Optimization and Engineering
https://doi.org/10.1007/s11081-018-9409-2

RESEARCH ARTICLE

A convex optimization approach to radiation treatment planning with dose constraints

Anqi Fu[1] · Barış Ungun[2] · Lei Xing[3] · Stephen Boyd[1]

Received: 30 March 2018 / Revised: 11 November 2018 / Accepted: 11 November 2018

Abstract
We present a method for handling dose constraints as part of a convex programming framework for inverse treatment planning. Our method uniformly handles mean dose, maximum dose, minimum dose, and dose-volume (i.e., percentile) constraints as part of a convex formulation. Since dose-volume constraints are non-convex, we replace them with a convex restriction. This restriction is, by definition, conservative; to mitigate its impact on the clinical objectives, we develop a two-pass planning algorithm that allows each dose-volume constraint to be met exactly on a second pass by the solver if its corresponding restriction is feasible on the first pass. In another variant, we add slack variables to each dose constraint to prevent the problem from becoming infeasible when the user specifies an incompatible set of constraints or when the constraints are made infeasible by our restriction. Finally, we introduce `ConRad`, a Python-embedded open-source software package for convex radiation treatment planning. `ConRad` implements the methods described above and allows users to construct and plan cases through a simple interface.

Keywords Optimization · Convex optimization · Radiation therapy · Treatment planning

1 Introduction

External beam radiation therapy is the treatment of diseased tissue with beams of ionizing radiation delivered from a source outside the patient. When radiation passes through the patient, it damages both healthy and diseased tissue. A treatment plan must be carefully designed to minimize harm to healthy organs, while delivering enough dose to ablate the targeted tissue. With recent hardware advances, delivered

Anqi Fu and Barış Ungun contributed equally to this paper.

✉ Anqi Fu
anqif@stanford.edu

Extended author information available on the last page of the article

Published online: 22 November 2018

beams can be positioned and shaped with growing sophistication, and clinicians rely increasingly on optimization techniques to guide their treatment decisions. In this paper, we focus on one part of the treatment planning process: the selection of an optimal intensity profile for every radiation beam.

In Shepard et al. (1999), the authors provide a comprehensive survey of several problem formulations for treatment planning, including linear (Rosen et al. 1991; Hölder 2003) and quadratic programming models (Bortfeld et al. 1990; Xing and Chen 1996; Xing et al. 1998). Generally, linear models minimize the weighted sum of doses or the maximum deviation from a prescribed dose, while quadratic models minimize the weighted sum of squared difference between actual and prescribed dose. These formulations incorporate linear bounds on the dose to each structure. Solutions can be rapidly found using various interior point methods, such as primal-dual (Aleman et al. 2010), projected gradient (Aleman et al. 2013), and interior point constraint generation (Oskoorouchi et al. 2011).

To address conflicting clinical goals, researchers have proposed models with multiple objectives and constraints. By varying the weight on each objective, one can produce a set of solutions on the Pareto frontier (Hamacher and Küfer 2002; Halabi et al. 2006a). A large number of treatment evaluation criteria can be transformed into convex criteria within this framework (Romeijn et al. 2004). Although multi-objective optimization offers flexibility, calculating thousands of points on the Pareto frontier proves computationally inefficient in practice, and expert judgment is still required to select a clinically acceptable plan from the set of mathematically optimal plans.

All the methods discussed so far hinge on a convex problem formulation. However, many clinically relevant constraints are non-convex. One such type is the dose-volume constraint, which bounds the dose delivered to a given percentage of a patient's anatomy (Zarepisheh et al. 2014). A review of some models for handling this class of constraints is given in Ehrgott et al. (2008). The simplest approach is to add a nonlinear, volume-sensitive penalty to the objective function (Cho et al. 1998; Spirou and Chui 1998). Then, a local search algorithm, such as the conjugate gradient method (Xing and Chen 1996; Xing et al. 1998; Shepard et al. 2000b) or simulated annealing (Webb 1989, 1992; Mageras and Mohan 1993), is used to solve the optimization problem. Unfortunately, since this formulation is non-convex, these algorithms often produce a local minimum, resulting in a sub-optimal treatment plan (Deasy 1997; Wu and Mohan 2002). Another method is to directly model the dose-volume constraint with a set of binary decision variables. Each variable indicates whether a voxel should be included in the fraction of a structure's volume that must fulfill the dose bound (Langer et al. 1990; Lee et al. 2000, 2003). Given the dimensions of patient data, this results in a large-scale mixed-integer programming problem, which is prohibitively expensive to compute for most clinical cases.

A more promising approach is to replace each dose-volume constraint with a convex approximation. This allows users to take advantage of large-scale convex optimization algorithms to quickly generate treatment plans. For instance, Halabi et al. (2006b) substitutes a ramp function for the indicator that a particular voxel violates its desired dose-volume threshold, then penalizes the total number of voxel violations in the objective. Other researchers have employed conditional

value-at-risk (CVaR), a metric that represents the average tail loss in a probability distribution (Rockafellar and Uryasev 2000). It is convex in the loss variable and thus offers a computationally suitable alternative to the dose-volume constraint. In the recent literature, CVaR has been used to formulate linear constraints on the average dose in the upper and lower tails of a structure's dose distribution, leading to significant improvements in treatment plans (Romeijn et al. 2003, 2006; Chan et al. 2014). However, CVaR functions are parametric, and implementations of this model require a heuristic search over the parameter space to obtain a good approximation of the dose-volume constraint (Ahmed et al. 2010).

Perhaps the method most similar to ours is Zarepisheh et al. (2013). In this paper, the authors propose constraining the dose moments to equal those of the desired dose-volume histogram curve. They derive a convex relaxation of these constraints, then solve their treatment planning problem in two phases: the first phase adds slack variables to the moment bounds, so a solution is always feasible, while the second phase tightens these bounds in order to improve upon plan quality whenever possible. Using only three moments, their technique is able to closely match the reference histogram curves in a prostate cancer case.

These results demonstrate the power of convex models: they provide more flexibility than linear or quadratic models, while avoiding the issue of multiple local optima in non-convex formulations and the intractability of a mixed-integer program. In this paper, we propose a new convex formulation of the fluence map optimization (FMO) problem with dose-volume constraints. Given a predetermined number of candidate beams, we construct a convex optimization problem around a set of clinical goals and solve for the radiation intensity pattern that best achieves these goals subject to restrictions on the dose distribution. Our algorithm is quick and efficient, allowing clinicians to rapidly compare trade-offs between different plans and select the best treatment for a patient. We provide a Python package, ConRad, that implements our method within a simple intuitive interface.

This paper is organized as follows: In Sect. 2, we review the radiation treatment planning problem. In Sect. 3, we describe the patient characteristics and constraints that clinicians must consider when selecting an optimal plan. In Sect. 4, we define a convex optimization framework for the basic treatment planning problem. Sect. 5 introduces dose constraints. We show how to incorporate dose-volume constraints via a convex restriction, which provides an approximation of the dose percentile. In Sect. 6, we present two extensions to our model. Sect. 7 describes the Python implementation of our algorithm, and Sect. 8 demonstrates its performance in several clinical cases. Finally, Sect. 9 concludes.

2 Problem description

During external beam radiation therapy, ionizing radiation travels through a patient, depositing energy along the beam paths. Radiation damages both diseased and healthy tissue, but clinicians aim to damage these tissues differentially, exploiting the fact that cancer cells typically have faulty cell repair mechanisms and exhibit a lower tolerance to radiation than healthy cells. The goal is to focus radiation beams

such that enough dose is delivered to kill diseased tissue, while avoiding as much of the surrounding healthy organs as possible. The clinician separates these structures into one or more planning target volumes (PTVs) to be irradiated at a prescribed dose level and several organs-at-risk (OARs) to spare from radiation.

Before treatment, the patient is positioned on a couch. Photon, electron, proton, or heavier particle beams are generated with a particle accelerator and coupled to a mechanized gantry that contains additional hardware components, which shape and focus the beams. The gantry typically rotates around one central axis (but may have additional rotational and translational degrees of freedom), so that by controlling the gantry and couch, beams can be delivered from almost any angle and location around the patient (Mackie et al. 1993; Glide-Hurst et al. 2013; Adler et al. 1998).

Delivery strategies vary from using a large number of apertures (beam shapes) delivered sequentially from a few beam angles, as in intensity-modulated radiation therapy (IMRT), to calculating a single optimal aperture at a large number of angles, as in volumetric modulated arc therapy (VMAT). For a given delivery strategy, the goal of treatment planning is to determine the optimal beam angles, shapes and intensities that most closely approximate a desired dose distribution to the targeted volumes. In this work, we consider the task of optimizing intensities for a given set of beams of known positions and shapes, i.e., calculating optimal beam weights. This is applicable to the fluence map optimization step in IMRT planning, the FMO step in direct aperture optimization for modalities such as VMAT, 4π, or SPORT (Bedford 2009; Dong et al. 2013; Li and Xing 2013), as well as inverse planning problems for other common modalities such as stereotactic radiosurgery (Shepard et al. 2000a; Schweikard et al. 2006) or proton beam therapy (Oelfke and Bortfeld 2001).

3 Clinical planning

3.1 Dose physics

Prior to treatment, medical images—such as CT, MRI and PET scans—are collected to form a three-dimensional image of the patient's anatomy. This representation is discretized into regular volume elements, or voxels. The anatomy is then delineated by clinicians into various structures, and the dose to each structure is considered during planning. Although the structure contours drawn by clinicians may overlap, in this work, we associate each voxel with a single structure.

Dose calculation algorithms range from analytical approximations to Monte Carlo simulations, but in all cases, they provide a model with a linear mapping from beam intensities to delivered voxel doses. The dose within each voxel is assumed to be uniform. For each candidate beam, we have an aperture shape that may be further subdivided, e.g., into regular rectangular subdivisions called beamlets. The intensities of these beams (or beamlets) are represented in a vector. A patient-specific dose deposition matrix maps this vector of radiation intensities to the vector of doses delivered per voxel.

3.2 Dose objectives

Given a fixed number of candidate beams, our goal is to determine the beam intensities that satisfy a clinical objective defined in terms of the dose delivered to each voxel in the patient anatomy.

Every PTV is prescribed a desired dose, which we wish to deposit uniformly throughout the target. Delivering too high or too low a dose of radiation has different clinical consequences, so we introduce separate underdose and overdose penalties for every PTV. In the case of OARs, a lower dose is always preferable, so we penalize any dose above zero and omit an underdose penalty term. In addition to clinical considerations, such as the patient's medical history and past courses of radiation therapy, different organs usually exhibit different levels of sensitivity to radiation. For these reasons, we allow the penalties for each OAR to be scaled independently, allowing the planner to adjust the relative importance of meeting the dose targets for each structure separately.

We apply the penalty associated with each structure to every voxel in that structure, and the objective function of our treatment planning problem tallies these dose penalties over all voxels in a patient's anatomy.

3.3 Dose constraints

In addition to dose penalties, we allow for hard constraints on the amount of radiation delivered to portions of the patient anatomy. For example, the clinician may only consider plans in which the spinal column receives a dose below a certain level because any more will increase the likelihood of injury beyond an acceptable limit. Basic constraints of this nature take the form of bounds on the mean, minimum, and maximum dose to a structure. More generally, bounds can be enforced on the dose to a fraction of the voxels in a structure. These dose-volume constraints restrict the relative volume that receives radiation beyond a particular threshold, giving the clinician precise control over the dose distribution. This is especially important when sparing OARs, since some organs are able to sustain high levels of uniform radiation, while others will fail unless the radiation is contained to a small fraction of the tissue.

Clinicians typically use a dose-volume histogram (DVH) to assess the quality of a treatment plan. For every structure, the DVH specifies the percentage of its volume that receives at least a certain dose. A point (x, y) on the curve indicates that $y\%$ of the total voxels in the structure receives a dose of at least x Gy. Ideally, we want our structures to receive exactly the prescribed dose throughout their volumes. If the prescription is d Gy, then our optimal DVH curve for the PTV is a step function with a drop at $(d, 100)$, and our optimal DVH for each OAR exhibits a drop at $(0, 100)$.

Dose constraints restrict the shape and location of points on the DVH curve. In Fig. 1, a lower dose-volume constraint, $D(90) \geq 60$, is represented by the right-facing arrow centered at (60, 90). This ensures that a minimum of 90% of the structure's volume receives at least 60 Gy, i.e., $y \geq 90$ along the vertical line $x = 60$. The PTV's DVH curve is pushed rightward by this type of constraint. Similarly, an upper dose-volume constraint, $D(33) \leq 12$, is labeled with a left-facing arrow at (12, 33), which pushes the OAR's DVH curve leftward, representing the restriction that $y \leq 33$ along

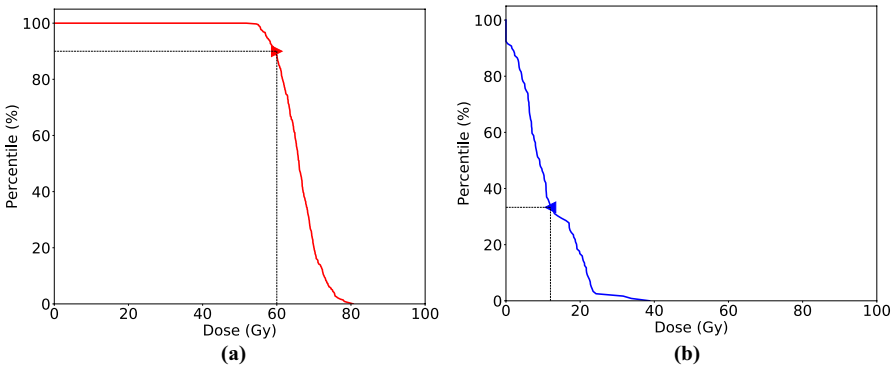

Fig. 1 **a** A lower DVH constraint ensures at least 90% of the structure's volume receives at least 60 Gy. The dotted line intersects the curve at (60, 90). **b** An upper DVH constraint allows at most 33% of the volume to receive at least 12 Gy

the line $x = 12$. Together, the DVH curves and their respective dose constraints enable the clinician to easily visualize trade-offs when formulating a treatment plan.

4 Convex formulation

Consider a case with m voxels inside a patient volume and n candidate treatment beams. Our goal is to determine the beam intensities $x \in \mathbf{R}^n_+$ that deliver a vector of voxel doses $y \in \mathbf{R}^m_+$, which meet a set of clinical objectives. We are given a case-specific dose influence matrix $A \in \mathbf{R}^{m \times n}_+$ that approximates the relationship between the beams and doses linearly as $y = Ax$. We refer to the rows of A as $a_i \in \mathbf{R}^n_+$ for $i = 1, \ldots, m$. The basic inverse treatment planning problem is of the form

$$\begin{aligned} \underset{x,y}{\text{minimize}} \quad & f(y) \\ \text{subject to} \quad & y = Ax \\ & x \geq 0, \end{aligned}$$

where $f : \mathbf{R}^m \to \mathbf{R}$ is a convex loss function chosen to penalize voxel doses based on the goals of the clinician. Here, the inequality on x is understood to be applied element-wise. In a typical case, a patient is prescribed a treatment plan that can be characterized by a vector of doses $d \in \mathbf{R}^m_+$ to each voxel. Our function f then penalizes the deviation of the calculated dose y from the prescribed dose d, taking into account the different structures inside a patient.

In our formulation, we consider a loss function $f(y) = \sum_{i=1}^m f_i(y_i)$ where $y_i = a_i^T x$ and each f_i is a piecewise-linear function

$$f_i(y_i) = w_i^-(y_i - d_i)_- + w_i^+(y_i - d_i)_+.$$

The parameters w_i^- and w_i^+ are the non-negative weights on the underdose and overdose, respectively. This penalty structure is common in the literature (Lim and Cao 2012; Chen et al. 2012) and provided the most efficient software implementation.

Prior to treatment planning, the m voxels in a patient volume are grouped into S distinct, non-overlapping sets representing the planning target volume (PTV), organs-at-risk (OARs), and generic non-target tissue (often labeled "body"). Each set $\mathscr{V}_s$ contains all the voxel indices i within a corresponding internal structure with index s. Together, $\{\mathscr{V}_s\}_1^S$ forms a partition of the patient volume, i.e., $\bigcup_1^S \mathscr{V}_s$ covers all voxel indices and $\mathscr{V}_{s_1} \bigcap \mathscr{V}_{s_2} = \emptyset$ for $s_1 \neq s_2$. We assume the indices are ordered such that $s = 1, \ldots, P \leq S$ are targets and the rest non-targets.

For simplicity, we choose our voxel doses and penalties to be uniform within each structure. We let d_s represent the prescribed dose, and w_s^- and w_s^+ the underdose and overdose penalties for all voxels $i \in \mathscr{V}_s$. The loss function for our basic inverse planning problem is $f(y) = \sum_{s=1}^{S} f_s(y_i)$ where

$$f_s(y_i) = \sum_{i \in \mathscr{V}_s} f_i(y_i) = \sum_{i \in \mathscr{V}_s} \left\{ w_s^-(y_i - d_s)_- + w_s^+(y_i - d_s)_+ \right\}.$$

A non-target structure s is always prescribed a dose of zero, and since $y \geq 0$, its individual loss simplifies to $f_i(y_i) = w_s^+ y_i$. Thus, its only contribution to the objective is through its total dose. An example of these loss functions is given in Fig. 2. We can collapse the sum of non-target losses into a single linear term,

$$\sum_{s=P+1}^{S} f_s(y_i) = \sum_{s=P+1}^{S} w_s^+ \left(\sum_{i \in \mathscr{V}_s} y_i \right) = \sum_{s=P+1}^{S} w_s^+ z_s = c^T z,$$

where $c = (w_{P+1}^+, \ldots, w_S^+)$ and $z = \left(\sum_{i \in \mathscr{V}_{P+1}} y_i, \ldots, \sum_{i \in \mathscr{V}_S} y_i \right)$. Our objective is then

$$f(y) = \sum_{s=1}^{P} \sum_{i \in \mathscr{V}_s} \left\{ w_s^-(y_i - d_s)_- + w_s^+(y_i - d_s)_+ \right\} + c^T z.$$

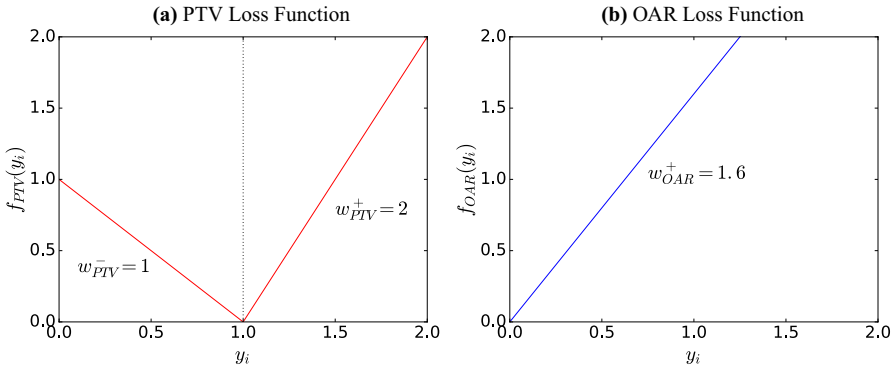

Fig. 2 **a** The loss function for a PTV prescribed $d_s = 1$ with penalties $w_s^- = 1$ and $w_s^+ = 2$, and **b** the loss function for an OAR with penalty $w_{s'}^+ = 1.6$

This formulation is closely related to quantile regression (Davino et al. 2013). In the latter, we minimize $\sum_i \phi(y_i - v - d_i)$ with respect to (x, v) where

$$\phi(u) = \tau(u)_+ + (1-\tau)(u)_- = \frac{1}{2}|u| + \left(\tau - \frac{1}{2}\right)u$$

is the tilted ℓ_1 penalty with $\tau \in (0,1)$. For our inverse planning problem, the residual $r_i := y_i - v - d_i$ can be interpreted as the difference between calculated and desired doses, allowing for a uniform dose offset $v \in \mathbf{R}$ within each structure. We rewrite our individual loss as

$$\begin{aligned} f_i(y_i - v) &= w_i^-(r_i)_- + w_i^+(r_i)_+ \\ &= \left(w_i^- + w_i^+\right)\left(\tfrac{w_i^-}{w_i^- + w_i^+}(r_i)_- + \tfrac{w_i^+}{w_i^- + w_i^+}(r_i)_+\right) \\ &= \left(w_i^- + w_i^+\right)\left(\tfrac{1}{2}|r_i| + \left(\tfrac{w_i^+}{w_i^- + w_i^+} - \tfrac{1}{2}\right)r_i\right), \end{aligned}$$

and the loss function becomes

$$f_s(y_i - v) = \sum_{i \in \mathscr{V}_s} f_i(y_i - v) = (w_s^- + w_s^+) \sum_{i \in \mathscr{V}_s} \left(\frac{1}{2}|r_i| + \left(\tau_s - \frac{1}{2}\right)r_i\right),$$

where $\tau_s := \frac{w_s^+}{w_s^- + w_s^+} \in (0,1)$. For $r_i \neq 0$, the first order condition with respect to v is

$$\frac{\partial f_s(y_i - v)}{\partial v} = \tau_s |\{i : r_i > 0\}| - (1 - \tau_s)|\{i : r_i < 0\}| = 0,$$

which implies $\tau_s|\mathscr{V}_s| = |\{i : r_i < 0\}|$, i.e., in a given structure s, the τ_s-quantile of optimal residuals is zero. Although our original loss does not include v, we can use this as a rule of thumb for selecting relative dose penalties (w_s^-, w_s^+).

5 Dose constraints

5.1 Percentile

A percentile constraint, otherwise known as a dose-volume constraint, bounds the dose delivered to a given percentile of a patient structure. This allows us to set a limit on the fraction of total voxels that are under- or overdosed with respect to a user-provided threshold. For clinicians, this provides a way to shape the dose-volume histogram directly rather than by searching through combinations of objective weights to achieve desired dose statistics. Given a structure s and dose vector y, let $D_s(p, y)$ represent the minimum dose delivered to p percent of all voxels in s, i.e., $D_s(p, y)$ is the greatest lower bound on the dose received by $p\%$ of the tissue.

To formalize this notion, we define an exact value count function $v_s : \mathbf{R}_+^m \times \mathbf{R}_+ \to \mathbf{Z}_+$, which computes the total number of voxels $i \in \mathscr{V}_s$ that receive a dose above $b \in \mathbf{R}_+$. Let $g(u) = \mathbb{1}\{u \geq 0\}$, then

$$v_s(y, b) = \sum_{i \in \mathscr{V}_s} \mathbb{1}\{y_i \geq b\} = \sum_{i \in \mathscr{V}_s} g(y_i - b)$$

and our p-th percentile dose is

$$D_s(p, y) = \max\{b \in \mathbf{R}_+ : v_s(y, b) \geq \phi_s(p)\} \quad \text{where} \quad \phi_s(p) := \frac{p}{100}|\mathscr{V}_s|.$$

Observe that $D_s(p, y) \geq 0$ is finite and weakly decreasing in p.

Our goal is to bound $D_s(p, y)$. For example, we may want at least 30% of the voxels in structure s to receive a dose above 25 Gy; this is identical to $D_s(30, y) \geq 25$. Let $\ell < u$ be non-negative scalar values. An lower dose-volume constraint, $D_s(p, y) \geq \ell$, requires the number of voxels in s that receive a dose above ℓ to be at least p% of the total voxels in the structure. Similarly, an upper dose-volume constraint, $D_s(p, y) \leq u$, requires the number of voxels $i \in \mathscr{V}_s$ with a dose above u to be at most p% of voxels in $\mathscr{V}_s$, or equivalently, at least $100 - p$% of the voxels to receive a dose under u. Thus, the inequalities

$$D_s(p, y) \leq u \quad \Leftrightarrow \quad v_s(y, u) \leq \phi_s(p) \quad \Leftrightarrow \quad v_s(-y, -u) \geq \phi_s(100 - p)$$

are equivalent, as are

$$D_s(p, y) \geq \ell \quad \Leftrightarrow \quad v_s(y, \ell) \geq \phi_s(p) \quad \Leftrightarrow \quad v_s(-y, -\ell) \leq \phi_s(100 - p).$$

In general, this is a hard combinatorial problem: the brute force approach for a single upper dose-volume constraint, for example, is to solve all $\binom{|\mathscr{V}_s|}{\phi}$ convex problems obtained by choosing subsets of $\phi = \lceil \phi_s(p) \rceil$ voxels to constrain below u, which is prohibitively large given the size of patient geometries.

5.2 Mean, minimum, and maximum

In a few special cases, we can set convex constraints on the dose. Let the average, minimum, and maximum dose delivered to all voxels in structure s be

$$D_s^{\text{avg}}(y) = \frac{1}{|\mathscr{V}_s|} \sum_{i \in \mathscr{V}_s} y_i, \quad D_s^{\min}(y) = \min_{i \in \mathscr{V}_s}\{y_i\}, \quad D_s^{\max}(y) = \max_{i \in \mathscr{V}_s}\{y_i\}.$$

A lower bound $b \in \mathbf{R}_+$ on the minimum dose is equivalent to requiring $y_i \geq b$ for all $i \in \mathscr{V}_s$, and similarly for an upper bound on the maximum dose. Thus, we can enforce linear constraints on these dose statistics in our problem. Our non-convex formulation with exact dose-volume constraints is

$$
\begin{array}{lll}
\underset{x,y}{\text{minimize}} & f(y) & \\
\text{subject to} & y = Ax & \\
& x \geq 0 & \\
& \ell_{s,k} \leq D_s(p_{s,k}, y) \leq u_{s,k}, & k = 1, \ldots, K_s, \quad s = 1, \ldots, S \\
& \ell_s^{\text{avg}} \leq D_s^{\text{avg}}(y) \leq u_s^{\text{avg}}, & s = 1, \ldots, S \\
& D_s^{\min}(y) \geq \ell_s^{\min}, & s = 1, \ldots, S \\
& D_s^{\max}(y) \leq u_s^{\max}, & s = 1, \ldots, S
\end{array} \tag{1}
$$

where for each structure s, we index the parameters of its dose-volume constraints with $k = 1, \ldots, K_s$.

5.3 Convex restriction

To address the non-convexities in Prob. 1, we introduce a convex restriction that provides an effective heuristic for satisfying the dose-volume constraints. Our restricted constraint overestimates the number of voxels that are underdosed with respect to d by replacing the indicator g in v with a family of hinge loss functions

$$\hat{g}_\lambda(u) = (1 + \lambda u)_+ = \max(1 + \lambda u, 0),$$

parametrized by $\lambda > 0$, giving us a restricted value count for structure s of

$$\hat{v}_s(y, b; \lambda) = \sum_{i \in \mathcal{V}_s} \hat{g}_\lambda(y_i - b) = \sum_{i \in \mathcal{V}_s} (1 + \lambda(y_i - b))_+.$$

If $u > 0$, then $g(u) = 1 < 1 + \lambda u = \hat{g}_\lambda(u)$, and if $u \leq 0$, then $g(u) = 0 \leq \hat{g}_\lambda(u)$. Hence, $g(u) \leq \hat{g}_\lambda(u)$ for all $u \in \mathbf{R}$ and $\lambda > 0$ (Fig. 3). Evaluating at $u_i = y_i - b$ and summing over all voxels $i \in \mathcal{V}_s$, we obtain

$$v_s(y, b) = \sum_{i \in \mathcal{V}_s} g(y_i - b) \leq \sum_{i \in \mathcal{V}_s} \hat{g}_\lambda(y_i - b) = \hat{v}_s(y, b; \lambda).$$

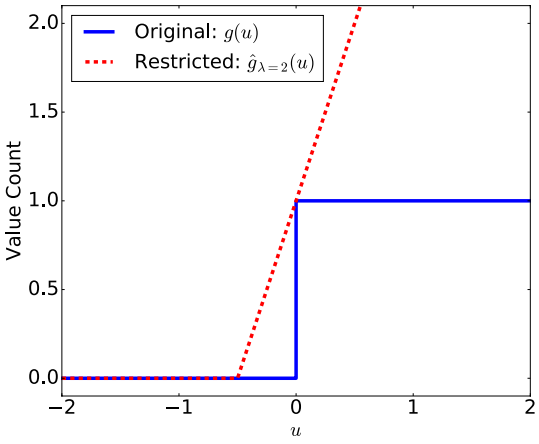

Fig. 3 The indicator function $g(u)$ (solid) and hinge loss $\hat{g}_\lambda(u)$ with $\lambda = 2$ (dashed). Note that $\hat{g}_\lambda(u) \geq g(u)$ for all $u \in \mathbf{R}$, so the hinge loss provides a convex restriction on the dose-volume constraint

An upper bound on the restricted value count at a given point thus ensures the exact value count is bounded above as well.

We can guarantee our dose-volume constraints hold by enforcing specific limits on $\hat{v}$. If $\hat{v}_s(y,u;\lambda) \le \phi_s(p)$, then $v_s(y,u) \le \phi_s(p)$, and the upper dose-volume constraint, $D_s(p,y) \le u$, is satisfied. Similarly, $\hat{v}_s(-y,-\ell;\lambda) \le \phi_s(100-p)$ implies that $D_s(p,y) \le \ell$. To simplify notation, we rewrite $\hat{v}_s(y,b;\lambda) \le \phi$ as

$$\sum_{i\in\mathscr{V}_s}(1+\lambda(y_i-b))_+ \le \phi.$$

Since $\lambda > 0$, we can divide both sides of the inequality by λ. Letting $\alpha := \frac{1}{\lambda}$ and gathering all terms on the left-hand side, we obtain the inequality

$$\sum_{i\in\mathscr{V}_s}(\alpha+(y_i-b))_+ - \alpha\phi \le 0.$$

The left-hand side of this inequality is a sum of convex functions of (α, y), and hence convex. Note that while λ was a parameter of our restricted value count function, $\alpha > 0$ can be an optimization variable, since the left-hand term is jointly convex in α and y. Additionally, we can replace the constraint $\alpha > 0$ with $\alpha \ge 0$ because when $\alpha = 0$, the constraint simplifies to $(y_i - b)_+ \le 0$, which is equivalent to $y_i \le b$ for all $i \in \mathscr{V}_s$. Certainly in this case, the condition $D_s(p,y) \le b$ holds. Thus, by defining the functions

$$\hat{D}_s^+(p,y,b,\alpha) = \sum_{i\in\mathscr{V}_s}(\alpha+(y_i-b))_+ - \alpha\phi_s(p)$$

for upper constraints and

$$\hat{D}_s^-(p,y,b,\alpha) = \sum_{i\in\mathscr{V}_s}(\alpha-(y_i-b))_+ - \alpha\phi_s(100-p),$$

for lower constraints, each convex restriction can be represented by inequalities in terms of these functions: for $\alpha \ge 0$, $\hat{D}_s^+(p,y,u,\alpha) \le 0$ implies $D_s(p,y) \le u$, and $\hat{D}_s^-(p,y,\ell,\alpha) \le 0$ implies $D_s(p,y) \ge \ell$. Our convex formulation with restricted dose-volume constraints is

$$\begin{array}{lll}
\underset{(x,y,\alpha)}{\text{minimize}} & f(y) & \\
\text{subject to} & y = Ax & \\
& x \ge 0, \alpha \ge 0 & \\
& \hat{D}_s^+\left(p_{s,k}, y, u_{s,k}, \alpha_{s,k}^{(u)}\right) \le 0, & k = 1,\dots,K_s^{(u)}, \quad s = 1,\dots,S \\
& \hat{D}_s^-\left(p_{s,k}, y, \ell_{s,k}, \alpha_{s,k}^{(\ell)}\right) \le 0, & k = 1,\dots,K_s^{(\ell)}, \quad s = 1,\dots,S \\
& \ell_s^{\text{avg}} \le D_s^{\text{avg}}(y) \le u_s^{\text{avg}}, & s = 1,\dots,S \\
& D_s^{\min}(y) \ge \ell_s^{\min}, & s = 1,\dots,S \\
& D_s^{\max}(y) \le u_s^{\max}, & s = 1,\dots,S,
\end{array} \tag{2}$$

where for every structure s, we index the parameters of its upper dose-volume constraints with $k = 1, \ldots, K_s^{(u)}$, and its lower dose-volume constraints with $k = 1, \ldots, K_s^{(\ell)}$. We include a separate optimization variable, $\alpha_{s,k}$, in each dose-volume constraint to represent the inverse slope of its convex restriction and stack these variables in a vector $\alpha := (\alpha^{(\ell)}, \alpha^{(u)})$. Optimizing over α in addition to (x, y) ensures we obtain the best hinge loss approximation to the value count function. The above formulation is a restriction of our original problem: if (x, y, α) is feasible for Prob. 2, then (x, y) is feasible for Prob. 1.

6 Refinements

6.1 Two-pass refinement

A solution (x^*, y^*, α^*) to Prob. 2 satisfies our restricted dose-volume constraints, so it is feasible for our original Prob. 1 with exact dose-volume constraints. However, since the convex restriction enforces an upper bound on the restricted value count function $\hat{v}$, the feasible set of Prob. 2 is a subset of the feasible set of Prob. 1, and (x^*, y^*) may not be optimal for the latter. One way to improve our solution is to bound only the minimum number of voxels in each structure required to satisfy the dose-volume constraint. A good heuristic is to select those voxels i that receive a dose y_i^*, which satisfies the associated dose-volume bound by the largest margin, and re-solve the problem with the convex restriction replaced by bounds on just these voxels. The solution of this second pass, (x^{**}, y^{**}), will achieve an objective value $f(y^{**}) \leq f(y^*)$ while still satisfying our exact dose-volume constraints.

To make this precise, consider the lower dose-volume constraint $D_s(p, y) \geq \ell$. This is equivalent to $y_i \geq \ell$ for at least $\phi_s(p)$ voxels in structure s. Given y^* from our first pass optimization, we compute the margin $\xi_i^* = (y_i^* - \ell)$ and select the $q_s = \lceil \phi_s(p) \rceil$ voxels $i \in \mathcal{V}_s$ with the largest values of ξ_i^*. Call this subset $\mathcal{Q}_s^- \subseteq \mathcal{V}_s$. Now, we replace $D_s(p, y) \geq \ell$ in Prob. 1 with the precise voxel constraints $y_i \geq \ell$ for all $i \in \mathcal{Q}_s^-$. On the second pass,

$$v_s(y, \ell) = \sum_{i \in \mathcal{V}_s} \mathbb{1}\{y_i \geq \ell\} \geq \sum_{i \in \mathcal{Q}_s^-} \mathbb{1}\{y_i \geq \ell\} = q_s \geq \phi_s(p),$$

so our upper dose-volume constraint is satisfied. An analogous argument with $q_s = \lceil \phi_s(100 - p) \rceil$ and $\xi_i^* = (u - y_i^*)$ produces the subset $\mathcal{Q}_s^+$ for an upper dose-volume constraint $D_s(p, y) \leq u$. Given a solution (x^*, y^*, α^*) to Prob. 2, we repeat this process with every such constraint to obtain the second-pass problem formulation

$$
\begin{array}{lll}
\underset{x,y}{\text{minimize}} & f(y) & \\
\text{subject to} & y = Ax & \\
& x \geq 0 & \\
& y_i \leq u_{s,k} \ \forall i \in \mathcal{Q}^{+}_{s,k}, & k = 1, \ldots, K^{(u)}_s, \quad s = 1, \ldots, S \\
& y_i \geq \ell_{s,k} \ \ \forall i \in \mathcal{Q}^{-}_{s,k}, & k = 1, \ldots, K^{(\ell)}_s, \quad s = 1, \ldots, S \\
& \ell^{\text{avg}}_s \leq D^{\text{avg}}_s(y) \leq u^{\text{avg}}_s, & s = 1, \ldots, S \\
& D^{\text{min}}_s(y) \geq \ell^{\text{min}}_s, & s = 1, \ldots, S \\
& D^{\text{max}}_s(y) \leq u^{\text{max}}_s, & s = 1, \ldots, S,
\end{array}
\tag{3}
$$

where the voxel subsets are indexed with $k = 1, \ldots, K^{(u)}_s$ for upper dose-volume constraints, and $k = 1 \ldots, K^{(\ell)}_s$ for lower dose-volume constraints. We can warm start our solver at (x^*, y^*) to speed up the second pass optimization.

Algorithm 6.1 *Two-pass algorithm.*

given a dose matrix $A \in \mathbf{R}^{m \times n}$, a prescribed dose vector $d \in \mathbf{R}^m$,
and a set of dose-volume constraints $\mathscr{C}$.
1. *First pass.* Obtain the solution (x^*, y^*, α^*) to Prob. 2.
for each $(\ell, p, s) \in \mathscr{C}$ **do**
2a. *Compute margins.* Calculate $\xi^*_i = y^*_i - \ell$ for all $i \in \mathscr{V}_s$.
2b. *Sort margins.* Sort $\{\xi^*_i\}_{i \in \mathscr{V}_s}$ in ascending order to form a set ξ_s.
2c. *Identify voxel subset.* Select the $\lceil \phi_s(p) \rceil$ largest values $\xi_i \in \xi_s$ and include their indices i in $\mathscr{Q}^{-}_{s,k}$.
end for
for each $(u, p, s) \in \mathscr{C}$ **do**
3a. *Compute margins.* Calculate $\xi^*_i = u - y^*_i$ for all $i \in \mathscr{V}_s$.
3b. *Sort margins.* Sort $\{\xi^*_i\}_{i \in \mathscr{V}_s}$ in ascending order to form a set ξ_s.
3c. *Identify voxel subset.* Select the $\lceil \phi_s(100 - p) \rceil$ largest values $\xi_i \in \xi_s$ and include their indices i in $\mathscr{Q}^{+}_{s,k}$.
end for
4. *Second pass.* Obtain the solution (x^{**}, y^{**}) to Prob. 3 using (x^*, y^*) as a warm start point.

6.2 Dose constraints with slack

If our dose constraints are too strict, Prob. 2 may not have a solution. This can arise even if the feasible set for our original Prob. 1 is non-empty, since our convex restriction enforces more stringent bounds on the dose distribution. To ensure the first pass of our algorithm always supplies a solution, we introduce a slack variable $\delta \in \mathbf{R}_+$ to the bounds of each dose constraint, mapping lower bounds $\ell \mapsto (\ell - \delta)$ and upper bounds $u \mapsto (u + \delta)$. This creates soft constraints that need not be met precisely by the solution. Our problem reformulated with restricted dose-volume constraints and slack is

$$
\begin{array}{lll}
\underset{(x,y,\alpha,\delta)}{\text{minimize}} & f(y) & \\
\text{subject to} & y = Ax & \\
& x \geq 0, \alpha \geq 0, \delta \geq 0 & \\
& \hat{D}_s^+\left(p_{s,k}, y, u_{s,k} + \delta_{s,k}^{(u)}, \alpha_{s,k}^{(u)}\right) \leq 0, & k = 1, \ldots, K_s^{(u)}, \quad s = 1, \ldots, S \\
& \hat{D}_s^-\left(p_{s,k}, y, \ell_{s,k} - \delta_{s,k}^{(\ell)}, \alpha_{s,k}^{(\ell)}\right) \leq 0, & k = 1, \ldots, K_s^{(\ell)}, \quad s = 1, \ldots, S \\
& \ell_s^{\text{avg}} - \delta_s^{\text{avg},(\ell)} \leq D_s^{\text{avg}}(y) \leq u_s^{\text{avg}} + \delta_s^{\text{avg},(u)}, & s = 1, \ldots, S \\
& D_s^{\min}(y) \geq \ell_s^{\min} - \delta_s^{\min}, & s = 1, \ldots, S \\
& D_s^{\max}(y) \leq u_s^{\max} + \delta_s^{\max}, & s = 1, \ldots, S,
\end{array}
\tag{4}
$$

Note that δ is a variable in the optimization, and the value of each $\delta_{s,k}$ indicates the amount (in units of delivered dose, e.g., Gy) by which each bound is weakened in the solution.

We can incorporate soft constraints into the two-pass algorithm as well. On the first pass, we solve Prob. 4 to obtain the optimal variables (x^*, y^*, α^*) and the optimal slacks δ^*. Our margin for selecting $\mathcal{Q}_s$ is now computed with respect to the slack bound, i.e., $\xi_i^* = (y_i^* - \ell - \delta^*)$ for lower-volume dose constraints, and $\xi_i^* = (u + \delta^* - y_i^*)$ for upper dose-volume constraints. Finally, we weaken the bounds in problem (3) by δ^*, giving us the reformulated second pass optimization with slack dose-volume constraints

$$
\begin{array}{lll}
\underset{x,y}{\text{minimize}} & f(y) & \\
\text{subject to} & y = Ax & \\
& x \geq 0 & \\
& y_i \leq u_{s,k} + \delta_{s,k}^{(u)*} \ \forall i \in \mathcal{Q}_{s,k}^+, & k = 1, \ldots, K_s^{(u)}, \quad s = 1, \ldots, S \\
& y_i \geq \ell_{s,k} - \delta_{s,k}^{(\ell)*} \ \forall i \in \mathcal{Q}_{s,k}^-, & k = 1, \ldots, K_s^{(\ell)}, \quad s = 1, \ldots, S \\
& \ell_s^{\text{avg}} - \delta_s^{\text{avg},(\ell)*} \leq D_s^{\text{avg}}(y) \leq u_s^{\text{avg}} + \delta_s^{\text{avg},(u)*}, & s = 1, \ldots, S \\
& D_s^{\min}(y) \geq \ell_s^{\min} - \delta_s^{\min*}, & s = 1, \ldots, S \\
& D_s^{\max}(y) \leq u_s^{\max} + \delta_s^{\max*}, & s = 1, \ldots, S,
\end{array}
\tag{5}
$$

Algorithm 6.2 *Two-pass algorithm with slack.*

given a dose matrix $A \in \mathbf{R}^{m\times n}$, a prescribed dose vector $d \in \mathbf{R}^m$,
and a set of dose-volume constraints $\mathscr{C}$.
1. *First pass.* Obtain the solution $(x^*, y^*, \alpha^*, \delta^*)$ to Prob. 4.
for each $(\delta^*, \ell, p, s) \in \mathscr{C}$ **do**
  2a. *Compute margins.* Calculate $\xi_i^* = y_i^* - \ell - \delta^*$ for all $i \in \mathscr{V}_s$.
  2b. *Sort margins.* Sort $\{\xi_i^*\}_{i\in\mathscr{V}_s}$ in ascending order to form a set ξ_s.
  2c. *Identify voxel subset.* Select the $\lceil \phi_s(p) \rceil$ largest values $\xi_i \in \xi_s$
    and include their indices i in $\mathscr{D}_{s,k}^-$.
end for
for each $(\delta^*, u, p, s) \in \mathscr{C}$ **do**
  3a. *Compute margins.* Calculate $\xi_i^* = u + \delta^* - y_i^*$ for all $i \in \mathscr{V}_s$.
  3b. *Sort margins.* Sort $\{\xi_i^*\}_{i\in\mathscr{V}_s}$ in ascending order to form a set ξ_s.
  3c. *Identify voxel subset.* Select the $\lceil \phi_s(100-p) \rceil$ largest values $\xi_i \in \xi_s$
    and include their indices i in $\mathscr{D}_{s,k}^+$.
end for
4. *Second pass.* Obtain the solution (x^{**}, y^{**}) to Prob. 5 using (x^*, y^*) as a warm start point.

7 Implementation

We implement our radiation treatment planning methodology with ConRad, a Python-embedded open-source software package based on the convex programming library, CVXPY (Diamond and Boyd 2016), using the convex solvers SCS (O'Donoghue et al. 2016) and ECOS (Domahidi et al. 2013). `ConRad` provides a simple, intuitive interface for ingesting patient data, constructing plans based on a clinical prescription, and visualizing the dose-volume histograms of the result. It allows the user to add dose constraints using syntax familiar to clinicians. Since `ConRad` is an ordinary Python library, it can be easily integrated into existing data processing pipelines.

The following code imports a prescription, solves for the optimal treatment plan without dose constraints, and plots the DVH curves for all the patient structures. The $m \times n$ dose-influence matrix `A` can be encoded as a NumPy `ndarray` or any of several sparse representations in Python. The m-length vector `voxel_labels` enumerates the index of the assigned structure for each voxel in the patient volume.

```
import conrad

# Construct the case with no dose constraints
case = conrad.Case()
case.prescription = "/Documents/prescriptions/rx_patient_01.yaml"
case.physics.dose_matrix = A
case.physics.voxel_labels = voxel_labels
graphics = conrad.CasePlotter(case)

# Solve with a single pass and no slack
status, run = case.plan(solver="ECOS", use_slack=False, use_2pass=False)

print("Problem feasible?:\n{}".format(status))
print("Dose summary:\n{}".format(case.dose_summary_string))

# Display color-coded plot of all DVH curves
graphics.plot(run, show=True)
```

A `Case` object comprises `Anatomy`, `Physics`, `Prescription` and `PlanningProblem` objects. Prior to planning, the case's `Anatomy` and `Physics` objects must be built. The user can either build the case's `Anatomy` by adding structures programmatically (with data on each structure's name, index, identity as target/non-target, and desired dose) or by ingesting a prescription, which can be supplied as a Python dictionary or as a YAML or JSON file formatted for `ConRad`'s parser. The minimum information required for the case's `Physics` object are the $m \times n$ dose matrix and a m-length vector of voxel labels. The case's `Prescription` object can be used to populate the `Anatomy` object or to keep track of clinical guidelines and objectives. It can also be left empty.

The case's `PlanningProblem` object builds and solves optimization problems based on the structures in the case anatomy and any constraints assigned to those structures. The `PlanningProblem` is not exposed to the user. Instead, users form a treatment plan by calling the case's `plan()` method, which returns a `bool` status indicating whether the specified problem was feasible, along with a `RunRecord` object that carries solver performance data, optimal variables, and DVH curves.

Before planning a case, the user can add, remove, or modify dose constraints to any structure. Thus, even when a case has an assigned prescription, the dose constraints attached to each structure in the case anatomy may differ from the constraints specified in the prescription. For example, the prescribed constraints may correspond to clinical guidelines, while the constraints used during planning may be chosen arbitrarily by the user to obtain plans with desirable dose properties.

After planning a case, users can plot the DVH curves, retrieve and print summaries of dose statistics for each structure, and when applicable, display a report of whether the current plan satisfies each constraint listed in the prescription. A case can be re-planned with different objective weights or dose constraints on any structure. The `ConRad` library provides a `PlanningHistory` object to retain and manage results from prior runs.

The following code adds a dose-volume constraint to the PTV from our previous case, allowing at most 20% of the PTV's voxels to receive more than 70 Gy. Algorithm 6.1 is then applied to obtain an optimal beam output.

```
# Constraint allows at most 20% of PTV voxels to receive dose above 70 Gy
case.anatomy["PTV"].constraints += D(20) <= 70 * Gy

# Solve with two-pass algorithm and no slack
_, run = case.plan(solver="ECOS", use_slack=False, use_2pass=True)
print("x PASS 1: {}".format(run.x_pass1))
print("x PASS 2: {}".format(run.x_pass2))

# Plot DVH curves from first (dashed) and second pass (solid)
graphics.plot(run, show=False, ls="--")
graphics.plot(run, second_pass=True, show=True, clear=False, legend=True)
```

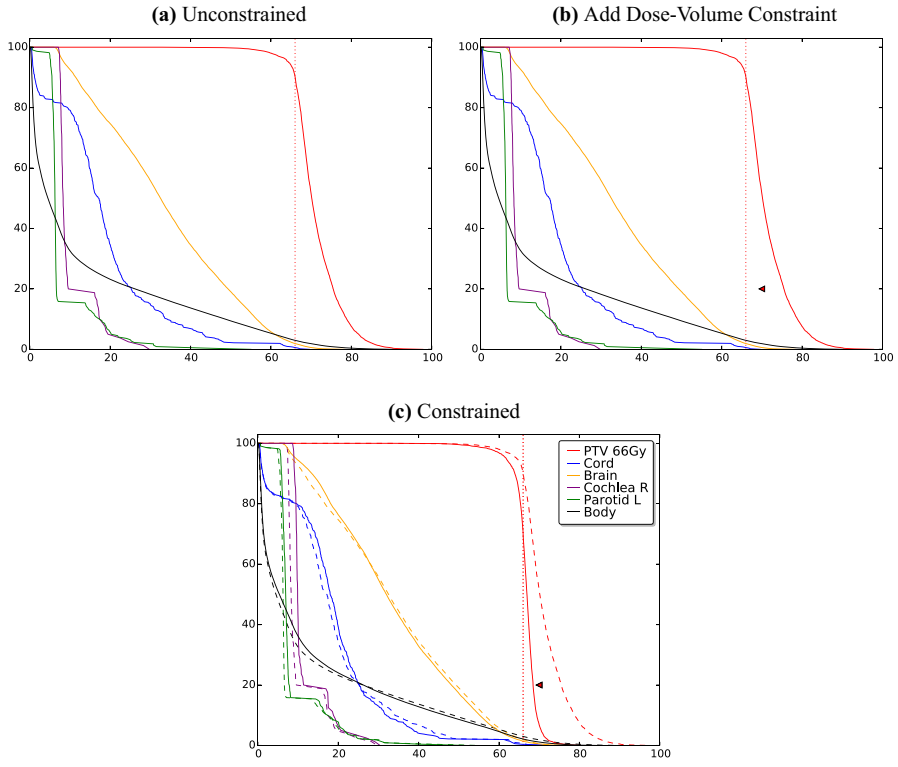

Fig. 4 DVH curves for PTV and several OARs from the 4-arc VMAT head-and-neck case. **a** Plan without any dose constraints. **b** Add a dose-volume constraint $D(20) \leq 70\,\text{Gy}$. **c** Re-plan with the new constraint. The unconstrained plan is shown in dashed lines

8 Examples

8.1 Basic functionality

We present results illustrating the methods described in this paper: approximating dose-volume constraints via convex restrictions, two-pass refinement of plans with dose-volume constraints, and handling incompatible constraints with slack variables.

Problem instances We demonstrate the basic functionality on a head-and-neck case expressed as a VMAT aperture re-weighting problem. The case contains 360 apertures in four arcs, 270,000 voxels distributed across 17 planning structures, including the PTV treated to 66 Gy, two auxiliary targets treated to 60 Gy, several OARs, and generic body voxels.

To test the handling of dose-volume constraints, we plan the case with no dose constraints and then re-plan with a single dose-volume constraint applied to the PTV, namely $D(20) \leq 70$ Gy. We run the two-pass algorithm and compare the plans obtained by applying restricted and exact versions of the aforementioned dose-volume constraint. Finally, we test the slack method by planning the case with two incompatible dose-volume constraints: $D(98) \geq 66$ Gy on the PTV and $D(20) \leq 20$ Gy on the spinal cord. We compare the results from enforcing the PTV constraint alone, both constraints without slack, and both constraints with slack allowed.

Computational details The size of the dose matrix passed from `ConRad` to the convex solvers in the backend varied depending on the dose constraints. When no minimum, maximum, or dose-volume constraints were applied to a non-target structure, the submatrix for that structure was replaced with a mean dose representation, thereby eliminating $|\mathcal{V}_s| - 1$ rows from the problem matrix. In particular, the matrix representing the full dose on the targets and mean dose for non-targets had dimensions $11{,}141 \times 360$, and the matrix including the full dose on the spinal cord was $15{,}000 \times 360$. The `ConRad` problem request was formulated as a convex program in CVXPY and passed to a GPU-based implementation of the convex solver

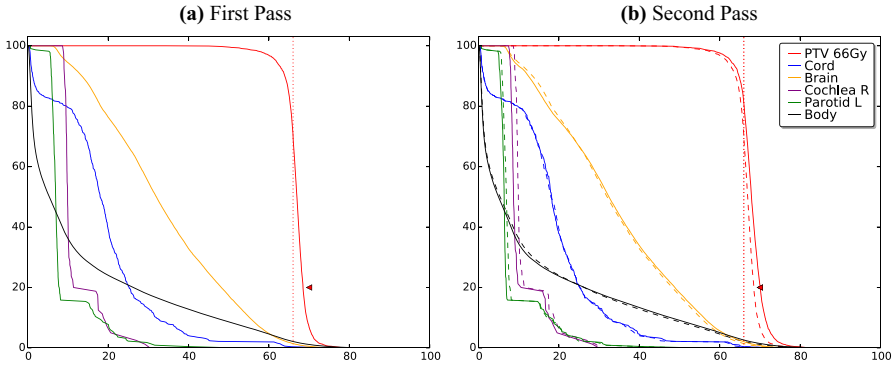

Fig. 5 DVH curves for a two-pass algorithm with a single dose-volume constraint on the PTV. **a** On the first pass, the constraint is met with a margin of about 0.5 Gy. **b** On the second pass, the constraint is met tightly with small gains elsewhere

SCS. Calculations were performed on a cluster with 32-core, 2.20 GHz Intel Xeon E5-4620 CPU and a nVidia TitanX graphics card.

Clinical results: dose-volume constraints Figure 4a depicts the DVH curves of the plan produced without any dose constraints. The PTV curve is shown in red with a dotted vertical line marking its prescribed dose of 66 Gy. The rest are DVHs for the OARs and generic body voxels. This solution to the unconstrained problem already gives a fairly good treatment plan. The DVH of the right cochlea and left parotid are pushed far left, so only 15–20% of their volume exceed 10 Gy, and almost no voxels are dosed above 30 Gy. The spinal cord receives somewhat more radiation, while the worst case is the brain with a slow, nearly linear drop-off to about 75 Gy.

The PTV curve begins to fall at 66 Gy, but does not reach zero until nearly 95 Gy. To reduce this overdosing, we add a dose-volume constraint that limits no more than 20% of the PTV to receive over 70 Gy, as indicated by the red, left-pointing arrow in Fig. 4b, and re-plan the case. The resulting DVH curves are depicted as solid lines in Fig. 4c, while the dashed curves represent the original plan. Under the new plan, the PTV curve has been pushed left at the arrow, and its drop-off around

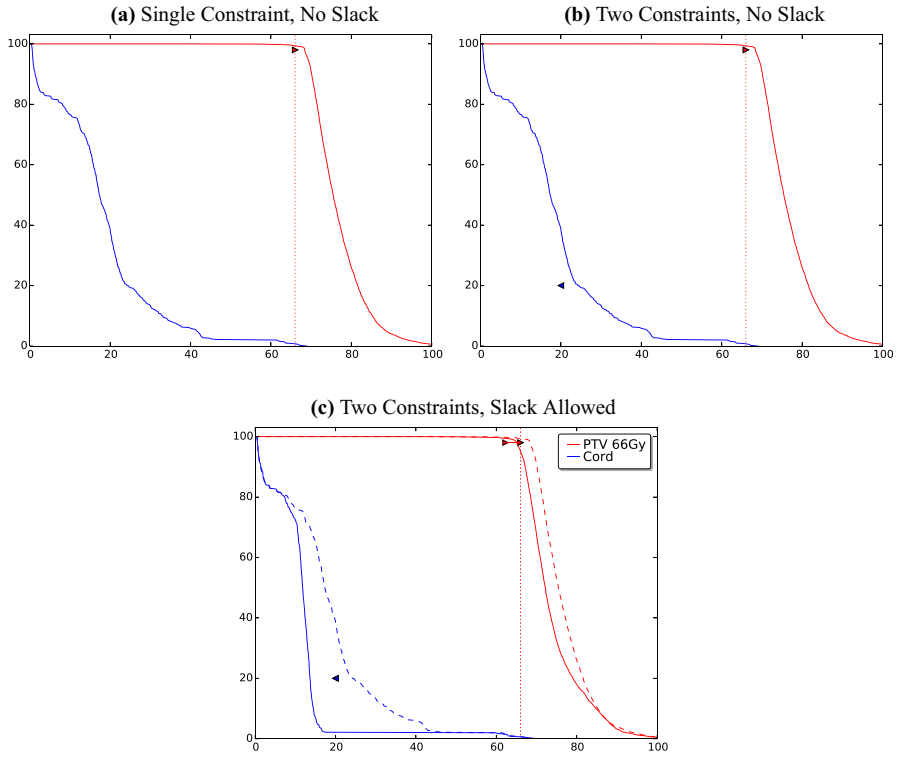

Fig. 6 DVH curves for the PTV and spinal cord. **a** Plan without slack, constraining $D(98) \geq 66$ Gy in the PTV. **b** A constraint $D(20) \leq 20$ Gy is added to the spinal cord, rendering the problem infeasible. **c** Re-plan with slack allowed. The spinal cord constraint is met, but the PTV constraint relaxes by about 3 Gy

66 Gy is steeper, meaning all voxel doses are closer to the prescription. Moreover, our OARs are minimally affected. We have reduced overdosing to the PTV without significantly increasing radiation to other organs.

Clinical results: two-pass algorithm A close inspection of Fig. 4c reveals a gap between the PTV curve and the DVH constraint arrow. This is due to the conservative nature of the convex restriction, which overestimates the number of voxel violations. We can eliminate this gap and improve our overall objective with the two-pass algorithm. Figure 5a shows the plan from the first pass. There is a margin of about 0.5 Gy between the PTV curve and arrow, meaning at most 20% of the PTV receives over 69.5 Gy, a more restrictive solution than required by our constraint $D(20) \leq 70$ Gy.

This margin disappears in Fig. 5b. Here, the dashed lines depict the first pass solution, and the solid lines come from the second pass. The second pass PTV curve falls precisely on the center of the left-facing arrow, meaning the dose-volume constraint is tight. In addition, the DVH curves for the right cochlea and left parotid have shifted leftward, indicating they now receive less radiation. By replacing our convex restriction with exact voxel constraints, we are able to make gains in our OAR clinical objectives while still fulfilling the DVH constraint.

Clinical results: constraints with slack So far, we have specified only one dose constraint. A problem may become infeasible when multiple constraints are enforced, either because the user-supplied bounds are too extreme or the convex restriction too severe in its overestimation of the voxel count. In such cases, we can still produce a plan that approximately conforms to the desired specifications by enabling slack constraints.

Figure 6a depicts a plan created without slack. The single PTV constraint, $D(98) \geq 66$ Gy, is met with margin. However, when we add the OAR constraint $D(20) \leq 20$ Gy, as symbolized by the blue arrow in Fig. 6b, and re-plan the case, the optimizer tells us that the problem is infeasible. It is impossible to meet both (convex restricted) dose-volume constraints exactly. We thus re-plan allowing

Table 1 Prostate FMO prescription

Structure	Target?	Dose (Gy)	Constraints (Gy)
Prostate	Yes	75.6	$D^{\text{avg}} \geq 75.6$
Urethra	No	0	$D^{\text{avg}} < 52.5$
Bladder	No	0	$D(85) < 80$
			$D(75) < 75$
			$D(65) < 70$
			$D(50) < 65$
Rectum	No	0	$D(90) < 75$
			$D(85) < 70$
			$D(50) < 65$
L. Femoral head	No	0	$D(95) < 50$
R. Femoral head	No	0	$D(95) < 50$
Body	No	0	$D^{\text{avg}} < 52.5$

for slack bounds on these constraints. The resulting DVH curves are plotted in Fig. 6c with dotted lines representing the original plan and solid lines for the new plan. The PTV constraint has been relaxed by about 3 Gy, as symbolized by the red arrow shifting left behind the solid red curve to (98, 63). This small concession allows us to satisfy the OAR constraint by a wide margin.

8.2 Problem scaling

Problem instances We assess the performance of our algorithm on a larger prostate FMO problem. This case contains 74,453 voxels × 34,848 beamlets, encompassing a single PTV treated to 75.6 Gy, five OARs with various dose constraints, and generic body voxels. Approximately 226 million (roughly 10.6%) of the entries in the dose matrix are non-zero. In our experiments, we used only a subset of 10,000 beamlets from this matrix.

We plan the case with the prescription detailed in Table 1, which is adapted from the QUANTEC guidelines (Marks et al. 2010). The computational details are the same as in the head-and-neck case. As before, we analyze the results from a single pass and two-pass algorithm with and without slack allowed. We then re-plan the case with only the PTV dose constraint and compare its runtime and OAR overdose to the plans produced from the full prescription.

Timing results Our algorithm produces a plan that satisfies all dose constraints using a single pass with slack enabled. The optimization finishes in 426.9 s, about 2.5× the runtime of the head-and-neck case. A second pass takes approximately the same amount of time and does not result in significantly larger constraint margins. The presence or absence of slack also has little impact on the runtime, which varies by at most 7 s.

If we drop all except the mean dose constraints, the problem collapses into a linear program, greatly decreasing the runtime. The size of this reduction depends on the characteristics of the affected structures and the dose influence matrix. For example, in the head-and-neck case, the runtime falls by 87% to a mere 24 s. Conversely, when adding dose-volume constraints, the initial constraint on a structure will have a greater impact on runtime than subsequent additions.

9 Conclusion

We have developed a convex formulation for the FMO problem that incorporates dose-volume constraints. Our model replaces each exact dose-volume constraint with a convex restriction, which overestimates the number of voxels that violate the clinician's desired threshold. This allows us to solve the problem quickly and efficiently using standard convex optimization algorithms. We also introduce two refinements: a two-pass algorithm and a model with slack. In the former, we improve our initial solution by re-optimizing with the restrictions replaced by bounds on a subset of voxels, enabling us to achieve a better objective that still satisfies the dose-volume constraints. The latter allows for soft bounds and is useful if the restricted

constraints render the problem infeasible. We demonstrate the efficacy of our method on a VMAT head-and-neck case and a prostate case. Our algorithm consistently produces good treatment plans that fulfill all dose constraints when feasible. In problems with infeasible constraints, we are able to generate plans that minimize the dose violation while taking into account clinical goals, allowing clinicians to easily visualize trade-offs and select the plan that is best for the patient.

A variety of extensions to our two-pass algorithm are possible. For instance, we could rewrite the original problem as a mixed-integer linear program and use the solution of the convex restriction to warm start a branch-and-bound solver. More broadly, we could apply this starting point to accelerate any number of iterative approaches in the literature. Dose-volume constraints are often assigned different priorities in practice, and our algorithm may be easily adapted to accommodate such user-defined preferences, either through new penalties, changes in the slack, or additional passes that impose the constraints in a lexicographic order. These hybrid methods, which combine convex approximations with non-convex solution methods, offer an important avenue for future research.

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. This research was supported by the Stanford Graduate Fellowship, Stanford Bio-X Bowes Fellowship, and NIH Grant 5R01CA176553.

References

Adler JR Jr., Chang SD, Murphy MJ, Doty J, Geis P, Hancock SL (1998) The cyberKnife: a frameless robotic system for radiosurgery. Stereotact Funct Neurosurg 69(1–4):124–128

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, Glaser D, Romeijn HE, Dempsey JF (2010) Interior point algorithms: guaranteed optimality for fluence map optimization IMRT. Phys Med Biol 55(18):5467–5482

Aleman DM, Mišić VV, Sharpe MB (2013) Computational enhancements to fluence map optimization for total marrow irradiation using IMRT. Comput Oper Res 40(9):2167–2177

Bedford JL (2009) Treatment planning for volumetric modulated arc therapy. Med Phys 36(11):5128–5138

Bortfeld T, Bürkelbach J, Boesecke R, Schlegel W (1990) Methods of image reconstruction from projections applied to conformation radiotherapy. Phys Med Biol 35(10):1423–1434

Chan TCY, Mahmoudzadeh H, Purdie TG (2014) A robust-CVaR optimization approach with applications to breast cancer therapy. Eur J Oper Res 238(3):876–885

Chen W, Unkelbach J, Trofimov A, Madden T, Kooy H, Bortfeld T, Craft D (2012) Including robustness in multi-criteria optimization for intensity-modulated proton therapy. Phys Med Biol 57(3):591–608

Cho PS, Lee S, Marks RJ II, Oh S, Sutlief SG, Phillips MH (1998) Optimization of intensity modulated beams with volume constraints using two methods: cost function minimization and projections onto convex sets. Med Phys 25(4):435–443

Davino C, Furno M, Vistocco D (2013) Quantile regression: theory and applications. Wiley, New York

Deasy JO (1997) Multiple local minima in radiotherapy optimization problems with dose-volume constraints. Med Phys 24(7):1157–1161

Diamond S, Boyd S (2016) CVXPY: a Python-embedded modeling language for convex optimization. J Mach Learn Res 17(83):1–5

Domahidi A, Chu E, Boyd S (2013) ECOS: an SOCP solver for embedded systems. In: European control conference, pp 3071–3076

Dong P, Lee P, Ruan D, Long T, Romeijn HE, Yang Y, Low D, Kupelian P, Sheng K (2013) 4π non-coplanar liver sbrt: a novel delivery technique. Int J Radiat Oncol Biol Phys 85(5):1360–1366

Ehrgott M, Güler Ç, Hamacher HW, Shao L (2008) Mathematical optimization in intensity modulated radiation therapy. 4OR 6(3):199–262

Glide-Hurst C, Bellon M, Foster R, Altunbas C, Speiser M, Altman M, Westerly D, Wen N, Zhao B, Miften M (2013) Commissioning of the Varian TrueBeam linear accelerator: a multi-institutional study. Med Phys 40(3):031719

Halabi T, Craft D, Bortfeld T (2006a) Dose-volume objectives in multi-criteria optimization. Phys Med Biol 51(15):3809–3818

Halabi T, Craft D, Bortfeld T (2006b) Dose-volume objectives in multi-criteria optimization. Phys Med Biol 51:3809–3818

Hamacher HW, Küfer KH (2002) Inverse radiation therapy planning—a multiple objective optimization approach. Discrete Appl Math 118(1):145–161

Hölder A (2003) Designing radiotherapy plans with elastic constraints and interior point methods. Health Care Manag Sci 6(1):5–16

Langer M, Brown R, Urie M, Leong J, Stracher M, Shapiro J (1990) Large scale optimization of beam weights under dose-volume restrictions. Int J Radiat Oncol Biol Phys 18(4):887–893

Lee E, Fox T, Crocker I (2000) Optimization of radiosurgery treatment planning via mixed integer programming. Med Phys 27(5):995–1004

Lee E, Fox T, Crocker I (2003) Integer programming applied to intensity-modulated radiation therapy treatment planning. Ann Oper Res 119(1–4):165–181

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

Lim G, Cao W (2012) A two-phase method for selecting IMRT treatment beam angles: Branch-and-Prune and local neighborhood search. Eur J Oper Res 217(3):609–618

Mackie TR, Holmes T, Swerdloff S, Reckwerdt P, Deasy JO, Yang J, Paliwal B, Kinsella T (1993) Tomotherapy: a new concept for the delivery of dynamic conformal radiotherapy. Med Phys 20(6):1709–1719

Mageras GS, Mohan R (1993) Application of fast simulated annealing to optimization of conformal radiation treatments. Med Phys 20(3):639–647

Marks LB, Yorke ED, Jackson A, Haken RKT, Constine LS, Eisbruch A, Bentzen SM, Nam J, Deasy JO (2010) Use of normal tissue complication probability (NTCP) models in the clinic. Int J Radiat Oncol Biol Phys 76(3 Suppl):S10–S19

O'Donoghue B, Chu E, Parikh N, Boyd S (2016) Conic optimization via operator splitting and homogeneous self-dual embedding. J Optim Theory Appl 169(3):1042–1068

Oelfke U, Bortfeld T (2001) Inverse planning for photon and proton beams. Med Dosim 26(2):113–124

Oskoorouchi MR, Ghaffari HR, Terlaky T, Aleman DM (2011) An interior point constraint generation algorithm for semi-infinite optimization with health-care application. Oper Res 59(5):1184–1197

Rockafellar R, Uryasev S (2000) Optimization of conditional value-at-risk. J Risk 2:21–42

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 Med Biol 48(21):3521–3542

Romeijn HE, Dempsey J, Li J (2004) A unifying framework for multi-criteria fluence map optimization models. Phys Med Biol 49(10):1991–2013

Romeijn HE, Ahuja RK, Dempsey JF, Kumar A (2006) A new linear programming approach to radiation therapy treatment planning problems. Oper Res 54(2):201–216

Rosen II, Lane RG, Morrill SM, Belli JA (1991) Treatment plan optimization using linear programming. Med Phys 18(2):141–152

Schweikard A, Schlaefer A, Adler J Jr. (2006) Resampling: an optimization method for inverse planning in robotic radiosurgery. Med Phys 33(11):4005–4011

Shepard DM, Ferris MC, Olivera GH, Mackie TR (1999) Optimizing the delivery of radiation therapy to cancer patients. SIAM Rev 41(4):721–744

Shepard DM, Ferris MC, Ove R, Ma L (2000a) Inverse treatment planning for Gamma Knife radiosurgery. Med Phys 27(9):2146–2149

Shepard DM, Olivera GH, Reckwerdt PJ, Mackie TR (2000b) Iterative approaches to dose optimization in tomotherapy. Phys Med Biol 45(1):69–90

Spirou SV, Chui C (1998) A gradient inverse planning algorithm with dose-volume constraints. Med Phys 25(3):321–333

Webb S (1989) Optimization of conformal radiotherapy dose distribution by simulated annealing. Phys Med Biol 34(10):1349–1370

Webb S (1992) Optimization by simulated annealing of three-dimensional, conformal treatment planning for radiation fields defined by a multileaf collimator: II. inclusion of two-dimensional modulation of the X-ray intensity. Phys Med Biol 37(8):1689–1704

Wu Q, Mohan R (2002) Multiple local minima in IMRT optimization based on dose-volume criteria. Med Phys 29(7):1514–1527

Xing L, Chen GTY (1996) Iterative methods for inverse treatment planning. Phys Med Biol 41(10):2107–2123

Xing L, Hamilton RJ, Spelbring D, Pelizzari CA, Chen GTY, Boyer AL (1998) Fast iterative algorithms for three-dimensional inverse treatment planning. Med Phys 25(10):1845–1849

Zarepisheh M, Shakourifar M, Trigila G, Ghomi PS, Couzens S, Abebe A, Noreña L, Shang W, Jiang SB, Zinchenko Y (2013) A moment-based approach for DVH-guided radiotherapy treatment plan optimization. Phys Med Biol 58(6):1869–1887

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

Affiliations

Anqi Fu[1] · Barış Ungun[2] · Lei Xing[3] · Stephen Boyd[1]

Barış Ungun
ungun@stanford.edu

Lei Xing
lei@stanford.edu

Stephen Boyd
boyd@stanford.edu

[1] Department of Electrical Engineering, Stanford University, 350 Serra Mall, Stanford, CA 94305, USA

[2] Department of Bioengineering, Stanford University, 443 Via Ortega, Stanford, CA 94305, USA

[3] Department of Radiation Oncology, Stanford School of Medicine, 875 Blake Wilbur Drive, Stanford, CA 94305, USA