Search for IMRT inverse plans with piecewise constant fluence maps using compressed sensing techniques

Lei Zhu^{a)} and Lei Xing

Department of Radiation Oncology, Stanford University, Stanford, California 94305

(Received 3 October 2008; revised 5 February 2009; accepted for publication 10 March 2009; published 27 April 2009)

An intensity-modulated radiation therapy (IMRT) field is composed of a series of segmented beams. It is practically important to reduce the number of segments while maintaining the conformality of the final dose distribution. In this article, the authors quantify the complexity of an IMRT fluence map by introducing the concept of sparsity of fluence maps and formulate the inverse planning problem into a framework of compressing sensing. In this approach, the treatment planning is modeled as a multiobjective optimization problem, with one objective on the dose performance and the other on the sparsity of the resultant fluence maps. A Pareto frontier is calculated, and the achieved dose distributions associated with the Pareto efficient points are evaluated using clinical acceptance criteria. The clinically acceptable dose distribution with the smallest number of segments is chosen as the final solution. The method is demonstrated in the application of fixed-gantry IMRT on a prostate patient. The result shows that the total number of segments is greatly reduced while a satisfactory dose distribution is still achieved. With the focus on the sparsity of the optimal solution, the proposed method is distinct from the existing beamlet- or segment-based optimization algorithms. © 2009 American Association of Physicists in Medicine. [DOI: [10.1118/1.3110163](https://doi.org/10.1118/1.3110163)]

Key words: radiation therapy, inverse planning, compressed sensing

I. INTRODUCTION

In intensity-modulated radiation therapy (IMRT), the treatment plan is selected from a large pool of physically feasible solutions by optimization of an objective function. The final solution depends on the choice of objective function and constraints applied to the optimization. Two commonly used approaches are beamlet- and segment-based optimizations. In the traditional beamlet-based algorithms for the step-and-shoot IMRT, each beamlet intensity is an independent and continuous variable. For a fast calculation, the nonconvex physical constraints of the dose delivery are not included in the optimization. As a result, the optimized beamlet intensity map has a high complexity, and the number of segments for dose delivery is usually large after leaf sequencing. A large number of segments reduces not only treatment efficiency but also treatment accuracy due to increased patient motion during beam delivery and the involvement of irregularly shaped segments. Many attempts have been made to reduce the fluence map complexity by using various data smoothing techniques.^{1–6} These algorithms smooth the edges and help get rid of spiky behaviors of fluence maps. However, the overall shapes of the final fluence maps remain the same and, as thus, the solution so obtained represents only a small perturbation to the original unsmoothed plan and the reduction of the number of segments is usually rather limited. Segment-based methods tackle the problem from the delivery aspect typically by enforcing a prechosen (often unjustified) number of segments for each incident beam and then optimizing the shapes and weights of the apertures.^{7–14} However, searching for an optimal solution by using segment-based optimization is inherently complicated because of the

highly nonconvex dependence of the objective function on the multi-leaf collimator (MLC) coordinates and the optimality of the final solution is not always guaranteed when an iterative algorithm is used.

An important characteristic that has not been utilized in most of inverse planning methods is that the IMRT solution space is highly degenerated in the sense that there are usually a large number of IMRT plans for the same prescription. While these plans yield similar dose distributions satisfying the prescription and constraints, the fluence maps of the plans can be dramatically different. Therefore, it is possible to stipulate constraints in the search of the optimal beamlet intensity such that the resultant number of segments is greatly reduced while the dose distribution is not severely deteriorated. In this work, instead of directly including the nonconvex physical constraints in the optimization, which is computationally intensive and increases the probability of being trapped in local optimal solutions, we propose an efficient method to achieve a global optimal solution only in a sparse space of fluence maps where the physical constraints are implied. The derivation is based on the fact that a beamlet intensity map which can be delivered using a small number of segments must be piecewise constant and its derivative is sparse. The problem is formulated as a multiobjective optimization, with an L-1 norm to enforce the sparsity of the solution, such that the number of beam segments is minimized, and a quadratic term to quantify the dose performance. Pareto efficient solutions are calculated, among which the clinically acceptable solution with the smallest number of beam segments is selected as the final solution. The performance of the proposed method is demonstrated using a prostate patient study.

The proposed algorithm can be regarded as an application of compressed sensing method in signal processing. Briefly, compressed sensing is a technique for acquiring and reconstructing a signal that is known to be sparse or compressible. A mathematical manifestation of a sparse signal is that it contains many coefficients close to or equal to zero when represented in some domain. Effective utilization of this prior knowledge of the system (i.e., the sparsity of the signal to be processed) can potentially reduce the required number of measurement samples (typically, this is determined by the classical Shannon-Nyquist theorem). Mathematically, IMRT inverse planning is analogous to the signal processing problem with the fluence maps being the “signal” to be detected for the given prescription doses. As mentioned above, inverse planning is an underdetermined problem and there are usually numerous fluence maps that are capable of yielding a clinically acceptable dose distribution. In this application, the sparsity of the derivative of the fluence maps makes compressed sensing a viable solution to the treatment planning. In reality, recovering or reconstructing sparse signals is generally a nonconvex problem, and therefore the computation is intense. However, recent development in the field of inverse problem shows that a mathematically heuristic sparse solution can be obtained using a convex optimization of an L-1 norm.^{15,16}

II. METHOD

II.A. Dose optimization without delivery constraints

The conventional beamlet-based optimization for inverse treatment planning is based on the linear relationship between the delivered dose distribution on the patient, d , and the intensity of the beamlets, x :

$$d = Ax, \quad (1)$$

where d is a vectorized dose distribution for a three-dimensional volume, and the beamlet intensity x is a one-dimensional vector that consists of row-wise concatenations of beamlet intensities for all fields. Each column of the matrix A is a beamlet kernel, corresponding to the dose distribution achieved by one beamlet with unit intensity. The beamlet kernels are precomputed based on the CT images of the patient, the treatment machine settings, and the beam geometry. In this work, we used the voxel-based Monte Carlo algorithm (VMC) as our dose calculation engine.¹⁷

For an efficient calculation, a convex function is usually used as an objective function in the optimization. If we use $\phi_1(x)$, the square of the L-2 norm of the difference between the delivered dose and the target dose as the objective function of x , the treatment planning problem can now be expressed as follows:

minimize

$$\phi_1(x) = \sum_i \lambda_i (A_i x - d_i)^T (A_i x - d_i) \quad (2)$$

subject to

TABLE I. Variable glossary.

$A(A_i)$	Matrix that relates the beamlet intensity to delivered dose
$d(d_i)$	Delivered dose
N	Total number of beamlets, $N=N_u N_v N_f$
N_f	Number of fields
N_t	Total number of segments of all fields
N_u	Number of MLC leaf positions for each leaf
N_v	Number of MLC leaf pairs per field
x	Beamlet intensity, the decision variable in the optimization
λ_i	Importance factor associated with the i th structure

$$x \geq 0,$$

where the index i denotes different structures, λ_i is the relative importance factor;^{4,18} each column of the matrix A_i is the beamlet kernel corresponding to the i th structure, and d_i is the prescribed dose. The main variables used in this paper are summarized in Table I for readers' reference.

II.B. Sparsity of fluence maps

The above optimization problem (2) does not consider dose delivery constraints of treatment machines. For MLC based IMRT delivery, two types of constraints on the segmented apertures are important. The first is the *uniformity constraint*, i.e., the intensity map of one beam aperture is uniform inside the MLC open area and zero elsewhere. The second is the *connectivity constraint*, i.e., the nonzero intensity areas of one beam aperture are connected in the direction of MLC leaf pairs.

The essence of compressed sensing methods is to utilize the prior knowledge that the signals of interest are sparse when represented in some domain. A fluence map is a summation of contributions from a series of segmented fields. If all the possible segments with different shapes are considered as the basis functions in a linear space, a fluence map with a small number of segments is a sparse presentation in such a space. Now the challenge is how to describe this sparsity mathematically and use it as an objective in the optimization. Fortunately, the sparsity of an actual fluence map can be easily quantified based on the uniformity constraint of apertures. As a summation of uniform intensity maps with different shapes, an actual fluence map is a piecewise constant function, which can be easily “sparsified” by taking derivatives. Define a gradient operator as

$$\nabla_{u,v} x(u,v) = |x_{u,v} - x_{u-1,v}| + |x_{u,v} - x_{u,v-1}|, \quad (3)$$

where the variables u (v) is the row (column) index of the beam intensity for each field.

The sparsity of a fluence map can be evaluated as the summation of the absolute values of the gradients, defined as

$$\phi_2(x) = \sum_{f=1}^{N_f} \sum_{u=2}^{N_u} \sum_{v=2}^{N_v} |\nabla_{u,v} x(u,v,f)|, \quad (4)$$

where the beamlet intensity map x is parametrized by the variables u , v , and f . The variable f is the field index. N_u is the total number of possible MLC leaf positions for each leaf; N_v is the total number of MLC leaf pairs per field; N_f is the number of fields. For simplicity, we assume that each treatment field has a rectangular shape when it is fully open, and N_u and N_v do not change for different fields. Note that $\phi_2(x)$ is the L-1 norm of the gradient, i.e., a total-variation function, which is commonly used as an objective function in many optimization applications to encourage a piecewise constant solution.^{19,20}

II.C. Search for an optimal solution in a sparse space using a multiobjective optimization

The aperture constraints are nonconvex and not included in the optimization step of the traditional beamlet-based methods, resulting in a large number of beam segments. In this paper, we reduce the number of segments without compromising the dose distribution by searching for solutions only in a sparse space of intensity maps. The sparsity of the intensity map is well correlated with the corresponding number of segments. The more sparse the optimized intensity map is, the less segments the leaf-sequencing algorithm derives. To enforce the sparsity on the optimized solution and therefore to reduce the number of segments, we include $\phi_2(x)$ as defined in Eq. (4) as a second objective function and reformulate the problem as a multiobjective optimization as follows:

$$\begin{aligned} & \text{minimize} \\ & \begin{cases} \phi_1(x) = \sum_i \lambda_i (A_i x - d_i)^T (A_i x - d_i), \\ \phi_2(x) = \sum_{f=1}^{N_f} \sum_{u=2}^{N_u} \sum_{v=2}^{N_v} |\nabla_{u,v} x(u,v,f)| \end{cases} \end{aligned} \quad (5)$$

subject to

$$x \geq 0.$$

Using an L-1 norm $\phi_2(x)$ as an objective, in fact, we solve the problem using compressed sensing techniques which are able to find heuristic sparse solutions.^{15,16}

The above formulation (5) is the main optimization framework proposed in this paper. The optimized beamlet intensity map, however, is close to but not exactly piecewise constant. Furthermore, the connectivity constraint due to the MLC hardware is not applied in the algorithm. A leaf-sequencing algorithm as in beamlet optimization is therefore needed to finally generate deliverable beam segments. Our multiobjective optimization does not post special requirements on the leaf-sequencing step and any existing leaf-sequencing algorithms can be used in combination with the proposed method.

II.D. Calculation of the Pareto frontier

The optimization of the multiobjective problem (5) is a trade-off between the dose performance and the total number of segments. If an upper limit constraint p is imposed on the first objective $\phi_1(x)$ and the minimization is carried out only on the second objective $\phi_2(x)$, the optimized solution gives the minimum number of segments that is required to achieve the dose performance defined by p . As the constraint p is relaxed or strengthened, the achieved minimum number of segments reduces or increases.

In order to obtain a final solution of the multiobjective optimization problem (5), we choose to first calculate the Pareto frontier and then select the solution which satisfies the clinical acceptance criteria with the smallest number of segments. The main reason is that some of the clinical goals are nonconvex and difficult to be included in the optimization as constraints.²¹ It is also difficult to find a proper value of upper limit p on the first objective $\phi_1(x)$, which is able to represent the clinical acceptance criteria. Visual inspections on the dose volume histograms (DVHs) and the dose distributions are therefore used to judge whether a certain plan is clinically acceptable.

The function $\phi_2(x)$ is not linear or quadratic. For an efficient calculation, we reformulate the optimization problem (5) into an equivalent form²² as follows:

$$\begin{aligned} & \text{minimize} \\ & \begin{cases} \phi_1(x) = \sum_i \lambda_i (A_i x - d_i)^T (A_i x - d_i), \\ \phi_2(x) = e^T t \end{cases} \end{aligned} \quad (6)$$

subject to

$$x \geq 0,$$

$$Bx - t \leq 0,$$

$$Bx + t \geq 0,$$

where e is an all-1 vector, with a size of $((N_u - 1)N_v N_f + N_u(N_v - 1)N_f) \times 1$, i.e., $e^T = (1, 1, 1, \dots, 1)$, $e \in R^{((N_u - 1)N_v N_f + N_u(N_v - 1)N_f) \times 1}$; the vector t is an intermediate

FIG. 1. The Pareto frontier of the multiobjective problem. T_1 and T_2 are the anchor points.

(a)

(b)

FIG. 2. The calculated Pareto frontiers of the prostate plans using different objectives. The derived number of segments (N_t) corresponding to each data point is marked on the plot. (a) The quadratic term on the dose distribution (ϕ_1) versus the L-1 norm of the fluence derivative (ϕ_2). (b) The quadratic term on the dose distribution (ϕ_1) versus the square of the L-2 norm of the fluence derivative (ϕ_3).

variable with the same size as e ; the matrix B is used to calculate the derivatives of x . Specifically,

$$B = \begin{bmatrix} B_u \\ B_v \end{bmatrix}, \quad (7)$$

where B_u is used to calculate the derivatives in the u direction:

$$B_u = \begin{bmatrix} C_1 & & & 0 \\ & C_2 & & \\ & & \ddots & \\ 0 & & & C_{N_v N_f} \end{bmatrix}. \quad (8)$$

C_i are identical, with a size of $(N_u - 1) \times N_u$:

(a)

(b)

FIG. 3. Optimized fluence maps for the fifth field. Optimization parameters are tuned such that both plans achieve the same dose distribution performance (roughly the same ϕ_1 values). (a) Using the L-1 norm of the fluence derivative (ϕ_2) in the optimization. Four segments are needed for this field. The total number of segments for all field is 35. (b) Using the square of the L-2 norm of the fluence derivative (ϕ_3) in the optimization. Twelve segments are needed for this field. The total number of segments for all field is 66.

$$C_i = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}. \quad (9)$$

B_v is used to calculate the derivatives in the v direction:

$$B_v = \begin{bmatrix} D_1 & & & 0 \\ & D_2 & & \\ & & \ddots & \\ 0 & & & D_{N_f} \end{bmatrix}. \quad (10)$$

D_i are identical, with a size of $N_u(N_v - 1) \times N_u N_v$:

$$D_i = \begin{bmatrix} -1 & 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & -1 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & \cdots & \cdots & 1 \end{bmatrix}. \quad (11)$$

On each row, -1 and 1 are separated by $N_u - 1$ zeros.

II.D.1. Calculation of anchor points

In order to obtain the Pareto frontier, we first fix $\phi_2(x)$ to a small value of s_1 and minimize $\phi_1(x)$ using the following quadratic optimization to obtain an objective value of p_1 :

minimize

$$\phi_1(x) = \sum_i \lambda_i (A_i x - d_i)^T (A_i x - d_i) \quad (12)$$

subject to

$$x \geq 0,$$

$$Bx - t \leq 0,$$

$$Bx + t \geq 0,$$

$$e^T t = s_1.$$

Repeat the optimization using a large $\phi_2(x)$ value of s_2 and obtain a minimized $\phi_1(x)$ value of p_2 . Thus, we find two anchor points on the Pareto frontier, T_1 and T_2 , as illustrated in Fig. 1.

The selection of s_1 and s_2 determines the search range of the Pareto frontier. In this work, these values are chosen empirically.

II.D.2. Calculation of the Pareto efficient points between anchor points

In order to calculate the complete Pareto frontier between the two anchor points, one solution is repeating the above optimization using different s values uniformly distributed between s_1 and s_2 . This approach, however, does not achieve uniformly distributed data points on the Pareto frontier due to its curvature. To calculate Pareto efficient points more uniformly on the Pareto frontier, we minimize the values of $\phi_1(x)$ and $\phi_2(x)$ along lines perpendicular to the line connecting T_1 and T_2 as shown in Fig. 1. Mathematically, the optimization is changed to be as follows:

minimize

$$\phi_2(x) = e^T t \quad (13)$$

subject to

$$x \geq 0,$$

$$Bx - t \leq 0,$$

$$Bx + t \geq 0,$$

$$\sum_i \lambda_i (A_i x - d_i)^T (A_i x - d_i) = g e^T t + h,$$

where the variable g is the slope of the lines perpendicular to the line $T_1 T_2$, $g = (s_2 - s_1) / (p_1 - p_2)$; the variable h is the intercept of these lines. Denote h_1 or h_2 as the intercept of the line passing through T_1 or T_2 , $h_{1,2} = p_{1,2} - g s_{1,2}$. The optimization is repeated for different values of h , which are chosen uniformly between h_1 and h_2 .

Note that the last constraint in the above formulation of optimization defines a nonconvex solution set, which makes the problem challenging. Fortunately, it can be verified that this constraint can be changed to be convex without affecting the solution. The optimization becomes a linear programming with linear and quadratic constraints, as follows:

minimize

$$\phi_2(x) = e^T t \quad (14)$$

subject to

$$x \geq 0,$$

$$Bx - t \leq 0,$$

$$Bx + t \geq 0,$$

$$\sum_i \lambda_i (A_i x - d_i)^T (A_i x - d_i) \leq g e^T t + h.$$

II.E. Evaluation

The proposed algorithm has been tested on a prostate patient. The algorithm was implemented in MATLAB using the MOSEK optimization software package (<http://www.mosek.com>). The anchor points of the Pareto frontier were first calculated using a standard quadratic optimization routine provided in MOSEK with an interior-point optimizer according to the problem formulation (12). Other Pareto efficient points were calculated using a linear programming with linear and quadratic constraints as shown in Eq. (14).

Five fields were used at angles of 35° , 110° , 180° , 250° , and 325° , based on a standard clinical protocol for prostate patients. Each field targeted the center of the planning target volume (PTV) and contained 20×16 beamlets, with a beamlet size of $5 \times 5 \text{ mm}^2$ at the source-to-axis distance (SAD). To save computation, the CT data were downsampled in the dose calculation, and the voxel size was $3.92 \times 3.92 \times 2.5 \text{ mm}^3$. The rectum, bladder, and femoral heads were included as sensitive structures. All the plans are normalized such that 95% of the PTV volume receives 100% prescribed dose (78 Gy).

To demonstrate the advantage of the proposed method, we also compare with the existing beamlet-based planning algorithm using quadratic smoothing (L-2 norm regularization).^{1,3,23} For a fair comparison, we still implement the algorithm as a multiobjective optimization and substitute a quadratic term (the square of the L-2 norm) $\phi_3(x)$ for the L-1 norm $\phi_2(x)$. Mathematically, $\phi_3(x)$ is defined as

FIG. 4. DVHs for the prostate plan using different total numbers of segments (N_t). The data are shown with one separate plot for each structure of interest. Note that a zoom-in insert is also included in the DVH plot of the PTV (a). (a) PTV; (b) Bladder; (c) Rectum; (d) Left femoral head; (e) Right femoral head.

$$\phi_3(x) = \sum_{f=1}^{N_f} \sum_{u=2}^{N_u} \sum_{v=2}^{N_v} ((x_{u,v} - x_{u-1,v})^2 + (x_{u,v} - x_{u,v-1})^2). \quad (15)$$

The Pareto frontier is calculated in a similar way as in the proposed algorithm.

III. RESULTS

Figure 2 compares the calculated Pareto frontiers of the prostate plan. Using the proposed algorithm with an L-1 norm as one objective, each Pareto efficient point of Fig. 2(a) took about 2 min on average on a 3 GHz PC to compute.

FIG. 5. Actual fluence maps of the second field for the prostate plan using different total numbers of segments (N_t). (a) $N_t=18$; (b) $N_t=23$; (c) $N_t=29$; (d) $N_t=35$; (e) $N_t=45$; (f) $N_t=48$.

The number of segments (N_t) corresponding to each Pareto efficient point after applying a leaf-sequencing algorithm is also marked on the plot. As discussed earlier, in general, a small (large) $\phi_2(x)$ value on the Pareto frontier achieves a small (large) number of segments, while the dose distribution

is degraded (improved), as indicated by the increase (decrease) in the $\phi_1(x)$ value. However, since the L-1 norm objective in our algorithm only implies the uniformity constraint of the apertures and the connectivity constraint is enforced by the subsequent leaf sequencing, the above relation-

TABLE II. Prostate plan goals and results. % vol > x Gy: percentage of the volume that receives more than x Gy dose; vol > x Gy: size of the volume that receives more than x Gy dose.

Regions	Acceptance criteria	$N_t=18$	$N_t=23$	$N_t=29$	$N_t=35$	$N_t=45$	$N_t=48$
PTV	% vol > 78 Gy ≥ 95	95.0	95.0	95.0	95.0	95.0	95.0
Rectum	% vol > 40 Gy ≤ 35	56.5	44.3	38.2	33.3	31.0	30.0
	% vol > 65 ≤ 17	13.8	12.9	10.7	9.8	9.7	9.4
	vol > 79.6 Gy ≤ 1 cc	0.50 cc	1.42 cc	1.27 cc	0.54 cc	0.81 cc	0.87 cc
Bladder	% vol > 40 Gy ≤ 50	46.5	38.8	29.1	24.3	21.3	19.1
	% vol > 65 Gy ≤ 25	11.1	9.3	8.1	7.9	7.5	6.9
Femoral heads	% vol > 45 Gy ≤ 1	0.08	0.30	0.20	0.15	0.03	0
Body	vol > 82.7 Gy ≤ 1 cc	0.65 cc	0.46 cc	1.61 cc	0.73 cc	0.96 cc	0.85 cc
	MUs (2 Gy fr)	334	334	342	343	347	350

ship is not exactly monotonic. As shown in Fig. 2(a), in some local areas (where $N_t=45, 43$), a larger $\phi_2(x)$ value achieves a smaller number of segments.

The calculated Pareto frontier using an L-2 norm square ($\phi_3(x)$) as one objective is shown in Fig. 2(b). Each Pareto efficient point is equivalent to a beamlet-based optimal plan using quadratic smoothing. For a better comparison, the algorithm parameters are tuned such that the Pareto frontiers shown in Fig. 2 have roughly the same range of $\phi_1(x)$ values. It is seen that, while the dose distribution performance is similar (as indicated by the close $\phi_1(x)$ values), the proposed algorithm using an L-1 norm is able to achieve a total number of segments much smaller than that using an L-2 norm. It is also worth noting that as the quadratic smoothing gets stronger ($\phi_3(x)$ values get smaller), the total number of segments of the optimized plan does not decrease. This indicates that although quadratic smoothing is able to reduce the complexity of the fluence maps, it smoothes the edges of the maps and does not efficiently reduce the number of beam segments. To further support the above argument, Fig. 3 shows the optimized fluence maps for the fifth field using the L-1 norm and the L-2 norm in the optimization. Both plans achieve almost the same dose distribution performance. However, using an L-1 norm as one objective achieves a nearly piecewise constant fluence map and only four segments are needed for this field. Instead, using an L-2 norm achieves a much smoothed fluence map and for this field, the resultant number of segments after leaf sequencing is 12.

Figure 4 shows the DVHs of the prostate plans corresponding to every other Pareto efficient point in Fig. 2(a). Each subfigure shows the DVH for one structure as N_t changes. Since the plans are normalized based on the dose distribution on the PTV, the DVHs of the PTV are very similar for different N_t . However, more organ at risk (OAR) volume is spared as N_t increases. Figure 5 shows the actual fluence maps of the second field for different total numbers of segments. As the number of segments increases, the complexity of the actual fluence map increases and the plan performance, especially the avoidance of the OARs, improves. The improvement slows down when the number of segments reaches a certain level. These plans are evaluated using clinical acceptance criteria and the results are summarized in Table II. The monitor units (MUs) per 2 Gy fraction are also

listed for each plan. The plan is satisfactory when the segment number is not less than 35, and the result using 35 segments is chosen as the final solution. Using the Eclipse planning system on the same patient data, the total number of segments is 61. Our method significantly reduces the number of segments without compromising the clinical performance of the treatment plan. The isodose distributions using different numbers of segments are shown in Fig. 6.

The total number of Pareto efficient points is mainly determined by the user-defined values of s_1 and s_2 as shown in Fig. 1. Note that, for a better illustration, the complete Pareto frontier was calculated in Fig. 2(a). In reality, it is not necessary to compute all the Pareto efficient points, and calculations of many clinically unacceptable Pareto efficient points can be avoided to improve the computation efficiency. For example, if the Pareto efficient points between the anchor points are calculated from a small ϕ_1 to a large ϕ_1 , the multiobjective optimization can stop when the plan first becomes clinically unacceptable, i.e., when $N_t=31$.

IV. DISCUSSION AND CONCLUSIONS

IMRT inverse planning is to obtain the best possible fluence profiles/maps that produce a desired/prescribed dose distribution. This is inherently an underdetermined problem and thus has no unique solution.²⁴ Indeed, in inverse planning, a clinically satisfactory dose distribution for a given case can generally be achieved using different sets of fluence maps. In other words, there are many fluence maps that can yield a sensible IMRT treatment plan. Each of these “optimal” solutions has its pros and cons. A practical challenge is to find the solution that best balances the conformality of the final dose distribution and the sparsity of the fluence maps. This paper provides an effective way of finding the optimal IMRT solutions with sparse or piecewise constant fluence maps.

Using compressed sensing, we model the planning as a multiobjective optimization problem, with one objective to quantify the dose performance and the other one to measure the sparsity of the solution. The algorithm has a form of convex optimization, with an ability to optimize the number of segments without compromising the dose performance in radiation therapy treatment. A method of calculating the

FIG. 6. Dose distributions of the prostate plan using different total numbers of segments (N_t). The isodose lines correspond to 95%, 65%, and 30% of the prescribed dose (78 Gy). The PTV and the sensitive structures (bladder, rectum, and femoral heads) are patched using different colors. The hotspots are marked using crosses. (a) $N_t=18$; (b) $N_t=23$; (c) $N_t=29$; (d) $N_t=35$; (e) $N_t=45$; (f) $N_t=48$.

Pareto frontier is also designed. Pareto efficient solutions are evaluated using clinical acceptance criteria, and the satisfactory plan with the smallest number of segments is chosen as the final solution. The performance of the algorithm is demonstrated using a prostate study. The result shows that the proposed method greatly reduces the number of segments without compromising the clinical performance of the treatment plan.

Calculation of the Pareto frontier is one solution to a multi-objective optimization problem. Other standard methods can also be used here. For example, we can combine the two objectives and consider the L-1 norm as a regularization term with a user-defined penalty weight of β .²⁵ The optimization problem is then converted to a quadratic programming. As shown in Fig. 7, the optimal solution obtained using this method is the Pareto efficient point on the Pareto frontier at

FIG. 7. The optimal point on the Pareto frontier if the optimization is solved using a regularization-based method.

which the tangent has a slope of $-\beta$. The optimal value of β can be determined by balancing the trade-off between the objectives. For example, some researchers use an L-curve analysis to first calculate the point of maximum curvature in Fig. 7 and then find the corresponding β as the optimal value.^{26,27} The multiobjective approach proposed in this paper provides a more general solution without introducing the parameter β .²¹

The traditional beamlet-based optimization method is sensible from a mathematical point of view, as it is conceptually intuitive, computationally tractable, and yields the best possible dose distribution for a given objective function. However, because of the complete ignorance of issues related to the MLC-based dose delivery, this approach usually results in a large number of segments and leads to a plan inefficient to delivery. The large number of segments using a traditional beamlet-based method is due to the high complexity of the optimized beamlet intensity map. In the literature, many algorithms have been proposed to ameliorate this problem using smoothing techniques.¹⁻⁶ Typical examples use an additional term of sum of derivative squares,^{1,3,23} which are often referred to as quadratic smoothing or regularization in the theory of convex optimization. Although these algorithms suppress the complexity of the beamlet intensity map, they do not achieve piecewise constant beamlet intensity maps. As shown in the comparisons of Figs. 2 and 3, the smoothing on the sharp edges at the aperture boundaries makes it difficult to further reduce the number of segments. In this paper, we formulate a general framework of multiobjective optimization with a focus on the piecewise constant feature of an actual fluence map and relate the number of segments to the sparsity of the derivative of a beamlet intensity map. Compressed sensing techniques are used to solve the problem, since it is able to achieve a heuristic sparse solution.^{15,16}

Segment-based optimization algorithms achieve small numbers of segments by imposing the physical constraints of beam apertures in the optimization.^{7-14,28} In a sense, this is similar to what many investigators have done in the context of 3D conformal therapy plan optimization, where the machine related parameters such as the beam weights and wedge angles are optimized. These algorithms eventually search in a space of all possible segments for a sparse opti-

mal solution. Since such a space is nonconvex, random search algorithms, such as simulated annealing, are commonly employed. The computation is therefore intensive and a global optimal solution is not always guaranteed. Furthermore, most of the segment-based methods prefix the total number of segments to limit the size of search space and increase the search efficiency. Roughly speaking, these methods calculate only one point on the Pareto frontier as shown in Fig. 1, of which the $\phi_2(x)$ value is defined by the prefixed number of segments. As a result, the solution is most likely not clinically optimal. In our method, we use compressed sensing to encourage a sparse solution and formulate a multiobjective optimization problem. The optimization is still convex, and therefore a Pareto optimal solution can always be obtained with a high computation efficiency.

In summary, a compressed sensing based inverse planning technique is proposed for IMRT planning. The main features of the approach include (1) the inclusion of a sparsity objective and (2) the use of a convex optimization algorithm. Without compromising the dose performance, IMRT solutions with piecewise constant fluence maps can be easily obtained using the proposed approach. The reduction of the number of segments in IMRT delivery reduces the total treatment time and therefore increases the clinical throughput. In addition, the fast dose delivery can also potentially improve the beam targeting by reducing the adverse influence of the patient organ motion during the treatment. As such, the proposed algorithm provides a practically attractive way to plan IMRT treatments.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Alexander Schlaefter for insightful discussion on multiobjective optimization. They also wish to thank Professor Zuwei Shen of Singapore National University (SNU) and Professor Stanley Osher of University of California, Los Angeles, for useful discussion during the Workshop on Mathematical Imaging and Digital Media held in June 2008 at the SNU. This project is supported in part by Grant No. 5R01CA98523 of National Cancer Institute and Grant No. PC040282 of Department of Defense.

^{a)}Electronic mail: leizhu@stanford.edu

¹M. Alber and F. Nsslin, "Intensity modulated photon beams subject to a minimal surface smoothing constraint," *Phys. Med. Biol.* **45**(5), N49-N52 (2000).

²L. Ma, "Smoothing intensity-modulated treatment delivery under hardware constraints," *Med. Phys.* **29**(12), 2937-2945 (2002).

³S. V. Spirou, N. Fournier-Bidoz, J. Yang, C. S. Chui, and C. C. Ling, "Smoothing intensity-modulated beam profiles to improve the efficiency of delivery," *Med. Phys.* **28**(10), 2105-2112 (2001).

⁴S. 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).

⁵X. 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).

⁶Y. 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).

- ⁷D. M. Shepard, M. A. Earl, X. A. Li, S. Naqvi, and C. Yu, "Direct aperture optimization: A turnkey solution for step-and-shoot imrt," *Med. Phys.* **29**(6), 1007–1018 (2002).
- ⁸D. Michalski, Y. Xiao, Y. Censor, and J. M. Galvin, "The dose-volume constraint satisfaction problem for inverse treatment planning with field segments," *Phys. Med. Biol.* **49**(4), 601–616 (2004).
- ⁹C. Cotrutz and L. Xing, "Segment-based dose optimization using a genetic algorithm," *Phys. Med. Biol.* **48**(18), 2987–2998 (2003).
- ¹⁰B. van Asselen, M. Schwarz, C. van Vliet-Vroegindeweij, J. V. Lebesque, B. J. Mijnheer, and E. M. F. Damen, "Intensity-modulated radiotherapy of breast cancer using direct aperture optimization," *Radiother. Oncol.* **79**(2), 162–169 (2006).
- ¹¹J. L. Bedford and S. Webb, "Direct-aperture optimization applied to selection of beam orientations in intensity-modulated radiation therapy," *Phys. Med. Biol.* **52**(2), 479–498 (2007).
- ¹²A. M. Bergman, K. Bush, M.-P. Milette, I. A. Popescu, K. Otto, and C. Duzenli, "Direct aperture optimization for imrt using monte carlo generated beamlets," *Med. Phys.* **33**(10), 3666–3679 (2006).
- ¹³A. Mestrovic, M.-P. Milette, A. Nichol, B. G. Clark, and K. Otto, "Direct aperture optimization for online adaptive radiation therapy," *Med. Phys.* **34**(5), 1631–1646 (2007).
- ¹⁴H. E. Romeijn, R. K. Ahuja, J. F. Dempsey, and A. Kumar, "A column generation approach to radiation therapy treatment planning using aperture modulation," *SIAM J. Optim.* **15**(3), 838–862 (2005).
- ¹⁵D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory* **52**, 1289–1306 (2006).
- ¹⁶E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory* **52**(2), 489–509 (2006).
- ¹⁷I. Kawrakow, "Improved modeling of multiple scattering in the voxel monte carlo model," *Med. Phys.* **24**(4), 505–517 (1997).
- ¹⁸S. Breedveld, P. R. M. Storchi, M. Keijzer, and B. J. M. Heijmen, "Fast, multiple optimizations of quadratic dose objective functions in IMRT," *Phys. Med. Biol.* **51**(14), 3569–3579 (2006).
- ¹⁹K. T. Block, M. Uecker, and J. Frahm, "Undersampled radial mri with multiple coils. Iterative image reconstruction using a total variation constraint," *Magn. Reson. Med.* **57**(6), 1086–1098 (2007).
- ²⁰V. Kolehmainen, A. Vanne, S. Siltanen, S. Järvenp, J. P. Kaipio, M. Lassas, and M. Kalke, "Parallelized bayesian inversion for three-dimensional dental x-ray imaging," *IEEE Trans. Med. Imaging* **25**(2), 218–228 (2006).
- ²¹A. Schlaefer and A. Schweikard, "Stepwise multi-criteria optimization for robotic radiosurgery," *Med. Phys.* **35**(5), 2094–2103 (2008).
- ²²S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, 2004).
- ²³M. M. Matuszak, E. W. Larsen, and B. A. Fraass, "Reduction of imrt beam complexity through the use of beam modulation penalties in the objective function," *Med. Phys.* **34**(2), 507–520 (2007).
- ²⁴Y. Yang and L. Xing, "Clinical knowledge-based inverse treatment planning," *Phys. Med. Biol.* **49**(22), 5101–5117 (2004).
- ²⁵L. Zhu, L. Lee, Y. Ma, Y. Ye, R. Mazzeo, and L. Xing, "Using total-variation regularization for intensity modulated radiation therapy inverse planning with field-specific numbers of segments," *Phys. Med. Biol.* **53**(23), 6653–6672 (2008).
- ²⁶A. V. Chvetsov, "L-curve analysis of radiotherapy optimization problems," *Med. Phys.* **32**(8), 2598–2605 (2005).
- ²⁷A. V. Chvetsov, J. F. Dempsey, and J. R. Palta, "Optimization of equivalent uniform dose using the l-curve criterion," *Phys. Med. Biol.* **52**(19), 5973–5984 (2007).
- ²⁸J. Dai and W. Que, "Simultaneous minimization of leaf travel distance and tongue-and-groove effect for segmental intensity-modulated radiation therapy," *Phys. Med. Biol.* **49**(23), 5319–5331 (2004).