [4].

5. Approximation of the Hessian

We are interested in considering the optimization problem as a problem in Ξ only, i.e., on the form (2.3). The max(min) dose function is continuously differentiable, but not twice

[1] ORBIT is a product of RaySearch Laboratories.

[2] Pinnacle^3® is a registered trademark of Philips Medical Systems.

[3] NPSOL® is a registered trademark of Stanford University.

continuously differentiable, see (5.2). This discontinuity of the Hessian can be removed by adding slack variables. With objective functions given by uniform dose or max(min) dose, (2.3) may be reformulated as a pure quadratic programming problem by introducing slack variables. The number of voxels is however too large for this to be a viable strategy, if an active-set type strategy is to be used. As the tool available in this study is based on NPSOL, which in turn is based on an active-set type quadratic program solver, we stick to the formulation (2.3).

We denote by $H(\Xi)$ the second derivative matrix of the objective function with respect to Ξ, i.e.,

$$H(\Xi) = \nabla^2_{\Xi\Xi} F(d(\Xi)) = P^T \nabla^2_{dd} F(d(\Xi)) P. \tag{5.1}$$

To simplify the notation, we denote $\nabla^2_{dd} F(d(\Xi))$ by F''.

For the max(min) dose objective function, given by (2.5), the second-derivative matrix is not defined when $d_i(\Xi) = d^k$ for some i. If $d_i(\Xi) \neq d^k$, $i = 1, \ldots, m$, we obtain

$$F''_{ij} = \begin{cases} \sum_k (w^k \Delta v_i^k)/(d^k)^2 & \text{if } i = j \text{ and } f(d_i, d^k) > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{5.2}$$

It follows from (5.2) that F'' is a diagonal matrix, which is discontinuous if $d_i = d^k$ for some i, $i = 1, \ldots, m$. Expression (5.2) describes F'' for the max(min) DVH objective as well, as long as the DVH function is approximated with a local penalty method in every iteration.

Our approach for approximating the Hessian of the objective function is to form the matrix $P^T DP$, where D is an approximation of F'', where an estimate of significant voxels is made. This estimation is based on the observation that the vast majority of the voxels i in objective k fulfill $f(d_i, d^k) = 0$ after a few iterations, i.e. their dose prescriptions are fulfilled and they do not contribute to the composite objective function. This leads to many zeros in the diagonal of F'', especially near the optimum. On the other hand, voxels located in areas in or near the intersection between tumors and the critical OARs, will violate their prescribed dose levels more frequently. The diagonal elements of D for such voxels are given by the non-zero elements of (5.2). The D matrix contains much less nonzero elements than F'', making it faster to compute $P^T DP$ than $P^T F'' P$. The curvature information in the Hessian is however reduced when using D instead of F''. The choice of the non-zero diagonal elements in D is therefore a tradeoff between computational time and preserving vital curvature information. Including the voxels in the tumor and in two or three OARs turned out to be a viable strategy when building D.

With physical least-square objectives, the structure of $P^T DP$ and $P^T P$ will be similar since D is diagonal. The $P^T P$ matrix, which describes the overlap of the kernels, is band-diagonal. The main diagonal corresponds to overlap between kernels within one beam, while the off-diagonal bands correspond to the interaction between kernels from different beams. The number of bands in $P^T P$ therefore equals the number of incident beams. Furthermore are the elements along the diagonal in the $P^T P$ matrix much larger than the off-diagonal elements, since the overlap between bixels within one beam includes more voxels. The left part of Figure 2 shows this distinctive band-diagonal structure of $P^T DP$. Note that the number of beams for the patient plan, which is seven, equals the number of elliptic bodies along the main diagonal.

In the right part of Figure 2 it can be seen that an elliptic body, arising from the interaction between two beams, is composed by small non-zero regions arranged in a symmetric pattern. The size of such a region quantifies the interaction between kernels from two bixel

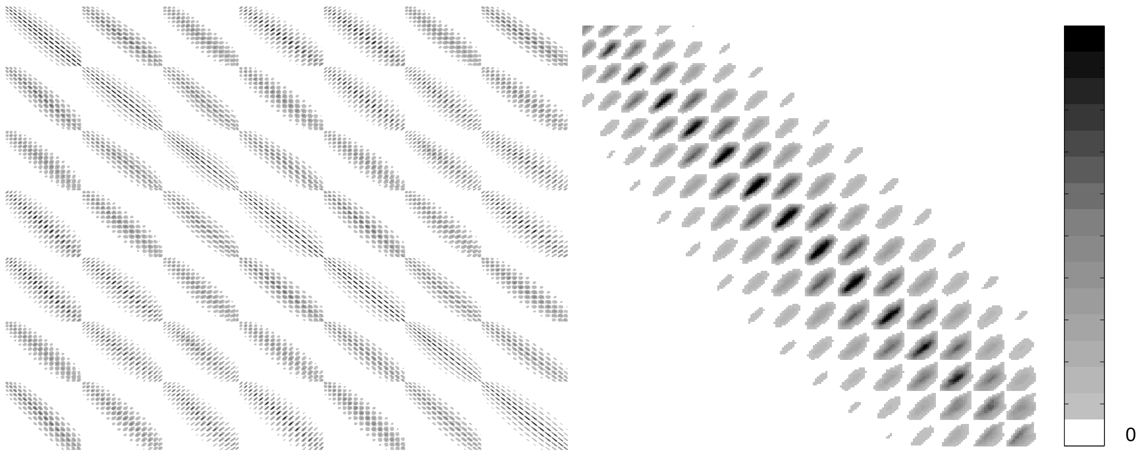

Figure 2: Left: The approximation of the Hessian showing the band-diagonal structure. Each elliptic body originates from the interaction between two beams. Right: Details of one of the off-diagonal elliptic bodies.

rows, and the orientation and shape of a region indicate the relative angle between the kernels, i.e. beams, interacting. It turns out that a circular region corresponds to perpendicular beams while an oval transverse region corresponds to opposite beams. As seen in the diagonal, an oval region along the diagonal corresponds to the interaction between bixels within one beam (beams parallel). This clear structure of P^TDP indicates that it might be possible to construct the P^TDP matrix without having to calculate the interaction between different kernels.

6. Reduction of dimension

From above, we see that the approximate Hessian may be expressed as P^TDP, where D is a positive semidefinite diagonal $m \times m$ matrix with zero diagonal elements for non-significant voxels.

The $n \times n$ matrix P^TP describes the kernel overlap in the voxels. Since a kernel is spread in the tissue, both kernels within one beam as well as kernels from different beams will overlap in the voxels. A small change in energy in one kernel can then be compensated by changing the energy in another or many other kernels. This redundancy makes the P^TP matrix degenerate. It was observed in [2] that the Hessian often has few significant eigenvalues at the optimum and that the Hessian at optimum is more degenerated than P^TP itself. The objective function thus increases the amount of degeneracy of the problem.

We suggest using this information to form a lower-dimensional problem based on the the eigenvalue decomposition of P^TDP, or equivalently, the singular value decomposition of $D^{1/2}P$, i.e.,

$$D^{1/2}P = U\Sigma V^T, \tag{6.1}$$

where the columns of V (eigenvectors) correspond to fluence shapes and the columns of U to dose shapes. The singular value matrix Σ is rectangular, $\Sigma = (S^T\, 0^T)^T$, where S is a diagonal $n \times n$ matrix containing all the eigenvalues. Both U and V are orthogonal according to the SVD [5]. The left part of Figure 3 shows the eigenvalues of $D^{1/2}P$. It turns out that more than 80% of the eigenvalues are smaller than one tenth of the largest one.

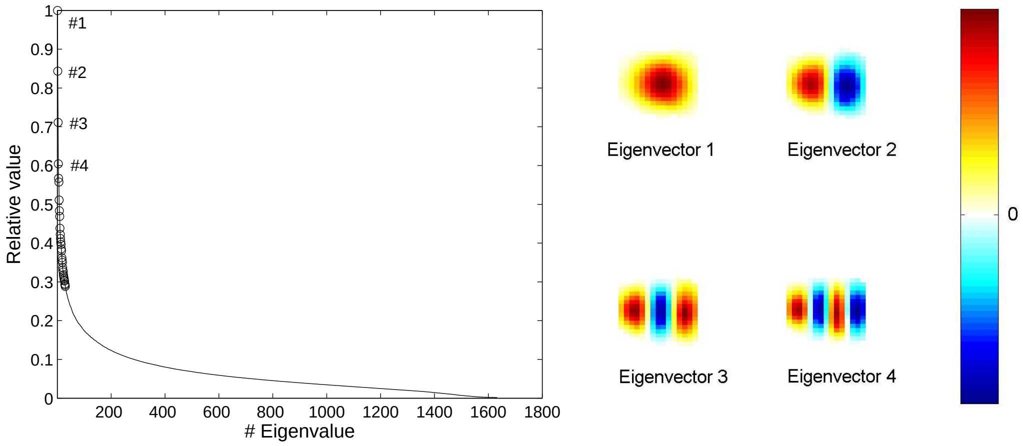

Figure 3: Left: All eigenvalues of $D^{1/2}P$ scaled relative the largest eigenvalue. Right: The beam profiles of the four dominant eigenvectors of $D^{1/2}P$.

The right half of Figure 3 indicates that the eigenvectors have positive components as well as negative components. An exception is the eigenvector corresponding to the largest eigenvalue, denoted by Eigenvector 1 in Figure 3. This eigenvector has a Gaussian-like shape. The next eigenvector, corresponding to the second largest eigenvalue, has two extreme points. We observe that the shape of the dominant eigenvectors are smooth and remind of basic tones, while the non-dominant eigenvectors show high-frequency patterns. The dominant eigenvectors furthermore solve the main conflicts between the objective functions as was discussed in [2].

These observations indicate that the eigenvectors could be used as optimization parameters. We introduce variables ξ that represent the weight of the eigenvectors. Any fluence shape Ξ is then given by

$$\Xi = V\xi. \tag{6.2}$$

The eigenvector matrix V acts as a transformation matrix between the two variable sets. With the new variables ξ, the curvature of the approximate Hessian is given by the diagonal matrix

$$H(\xi) = V^T P^T D P V = S^T S. \tag{6.3}$$

The bixel weight optimization problem (2.3) can be transformed with (6.2) to the eigenvector weight optimization problem

$$\begin{array}{ll} \underset{\xi \in \mathbb{R}^n}{\text{minimize}} & F(PV\xi) \\ \text{subject to} & V\xi \geq 0. \end{array} \tag{6.4}$$

The choice of D affects V, but since all columns of V are used in (6.4), this dependence is irrelevant. Now the n bound constraints on Ξ have been replaced by n linear constraints on ξ, to assure non-negative fluence. The linear inequalities are harder to treat for NPSOL than bounds, so this problem formulation is less useful than (2.3). To improve the usefulness of

(6.4), we need to reduce the problem dimension. We study the curvature of the Hessian to motivate such a reduction.

The degeneracy of the IMRT problem implies that the optimization problem can be approximated by p terms, with $p \ll n$. We get an approximation to the curvature with

$$H(\xi) = V_p^T P^T D P V_p = S_p^T S_p, \tag{6.5}$$

where V_p is a matrix consisting of the p dominant eigenvectors and S_p is a $p \times p$ diagonal matrix, with the p dominant eigenvalues along the diagonal. The approximation in (6.5) is accurate as long as the $n - p$ smallest eigenvalues to $D^{1/2}P$ are much smaller than the p dominant ones. This indicates that it might be possible to reduce the problem dimension and optimize p eigenvector weights without deteriorating the solution significantly. We suggest selecting the p first columns of V to obtain the p-dimensional optimization problem

$$\begin{array}{ll} \underset{\xi_p \in I\!R^p}{\text{minimize}} & F(PV_p\xi_p) \\ \text{subject to} & V_p\xi_p \geq 0. \end{array} \tag{6.6}$$

In this situation, the properties of V_p depend on the choice of the weight matrix D as well as the dimension p. It is important to have a good choice of D for (6.6) to be a suitable approximate problem to (2.3) and (6.4). We expect that the reduction of the number of variables will compensate for the increase in optimization time induced by the general linear constraints. With a low value of p, each iteration should be at least as fast for (6.6) as for (2.3).

The focus will be on solving (6.6) for different values of p and comparing the solutions to the one obtained for the bixel weight optimization problem (2.3).

7. Eigenvector weight optimization

An advantage when optimizing eigenvector weights is that the Hessian of the reduced dimension is given directly from (6.5). The quasi-Newton SQP method can therefore be initialized with a good Hessian approximation with no extra effort. When optimizing bixel weights, the initial Hessian approximation is set by NPSOL as a multiple of the identity matrix. The curvature approximation is therefore not as accurate when solving (2.3) as when solving (6.6). Work is in progress to solve (2.3) with a more accurate Hessian, but in this study the identity matrix was used as initial Hessian when solving (2.3).

7.1. Initializing the eigenvector weights

When solving (2.3), the general idea is to set the initial value of all bixel weights, Ξ_0, to the same value, i.e. to start with a homogeneous fluence. This value is scaled so that the average dose in the tumor equals the largest prescribed dose level. Our strategy when setting the initial eigenvector weights, ξ_0, is to generate a starting point that lies as close to the bixel weight starting point as possible. This is done by solving the system

$$V_p\xi_0 = \Xi_0. \tag{7.1}$$

The system (7.1) is over-determined when $p < n$, so the initial point $V_p\xi_0$ does not equal Ξ_0. Our experiences however showed that the starting point ξ_0, obtained from (7.1), was comparable in the sense that the initial value of the objective function was similar when using $V_p\xi_0$ and Ξ_0. When $p < n$, it is important to check that ξ_0 is feasible, i.e. that $V_p\xi_0 \geq 0$.

7.2. Results

Table 2 shows the objective values f and the optimization times t for different values of p after 25 iterations of (6.6) with the SQP method. The values in the table are related to the objective value f_{bixel} and optimization time t_{bixel} of (2.3) after 25 iterations. The upper limit of iterations is set to 25, since 25 iterations is the default setting in ORBIT for similar plans. This default setting is due to the empiricial observation that 25 iterations are sufficient for ORBIT to produce a satisfactory plan for such problems. The notation "*" in the table means that the optimal solution was found in less than 25 iterations. If no "*" is present next to the figure, the optimization went 25 iterations without reaching optimum and the value shown is the value obtained after 25 iterations. It should be pointed out that the optimization times given in Table 2 and Figure 5 are the times it takes from the beginning of the first iteration until optimum or 25 iterations is reached. The time to calculate the approximation of the Hessian ($20s$) and to extract the eigenvalues from this Hessian ($25s$), are not included. The pre-computational time is thus 45 seconds longer when solving (6.6) than when solving (2.3). It is clear in Table 2 that there is no p such that both f/f_{bixel} and t/t_{bixel} are less than one.

p	20	30	50	75	100	150	200
f/f_{bixel}	104*	58.5*	32.8*	20.4*	12.8*	6.55*	4.40*
t/t_{bixel}	0.29	0.33	0.48	0.51	0.71	0.81	0.90
p	300	400	500	600	750	1000	1633
f/f_{bixel}	2.36	1.82	1.34	1.04	0.77	0.57	0.39
t/t_{bixel}	1.22	1.34	1.32	1.42	1.61	2.13	5.65

Table 2: The objective values and optimization times for eigenvector weight optimization after 25 iterations relative f_{bixel} and t_{bixel} respectively. The * notation points out optimizations where optimum was found in less than 25 iterations.

To reach a lower objective value in 25 iterations with (6.6) than with (2.3), the value of p must be large, leading to a longer optimization time. When p is large, the linear inequalities will affect the computational time considerably. If p is small the optimization is fast, but the objective values obtained for the reduced dimension problems are then much higher than f_{bixel}. The information excluded when reducing the problem dimension seems to be vital to be able to generate a good dose distribution. When optimizing all eigenvector weights, the objective value is lower than the value for the bixel weight optimization. The reason is that a more accurate Hessian is used when solving the former problem.

Figure 4 shows the DVH after 25 iterations for the PTV, rectum and bladder for the bixel weight optimization and for the eigenvector weight optimization with $p = 100$ and $p = 1633$ (all eigenvectors) respectively. When using all eigenvectors, the DVH for the eigenvector optimization is better than the DVH for the bixel weight optimization. The risk organs get similar dose distributions, although slightly lower dose is delivered with the eigenvectors approach, but the dose distribution in the PTV is much more conformal when optimizing the eigenvector weights. With only 100 eigenvectors, the dose distribution achieved is much worse than the one obtained with bixel weight optimization. The bladder receives more high dose and rectum has much more voxels receiving a dose of 50 Gy or higher with the eigenvector approach. Furthermore, the dose to the PTV is much less conformal than with the bixel weight approach. Apparently, 100 eigenvectors cannot generate a satisfactory dose distribution. An observation in Figure 4 is that relatively many voxels in the bladder receives

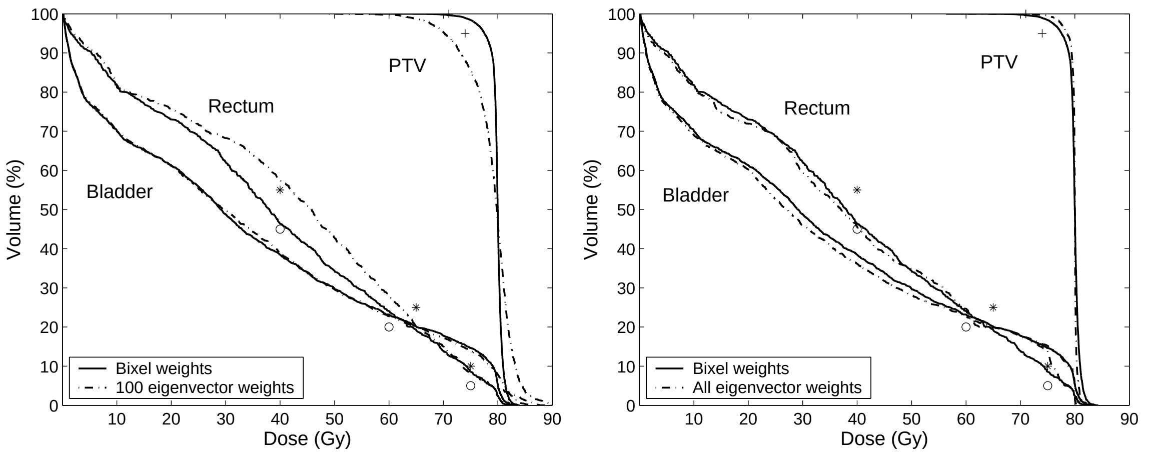

Figure 4: Dose volume histogram for the bixel weight optimization compared to an eigenvector weight optimization. The $+$, o and $*$ marks in the figures point out the objective functions for the PTV, rectum and the bladder respectively. Left: With 100 eigenvector weights. Right: With all eigenvector weights.

high dose. This is due to an overlap between the bladder and the PTV, leading to a conflict in the objective for voxels included in both these ROIs.

The results so far indicate that eigenvectors cannot both speed up the optimization and improve the dose distribution at the same time. This is visualized in the left half of Figure 5, which shows the objective value as a function of optimization time for (2.3) and for (6.6) with $p = 100$, $p = 400$ and $p = 750$. A small p generates a steep decline in the objective in the beginning, but then the curve flattens out and the objective stops decreasing after a few iterations. For a bigger value of p, the situation is the opposite. The curve lies above the bixel weight curve initially, but decreases faster and the objective is in fact lower after about two minutes. The curve corresponding to $p = 400$ lies below the bixel weight curve the first two minutes, but then it flattens out and lies above the bixel weight curve.

This behavior of fast initial decline motivates the use of both eigenvector weights and bixel weights as variables to solve the same optimization problem. This approach and the results obtained are described in the next section.

8. Using both bixel weights and eigenvector weights as variables

The idea is to solve (6.6) with a relatively small value of p in a few (five) iterations, and then use this solution as an initial estimate for solving the original problem (2.3). The right half of Figure 5 shows the objective value as a function of optimization time for (2.3) with a conventional starting point and with three improved starting points. These new starting points equal the solution after five iterations of (6.6) with $p = 300$, $p = 400$ and $p = 500$ respectively. Again, the calculation time of the eigenvectors is excluded in the figure. The small bumps in the three curves with the new starting points indicate where the optimization parameters are changed. When changing from (6.6) to (2.3), the curvature information will change and the bixel weights will have some problems finding a good search direction in the first iteration after the change, resulting in a small bump in the curve. The three new

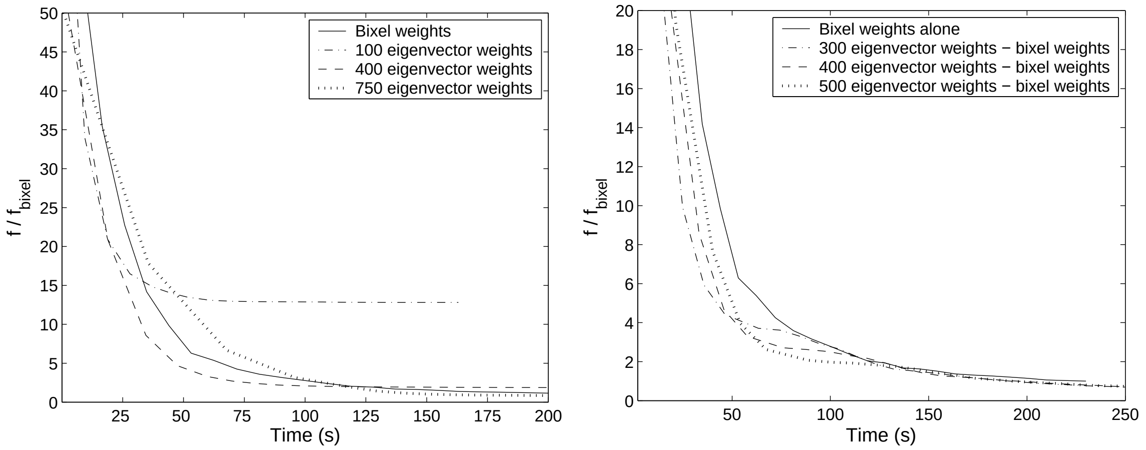

Figure 5: The relative objective value as a function of optimization time. Left: For the bixel weight optimization and three optimizations with eigenvector weights. Right: For the bixel weight optimization with a conventional starting point and with three starting points generated with eigenvector weight optimization.

starting points all generate curves that lie below the curve obtained with the conventional starting guess. The objective value f_{bixel} is passed with the new approach after $83 - 85\%$ of t_{bixel} for all three starting points. The combination of eigenvector weights and bixel weights thus improves the optimization by finding a good dose distribution faster.

Although our approach gives a faster decline in objective function value, the conventional approach is still superior due to our intial computational cost for the approximate Hessian and the associated singular values. The computation of the singular values is made in Matlab. For our approach to become superior, this computational time would have to be reduced, either by computing only the relevant part of the singular value decomposition, or by approximating the relevant singular vectors.

9. Summary and discussion

An approximation of the Hessian to the IMRT problem with physical least-square objective functions has been calculated and a spectral decomposition extracting the eigenvalues and the eigenvectors of the approximation has been performed. The considered IMRT problem is degenerate in the sense that a majority of the eigenvalues are much smaller than the dominant ones. The dominant eigenvectors, i.e. the eigenvectors corresponding to large eigenvalues, are smooth and remind of basic tones with few extreme points.

The degeneracy of the problem was used to reduce the problem dimension. By introducing eigenvector weights as variables to the dominant eigenvectors, the dimension of the bixel weight optimization problem was reduced and the problem was reformulated. The IMRT problem of a prostate patient case was then optimized with the SQP solver NPSOL, and the results for eigenvector weight optimization and bixel weight optimization were compared. With few eigenvectors present, the resulting optimization problem was solved faster than the original one. However, the resulting dose distribution was not entirely satisfactory. With many eigenvector weights included in the parameter set, the objective value was lower than the one obtained with bixel weights, but the optimization time was increased due to the introduction of computationally heavy linear inequalities.

Finally, an idea of combining bixel weights and eigenvector weights was tested. The strategy was to take advantage of the fast initial decline in the objective function that was achieved with eigenvector weights, and then use the solution as starting point to the bixel weight optimization problem. With this approach, the objective value declined faster than with the original starting guess. The total calculational time was however increased, since the pre-computation of the approximation of the Hessian and of the spectral decomposition was longer than the time gained.

For our approach to become a competitive alternative, the pre-computational time for the approximate Hessian and its associated singular values would have to be reduced. In our experiments, we have used a straightforward computation of the singular values in Matlab. We find it highly encouraging that the approximate optimization problem gives a faster initial decline in objective function value. This implies that finding alternative optimization parameters, other than bixel weights, is an area worth further study. In addition, one may consider less computationally expensive alternatives to the singular value decomposition for computing a basis for the relevant subspace. In our forthcoming research, we intend to investigate such alternatives, also in conjunction with other types of optimization methods, e.g., interior methods.

References

[1] A. Ahnesjö and M. M. Aspradakis. Dose calculations for external photon beams in radiotherapy. *Physics in Medicine and Biology*, 44(11):R99–R155, 1999.

[2] M. Alber, G. Meedt, F. Nusslin, and R. Reemtsen. On the degeneracy of the IMRT optimization problem. *Medical Physics*, 29, 2002.

[3] J. O. Deasy. Multiple local minima in radiotherapy optimization problems with dose–volume constraints. *Medical Physics*, 24, 1997.

[4] P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright. User's guide for NPSOL (version 4.0): a fortran package for nonlinear programming. Stanford University SOL 86-2, 1986.

[5] G. H. Golub and C. F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, Baltimore, Maryland, third edition, 1996. ISBN 0-8018-5414-8.

[6] K.-H. Küfer, A. Scherrer, M. Monz, F. Alonso, H. Trinkaus, T. Bortfeld, and C. Thieke. Intensity-modulated radiotherapy - a large scale multi-criteria programming problem -. *OR Spectrum*, 25:223–249, 2003.

[7] J. Löf. *Development of a general framework for optimization for optimization of radiation therapy.* PhD thesis, 2000, Stockholm University.

[8] J. Markman, D. Low, A. Beavis, and J. Deasy. Beyond bixels: Generalizing the optimization parameters for intensity modulated radiation therapy. *Medical Physics*, 29, 2002.

[9] H. E. Romeijn, R. K. Ahuja, J. F. Dempsey, A. Kumar, and J. G. Li. A novel linear programming approach to fluence map optimization for intensity modulated radiation therapy treatment planning. *Physics in Medicine and Biology*, 48(21):3521–3542, 2003.

[10] C. G. Rowbottom and S. Webb. Configuration space analysis of common cost functions in radiotherapy beam-weight optimization algorithms. *Physics in Medicine and Biology*, 47, 2002.

[11] Q. Wu and R. Mohan. Multiple local minima in IMRT optimization based on dose–volume criteria. *Medical Physics*, 29, 2002.