A fast optimization algorithm for multicriteria intensity modulated proton therapy planning

Wei Chen[a)]
Department of Computer Science, Graduate Center, City University of New York, New York, New York 10016

David Craft,[b)] Thomas M. Madden, Kewu Zhang, and Hanne M. Kooy
Department of Radiation Oncology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114

Gabor T. Herman
Department of Computer Science, Graduate Center, City University of New York, New York, New York 10016

(Received 29 March 2010; revised 28 June 2010; accepted for publication 16 July 2010; published 26 August 2010)

Purpose: To describe a fast projection algorithm for optimizing intensity modulated proton therapy (IMPT) plans and to describe and demonstrate the use of this algorithm in multicriteria IMPT planning.
Methods: The authors develop a projection-based solver for a class of convex optimization problems and apply it to IMPT treatment planning. The speed of the solver permits its use in multicriteria optimization, where several optimizations are performed which span the space of possible treatment plans. The authors describe a plan database generation procedure which is customized to the requirements of the solver. The optimality precision of the solver can be specified by the user.
Results: The authors apply the algorithm to three clinical cases: A pancreas case, an esophagus case, and a tumor along the rib cage case. Detailed analysis of the pancreas case shows that the algorithm is orders of magnitude faster than industry-standard general purpose algorithms (MOSEK's interior point optimizer, primal simplex optimizer, and dual simplex optimizer). Additionally, the projection solver has almost no memory overhead.
Conclusions: The speed and guaranteed accuracy of the algorithm make it suitable for use in multicriteria treatment planning, which requires the computation of several diverse treatment plans. Additionally, given the low memory overhead of the algorithm, the method can be extended to include multiple geometric instances and proton range possibilities, for robust optimization. [DOI: 10.1118/1.3481566]

Key words: projection method, multi-criteria, optimization, numerical evaluation

I. INTRODUCTION

Multicriteria optimization (MCO) is a powerful technique for optimizing intensity modulated radiation therapy (IMRT) treatment plans. The standard single criterion approach to treatment planning uses weights and dose volume histogram (DVH) control points to steer toward desired plans. The results of the underlying optimizations are, however, hard to control and often result in time-consuming iterations among the treatment planner, the treatment planning system, and the physicians.[1] In contrast, MCO uses a precalculated database of treatment plans that span the viable treatment planning options[2] and a user interface that allows the treatment planner to navigate the approximated Pareto surface to select the desirable plan from a blend of database plans that suits the treatment goals for the patient.[3–5] Such a system for IMRT planning is currently under evaluation at the Massachusetts General Hospital.[6]

The focus of the present work is a new algorithm that is very time and memory efficient and is thus suitable for MCO treatment planning for intensity modulated proton therapy (IMPT). Structurally, the mathematical problems of IMRT and IMPT planning are the same. Both require large-scale optimizers to choose the optimal set of beamlet intensities that map to voxel doses. An IMPT optimization problem has, however, on the order of three to ten times more decision variables (beamlets) than the corresponding IMRT optimization problem, depending on the pencil-beam size and the number of energy layers delivered. Additionally, a Pareto surface based MCO requires the computation of multiple treatment plans to span the space of treatment possibilities.[3] Therefore, to apply MCO to this setting requires a fast and reliable optimization algorithm.

We use a linear problem formulation for the IMPT MCO implementation. It is often the case that algorithms for linear programming (LP) problems (or, more generally, for convex constrained optimization problems) are slower than gradient based penalty function methods applied to the related global optimization problems because LP algorithms store and utilize the exact geometry of a polyhedral feasible set and solve to exact provable optimality. In certain cases, depending on the problem structure and on the nature of the constraints,

linear formulations are solved rapidly with custom algorithms. This is the case for radiation therapy planning problems, where we demonstrate that a projection-based algorithm can solve linear instances fast and with little memory overhead.

Projection algorithms[7] are a family of algorithms for convex feasibility problems and convex constrained optimization problems. These algorithms iteratively project the current solution onto violated individual constraints. In linear settings, the constraints are hyperplanes (linear equalities), half spaces (linear inequalities), and hyperslabs (linear interval inequalities). The algebraic operation of a projection into or onto a linear constraint is simple and fast. Projection algorithms have been successfully used in areas such as image reconstruction from projections[8–10] and IMRT planning.[11–15]

The projection algorithm used in this paper, ART3+, is a finitely convergent sequential projection algorithm (finitely convergent means that it is guaranteed to converge to a feasible point in a finite number of iterations, as opposed to in the limit).[14,16] ART3 (Ref. 8) is an algebraic reconstruction technique for solving linear interval inequalities and ART3+ (Ref. 14) is a faster version of ART3, with the significant speed improvement stemming from a repetitive (noncyclic) control for selecting the constraints. In each iteration of either ART3 or ART3+, a constraint is selected and checked. If the constraint is violated, then the current iterate is projected into or onto the constraint, making the point feasible for that constraint. The difference between the algorithms is in their behavior if the selected constraint is not violated. In ART3+, such a constraint is temporarily removed from the list of constraints, and thus computer time is saved by not selecting and checking constraints that are not likely to be violated (see Refs. 14 and 16). Both ART3 and ART3+ are finitely convergent, provided that the feasible solution set is full-dimensional. Due to the speed of the ART3+ algorithm on IMPT planning problem instances, it is efficacious to also make use of it for optimization. For example, to minimize $c'x$ we convert the objective into a constraint $c'x \leq r$ and iteratively reduce r until convergence to a solution within a specified optimality tolerance. (As common, c' denotes the row vector that is the transpose of the column vector c.) We call the new optimization algorithm ART3+O, where "O" stands for "optimization."

ART3+O requires that problems have only linear constraints and either a linear objective or a certain type of convex objective, such as minimizing the maximum dose delivered to an organ at risk (OAR). Although this is restrictive, the nature of MCO allows us to be fairly flexible regarding the exact formulation of the problem (meaning the functional form of the constraints and objectives).[17,18] In a MCO setting, if two functions are correlated, such as equivalent uniform dose[19] (EUD) and normal tissue complication probability, it is mathematically equivalent to use either in a MCO formulation, and so the simpler one would be used. When two functions are effectively (but not mathematically) correlated, then the same reasoning applies and one can therefore choose the simpler function to optimize. For these reasons, the linear restrictions imposed by the ART3+O solver do not affect the quality of the treatment plans that get produced.

We compare the performance of our implementation of ART3+O to three commercially available state-of-the-art implementations of LP algorithms from MOSEK ApS (see http://www.mosek.com and Ref. 20 for background): An interior point algorithm, a primal simplex algorithm, and a dual simplex algorithm. ART3+O is shown to outperform all of these algorithms in speed and memory usage.

In order to apply the multicriteria radiation therapy planning approaches of Refs. 3 and 5, we need to generate a database of treatment plans that approximate the Pareto surface. This can be done using ART3+O as long as the database generation is based on constraints (as in Refs. 4 and 21) rather than on weighted sums of the underlying objectives (as in Refs. 3 and 22). The final step in the MCO treatment planning process, the selection of the convex combination of the database plans with which to treat the patient, is called the navigation step. We implement a procedure very similar to that given in Ref. 5 and refer the reader there for details.

II. METHODS

II.A. The optimization algorithm ART3+O

By a *linear feasibility problem* we mean here a problem of the following form: Given $B \in \mathbb{R}^{I \times J}$, $l \in (\mathbb{R} \cup \{-\infty\})^I$, and $u \in (\mathbb{R} \cup \{+\infty\})^I$, find an $x \in \mathbb{R}^J$ that satisfies the hyperslab constraints

$$l \leq Bx \leq u. \tag{1}$$

For $1 \leq i \leq I$, we use B_i to denote the column vector in $\mathbb{R}^J$ that is the transpose of the ith row of the matrix B and throughout this paper we assume that $\|B_i\| \neq 0$.

ART3+ solves the above feasibility problem (provided that its set of solutions is full-dimensional[14]) by sequentially projecting the current iterate x onto or into the violated hyperslab constraints, one at a time. If x is farther away from the hyperslab than half the thickness of the hyperslab, then x is projected to the center hyperplane of the hyperslab; otherwise, x is reflected across the nearer of the two bounding hyperplanes of the hyperslab. More precisely, for a constraint $l_i \leq B_i'x \leq u_i$, the ART3 projection operator $P_i(x)$, which provides us the next iterate, is

$$P_i(x) = x - \begin{cases} \dfrac{B_i'x - (l_i+u_i)/2}{\|B_i\|^2}B_i, & \text{if } B_i'x < l_i - (u_i-l_i)/2 \text{ or } u_i + (u_i-l_i)/2 < B_i'x, \\ 2\dfrac{B_i'x - l_i}{\|B_i\|^2}B_i, & \text{if } l_i - (u_i-l_i)/2 \le B_i'x < l_i, \\ 2\dfrac{B_i'x - u_i}{\|B_i\|^2}B_i, & \text{if } u_i < B_i'x \le u_i + (u_i-l_i)/2, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

While ART3 visits all the constraints cyclically, ART3+ has a more complicated mechanism to determine which constraint to pick in each iteration in order to speed up the finite convergence. In ART3+, an ordered set of violated constraints is maintained by removing from it the currently picked constraint if it is satisfied by the current iterate x. When this ordered constraint set becomes empty, it is filled up by the complete set of constraints, unless all the constraints are satisfied, in which case ART3+ stops. Full details of the ART3+ algorithm for solving the linear feasibility problem without infinite bounds can be found in Refs. 14 and 16. The proof of finite convergence of ART3+ therein can be generalized easily to apply to the linear feasibility problem with infinite bounds.

In order to take care of convex objectives (such as minimization of the maximum dose delivered to an OAR) that occur in IMPT planning, we specify below a convex optimization problem, which has the linear optimization problem with an objective "minimize $c'x$" as a special case. In order to do this, we need to introduce some notation. For any $C \in \mathbb{R}^{K\times J}$ and $x \in \mathbb{R}^J$, we define

$$f_C(x) = \max\{C_1'x, C_2'x, \ldots, C_K'x\}. \tag{3}$$

Here, for $1 \le k \le K$, C_k is the column vector in $\mathbb{R}^J$ that is the transpose of the kth row of the matrix C and we assume throughout this paper that $\|C_k\| \neq 0$. When $K=1$, the objective $f_C(x)=C_1'x$ is a linear function. If an OAR has K voxels, then C can be defined so that $C_k'x$ is the dose delivered to the kth of these K voxels and so that $f_C(x)$ is the maximum dose delivered to the OAR. (Since these are doses, we point out that throughout this paper, we use Gy as the unit of dose.) Then the convex optimization problem is of the following form: Given $B \in \mathbb{R}^{I\times J}$, $l \in (\mathbb{R}\cup\{-\infty\})^I$, $u \in (\mathbb{R}\cup\{+\infty\})^I$, and $C \in \mathbb{R}^{K\times J}$, find an $x \in \mathbb{R}^J$ that minimizes $f_C(x)$, subject to $l \le Bx \le u$.

To solve this problem with ART3+ we define, for every $r \in \mathbb{R}$, $R(r)$ to be the linear feasibility problem: Given $B \in \mathbb{R}^{I\times J}$, $l \in (\mathbb{R}\cup\{-\infty\})^I$, $u \in (\mathbb{R}\cup\{+\infty\})^I$, and $C \in \mathbb{R}^{K\times J}$, find an $x \in \mathbb{R}^J$ such that

$$C_1'x \le r,$$

$$C_2'x \le r,$$

$$\cdots$$

$$C_K'x \le r,$$

$$l \le Bx \le u. \tag{4}$$

Note that a solution x of $R(r^*)$, where r^* is the smallest r for which $R(r)$ has a solution, is in fact a solution of the convex optimization problem of the previous paragraph. We assume that there exists an $r_{\min}$ for which $R(r_{\min})$ does not have a solution and that such an $r_{\min}$ is available to us. In Sec. II B, we explain why such an assumption is valid and how to find $r_{\min}$ in practice. A small positive real number ϵ is specified as the desired optimality tolerance. We also use the parameter Q to denote the maximum number of iterations in any execution of ART3+.

The algorithm ART3+O is the following procedure:

Initialization. Use ART3+ to solve $l \le Bx \le u$. If a solution is found in Q or fewer iterations, we assign that solution to x^0 and x^* and let $r_{\max}=f_C(x^0)$ and $t=0$. Otherwise, we consider the optimization problem unsolvable.

ART3+. Let $r^t=(r_{\min}+r_{\max})/2$. Run ART3+ on $R(r^t)$ starting with x^t.

(i) If ART3+ finds a solution in Q or fewer iterations, then assign the solution to x^{t+1} and x^* and let $r_{\max}=f_C(x^{t+1})$.
(ii) Otherwise, let $r_{\min}=r^t$ and x^{t+1} be the Qth iterate of ART3+.

Termination. If $r_{\max}-r_{\min} \le \epsilon$, then return x^* as the solution to the optimization problem, otherwise increase t by 1 and repeat ART3+.

The idea of the algorithm is to search for the r^* as defined below Eq. (4). We first find a range $[r_{\min}, r_{\max}]$ that is big enough to contain r^*. After this we apply a bisection search on the decreasing range $[r_{\min}, r_{\max}]$ until the length of this range is ϵ or less, which will happen for a $t \le 2\lceil \log_2(r_{\max}-r_{\min})/\epsilon \rceil$. ART3+ will return a solution x for which $f_C(x)-r^* \le \epsilon$, provided that in each call to it, ART3+ stops in at most Q iterations for the problem $R(r)$ that has a solution.

To achieve the same goal in MOSEK, we have to take r in Eq. (4) as an auxiliary unknown variable and solve the LP problem: Minimize r, subject to Eq. (4).

II.B. Clinical MCO formulation and database generation process

Let $D \in \mathbb{R}^{H\times J}$ denote the dose-influence matrix. Thus, D_{hj} is the dose contribution to voxel h $(1 \le h \le H)$ from a unit intensity in beamlet j $(1 \le j \le J)$. If $x \in \mathbb{R}^J$ denotes the beamlet intensity vector, then the dose vector is given by Dx. For MCO database generation for IMPT, the problems we solve are of the following form:

$$\begin{aligned} &\text{minimize} \quad g(Dx), \\ &\text{subject to} \quad l \le Dx \le u, \\ &\qquad\qquad x \ge 0, \end{aligned} \tag{5}$$

where the objective g can be the mean dose to a structure or a set of structures, the negative mean dose to a target or a set of targets (maximizing a function is the same as minimizing the negative of that function), the maximum dose to a structure or a set of structures, or the negative minimum dose to a target or a set of targets. In all the cases, we can find a $K \in \mathbb{Z}$ and a $C \in \mathbb{R}^{K\times J}$ such that $f_C(x)=g(Dx)$. We call a plan x feasible if it satisfies the inequalities in Eq. (5).

The $r_{\min}$ in our algorithm can be found by considering the problem and the constraints at hand. For example, if $g(Dx)$ is the mean dose to an OAR, we choose $r_{\min}=-0.01$. The upper bound of the number of calls to ART3+ is $2\lceil \log_2(r_{\max} - r_{\min})/\epsilon \rceil$, which is a very modest number since the optimality tolerance ϵ required for radiation therapy planning is not very small. For each call of ART3+, we use $Q=2\times 10^7$.

For multicriteria radiation therapy planning,[3,5] we first specify N objective functions g and for each, we find an optimal plan x, as specified by Eq. (5). We generate a database (to be used in the navigation step[5]) that contains these plans and some additional ones. The additional plans are determined by a variant, which we specify in the next paragraph, of the bounded objective function method.[23] Since each additional plan is obtained by solving a problem that is of the same form as Eq. (5), we again make use of ART3+O.

Let $\bar{x}$ denote the average of the N optimal plans. This averaged plan is used as the source of N additional constraints of the form $g(Dx) \le g(D\bar{x})$ (one for each of the original objective functions) for all the subsequent optimization tasks. This extended set of constraints is feasible, since $\bar{x}$ satisfies it. The additional plans in the database are obtained by solving, subject to the extended set of constraints, the following optimization problems:

(i) Minimize the sum of the mean doses of the structures that appeared in all of the original "minimize mean dose"-type objectives;
(ii) Maximize the sum of the mean doses of the targets that appeared in all of the original "maximize mean dose"-type objectives;
(iii) Repeat all of the original "minimize maximum dose" optimizations and "maximize minimum dose" optimizations.

The resulting database contains more than N but no more than $2N$ plans.

III. RESULTS

III.A. Numerical results on the algorithm

We use a pancreas case to demonstrate the advantages of ART3+O over the three general purpose LP algorithms implemented in MOSEK for the problem in the last paragraph of Sec. II A. The patient volume (302 491 voxels) consists of liver, stomach, left kidney, right kidney, the planning target volume (PTV), and skin (all remaining voxels). We select Rx (the prescription dose) to be 59.4 Gy (all Gy values reported include a relative biological effectiveness, factor of 1.1), the maximum dose to all structures to be 1.12Rx, and the minimum dose for the PTV to be 0.95Rx. We use three proton beams (the energy layers, set by in-house IMPT software, yield Bragg peaks approximately 5 mm apart, depending on the depth) yielding 13 734 beamlets in total. The dose-influence matrix D has 62 226 127 nonzero elements. After preprocessing, we use 144 411 constraints in our optimization tasks. We perform the following optimizations for the numerical studies: [Task0] minimize mean skin dose; [Task1] maximize minimum PTV dose; [Task2] minimize mean liver dose; [Task3] minimize mean stomach dose; [Task4] minimize mean left kidney dose; [Task5] minimize mean right kidney dose; [Task6] set the minimum PTV dose to Rx and minimize the sum of the mean doses of the liver, stomach, left kidney, and right kidney; and [Task7] set the minimum PTV dose to Rx and minimize the overall maximum dose. We use $\epsilon=0.1$ Gy for the optimality tolerance, which is well within radiation delivery precision. (For Task6 and Task7, we have tightened the constraint set by setting the PTV minimum dose to Rx. This was done for historical reasons and it does not matter for the comparison of the performance of algorithms reported in this subsection. For the clinical demonstrations in the Sec. III B, we use the procedure exactly as described in Sec. II.) We use [NoTask] to refer to the initial feasibility run. Although we only optimize the mean dose to the OARs in this pancreas case, minimizing the maximum dose to a serial critical organ can also be done by an appropriate choice of the function g in Eq. (5).

All the experiments run on a 64-bit Linux computer with 2.66 GHz Quad core Intel Xeon CPU and 16 Gbyte memory (but only one core of the CPU is used). To demonstrate the search process of ART3+O, we give as an example the run of Task6; the results are shown in Table I.

The timing results of ART3+O and the three standard LP algorithms, for the feasibility run and all eight tasks, are shown in Table II. ART3+O is far faster than the three algorithms in MOSEK. Typically, for each task, ART3+O uses about 1–2 min and MOSEK uses 1 h to several hours. These results are not biased by the fact that ART3+O finds an ϵ-optimal value, while the algorithms in MOSEK find a better approximation to the true optimum for the following reasons: First, it is hard for MOSEK to stop earlier at an ϵ-optimal value and, at the same time, find an exactly feasible solution like ART3+O does; second, even if we only consider the

TABLE I. An example run of ART3+O with Task6 of the pancreas case.

t	$[r_{\min}, r_{\max}]$	r^t	Timing (s)	Find a solution in Q iterations
			1.36	Yes
0	[00.000, 43.809]	21.904	31.35	No
1	[21.904, 43.809]	32.857	4.08	Yes
2	[21.904, 24.972]	23.438	11.55	Yes
3	[21.904, 23.437]	22.671	31.55	No
4	[22.671, 23.437]	23.054	14.04	Yes
5	[22.671, 23.051]	22.861	31.08	No
6	[22.861, 23.051]	22.956	2.72	Yes
	[22.861, 22.927]			

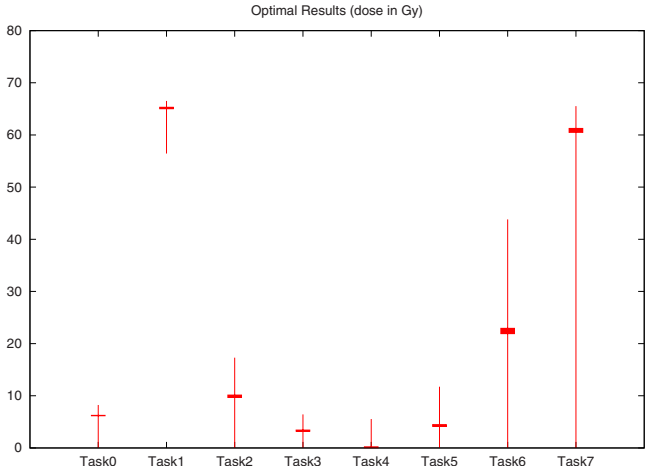

FIG. 1. The vertical line segments are the initial ranges of r. The thickness of the horizontal segments is the gap between the optimal r found by ART3+O and the optimal r given by MOSEK.

optimality, ART3+O still arrives at an ϵ-optimal value much faster than the algorithms in MOSEK (e.g., for Task0, ART3+O gets to 6.25 Gy in 90.75 s, the interior point method of MOSEK gets to 6.76 Gy in 2099.99 s and it gets to 6.17 Gy, which it returns as the optimum, in 4718.72 s). The total time to generate the plan database using ART3+O is on the order of 10 min. (We note, by the way, that the speedup ratio from ART3 to ART3+ is about 2–7 in our experiments, which is much greater than what is reported in Ref. 14 due to the larger size of our current problems.)

Memory usage is an important consideration for radiation therapy planning algorithms because the advanced models that take into consideration uncertainty and organ motion require much larger optimization instances. The ART3+O algorithm has almost no memory overhead: Its memory usage is just slightly over the memory required to hold the dose-influence matrix D and the (much smaller) dose bounds l and u. The general purpose algorithms come with a large memory overhead, as shown in Table II: ART3+O uses less than a tenth of the memory needed by the other algorithms.

Given the fact that the planned treatment dose will not be delivered precisely (most institutions assume at least a 2% error in planned versus delivered dose on a per voxel basis), there is no need to run the optimizations to a high degree of optimality. Furthermore, running algorithms to suboptimality is typically much faster than running to exact optimality. Using the results from the MOSEK algorithms as the "true" optimum, Fig. 1 displays the ART3+O results in comparison. The results are well within the delivery precision for protons and could be relaxed further for additional computation speed improvements.

III.B. Clinical case demonstrations

We demonstrate the full system (database generation and navigation) on two clinical cases, and in doing so we show how to choose constraints and objectives in our MCO setting that allows only hard constraints and certain kinds of objective functions. Since treatment planning today is typically done by nonlinear algorithms that do not use true hard constraints, treatment planners put as constraints what they would like to achieve, even if such constraints cannot be satisfied simultaneously, such as having the maximal spinal cord dose below 45 Gy and the minimal dose to a very close

TABLE II. Timing (in hh:mm:ss) and memory usage (in Gbyte) of the four methods for the feasibility run and the eight optimization tasks. IP, PS, and DS are short for interior point, primal simplex, and dual simplex, respectively.

	Time (hh:mm:ss)				Memory (Gbyte)			
	ART3+O	IP	PS	DS	ART3+O	IP	PS	DS
NoTask	2	32:55	43:18	3:46	0.5	8.0	5.8	5.8
Task0	1:31	1:17:99	7:46	7:47	0.5	8.2	5.9	5.9
Task1	1:57	2:11:11	8:59:45	10:19:56	0.5	9.5	6.8	6.6
Task2	1:05	1:41:13	53:54	2:06:20	0.5	8.2	5.9	5.7
Task3	1:15	6:46:04	1:08:15	2:07:19	0.5	8.2	6.0	5.7
Task4	1:03	4:50:49	44:41	1:16:27	0.5	8.2	5.8	5.7
Task5	1:45	2:42:17	1:32:26	2:56:20	0.5	8.2	6.0	5.7
Task6	2:07	1:18:30	33:59	4:34:50	0.5	8.1	6.0	5.7
Task7	39	1:42:28	16:07:55	12:20:43	0.5	13.0	9.7	9.7

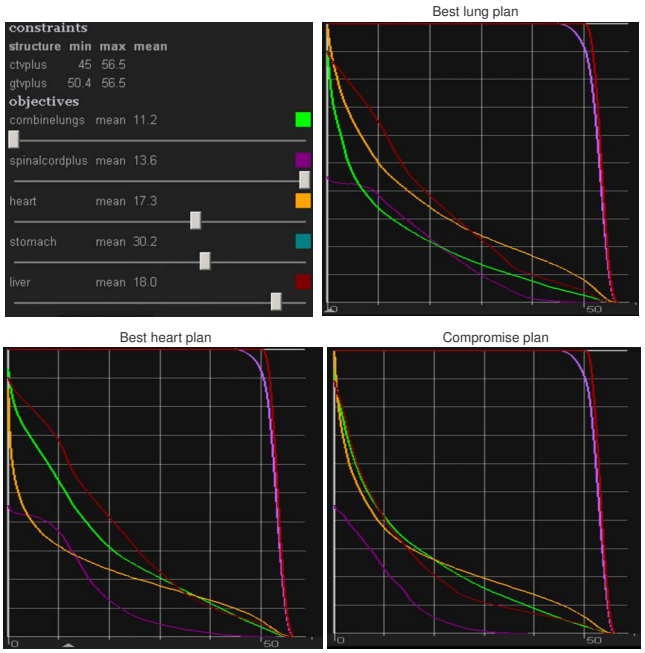

FIG. 2. IMPT MCO planning system showing the esophagus case. The top left panel displays the constraints and the objectives, and shows the navigation sliders in positions for the best lung plan, with the accompanying DVHs shown in top right panel. On the bottom left we show the DVHs for the best heart plan and on the bottom right we display the DVHs for a plan that was navigated to in order to attempt to spare all of the structures. (Patient data courtesy of Dr. T. Hong.)

target to be above 50.4 Gy. In our MCO setting with true hard constraints, we would formulate this as a hard constraint that the maximal spinal cord dose be no more than 45 Gy and an objective to maximize the minimum target dose. Nevertheless, our use of hard constraints is "softened" by the fact that we generate a large diversity of treatment plans and allow them to be mixed during navigation. Clinically, we only use hard constraints when the physician really will not tolerate any violation from that constraint. Everything else becomes an objective and this allows for an exploration of the entire clinically meaningful dose space. This has been discussed more fully in previous publications including Refs. 17 and 24.

The first case we present is an esophagus case. The esophagus is a difficult treatment site due to the proximity of several critical structures, namely, the lungs, spinal cord, heart, stomach, and liver. The target coverage is fixed by setting hard constraints of 45 Gy as the clinical target volume minimum dose and 50.4 Gy as the gross tumor volume minimum dose. All doses are kept below 112% of 50.4 Gy. The five objectives are minimizing the mean dose to the combined lungs, spinal cord, heart, stomach, and liver. The tradeoff between the heart dose and the combined lung dose has been found to be the most challenging compromise for many of the esophagus patients at Massachusetts General Hospital, so we display the range of possibilities for their DVHs in Fig. 2.

The second case is a tumor along the right rib cage and thus the right lung is a particular sensitive critical structure. Rather than specifying lower constraints for target coverage as in the esophagus case, here we are interested in assessing how target coverage affects the sparing of the lung, and so minimizing right lung mean dose and maximizing target minimum dose are put in as objectives. The other indicated objectives are minimizing the mean doses to the spinal cord and the heart. Figure 3 shows two solutions: One that forces the coverage of the target to be everywhere near its upper constraint of 70 Gy and one that minimizes the exposure of the right lung. Note the complete sparing of the left lung as proton beams have no exit dose.

IV. DISCUSSION AND CONCLUSION

The implementation of ART3+O is simple, fast, and memory efficient. The underlying optimizations produce feasible plans (i.e., the resulting dose distributions satisfy all the specified constraints exactly) for which the objective function is within a distance from that of the truly optimal feasible plan that is controlled by the optimality tolerance, for

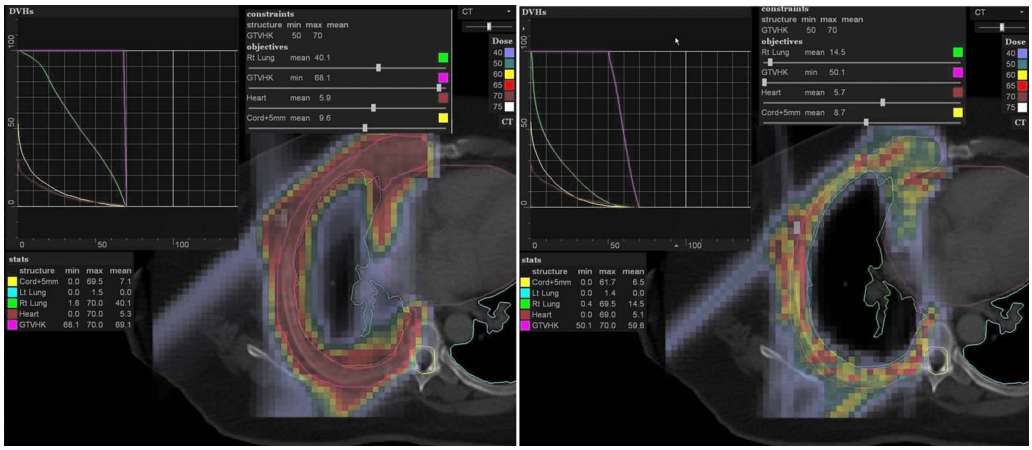

FIG. 3. Two plans shown from the lung/chest-wall target case. The plans use three fields: Anterior, posterior, and posterior oblique. The left plan forces complete target volume coverage to be near its maximum allowed value of 70 Gy, while the right plan achieves maximal lung sparing. (Patient data courtesy of Dr. B. Eden.)

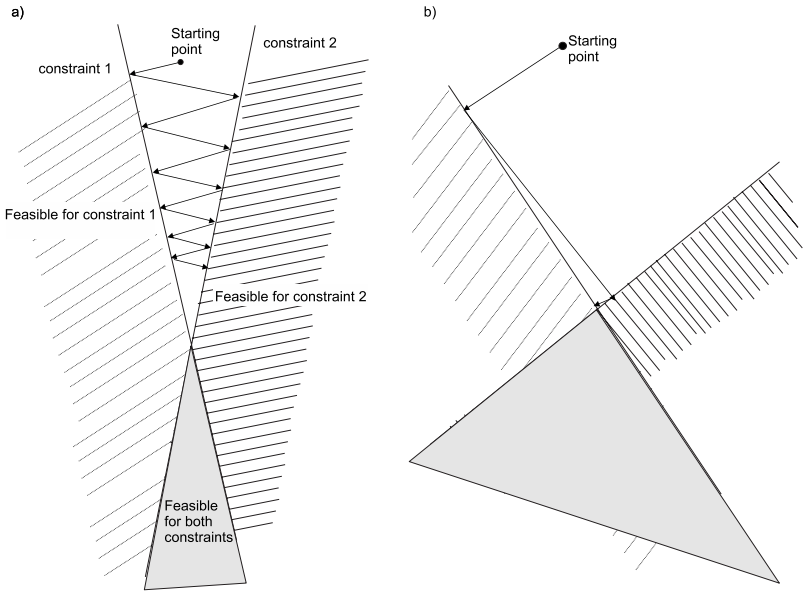

FIG. 4. Projections onto half spaces approach the feasible set faster when the half spaces are almost orthogonal.

which we use ϵ=0.1 Gy. However, extra experiments also show that the time ART3+O takes to find an ϵ-optimal solution increases dramatically when ϵ is set to be an order of magnitude smaller, e.g., 0.01 Gy, which is due to the nature of the underlying projection algorithm ART3+, which loses its efficiency for finding a feasible point when the feasible set is very small. Since all of our objectives and constraints are in units of Gy, this value of ϵ is directly interpretable and 0.1 Gy is much smaller than the precision of radiation delivery. Additionally, the point of database generation is to span the space of dose distribution possibilities by populating the approximated Pareto surface with diverse plans. With this in mind, it is clear that it is unnecessary to optimize each plan to a high numerical precision. Additional speed gains for database generation are achievable by performing the optimization tasks in parallel or implementing ART3+O on a graphics processing unit.[25]

ART3+ is a projection algorithm and such algorithms converge rapidly when the matrix used in the constraints, the D of Eq. (5), is sparse. Sparsity of our D matrix implies that most of the time, the normal vectors associated with a pair of constraints are mutually nearly orthogonal. Consecutive projections onto the two constraints with orthogonal normal vectors will produce a point that satisfies both. Therefore, by sequentially projecting onto the mostly mutually orthogonal constraints, one quickly finds a solution that satisfies all of the constraints. See Fig. 4 for an illustration of why projections onto half spaces approach the feasible set much more rapidly when the half spaces are almost orthogonal. In the IMPT setting, the D matrix has on the order of 10% of its entries nonzero and thus is fairly sparse. Since this is similar to IMRT, ART3+O is also useful for IMRT beamlet optimizations.

While in this work we restrict ourselves to controlling mean, minimum, and maximum structure doses, we do not view this as overly restrictive. Dose distributions are limited by the shape of the pencil beams and we have found that mean dose control is good for moving dose in and out of an organ, while minimum and maximum dose control is good for controlling cold-spots and hot-spots, respectively. Nevertheless, we will investigate using more general convex functions, in particular the EUD model, since such functions might be desirable to model certain clinical end points.[26] To this end, projection algorithms exist for optimizing quadratic functions or convex functions in general (see Refs. 10 and 27 or Ref. 9, Chapter 11) and it seems quite likely that a fast solution based on these algorithms is possible. Similarly, convex approximations of dose volume constraints such as cVAR (Ref. 28) and stochastic dominance[29] make it possible to include DVH control into this framework.

An important consideration in the optimization of proton therapy plans is robustness. Small uncertainties in the location of the Bragg peaks of the proton beams can lead to large voxel dose uncertainties. Optimizers that do not account for this uncertainty might produce plans that are highly sensitive to errors in the proton beam range. Similarly, changes in patient geometry that occur during treatment, especially in areas that are highly heterogeneous, such as nasal cavities, can lead to highly variable dose distributions. Accounting for range and patient geometry uncertainties can be done by using multiple instances of the D matrix.[30] Since the memory requirement of ART3+O is essentially only what is needed to store D, ART3+O should prove to be an ideal platform for robust IMPT optimization. Even though there are as yet no theoretical or empirical results on the complexity of ART3+O, we note that the multiple D matrices in the robust setting will be correlated, which amounts to having redundant (or close to redundant) constraints. The larger robust IMPT planning project is currently under development and we hope to be able to report on computational results on it in the near future.

ACKNOWLEDGMENTS

The authors thank Thomas Bortfeld and Yair Censor for their support of this work. Also, the authors are grateful for the insightful suggestions of the anonymous referees. This work was supported in part by NCI Grant No. 1 R01 CA103904-01A1, Multicriteria IMRT Optimization and Award No. R01HL070472 from the National Heart, Lung And Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung And Blood Institute or the National Institutes of Health.

[a]Much of Wei Chen's work on this paper was done during his visit to Massachusetts General Hospital and Harvard Medical School in 2009.
[b]Electronic mail: dcraft@partners.org

[1]L. Xing, J. Li, S. Donaldson, Q. Le, and A. Boyer, "Optimization of importance factors in inverse planning," Phys. Med. Biol. **44**, 2525–2536 (1999).
[2]C. Thieke, K.-H. Küfer, M. Monz, A. Scherrer, F. Alonso, S. Nill, C. Thilmann, and T. Bortfeld, "Beyond weight factors: New concepts for defining and analysing dose optimisation," Radiother. Oncol. **73**, S75–S75 (2004).
[3]D. Craft, T. Halabi, H. Shih, and T. Bortfeld, "Approximating convex Pareto surfaces in multi-objective radiotherapy planning," Med. Phys. **33**, 3399–3407 (2006).
[4]A. Messac, A. Ismail-Yahaya, and C. A. Mattson, "The normalized normal constraint method for generating the Pareto frontier," Struct. Multidiscip. Optim. **25**, 86–98 (2003).
[5]M. Monz, K.-H. Küfer, T. Bortfeld, and C. Thieke, "Pareto navigation—Algorithmic foundation of interactive multi-criteria IMRT planning," Phys. Med. Biol. **53**, 985–998 (2008).
[6]D. Craft, F. Carlsson, T. Bortfeld, and H. Rehbinder, "Multi-objective IMRT planning which produces deliverable plans," Med. Phys. **35**, 2846–2846 (2008).
[7]H. H. Bauschke and J. M. Borwein, "On projection algorithms for solving convex feasibility problems," SIAM Rev. **38**, 367–426 (1996).
[8]G. T. Herman, "A relaxation method for reconstructing objects from noisy x-rays," Math. Program. **8**, 1–19 (1975).
[9]G. T. Herman, *Fundamentals of Computerized Tomography: Image Reconstruction from Projections*, 2nd ed. (Springer, London, 2009).
[10]G. T. Herman and A. Lent, "A family of iterative quadratic optimization algorithms for pairs of inequalities, with application in diagnostic radiology," Mathematical Programming Studies **9**, 15–29 (1978).
[11]Y. Censor, M. D. Altschuler, and W. D. Powlis, "On the use of Cimmino's simultaneous projections method for computing a solution of the inverse problem in radiation therapy treatment planning," Inverse Probl. **4**, 607–623 (1988).
[12]W. Chen, G. T. Herman, and Y. Censor, in *Mathematical Methods in Biomedical Imaging and Intensity-Modulated Radiation Therapy (IMRT)*, edited by Y. Censor, M. Jiang, and A. K. Louis (Edizioni della Normale, Pisa, 2008), pp. 97–106.
[13]P. S. Cho, S. Lee, R. J. Marks, S. Oh, S. G. Sutlief, and M. H. Phillips, "Optimization of intensity modulated beams with volume constraints using two methods: Cost function minimization and projection onto convex sets," Med. Phys. **25**, 435–443 (1998).
[14]G. T. Herman and W. Chen, "A fast algorithm for solving a linear feasibility problem with application to intensity-modulated radiation therapy," Linear Algebr. Appl. **428**, 1207–1217 (2008).
[15]L. Xing, R. J. Hamilton, D. Spelbring, C. A. Pelizzari, G. T. Y. Chen, and A. L. Boyer, "Fast iterative algorithms for three-dimensional inverse treatment planning," Med. Phys. **25**, 1845–1849 (1998).
[16]W. Chen and G. T. Herman, "Efficient controls for finitely convergent sequential algorithms," ACM Trans. Math. Softw. 37, Article No. 14 (2010).
[17]D. Craft, T. Halabi, and T. Bortfeld, "Exploration of tradeoffs in intensity-modulated radiotherapy," Phys. Med. Biol. **50**, 5857–5868 (2005).
[18]H. E. Romeijn, J. Dempsey, and J. Li, "A unifying framework for multi-criteria fluence map optimization models," Phys. Med. Biol. **49**, 1991–2013 (2004).
[19]A. Niemierko, "Reporting and analysing dose distributions: A concept of equivalent uniform dose," Med. Phys. **24**, 103–110 (1997).
[20]E. D. Andersen and K. D. Andersen, in *High Performance Optimization*, edited by H. Frenk, K. Roos, T. Terlaky, and S. Zhang (Kluwer Academic, Dordrecht, The Netherlands, 2000), pp. 197–232.
[21]S. Breedveld, P. R. M. Storchi, M. Keijzer, A. W. Heemink, and B. J. M. Heijmen, "A novel approach to multi-criteria inverse planning for IMRT," Phys. Med. Biol. **52**, 6339–6353 (2007).
[22]K.-H. Küfer, A. Scherrer, M. Monz, F. Alonso, H. Trinkaus, T. Bortfeld, and C. Thieke, "Intensity modulated radiotherapy—A large scale multi-criteria programming problem," OR Spectrum **25**, 223–249 (2003).
[23]R. T. Marler and J. S. Arora, "Survey of multi-objective optimization methods for engineering," Struct. Multidiscip. Optim. **26**, 369–395 (2004).
[24]D. Craft, T. Halabi, H. Shih, and T. Bortfeld, "An approach for practical multiobjective IMRT treatment planning," Int. J. Radiat. Oncol., Biol., Phys. **69**, 1600–1607 (2007).
[25]J. M. Elble, N. V. Sahinidis, and P. Vouzis, "GPU computing with Kaczmarz's and other iterative algorithms for linear systems," Parallel Comput. **36**, 215–231 (2010).
[26]M. Söhn, D. Yan, J. Liang, E. Meldolesi, C. Vargas, and M. Alber, "Incidence of late rectal bleeding in high-dose conformal radiotherapy of prostate cancer using equivalent uniform dose-based and dose-volume-based normal tissue complication probability models," Int. J. Radiat. Oncol., Biol., Phys. **67**, 1066–1073 (2007).
[27]A. R. De Pierro and A. N. Iusem, "A finitely convergent row-action method for the convex feasibility problem," Appl. Math. Optim. **17**, 225–235 (1988).
[28]H. E. Romeijn, R. K. Ahuja, J. F. Dempsey, A. Kumar, and J. G. Li, "A novel linear programming approach to fluence map optimization for intensity modulated radiation therapy treatment planning," Phys. Med. Biol. **48**, 3521–3542 (2003).
[29]N. Noyan, G. Rudolf, and A. Ruszczynski, "Relaxations of linear programming problems with first order stochastic dominance constraints," Oper. Res. Lett. **34**, 653–659 (2006).
[30]J. Unkelbach, T. C. Y. Chan, and T. Bortfeld, "Accounting for range uncertainties in the optimization of intensity modulated proton therapy," Phys. Med. Biol. **52**, 2755–2773 (2007).