Iterative regularization in intensity-modulated radiation therapy optimization

Fredrik Carlsson[a)]
RaySearch Laboratories, Sveav. 25, SE-111 34 Stockholm, Sweden; and Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology (KTH), SE-100 44 Stockholm, Sweden

Anders Forsgren
Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology (KTH), SE-100 44 Stockholm, Sweden

(Received 1 June 2005; revised 11 November 2005; accepted for publication 12 November 2005; published 28 December 2005)

A common way to solve intensity-modulated radiation therapy (IMRT) optimization problems is to use a beamlet-based approach. The approach is usually employed in a three-step manner: first a beamlet-weight optimization problem is solved, then the fluence profiles are converted into step-and-shoot segments, and finally postoptimization of the segment weights is performed. A drawback of beamlet-based approaches is that beamlet-weight optimization problems are ill-conditioned and have to be regularized in order to produce smooth fluence profiles that are suitable for conversion. The purpose of this paper is twofold: first, to explain the suitability of solving beamlet-based IMRT problems by a BFGS quasi-Newton sequential quadratic programming method with diagonal initial Hessian estimate, and second, to empirically show that beamlet-weight optimization problems should be solved in relatively few iterations when using this optimization method. The explanation of the suitability is based on viewing the optimization method as an iterative regularization method. In iterative regularization, the optimization problem is solved approximately by iterating long enough to obtain a solution close to the optimal one, but terminating before too much noise occurs. Iterative regularization requires an optimization method that initially proceeds in smooth directions and makes rapid initial progress. Solving ten beamlet-based IMRT problems with dose-volume objectives and bounds on the beamlet-weights, we find that the considered optimization method fulfills the requirements for performing iterative regularization. After segment-weight optimization, the treatments obtained using 35 beamlet-weight iterations outperform the treatments obtained using 100 beamlet-weight iterations, both in terms of objective value and of target uniformity. We conclude that iterating too long may in fact deteriorate the quality of the deliverable plan. *© 2006 American Association of Physicists in Medicine*. [DOI: 10.1118/1.2148918]

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

I. INTRODUCTION

Inverse treatment planning aims at satisfying certain user-specified criteria by choosing the machine settings. In practice, an optimization problem is formulated and the machine settings are given by the optimal solution to this optimization problem, either directly or after a conversion process. In a direct approach, the physical restrictions of the delivery system are incorporated as constraints in the optimization problem and the solution can be delivered directly.[1–4] Although not as straightforward, currently most IMRT plans are created after a conversion process via a beamlet-based approach.[5,6] In this study we solely focus on a beamlet-based approach and employ it by a three-step procedure. In the first step, a beamlet-weight optimization problem is solved. Then the fluence profiles are converted, via an in-house leaf-sequencing algorithm, into step-and-shoot segments. Finally, segment-weight optimization is performed to improve the deliverable plan.

In Ref. 7, it was observed that beamlet-weight IMRT optimization problems are *degenerate* in the sense that the Hessian of their objective function has a large number of small eigenvalues and rather few large eigenvalues. The Hessian thus has a large *condition number* and the problem is *ill-conditioned*. This leads to numerical instabilities and solutions sensitive to high-frequency perturbations. The optimal fluence profiles are therefore, in general, very jagged.

Even though the leaf-sequencing process is complex, one could in general say that smooth profiles are easier to convert into step-and-shoot segments than jagged profiles. The degradation of the treatment quality will therefore be larger when converting jagged profiles than smooth ones. Furthermore, jagged profiles produce plans that are more sensitive to geometric uncertainties[8] and tend to increase the contribution of scattered radiation.[9]

Apparently, there is a conflict between solving the beamlet-weight problem to optimum and keeping the degradation in plan quality small in the conversion step. To solve this conflict we need to find smooth profiles that produce a high-quality, but not necessarily optimal, dose distribution before conversion. This can be done by *regularizing* the beamlet-weight problem. For a regularization approach to be viable for clinical use, some requirements have to be met. 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 fulfilling these requirements is *iterative regularization*.[10–12] This regularization scheme requires an optimization method that initially proceeds in smooth directions and makes rapid initial progress, e.g., a conjugate gradient (CG) method. The regularization is performed by optimizing long enough to obtain high-quality profiles, but terminating the optimization before high-frequency amplification occurs.

It has been observed that a quasi-Newton (QN) method with a diagonal matrix as initial Hessian estimate often gives solutions to the beamlet-weight optimization problem that are smooth and of high quality in relatively few iterations. Although very widely spread, for simplicity, we refer to this optimization method as *our QN approach* in this paper. Our key observation is that such an approach has properties suitable for performing iterative regularization when applied to beamlet-weight optimization problems. We demonstrate its appealing properties by solving two problem sets. First, a beamlet-weight problem of a simplified prostate case formulated as an unconstrained quadratic programming (QP) problem is solved. Then, ten real IMRT cases are solved to illustrate the importance of not "over-optimizing" the beamlet-weight problem when using our QN approach, i.e., to avoid performing too many beamlet-weight iterations prior to conversion.

This paper is organized as follows. In Sec. II we formulate the inverse treatment planning problem in continuous and discrete form, and discuss the ill-conditioning of the problem. Section III gives an introduction to QN and CG methods applied to unconstrained quadratic programming problems. The general IMRT problem is formulated and the considered optimization functions are introduced in Sec. IV. In Sec. V, various regularization techniques for IMRT problems are discussed. The optimization and conversion methods used in this study are introduced in Sec. VI. The patient cases are described in Sec. VII and numerical results, for a simplified and ten clinical cases, appear in Sec. VIII.

II. A MATHEMATICAL FORMULATION OF TREATMENT PLANNING PROBLEMS

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.[13] For a certain irradiation density $x(\xi)$, the dose d in $r \in V$ is given by

$$d(r) = \int_S x(\xi)p(r,\xi)d\xi. \tag{2.1}$$

Calculating d for a given x is a forward problem encountered in conventional treatment planning. Conversely, in inverse treatment planning, the desired dose distribution $\hat{d}(r)$ is given and the task is to find the non-negative irradiation density $x(\xi)$ that solves

$$\hat{d}(r) = \int_S x(\xi)p(r,\xi)d\xi, \tag{2.2}$$

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 $\hat{d}$, e.g., jumps in the prescribed dose at the boundaries between the planning target volume (PTV) and the healthy tissue.[14]

The natural approach to solve (2.2) is to discretize and formulate the problem as a least-squares problem.[15] We discretize V into m voxels and S into n beamlets. 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

$$\begin{aligned} &\underset{x\in\mathbb{R}^n}{\text{minimize}} && \tfrac{1}{2}\|Px-\hat{d}\|_2^2 \\ &\text{subject to} && x \geqslant 0. \end{aligned} \tag{2.3}$$

The ill-posedness of (2.2) is inherited in the discretized form (2.3) in 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.[7] In opposition to (2.2), (2.3) always 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.[16,17] For example, iterative regularization approaches incorporating bounds, similar to the approach we are considering, have been studied in astronomic image reconstruction.[10,18]

III. SOLUTION APPROACHES FOR UNCONSTRAINED QUADRATIC PROGRAMMING

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

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \tfrac{1}{2}x^THx + c^Tx, \tag{3.1}$$

where $H=P^TP$ 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 it turns out that the solution to this problem has many similarities to solutions for real IMRT problems.

The solution to problem (3.1) is given by the single linear system $x=-H^{-1}c$. 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 tends to generate smooth iterates as the optimal solution is approached. Two methods fulfilling these requirements are CG methods and QN methods. They only access the matrix H through matrix-vector products on the form Hv, where v is any n-dimensional vector. These methods are equivalent when solving (3.1) with exact line-search.[19]

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., $p_k^THp_l=0$ 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 $K_{k+1}(H,g_0)=\{g_0,Hg_0,\ldots,H^kg_0\}$, 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.

A key feature of CG methods is that the initial iterations tend to proceed in directions corresponding to the smooth dominant singular vectors, see, e.g., Ref. 20. This results in fast decrease of the objective and smooth profiles during the first iterations. This is the basis for our regularization approach.

IV. FORMULATION OF IMRT PROBLEMS

In order to model the treatment goals more accurately, we extend (2.3) by allowing a nonquadratic objective function $F(d(x))$, where $d(x)=Px$. The objective function is a composite of nonquadratic optimization functions, where each optimization function is defined for a subset of the patient volume where similar treatment goals are present. These subvolumes are denoted by regions of interests (ROIs) and we allow more than one optimization function to be defined for each ROI.

The optimization functions often have conflicting goals. One way of dealing with this is to use a multi-criteria solution scheme, see, e.g., Ref. 21. We have chosen to formulate the problem as a scalar-valued problem with the objective function being a weighted sum of the optimization functions, with weights set to reflect their significance to the treatment outcome.

In this study, we use the optimization functions *uniform dose, min (max) dose*, and *min (max) dose-volume*. They all penalize deviation from the prescribed dose level quadratically, but differ in how the penalized voxels are chosen. The dose-volume function is handled by forming a penalty function at each iteration to avoid introducing integer variables. Although this heuristic handling of dose-volume functions may lead to nonsmooth functions, it has proven successful in practice. For a description of these optimization functions and for mathematical expressions, see Ref. 22.

The IMRT optimization problem is given by

$$\begin{aligned}&\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad F(d(x))\\ &\text{subject to} \quad x\geqslant 0.\end{aligned} \tag{4.1}$$

With our choice of optimization functions, the objective function F has continuous first derivatives and diagonal $\nabla^2_{dd}F(d)$, i.e., $F(d)$ is separable in voxels. The structure of the Hessian to (4.1), $\nabla^2_{xx}F(d(x))=P^T\nabla^2_{dd}F(d)P$, has similar structure to P^TP since $\nabla^2_{dd}F(d)$ is diagonal. The Hessian to (4.1) therefore has large condition number and (4.1) is ill-conditioned. Often, $\nabla^2_{dd}F(d)$ has many zeros along its diagonal, which may increase the condition number of the Hessian to (4.1) even more, see Ref. 22 for details. In general, $\nabla^2_{dd}F(d)$ is discontinuous, which implies that the Hessian to (4.1) is discontinuous.

V. REGULARIZATION APPROACHES

Several approaches to generate smooth profiles when solving IMRT optimization problems have been proposed. Their common goal is to filter out the high-frequency components associated with the small singular values. All approaches require a regularization parameter that specifies the trade-off between complexity in the intensity patterns and quality in the solution. This parameter is in practice chosen *a priori* since it in general is very costly to tune the regularization parameter during the optimization process.[11,23]

Many of the regularization approaches belong to the following three categories; variational methods, filtering methods, and iterative methods. The regularization parameter for these approaches is the weight of the stabilizing functional, the cutoff of the filter, and the number of iterations, respectively.

The variational methods are based on adding a term to the objective function penalizing nonsmooth intensity patterns. A well-known approach among these is Tikhonov regularization,[24] which has been applied to the beamlet-weight optimization problem in Ref. 14. An approach to calculate the regularization parameter in Tikhonov regularization by using an L-curve method is discussed in Ref. 25. Variational methods incorporating terms different from the Tikhonov stabilizing functional in the objective function are studied in Refs. 26 and 27.

The filtering methods solve the original problem and filters out the high-frequency elements in the profiles, either during the optimization,[27,28] or when the optimal profiles are found.[29,30]

Other regularization approaches include an algorithm with inherent smoothing effects, which was proposed in Ref. 8, and the introduction of upper limits on the beamlet intensities, which was studied in Ref. 31.

The idea of iterative regularization is to solve the original problem directly, iterating long enough to find a solution with objective value close to the optimal objective value, but terminating the optimization process before the profiles get too jagged. This approach requires an optimization method that initially proceeds in the dominant singular values, e.g., a CG method. When the optimization problem is ill-conditioned, we expect to find a suitable solution in a number of iterations which is very small compared to the problem size. Iterative regularization is therefore often appropriate for large-scale problems.

VI. METHOD

To illustrate the properties of our QN approach and to verify the equivalence with a CG method, we begin with solving a simplified IMRT problem formulated as an unconstrained QP (3.1) using exact line-search. We then consider ten realistic IMRT problems, with nonquadratic objectives and bounds on the variables (4.1), to empirically show that our QN approach has suitable properties for iterative regularization on IMRT problems.

Each real IMRT problem is solved by a three-step procedure. First, the beamlet-weights are optimized, generating a treatment denoted by τ_x, where x denotes the number of completed iterations. This is where we perform iterative regularization with our QN approach, i.e., we terminate this optimization before reaching optimum. Then a leaf-sequencing is performed to convert the intensity patterns into segments. Finally, the segment-weights are optimized for 25 iterations to improve the treatment and hopefully obtain a dose distribution close to the one obtained in the beamlet-weight optimization. The segment-weight optimization problem has the form (4.1), where the variables are the segment weights. Having performed the segment-weight optimization, we have determined the final machine settings and the treatment, denoted by ξ_x, can be delivered. As above, x denotes the number of beamlet-weight iterations carried out prior to the conversion.

The real IMRT problems are solved by ORBIT (ORBIT is a product of RaySearch Laboratories),[1] coupled to the sequential quadratic programming solver NPSOL (NPSOL is a registered trademark of Stanford University).[32] More precisely, NPSOL is a BFGS QN method (see e.g., Ref. 33 Chap. 8.1), which we initialize with a diagonal matrix as initial Hessian estimate in the beamlet-weight and the segment-weight optimization problems. We envisage that similar behavior might be obtained using CG methods for solving IMRT problems with bounds (see Ref. 34).

TABLE I. The number of beams, the predefined total number of segments, the number of voxels (m), and the number of beamlets (n) for the considered patient cases.

Patient case	No. of beams	No. of segments	m	n
Head-and-neck	9	90	163 800	6387
Meningioma	5	50	246 738	494
Prostate A	9	90	531 852	3665
Prostate B	5	40	299 811	1736
Prostate C	7	50	320 682	2521
Prostate D	6	60	553 656	1637
Prostate E	5	50	248 320	1894
Spinal	6	60	340 470	1556
Tonsil A	7	70	173 420	3502
Tonsil B	9	100	173 420	4490

The in-house leaf-sequencing algorithm can be used in two ways, either by specifying the number of intensity levels that the beamlet intensities should be discretized into or by specifying the total number of segments over all beams. We performed tests using both approaches and obtained very similar results with regard to the performance of our QN approach as an iterative regularization scheme. For conciseness, we choose to display only the results obtained with a fixed total number of segments.

VII. PATIENT CASES

Our simplified IMRT problem is a simplified prostate case, which is generated by discretizing a prostate patient with a coarse voxel-grid ($1\times1\times1$ cm^3) and a coarse beamlet-grid (1×1 cm^2). The problem is formulated according to (3.1) by applying uniform dose optimization functions to the PTV (with prescribed dose of 75 Gy), bladder (40 Gy), rectum (40 Gy), and the femoral heads (50 Gy). All other voxels in the patient are neglected. The starting point is chosen as uniform fluence with intensity level such that the mean dose in the PTV equals the prescribed dose, i.e., 75 Gy. In total, the three beams have 214 beamlets. The calculation of eigenvectors and optimization of this case are performed in Matlab.

The ten clinical patient cases consists of five prostate cases (denoted by Prostate A,B,C,D,E), two tonsil cases (Tonsil A,B), one head-and-neck case, one meningioma case, and one spinal case. They are modeled on the form (4.1) with F as described in Sec. IV. The patients are treated with a 6 MV Varian linear accelerator together with a 120 leaf Varian collimator with 0.5 and 1.0 cm leaf widths. They all have 5 mm cubic voxels and 5 mm square beamlets. The number of beams and segments differ between the cases, since the cases have been created at different clinics. As for the simplified prostate case, we use uniform fluence as the

TABLE II. The objective function for the Prostate A case.

ROI	Function	Prescription	Weight
CTV	max dose	78 Gy	100
	min dose-volume	76 Gy in 99 %	100
PVT-CTV	max dose	76 Gy	80
	min dose	50 Gy	80
	max dose-volume	75 Gy in 50 %	10
PTV_+-CTV_+	max dose-volume	60 Gy in 60 %	60
Rectum	max dose-volume	64 Gy in 85 %	60
	max dose-volume	40 Gy in 65%	60
Bladder	max dose-volume	65 Gy in 90 %	70
	max dose-volume	40 Gy in 50 %	70
Femoral heads	max dose	50 Gy	1
Normal tissue	max dose	45 Gy	50

starting point. Table I shows the number of beams, together with the predefined total number of segments, the number of voxels, and the number of beamlets for the ten patient cases. Since we show more detailed results for the Prostate A case in Sec. VIII, we list the details of the objective function for this case in Table II. Note that no optimization functions are applied to the PTV directly. Instead, the planner has chosen to address the clinical target volume (CTV), the PTV-CTV (voxels in the PTV outside the CTV), and the PTV_+-CTV_+ (voxels in a slightly extended PTV outside a slightly extended CTV).

VIII. RESULTS

Starting with the simplified prostate case, we study the spectral decomposition of the Hessian and the performance of our QN method when solving (3.1). Using our QN approach on an ill-conditioned unconstrained QP, we expect that nonjagged profiles producing a dose distribution close to the optimal one should be obtained after a significantly smaller number of iterations than the problem size of 214.

The left part of Fig. 1 shows four eigenvectors, where eigenvector k corresponds to the kth largest eigenvalue. The jaggedness of the eigenvectors increases with increasing eigenvector number. Proceeding in search directions where the nondominant eigenvectors are included will therefore increase the jaggedness of the fluence profiles. We express the search directions as linear combinations of the eigenvectors to the Hessian. The right part of Fig. 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. Such a direction is undesirable since it will add noise to the profiles without improving the objective value significantly.

Studying the dose distributions in the PTV, it is impossible to, by eye, distinguish the dose distribution obtained after 20 iterations with the optimal dose distribution. Comparing these dose distributions numerically, the mean doses to the PTV are identical to four digits, while the max dose and the min dose in the PTV are slightly different (less than 1 Gy). The difference between these dose distributions is, by our judgment, negligible, meaning that it is a waste of effort to optimize for more than 20 iterations.

Turning to the real IMRT problems, we want to assess how suitable our QN approach is for performing iterative regularization on clinical cases. We hope to see that high-quality deliverable plans (after segment-weight optimization) can be obtained without having to perform many beamlet-weight iterations. To do this, we study how the complexity of the intensity patterns changes with iteration number and how

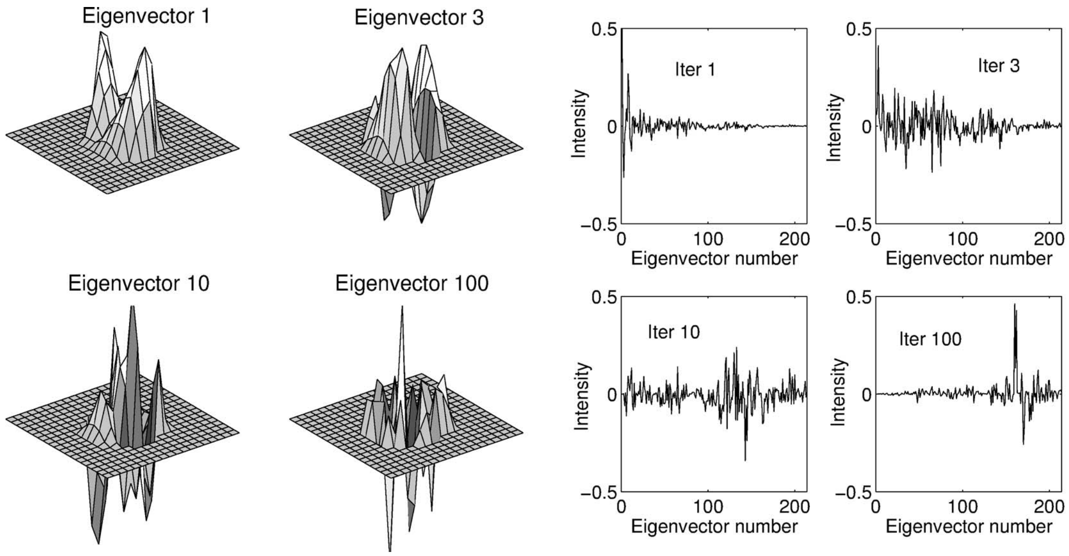

FIG. 1. Left: Eigenvectors to the Hessian of the QP. Right: Steps expressed as a linear combination of eigenvectors when solving the QP with a BFGS QN method with the identity matrix as initial Hessian estimate.

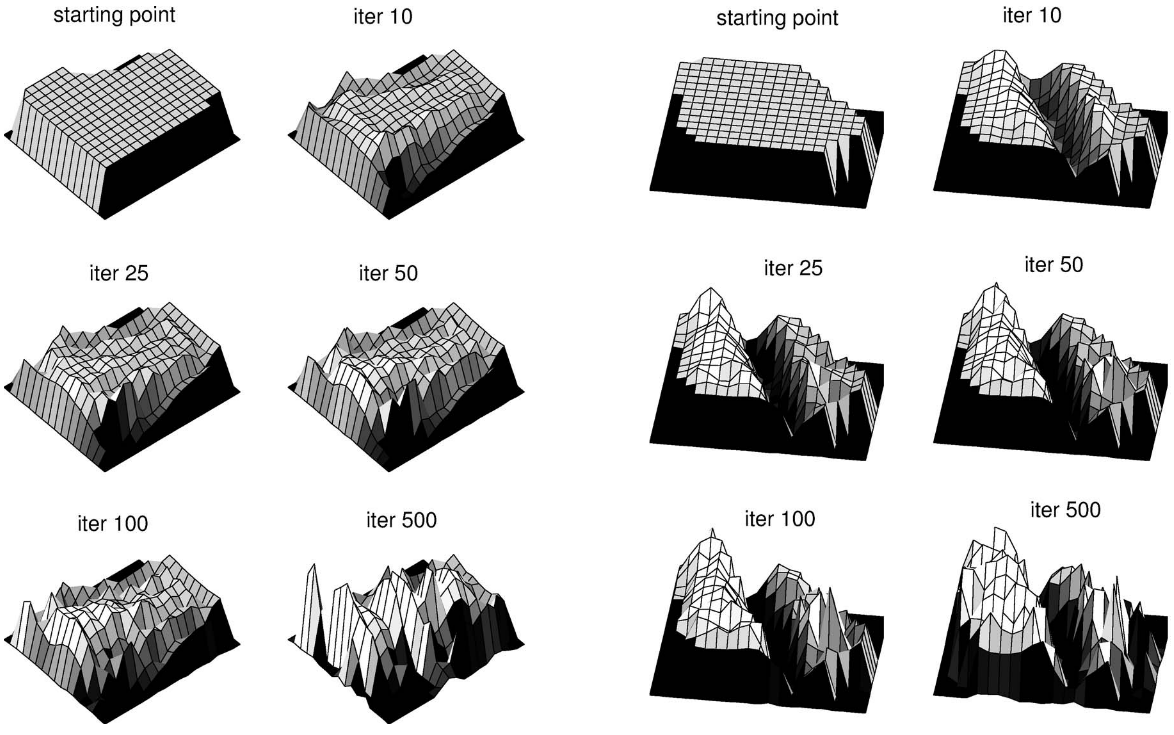

FIG. 2. The fluence profiles for one beam after different numbers of beamlet-weight iterations. Left: Prostate A. Right: Spinal.

the treatment quality after beamlet-weight optimization, after conversion, and after segment-weight optimization varies with beamlet-weight iteration number. We will assess the treatment quality in terms of objective value, dose-volume histograms (DVHs), and 2D dose distributions.

Figure 2 shows the fluence profiles of one of the beams after different numbers of iterations for the Prostate A and the Spinal cases. The jaggedness increases with iteration number and, to avoid jagged profiles, the optimization has to be terminated before reaching the optimal solution. Note that the main shapes of the profiles are obtained after 25 iterations. Reaching 100 iterations and beyond, the dose distribution is "fine-tuned" and high-frequency components are introduced into the profiles.

Figure 3 shows the objective values for the Prostate A and Spinal cases versus beamlet-weight iterations, for beamlet-weight optimization (left), after leaf-sequencing (center), and after segment-weight optimization (right). The objective values in the left and the right figures are normalized with respect to the initial beamlet-weight objective value. The center figure contains intermediate values, and hence it has been normalized with respect to its initial value.

Starting from the left with the beamlet-weight optimization, we see that the objective values decrease rapidly during the initial iterations. For both cases, the improvement in objective function value decays relatively rapidly and the dif-

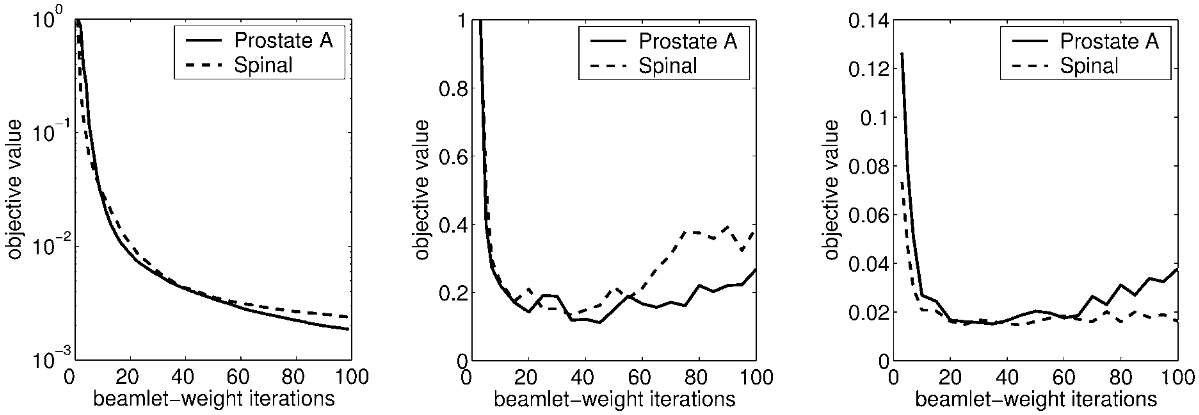

FIG. 3. The normalized objective value versus beamlet-weight iterations for the Prostate A and Spinal cases. The left and right figures are normalized with respect to the initial beamlet-weight objective function value. The center figure is normalized with respect to its initial value. Left: With beamlet-weight optimization. Center: After leaf-sequencing (prior to segment-weight optimization). Right: After segment-weight optimization.

TABLE III. Comparison of regularized and "over-optimized" plans after beamlet-weight optimization (τ_x) and after segment-weight optimization (ξ_x) in terms of objective value F, total optimization time T and MUs, where x is the number of beamlet-weight iterations.

Patient case	$\frac{F(\tau_{35})}{F(\tau_{100})}$	$\frac{F(\xi_{35})}{F(\xi_{100})}$	$\frac{T(\xi_{35})}{T(\xi_{100})}$	$\frac{\mathrm{MU}(\xi_{35})}{\mathrm{MU}(\xi_{100})}$	$\frac{F(\xi_{25})}{F(\xi_{100})}$	$\frac{F(\xi_{45})}{F(\xi_{100})}$
Head-and-neck	9.55	0.86	0.45	1.01	0.60	0.96
Meningioma	1.42	0.71	0.45	0.93	0.81	0.73
Prostate A	5.24	0.69	0.50	1.09	0.82	0.55
Prostate B	1.30	0.65	0.52	0.91	0.53	0.67
Prostate C	2.54	0.40	0.52	0.91	0.42	0.50
Prostate D	1.80	0.78	0.52	1.00	0.75	0.74
Prostate E	1.22	0.83	0.45	1.03	0.99	0.92
Spinal	2.07	1.00	0.53	0.98	0.92	0.91
Tonsil A	4.96	1.05	0.44	0.97	1.41	1.08
Tonsil B	6.26	0.64	0.46	0.97	1.03	0.64
Mean	3.64	0.76	0.48	0.98	0.83	0.77
Std	2.78	0.19	0.04	0.06	0.28	0.19

ference in objective value between iteration 50 and iteration 100 is very small compared to the objective value at the starting point.

Moving to the center part of Fig. 3, we see that the objective value after leaf-sequencing attains a minimum after 45 and 35 iterations for the Prostate A and Spinal cases, respectively. Note that no segment-weight optimization has been performed. Gradually, as the number of iterations increases, the profiles get jagged, and the conversion algorithm runs into difficulties reconstructing them without adding more segments. This leads to a deterioration of the dose distribution and an increase in the objective value.

The right part of Fig. 3 shows the objective values after segment-weight optimization, i.e., the objective values of the final plans. The objective values again tend to increase after a certain number of beamlet-weight iterations, although the curves are slightly flatter after segment-weight optimization than after leaf-sequencing. For the two cases, terminating the beamlet-weight optimization anywhere between 20 and 60 iterations seems to produce final plans with equal quality in terms of objective value. This is an important observation, since it indicates that it is not crucial to identify the optimal number of beamlet-weight iterations. Comparing with the objective values obtained with beamlet-weight optimization, we see that the best solution after segment-weight optimization, about 0.02, is obtained after only 10–15 iterations of beamlet-weight optimization. This can be explained by comparing the number of variables in these problems; it is easier to obtain a conform dose-distribution when using thousands of beamlets than when using less than 100 segments.

We want to compare plan quality of a treatment obtained by terminating the beamlet-weight optimization at an early stage with a treatment that has been "over-optimized." To produce the former, an estimation of the regularization parameter to our QN approach is made by finding the minimum in the objective value curve after the leaf-sequencing step for each patient case. We then set the number of iterations for early termination for all cases to the mean of these values, which is calculated to 35. The standard deviation is relatively large, 18, which will be commented on in the next paragraph. In comparison to the problem sizes, 35 iterations are very few, indicating a fast initial decrease in objective value for our QN approach. The "over-optimized" plan is produced by performing the leaf-sequencing after 100 beamlet-weight iterations. We argue that performing 100 iterations captures the features of "over-optimization," since no plans are optimized for this long using our QN approach in practice.

We thus compare the treatments obtained after 35 and 100 iterations of beamlet-weight optimization (denoted by τ_{35} and τ_{100}), with the treatments obtained after conversion and segment-weight optimization of τ_{35} and τ_{100} (denoted by ξ_{35} and ξ_{100}). Table III compares these treatments in terms of objective value F, total optimization time T, and amount of monitor units (MUs) for the clinical cases. The objective value of a single plan might not be a valid measure of the treatment quality. However, if an approach consistently finds lower objective values for given plans, it is preferable. Although the relative objective value is significantly larger with τ_{35} than with τ_{100}, the leaf-sequencing and segment-weight optimization change the situation so that the mean objective value for the ξ_{35} plans is 24% lower than the mean objective value for the ξ_{100} plans. This is obtained with half the optimization time and without increasing the amount of MUs, indicating that 100 beamlet-weight iterations are far too many. As mentioned above, the standard deviation of the regularization parameter is 18, which is relatively large. From the above observation that the final objective value of the right part of Fig. 3 has a flat minimum, we argue that it is not necessary to determine the regularization parameter precisely for each individual plan. To support this claim, we add two columns to Table III, showing the relative objective value for treatments obtained using 25 and 45 beamlet-weight iterations, respectively. We see that, in terms of mean objective value, the ξ_{35} treatment is similar to, but slightly

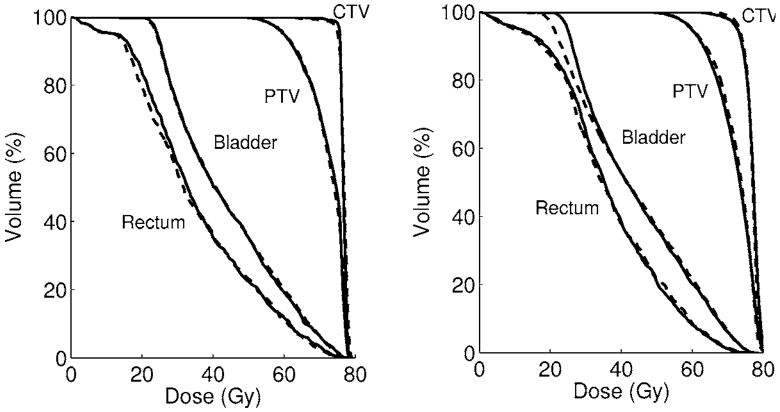

FIG. 4. Dose-volume histogram for the Prostate A case. Left+Dashed: τ_{35}. Left+Solid: τ_{100}. Right+Dashed: ξ_{35}. Right+Solid: ξ_{100}.

better than, both the ξ_{25} and ξ_{45} treatments. Hence, terminating the beamlet-weight problem in a relatively large interval around 35 is reasonable. If optimizing longer, the optimization time will increase without improving the objective value, and, if terminating earlier, the objective value will increase.

As a further measure of plan quality, we study DVHs and 2D dose distributions of the Prostate A case. The left part of Fig. 4 shows DVHs for τ_{35} (dashed) and τ_{100} (solid) for the CTV, PTV, bladder, and rectum. The uniformity in the CTV is very good for both treatments and, if any difference, it is slightly better with τ_{100}. There are very small differences between the DVHs of τ_{35} and τ_{100} for the bladder and rectum, especially above 40 Gy, which is the region the objective function will consider for these ROIs (see Table II). For the voxels outside the CTV but in the PTV, the minimum dose is allowed to be 50 Gy. This can be seen in the DVH for the PTV.

The right part of Fig. 4 shows the DVHs for the ξ_{35} (dashed) and ξ_{100} (solid) treatments for the same ROIs. The differences between the DVHs of ξ_{35} and ξ_{100} for the risk organs are very small, at least above 40 Gy. The maximum dose to the bladder is the same for both treatments, while the maximum dose to rectum is 0.7 Gy higher with ξ_{35} than with ξ_{100}. For the CTV, ξ_{35} seems to give slightly better uniformity than ξ_{100}. This is verified by the following comparisons. The percentage of the voxels in the CTV receiving more than 76 Gy is 1.9% higher with ξ_{35} than with ξ_{100}. The maximum dose to the CTV is 1.1 Gy lower and the minimum dose to the CTV is 1.3 Gy higher with ξ_{35} than with ξ_{100}. For the PTV, the maximum dose is 1.2 Gy lower and the minimum dose is 3.8 Gy lower with ξ_{35} compared to ξ_{100}.

Comparing all of the DVHs in Fig. 4, we see that the uniformity to the CTV is significantly better for the τ treatments than for the ξ treatments, while the DVHs for the risk organs are relatively similar.

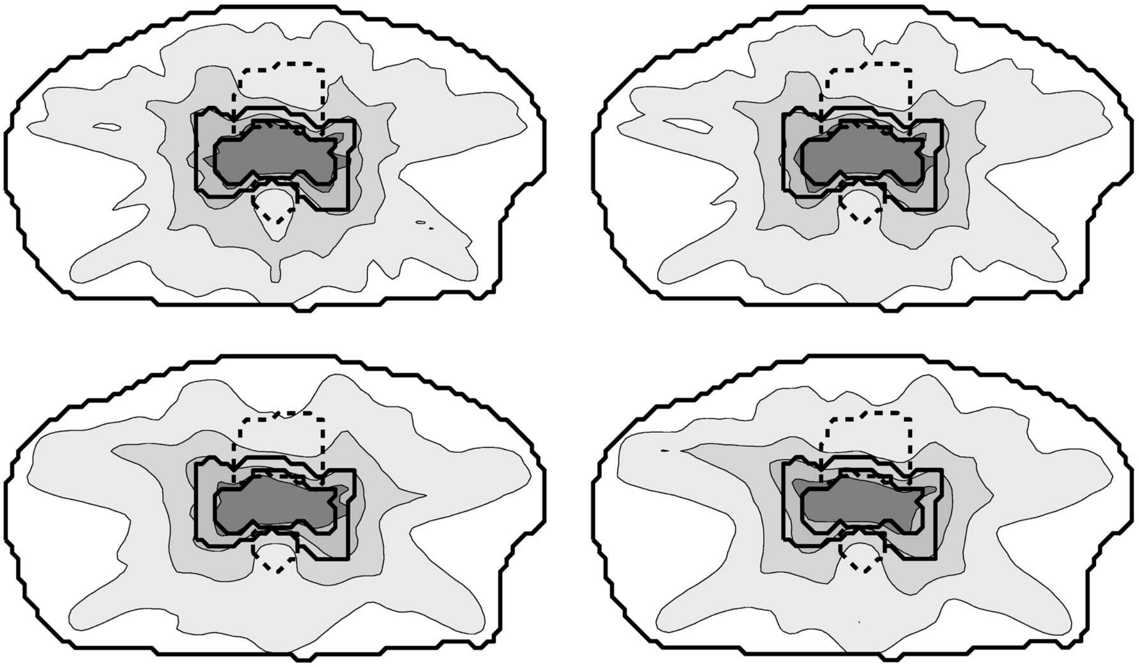

FIG. 5. Dose distribution for the Prostate A case, showing the CTV (inner solid), PTV (outer solid), bladder (upper dashed), and rectum (lower dashed). Contours at 20, 40, 64, and 74 Gy. Upper left: τ_{35}. Upper right: τ_{100}. Lower left: ξ_{35}. Lower right: ξ_{100}.

Figure 5 shows the dose distributions for the four treatments, together with the outline of the CTV (inner solid), PTV (outer solid), bladder (upper dashed), and rectum (lower dashed). The contours are placed at 20, 40, 64, and 74 Gy, which are dose levels that are relevant to the objective function (see Table II). Both τ_{35} (upper left) and τ_{100} (upper right) seem to produce very conform dose distributions, with sharp dose gradients outside the CTV. As expected, the conformity is slightly better for τ_{100}, which is reflected in the DVHs and the objective value (see Table III). Comparing the dose distribution of ξ_{35} (lower left) with ξ_{100} (lower right), the most obvious difference is the larger fraction of cold spots in the CTV with ξ_{100}. The differences are bigger when comparing the τ treatments (top figures) with the ξ treatments (bottom figures). The dose gradient around the CTV is smaller for the latter, indicating that it is harder to obtain a conform dose distribution when using segments than when using beamlets. We also see smoother contours in the bottom figures, more degrees of freedom introduce more jagged fluence profiles, and therefore more jagged isodose curves in the outer regions of the patient.

The results for the other nine clinical cases were very similar to the one reported above. Terminating the beamlet-weight problem early resulted in a slightly increased target uniformity of the deliverable plan. The other observed differences in the dose distributions of the ξ_{35} and ξ_{100} treatments were small.

IX. DISCUSSION AND CONCLUSION

We have discussed the suitability of a BFGS QN approach for solving beamlet-based IMRT problems with dose-volume objective functions in terms of iterative regularization. Empirically, the iterates tend to proceed along directions that initially solve the main conflicts in the plan, while keeping the intensity patterns smooth. Both these properties of the optimization method are crucial when performing iterative regularization. As with other regularization approaches, there is a regularization parameter to be determined. Our opinion is that the determination of our regularization parameter, the number of iterations, should be based on experience. We believe that it is not crucial to know the value of this parameter exactly since segment-weight optimization tends to diminish the differences in treatment quality between different beamlet-weight optimization iterates. This has been demonstrated on a set of clinical cases. Instead, we want to stress the importance of not "over-optimizing" the beamlet-weight optimization. An "over-optimization" effect can easily be obtained by warm-starting the plan over and over again, but utilizing a QN or CG method in this way will undoubtedly lead to jagged intensity patterns and deterioration of the final plan quality.

In Refs. 10 and 18, where the image reconstruction problem is considered, 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 the optimization time.

One way of avoiding the necessity of regularizing IMRT problems is to abandon the beamlet-based approach. By optimizing directly on the machine parameters and including the delivery constraints, each iterate can be delivered without any postprocessing and the ill-conditioning of the problem may be less problematic. Our future research will therefore focus on studying advanced aspects of already existing direct machine parameter optimization methods, e.g., including delivery time as a variable. Although this problem does not include beamlets explicitly, we believe that good understanding of the beamlet-based approach is of fundamental importance.

ACKNOWLEDGMENTS

We thank Lars Eldén, Philip Gill, Björn Hårdemark, Henrik Rehbinder, Göran Sporre, and Anders Szepessy for many helpful discussions and comments. We also thank the associate editor and the two anonymous referees for their constructive and insightful comments, which significantly improved the presentation. The research was supported by the Swedish Research Council (VR).

[a] Author to whom correspondence should be addressed.

[1] J. Löf, "Development of a general framework for optimization of radiation therapy," Ph.D. thesis, Stockholm University, 2000.

[2] M. Alber and F. Nüsslin, "Optimization of intensity modulated radiotherapy under constraints for static and dynamic MLC delivery," Phys. Med. Biol. **46**(12), 3229–3239 (2001).

[3] J. Seco, P. M. Evans, and S. Webb, "An optimization algorithm that incorporates IMRT delivery constraints," Phys. Med. Biol. **47**(6), 899–915 (2002).

[4] J. V. Siebers, M. Lauterbach, P. J. Keall, and R. Mohan, "Incorporating multi-leaf collimator leaf sequencing into iterative IMRT optimization," Med. Phys. **29**(6), 952–959 (2002).

[5] A. Brahme, "Optimization of stationary and moving beam radiation therapy techniques," Radiother. Oncol. **12**(2), 129–140 (1988).

[6] S. Webb, "Optimizing the planning of intensity-modulated radiotherapy," Phys. Med. Biol. **39**(12), 2229–2246 (1994).

[7] M. Alber, G. Meedt, F. Nüsslin, and R. Reemtsen, "On the degeneracy of the IMRT optimization problem," Med. Phys. **29**(11), 2584–2589 (2002).

[8] Y. Xiao, D. Michalski, Y. Censor, and J. M. Galvin, "Inherent smoothness of intensity patterns for intensity modulated radiation therapy generated by simultaneous projection algorithms," Phys. Med. Biol. **49**(14), 3227–3245 (2004).

[9] R. Mohan, M. Arnfield, S. Tong, Q. Wu, and J. Siebers, "The impact of fluctuations in intensity patterns on the number of monitor units and the quality and accuracy of intensity modulated radiotherapy," Med. Phys. **27**(6), 1226–1237 (2000).

[10] J. M. Bardsley, "A limited-memory, quasi-Newton preconditioner for nonnegatively constrained image reconstruction," J. Opt. Soc. Am. A **21**, 724–731 (2004).

[11] M. Hanke, "Iterative regularization techniques in image reconstruction," in *Surveys on Solution Methods for Inverse Problems* (Springer, Vienna, 2000), pp. 35–52.

[12] P. C. Hansen, "Numerical tools for analysis and solution of Fredholm integral equations of the first kind," Inverse Probl. **8**(6), 849–872 (1992).

[13] A. Ahnesjö and M. M. Aspradakis, "Dose calculations for external photon beams in radiotherapy," Phys. Med. Biol. **44**(11), R99–R155 (1999).

[14] A. V. Chvetsov, D. Calvetti, J. W. Sohn, and T. J. Kinsella, "Regularization of inverse planning for intensity-modulated radiotherapy," Med. Phys. **32**(2), 501–514 (2005).

[15] B. K. Lind, "Properties of an algorithm for solving the inverse problem in radiation therapy," Inverse Probl. **6**(3), 415–426 (1990).

[16] T. Bortfeld, J. Bürkelbach, R. Boesecke, and W. Schlegel, "Methods of image reconstruction from projections applied to conformation radiotherapy," Phys. Med. Biol. **35**(10), 1423–1434 (1990).

[17] L. Xing and G. T. Y. Chen, "Iterative methods for inverse treatment plan-

ning," Phys. Med. Biol. **41**(10), 2107–2123 (1996).

[18] J. M. Bardsley and C. R. Vogel, "A nonnegatively constrained convex programming method for image reconstruction," SIAM J. Sci. Comput. **25**(4), 1326–1343 (2003).

[19] L. Nazareth, "A relationship between the BFGS and conjugate gradient algorithms and its implications for new algorithms," SIAM J. Numer. Anal. **16**(5), 794–800 (1979).

[20] L. Eldén, "Partial least-squares vs. Lanczos bidiagonalization. I. Analysis of a projection method for multiple regression," Comput. Stat. Data Anal. **46**(1), 11–31 (2004).

[21] K.-H. Küfer, A. Scherrer, M. Monz, F. Alonso, H. Trinkaus, T. Bortfeld, and C. Thieke, "Intensity-modulated radiotherapy—a large scale multicriteria programming problem," OR Spectrum **25**, 223–249 (2003).

[22] F. Carlsson, A. Forsgren, H. Rehbinder, and K. Eriksson, "Using eigenstructure of the Hessian to reduce the dimension of the intensity modulated radiation therapy optimization problem," Report TRITA-MAT-2004-OS1, Department of Mathematics, Royal Institute of Technology, Stockholm, Sweden, 2004.

[23] M. Hanke, J. Nagy, and R. Plemmons, "Preconditioned iterative regularization for ill-posed problems," in *Numerical Linear Algebra*, (Kent, OH, 1992) (de Gruyter, Berlin, 1993), pp. 141–163.

[24] A. N. Tikhonov, "Regularization of incorrectly posed problems," Sov. Math. Dokl. **4**, 1624–1627 (1963).

[25] A. V. Chvetsov, "L-curve analysis of radiotherapy optimization problems," Med. Phys. **32**(8), 2598—2605 (2005).

[26] M. Alber and F. Nüsslin, "Intensity modulated photon beams subject to a minimal surface smoothing constraint," Phys. Med. Biol. **45**(5), N49–N52 (2000).

[27] S. V. Spirou, N. Fournier-Bidoz, J. Yang, C. Chui, and C. C. Ling, "Smoothing intensity-modulated beam profiles to improve the efficiency of delivery," Med. Phys. **28**, 2105–2112 (2001).

[28] S. Webb, D. J. Convery, and P. M. Evans, "Inverse planning with constraints to generate smoothed intensity-modulated beams," Phys. Med. Biol. **43**(10), 2785–2794 (1998).

[29] J. Llacer, N. Agazaryan, T. D. Solberg, and C. Promberger, "Degeneracy, frequency response and filtering in IMRT optimization," Phys. Med. Biol. **49**(13), 2853–2880 (2004).

[30] X. Sun and P. Xia, "A new smoothing procedure to reduce delivery segments for static MLC-based IMRT planning," Med. Phys. **31**(5), 1158–1165 (2004).

[31] M. M. Coselmon, J. M. Moran, J. D. Radawski, and B. A. Fraass, "Improving IMRT delivery efficiency using intensity limits during inverse planning," Med. Phys. **32**(5), 1234–1245 (2005).

[32] 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," Report SOL 86-2, Department of Operations Research, Stanford University , 1986.

[33] J. Nocedal and S. J. Wright, *Numerical Optimization* (Springer, New York, 1999).

[34] M. Alber and R. Reemtsen, "Intensity modulated radiotherapy treatment planning by use of a barrier-penalty multiplier method," to appear in Optim. Methods Softw.