A conjugate-gradient based approach for approximate solutions of quadratic programs*

Fredrik CARLSSON† and Anders FORSGREN‡

Technical Report TRITA-MAT-2008-OS2
Department of Mathematics
Royal Institute of Technology
February 2008

Abstract

This paper deals with numerical behaviour and convergence properties of a recently presented column generation approach for optimization of so called step-and-shoot radiotherapy treatment plans. The approach and variants of it have been reported to be efficient in practice, finding near-optimal solutions by generating only a low number of columns.

The impact of different restrictions on the columns in a column generation method is studied, and numerical results are given for quadratic programs corresponding to three patient cases. In particular, it is noted that with a bound on the two-norm of the columns, the method is equivalent to the conjugate-gradient method. Further, the above-mentioned column generation approach for radiotherapy is obtained by employing a restriction based on the infinity-norm and non-negativity.

The column generation method has weak convergence properties if restricted to generating feasible step-and-shoot plans, with a "tailing-off" effect for the objective values. However, the numerical results demonstrate that, like the conjugate-gradient method, a rapid decrease of the objective value is obtained in the first few iterations. For the three patient cases, the restriction on the columns to generate feasible step-and-shoot plans has small effect on the numerical efficiency.

Key words. column generation, conjugate-gradient method, intensity-modulated radiation therapy, step-and-shoot delivery

1. Introduction

Optimization is an indispensable tool when planning cancer treatments with *intensity-modulated radiation therapy* (IMRT). The objective of IMRT is to determine the values of a set of treatment parameters associated with the delivery system such that the dose distribution generated in the patient meets the specified treatment goals. This is closely related to an inverse problem that can be formulated as a Fredholm equation of the first kind. The IMRT optimization problems are therefore ill-conditioned with a few dominating degrees of freedom [2, 7]. In practice, approximate solutions to the IMRT optimization problems suffice; there are numerous non-negligible

*Research supported by the Swedish Research Council (VR).

†Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology (KTH), SE-100 44 Stockholm, Sweden, (fcar@kth.se); and RaySearch Laboratories, Sveavägen 25, SE-111 34 Stockholm, Sweden.

‡Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology (KTH), SE-100 44 Stockholm, Sweden (andersf@kth.se).

uncertainties and sources of errors in the treatment planning process that make the search for the optimal solutions unmotivated. In particular, a near-optimal solution that corresponds to a "simple" treatment plan that can be delivered efficiently and with high accuracy is often preferred to the complex optimal solution. These characteristics motivate the use of conjugate-gradient like methods for IMRT optimization [5, 8].

One delivery technique for IMRT treatment plans is called *step-and-shoot* delivery. A column generation approach for optimization of plans to be delivered with this technique was introduced in [16]. Numerical results for this and related approaches demonstrate their capability of finding near-optimal solutions that utilize only a fraction of the disposable columns [4, 6, 16]. Such solutions are appealing since it turns out that they represent "simple" treatment plans.

In many approaches to IMRT, so called *fluence map optimization* problems are formulated and solved. The solutions to these problems are not feasible with respect to the treatment parameters so they must be transformed into feasible hardware settings *a posteriori*, with a potentially degradation of plan quality. In contrast, *direct aperture optimization* approaches to IMRT incorporate the delivery requirements in the problem formulation and thus generate solutions that can be delivered without any post-processing; see, e.g., [10, 18, 19]. The column generation approach belongs to this second group of approaches since the generated columns can be restricted to correspond to feasible hardware settings. An advantage of the column generation approach to the other IMRT optimization approaches is that the complexity of the treatment plan can be controlled through the number of columns. This makes the approach ideal for investigating the non-trivial trade-off between plan quality and treatment complexity.

To our knowledge, no efforts have been conducted to mathematically explain the promising numerical results obtained with the column generation approaches on IMRT problems. The goal of this paper is to gain an understanding of why the column generation method works so well in practice for step-and-shoot optimization problems. This is done by studying the numerical performance and the convergence properties of the method on quadratic programs (QP) and by comparing these characteristics to the conjugate-gradient method. Real-life IMRT optimization problems are often non-quadratic and may even be nonconvex. We limit this paper to quadratic programs since they capture the core of the structure of the IMRT optimization problems while being well-studied in the context of conjugate-gradient methods.

Throughout this paper, the objective function is given by $f(x) = \frac{1}{2}x^THx + c^Tx$, where H is an $n \times n$ positive definite symmetric matrix and c is an n-dimensional vector. We denote the gradient of f at a point x by $g(x)$, i.e., $g(x) = Hx + c$, the kth iterate by x_k and the gradient at x_k by g_k.

2. IMRT application

This section is devoted to introducing basic concepts in IMRT to readers not familiar with the field. Comprehensive introductions to IMRT can be found in, e.g., [1, 20].

The goal of radiation therapy is to generate a *dose distribution* with high dose concentrated to the tumor region(s) while sparing the healthy tissue as much as possible. The dose distribution in the patient is a result of high energy photon *fluences* generated by a *linear accelerator*. The photon fluences emanate from the *gantry head* which is positioned at different angles relative the patient. The fluence delivered from one angle is denoted by a *beam*. For many patient cases, it is beneficial to use IMRT, i.e., to spatially modulate the fluences of the beams, to generate high-quality dose distributions. The high dose regions can then be shaped to conform closely to

the tumor volume, even if nonconvex. A common tool for realizing this modulation is a *multileaf collimator* (MLC); see, e.g., [9] for a detailed description. This device is mounted on the gantry head of the accelerator and consists of several opposed tungsten leaves that block the incident radiation. The leaves can be positioned with high accuracy to form an aperture, also called a *segment*, that shapes the fluence. This paper considers step-and-shoot IMRT, where the fluence for a beam is given by a weighted sum over the fluences of a few segments. We denote the segment weights by w. These weights are proportional to the time the segments are exposed to the photon beam. We say that a step-and-shoot plan is "simple" if it is composed by few, large and regular segments. Such plans have non-jagged fluence profiles.

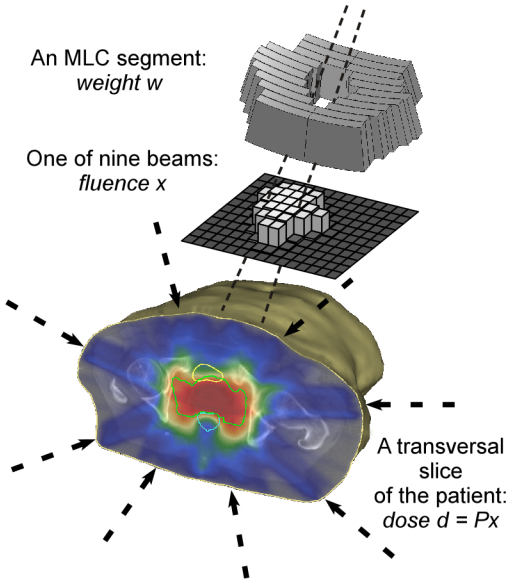

Figure 1: The delivery of step-and-shoot treatment plans. The figure illustrates an MLC segment, the transmitted fluence of the segment and the total dose distribution in one transversal slice of a prostate case. Note that the dose distribution is the result of nine beam fluences. The fluence for each beam is given by a weighted sum of the transmitted fluences over the segments of the beam.

For each beam, the fluence is discretized into a two-dimensional grid of beam elements (*bixels*) aligned with the MLC leaves. We denote the fluences of the n bixels of all beams by x. The bixel widths are set to the width of the MLC leaves to ensure that every leaf pair covers exactly one row of bixels. By restricting the leaf positions to the bixel boundaries, the transmission of the fluence through the MLC onto the bixels is essentially binary; for each segment, a bixel is either exposed to the incident fluence or covered by an MLC leaf. Note that this binary representation of the transmitted fluence is an approximation, where leakage and scatter effects induced by the MLC are neglected. The patient volume is discretized into m cubic volume elements (*voxels*) and the dose in voxel i is denoted by d_i. The dose distribution $d \in \mathbb{R}^m$ is given by $d = Px$, where P is the $m \times n$ *dose kernel matrix*. This matrix relates the fluence of each bixel to a dose distribution in the patient and it is precalculated with a *pencil beam algorithm* [12]. The dose kernel matrix has non-negative components and full column rank, and is typically ill-conditioned.

Figure 1 illustrates the delivery of step-and-shoot plans, with a schematic picture of an MLC segment, the transmitted fluence and the total dose distribution in a transversal slice of a prostate case. IMRT optimization problems have a multiobjective nature, with optimization functions describing the often conflicting treatment goals defined on patient specific *regions of interest* (ROIs). The ROIs specify groups of voxels of specific interest for the treatment such as the tumor or healthy organs. In Figure 1, the tumor ROI is outlined in the center of the patient slice. It is surrounded by the bladder ROI above and the rectum ROI underneath. Note how the IMRT treatment plan manages to conform a high dose region (red) to the tumor ROI while avoiding excessive dose to the bladder and rectum.

3. A column generation method

This section describes a column generation method that can be tailored to step-and-shoot optimization problems. First, unconstrained QP problems are considered. The conjugate-gradient method is briefly described and an equivalent column generation method is presented. Motivated by the IMRT application, non-negativity bounds are then added and the generated columns are restricted to be binary vectors. An illustrative example of the solution process is presented and the convergence properties of the column generation method are discussed. Finally, a version of the method that generates feasible step-and-shoot plans is presented.

3.1. The conjugate-gradient method

Consider the unconstrained QP problem

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

with the optimal solution $\widehat{x}$ satisfying $g(\widehat{x}) = 0$. A popular approach for solving large-scale instances of (3.1) is the conjugate-gradient method; see, e.g., [17, chapter 6] for a thorough discussion. The method proceeds in conjugate and linearly independent search directions. For any iteration k, the point x_{k+1} minimizes $f(x)$ over the Krylov subspace $\mathcal{K}_{k+1} = \text{span}\{g_0, Hg_0, \ldots, H^kg_0\}$, which equals the set $\text{span}\{g_0, g_1, \ldots, g_k\}$. This is accomplished by minimizing $f(x)$ in a direction that is a linear combination of g_k and the preceding search direction $p_{k-1} \in \mathcal{K}_k$. The method converges to the solution of (3.1) in at most n iterations and tends to minimize the objective function along the directions corresponding to the dominant eigenvalues of H first [8]. When H has a few large eigenvalues and many small eigenvalues, as often is the case for IMRT optimization problems [2], a near-optimal and regular (smooth) solution can be found in few iterations. This behaviour is desirable for IMRT optimization problems since such solutions are preferable from a practical viewpoint to the nonsmooth optimal solutions.

3.2. The conjugate-gradient method formulated as a column generation method

For any $n \times p$ matrix Q with full row rank, the problem

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n,\, w\in\mathbb{R}^p}{\text{minimize}} & \tfrac{1}{2}x^THx + c^Tx \\ \text{subject to} & x - Qw = 0, \end{array} \tag{3.2}$$

is equivalent to (3.1). The reason for considering a formulation on the form (3.2) is that it allows for including MLC requirements into the problem. With the columns of Q representing feasible

MLC segments, p may be in the order of 10^{17} for realistic IMRT problems [16]. It is therefore unreasonable to form Q explicitly. It is not in our interest to solve (3.2) exactly, but rather to generate a high-quality solution formed by a subset of the columns of Q.

The idea of column generation is to successively include columns of Q that have potential to improve the objective function. The column generation method proceeds by alternatively solving a *master problem* and a *subproblem*. The master problem is a restricted version of the original problem and the purpose of the subproblem is to detect the most promising column of Q not yet included. This is done by utilizing the current dual variables, here given by the gradient g. The method may be started with Q empty or with Q consisting of a few predetermined columns, and it terminates when the optimal value of the subproblem is non-negative. Then, none of the candidate columns can decrease the objective value (since f is convex), and the optimal solution to the master problem is optimal also to the original problem. Since the solution should include only a fraction of the feasible columns, the order of inclusion of the columns is crucial. For a review of column generation applied to linear programs, see [15].

The column generation method is described in Algorithm 3.1. For the problem (3.2), the master problem at step k is given by

$$(MASTER_k) \qquad \begin{array}{ll} \underset{x\in\mathbb{R}^n,\, w\in\mathbb{R}^k}{\text{minimize}} & \frac{1}{2}x^THx + c^Tx \\ \text{subject to} & x - Q_k w = 0, \end{array} \tag{3.3}$$

where Q_k is an $n \times k$ matrix and w is a k-dimensional vector. The optimality criteria (KKT conditions [13, 14]) to (3.3) are given by $Q_k^T g_k = 0$ and $x_k - Q_k w_k = 0$, with $g_k = Hx_k + c$. Note

Algorithm 3.1. *The column generation method.*

$g_0 \leftarrow c$;
Solve SUB_0 to get q_0;
$k \leftarrow 0$;
while $q_k^T g_k < 0$
  Let $Q_{k+1} = (Q_k\, q_k)$;
  $k \leftarrow k + 1$;
  Solve $MASTER_k$ to get x_k, w_k, g_k;
  Solve SUB_k to get q_k;
end while

that Algorithm 3.1 is designed for starting with an empty Q_0, i.e., $x_0 = 0$. If Q_0 is nonempty, the column generation method starts by solving the master problem instead. The subproblem at step k is given by

$$(SUB_k) \qquad \begin{array}{ll} \underset{q\in\mathbb{R}^n}{\text{minimize}} & q^Tg_k \\ \text{subject to} & q \in \mathcal{Q}, \end{array} \tag{3.4}$$

where $\mathcal{Q}$ defines the set of feasible columns. Due to the termination criteria of Algorithm 3.1, $\mathcal{Q}$ must be defined such that $g_k \neq 0 \;\Leftrightarrow\; q_k^T g_k < 0$ when the column generation method is applied to problem (3.1).

With $\mathcal{Q}_2 = \{q : ||q||_2 \leq 1\}$, the optimal solution q_k of (3.4) is given by $q_k = -g_k / ||g_k||_2$. Obviously, $g_k \neq 0 \;\Leftrightarrow\; q_k^T g_k < 0$. Analogous to the conjugate-gradient method, each iterate x_k

minimizes $f(x)$ over the subspace spanned by the current negative gradient and the previous negative gradients, implying that $x_k \in \mathcal{K}_k$. The solutions x_k of (3.3) are thus identical to the iterates of the conjugate-gradient method applied to (3.1), which results in that the weights w_k are non-negative for any k.

For any $\mathcal{Q}$ fulfilling $g_k \neq 0 \;\Leftrightarrow\; q_k^T g_k < 0$, Algorithm 3.1 converges to $\widehat{x}$ in at most n steps (although it might generate different iterates than the conjugate-gradient method). The reason is that any new column q_k satisfying $q_k^T g_k < 0$ is linearly independent of the columns of Q_k since $Q_k^T g_k = 0$. Then, if not finished in less than n steps, Q_n has full rank, which implies that the solution of (3.3) also solves (3.1). One such set is $\mathcal{Q}_\infty = \{q : ||q||_\infty \leq 1\}$, which results in integer solutions of the subproblem with $(q_k)_i = 1$ if $(g_k)_i < 0$ and $(q_k)_i = -1$ if $(g_k)_i > 0$. A numerical comparison of the column generation method applied to $\mathcal{Q}_\infty$ and $\mathcal{Q}_2$ is presented in Section 5. In this paper, the master problems are solved with an active-set quasi-Newton method, which is warm-started in every step with the solution of the previous master problem.

3.3. Inclusion of bound constraints on x

Motivated by the radiation therapy application, non-negativity bounds on x are added and we get the bound constrained QP problem

$$\begin{array}{ll} \underset{x \in I\!R^n}{\text{minimize}} & \frac{1}{2} x^T H x + c^T x \\ \text{subject to} & x \geq 0, \end{array} \tag{3.5}$$

which describes the fluence map optimization problem with a quadratic objective function. The unique solution x^* of (3.5) is defined by $x^* \geq 0$, $g(x^*) \geq 0$ and $(x^*)^T g(x^*) = 0$. The master problem is modified accordingly,

$$(MASTER_k) \qquad \begin{array}{ll} \underset{x \in I\!R^n,\, w \in I\!R^k}{\text{minimize}} & \frac{1}{2} x^T H x + c^T x \\ \text{subject to} & x - Q_k w = 0, \\ & x \geq 0. \end{array} \tag{3.6}$$

The optimality criteria of (3.6) are given by $Q_k^T(g_k - z_k) = 0$, $z_k \geq 0$ and $x_k^T z_k = 0$, together with the feasibility criteria $x_k - Q_k w_k = 0$ and $x_k \geq 0$. Solving (3.5) with the column generation method, $\mathcal{Q}$ must be chosen such that the solution q_k of the subproblem fulfills

$$g_k \geq 0 \;\Leftrightarrow\; q_k^T g_k \geq 0. \tag{3.7}$$

Recall from Section 2 that the transmitted fluence of an MLC segment with respect to the bixels is essentially binary. This restriction on x can be handled in the subproblem by the feasible set

$$\mathcal{Q}_{greedy} = \{q : ||q||_\infty \leq 1 \text{ and } q \geq 0\}, \tag{3.8}$$

which can be viewed as a restricted version of $\mathcal{Q}_\infty$. The binary solution of the subproblem with $\mathcal{Q}_{greedy}$ is given by $(q_k)_i = 1$ if $(g_k)_i < 0$ and $(q_k)_i = 0$ if $(g_k)_i \geq 0$. Hence, (3.7) is satisfied. The components of q_k with values equal to one correspond to exposed bixels, i.e., bixels with no MLC leaf covering the incident photon beam, while the zero components correspond to bixels covered by an MLC leaf. To generate feasible MLC segments, additional requirements on the columns must be added. This will be further discussed in Section 3.5.

The column generation method with $\mathcal{Q}_{greedy}$ does not proceed in Krylov subspaces since q_k is not parallel to g_k in general, but it still converges to x^* in n steps if the master problem is given by (3.6).

Proposition 3.1. *Algorithm* 3.1 *applied to* (3.6) *and* (3.4) *with* $\mathcal{Q}_{greedy}$ *solves* (3.5) *in at most* n *iterations.*

Proof. For any k, the solution of (3.6) fulfills $x_k \geq 0$ and $x_k^T g_k = x_k^T(g_k - z_k) = w_k^T Q_k^T(g_k - z_k) = 0$. The feasibility and complementarity conditions of (3.5) are thus always fulfilled. It remains to show that $g_k \geq 0$ is obtained in at most n steps.

Now assume that Q_k has full column rank. We want to show that, if $q_k^T g_k < 0$, then Q_{k+1} has full column rank. The solution x_k of (3.6) is the unique minimizer in $\text{span}\{q_0, \ldots, q_{k-1}\} \cap \{x : x \geq 0\}$ since Q_k has full column rank and since f is strictly convex. There exists a $q_k \in \mathcal{Q}_{greedy}$ such that $q_k^T g_k < 0$, which implies that $q_k \notin \text{span}\{q_0, \ldots, q_{k-1}\}$ since $q_k \geq 0$. Extending Q_k by q_k thus increases the rank by one. The rank of Q_1 is clearly one if $g_0 \ngeq 0$ and by induction, we get that Q_{k+1} has full column rank if $q_k^T g_k < 0$.

In at most n steps, q_k is linearly dependent of the columns of Q_k since Q_n is a square matrix. But $Q_{k+1} = (Q_k\, q_k)$ not having full column rank implies that $q_k^T g_k \geq 0$ and by (3.7), $g_k \geq 0$. ∎

3.4. Inclusion of bound constraints on w

To generate feasible step-and-shoot plans, w must be non-negative since the segment weights cannot be negative. With the master problem given by (3.6), this cannot be guaranteed. For instance, with $x = [0\ \ 1]^T$, $q_0 = [1\ \ 1]^T$ and $q_1 = [1\ \ 0]^T$, the weight vector becomes $w = [1\ \ -1]^T$. This motivates the following master problem with lower bounds on w,

$$(MASTER_k) \qquad \begin{array}{ll} \underset{x \in \mathbb{R}^n,\, w \in \mathbb{R}^k}{\text{minimize}} & \frac{1}{2} x^T H x + c^T x \\ \text{subject to} & x - Q_k w = 0, \\ & w \geq 0. \end{array} \tag{3.9}$$

By not allowing negative components in Q_k, $w \geq 0$ implies that $x \geq 0$ and feasibility of (3.5) is ensured. The optimality criteria of (3.9) are given by $Q_k^T g_k = y_k$, $y_k \geq 0$ and $w_k^T y_k = 0$, together with the feasibility criteria $x_k - Q_k w_k = 0$ and $w_k \geq 0$. A consequence of these optimality criteria is that as long as there is a negative component in the gradient g_k, the columns of Q_{k+1} are positively independent. To see this, let q be positively dependent on the columns of Q_k, i.e., let $q = Q\lambda$ for a $\lambda \geq 0$. Then $q^T g_k = \lambda^T Q_k^T g_k \geq 0$ and by (3.7), $g_k \geq 0$.

Let us define another feasible set of the subproblem,

$$\mathcal{Q}_{unit} = \{e_1, \ldots, e_n\}, \tag{3.10}$$

where e_i is the ith unit vector. With $\mathcal{Q}_{unit}$, the solution of the subproblem is given by $q_k = e_p$, where $p = \text{argmin}_i\{(g_k)_i\}$. It is clear that $\mathcal{Q}_{unit} \subset \mathcal{Q}_{greedy}$ and that (3.7) holds also for $\mathcal{Q}_{unit}$. In the context of column generation, $\mathcal{Q}_{unit}$ may be interpreted as $\{q : ||q||_1 \leq 1 \text{ and } q \geq 0\}$, i.e., as $\mathcal{Q}_{greedy}$ with the infinity-norm replaced by the one-norm.

Proposition 3.2. *Algorithm* 3.1 *applied to* (3.9) *and* (3.4) *with* $\mathcal{Q}_{unit}$ *solves* (3.5) *in at most* n *iterations.*

Proof. For any k, the solution of (3.9) fulfills $x_k \geq 0$ and $x_k^T g_k = w_k^T Q_k^T g_k = w_k^T y_k = 0$. The feasibility and complementarity conditions of (3.5) are thus always fulfilled. It remains to show that $g_k \geq 0$ is obtained in at most n steps.

In at most n steps, q_k is linearly dependent on the columns of Q_k since Q_k has n rows. With unit vectors as columns, linear dependence implies that not all columns are unique and hence q_k is positively dependent on the columns of Q_k. As discussed above, this implies that $g_k \geq 0$. ∎

We denote the approach of solving (3.5) by applying Algorithm 3.1 to (3.9) and (3.4) with $\mathcal{Q}_{greedy}$ and with $\mathcal{Q}_{unit}$ for the *greedy strategy* and the *unit strategy*, respectively.

An illustrative example of the solution process of the greedy strategy is the following *stairway* problem: Let $H = I$ and let $c = -[1\ \ 2\ \ \ldots\ \ 8]^T$, i.e., $f(x) = \frac{1}{2}\,||x + c||^2$. The following gradients and Q_4 matrix are generated by the greedy strategy,

$$G = \begin{bmatrix} -1 & 7/2 & 3/2 & 1/2 & 0 \\ -2 & 5/2 & 1/2 & -1/2 & 0 \\ -3 & 3/2 & -1/2 & 1/2 & 0 \\ -4 & 1/2 & -3/2 & -1/2 & 0 \\ -5 & -1/2 & 3/2 & 1/2 & 0 \\ -6 & -3/2 & 1/2 & -1/2 & 0 \\ -7 & -5/2 & -1/2 & 1/2 & 0 \\ -8 & -7/2 & -3/2 & -1/2 & 0 \end{bmatrix} \quad \text{and} \quad Q_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \tag{3.11}$$

where $G = [g_0\ \ \ldots\ \ g_4]$. The optimal solution x^* is, of course, given by $x^* = -c = [1\ \ 2\ \ \ldots\ \ 8]^T$, with the final weights given by $w = [1\ \ 4\ \ 2\ \ 1]^T$. Figure 2 illustrates the solution process. In

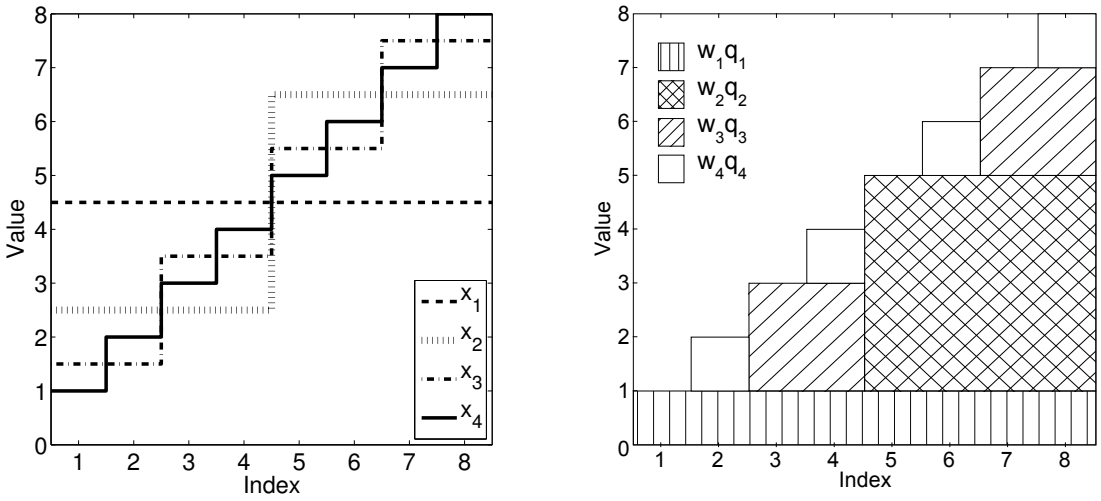

Figure 2: The greedy strategy applied to the stairway problem. Left: The four iterates with the optimal solution given by the solid line. Right: The optimal solution displayed as the $w_i q_i$ blocks.

each step, the method strives to fulfill $Q_k^T(x_k + c) = 0$. In other words, the sum over some of the components in x_k determined by the columns of Q_k should agree with the sum over the same components in $-c$. Although one might expect that the greedy strategy would require n steps to reach x^*, it takes advantage of the symmetry of the problem as can be seen in the right part of Figure 2. Instead of generating all eight components of x^* separately, as the unit strategy would,

the greedy strategy generates x^* with four "blocks". The reason that each block can include several components of x is the symmetry in c. If the components in c are assigned random numbers, one expects the greedy strategy to need all n steps. One would, however, expect the final steps to be devoted to "fine-tuning" the solution.

For general convex QP problems, the greedy strategy is not guaranteed to converge to x^* in n steps. The reason is that there might be a point $\widetilde{x}_k \in \text{span}\{q_0, \ldots, q_{k-1}\} \cap \{x : x \geq 0\}$ such that $f(\widetilde{x}_k) < f(x_k)$ for any $x_k \in \text{pos}\{q_0, \ldots, q_{k-1}\}$, where "pos" denotes the positive cone. A vector q that is positively independent of but linearly dependent on the columns of Q_k, while satisfying $q^T g_k < 0$, might therefore be generated in the subproblem. This makes Q_n rank-deficient, which could result in $g_n \not\geq 0$. This cannot happen with $\mathcal{Q}_{unit}$ since positive independence implies linear independence for unit vectors.

To construct a problem that requires more than n steps to converge to the optimal solution x^* of (3.5) with the greedy strategy, one has to make sure that at least one component of w is zero. As long as the weight vector is positive, the dual variable y_k must be zero to ensure $w_k^T y_k = 0$. Then, a new column q_k fulfilling $q_k^T g_k < 0$ is linearly independent of the columns of Q_k since $Q_k^T g_k = y_k = 0$, and the optimal solution is found in at most n steps. An example of a convex QP requiring $n+1$ steps to converge using the greedy strategy is the following three-dimensional problem (which was found numerically):

$$H = \begin{bmatrix} 0.8 & 6.6 & 5.2 \\ 6.6 & 60.0 & 54.7 \\ 5.2 & 54.7 & 60.0 \end{bmatrix} \quad \text{and} \quad c = \begin{bmatrix} -14 \\ -122 \\ -118 \end{bmatrix}. \tag{3.12}$$

The greedy strategy generates the following iterates,

$$Q_4 = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \quad W = \begin{bmatrix} 1.0 & 0.7 & 0.6 & 0 \\ - & 0.6 & 0 & 0 \\ - & - & 8.0 & 10.8 \\ - & - & - & 1.0 \end{bmatrix} \quad \text{and} \quad x^* = \begin{bmatrix} 10.8 \\ 0 \\ 1.0 \end{bmatrix}, \tag{3.13}$$

where the columns of W are the weight vectors $w_1, \ldots, w_4$. Note that the weights of the first two columns are zero at the optimum. The reason is that the second component of x^* is zero. The method has been "fooled" to proceed in directions including the second component of x in the first two iterations, resulting in that two more iterations are required to reach x^*. The greedy strategy might not converge to x^* in n steps even if H is the identity matrix, i.e., with $f(x) = \frac{1}{2}\,||x + c||^2$. For instance, with $c = -[1 \;\; 1.2 \;\; 2.7 \;\; 3.5 \;\; 4.9]^T$, the greedy strategy requires $n+1$ steps to converge. The reason is that the weight of the first column, $q_0 = [1 \;\; \ldots 1]^T$, is set to zero after four steps since all components of x are "covered" by the other columns of Q_4.

The number of steps required for the greedy strategy to converge is potentially exponential, with $2^n - 1$ steps as the upper limit. Whether there exist convex QP problems that actually require an exponential number of steps is unknown to us. For the IMRT application, however, it is the performance of the column generation method in the first few iterations that is of interest. For practical purposes, only a low number of columns (segments) are allowed in the solution. Typically, this number is of orders of magnitude less than n.

The initial performance of the unit strategy and the greedy strategy depends on the problem structure. With $x^* = e_p$ and $c < 0$, the unit strategy will converge directly if $p = \text{argmin}_i\{c_i\}$, but the greedy strategy will not. On the contrary, with a uniform and positive x^*, the greedy

strategy converges in one step while the unit strategy needs all n iterations. The numerical performance of these two strategies on IMRT problems is demonstrated in Section 5.

3.5. Generating feasible step-and-shoot plans

A solution $x_k = Q_k w_k$ to (3.9) represents a feasible step-and-shoot plan if the columns of Q_k correspond to feasible MLC segments. A segment is generated from a (subset) of a binary column of Q_k in two steps. First, the exposed bixels specified by the column are identified for every bixel row (leaf pair). Then, for each bixel row, the leaf positions are set to cover all non-exposed bixels. This implies that the exposed bixels of every bixel row must be contiguous to be realizable by a leaf pair. We call this requirement the *contiguous rows* criterion and denote the feasible set of the subproblem stipulated by this criterion by $\mathcal{Q}_{rows}$. With $\mathcal{Q}_{rows}$, the subproblem can be separated into smaller problems, where each problem only contains the bixels of one bixel row. These problems can be solved efficiently by the algorithm presented in [16]. The criteria set by $\mathcal{Q}_{rows}$ are sufficient for some of the MLCs used clinically. Other MLCs require additional criteria that are not separable in the bixel rows. Then, the subproblem can be formulated as a shortest-path problem for each beam and feasible segments are generated from the solutions of these network problems [3, 16].

The unit strategy and the greedy strategy discussed in the previous section "almost" produce feasible step-and-shoot plans. The unit vector columns produced by the unit strategy results in feasible segments (in theory), but, in practice, this approach is not viable since too many segments must be included to generate a solution with many nonzero components. Further, the transmitted fluence of small segments is hard to model, which results in large uncertainties in the delivered dose. In contrast, the greedy strategy allows for columns with more than one exposed bixel. These columns do not, however, correspond to feasible MLC segments in general since it is not guaranteed that they satisfy the contiguous rows criterion.

Let the method of solving (3.5) by applying Algorithm 3.1 to (3.9) and (3.4) with $\mathcal{Q}_{rows}$ be denoted by the *rows strategy*. The rows strategy can be seen as a limited version of the greedy strategy and we do not expect the rows strategy to converge to x^* in n steps in general. Note that the rows strategy, which is similar to the column generation approach presented in [16], is the only approach discussed in this paper that generates feasible step-and-shoot plans.

4. Test problems

Three patient cases are studied; a pancreas case (with 5 beams, 335 bixels and 68040 voxels), a prostate case (3, 214, 113738) and a head-and-neck case (9, 365, 26775). These cases originate from clinical cases, but have coarser discretizations of the fluences and the patient volumes. For each case, the patient volume is partitioned into three ROIs: a *target region* $\mathcal{T}$, an *organs-at-risk region* $\mathcal{O}$ and a *normal tissue region* $\mathcal{N}$.

A simple yet reasonable objective function for IMRT is given by the quadratic function

$$f(x) = \tfrac{1}{2} \left|\left| D^{(1/2)}(Px - d^{pres}) \right|\right|_2^2, \tag{4.1}$$

where $d^{pres} \in \mathbb{R}^m$ is the prescribed dose distribution and D is an $m \times m$ diagonal weight matrix with positive diagonal entries. The Hessian of (4.1), $H = P^T D P$, is positive definite and typically ill-conditioned with many eigenvalues clustered near zero.

Let $d_i^{pres} = 1$ for $i \in \mathcal{T}$ and $d_i^{pres} = 0$ for $i \in \mathcal{O} \cup \mathcal{N}$, i.e., prescribe nonzero uniform dose to the target region and zero dose to the healthy tissue. It is impossible to replicate this d^{pres} exactly since depositing dose to the target will, inevitably, lead to nonzero dose in the surrounding healthy tissue. To prioritize the target and the organs-at-risk, we let $D_{ii} = 10$ for $i \in \mathcal{T}$, $D_{ii} = 5$ for $i \in \mathcal{O}$ and $D_{ii} = 1$ for $i \in \mathcal{N}$. We argue that this objective function grasps the mathematical properties of IMRT problems despite being idealized and not focused on clinical relevance. For a numerical evaluation of the column generation method on test problems with more emphasis on clinical relevance, we refer to [4].

5. Numerical results

Figure 3 displays the objective values versus iteration number for different subproblem restrictions on each of the three patient cases with the objective function defined by (4.1). The objective value is scaled with the initial objective value for each case.

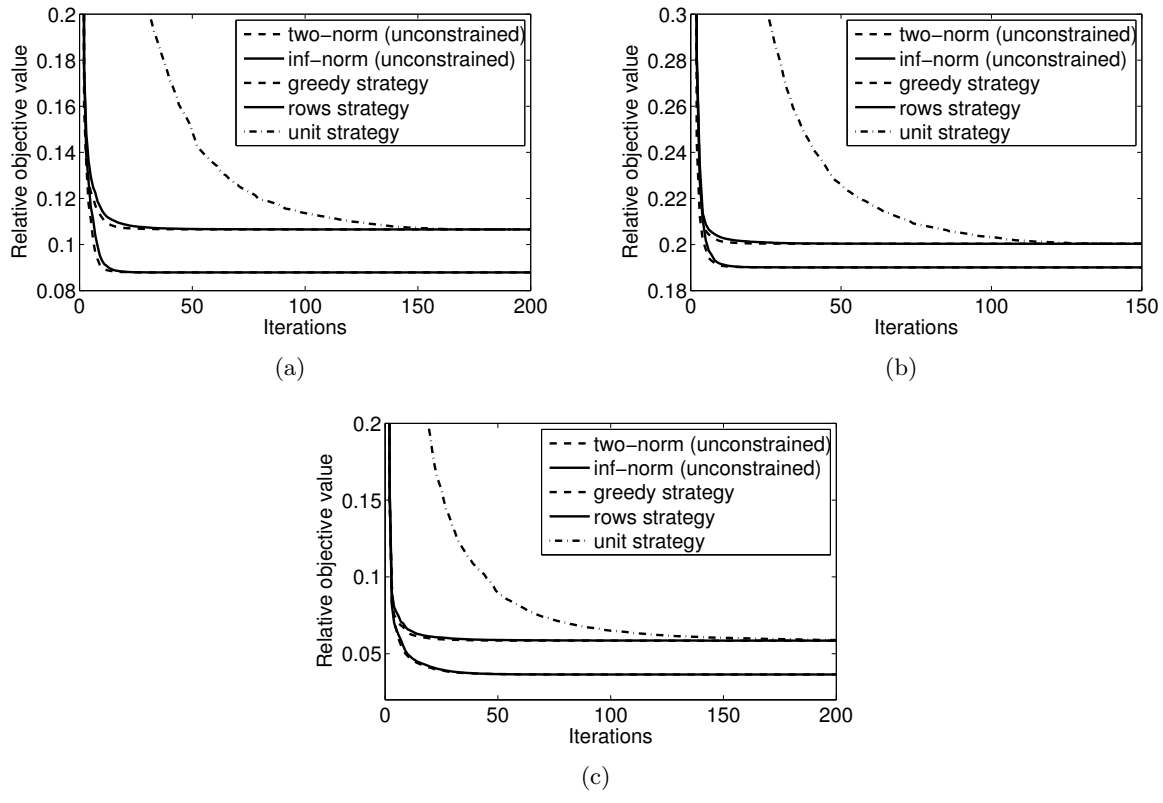

Figure 3: The relative objective values versus iteration number for the column generation method with five different subproblem strategies applied to IMRT optimization problems for a pancreas case (a), a prostate case (b) and a head-and-neck case (c). The curves with the lowest objective values correspond to the first two strategies, which are applied to the unconstrained problem (3.1). The other three strategies are applied to the bound constrained problem (3.5).

The first two strategies, denoted by *two-norm* and by *inf-norm* in the figure, correspond to the column generation method applied to the unconstrained QP problem (3.1) with the feasible

set of the subproblems defined by $\mathcal{Q}_2$ and $\mathcal{Q}_\infty$, respectively. The objective values are shown as the curves with the lowest objective values, with dashed lines for the two-norm strategy and with solid lines for the inf-norm strategy. As discussed in Section 3.2, the two-norm strategy is equivalent to the conjugate-gradient method. Note that this formulation allows for negative fluence, which is non-physical.

The objective values of the other three strategies, introduced in Sections 3.4 and 3.5, are shown with dashed lines for the greedy strategy, solid lines for the rows strategy and dash-dotted lines for the unit strategy. These strategies are applied to the bound constrained QP problem (3.5). As can be seen in Figure 3, the optimal values are higher with bounds on x. The reason is that (3.1) is a relaxation of (3.5) and that the optimal solution $\widehat{x}$ of (3.1) is not feasible with respect to (3.5) for any of the patient cases. With lower bounds on x, some bixels get zero fluence at the optimal solutions. The unit strategy is therefore able to converge to x^* in less than n steps. This is also the case for the greedy strategy, while the rows strategy fails to reach x^* in n steps for all the three cases.

Comparing the objective values in the first few iterations, it is apparent that the unit strategy initially is inferior to the other strategies. In contrast, the objective values of the solutions obtained with the rows strategy are very close to the optimal values after rather few iterations. There is clearly not much to gain after a certain number of iterations, the improvement in objective value is negligible and the number of columns (segments) will increase if continuing. The impact of restricting the feasible set of the subproblem from $\mathcal{Q}_{greedy}$ to $\mathcal{Q}_{rows}$ to generate feasible segments is very small in terms of the objective values. For these two feasible sets, similar results were obtained when a steepest-edge like strategy was applied in the subproblems, i.e., when only significantly negative elements of the gradient vector were considered in the subproblem.

The shape of the objective value curves are similar for the rows strategy and the conjugate-gradient method (two-norm strategy). The appealing properties of the conjugate-gradient method on ill-conditioned problems seem to be inherited in the column generation approach with $\mathcal{Q}_{rows}$. Near-optimal solutions are found in very few iterations compared to the problem size. This result falls well in line with the more clinical results reported in [4,6], where few iterations seem to suffice in order to generate high-quality step-and-shoot plans. As a side note, column generation for linear programs also exhibits slow convergence combined with the ability of reaching near-optimal solutions fast. This characteristic is known in the literature as the *tailing-off effect* [15].

6. Summary and discussion

We have discussed numerical behaviour and convergence properties for a column generation method applied to convex quadratic programs arising in IMRT. In particular, the impact of different subproblem restrictions was studied. With a two-norm based restriction, a method equivalent to the conjugate-gradient method was obtained. Infinity-norm based restrictions were introduced motivated by the IMRT application. In terms of initial decrease of the objective value, the column generation method performs similarly to the conjugate-gradient method if the restriction of the subproblem is based on the infinity-norm. Further restrictions on the subproblem to generate feasible step-and-shoot IMRT treatment plans, called the rows strategy approach, has small effect on the efficiency of the method.

One may view the rows strategy approach as a fluence map optimization of "groups of bixels" defined by the generated columns, with the property that each "group" can be realized by an

MLC segment. We believe that the efficiency of the rows strategy approach is related to that non-jagged fluence profiles can generate near-optimal solutions to fluence map optimization problems.

Additional heuristics to produce clinically more relevant and thus acceptable plans may also be used. Such heuristics include letting the lower bound of w be positive, removing columns with associated weights at the lower bound and using a perturbed g_k in the subproblem to generate clinically preferable segment shapes. Further, the optimization of the master problem may be prematurely terminated [11]. Accounting for "second-order" MLC transmission effects such as leakage and scatter is essential. A sound strategy is to employ an accurate transmission model when generating a new segment from a binary column. Then, the columns of Q are non-binary and the dose distribution can be computed from an accurate fluence in the master problem.

Despite incorporating all these heuristic extensions and using non-quadratic objective functions and MLC constraints that do not separate in bixel rows, the numerical results of [4] demonstrate that high-quality solutions are still obtained in a low number of iterations. Further, in [6], it is reported that the impact of different MLC constraints on the numerical performance of the column generation method is small. The column generation method thus performs remarkably well on step-and-shoot optimization problems even if "perturbed" by heuristics and the numerical performance is robust with respect to different restrictions on the subproblem. The likeness to the conjugate-gradient method is, in our opinion, an important explanation for the ability of the column generation method to take advantage of the fundamental nature of IMRT optimization problems to find near-optimal step-and-shoot plans with few segments.

7. Acknowledgements

We thank Göran Sporre and Henrik Rehbinder for their careful reading of the manuscript and for a number of helpful comments.

References

[1] A. Ahnesjö, B. Hårdemark, U. Isacsson, and A. Montelius. The IMRT information process—mastering the degrees of freedom in external beam therapy. *Physics in Medicine and Biology*, 51(13):R381–R402, 2006.

[2] M. Alber, G. Meedt, F. Nüsslin, and R. Reemtsen. On the degeneracy of the IMRT optimization problem. *Medical Physics*, 29(11):2584–2589, 2002.

[3] N. Boland, H. W. Hamacher, and F. Lenzen. Minimizing beam-on time in cancer radiation treatment using multileaf collimators. *Networks*, 43(4):226–240, 2004.

[4] F. Carlsson. Combining segment generation with step-and-shoot optimization in intensity-modulated radiation therapy. Technical report, Report TRITA-MAT-2007-OS3, Department of Mathematics, Royal Institute of Technology, 2007.

[5] F. Carlsson and A. Forsgren. Iterative regularization in intensity-modulated radiation therapy optimization. *Medical Physics*, 33(1):225–234, 2006.

[6] Z. C. T. Chunhua Men, H Edwin Romeijn and J. F. Dempsey. An exact approach to direct aperture optimization in IMRT treatment planning. *Physics in Medicine and Biology*, 52(24):7333–7352, 2007.

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

[8] H. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*. Kluwer Academic Publishers, Dordrecht, 2000.

[9] J. Galvin. The multileaf collimator: a complete guide. In *Proc. AAPM Annual Meeting*, 1999.

[10] W. D. Gersem, F. Claus, C. D. Wagter, B. V. Duyse, and W. D. Neve. Leaf position optimization for step-and-shoot IMRT. *International Journal of Radiation Oncology, Biology, Physics*, 51(5):1371–1388, 2001.

[11] J. Gondzio and R. Sarkissian. Column generation with a primal-dual method. Technical report, Logilab Technical Report 96.6, Department of Management Studies, University of Geneva, Switzerland, 1996.

[12] A. Gustafsson, B. K. Lind, and A. Brahme. A generalized pencil beam algorithm for optimization of radiation therapy. *Medical Physics*, 21(3):343–356, 1994.

[13] W. Karush. Minima of functions of several variables with inequalities as side constraints. Master's thesis, Department of Mathematics, University of Chicago, 1939.

[14] H. W. Kuhn and A. W. Tucker. Nonlinear programming. In J. Neyman, editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492, Berkeley, 1951. University of California Press.

[15] M. E. Lübbecke and J. Desrosiers. Selected topics in column generation. *Operations Research*, 53(6):1007–1023, 2005.

[16] H. E. Romeijn, R. K. Ahuja, J. F. Dempsey, and A. Kumar. A column generation approach to radiation therapy treatment planning using aperture modulation. *SIAM Journal on Optimization*, 15(3):838–862, 2005.

[17] Y. Saad. *Iterative methods for sparse linear systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2003.

[18] D. M. Shepard, M. A. Earl, X. A. Li, S. Naqvi, and C. Yu. Direct aperture optimization: A turnkey solution for step-and-shoot IMRT. *Medical Physics*, 29(6):1007–1018, 2002.

[19] J. Tervo, P. Kolmonen, T. Lyyra-Laitinen, J. Pintér, and T. Lahtinen. An optimization-based approach to the multiple static delivery technique in radiation therapy. *Annals of Operations Research*, 119:205–227, 2003.

[20] S. Webb. The physical basis of IMRT and inverse planning. *The British Journal of Radiology*, 76(910):678–689, 2003.