http://pubsonline.informs.org/journal/ijoc/

INFORMS JOURNAL ON COMPUTING
Vol. 31, No. 3, Summer 2019, pp. 544–558
ISSN 1091-9856 (print), ISSN 1526-5528 (online)

Real-Time Radiation Treatment Planning with Optimality Guarantees via Cluster and Bound Methods

Baris Ungun,[a] Lei Xing,[b] Stephen Boyd[c]

[a] Department of Bioengineering, Stanford University, Stanford, California 94305; [b] Department of Radiation Oncology, Stanford University School of Medicine, Stanford, California 94305; [c] Department of Electrical Engineering, Stanford University, Stanford, California 94305

Contact: ungun@stanford.edu, https://orcid.org/0000-0003-4131-9352 (BU); lei@stanford.edu, https://orcid.org/0000-0003-2536-5359 (LX); boyd@stanford.edu (SB)

Received: October 20, 2016
Revised: August 11, 2017; April 20, 2018; May 19, 2018
Accepted: May 24, 2018
Published Online in Articles in Advance: May 8, 2019

https://doi.org/10.1287/ijoc.2018.0841

Abstract. Radiation therapy is widely used in cancer treatment; however, plans necessarily involve tradeoffs between tumor coverage and mitigating damage to healthy tissue. Although current hardware can deliver custom-shaped beams from any angle around the patient, choosing (from all possible beams) an optimal set of beams that maximizes tumor coverage while minimizing collateral damage and treatment time is intractable. Furthermore, even though planning algorithms used in practice consider highly restricted sets of candidate beams, the time per run combined with the number of runs required to explore clinical tradeoffs results in planning times of hours to days. We propose a suite of cluster and bound methods that we hypothesize will (1) yield higher-quality plans by optimizing over much (i.e., 100-fold) larger sets of candidate beams, and/or (2) reduce planning time by allowing clinicians to search through candidate plans in real time. Our methods hinge on phrasing the treatment-planning problem as a convex problem. To handle large-scale optimizations, we form and solve compressed approximations to the full problem by clustering beams (i.e., columns of the dose deposition matrix used in the optimization) or voxels (rows of the matrix). Duality theory allows us to bound the error incurred when applying an approximate problem's solution to the full problem. We observe that beam clustering and voxel clustering both yield excellent solutions while enabling a 10- to 200-fold speedup.

History: Accepted by Allen Holder, Applications in Biology, Medicine, & Healthcare.
Funding: Financial support from the National Institutes of Health [Grant 5R01CA176553] and Stanford Bio-X Bowes Graduate Fellowship is gratefully acknowledged.

Keywords: convex optimization • radiation therapy • treatment planning • clustering • optimality bounds • matrix factorization • dimensionality reduction

1. Introduction

1.1. Problem Description

In external beam radiation therapy, clinicians seek treatment plans that balance the competing objectives of maximizing tumor coverage, minimizing radiation to nontarget tissue, and achieving short treatment times. Treatments may be planned for delivery in several sessions, or fractions, and the planning process may try to account for uncertainties in the radiation delivery process.

In the broadest terms, for each treatment session, the planning process takes as inputs an estimate of the patient's anatomy and a model of the dose delivery physics and outputs a treatment design, or plan, consisting of a trajectory in the parameter space of the beam delivery hardware and an estimate of the resulting dose imparted to the patient, with the goal being to make this design optimal with respect to the aforementioned objectives (Craft et al. 2012b).

Even for the optimistic scenarios in which the uncertainties in patient anatomy and radiation delivery are not considered, two major categories of problems remain. The first is the large search space of treatment plans, stemming from the flexibility of modern radiation hardware in delivering shaped beams from nearly any angle around the patient. The size of the search space is exacerbated by the fact that for many common treatment modalities, the dose delivered to a point inside the patient is a nonconvex function of the machine parameters (Craft 2007). Consequently, obtaining an exact, globally optimal solution to the planning problem over the full reachable space of the hardware is either computationally intractable or impractical within the time constraints of the clinic.

Nearly all formulations of the planning problem use a discrete representation of the space of hardware parameters. As part of this discretization, the machine parameters are often not used directly as the optimization variables and are instead replaced with an abstraction: discrete radiation sources ("beams") parameterized by position, shape, and intensity (i.e., *fluence*, a product of the dose rate and delivery duration). Approaches to render the problem computationally tractable involve

restricting the search space into some manageable set of candidate beam and addressing the problem in three stages: (1) determining the geometric setup of candidate beams, (2) optimizing intensity profiles for each candidate beam with respect to some clinical objectives, and (3) generating a sequence of hardware parameters that (approximately) delivers the optimized intensities from the specified locations. Because candidate beams are an abstraction, the exact sense depends on the radiation type (e.g., electron, photon, or proton) and delivery modality under consideration [e.g., intensity-modulated radiation therapy (IMRT), volumetric-modulated arc therapy (VMAT), or tomotherapy]. Concrete examples include an IMRT field, which can be further subdivided into beamlets for fluence map optimization (FMO); a single beamlet from such a fluence map; a VMAT aperture, the geometric setup of which may have been obtained by refining the solution to a FMO problem, as per the method described in Craft et al. (2012b); and a pencil beam in intensity-modulated proton therapy (IMPT).

The three stages of planning can be performed sequentially or jointly, depending on the choice of mathematical formulation. For instance, an intensity optimization problem nominally targeted at the second stage can include constraints that enforce (or regularization terms that promote) machine deliverability, thereby easing the sequencing step. As another example, an objective that promotes sparsity in beam intensities can be used to jointly optimize intensities and select sources from a given set of candidate beams.

The large body of work on planning algorithms spans both convex formulations, such as fluence map optimization problems used in IMRT planning (Romeijn et al. 2003, Aleman et al. 2010), which are paired with a set of small mixed integer programs to decompose fluence maps into deliverable apertures (Baatar et al. 2009, Ernst et al. 2009), and nonconvex formulations, such as direct machine parameter optimization used for VMAT planning (Peng et al. 2012) or robust optimization of beam angles and intensities (Bertsimas et al. 2010). However, all of these methods involve somewhat arbitrary choices of parameters (such as plan isocenters, beam positions, or arc angles) that have a major impact on plan quality and generally constitute significant restrictions of the search space of candidate beams (Dong et al. 2013a, Li and Xing 2013, Li et al. 2014, Zarepisheh et al. 2014a).

In addition to the problem of trying to optimally utilize the delivery hardware, a second source of major clinical and computational challenges is the inherent multiobjective nature of the treatment planning problem. Clinicians must balance several clinical objectives—typically at least one per anatomical structure in the plan, so at least 10–20 objectives for most planning cases. Finding an acceptable plan often involves significant time spent generating and comparing plans optimal for different objective tradeoffs, which can be interpreted as populating and navigating a Pareto surface for a multiobjective optimization problem (Craft and Bortfeld 2008, Küfer et al. 2009, Chan et al. 2014).

In this paper, we address large-scale intensity optimization problems in which we assume the geometric configuration of the candidate beams to be given. The methods we present do not depend on the radiation physics or treatment modality; they apply directly to beamlet and aperture intensity optimization problems and can therefore be used as is in IMRT or IMPT planning or to accelerate more complex planning algorithms that involve an intensity optimization phase.

In tandem with a voxel-separable convex formulation that allows the use of state-of-the-art distributed optimization methods, we propose cluster and bound methods that allow an intensity optimization problem of a given size to be approximated by one 20–100 times smaller. These methods allow for a dramatic reduction in the per-solve computational cost, which serves two primary goals. The first is to enable plans to be optimized over much larger sets of candidate beams in reasonable time. The second is to allow (for modestly sized problems) plans to be generated in hundredths to tenths of a second, which would enable clinicians to navigate clinical tradeoffs in real time, or for a library with several hundred or a few thousand Pareto-optimal plans (or nearly optimal plans) to be populated in a matter of minutes.

We find that the clustered problems generate solutions that are close to those of their corresponding full problems; however, because comparisons against true optima cannot be performed in practice, we use lower bounds obtained from the dual of the treatment planning problem to bound the maximum suboptimality of plans generated through clustered approximations.

1.2. Outline

This paper is structured as follows. In Section 1.3, we examine previous work related to solving large-scale intensity optimization problems in treatment planning. In Section 2, we introduce the class of convex treatment planning problems compatible with the methods detailed in this work, as well as their associated dual problems. In Section 3, we describe two approximation methods, voxel clustering and voxel collapse, that form relaxations of the planning problem at dramatically decreased computational cost. We further present optimality bounds for plans generated by these approximation methods. In Section 4, we describe an approximation method, column clustering, that allows a restriction of the planning problem to be solved at significantly lower computational cost and a paired method for generating optimality bounds on plans generated in this fashion. In Section 5, we present examples using these methods, including a fluence map optimization of a prostate

IMRT case and an aperture reweighting of a head and neck VMAT case.

1.3. Related Work

1.3.1. Convex Programming in Treatment Planning. Nonconvexity in treatment planning can arise from the use of certain clinical objectives or phrasing the problem in terms of machine parameters (specifically, leaflet positions of a multileaf collimator as detailed by Küfer et al. 2009) instead of optimizing over the intensities of predetermined apertures or beamlets, or using integer decision variables to select candidate beams. To work around these challenges, a column-generation approach that alternates between solving an aperture intensity optimization problem and another convex pricing problem to incorporate new candidate apertures was developed (Peng et al. 2012, Zarepisheh et al. 2014a).

The efforts in beam angle optimization (Zhang et al. 1999, Craft 2007, Lim et al. 2007, Aleman et al. 2008, Ahmed et al. 2010), nonuniform arc therapy (Li and Xing 2013, Zarepisheh et al. 2014a), noncoplanar planning (Dong et al. 2013a, b), and optimized isocenter selection (Li et al. 2014) all highlight the limitations to plan quality incurred by conventional methods that only consider restrictions of the planning space to a small number of intensity-modulated fields or to apertures distributed uniformly along coplanar arcs with manually chosen orientations and isocenters. In other words, conventional IMRT and uniformly sampled coplanar VMAT planning methods suffer from undersampling the treatment hardware's search space. The results from these studies reinforce the potential value of methods that can optimize intensities of many more fields or apertures.

While some clinical objectives, such as the dose volume constraints widely used as metrics in plan evaluation, are neither separable nor convex in the optimization variables, Romeijn et al. (2003) and Kessler et al. (2005) propose that good convex approximations exist for all clinically interesting objectives; furthermore, many of these take on the simple form of fully separable piecewise linear functions. We follow a strongly related approach in which we restrict our formulation to consider fully separable convex functions.

In Parikh and Boyd (2014), the authors describe how intensity optimization problems in treatment planning can be phrased as graph form problems so as to benefit from highly parallel optimization algorithms; the freely available open-source Proximal Operator Graph Solver (POGS) that implements graph form Alternating Direction Method of Multipliers (ADMM), which we use in this work, is described in Fougner and Boyd (2015).

1.3.2. Planning Tradeoff Navigation. Besides the per-solve cost, much of the computational burden in planning comes from the need to iterate through many plans that correspond to different tradeoffs between the multiple clinical objectives. Formulating a planning problem with convex objectives and constraints ensures that the set of achievable plans will be convex, which simplifies the task of finding Pareto-optimal plans that lie on the boundary of this set; by contrast, when the set of achievable plans is nonconvex, there may be Pareto-optimal plans that are not attainable by scalarization methods commonly used in multiobjective optimization (Boyd and Vandenberghe 2004). Nevertheless, even when it is straightforward to find an optimal plan for a given clinical preference, the task of solving multiple optimization problems remains, because tradeoffs between optimal plans can only be resolved by the planner's clinical judgment.

Major research efforts in this area include multicriterion optimization (MCO) approaches that generate libraries of optimal points along the Pareto surface and methods to approximate this surface (Craft et al. 2006; Chen et al. 2010; Siem et al. 2011; Craft 2012a, b; Bokrantz and Forsgren 2013; Rennen et al. 2013), approximation of the Pareto surface, automated planning and Pareto surface navigation (Li et al. 2013, Zarepisheh et al. 2014b), and statistical learning (from previously planned cases) of clinical preferences and anatomically driven predictions of feasible designs, or even favorable beam directions (Pugachev and Xing 2002, Appenzoller et al. 2012, Lee et al. 2013, Moore et al. 2014, Boutilier et al. 2016).

Another approach discussed by Otto (2014) is to solve for or estimate feasible dose distributions at rates upwards of 20 times a second, allowing for real-time navigation of the clinical tradeoffs.

1.3.3. Voxel Clustering. Reducing the dose grid resolution is commonly done to lower the cost of dose calculations and plan optimization, but studies such as Scherrer et al. (2005) and Martin et al. (2007) have demonstrated that random voxel sampling (which, in expectation, approaches the voxel clustering problem) or adaptive hierarchical clustering methods can lead to dramatic reductions in computational cost with clinically acceptable approximation error.

Such approximation methods provide one way to address the challenges discussed in Sections 1.3.1 and 1.3.2: reducing per-plan solve time allows clinicians to explore a greater number of clinical tradeoffs, or sample a greater portion of the delivery hardware's reachable space, while keeping total planning time fixed.

1.3.4. Beam Clustering. We propose to efficiently optimize over large numbers of beams by clustering them based on their numerical similarity. Related techniques (that avoid the overhead associated with the clustering calculation) include optimizing over subsets of a pool of available beams, as in the column generation approach formulated by Peng et al. (2012). Lu et al. (2008) take

a similar approach of doing some intensive computation up-front in order to estimate the principal components spanning the space of clinically relevant tradeoffs. In Banger and Oelfke (2010), the authors score effects of beams on each voxel and cluster them by score vectors to optimize beam angle choices, while optimal intensities of beams and interbeam similarities are used to cluster and select beam angles in Lim et al. (2009).

Both the beam and voxel clustering approaches are special cases of nonnegative matrix factorization (NNMF); the broader class of NNMF algorithms could be pertinent here because they would preserve the physical sense of the entries of the dose matrix. See, for example, Udell et al. (2016) for a detailed discussion of a broad class of low rank approximation methods or Tropp (2004) for random matrix algorithms used for dimensionality reduction in optimization.

2. Convex Treatment Planning

2.1. Formulation

For a case with m voxels inside a patient volume and n candidate treatment beams, we consider the class of inverse treatment planning problems of the form

$$\begin{array}{ll}\text{minimize} & f(y)\\ \text{subject to} & y = Ax, \quad x \geq 0,\end{array} \tag{1}$$

where the vectors of voxel doses, $y \in \mathbf{R}^m$, and beam intensities, $x \in \mathbf{R}^n$, are the optimization variables and $A \in \mathbf{R}^{m\times n}$ is a case-specific dose deposition matrix with nonnegative entries. (In the treatment planning literature, this matrix is also termed the "dose influence matrix" or "dose information matrix.")

The constraint $y = Ax$ expresses the physical relationship between beam intensities and delivered dose. The (element-wise) inequality constraint on x corresponds to the fact that it is physically impossible to deliver beams of negative intensity. The function $f : \mathbf{R}^m \to \mathbf{R}$ is assumed to be convex, and is constructed to penalize voxel doses according to clinical objectives.

Any desired treatment plan can be characterized (at least partially) by a vector of nonnegative doses $d \in \mathbf{R}^m_+$ prescribed to each voxel. Convex objectives used in the literature typically penalize the deviation of the calculated dose y from the prescribed dose d, or calculate a penalty on y in relation to some dose statistics. Common examples include one-sided and piecewise-quadratic penalties, piecewise-linear penalties, and conditional value at risk penalties; we refer the reader to Romeijn et al. (2003) and Kessler et al. (2005) for comprehensive surveys of objective functions in treatment planning.

In this work we consider the case of a fully separable objective given by

$$f(y) = \sum_{i=1}^{m} w_i f_i(y_i),$$

where each $f_i : \mathbf{R} \to \mathbf{R}$ is a convex function parametrized by a target dose d_i and $w_i > 0$ is a nonnegative weight. We take $d_i = 0$ for indices i corresponding to nontarget voxels and $d_i > 0$ according to a clinical prescription for target voxels.

The m voxels of the treatment plan are grouped into N delineated structures, such as the planning target volume (PTV), various sensitive structures termed organs at risk (OARs), and unlabeled tissue. We assume that each voxel index i is assigned uniquely to a set S_s such that $\bigcup_{s=1}^{N} S_s$ covers all voxel indices and $S_s \bigcap S_{s'} = \emptyset$ for $s \neq s'$. Structures can be prioritized to resolve the identity of voxels assigned to multiple structures during the clinical contouring process. We choose our voxel penalties to be uniform within structures: for each structure index s we have $d_i = d_{i'}$, $w_i = w_{i'}$, and $f_i = f_{i'}$ for $i, i' \in S_s$.

2.1.1. Optimality.

We denote the optimal value of (1) as $p^\star$. Any point (x, y) for which $y = Ax$ and $x \geq 0$ hold is said to be feasible. If we further have that $f(y) = p^\star$, then the point is optimal. We can express the suboptimality of any feasible point (x, y) as

$$\frac{f(y) - p^\star}{f(y)}.$$

Because the objective f is a weighted sum of objectives concerning each structure, the problem (1) can be interpreted as a linear scalarization of a multiobjective optimization; the particular scalarization is given by the choice of weights w as described above. Each choice of w represents a different tradeoff between the structure objectives, and solving (1) for that w yields a Pareto-optimal treatment plan for different clinical tradeoffs.

2.2. Dual Problem

We now derive the dual to our treatment planning problem. The Lagrangian of the problem (1) is given by

$$L(x, y, \nu, \lambda) = f(y) + \nu^T(y - Ax) - \lambda^T x,$$

with $\nu \in \mathbf{R}^m$ as the dual variable associated with the constraint $y = Ax$, and $\lambda \in \mathbf{R}^n_+$ as the nonnegative dual variable associated with the inequality constraint $x \geq 0$. The dual objective is defined as $g(\nu, \lambda) = \inf_{x,y} L(x, y, \nu, \lambda)$. Applying this definition, we obtain the dual problem

$$\begin{array}{ll}\text{maximize} & -f^*(\nu)\\ \text{subject to} & A^T\nu \geq 0,\end{array} \tag{2}$$

where f^* is the convex conjugate of f. This formulation implicitly carries the constraint $\nu \in \mathbf{dom}\,(f^*)$.

2.2.1. Dual Optimality and Suboptimality Bounds.

We denote the optimal value of (2) as $d^\star$, and we have $d^\star = p^\star$ when strong duality holds. (This condition holds for

all the examples of clinically relevant convex objectives discussed in Section 2.1.) Any $\nu \in \mathbf{dom}\,(f^*)$ for which the constraint $A^T\nu \geq 0$ hold is said to be dual feasible and

$$-f^*(\nu) \leq p^\star,$$

for all dual feasible ν. Consequently, for some feasible pair of variables (x, y), given *any* dual feasible ν, $-f^*(\nu)$ is a lower bound on $p^\star$ and when $-f^*(\nu)$ is nonnegative we can certify that the suboptimality of (x, y) as a solution to (1) is at most

$$\frac{f(y) - p^\star}{f(y)} \leq \frac{f(y) + f^*(\nu)}{f(y)}. \tag{3}$$

2.3. Example Objective Function

The techniques presented in this work are applicable to any fully separable convex objective $f(y) = \sum w_i f_i(y_i)$, and the exposition throughout the sequel is developed for this general class except where noted otherwise. Here, we introduce a specific objective function that we later use in our numerical experiments, and restate the primal and dual problems for this choice of objective. Specifically, we let f_i be the piecewise-linear function

$$f_i(y_i) = \frac{w_i^-}{w_i}(y_i - d_i)_- + \frac{w_i^+}{w_i}(y_i - d_i)_+,$$

where the scalar operation $(\cdot)_+$ is shorthand for $\max(0, \cdot)$ and similarly $(\cdot)_-$ is shorthand for $-\min(0, \cdot)$. This voxel objective imposes linear penalties on both underdose and overdose to the ith voxel, and we have introduced positive parameters w_i^- and w_i^+ to represent the relative weights of the underdose and overdose terms.

The product $w_i f_i(y_i)$ can also be written as

$$w_i f_i(y_i) = b_i|y_i - d_i| + c_i y_i + e_i,$$

where $b_i = (w_i^+ + w_i^-)/2$, $c_i = (w_i^+ - w_i^-)/2$, and $e_i = -c_i d_i$. Summing the weighted objective contributions and applying the constraints from (1) yields the following intensity optimization problem:

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{m} b_i|y_i - d_i| + c_i y_i + e_i \\ &\text{subject to} && y = Ax, \quad x \geq 0. \end{aligned} \tag{4}$$

This choice of f yields the dual problem

$$\begin{aligned} &\text{maximize} && -d^T\nu \\ &\text{subject to} && A^T\nu \geq 0, \quad |\nu - c| \leq b, \end{aligned} \tag{5}$$

where the absolute value and inequalities are understood to hold elementwise.

We note (for later reference) that the choice of $d_i = 0$ for nontarget voxels, in conjunction with the use of piecewise linear objectives and the constraint $x \geq 0$, implies that we can equivalently use a linear objective for nontarget voxels—that is, $w_i f_i(y_i) = w_i y_i$. We further note the incidental and possibly beneficial property that this penalty promotes sparsity in nontarget voxel doses and, by the relationship $y = Ax$, sparsity in the beam intensities.

2.4. Large-Scale Treatment Planning

Because solving (1) alone is sufficient to produce a treatment plan, we now explain the value of the dual problem (2) and the suboptimality bound (3) in large-scale treatment planning. Suppose we have a treatment planning problem that, despite the use of modern hardware and the fastest available optimization methods, is too large to be solved in a clinically acceptable timeframe—for example, a plan in which we consider tens of thousands of candidate beams and a dose grid of several hundred thousand voxels. Although we can write an optimization of the form (1) to represent our problem, we cannot solve that exact problem in the available time.

However, we may instead choose to solve smaller, computationally tractable approximations to this problem that can still be phrased in the form given by (1). In Sections 3 and 4, we will discuss two methods for generating such approximations. When we solve an optimization problem, we get both a primal optimal and a dual optimal point. Thus, upon solving one of our proposed approximations, we obtain a solution $(\tilde{x}^\star, \tilde{y}^\star, \tilde{\nu}^\star)$ (optimal for that reduced problem), from which we can construct a solution $(\hat{x}, \hat{y}, \hat{\nu})$ that is feasible for the large-scale problem. By virtue of being primal feasible, this solution will be physically achievable; by virtue of being dual feasible, we can use (3) to mathematically guarantee a maximum suboptimality for this solution with respect to the large-scale problem. In other words, if the suboptimality bound is P%, we can guarantee that the treatment plan obtained by solving the reduced problem is at most P% worse than the best achievable plan for the full problem—without ever paying the full computational cost of solving the large-scale problem.

3. Voxel Clustering

3.1. Formulation

We consider approximations to the dose-deposition matrix A obtained by clustering voxels (i.e., clustering rows of the matrix). The approximate dose-deposition matrix A^{vclu} can be written as the product of an upsampling matrix $U \in \mathbf{R}^{m\times k}$ and a voxel-clustered matrix $A_{\mathcal{R}} \in \mathbf{R}^{k\times n}$,

$$A^{\text{vclu}} = U A_{\mathcal{R}} \approx A.$$

From the above equation, we can see that voxel clustering is a special case of approximate matrix factorization. In particular, we have that $A_{\mathcal{R}}$ represents in k

rows (or voxel clusters) an approximation of the information contained in the m rows (or voxels) of A, whereas U maps each cluster to its associated voxels. The entries of U are given as

$$U_{i\kappa} = \begin{cases} 1, \text{ voxel } i \text{ assigned to cluster } \kappa \\ 0, \text{ otherwise,} \end{cases}$$

implying that U contains exactly one nonzero entry per row. For each cluster κ, the corresponding set $C_\kappa = \{i \,|\, \text{voxel } i \in \text{cluster } \kappa\}$ contains the indices of the voxels assigned to that cluster.

For a given set of voxel-to-cluster assignments given by U, if we choose to represent the rows assigned to each cluster by their mean, the explicit formula to construct clustered matrix $A_\mathcal{R}$ is

$$A_\mathcal{R} = (U^T U)^{-1} U A.$$

We define a vector $\omega = U^T \mathbf{1} = \mathbf{diag}(U^T U)$ whose entries give the number of voxels assigned to each cluster—that is, $\omega_\kappa = |C_\kappa|$. To avoid ambiguities regarding the mapping of the voxel clusters to structures, we restrict each cluster to contain only voxels from the same planning structure.

We can write an approximation of our full problem (1) as a smaller problem defined in terms of our clustered matrix $A_\mathcal{R}$,

$$\begin{aligned} &\text{minimize} && \sum_{\kappa=1}^{k} \tilde{w}_\kappa f_\kappa(y_\kappa) \\ &\text{subject to} && y_\mathcal{R} = A_\mathcal{R} x_\mathcal{R}, \quad x_\mathcal{R} \geq 0, \end{aligned} \tag{6}$$

with optimization variables $x_\mathcal{R} \in \mathbf{R}^n$, $y_\mathcal{R} \in \mathbf{R}^k$, weight vector $\tilde{w} \in \mathbf{R}^k$, and (implicitly) prescription $\tilde{d} \in \mathbf{R}^k$. The entries of $\tilde{d}$ are given by $\tilde{d}_\kappa = d_i$ for voxel i in cluster κ, which is uniquely defined for the reasons that $d_i = d_{i'} = d_s$ for $i, i' \in S_s$ and that, by choice, the clustering respects structure boundaries. Similarly, we have $f_\kappa = f_i$ and we choose $\tilde{w}_\kappa$ such that $\tilde{w}_\kappa = |C_\kappa| w_i = |C_\kappa| w_s$ for voxel i in cluster κ and structure s. Under these definitions, if $U A_\mathcal{R} = A$ holds exactly, then the problems (1) and (6) are equivalent. Otherwise, by Jensen's inequality, for any $x \geq 0$ we have

$$\begin{aligned} \sum_{\kappa=1}^{k} \tilde{w}_\kappa f_\kappa(y_\kappa) &= \sum_{\kappa=1}^{k} w_s |C_\kappa| f_\kappa\left(\frac{1}{|C_\kappa|} \sum_{i \in C_\kappa} \tilde{a}_i^T x\right) \\ &\leq \sum_{\kappa=1}^{k} w_s \sum_{i \in C_\kappa} f_\kappa(\tilde{a}_i^T x) = \sum_{i=1}^{m} w_i f_i(y_i), \end{aligned}$$

where $\tilde{a}_i \in \mathbf{R}^n$ is the ith row of A. This shows that (6) is a relaxation of (1).

Solving the reduced problem (6) instead of (1) will produce a feasible, but not necessarily optimal, vector of beam intensities x. However, by choosing $k \ll m$ we make the voxel-clustered planning problem much smaller than the original, and obtain a commensurate reduction in planning time.

3.2. Bounding Procedure

Given a primal optimal point $(x^\star_\mathcal{R}, y^\star_\mathcal{R})$ for which the voxel-clustered problem attains its optimal value $p^\star_\mathcal{R}$, we seek upper and lower bounds on $p^\star$. To obtain an upper bound, we set $\hat{x} = x^\star_\mathcal{R}$. Because $x^\star_\mathcal{R} \geq 0$, $\hat{x}$ is feasible for (1). We define $\hat{y} = A x^\star_\mathcal{R}$, and an upper bound is given simply by

$$p_{\text{ub}} = f(\hat{y}) = \sum_{i=1}^{m} w_i f_i(\hat{y}_i) \geq p^\star.$$

To obtain a lower bound, it is sufficient to recall that (6) is a relaxation of (1); hence,

$$p_{\text{lb}} = p^\star_\mathcal{R} \leq p^\star.$$

We can therefore guarantee the suboptimality of the solution given by $(\hat{x}, \hat{y})$ to be bounded by the expression

$$\frac{f(\hat{y}) - p^\star_\mathcal{R}}{f(\hat{y})}. \tag{7}$$

3.3. Voxel Collapse for Nontarget Structures

We present a special case of voxel clustering that applies to choices of f that impose linear penalties on nontarget voxels, as it provides an opportunity for significant computational savings.

For the piecewise linear objective used in (4) and our choice of prescribed dose $d_i = 0$ for nontarget voxels, as noted in Section 2.3, the objective contribution of nontarget voxels is simply the linear term $w_i y_i$. Because $w_i = w_{i'} = w_s$ for $i, i' \in S_s$, with trivial rearrangement, the objective contribution of nontarget structure s can be written as a linear function of $\bar{y}_s$, the mean dose to that structure:

$$\sum_{i \in S_s} w_i y_i = w_s \sum_{i \in S_s} y_i = w_s |S_s| \bar{y}_s.$$

Let $A_s \in \mathbf{R}^{|S_s| \times n}$ be the submatrix formed by gathering the rows $\tilde{a}_i$ of A corresponding to voxels in structure s. If we denote as $\bar{a}_s$ the average of the rows of A_s, we have $\mathbf{1}^T A_s = |S_s| \bar{a}_s$. For a given x, the product $\bar{a}_s^T x = \bar{y}_s$ is simply the mean dose on structure s for a given x. Thus, letting $\mathcal{T}$ be the set of target structures, $\mathcal{N}$ be the set of nontarget structures, and $A_{\text{target}} \in \mathbf{R}^{m_t \times n}$ and $y_{\text{target}} \in \mathbf{R}^{m_t}$ be the submatrix of A and subvector of y formed by gathering all target voxel rows, respectively, the problem (4) can be written as the smaller problem,

$$\begin{aligned} &\text{minimize} && \sum_{s \in \mathcal{T}} \sum_{i \in S_s} (b_s |y_i - d_s| + c_s y_i - e_s) + \sum_{s \in \mathcal{N}} w_s |S_s| \bar{y}_s \\ &\text{subject to} && y_{\text{target}} = A_{\text{target}} x, \\ &&& \bar{y}_s = \bar{a}_s^T x, \quad s \in \mathcal{N} \end{aligned}$$

with no approximation involved. (Here, we have also applied the definitions $b_i = b_s$, $c_i = c_s$, and $e_i = e_s$.) This substitution is effectively an $|S_s| : 1$ voxel clustering for each nontarget structure s, and it reduces the problem dimension from $\mathbf{R}^{m \times n}$ to $\mathbf{R}^{(m_t + |\mathcal{N}|) \times n}$, where $m_t = \sum_{s \in \mathcal{T}} |S_s|$ is the total number of target voxels and $|\mathcal{N}|$ is

the number of nontarget structures. Of course, most cases have many more target voxels than nontarget structures, so $m_t \gg |\mathcal{N}|$ usually holds, so we expect a speed-up in solve times that is proportional to m/m_t while yielding the exact solution.

When using the piecewise linear objective specified in (4), to minimize approximation error while maximizing computational speed-up, voxel clustering should be used for target structures, whereas voxel collapse should be applied to nontarget structures.

3.4. Clustering Procedure

In Section 3.1, the cluster assignments represented by the matrix U were assumed to be given, and we understood $U A_{\mathcal{R}} \approx A$ to hold without specifying the sense in which the product $U A_{\mathcal{R}}$ was to approximate A.

Several decisions are involved in the clustering process, most notably: a rule for generating cluster assignments, given a set of vectors; a rule for computing a prototype vector to represent each cluster; and an algorithm (typically a heuristic) that carries out the two rules. For instance, the well-known k-means clustering seeks to partition m vectors into k clusters, where each cluster is associated with a centroid defined as the mean of its assigned vectors, and each vector is assigned to the cluster that has the centroid that is nearest in the ℓ_2-norm. (In other words, we seek U and $A_{\mathcal{R}}$ that minimize $|A - U A_{\mathcal{R}}|_2^2$, subject to the constraint that each row of U must contain exactly one nonzero entry with value 1.) The problem is nondeterministic polynomial time-hard and sensitive to the initial choice of centroids; and a commonly used heuristic is Lloyd's algorithm, which can be summarized as the following algorithm:

Algorithm 3.1 (Lloyd's Algorithm for k-Means (Lloyd 1982))

given points $p_i \in \mathbf{R}^r, i = 1, \ldots, q$, cluster number k, uninitialized centroids $c_\kappa \in \mathbf{R}^q, \kappa = 1, \ldots, k$, and point to cluster assignments represented as a vector $u \in \mathbf{Z}^m$, with $u_i = \kappa$ if point i assigned to cluster κ.

repeat

1. *Calculate centroids.*

   repeat

   $$c_\kappa = \sum_i^m a_i \cdot (u_i = k) / \sum_i^q (u_i = k)$$

   for $\kappa = 1, \ldots, k$

2. *Update assignments.*

   repeat

   $$u_i = \arg\min_\kappa |p_i - c_\kappa|_2^2$$

   for $i = 1, \ldots, q$

until assignments stable or an iteration limit is reached

return assignments u, centroids $\{c_1, \ldots, c_k\}$.

For a thorough treatment of many popular and relevant clustering methods, a comparison of their strengths and drawbacks, as well as of their computational complexities, we refer the reader to the reviews Jain et al. (1999) and Xu and Wunsch (2005, 2010).

The clustering method used may generate cluster assignments, given by U, and cluster prototypes, given by $A_{\mathcal{R}}$, jointly. Alternatively, the method may simply provide cluster assignments (such as grouping voxels in regularly sized clusters based on geometric adjacency) which may be used to construct the rows of $A_{\mathcal{R}}$. In this work, we set $A_{\mathcal{R}} = (U^T U)^{-1} U^T A$, which corresponds to averaging the elements in each cluster, as stated in Section 3.1. Given the cluster assignments encoded in U, however, one could equally well calculate the mean, median, or any convex combination of the clustered elements. In particular, choosing a single element to represent the cluster is strongly related to work on random sampling and importance sampling, both of which have been studied in the context of dimensionality reduction for radiation treatment planning (Martin et al. 2007).

Although the choice of clustering method (and its parameters and initialization) likely influences the quality of the approximate solutions obtained by solving (6), the choice of objective function f is also likely to play a large role. We leave the interesting—and possibly complicated—interplay between clustering methods and objective functions as a topic for future investigation; the focus of the present work is on methods to form and solve approximate planning problems given a clustered approximation to the dose matrix, as well methods to obtain case-, objective-, and approximation-specific optimality bounds.

In this work, we elect to use k-means clustering applied block-wise to each anatomical structure and implement a vectorized version of Lloyd's algorithm (Algorithm 3.2):

Algorithm 3.2 (Vectorized k-Means)

given data matrix $P \in \mathbf{R}^{q \times r}$, cluster number $k < q$, uninitialized centroid matrix $C \in \mathbf{R}^{k \times r}$, uninitialized distance matrix $D \in \mathbf{R}^{q \times k}$, and initialized point to cluster assignments represented as a matrix $U \in \{0,1\}^{m \times k}$, with $u_{i\kappa} = 1$ if point i assigned to cluster κ, and 0 otherwise.

repeat

1. *Calculate centroids.* $C = (U^T U)^{-1} U^T P$
2. *Update assignments.*

$$D = -2PC^T + \mathbf{1}\mathbf{diag}(C^T C)^T$$

$$u_{i\kappa} = \begin{cases} 1 & \kappa = \arg\min_{\kappa'} \{d_{i\kappa'}\} \\ 0 & \text{otherwise} \end{cases}, \quad i = 1, \ldots q, \quad \kappa = 1, \ldots, k$$

until assignments stable or an iteration limit is reached

return assignments U, centroids C.

(Note the modification to Lloyd's algorithm implied by the update rule for D: Because the cluster assignment for each i is determined by choosing the κ that minimizes $|p_i - c_\kappa|_2^2$, the contribution made to distance

$d_{i\kappa}$ by term $p_i^T p_i$ can be neglected without changing the minimizer.) When clustering the dose matrix by rows (voxel clustering), we use $p = m$, $q = n$, $P = A$, choose k to be appreciably smaller than m, and the centroid matrix C then corresponds to the matrix $A_{\mathscr{R}}$ introduced above. To perform voxel clustering by structure, we make the substitution $P = A_s \in \mathbf{R}^{m_s \times n}$ for each structure-specific submatrix A_s and choose a proportionally smaller k (e.g., $k_s = k \cdot m_s/m$). The centroid matrices C_s returned for each structure s would then be vertically concatenated to form $A_{\mathscr{R}}$. In order to cluster the dose matrix by columns (beam clustering), we substitute $p = n$, $q = m$, $P = A^T$, and choose k to be appreciably smaller than n.

4. Beam Clustering

4.1. Formulation

We turn to subproblems formed by clustering *columns* (i.e., beams, beamlets, or apertures) of our full dose deposition matrix A. The approximate dose matrix A^{bclu} can be written as the product of a column-clustered matrix $A_{\mathscr{C}} \in \mathbf{R}^{m \times k}$, and an upsampling matrix $V \in \mathbf{R}^{n \times k}$,

$$A^{\text{bclu}} = A_{\mathscr{C}} V^T \approx A.$$

As with the upsampling matrix U defined in (3), the entries of matrix V are given as

$$V_{j\kappa} = \begin{cases} 1 & \text{beam } j \text{ assigned to cluster } \kappa \\ 0 & \text{otherwise.} \end{cases}$$

Then, for a given set of beam-to-cluster assignments given by V, a clustered matrix $A_{\mathscr{C}}$ can be constructed as:

$$A_{\mathscr{C}} = AV(V^T V)^{-1}.$$

We can write an approximation of (1) as a smaller problem in terms of our clustered matrix $A_{\mathscr{C}}$,

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{m} w_i f_i(y_{\mathscr{C}i}) \\ \text{subject to} & y_{\mathscr{C}} = A_{\mathscr{C}} x_{\mathscr{C}}, \quad x_{\mathscr{C}} \geq 0, \end{array} \tag{8}$$

with optimization variables $x_{\mathscr{C}} \in \mathbf{R}^k$, $y_{\mathscr{C}} \in \mathbf{R}^m$. Here, the functions f_i, their prescription parameters d_i, and the weights w_i are the same as those used in (1). Solving (8) is equivalent to solving (1) with the added constraints

$$x_j = x_{j'}, \quad j, j' \in C_\kappa, \quad \kappa = 1, \ldots, k,$$

that is, the added condition that beam intensities must be equal for beams assigned to the same cluster. From this, it is clear that (8) is a restriction of (1). We label the optimal value of (8) as $p^\star_{\mathscr{C}}$, and its dual problem has the form

$$\begin{array}{ll} \text{maximize} & -f^*(\nu_{\mathscr{C}}) \\ \text{subject to} & A_{\mathscr{C}}^T \nu_{\mathscr{C}} \geq 0, \end{array} \tag{9}$$

for the dual variable $\nu_{\mathscr{C}} \in \mathbf{dom}\,(f^*) \subseteq \mathbf{R}^m$. Because this is a dual of a restriction of the full primal problem, it is a relaxation of the full dual problem.

4.2. Bounding Procedure

Given a solution $(x^\star_{\mathscr{C}}, y^\star_{\mathscr{C}}, \nu^\star_{\mathscr{C}})$ for which the column-clustered problem attains its optimal value $p^\star_{\mathscr{C}}$, we seek upper and lower bounds on $p^\star$.

Because we have $x^\star_{\mathscr{C}} \geq 0$, choosing $\hat{x} = V(V^T V)^{-1} x^\star_{\mathscr{C}}$ and $\hat{y} = A\hat{x} = A_{\mathscr{C}} x^\star_{\mathscr{C}}$ yields a pair of variables $(\hat{x}, \hat{y})$ that are feasible for (1). (This choice can be interpreted as evenly distributing the optimal intensity $x^\star_{\mathscr{C}\kappa}$ assigned to beam cluster κ among its constituent beams.) Thus, an upper bound to the value of (1) is given by

$$p^{\text{ub}} = f(\hat{y}) = p^\star_{\mathscr{C}}.$$

To obtain a lower bound, we seek a $\hat{\nu}$ that is dual feasible for (2). Because $\nu^\star_{\mathscr{C}}$ is dual feasible for (9), we have $A_{\mathscr{C}}^T \nu^\star_{\mathscr{C}} = (V^T V)^{-1} V^T A^T \nu^\star_{\mathscr{C}} \geq 0$, but not necessarily $A^T \nu^\star_{\mathscr{C}} \geq 0$. To obtain a feasible $\hat{\nu}$ at reasonable computational cost, we propose solving a problem that takes advantage of our infeasible estimate $\nu^\star_{\mathscr{C}}$.

Let $\nu^{(0)} = \nu^\star_{\mathscr{C}}$. Because the entries of A are nonnegative, we have $A^T\delta \geq 0$ for any $\delta \geq 0$ and $A^T(\nu^{(0)} + \delta) \geq 0$ for δ sufficiently large. We seek the smallest such δ [in the sense that $|-f^*(\delta)|$ is small] that we can add to the optimal solution of (9) to make it feasible on (2). In other words, we desire the solution to

$$\begin{array}{ll} \text{maximize} & -f^*(\nu^{(0)} + \delta) \\ \text{subject to} & A^T(\nu^{(0)} + \delta) \geq 0, \quad \delta \geq 0, \quad \nu^{(0)} + \delta \in \mathbf{dom}\,(f^*). \end{array}$$

However, because this problem has the same dimension as the full planning problem, we propose to solve one or more problems that do not exceed the dimension (i.e., $m \times k$) or complexity of the clustered problem.

Let $\mathcal{J} = \{a_j \,|\, a_j^T \nu^{(0)} < 0\}$ be the subset of the columns a_j of A that are associated with infeasible dual constraints. If $|\mathcal{J}|$ exceeds the clustered dimension k, we form a matrix $\hat{A}^{(1)}_{\mathscr{C}} \in \mathbf{R}^{m \times k}$ from the top k columns with the largest margins of violation—that is, the k columns a_j with the most negative values of $a_j^T \nu^{(0)}$. We then solve the problem

$$\begin{array}{ll} \text{maximize} & -f^*(\nu^{(0)} + \delta) \\ \text{subject to} & \hat{A}^{(1)T}_{\mathscr{C}}(\nu^{(0)} + \delta) \geq 0, \quad \delta \geq 0, \\ & \nu^{(0)} + \delta \in \mathbf{dom}\,(f^*), \end{array} \tag{10}$$

and ignore the remaining columns of A because it is guaranteed by the nonnegativity of δ that any feasible entries of $A^T \nu^{(0)}$ will only become feasible with greater margin upon the addition of δ.

In other words, we take a greedy approach to estimating the k columns a_j for which the constraints $a_j^T \nu \geq 0$ are the most restrictive, in the hopes of solving a problem (of the same size and cost as the clustered problem) whose solution will satisfy $a_j^T \nu \geq 0$ for all n constraints. Of course, when $|\mathcal{J}| \leq k$, solving (10) satisfies all remaining constraints directly.

We check whether the optimal $\delta^{\star(1)}$ produced by (10) satisfies $A^T(\nu^{(0)} + \delta^{\star(1)}) \geq 0$. If this constraint holds, our

task is complete. If the constraint fails to hold, we set $\nu^{(t)} = \nu^{(t-1)} + \delta^{\star(t)}$ and repeat the above procedure. We do this for T such iterations, until

$$A^T\left(\nu^{(0)} + \sum_{t=1}^{T} \delta^{\star(t)}\right) \geq 0,$$

holds. At this point, we take $\hat{\nu} = \nu^{(0)} + \sum_{t=1}^{T} \delta^{\star(t)}$ to be our feasible dual variable, and the corresponding lower bound is given by

$$p^{\text{lb}} = -f^*\left(\nu^{(0)} + \sum_{t=1}^{T} \delta^{\star(t)}\right).$$

Because we require δ nonnegative in each subproblem, we are guaranteed to reduce the number of infeasible constraints by at least k on each solve t. In theory, the number T of subproblems we are required to solve to obtain a feasible $\hat{\nu}$ could approach n/k; in practice, we find that a single subproblem is sufficient. We summarize the full procedure in Algorithm 4.1.

Algorithm 4.1 (Beam Clustering Lower Bound)
given an initial point $\nu^{\star}_{\mathcal{C}}$ optimal for (9).
$\nu^{(0)} := \nu^{\star}_{\mathcal{C}}$.
$t := 1$.
repeat
1. *Fix dimension.* Form $\mathcal{J} = \{a_j \mid a_j^T \nu^{(t-1)} < 0\}$. Then, $k^{(t)} := \min(k, |\mathcal{J}|)$.
2. *Approximate.* **repeat**
$\hat{a}_\kappa = \arg\min\{a_j^T \nu^{(t-1)} \mid a_j \in \mathcal{J}\}$
$\mathcal{J} := \mathcal{J} \backslash \{\hat{a}_\kappa\}$
for $\kappa = 1, \ldots, k^{(t)}$.
3. *Solve.* Set the value of $\delta^{\star(t)}$ to a solution of the convex problem
$\underset{\delta}{\text{minimize}} \quad f^*(\nu^{(t)} + \delta)$
subject to $\hat{A}_{\mathcal{C}}^{(t)T}(\nu^{(t-1)} + \delta) \geq 0$.
4. *Update dual variable.* $\nu^{(t)} := \nu^{(t-1)} + \delta^{\star(t)}$.
5. *Update iteration.* $t := t + 1$.
until $A^T \nu^{(t)} \geq 0$.
return lower bound $-f^*(\nu^{(t)})$.

5. Examples

5.1. Voxel Collapse

The submatrices of A corresponding to nontarget structures were averaged to form $A_{\text{collapsed}}$, defined as

$$A_{\text{collapsed}} = \begin{bmatrix} A_{\text{target}} \\ A_{\text{non-target collapsed}} \end{bmatrix} = \begin{bmatrix} A_{\text{target}} \\ (1/|S_1|)\mathbf{1}^T A_{s_1} \\ \vdots \\ (1/|S_N|)\mathbf{1}^T A_{s_N} \end{bmatrix},$$

where $|S_t|$ is the number of voxels in structure s_t.

5.1.1. Small Problem Instance. A 268,228-voxel, 360-aperture VMAT head and neck case was used for reoptimization of the aperture intensities. The plan comprised three target regions: the PTV, treated to 66 Gray; a second lesion treated to 60 Gray; and lymph nodes also treated to 60 Gray. The plan also contained 14 other structures, including the brain, brain stem, spinal cord, optic nerve, optic chiasm, cochlea, and parotid gland. Unlabeled tissue was also included in the objective.

The optimization was formulated to solve (4) using the piecewise linear objectives introduced in Section 2.3. We have found that a good default setting for objective weights is to set the underdosing penalty to $w_i^- = 1$ and overdosing penalty to $w_i^+ = 1/20$ for target structures. We set $w_i = 1/30$ for nontarget structures. We subsequently normalize all weights by the number of voxels in its corresponding structure. Unless mentioned otherwise, we use these weights by default throughout our experiments. For this and all other examples, we graded the resulting doses to each structure against Quantitative Analyses of Normal Tissues Effects in the Clinic (QUANTEC) reference guidelines (Bentzen et al. 2010) as a first cut for obtaining clinically reasonable plans.

For this problem instance, we modified the default weights for the spinal cord ($w_i = 1.2$), spinal canal ($w_i = 1.3$), and brainstem ($w_i = 3.5$), as well as the overdosing penalty for the primary target ($w_i^+ = 0.9$) to meet QUANTEC guidelines.

We then compressed the dose matrix for this case in a lossless manner (relative to the choice of piecewise linear objective), yielding a dose matrix of 11,253 voxels and 360 apertures, or 24-fold compression. Planning was performed at the nominal objective weights introduced above, and the weight for each collapsed (mean dose) term was multiplied by the size of the corresponding structure, $|S_s|$, as specified in Section 3.3, so that the objective value coincided exactly with that of (4).

Additionally, we performed 61 warm-start trials by reoptimizing for different objective weights while using optimal variables generated from the previous solve to initialize the optimizer. We explored weights around the nominal set of weights by choosing one structure and scaling its objective weight by a fixed factor (e.g., a 20% increase) until the resulting optimal dose vector $y^\star$ failed to meet QUANTEC dosing guidelines. After each such failure, we reset the chosen structure's objective weight back to its nominal weight and repeated the procedure for another (unvisited) planning structure until all structures had been visited. For target structures, we scaled the overdose penalty while leaving the underdose penalty at its nominal value. This procedure, which samples a portion of the Pareto surface for the head and neck case, yielded 61 warm-start plans when using a fixed scaling factor of 1.2.

5.1.2. Large Problem Instance. We also planned a much larger case, a 589,467 voxel by 68,208 beamlet prostate FMO problem. This matrix contains 865 million nonzero

Table 1. Timing Results for Voxel Collapse, CPU

Case	State	Dimensions	Setup time(s)	Solve time(s)	Iterations
Head and neck	Full	268,228 × 360	28.1	267.4	3,117
Head and neck	Collapsed	11,253 × 360	0.4	1.0	203
Prostate	Full	589,467 × 68,208	1,500	*	*
Prostate	Collapsed	6,054 × 68,208	260.0	190.1	258

*Unconverged after 7 hours, 16 iterations.

entries, which occupies 19 GB of storage in column compressed sparse format. We were not able to run experiments to completion with the full version of this matrix on central processing unit (CPU), and the matrix did not fit on a single graphics processing unit (GPU).

Collapsing the nontarget structures yielded a dose matrix of 6,054 voxels and 68,208 beamlets, or 3.1 GB when stored as a dense matrix. Planning was performed by using the default weights introduced above; no modifications were needed to meet QUANTEC dosing guidelines.

In addition, we iteratively replanned the case for different objective weights using the same warm-starting procedure described for the head and neck case. This yielded 207 warm-start plans when using a fixed scaling factor of 1.2.

5.1.3. Computational Details. Optimizations were performed in the Python interface to POGS (Fougner and Boyd 2015), which calls a C or CUDA solver. The CPU version was implemented with OpenMP and was run on 32 threads on a 32-core/64-thread, 2.20GHz Intel Xeon CPU E5-4620; the GPU version was executed on a nVidia TitanX. The same hardware was used for all examples below.

A free, open-source Python implementation of the clustered (and full) intensity optimization problems and bounding methods described in this paper is available at https://github.com/bungun/rad_cluster. The repository includes one example each for the voxel-clustered and beam-clustered methods; the scripts are identical to those used to generate the results presented in the sequel, except that the clinical dose matrices are replaced with randomly generated synthetic data as placeholders.

5.1.4. Results. For the head and neck case, voxel collapse resulted in a 200-fold speedup when working on the CPU and an 11-fold speedup on the GPU, as documented in Tables 1 and 2. We note that the GPU was about 15 times faster to begin with and that the setup (which included matrix equilibration and Cholesky factorization) plus solution times became comparable for the reduced size problems; however, with the larger collapsed matrix in the prostate case, we continue to observe a 20-fold speed advantage on the GPU.

For the prostate case objective weight sweeps, we obtained median solve times (and ranges) of 18.5 s (4.9–128.3 s) on the CPU and 1.1 s (0.2–6.3 s) on the GPU.

5.2. Voxel Clustering

5.2.1. Clustering. Vectors corresponding to the rows of A_{target} were clustered by using k-means clustering, whereas voxel collapse was applied to the voxels of each nontarget structure. Clustering was performed separately for the rows (voxels) of each target structure's submatrix A_s.

Although we implemented a naive version of the k-means clustering as described in Algorithm 3.2, accelerated variants exist, such as minibatch k-means (Scherrer et al. 2005, Sculley 2010, Ganage et al. 2013, Boutsidis et al. 2015).

5.2.2. Sketched k-Means. Although voxel clustering is intended to be used in cases when the dimension m is large, the dimension n may also be large if many candidate beams are under consideration. In such situations, it may be prohibitively slow to run the k-means algorithm that produces a smaller $A_{\mathcal{R}}$ that would enable efficient treatment planning. In such cases, we propose

Table 2. Timing Results for Voxel Collapse, GPU

Case	State	Dimensions	Setup time(s)	Solve time(s)	Iterations
Head and neck	Full	268,228 × 360	7.8	15.5	1,210
Head and neck	Collapsed	11,253 × 360	1.7	0.3	203
Prostate	Full	589,467 × 68,208	*	*	*
Prostate	Collapsed	6,054 × 68,208	15.6	8.9	258

*Case does not fit on a single nVidia TitanX GPU.

Table 3. Timing and Suboptimality Results for Voxel Clustering, CPU

Compression	Dimensions	Suboptimality bound (%)	Setup+solve time (s)	Mean solve time, warm start (s)
(Collapsed)	11,253 × 360	—	1.25	0.55
10	1,036 × 360	1.3	0.22	0.13
20	534 × 360	2.0	0.22	0.08
30	364 × 360	2.2	0.17	0.06
50	221 × 360	3.4	0.25	0.09
100	111 × 360	5.6	0.05	0.03

sketching the matrix A by multiplication with a random matrix $\Omega \in \mathbf{R}^{n\times r}$,

$$A_{\text{sketch}} = A\Omega,$$

to obtain a smaller matrix $A_{\text{sketch}} \in \mathbf{R}^{m\times r}$. Clustering is then performed on A_{sketch} to obtain upsampling matrix $U \in \mathbf{R}^{m\times k}$ and clustered matrix $B \in \mathbf{R}^{k\times r}$ such that $A \approx UB$. This U is then used to form $A_{\mathcal{R}}$. In our experience, drawing the entries of Ω from the normal distribution $\mathcal{N}(0,1)$, choosing $r = \max(k, n/20)$, and running k-means on the sketched rows yields results comparable to running k-means on the original rows.

5.2.3. Problem Instance. Voxel clustering was performed on the dose matrix for the head and neck case introduced in Section 5.1. Clustering was performed to approximately 10-, 20-, 30-, 50-, and 100-fold compression levels, yielding compressed matrices of sizes 1,036 × 360, (534 × 360), (364 × 360), (221 × 360), and (111 × 360).

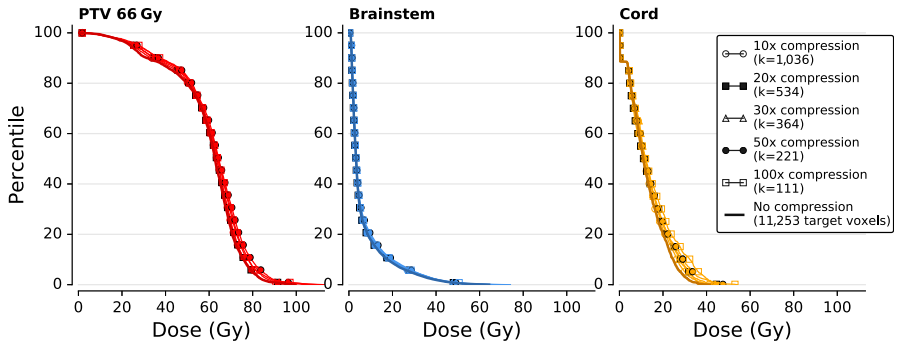
For each instance, we solved (6) using the piecewise linear objective discussed in Section 2.3, using our default objective weights (Section 5.1), with the modification that the weights for each clustered (or collapsed) metavoxel were multiplied by the number of elements in its cluster (or structure) so that the objective value of (6) coincided approximately with that of (1). The objective weight sweep carried out in Section 5.1 to generate plans sampling the Pareto surface was repeated for the clustered problem instances. (Because these plans were solutions to approximations of the full problem, we were in fact sampling the feasible region near the Pareto surface.)

5.2.4. Computational Details. Clustering was performed on a 32-core, 2.20-GHz Intel Xeon CPU E5-4620 processor in Julia (Bezanson et al. 2014), with point-to-cluster distance calculations and comparisons vectorized as described in Algorithm 3.2 and cast as BLAS operations (Lawson et al. 1979; Dongarra et al. 1988, 1990); we used the Julia language-default of eight parallel threads for BLAS operations.

5.2.5. Results. Results for the voxel clustering approximations are summarized in Table 3. Encouragingly, the largest maximum-suboptimality gap was 6% for the 100-fold compressed approximation, whereas all other approximations yielded solutions that were guaranteed to be within 1%–4% of the true optimal value.

In Figure 1, we observe that the dose-volume histograms (DVHs) for the voxel-clustered plans are nearly identical across compression levels (with the 100-fold compression plan deviating the most from the others). The voxel-clustered plans are dosimetrically comparable to the plan obtained using the uncompressed dose matrix. In particular, we observed that for a clustered plan with s% suboptimality, the doses

Figure 1. (Color online) Dose Volume Histograms for Head and Neck Treatment Plans Generated Using Voxel Clustering



Notes. Results are shown for five levels of k-means compression applied to the target structures, as well as the uncompressed plan (solid line) shown for reference. Voxel collapse was used for nontarget structures in all plans. The same objective weights were used to generate each solution.

Table 4. Timing and Suboptimality Results for Beam Clustering, CPU

Compression	Dimensions	Suboptimality bound (%)	Primal solve time (s)	Dual solve time (s)
(Collapsed)	6,054 × 68,208	—	121.6	—
10	6,054 × 6,821	100.0	15.2	0.0
20	6,054 × 3,410	10.8	6.7	0.0
30	6,054 × 2,274	29.2	1.9	3.0
50	6,054 × 1,364	37.7	0.7	4.3
100	6,054 × 682	53.7	0.2	1.7

achieved for the clustered problems at the percentiles specified by the QUANTEC guidelines were within s% of the doses achieved for the full problem.

Solve times (cold start) ranged from 0.05 s at maximum compression to 0.22 s at minimum compression on the CPU, representing a 6- to 25-fold speedup. Solve times on the GPU averaged 0.52 s across compressions, or about the same as the nonclustered, collapsed problem—the clustered approximations of this problem are effectively too small to benefit from GPU acceleration. GPU and CPU suboptimality bounds agreed within 0.1%.

Warm-start solve times for the objective sweep averaged in the hundredths to low tenths of seconds, making it conceivable to sample thousands of points on the Pareto surface and thereby form an MCO planning library in a few minutes.

5.3. Beam Clustering

5.3.1. Clustering. Vectors corresponding to the columns of $A_{\text{collapsed}}$ (as defined in Section 5.1) were clustered into k column clusters (i.e., aggregate beams) by using k-means clustering. For a desired compression factor ϕ, $k = \lceil n/\phi \rceil$ initial clusters were generated by assigning approximately $n_{\text{clu}} = \lceil n/k \rceil$ columns to each cluster. Because sequentially indexed columns of A correspond to the dose deposition data for candidate beams that are usually "nearby" in some sense (e.g., apertures on the same arc with small angular separation or adjacent beamlets in a fluence map), we would expect the numerical content of such columns to be similar, so taking sequential blocks of width n_{clu} is a reasonable initialization for the clusters.

5.3.2. Problem Instance. The prostate case introduced in Section 5.1 was clustered to approximate compression levels of 10-, 20-, 30-, 50-, and 100-fold compression, yielding clustered matrices sized 6,054 × 6,821, 6,054 × 3,410, 6,054 × 2,274, 6,054 × 1,364, and 6,054 × 682.

Planning was performed at the default objective weights, and the objective weight sweep performed for the voxel-collapsed version of the prostate case was repeated for each instance of the clustered approximations to the case.

5.3.3. Results. For cold-start problems, in addition to the unmeasured speedup obtained through voxel collapse of nontarget structures, beam clustering resulted in 8- to 64-fold speed gains on the CPU to solve the clustered problem and find a bound and 8- to 600-fold speedup for primal solve alone. On the GPU, we obtained smaller speedups of 4- to 7-fold when considering the time for both the primal solve and the bounding procedure or 6- to 45-fold when considering the primal solve time only. For both hardware configurations, the bounds ranged from 11%–54% without a strong correlation to the degree of clustering; for the 10-fold compression level, the bounding procedure produced a feasible dual variable with a negative objective value; in this case, we took zero as a trivial lower bound (because we know our objective value is nonnegative), and had a suboptimality bound of 100%; we discuss this failure of our bounding procedure in the sequel. The bounds obtained and timing results are summarized in Tables 4 and 5.

On both CPU and GPU, the lower bounding procedure failed to produce a nontrivial bound for the cold-start runs of the 10-fold compression level; similar failures occurred for about 20% of the warm-start runs at all compression levels. Recalling that the column-clustered dual problem is a relaxation of the full dual (as the column-clustered primal problem is a restriction

Table 5. Timing and Suboptimality Results for Beam Clustering, GPU

Compression	Dimensions	Suboptimality bound (%)	Primal solve time (s)	Dual solve time (s)
(Collapsed)	6,054 × 68,208	—	9.0	—
10	6,054 × 6,821	100.0	1.3	0.0
20	6,054 × 3,410	11.4	0.7	0.0
30	6,054 × 2,274	29.1	0.3	1.4
50	6,054 × 1,364	37.7	0.2	1.1
100	6,054 × 682	53.7	0.2	2.0

Table 6. Mean Timing and Suboptimality Results for Objective Sweep, Column-Clustered Prostate Case, 207 Warm Start Solves, GPU

Compression	Dimensions	Average gap (%)	Average true error (%)[1]	Average primal solve time (s)	Average dual solve time (s)
(Collapsed)	6,054 × 68,208	—	—	2.0	—
20	6,054 × 3,410	26.4	21.1	0.07	1.4
30	6,054 × 2,274	28.2	9.9	0.10	1.7
50	6,054 × 1,364	40.5	15.2	0.07	0.3

Note. Percent true error $= 100 \cdot (p^{\text{ub}} - p^\star)/p^{\text{ub}}$, $p^\star$ is solution obtained in Section 5.1.

of the full primal problem, with the added stipulation that intensities assigned to the same cluster be equal), we remark that $f^*(v_{\mathscr{C}}^\star) \le p^\star$ need not hold because we have no guarantee that $v_{\mathscr{C}}^\star$ be feasible on the full problem. In practice, our bounding procedure produces a feasible dual variable $\hat{v}$ when the constraint $A^T v \ge 0$ is enforced to a sufficiently tight numerical tolerance; choosing this tolerance turns out to influence the quality of the lower bound we obtain. As we tighten the tolerance, the lower bound obtained by evaluating $f^*(\hat{v})$ decreases, and can even become negative, which results in the failure we described above. If the tolerance is too wide, then $A^T\hat{v} \ge -\epsilon$ can be satisfied even by the output of the clustered problem, $v_{\mathscr{C}}^\star$, and yield an invalid lower bound. As a rule of thumb, we take the tolerance achieved for the constraint $A_{\mathscr{C}}^T v \ge 0$ when solving the clustered primal problem and require that our bounding procedure produce a dual variable that satisfies $A^T v \ge 0$ to this same tolerance. In the future, we may be able to obtain tighter suboptimality bounds by more judicious choices of tolerances for our bounding procedure.

It appears that the bounding procedure takes relatively longer on the smaller cases; this probably corresponds to the smaller problems being deeper relaxations than the less compressed instances and yielding points $v_{\mathscr{C}}^\star$ that are further from being feasible on the full dual. This trend is corroborated by the data from the objective sweep (warm start) solves, summarized in Table 6. Although the warm-start solve times for the primal clustered problems are *extremely* fast (again, offering the potential for real-time interactive planning or rapid, dense population of MCO plan libraries), the effort required to bound the solution is comparable to that needed for a warm-start solution of the unclustered, voxel-collapsed problem. The true errors are also 10%–20% on average, so the loose bounds are not overly pessimistic.

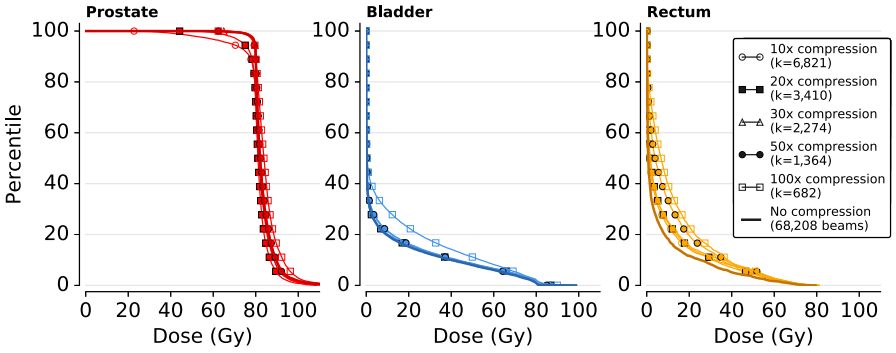
However, when we examine the DVHs generated by the beam-clustered plans, shown in Figure 2, we observe that the plans are highly dosimetrically similar to the uncompressed plan, despite the loose suboptimality bounds.

6. Summary

In this paper, we have presented theory and examples for three methods for approximately (or exactly, in the case of voxel collapse) solving treatment planning problems at significantly reduced computational cost. In addition to the gains realized by using the presented cluster and bound methods, a significant portion of the planning speed is due to the highly parallelized implementation of ADMM implemented by the POGS solver, which is available as a tool based on our choice of fully separable convex planning objectives.

Given our observations (Dong et al. 2016) that linear penalties on OAR structures can produce clinically satisfactory plans (in tandem with piecewise linear penalties on target structures), the voxel collapse method is an obvious win, providing at least an order-of-magnitude

Figure 2. (Color online) Dose Volume Histograms for Prostate Treatment Plans Generated Using Beam Clustering



Notes. Results are shown for three levels of k-means compression applied to the beams, as well as the uncompressed plan (solid line) shown for reference. Default objective weights were used to generate each solution.

speedup on CPU or GPU. These gains compound with another order-of-magnitude (or greater) speedup obtained by the clustering methods to yield planning speeds that would be sufficiently fast for a real-time, interactive planning environment.

Although the voxel collapse method effectively restricts planning to the use of mean dose penalties on OAR structures, a promising option would be to use this technique to rapidly form many plans on the Pareto surface; linear combinations of these plans (based on the full voxel content) could then be optimized to satisfy more complex constraints—for example, dose volume constraints.

The row (voxel) version of the cluster and bound method produces fairly tight bounds at essentially no added computational cost, so this tool could work well to accelerate cases where finer dose grid resolution is desired or could be used to compensate for the enlarged column dimension in cases with many beams.

Because the bounds achieved in the column (beam) version of the cluster and bound method were not particularly tight, we will look to improve the bounding procedure as well as the initialization of the clustering process, because k-means is nonconvex and sensitive to the choice of initialization.

Extrapolating from the very rapid performance we observed on the 3-GB prostate case, we estimate that for dose matrices that can fit on a single GPU (i.e., smaller than 12 GB, currently), each planning run will cost no more than a few seconds, and often less than a second.

Because we have shown that for dose matrices that can fit on a single GPU, each planning run is on the order of tens of milliseconds to (low) tens of seconds or less, and that dose matrices can be compressed quite significantly while still yielding reasonable plans, we can now perform efficient planning with cases with dose matrices that are an order of magnitude too large to fit on a GPU when represented in their entirety, as illustrated by the prostate case.

Although this work does not address beam deliverability, regularization terms (such as total variation penalty on beamlet or aperture intensities) can be added while maintaining a separable formulation compatible with the POGS solver. Similarly, with the incorporation of intensity-sparsifying objectives, our work on large-scale intensity optimization could be of some use in the problem of beam angle selection: some of the computation in the geometric setup phase could be deferred to the intensity optimization phase by selecting an overcomplete set of radiation sources and allowing this pool to be narrowed while intensities are optimized.

Furthermore, we imagine that a robust approach going forward would be to optimize over tens of thousands of apertures (i.e., for nonuniform arc therapy, station parameter optimized radiation therapy, or 4π planning) in lieu of the same number of beamlets. If we can efficiently handle large-scale planning problems, it mitigates the need for apertures to be carefully chosen; instead, a very large number of reasonable apertures can be generated through some heuristic (e.g., one statistically learned from previous treatment plans), and the active apertures can be sparsified during planning.

Acknowledgments

The authors thank Michael Folkerts for providing the anonymized data set for the head and neck VMAT reweighting case and Peng Dong for the anonymized data set for the prostate IMRT case. The authors thank their anonymous reviewers for their extremely helpful feedback that improved the technical and clinical aspects of this paper.

References

Ahmed S, Gozbasi O, Savelsbergh M, Crocker I, Fox T, Schreibmann E (2010) An automated intensity-modulated radiation therapy planning system. *INFORMS J. Comput.* 22(4):568–583.

Aleman DM, Romeijn HE, Dempsey JF (2008) A response surface approach to beam orientation optimization in intensity-modulated radiation therapy treatment planning. *INFORMS J. Comput.* 21(1):62–76.

Aleman DM, Glaser D, Romeijn HE, Dempsey JF (2010) Interior point algorithms: Guaranteed optimality for fluence map optimization in IMRT. *Phys. Medicine Biol.* 55(18):5467–5482.

Appenzoller LM, Michalski JM, Thorstad WL, Mutic S, Moore KL (2012) Predicting dose-volume histograms for organs-at-risk in IMRT planning. *Medical Phys.* 39(12):7446–7461.

Baatar D, Boland N, Johnston R, Hamacher HW (2009) A new sequential extraction heuristic for optimizing the delivery of cancer radiation treatment using multileaf collimators. *INFORMS J. Comput.* 21(2): 224–241.

Banger M, Oelfke U (2010) Spherical cluster analysis for beam angle optimization in intensity-modulated radiation therapy treatment planning. *Phys. Medicine Biol.* 55(19):6023–6037.

Bentzen SM, Constine LS, Deasy JO, Eisbruch A, Jackson A, Marks LB, Haken T, Randall K, Yorke ED (2010) Quantitative analyses of normal tissue effects in the clinic (QUANTEC): An introduction to the scientific issues. *Internat. J. Radiation Oncology Biol. Phys.* 76(3):S3–S9.

Bertsimas D, Nohadani O, Teo KM (2010) Nonconvex robust optimization for problems with constraints. *INFORMS J. Comput.* 22(1):44–58.

Bezanson J, Edelman A, Karpinski S, Shah VB (2014) Julia: A fresh approach to numerical computing. *SIAM Rev.* 59(1):65–98.

Bokrantz R, Forsgren A (2013) An algorithm for approximating convex pareto surfaces based on dual techniques. *INFORMS J. Comput.* 25(2):377–292.

Boutilier JJ, Craig T, Sharpe MB, Chan TCY (2016) Sample size requirements for knowledge-based treatment planning. *Medical Phys.* 43(3):1212–1221.

Boutsidis C, Zouzias A, Mahoney MW, Drineas P (2015) Randomized dimensionality reduction for$k$-means clustering. *IEEE Trans. Inform. Theory* 61(2):1045–1062.

Boyd S, Vandenberghe L (2004) *Convex Optimization* (Cambridge University Press, Cambridge, UK).

Chan TCY, Craig T, Lee T, Sharpe MB (2014) Generalized inverse multiobjective optimization with application to cancer therapy. *Oper. Res.* 62(3):680–695.

Chen W, Craft D, Madden TM, Zhang K, Kooy HM, Herman GT (2010) A fast optimization algorithm for multicriteria intensity modulated proton therapy planning. *Medical Phys.* 37(9): 4938–4945.

Craft D (2007) Local beam angle optimization with linear programming and gradient search. *Phys. Medicine Biol.* 52(7):N127–N135.

Craft D, Bortfeld TR (2008) How many plans are needed in an IMRT multi-objective plan database? *Phys. Medicine Biol.* 53(11): 2785–2796.

Craft D, Halabi TF, Shih HA, Bortfeld TR (2006) Approximating convex Pareto surfaces in multiobjective radiotherapy planning. *Medical Phys.* 33(9):3399–3407.

Craft D, Hong TS, Shih HA, Bortfeld TR (2012a) Improved planning time and plan quality through multicriteria optimization for intensity-modulated radiotherapy. *Internat. J. Radiation Oncology Biol. Phys.* 82(1):e83–e90.

Craft D, McQuaid D, Wala J, Chen W, Salari E, Bortfeld T (2012b) Multicriteria VMAT optimization. *Medical Phys.* 39(2):686–696.

Dong P, Ungun B, Boyd S, Xing L (2016) Optimization of rotational arc station parameter optimized radiation therapy. *Medical Phys.* 43(9):4973–4982.

Dong P, Lee P, Ruan D, Long T, Romeijn HE, Yang Y, Low D, Kupelian P, Sheng K (2013a) 4πnon-coplanar liver SBRT: A novel delivery technique. *Internat. J. Radiation Oncology Biol. Phys.* 85(5):1360–1366.

Dong P, Lee P, Ruan D, Long T, Romeijn E, Low DA, Kupelian P, Abraham J, Yang Y, Sheng K (2013b) 4πnoncoplanar stereotactic body radiation therapy for centrally located or larger lung tumors. *Internat. J. Radiation Oncology Biol. Phys.* 86(3):407–413.

Dongarra JJ, Du Croz J, Hammarling S, Duff IS (1990) A set of level 3 basic linear algebra subprograms. *ACM Trans. Math. Software* 16(1):1–17.

Dongarra JJ, Du Croz J, Hammarling S, Hanson RJ (1988) An extended set of fortran basic linear algebra subprograms. *ACM Trans. Math. Software* 14(1):1–17.

Ernst AT, Mak VH, Mason LR (2009) An exact method for the minimum cardinality problem in the treatment planning of intensity-modulated radiotherapy. *INFORMS J. Comput.* 21(4):562–574.

Fougner C, Boyd S (2018) Parameter selection and pre-conditioning for a graph form solver. Tempo R, Yurkovich S, Misra P, eds. *Emerging Applications of Control and System Theory*, Lecture Notes in Control and Information Sciences (Springer, Cham, Switzerland), 41–62.

Ganage SA, Thool RC, Basit HA (2013) Heterogeneous computing based k-means clustering using Hadoop-MapReduce framework. *Internat. J. Adv. Res. Comput. Sci. Software Engrg.* 3(6):585–592.

Jain AK, Murty MN, Flynn PJ (1999) Data clustering: A review. *ACM Comput. Surveys* 31(3):264–323.

Kessler ML, Mcshan DL, Epelman MA, Vineberg KA, Eisbruch A, Lawrence TS, Fraass BA (2005) Costlets: A generalized approach to cost functions for automated optimization of IMRT treatment plans. *Optim. Engrg.* 6(4):421–448.

Küfer KH, Monz M, Scherrer A, Süss P (2009) Multicriteria optimization in intensity modulated radiotherapy planning. Pardalos PM, Romeijn HE, eds. *Handbook of Optimization in Medicine* (Springer, New York), 1–45.

Lawson CL, Hanson RJ, Kincaid DR, Krogh FT (1979) Basic linear algebra subprograms for fortran usage. *ACM Trans. Math. Software* 5(3):308–323.

Lee T, Hammad M, Chan TCY, Craig T, Sharpe MB (2013) Predicting objective function weights from patient anatomy in prostate IMRT treatment planning. *Medical Phys.* 40(12):121706.

Li R, Xing L (2013) An adaptive planning strategy for station parameter optimized radiation therapy (SPORT): Segmentally boosted VMAT. *Medical Phys.* 40(5):050701.

Li R, Xing L, Horst KC, Bush K (2014) Nonisocentric treatment strategy for breast radiation therapy: A proof of concept study. *Internat. J. Radiation Oncology Biol. Phys.* 88(4):920–926.

Li N, Zarepisheh M, Uribe-Sanchez A, Moore K, Tian Z, Zhen X, Graves YJ, et al. (2013) Automatic treatment plan re-optimization for adaptive radiotherapy guided with the initial plan DVHs. *Phys. Medicine Biol.* 58(24):8725–8738.

Lim GJ, Holder A, Reese J (2009) A clustering approach for optimizing beam angles in IMRT planning. *Mathematical Sciences Technical Reports (MSTR)* 14. Accessed November 19, 2018, https://scholar.rose-hulman.edu/math_mstr/14.

Lim GJ, Ferris MC, Wright SJ, Shepard DM, Earl MA (2007) An optimization framework for conformal radiation treatment planning. *INFORMS J. Comput.* 19(3):366–380.

Lloyd S (1982) Least squares quantization in PCM. *IEEE Trans. Inform. Theory* 28(2):129–137.

Lu R, Radke RJ, Yang J, Happersett L, Yorke E, Jackson A (2008) Reduced-order constrained optimization in IMRT planning. *Phys. Medicine Biol.* 53(23):6749–6766.

Martin BC, Bortfeld TR, Castañon DA (2007) Accelerating IMRT optimization by voxel sampling. *Phys. Medicine Biol.* 52(24): 7211–7228.

Moore KL, Appenzoller LM, Tan J, Michalski JM, Thorstad WL, Mutic S (2014) Clinical implementation of dose-volume histogram predictions for organs-at-risk in IMRT planning. *J. Phys. Conf. Ser.* 489(1):012055.

Otto K (2014) Real-time interactive treatment planning. *Phys. Medicine Biol.* 59(17):4845–4859.

Parikh N, Boyd S (2014) Block splitting for distributed optimization. *Math. Programming Comput.* 6(1):77–102.

Peng F, Jia X, Gu X, Epelman MA, Romeijn HE, Jiang SB (2012) A new column-generation-based algorithm for VMAT treatment plan optimization. *Phys. Medicine Biol.* 57(14):4569–4588.

Pugachev A, Xing L (2002) Incorporating prior knowledge into beam orientation optimization in IMRT. *Internat. J. Radiation Oncology Biol. Phys.* 54(5):1565–1574.

Rennen G, van Dam ER, den Hertog D (2013) Enhancement of sandwich algorithms for approximating higher-dimensional convex pareto sets. *INFORMS J. Comput.* 23(4):493–517.

Romeijn HE, Ahuja RK, Dempsey JF, Kumar A, Li JG (2003) A novel linear programming approach to fluence map optimization for intensity modulated radiation therapy treatment planning. *Phys. Medicine Biol.* 48(21):3521–3542.

Scherrer A, Küfer KH, Bortfeld T, Monz M, Alonso F (2005) IMRT planning on adaptive volume structures—a decisive reduction in computational complexity. *Phys. Medicine Biol.* 50(9):2033–2053.

Sculley D (2010) Web-scale k-means clustering. *WWW '10: Proc. 19th Internat. Conf. World Wide Web* (Association for Computing Machinery, New York), 1177–1178.

Siem AYD, den Hertog D, Hoffmann AL (2011) A method for approximating univariate convex functions using only function value evaluations. *INFORMS J. Comput.* 23(4):591–604.

Tropp JA (2004) Topics in sparse approximation. PhD thesis, University of Texas at Austin, Austin.

Udell M, Horn C, Zadeh R, Boyd S (2016) Generalized low rank models. *Foundations Trends Mach. Learn.* 9(1):1–118.

Xu R, Wunsch DC (2005) Survey of clustering algorithms. *IEEE Trans. Neural Networks* 16(3):645–678.

Xu R, Wunsch DC (2010) Clustering algorithms in biomedical research: A review. *IEEE Rev. Biomedicine Engrg.* 3:120–154.

Zarepisheh M, Ye Y, Boyd S, Li R, Xing L (2014a) SU-E-T-295: Simultaneous beam sampling and aperture shape optimization for station parameter optimized radiation therapy (SPORT). *Medical Phys.* 41(6):292.

Zarepisheh M, Long T, Li N, Tian Z, Romeijn HE, Jia X, Jiang SB (2014b) A DVH-guided IMRT optimization algorithm for automatic treatment planning and adaptive radiotherapy replanning. *Medical Phys.* 41(6):061711.

Zhang HH, Shi L, Meyer R, Nazareth D, D'Souza W (1999) Solving beam-angle selection and dose optimization simultaneously via high-throughput computing. *INFORMS J. Comput.* 21(3): 427–444.