ITERATIVE REGULARIZATION IN INTENSITY-MODULATED RADIATION THERAPY OPTIMIZATION

Fredrik CARLSSON† and Anders FORSGREN‡

Department of Mathematics
Royal Institute of Technology
June 2005

Abstract

In beamlet-based intensity-modulated radiation therapy (IMRT) optimization, the machine settings are determined from a two step procedure: first finding the optimal fluence profiles to the pencil-weight optimization problem, followed by a conversion of the optimal fluence profiles, with a possible post-optimization of the segment weights. The inherent ill-conditioning of the pencil-weight optimization problem, resulting in jagged optimal profiles, is a disadvantage in the conversion step.

In this paper, we demonstrate the suitability of an approach based on solving the pencil-weight problem approximately by a BFGS quasi-Newton sequential quadratic programming method with diagonal initial Hessian estimate. Such an approach tends to give smooth pencil-weight solutions that are well suited for the conversion step, so that the overall deliverable treatment is of high quality and obtained at less expense than if the pencil-weight problem was solved exactly. In addition, we demonstrate that there is no big potential in running the pencil-weight optimization too long, in that the jagged profiles typically inhibit further improvement.

We justify the appealing properties of our quasi-Newton approach for the pencil-weight optimization problem based on a discussion on iterative regularization for quadratic ill-conditioned inverse problems. Iterative regularization means utilizing a method that gives fast decrease in the objective function value by taking initial steps in directions that tend to correspond to large eigenvalues, e.g. the conjugate gradient method. The one-to-one correspondence between a quasi-Newton approach and a conjugate gradient method on a quadratic unconstrained problem is used to predict the behavior of our approach.

We study the performance of our method on five patient cases with dose-volume objectives and bounds on the beamlet weights. For all cases, the jaggedness of the beam profiles increases with iteration number while the objective value decreases rapidly in the first iterations. Further, for these cases, the objective value after conversion attains a minimum after less than 50 iterations, indicating that further optimization, apart from increasing the calculation time, deteriorates the plan quality.

Key words. intensity-modulated radiation therapy, quasi-Newton method, conjugate gradient method, regularization, iterative regularization

*Research supported by the Swedish Research Council (VR).
†RaySearch Laboratories, Sveavägen 25, SE-111 34 Stockholm, Sweden; and Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology (KTH), SE-100 44 Stockholm, Sweden (fredrik.carlsson@raysearchlabs.com, fcar@math.kth.se).
‡Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology (KTH), SE-100 44 Stockholm, Sweden (andersf@kth.se).
1. Introduction

In inverse treatment planning, the prescribed dose distribution in the patient is specified by
the clinician and the task is to find the optimal machine settings, i.e. the settings delivering a
plan as close as possible to the prescribed one. In practice, the prescribed plan is formulated
through an optimization problem, where optimization functions are assigned to important
regions of the patient. The machine settings are given by the optimal solution to this optimi-
zation problem, either directly or after a conversion process, depending on the choice of
optimization variables.

Ideally, the physical restrictions of the delivery system are introduced as constraints in the
optimization problem, see e.g. [4, 23], and the machine parameters are used directly as opti-
mization variables. Such machine parameters can be the number of beams, the gantry angles,
the couch angles, the leaf positions of the multi-leaf collimator (MLC), the jaw positions and
the segment weights. Although there are modules in commercial treatment planning systems
available where the MLC settings are optimized directly, e.g. [17], currently most intensity-
modulated radiation therapy (IMRT) plans are created by solving the inverse problem in two
steps via the beamlet-based approach. We believe that direct machine parameter optimiza-
tion is the future direction to go, but we think that the pencil-weight problem will still be
present in an initial phase or as a benchmark.

The first step in beamlet-based IMRT involves solving a pencil-weight optimization prob-
lem, where the beamlet weights constitute the optimization variables. In the second step,
the optimal profiles are converted into machine settings to obtain a deliverable plan. The
conversion step alters the dose distribution since the fluence profiles cannot be reconstructed
exactly with a reasonable amount of segments. Often, the quality of the dose distribution
is degraded, and post-optimization of the weights of each segment, so called segment-weight
optimization, is performed.

Even though the conversion process is complex, one could in general say that smooth
profiles are easier to convert than jagged profiles. More specifically, for a fixed number of
segments, the discrepancy between a reconstructed fluence profile and the corresponding
original profile is larger when converting a jagged profile than when converting a non-jagged
profiles. Conversely, if the discrepancy tolerance is fixed, more segments are needed to
reconstruct a jagged profile than a non-jagged one, resulting in more MUs and longer delivery
time. In addition to facilitating the conversion process, smooth profiles are less sensitive to
geometric uncertainties than jagged profiles [28]. Furthermore, as the jaggedness increase, so
does the contribution of scattered radiation [20].

In [2], it was observed that the beamlet-based IMRT optimization problem is degenerate
in the sense that the Hessian of the objective function has a large number of small eigenvalues
and rather few large eigenvalues. This leads to numerical instabilities and solutions sensitive
to high-frequency perturbations. The optimal fluence profiles are therefore, in general, very
jagged. Combining this with the delivery constraints included in the conversion step produces
a conflict; the more accurately we solve the pencil-weight problem the harder it will be to
find machine settings that yield a high-quality treatment.

The above discussion implies that the optimal fluence profiles do not generate the optimal
machine settings after conversion. Instead of finding the jagged optimal profiles, a sensible
approach is to find smooth profiles suitable for conversion that produce a high-quality, but
not necessarily optimal, dose distribution before conversion. It has been observed that quasi-
Newton (QN) methods with a diagonal matrix as initial Hessian estimate produces such solutions. The goal of this study is to explain why this method, which we refer to as our QN approach although it is very wide spread, has these appealing properties. To do this, we will discuss regularization of the IMRT problem.

A regularized IMRT problem has smooth solutions and the numerical instabilities present when solving the original problem have vanished. To make a regularization approach viable for clinical use, some requirements have to be met by the regularization technique. It should not be computationally heavy and it should be integrated into the iterative process. Further, the regularization scheme should be functional for problems with nonlinear objective functions and bounds.

For ill-conditioned large-scale problems, a well known regularization technique is iterative regularization \[6, 13, 15\]. In iterative regularization, the optimization is stopped before high-frequency amplification occurs. This approach requires an optimization method that initially proceeds in smooth directions and makes rapid initial progress. The conjugate gradient (CG) method fulfills these requirements. It initially tends to proceed in the dominant directions [11], corresponding to the large eigenvalues. The initial decrease of the objective function will be rapid since the dominant directions affect the objective value the most. Further, the fluence profiles will stay smooth during the initial iterations since the dominant singular vectors correspond to smooth profiles. Our QN approach has similar behavior to a CG method when solving IMRT problems. In addition, it is preferable when including constraints. For other applications of QN regularization related to ill-posed inverse problems, see, e.g. [16].

This paper is organized as follows. In section 2 we formulate the inverse treatment planning problem in continuous and discrete form, and we discuss the ill-posedness of the problem. Section 3 gives an introduction to, and a discussion on some methods to solve the unconstrained quadratic programming (QP) problem. In section 4, various regularization techniques for ill-posed problems are discussed. The general IMRT problem is formulated in Section 5 and the considered optimization functions are introduced. Numerical results, for a simplified and five real IMRT problems, appear in section 6.

2. Problem statement

The calculation of delivered dose \(d(r)\) at a point \(r\) in the patient volume \(V\) is performed by integrating the irradiation density \(x(\xi)\) over the isocenter planes of the beams \(S\) with the elementary pencil beam kernel \(p(r, \xi)\). The elementary pencil beam kernel, describing how the dose is spread in the patient volume due to interactions between the incident particles and the tissue, is calculated through Monte Carlo simulations [1]. For a certain irradiation density \(x(\xi)\), the dose \(d\) in \(r \in V\) is given by

\[
d(r) = \int_S x(\xi) p(r, \xi) d\xi.
\]

Calculating \(d\) for a given \(x\) is a forward problem encountered in conventional treatment planning. Conversely, in inverse treatment planning, the prescribed dose distribution \(\hat{d}(r)\) is given by the planner at every \(r \in V\) [18]. The goal is to find the non-negative irradiation
density $x(\xi)$ that solves

$$d(r) = \int_S x(\xi) p(r, \xi) \, d\xi,$$

(2.2)

which is a Fredholm equation of the first kind. The inverse problem in (2.2) is inherently ill-posed since it in general has no solution. The ill-posedness is associated with the smoothing effect the kernels have on $x$ in the sense that high-frequency components in $x$ are removed by the integration. Computing $x$ from $\hat{d}$ will tend to amplify high-frequency components in $\hat{d}$ [15], e.g. jumps in the prescribed dose at the boundaries between the PTV and the healthy tissue [9].

The natural approach to solve (2.2) is to discretize and formulate the problem as a least-squares problem. We discretize $V$ into $m$ voxels and $S$ into $n$ bixels. The goal is to minimize the discrepancy between $d = Px$ and $\hat{d}$, where $P$ is the $m \times n$ dose kernel matrix corresponding to $p(r, \xi)$, $d$ is the $m$-dimensional calculated dose vector, $x$ is the $n$-dimensional beamlet weight vector and $\hat{d}$ is the $m$-dimensional prescribed dose vector. This can be formulated as

$$\minimize_{x \in \mathbb{R}^n} \frac{1}{2} \|Px - \hat{d}\|^2_2$$

subject to $x \geq 0$.  

(2.3)

The ill-posedness of (2.2) is inherited in the discretized form (2.3) in the sense that the singular values of $P$ quickly decay to zero, i.e. the condition number of $P$ is very large. The result is a degenerate problem, where many solutions produce almost identical objective values [2]. In opposition to (2.2), (2.3) has a solution, but it is susceptible to high-frequency perturbations originating from the jagged singular vectors corresponding to the small singular values of $P$. The result is a very jagged optimal $x$.

Inverse problems and Fredholm equations of the first kind are encountered in many applications. One example is image reconstruction, where the true image is to be reconstructed given the received data and a model for how the light is spread between the source and the detector. The inverse treatment planning problem has many similarities to the image reconstruction problem and much knowledge can be gained by studying the methods from this field [7, 29]. For example, iterative regularization approaches incorporating bounds, similar to the one we are considering, have been studied in astronomic image reconstruction [6, 5].

### 3. Solution approaches for unconstrained quadratic programming

To demonstrate some properties of our QN approach, we consider the unconstrained quadratic programming (QP) problem. We assume that $P$ has full column rank, which is reasonable as long as the bixel grid is not much finer than the voxel grid. Neglecting the constant term $\frac{1}{2}d^T\hat{d}$ and the bounds on $x$, (2.3) can be re-written as

$$\minimize_{x \in \mathbb{R}^n} \frac{1}{2} x^T H x + c^T x,$$

(3.1)

where $H = P^T P$ is the symmetric, positive definite Hessian and $c = -P^T \hat{d}$.

The formulation in (3.1), where each voxel of the patient is given identical importance and negative fluence is allowed, has no clinical meaning. However, studying and solving (3.1) is of interest since empirical results for real IMRT problems closely follow the obtained, and by theory predicted, results for (3.1).
Problem (3.1) can be solved in one Newton iteration by calculating $x = -H^{-1}c$. For large-scale problems, e.g. real IMRT problems, this method is impractical since it is too time-consuming to calculate $H$. In addition, the solution is too jagged and of no practical interest due to the ill-conditioning of the problem. Instead, we want to use a method where $H$ does not need to be explicitly known and where the solver can find smooth solutions. Two methods fulfilling these requirements are the conjugate gradient (CG) method and the quasi-Newton (QN) method. They only access the matrix $H$ through matrix-vector products on the form $Hv$, where $v$ is any $n$-dimensional vector. We give a brief introduction to CG and QN methods during the rest of this section.

Both CG and QN methods try to accelerate the slow convergence of steepest descent while avoiding the information requirements associated with the Hessian in Newton's method. The CG methods proceed in conjugate directions, i.e. $p_k^T Hp_l = 0$ for $k \neq l$, where $p_k$ denotes the search direction in iteration $k$. Furthermore, both the gradient $g_k$ and the search direction $p_k$ lie in the Krylov subspace $K_{k+1}(H, g_0) = \{g_0, Hg_0, \ldots, H^k g_0\}$, where the brackets mean linear span of the given columns. This means that the number of iterations in exact arithmetic to reach optimum always is finite and equals the number of distinct eigenvalues of the Hessian.

In this study, we use a BFGS QN method, see e.g. [22, Ch. 8.1], with a diagonal matrix as initial Hessian estimate. It is well-known that such method is equivalent to a CG method for (3.1) with exact line-search [21]. Our approach therefore behaves like a CG method, where the initial iterations tend to proceed in directions corresponding to the smooth dominant singular vectors, see e.g. [11]. This results in fast decrease of the objective and smooth profiles during the first iterations. Subsequent iterations amplify the influence of the small singular values, resulting in an increased amount of jaggedness in the profiles.

4. Regularization approaches

As discussed in the introduction, we prefer smooth profiles to jagged ones because of the conversion step from fluence profiles to machine parameters. Quite surprisingly this means that we do not want to solve (2.3) to optimality.

Several approaches to generate smooth profiles have been proposed. Their common goal is to filter out the high-frequency components associated with the small singular values. A mathematically rigorous approach among these can be found in [9], where Tikhonov regularization is performed on (2.3). In Tikhonov regularization, a term is added to the objective function penalizing non-smoothness of $x$ [26]. The term is weighted with the regularization parameter $\alpha$, which in practice is chosen a priori since the cost for tuning it is too high [13].

Methods incorporating a term, different from the Tikhonov regularization term, in the objective function penalizing jagged profiles are studied in [3, 24]. Another approach is to filter the high-frequency elements in the profiles, either during the optimization [24, 27], or when the optimal profiles are found [19, 25]. An algorithm with inherent smoothing effects was proposed in [28] and the impact on IMRT delivery efficiency by introducing upper limits on the beamlet intensities was studied in [10]. Our opinion is that iterative regularization is faster and easier to perform than the Tikhonov regularization, and more rigorous than adding smoothing terms and filtering methods.

The idea of iterative regularization is to iterate long enough to find a solution with objective value close to the optimal objective value, but to terminate the optimization process before the profiles get too jagged. In iterative regularization, the number of iterations to
reach this turning point is the regularization parameter. We denote this parameter by \( \text{iter}^* \). For \( \text{iter}^* \) to be a meaningful parameter, the optimization solver should initially proceed in dominant directions, suggesting a CG method or similar methods to be suitable.

When the optimization problem is ill-conditioned, \( \text{iter}^* \) will be very small compared to the problem size. Iterative regularization is therefore often appropriate for large-scale problems. It is, as in Tikhonov regularization, non-trivial to choose the regularization parameter \( \text{iter}^* \) beforehand [14], and we will not address this issue. Instead, we list the aposteriori determined values of \( \text{iter}^* \) for five clinical cases to illustrate the magnitude and variation of \( \text{iter}^* \) in Table 2. The precise value of \( \text{iter}^* \) is not crucial in clinical practice when segment-weight optimization is performed, since this post-optimization tends to diminish the differences in treatment quality obtained in the conversion step.

The idea of iterative regularization with a QN method with diagonal initial Hessian estimate when solving (2.3) is similar to the approach described in [5, 6], where the image reconstruction problem is considered. For instance, a preconditioner for the CG method is proposed that improves the progress of optimization without increasing the high-frequency elements in the solution. Such preconditioning might be useful also in IMRT optimization to reduce \( \text{iter}^* \) and thus reduce the optimization time.

5. Iterative regularization for IMRT problems

Ensuring non-negative fluence by adding bounds to \( x \) in (3.1) results in the following bound-constrained QP problem,

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} x^T H x + c^T x \\
\text{subject to} & \quad x \geq 0,
\end{align*}
\]

which is equivalent to (2.3). The equivalence between our QN approach and the CG method no longer holds theoretically, but empirical results indicate similar behavior. We prefer QN methods to CG methods when adding bounds to the problem since reliable software already exists. Two widespread type of methods for handling the bounds rigorously with QN are interior-point methods and active-set methods.

For IMRT problems, each optimization function is defined for a subset of the patient volume where similar treatment goals are present. These sub-volumes are called regions of interest (ROIs). There can be more than one optimization function connected to each ROI in order to model the treatment goals more accurately. The objective function is a vector-valued function over all optimization functions and should ideally be solved as a multi-criteria problem. To simplify the solution process, we reformulate the problem as a scalar-valued problem with the objective function being a weighted sum of all optimization functions, with weights set to reflect their significance to the treatment outcome.

We extend (5.1) by allowing non-quadratic objective functions \( f(d(x)) \), where \( d(x) = Px \). The objective function \( f \) is required to have continuous first derivatives and diagonal \( \nabla^2_{dd} f(d) \), i.e. \( f(d) \) should be separable in voxels. The IMRT optimization problem is given by

\[
\begin{align*}
\text{minimize} & \quad f(d(x)) \\
\text{subject to} & \quad x \geq 0.
\end{align*}
\]

To fulfill the requirements on \( f \), we formulate the objective function as a weighted sum of the optimization functions uniform dose, min (max) dose and min (max) dose-volume, which
are quadratic penalty functions. The \textit{uniform dose} optimization function penalizes deviation from the prescribed dose level for all voxels in the considered ROI, while the \textit{min (max) dose} optimization function only penalizes voxels with dose levels below (above) the prescribed dose level. When a \textit{dose-volume} function is formulated, a fraction parameter is required, specifying the size of the subset of voxels in the ROI to be penalized in each iteration. The \textit{min (max) dose-volume} function picks out the specified fraction of the voxels in the ROI with highest (lowest) dose and penalizes the voxels in this subset with dose below (above) the prescribed dose level. The \textit{min (max) dose-volume} function is handled by forming a penalty function at each iteration, to avoid introducing integer variables.

The five patient cases studied, three prostate cases (denoted by Prostate A,B,C), one tonsil case (Tonsil) and one spinal case (Spinal), are modelled on the form (5.2) with $f$ as described above. We use ORBIT\textsuperscript{1} \cite{17}, coupled to the quasi-Newton sequential quadratic programming solver NPSOL\textsuperscript{2} \cite{12}, to solve these optimization problems. The patients are treated with a 6 MV Varian linear accelerator together with a 120 leaf Varian collimator with 1 and 0.5 cm leaf widths. The objective functions for the considered cases have similar structure, with \textit{uniform dose} and \textit{min dose/dose-volume} functions on the PTV and \textit{max dose/dose-volume} functions on the risk organs.

6. Results

To illustrate the properties of our QN approach and to verify the equivalence with a CG method, we begin with solving an IMRT problem formulated as a pure QP. We then consider realistic IMRT problems, with non-quadratic objectives and bounds on the variables, to empirically show that our QN approach has suitable properties for iterative regularization on IMRT problems.

The simplified IMRT problem was generated by discretizing a prostate patient with a coarse voxelgrid ($1\text{cm} \times 1\text{cm} \times 1\text{cm}$) and a coarse bixelgrid ($1\text{cm} \times 1\text{cm}$). The problem was formulated as a pure QP by applying a \textit{uniform dose} objective on the PTV and neglecting all other voxels in the patient. There were no bounds on the bixel weights and the starting point was chosen as uniform fluence with intensity level such that the mean dose in the PTV equaled the prescribed dose. In total, the five beams had 214 bixels and the PTV had 2356 voxels. The small size of the problem made it possible to calculate the Hessian and its eigenvectors. The left part of Figure 1 shows four eigenvectors, where eigenvector $k$ corresponds to the $k$th largest eigenvalue. The jaggedness of the eigenvectors increases with increasing eigenvector number. Proceeding in search directions where the non-dominant eigenvectors are included will therefore increase the jaggedness of the fluence profiles.

The optimal solution to the pure QP, calculated with Newton’s method, was very jagged. Despite the jaggedness of the optimal profiles, only three beamlets had negative optimal weight values. Adding bounds on the beamlet weights would therefore not change the optimal solution drastically, indicating that the theory valid for unconstrained QP problems also might be valid for constrained QP problems in beamlet-based IMRT.

For our QN approach, the theory predicts that the point obtained after iter* iterations corresponds to high-quality non-jagged profiles. To verify this, and that our QN approach proceeds in dominant directions initially, we solved the pure QP with our QN approach and

\textsuperscript{1}ORBIT is a product of RaySearch Laboratories.
\textsuperscript{2}NPSOL is a registered trademark of Stanford University.
expressed the search directions as linear combinations of the eigenvectors to the Hessian. The right part of Figure 1 shows the coefficients of the eigenvectors for such linear combinations after four different number of iterations. In iteration 1 and, to a large extent, in iteration 3, the method proceeds in dominant directions, meaning that the smoothness associated with the given starting point is preserved. After 10, and in particular, 100 iterations, the method proceeds in directions that mainly are spanned by the eigenvectors corresponding to small eigenvalues. This results in an increase of the jaggedness of the profiles. Studying the dose distributions in the PTV, it is impossible to distinguish the dose distribution obtained after 20 iterations with the optimal dose distribution, indicating that further optimization is a waste of effort.

Turning to real IMRT problems, we expect to see that our QN approach is suitable for iterative regularization. To verify this, we study the beam profiles and the objective values after conversion for the five patient cases described in Section 5. Figure 2 shows the fluence profiles of one of the beams after different number of iterations for the Spinal and the Prostate A cases. There is no doubt that the jaggedness increases with iteration number and to avoid jagged profiles, the optimization has to be terminated before reaching the optimal solution. Unless the smooth profiles produce high-quality dose distributions, our regularization scheme will not be successful. We measure the quality of the treatment by the corresponding objective value. Figure 3 shows the objective values, before and after conversion, for the Spinal and the Prostate A cases as a function of iterations. The non-converted objective values, given in log-scale in the left part of the figure, decrease rapidly and the clinical difference between the solution after 50 iterations and after 100 iterations is negligible. The objective values are normalized relative to the objective value in iteration 1 for both cases.

As shown in the right part of Figure 3, the converted objective values reach a minimum after 35 and 16 iterations for the Spinal and the Prostate A cases respectively. The increase after this optimal number of iterations, denoted by iter*, can be explained by the fixation of the total number of segments in the conversion process. After some iterations, the profiles
Figure 2: The fluence profiles for one beam after different number of iterations. Left: Spinal. Right: Prostate A.

get jagged, and the conversion algorithm runs into problems reconstructing them without adding more segments. This leads to a deterioration of the dose distribution and an increase in the objective value. The converted objective values are normalized relative the converted objective value after 3 pencil-weight iterations. The upper limit on the total number of segments, the number of beams and the number of beamlets for the five cases are given in Table 1. If no post-optimization of the converted solution is performed, the pencil-weight optimization should be terminated after iter* iterations to deliver the best possible treatment. Further optimization would, in addition to take valuable time from the planner, deteriorate the plan quality.

Often, segment-weight optimization is performed after conversion to improve the treatment and hopefully obtain a dose distribution close to or as good as the one obtained with pencil-weight optimization. Figure 4 shows the objective values for the Spinal and the Prostate A cases after 10 iterations (left) and after 25 iterations (right) of segment-
Figure 3: The normalized objective value versus pencil-weight iterations before (left) and after (right) conversion.

weight optimization. As in the right part of Figure 3, the x-axis represent the number of pencil-weight iterations before conversion. The objective values are normalized relative the objective value obtained after 10 (left) and 25 (right) iterations of segment-weight optimization of the converted solution after 3 pencil-weight iterations. Compared to the two curves in the right part of Figure 3, the curves in Figure 4 flatten out after relatively few number of pencil-weight iterations. This implies that when performing segment-weight optimization,
Results

the exact value of \( \text{iter}^* \) is not very important. It is however still preferable to terminate the pencil-weight optimization at an early stage since planning time can be shortened.

Table 2 lists \( \text{iter}^* \) and the objective values after \( \text{iter}^* \) and 100 pencil-weight iterations for the considered patient cases. Three types of objective values are shown; before conversion, after conversion and after 25 iterations of segment-weight optimization. The values are normalized as described above. The values of \( \text{iter}^* \) are small compared to the problem sizes, with a mean value of 28. When performing segment-weight optimization, four of the five cases have lower objective values if the pencil-weight optimization is terminated after \( \text{iter}^* \) iterations compared to if run for 100 iterations. For the Tonsil case, the situation is the opposite. An explanation could be that the large number of beamlets for this case prevent the profiles from getting very jagged during the 100 pencil-weight iterations.

Instead of having a fixed total number of segments, we change the conversion process and fix the number of beam levels to 5 for all beams in the considered patient cases. This should result in more segments when converting a jagged profile compared to when converting a non-jagged one. The increase in the converted objective value seen in the right part of Figure 3 will most likely vanish to the cost of more segments and more MUs with this conversion approach.

Figure 5 shows the total number of segments and the converted objective value versus pencil-weight iterations for the Spinal and the Prostate A cases. The number of segments for these cases are increasing with iteration number, but the converted objective values are relatively flat after the initial decrease. Table 3 shows similar trends for the other patient cases. We have chosen to compare the number of segments, the converted objective values and the objective values after 25 iterations of segment-weight optimization for conversion after 25 iterations with the corresponding values for conversion after 100 iterations. The values are normalized as in Table 2. For four of the five cases, the number of segments obtained when converting after 100 pencil-weight iterations are larger than when converting after 25 iterations. For the Prostate C case, the number of segments is the same. The converted objective values are higher in three cases and the objective values after segment-weight optimization are higher in two cases with 100 pencil-weight iterations compared to the corresponding values with 25 pencil-weight iterations. For these patient cases, it seems like there is no gain in optimizing the beamlet weights for too many iterations. In fact, more pencil-weight iterations produces more MUs and longer delivery times without improving the treatment. We

<table>
<thead>
<tr>
<th>Patient case</th>
<th>( \text{iter}^* )</th>
<th>Pencil weight</th>
<th>Converted</th>
<th>Segment weight</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \text{obj}(\text{iter}^*) )</td>
<td>( \text{obj}(100) )</td>
<td>( \text{obj}(\text{iter}^*) )</td>
<td>( \text{obj}(100) )</td>
</tr>
<tr>
<td>Prostate A</td>
<td>16</td>
<td>3.28e-2</td>
<td>9.18e-3</td>
<td>0.284</td>
</tr>
<tr>
<td>Prostate B</td>
<td>32</td>
<td>3.31e-2</td>
<td>2.48e-2</td>
<td>0.284</td>
</tr>
<tr>
<td>Prostate C</td>
<td>12</td>
<td>1.06e-1</td>
<td>5.09e-2</td>
<td>0.428</td>
</tr>
<tr>
<td>Spinal</td>
<td>35</td>
<td>4.97e-3</td>
<td>2.40e-3</td>
<td>0.134</td>
</tr>
<tr>
<td>Tonsil</td>
<td>44</td>
<td>1.88e-4</td>
<td>5.75e-5</td>
<td>0.0696</td>
</tr>
</tbody>
</table>

Table 2: The optimal number of iterations when performing iterative regularization, \( \text{iter}^* \), followed by the normalized objective values after \( \text{iter}^* \) and 100 iterations of pencil-weight optimization, after conversion and after 25 iterations of segment-weight optimization respectively.
Figure 5: Left: The total number of segments generated in the conversion versus the number of pencil-weight iterations to conversion. Right: The normalized objective value after conversion versus pencil-weight iterations to conversion.

believe that the pencil-weight optimization in beamlet-based IMRT optimization should be terminated somewhere between 25-40 iterations for regular sized cases.

<table>
<thead>
<tr>
<th>Patient case</th>
<th>Number of segments</th>
<th>Converted</th>
<th>Segment weight</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25 iter</td>
<td>100 iter</td>
<td>obj(25)</td>
</tr>
<tr>
<td>Prostate A</td>
<td>69</td>
<td>74</td>
<td>0.230</td>
</tr>
<tr>
<td>Prostate B</td>
<td>43</td>
<td>53</td>
<td>1.07</td>
</tr>
<tr>
<td>Prostate C</td>
<td>38</td>
<td>38</td>
<td>0.633</td>
</tr>
<tr>
<td>Spinal</td>
<td>53</td>
<td>65</td>
<td>0.201</td>
</tr>
<tr>
<td>Tonsil</td>
<td>60</td>
<td>67</td>
<td>0.127</td>
</tr>
</tbody>
</table>

Table 3: The total number of segments after 25 and 100 iterations, followed by the normalized objective values after conversion and after 25 iterations of segment-weight optimization for 25 and 100 pencil-weight iterations.

7. Discussion

In this paper, we have demonstrated the benefits of employing iterative regularization by a quasi-Newton method with diagonal initial Hessian estimate for solving the beamlet-based IMRT optimization problem. The method preserves the behavior of the conjugate gradient method for unconstrained problems, with fast initial decrease of the objective function while keeping the fluence profiles smooth. In addition, the quasi-Newton method is preferable when including constraints. By terminating the optimization after relatively few iterations, we generate smooth fluence profiles that, after conversion, outperform the plan obtained by converting the jagged optimal fluence profiles. We have previously considered an approach
where the initial Hessian has been approximated from a truncated singular valued decom-
position [8]. A main benefit of the iterative regularization approach is that we need not form
the Hessian explicitly.

Further refinement of the treatment plan would require direct optimization of the seg-
ments. This can in fact be done as of today in some commercial packages, such as OR-
BIT which was used in this study [17]. However, we believe that good understanding of
the beamlet-based approach is of fundamental importance for the development of specially
tailored methods for direct machine parameter optimization. Future research will include
incorporating delivery time into the direct machine parameter optimization problem, where
parameters such as the gantry angles, the collimator angles and the number of beams are
included as optimization variables.

Another interesting direction of future work is the inclusion of more complex models into
the regularization scheme. For example, biological objective functions appear more difficult
from an optimization perspective. Possibly, insights into the problem structure, similar to
what we have obtained here, could be of use also when considering biological optimization
functions.

8. Acknowledgments

We thank Lars Eldén, Philip Gill, Björn Härdermark, Henrik Rehbinder, Göran Sporre and
Anders Szepessy for many helpful discussions and comments.

References

[3] M. Alber and F. Nüsslin. Intensity modulated photon beams subject to a minimal surface smoothing
[7] T. Bortfeld, J. Bürkelbach, R. Boesecke, and W. Schlegel. Methods of image reconstruction from projec-
the dimension of the intensity modulated radiation therapy optimization problem. Report TRITA-
(Submitted to Annals of Operations Research).


