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

Barıs̡ Ungun, Lei Xing, and Stephen Boyd

June 6, 2018

Abstract

Radiation therapy is widely used in cancer treatment; however, plans necessarily involve tradeoffs between tumor coverage and mitigating damage to healthy tissue. While 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 (a) yield higher quality plans by optimizing over much (*i.e.*, 100-fold) larger sets of candidate beams, and/or (b) 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 10–200-fold speedup.

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 non-target tissue, and achieving short treatment times. Treatments may be planned for delivery in several sessions, or fractions, and the planing 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 [CMW+12].

Even for optimistic scenarios in which 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 [Cra07]. 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 beams 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 obectives, and (3) generating a sequence of hardware parameters that (approximately) delivers the optimized intensities from the specified locations. Since candidate beams are an abstraction, the exact sense depends on the radiation type (*e.g.*, electron, photon, proton) and delivery modality under consideration (*e.g.*, IMRT, VMAT, 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 [CMW+12]); a pencil beam in 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 of candidate beams.

The large body of work on planning algorithms spans both convex formulations, such as FMO problems used in IMRT planning [AGRD10, RAD+03] which are paired with a set of small mixed integer programs to decompose fluence maps into deliverable apertures [EMM09, BBJH09], and nonconvex formulations, such as direct machine parameter optimization used for VMAT planning [PJG+12] or robust optimization of beam angles and intensities [BNT10]. 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 [LXHB14, LX13, ZYB+14, DLR+13b].

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 multi-objective 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 [CCLS14, CB08, KMSS09].

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 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, since 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 §2 we examine previous work related to solving large-scale intensity optimization problems in treatment planning. In §3 we will introduce the class of convex treatment planning problems compatible with the methods detailed in this work, as well as their associated dual problems. In §4 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 §5 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 §6 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.

2 Related Work

2.0.1 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 [KMSS09]) instead of optimizing over the intensities of pre-determined apertures or beamlets, or using integer descision 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 [PJG$^+$12, ZYB$^+$14].

The efforts in beam angle optimization [ZSM$^+$99, Cra07, LFW$^+$07, ARD08, AGS$^+$10], non-uniform arc therapy [LX13, ZYB$^+$14], non-coplanar planning [DLR$^+$13a, DLR$^+$13b], and optimized isocenter selection [LXHB14] 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, the authors of [RAD$^+$03] and [KME$^+$05] 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 [PB14], the authors describe how intensity optimization problems in treatment planning can be phrased as graph form problems so as to benefit from highly parallel optimiza-

tion algorithms; the freely available open-source POGS solver that implements graph form ADMM, which we use in this work, is described in [FB15].

2.0.2 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 planing 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 multi-objective optimization [BV04]. Nevertheless, even when it is straightforward to find an optimal plan for a given clinical preference, the task of solving multiple optimization problems remains, since tradeoffs between optimal plans can only be resolved by the planner's clinical judgment.

Major research efforts in this area include multi-criterion optimization (MCO) approaches that generate libraries of optimal points along the Pareto surface and methods to approximate this surface [CHSB06, CCM$^+$10, SdHH11, CMW$^+$12, CHSB12, BF13, RvDdH13], approximation of the Pareto surface, automated planning and Pareto surface navigation [ZLL$^+$14, LZUS$^+$13], and statistical learning (from previously planned cases) of clinical preferences and anatomically driven predictions of feasible designs, or even favorable beam directions [LHC$^+$13, AMT$^+$12, MAT$^+$14, BCSC16, PX02].

Another approach discussed by [Ott14] 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.

2.0.3 Voxel Clustering

Reducing the dose grid resolution is commonly done to lower the cost of dose calculations and plan optimization, but studies such as [MBC07, SKB$^+$05] 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 §2.0.1 and §2.0.2: 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.

2.0.4 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 [PJG+12]. [LRY+08] 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 [BO10], 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 inter-beam similarities are used to cluster and select beam angles in [LHR09].

Both the beam and voxel clustering approaches are special cases of nonnegative matrix factorization; the broader class of NNMF algorithms could be pertinent here since they would preserve the physical sense of the entries of the dose matrix. See, for example, [UHZB16] for a detailed discussion of a broad class of low rank approximation methods, or [Tro04] for random matrix algorithms used for dimensionality reduction in optimization.

3 Convex Treatment Planning

3.1 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 CVaR penalties; we refer the reader to [RAD+03, KME+05] 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 non-target 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 $i, i' \in S_s$.

3.1.1 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

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

Since the objective f is a weighted sum of objectives concerning each structure, the problem (1) can be interpreted as a linear scalarization of a multi-objective 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.

3.2 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 $\nu \in \mathbf{dom}(f^*)$.

3.2.1 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 §3.1.) 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 $(\hat{x}, \hat{y})$, 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 $(\hat{x}, \hat{y})$ as a solution to (1) is at most

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

3.3 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:

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

This choice of f yields the dual problem

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

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

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

3.4 Large-Scale Treatment Planning

Since 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 which, despite the use of modern hardware and the fastest available optimization methods, is too large to be solved in a clinically acceptable timeframe—*e.g.*, a plan in which we consider tens of thousands of candidate beams and a dose grid of several hundred thousand voxels. While 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 §4 and §5, 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.

4 Voxel Clustering

4.1 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 up-sampling 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, while 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 non-zero entry per row. For each cluster κ, the corresponding set $C_\kappa = \{i \mid \text{voxel } i \in \text{cluster } \kappa\}$ 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, *i.e.*, $\omega_\kappa = |C_\kappa|$. 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 $A_\mathcal{R}$,

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

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

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

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.

4.2 Bounding Procedure

Given a primal optimal point $(x_\mathcal{R}^\star, y_\mathcal{R}^\star)$ for which the voxel-clustered problem attains its optimal value $p_\mathcal{R}^\star$, we seek upper and lower bounds on $p^\star$. To obtain an upper bound, we

set $\hat{x} = x^\star_{\mathcal{R}}$. Since $x^\star_{\mathcal{R}} \geq 0$, $\hat{x}$ is feasible for (1). We define $\hat{y} = Ax^\star_{\mathcal{R}}$, 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

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

4.3 Voxel Collapse for Non-Target Structures

We present a special case of voxel clustering that applies to choices of f that impose linear penalties on non-target 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 non-target voxels, as noted in §3.3, the objective contribution of non-target voxels is simply the linear term $w_i y_i$. Since $w_i = w_{i'} = w_s$ for $i, i' \in S_s$, with trivial rearrangement, the objective contribution of non-target 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 non-target 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,

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

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 non-target 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 non-target structures. Of course, most cases have many more target voxels than non-target 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 while voxel collapse should be applied to non-target structures.

4.4 Clustering Procedure

In §4.1, the cluster assignments represented by the matrix U were assumed to be given, and we understood $UA_{\mathcal{R}} \approx A$ to hold without specifying the sense in which the product $UA_{\mathcal{R}}$ was to approximate A.

Several choices 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 $|A - UA_{\mathcal{R}}|_2^2$, subject to the constraint that each row of U must contain exactly one nonzero entry with value 1.) The problem is NP-hard and sensitive to the initial choice of centroids; a commonly used heuristic is Lloyd's algorithm, which is summarized in Algorithm 4.1.

Algorithm 4.1 *Lloyd's algorithm for k-means [Llo82].*

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

   $$u_i = \operatorname*{argmin}_\kappa |p_i - c_\kappa|_2^2$$

   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 [JMF99, XW05, XW10].

The clustering method used may jointly generate cluster assignments, given by U, and cluster prototypes, given by $A_{\mathcal{R}}$. Alternatively, the method may simply provide cluster assignments (*e.g.*, by grouping voxels in regularly sized clusters based on geometric adjacency), leaving the question of how to construct the rows of $A_{\mathcal{R}}$. Given the cluster assignments encoded in U, one may 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 [MBC07].

While the choice of clustering method (including 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 treatment planning structure, and implement a vectorized version of Lloyd's algorithm as described in Algorithm 4.2.

Algorithm 4.2 *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 initalized point to cluster assignment matrix $U \in \{0,1\}^{m\times k}$, with entries $u_{i\kappa} = 1$ if point i assigned to cluster κ, and 0 otherwise.

repeat

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

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

$$u_{i\kappa} = \begin{cases} 1 & \kappa = \underset{\kappa'}{\operatorname{argmin}}\{d_{i\kappa'}\} \\ 0 & \text{otherwise} \end{cases}, \quad i = 1,\dots q, \quad \kappa = 1,\dots,k.$$

until assignments stable or an iteration limit is reached.

return assignments U, centroids C.

(Note the minor modification to Lloyd's algorithm implied by the update rule for D: since 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 $p_i^Tp_i$ 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 $A_{\mathcal{R}}$ introduced above. To perform voxel clustering by structure, we make the subtitution $P = A_s \in \mathbf{R}^{m_s\times n}$ for each structure-specific sub-matrix A_s, and choose a proportionally smaller k (specifically, $k_s = k \cdot m_s/m$). The centroid matrices C_s

returned for each structure s would then be vertically concatenated to form $A_{\mathcal{R}}$. 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.

5 Beam Clustering

5.1 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 $A_{\mathcal{C}} \in \mathbf{R}^{m\times k}$, and an up-sampling matrix $V \in \mathbf{R}^{n\times k}$,

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

As with the up-sampling matrix U defined in (4), 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 $A_{\mathcal{C}}$ can be constructed as:

$$A_{\mathcal{C}} = AV(V^TV)^{-1}.$$

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

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

with optimization variables $x_{\mathcal{C}} \in \mathbf{R}^k$, $y_{\mathcal{C}} \in \mathbf{R}^m$. 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

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

i.e., 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 $p^\star_{\mathcal{C}}$, and its dual problem has the form

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

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

5.2 Bounding Procedure

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

Since we have $x^\star_{\mathcal{C}} \geq 0$, choosing $\hat{x} = V(V^TV)^{-1}x^\star_{\mathcal{C}}$ and $\hat{y} = A\hat{x} = A_{\mathcal{C}}x^\star_{\mathcal{C}}$ 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 $x^\star_{\mathcal{C}_\kappa}$ assigned to beam cluster κ among its constituent beams.) Thus, an upper bound to the value of (1) is given by

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

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

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

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

and ignore the remaining columns of A since 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 $|\mathcal{I}| \leq k$, 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

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

Since 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 5.1.

Algorithm 5.1 *Beam clustering lower bound.*

given an initial point $\nu_{\mathcal{C}}^{\star}$ optimal for (9).
$\nu^{(0)} := \nu_{\mathcal{C}}^{\star}$.
$t := 1$.
repeat
1. *Fix dimension.* Form $\mathcal{I} = \{a_j \mid a_j^T \nu^{(t-1)} < 0\}$. Then, $k^{(t)} := \min(k, |\mathcal{I}|)$.
2. *Approximate.* **repeat**
   $\hat{a}_\kappa = \operatorname{argmin}\{a_j^T \nu^{(t-1)} \mid a_j \in \mathcal{I}\}$
   $\mathcal{I} := \mathcal{I} \setminus \{\hat{a}_\kappa\}$
   for $\kappa = 1, \ldots, k^{(t)}$.
3. *Solve.* Set the value of $\delta^{\star(t)}$ to a solution of the convex problem
   $\underset{\delta}{\text{minimize}} \quad f^*(\nu^{(t)} + \delta)$
   $\text{subject to} \quad \hat{A}_{\mathcal{C}}^{(t)T}(\nu^{(t-1)} + \delta) \geq 0.$
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 $-f^*(\nu^{(t)})$,

---

6 Examples

6.1 Voxel Collapse

The submatrices of A corresponding to non-target 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.

6.1.1 Small Problem Instance

A 268228 voxel, 360 aperture VMAT head and neck case was used for re-optimization 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 fourteen 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 §3.3. 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 non-target 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 QUANTEC reference guidelines [BCD$^+$10] 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 collapsed the non-target structures for this case, yielding a dose matrix of 11253 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 §4.3, so that the objective value coincides exactly with that of (4).

Additionally, we performed 61 warm-start trials by re-optimizing for different objective weights while using optimal variables generated from the previous solve to initialize the subsequent run. 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.

6.1.2 Large Problem Instance

We also planned a much larger case, a 589467 voxel by 68208 beamlet prostate FMO problem. This matrix contains 865 million nonzero 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 CPU, and the matrix did not fit on a single GPU.

Collapsing the non-target structures yielded a dose matrix of 6054 voxels and 68208

Table 1 Timing results for voxel collapse, CPU.

Case	State	Dimensions	Setup time(s)	Solve time(s)	Iterations
HN	Full	(268228×360)	28.1	267.4	3117
HN	Collapsed	(11253×360)	0.4	1.0	203
Prostate	Full	(589467×68208)	1500	*	**
Prostate	Collapsed	(6054×68208)	260.0	190.1	258

*, **: unconverged after 7 hours, 16 iterations

beamlets, or 3.1GB when stored as a dense matrix. Planning was performed using the default weights introduced above; no modifications were needed to meet QUANTEC dosing guidelines.

In addition, we iteratively re-planned 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.

6.1.3 Computational Details

Optimizations were performed in the Python interface to POGS [FB15], which calls a C or CUDA solver. The CPU version is 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.

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 includes 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 18.5s (4.9–128.3s) on the CPU, and 1.1s (0.2–6.3s) on the GPU.

Table 2 Timing results for voxel collapse, GPU.

Case	State	Dimensions	Setup time(s)	Solve time(s)	Iterations
HN	Full	(268228×360)	7.8	15.5	1210
HN	Collapsed	(11253×360)	1.7	0.3	203
Prostate	Full	(589467×68208)	*	**	***
Prostate	Collapsed	(6054×68208)	15.6	8.9	258

*, **, ***: case does not fit on a single nVidia TitanX GPU

6.2 Voxel Clustering

6.2.1 Clustering

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

While we implement a naive version of the k-means clustering as described in Algorithm 4.2, accelerated variants exist, such as mini-batch k-means [Scu10, GTB13, BZMD15, SKB+05].

6.2.2 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 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 up-sampling 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.

6.2.3 Problem Instance

Voxel clustering was performed on the dose matrix for the head and neck case introduced in §6.1. Clustering was performed to approximately 10-, 20-, 30-, 50-, and 100-fold compression levels, yielding compressed matrices of sizes (1036×360), (534×360), (364×360), (221×360), and (111×360).

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)	(11253×360)	-	1.25	0.55
10	(1036×360)	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

For each instance, we solved (6) using the piecewise linear objective discussed in §3.3, using our default objective weights (§6.1), with the modification that the weights for each clustered (or collapsed) metavoxel are multiplied by the number of elements in its cluster (or structure) so that the objective value of (6) is scaled to match that of (1). The objective weight sweep carried out in §6.1 to generate plans sampling the Pareto surface was repeated for the clustered problem instances. (Since these plans were solutions to approximations of the full problem, we were in fact sampling the feasible region near the Pareto surface.)

6.2.4 Computational Details

Clustering was performed on a 32-core, 2.20GHz Intel Xeon CPU E5-4620 processor in Julia [BEKS14], with point-to-cluster distance calculations and comparisons vectorized as described in Algorithm 4.2 and cast as BLAS operations [LHKK79, DDCHH88, DDCHD90]; we used the Julia language-default of 8 parallel threads for BLAS operations.

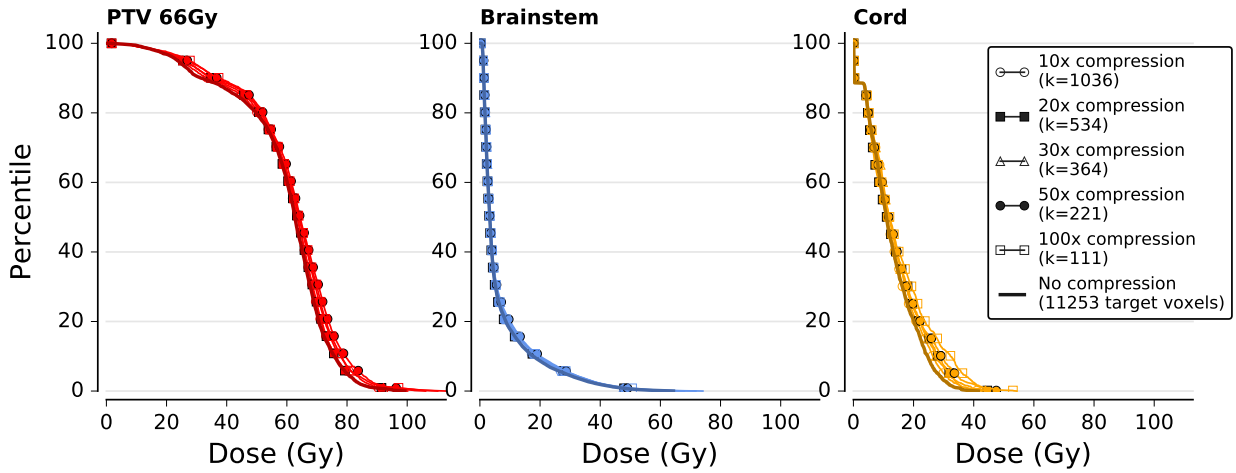
6.2.5 Results

Results for the voxel clustering approximations are summarized in Table 3. Encouragingly, the largest suboptimality bound was 6% for the 100-fold compressed approximation, while 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 acheived 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 0.05s at maximum compression to 0.22 at minimum compression on the CPU, representing a 6–25-fold speedup. Solve times on the GPU averaged 0.52s across compressions, or about the same as the non-clustered, collapsed problem—the

Figure 1: Dose volume histograms for head and neck treatment plans generated using voxel clustering.



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 non-target structures in all plans. The same objective weights were used to generate each solution.

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.

6.3 Beam Clustering

6.3.1 Clustering

Vectors corresponding to the columns of $A_{\text{collapsed}}$ (as defined in §6.1) were clustered into k column clusters (*i.e.*, aggregate beams) 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. Since 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.

Table 4 Timing and suboptimality results for beam clustering, CPU.

Compression	Dimensions	Suboptimality bound (%)	Primal solve time (s)	Dual solve time (s)
(collapsed)	(6054×68208)	-	121.6	–
10	(6054 × 6821)	100.0	15.2	0.0
20	(6054 × 3410)	10.8	6.7	0.0
30	(6054 × 2274)	29.2	1.9	3.0
50	(6054 × 1364)	37.7	0.7	4.3
100	(6054 × 682)	53.7	0.2	1.7

Table 5 Timing and suboptimality results for beam clustering, GPU.

Compression	Dimensions	Suboptimality bound (%)	Primal solve time (s)	Dual solve time (s)
(collapsed)	(6054×68208)	-	9.0	–
10	(6054 × 6821)	100.0	1.3	0.0
20	(6054 × 3410)	11.4	0.7	0.0
30	(6054 × 2274)	29.1	0.3	1.4
50	(6054 × 1364)	37.7	0.2	1.1
100	(6054 × 682)	53.7	0.2	2.0

6.3.2 Problem Instance

The prostate case introduced in §6.1 was clustered to approximate compression levels of 10-, 20-, 30-, 50-, and 100-fold compression, yielding clustered matrices sized (6054 × 6821), (6054 × 3410), (6054 × 2274), (6054 × 1364), and (6054 × 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.

6.3.3 Results

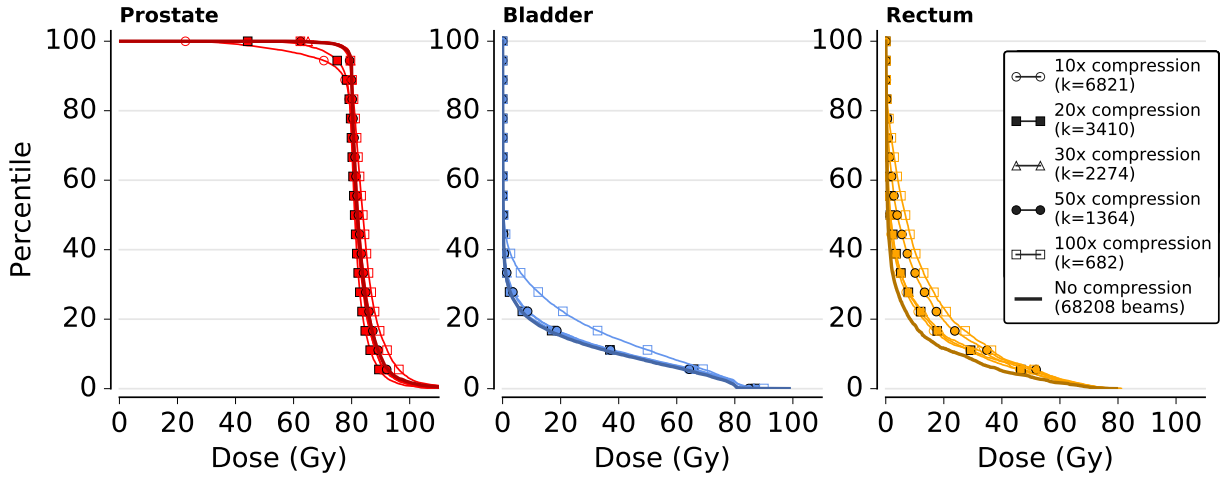
For cold-start problems, in addition to the unmeasured speedup obtained through voxel collapse of non-target structures, beam clustering resulted in 8–64-fold speed gains on the CPU to solve the clustered problem and find a bound, or 8–600-fold speedups for primal solves alone. On the GPU, we obtained smaller speedups of 4–7-fold when considering the time for both the primal solve and the bounding procedure, or 6–45-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 (since we know our objective value to be nonnegative) and consequently obtained a suboptimality bound of 100%; we discuss this failure of our bounding procedure in the sequel. The bounds obtained and timing results are

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)	(6054×68208)	-	-	2.0	-
20	(6054 × 3410)	26.4	21.1	0.07	1.4
30	(6054 × 2274)	28.2	9.9	0.10	1.7
50	(6054 × 1364)	40.5	15.2	0.07	0.3

[1]% true error= $100 \cdot (p^{\text{ub}} - p^\star)/p^{\text{ub}}$, $p^\star$ is solution obtained in §6.1.

Figure 2: Dose volume histograms for prostate treatment plans generated using beam clustering.



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.

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 of the full primal problem, owing to the added stipulation that intensities assigned to beams in the same cluster be equal), we remark that $f^*(\nu_{\mathcal{C}}^\star) \leq p^\star$ need not hold since we have no guarantee that $\nu_{\mathcal{C}}^\star$ be feasible on the full problem. In practice, our bounding procedure produces a feasible dual variable $\hat{\nu}$ when the constraint $A^T\hat{\nu} \geq 0$ is enforced to a sufficiently tight numerical tolerance, *i.e.*, $A^T\hat{\nu} \geq -\epsilon$ for some $\epsilon > 0$; choosing this tolerance ϵ turns out to influence the quality of the lower bound we obtain. As we tighten the tolerance, the lower bound obtained by evaluating $f^*(\hat{\nu})$ decreases, and can even become negative, which results in the failure we described above. If the tolerance is too large, then $A^T\hat{\nu} \geq -\epsilon$ can be satisfied even by the output of the clustered problem, $\nu_{\mathcal{C}}^\star$, and yield an invalid lower bound. As a heuristic, we take the largest numerical violation of the nonnegativity constraint achieved when solving (9) as the numerical tolerance to require for the dual variable produced by our bounding procedure with respect to the nonnegativity constraint in (2); *i.e.*, we set $\epsilon = \left|\min_j \left(A_{\mathcal{C}}^T \nu_{\mathcal{C}}^\star\right)_j\right|$. In the future, we may be able to obtain tighter suboptimality bounds by more judicious choices of tolerances for our bounding procedure.

There does not appear to be a clear relationship between problem compression level and dual solve time, as summarized in Table 6. While the warm start solve times for the primal clustered problems are *extremely* fast at all compression levels, (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.

7 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 [DUBX16] 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 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.

While 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, *e.g.*, 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.

Since 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, since k-means is non-convex and sensitive to the choice of initialization.

Extrapolating from the very rapid performance we observe on the 3GB prostate case, we estimate that for dose matrices that can fit on a single GPU (*i.e.*, smaller than 12GB, currently), each planning run will cost no more than a few seconds. Given our observations that dose matrices can be compressed quite significantly and still yield clinically reasonable plans, the techniques presented in this paper bring into reach efficient treatment planning for dose matrices that are an order of magnitude too large to fit on a GPU when represented in their entirety.

While 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 pruned 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 non-uniform arc therapy, SPORT, 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.

8 Acknowledgements

We thank Michael Folkerts for providing the anonymized dataset for the head and neck VMAT reweighting case, and Peng Dong for the anonymized dataset for the prostate IMRT

case. We are deeply grateful to our anonymous reviewers for their extremely helpful feedback that improved the technical and clinical aspects of this paper. This research was made possible by NIH Grant 5R01CA176553 and by the Stanford Bio-X Bowes Graduate Fellowship.

9 References

References

[AGRD10] D. M. Aleman, D. Glaser, H. E. Romeijn, and J. F. Dempsey. Interior point algorithms: guaranteed optimality for fluence map optimization in IMRT. *Physics in Medicine and Biology*, 55(18):5467–5482, 2010.

[AGS+10] S. Ahmed, O. Gozbasi, M. Savelsbergh, I. Crocker, T. Fox, and E. Schreibmann. An automated intensity-modulated radiation therapy planning system. *INFORMS Journal on Computing*, 22(4):568–583, 2010.

[AMT+12] L. M. Appenzoller, J. M. Michalski, W. L. Thorstad, S. Mutic, and K. L. Moore. Predicting dose-volume histograms for organs-at-risk in IMRT planning. *Medical Physics*, 39(12):7446–7461, 2012.

[ARD08] D. M. Aleman, H. E. Romeijn, and J. F. Dempsey. A response surface approach to beam orientation optimization in intensity-modulated radiation therapy treatment planning. *INFORMS Journal on Computing*, 21(1):62–76, 2008.

[BBJH09] D. Baatar, N. Boland, R. Johnston, and H. W. Hamacher. A new sequential extraction heuristic for optimizing the delivery of cancer radiation treatment using multileaf collimators. *INFORMS Journal on Computing*, 21(2):224–241, 2009.

[BCD+10] S. M. Bentzen, L. S. Constine, J. O. Deasy, A. Eisbruch, A. Jackson, L. B. Marks, T. Haken, K. Randall, and E. D. Yorke. Quantitative analyses of normal tissue effects in the clinic (QUANTEC): an introduction to the scientific issues. *International Journal of Radiation Oncology Biology Physics*, 76(3):S3–S9, 2010.

[BCSC16] J. J. Boutilier, T. Craig, M. B. Sharpe, and T. C. Y. Chan. Sample size requirements for knowledge-based treatment planning. *Medical Physics*, 43(3):1212–1221, 2016.

[BEKS14] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *arXiv preprint arXiv:1411.1607*, 2014.

[BF13] R. Bokrantz and A. Forsgren. An algorithm for approximating convex pareto surfaces based on dual techniques. *INFORMS Journal on Computing*, 25(2):377–292, 2013.

[BNT10] D. Bertsimas, O. Nohadani, and K. M. Teo. Nonconvex robust optimization for problems with constraints. *INFORMS Journal on Computing*, 22(1):44–58, 2010.

[BO10] M. Banger and U. Oelfke. Spherical cluster analysis for beam angle optimization in intensity-modulated radiation therapy treatment planning. *Physics in Medicine and Biology*, 55(19):6023–6037, 2010.

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[BZMD15] C. Boutsidis, A. Zouzias, M. W. Mahoney, and P. Drineas. Randomized dimensionality reduction for k-means clustering. *IEEE Transactions on Information Theory*, 61(2):1045–1062, 2015.

[CB08] D. Craft and T. R. Bortfeld. How many plans are needed in an IMRT multi-objective plan database? *Physics in Medicine and Biology*, 53(11):2785–2796, 2008.

[CCLS14] T. C. Y. Chan, T. Craig, T. Lee, and M. B. Sharpe. Generalized inverse multiobjective optimization with application to cancer therapy. *Operations Research*, 62(3):680–695, 2014.

[CCM+10] W. Chen, D. Craft, T. M. Madden, K. Zhang, H. M. Kooy, and G. T. Herman. A fast optimization algorithm for multicriteria intensity modulated proton therapy planning. *Medical Physics*, 37(9):4938–4945, 2010.

[CHSB06] D. Craft, T. F. Halabi, H. A. Shih, and T. R. Bortfeld. Approximating convex Pareto surfaces in multiobjective radiotherapy planning. *Medical Physics*, 33(9):3399–3407, 2006.

[CHSB12] D. Craft, T. S. Hong, H. A. Shih, and T. R. Bortfeld. Improved planning time and plan quality through multicriteria optimization for intensity-modulated radiotherapy. *International Journal of Radiation Oncology Biology Physics*, 82(1):e83–e90, 2012.

[CMW+12] D. Craft, D. McQuaid, J. Wala, W. Chen, E. Salari, and T. Bortfeld. Multicriteria VMAT optimization. *Medical Physics*, 39(2):686–696, 2012.

[Cra07] D. Craft. Local beam angle optimization with linear programming and gradient search. *Physics in Medicine and Biology*, 52(7):N127–N135, 2007.

[DDCHD90] J. J. Dongarra, Jeremy Du Croz, Sven Hammarling, and I. S. Duff. A set of level 3 basic linear algebra subprograms. *ACM Transactions on Mathematical Software*, 16(1):1–17, 1990.

[DDCHH88] Jack J. Dongarra, Jeremy Du Croz, Sven Hammarling, and Richard J. Hanson. An extended set of fortran basic linear algebra subprograms. *ACM Transactions on Mathematical Software*, 14(1):1–17, 1988.

[DLR+13a] P. Dong, P. Lee, D. Ruan, T. Long, E. Romeijn, D. A. Low, P. Kupelian, J. Abraham, Y. Yang, and K. Sheng. 4π noncoplanar stereotactic body radiation therapy for centrally located or larger lung tumors. *International Journal of Radiation Oncology Biology Physics*, 86(3):407–413, 2013.

[DLR+13b] P. Dong, P. Lee, D. Ruan, T. Long, H. E. Romeijn, Y. Yang, D. Low, P. Kupelian, and K. Sheng. 4π non-coplanar liver SBRT: A novel delivery technique. *International Journal of Radiation Oncology Biology Physics*, 85(5):1360–1366, 2013.

[DUBX16] P. Dong, B. Ungun, S. Boyd, and L. Xing. Optimization of rotational arc station parameter optimized radiation therapy. *Medical Physics*, 43(9):4973–4982, 2016.

[EMM09] A. T. Ernst, V. H. Mak, and L. R. Mason. An exact method for the minimum cardinality problem in the treatment planning of intensity-modulated radiotherapy. *INFORMS Journal on Computing*, 21(4):562–574, 2009.

[FB15] C. Fougner and S. Boyd. Parameter selection and pre-conditioning for a graph form solver. *arXiv preprint arXiv:1503.08366*, 2015.

[GTB13] S. A. Ganage, R. C. Thool, and H. A. Basit. Heterogeneous computing based k-means clustering using Hadoop-MapReduce framework. *International Journal of Advanced Research In Computer Science and Software Engineering*, 3(6):585–592, 2013.

[JMF99] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[KME+05] M. L. Kessler, D. L. Mcshan, M. A. Epelman, K. A. Vineberg, A. Eisbruch, T. S. Lawrence, and B. A. Fraass. Costlets: A generalized approach to cost functions for automated optimization of IMRT treatment plans. *Optimization and Engineering*, 6(4), 2005.

[KMSS09] K. H. Küfer, M. Monz, A. Scherrer, and P. Süss. Multicriteria optimization in intensity modulated rediotherapy planning. In P., M. Pardalos, and H. E. Romeijn, editors, *Handbook of optimization in medicine*, pages 1–45. Springer, New York, 2009.

[LFW+07] G. J. Lim, M. C. Ferris, S. J. Wright, D. M. Shepard, and M. A. Earl. An optimization framework for conformal radiation treatment planning. *INFORMS Journal on Computing*, 19(3):366–380, 2007.

[LHC+13] T. Lee, M. Hammad, T. C. Y. Chan, T. Craig, and M. B. Sharpe. Predicting objective function weights from patient anatomy in prostate IMRT treatment planning. *Medical Physics*, 40(12):121706, 2013.

[LHKK79] C. L. Lawson, R. J. Hanson, D. R. Kincaid, and F. T. Krogh. Basic linear algebra subprograms for fortran usage. *ACM Transactions on Mathematical Software*, 5(3):308–323, 1979.

[LHR09] G. J. Lim, A. Holder, and J. Reese. A clustering approach for optimizing beam angles in IMRT planning. In *IIE Annual Conference, Proceedings*, IERC 2009, pages 663–668, 2009.

[Llo82] S. Lloyd. Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[LRY+08] R. Lu, R. J. Radke, J. Yang, L. Happersett, E. Yorke, and A. Jackson. Reduced-order constrained optimization in IMRT planning. *Physics in Medicine and Biology*, 53(23):6749–6766, 2008.

[LX13] R. Li and L. Xing. An adaptive planning strategy for station parameter optimized radiation therapy (SPORT): Segmentally boosted VMAT. *Medical Physics*, 40(5):050701, 2013.

[LXHB14] R. Li, L. Xing, K. C. Horst, and K. Bush. Nonisocentric treatment strategy for breast radiation therapy: a proof of concept study. *International Journal of Radiation Oncology Biology Physics*, 88(4):920–926, 2014.

[LZUS+13] N. Li, M. Zarepisheh, A. Uribe-Sanchez, K. Moore, Z. Tian, X. Zhen, Y. J. Graves, Q. Gautier, L. Mell, L. Zhou, X. Jia, and S. Jiang. Automatic treatment plan re-optimization for adaptive radiotherapy guided with the initial plan DVHs. *Physics in Medicine and Biology*, 58(24):8725–8738, 2013.

[MAT+14] K. L. Moore, L. M. Appenzoller, J. Tan, J. M. Michalski, W. L. Thorstad, and S. Mutic. Clinical implementation of dose-volume histogram predictions for organs-at-risk in IMRT planning, 2014.

[MBC07] B. C. Martin, T. R. Bortfeld, and D. A. Castañon. Accelerating IMRT optimization by voxel sampling. *Physics in Medicine and Biology*, 52(24):7211–7228, 2007.

[Ott14] K. Otto. Real-time interactive treatment planning. *Physics in Medicine and Biology*, 59(17):4845–4859, 2014.

[PB14] N. Parikh and S. Boyd. Block splitting for distributed optimization. *Mathematical Programming Computation*, 6(1):77–102, 2014.

[PJG+12] F. Peng, X. Jia, X. Gu, M. A. Epelman, H. E. Romeijn, and S. B. Jiang. A new column-generation-based algorithm for VMAT treatment plan optimization. *Physics in Medicine and Biology*, 57(14):4569–4588, 2012.

[PX02] A. Pugachev and L. Xing. Incorporating prior knowledge into beam orientation optimization in IMRT. *International Journal of Radiation Oncology Biology Physics*, 54(5):1565–1574, 2002.

[RAD+03] H. E. Romeijn, R. K. Ahuja, J. F. Dempsey, A. Kumar, and J. G. Li. A novel linear programming approach to fluence map optimization for intensity modulated radiation therapy treatment planning. *Physics in Medicine and Biology*, 48(21):3521–3542, 2003.

[RvDdH13] G. Rennen, E. R. van Dam, and D. den Hertog. Enhancement of sandwich algorithms for approximating higher-dimensional convex pareto sets. *INFORMS Journal on Computing*, 23(4):493–517, 2013.

[Scu10] D. Sculley. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 1177–1178, New York, NY, USA, 2010.

[SdHH11] A. Y. D. Siem, D. den Hertog, and A. L. Hoffmann. A method for approximating univariate convex functions using only function value evaluations. *INFORMS Journal on Computing*, 23(4):591–604, 2011.

[SKB+05] A. Scherrer, K. H. Küfer, T. Bortfeld, M. Monz, and F. Alonso. IMRT planning on adaptive volume structures—a decisive reduction in computational complexity. *Physics in Medicine and Biology*, 50(9):2033–2053, 2005.

[Tro04] J. A. Tropp. *Topics in Sparse Approximation*. PhD thesis, University of Texas at Austin, Austin, TX, USA, 2004.

[UHZB16] M. Udell, C. Horn, R. Zadeh, and S. Boyd. Generalized low rank models. *Foundations and Trends in Machine Learning*, 9(1):1–118, 2016.

[XW05] R. Xu and D. C. Wunsch. Clustering algorithms in biomedical research: A review. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.

[XW10] R. Xu and D. C. Wunsch. Clustering algorithms in biomedical research: A review. *IEEE Reviews in Biomedical Engineering*, 3:120–154, 2010.

[ZLL+14] M. Zarepisheh, T. Long, N. Li, Z. Tian, H. E. Romeijn, X. Jia, and S. B. Jiang. A DVH-guided IMRT optimization algorithm for automatic treatment planning and adaptive radiotherapy replanning. *Medical Physics*, 41(6):061711, 2014.

[ZSM+99] H. H. Zhang, L. Shi, R. Meyer, D. Nazareth, and W. D'Souza. Solving beam-angle selection and dose optimization simultaneously via high-throughput computing. *INFORMS Journal on Computing*, 21(3):427–444, 1999.

[ZYB+14] M. Zarepisheh, Y. Ye, S. Boyd, R. Li, and L. Xing. SU-E-T-295: Simultaneous beam sampling and aperture shape optimization for station parameter optimized radiation therapy (SPORT). *Medical Physics*, 41(6):292–292, 2014.