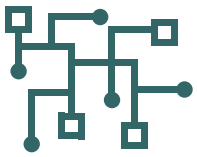


ISE

Industrial and Systems Engineering

An Adaptive Relaxation Technique for Mixed-integer Formulation of Proton Therapy Treatment Optimization

MOHAMMADHOSSEIN MOHAMMADISIAHROUDI[1], POUYA SAMPOURMAHANI[1], WEI ZOU[2], LEI DONG[2], YURIY ZINCHENKO[3], AND TAMÁS TERLAKY[1]

[1]Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA, USA

[2]SDepartment of Radiation Oncology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[3]Department of Mathematics and Statistics, University of Calgary, Calgary, Canada

ISE Technical Report 25T-008

LEHIGH UNIVERSITY

Noname manuscript No.
(will be inserted by the editor)

An Adaptive Relaxation Technique for Mixed-integer Formulation of Proton Therapy Treatment Optimization

Mohammadhossein Mohammadisiahroudi · Pouya Sampourmahani · Wei Zou · Lei Dong · Yuriy Zinchenko · Tamás Terlaky

Received: date / Accepted: date

Abstract Intensity modulated proton therapy (IMPT) using the pencil beam scanning (PBS) technique poses significant computational and optimization challenges due to the large number of decision variables involved and the high sensitivity of treatment outcomes to patient motion. Current commercial treatment planning systems rely heavily on manual, iterative adjustments by clinicians to satisfy dose-volume histogram (DVH) constraints, resulting in a time-consuming and resource-intensive process. Mixed-integer optimization (MIO) formulations, which encode DVH constraints directly, offer the highest plan quality but are often computationally intractable for practical proton therapy planning. In this paper, we propose an adaptive relaxation technique for the MIO formulation. The adaptive relaxation technique significantly reduces computational burden while maintaining high plan quality. Our method adaptively relaxes constraints based on sensitivity analysis, yielding treatment plans with high DVH conformity in just a few minutes of computation time. The proposed approach enables automated and efficient treatment planning with minimal clinical intervention, and so bridging the gap between plan quality and computational feasibility.

Keywords Proton Therapy · Treatment Planning · Mixed-integer Optimization

Mathematics Subject Classification (2020) MSC 90-00 · MSC 90C06 · 90C11 · 90C90

Mohammadhossein Mohammadisiahroudi
Lehigh University
E-mail: mom219@lehigh.edu

1 Introduction

Proton therapy has emerged as a highly precise form of radiotherapy, capable of delivering conformal dose distributions that spare healthy tissues more effectively than conventional photon-based treatments, see e.g., Merchant et al., 2008. This advantage is especially critical in cases like pediatric cancers and brain tumors, where reduced dose to normal tissues can lower the risk of long-term side effects and cognitive impairment.

Modern proton facilities increasingly employ pencil beam scanning (PBS) to modulate intensity across thousands of narrow beamlets (spots). This technique enables intensity-modulated proton therapy (IMPT) with fine-grained control, but it also introduces substantial optimization complexity due to the large number of decision variables (often tens of thousands of beamlet intensities) and the need to account for beam range uncertainties and patient motion.

In particular, PBS plans are highly sensitive to intra-fraction motion: interplay between organ motion and sequential spot delivery can degrade the realized dose distribution if not properly managed. Robust optimization methods have been developed to mitigate such motion and range uncertainties, albeit with increased computational burden, see Buti et al., 2020. As proton therapy becomes more widely adopted, as stated by Nystrom et al., 2020, there is a pressing need for treatment planning approaches that can achieve the full dosimetric potential of IMPT while meeting clinical constraints and remaining computationally efficient. This work addresses a central challenge toward that goal: how to satisfy complex dose-volume constraints in IMPT optimization in a tractable, automated way, without planner's interventions and sacrificing plan quality.

In radiation therapy treatment planning, clinical requirements are often expressed in terms of dose-volume histogram (DVH) constraints. These constraints limit the fraction (volume) of a structure receiving a dose above or below specified thresholds. For example, an organ-at-risk (OAR) constraint could require that no more than 20% of its volume is received beyond 30 Gy, while a target constraint could require that at least 95% of the tumor volume gets at least the prescription dose. Such DVH-based criteria have become standard in modern treatment planning and are employed in all major commercial planning systems (e.g. Varian Eclipse, RayStation) to ensure that plans are clinically acceptable. However, these constraints pose a non-convex optimization problem because they involve global dose distribution properties (volume percentages), rather than simple per-voxel dose limits.

Contemporary planning software typically handles DVH constraints indirectly by using soft constraints or penalty terms: the planner specifies a set of weighting parameters in a convex objective function to push the solution toward satisfying the DVH criteria. For instance, one can add large penalty costs for doses exceeding a threshold in a given organ, or use smooth approximations of DVH metrics in the objective function. Pioneering implementations of inverse planning optimized such weighted cost functions with gradient-based

or quasi-Newton solvers, requiring the user to iteratively adjust the weights to achieve an acceptable balance between tumor coverage and OAR sparing Wu and Mohan, 2000. In practice, this trial-and-error tuning of objectives is time-consuming and relies heavily on the planner's experience. Multiple optimization iterations are often needed, as planners adjust dose goals and rerun the solver until all DVH criteria (or as many as possible) are met. This manual process contributes significantly to treatment planning time and can lead to variability in plan quality between institutions and planners Hussein et al., 2018.

In short, while convex fluence map optimization (FMO) methods (e.g. linear or quadratic programming models) are fast and widely used, they do not inherently enforce DVH metrics, and achieving those metrics requires substantial user intervention in current clinical workflows. To reduce human effort and improve consistency, researchers have explored various automated planning strategies. One approach is knowledge-based planning (KBP), where models derived from past clinical plans predict achievable DVH targets or even complete dose distributions for new patients. This can guide the optimization process to appropriate solutions without extensive trial-and-error. Other approaches leverage multi-criteria optimization and heuristics to automate decision-making. For example, the iCycle system integrates beam angle selection with fluence optimization in a multicriterial framework, automatically producing high-quality IMRT plans that satisfy a prioritized list of clinical criteria, see Breedveld et al., 2012. More recently, fully automated pipelines combining machine learning and optimization have been developed; for instance, McIntosh et al., 2017 used a voxel-level dose prediction model to drive an automatic inverse planning process for head-and-neck cases with minimal planner input.

In proton therapy, Taasti et al., 2020 introduced an automated planning method based on constrained hierarchical optimization, which sequentially applies robust criteria to ensure target coverage and OAR sparing in a prioritized manner. These innovations demonstrate the potential of algorithmic automation; however, most still rely on solving a series of conventional optimization problems with soft constraints or approximations. Thus, they do not completely eliminate the possibility of unmet DVH constraints – rather, they aim to reduce the guesswork and speed up the planning process. A recent review by Hussein et al., 2018 highlights the significant progress in automation, but also notes that further improvements are needed for widespread clinical adoption. In particular, achieving truly automated DVH satisfaction remains challenging. An alternative paradigm is to incorporate DVH constraints directly into the optimization model as hard constraints. This leads to a MIO formulation, since one can introduce binary decision variables to model the selection of which voxels are allowed to exceed a given dose threshold. Early work in operations research showed that using such hard constraints in inverse planning can systematically produce plans that meet clinical goals without iterative re-planning: for example, Ferris et al., 2006 demonstrated that a properly for-

mulated MIO can eliminate the impractical trial-and-error tuning process and yield solutions meeting all prescription criteria by construction.

Several researchers have since applied MIO to radiotherapy problems to handle various planning aspects. Romeijn et al., 2006 presented a formulation for intensity-modulated radiation therapy (IMRT) that explicitly enforced dose–volume constraints via linear constraints and binary variables. Their results confirmed that an optimal solution of the MIO achieves superior or equal plan quality compared to conventional methods, whenever the latter one struggled to satisfy certain DVH limits. Subsequent studies refined and extended the approach, for instance, Rocha et al., 2012 used binary optimization to discretize optimal fluence into deliverable levels for IMRT, and Tuncel et al., 2012 derived strong valid inequalities that tighten the MIO formulation of fluence map optimization with DVH restrictions, improving solver performance. In the proton therapy domain, researchers have begun to investigate similar MIO-based techniques.

Zaghian et al., 2014 proposed an iterative linear programming approach that adjusts voxel constraints to approximate dose-volume limits, effectively satisfying DVH criteria after a few re-solves of a fluence optimization problem. More recently, in contrast to a conventional optimizer, Liu et al. (2017) incorporated dose-volume constraints directly into an IMPT planning model and showed that it can generate plans with excellent target coverage and OAR sparing without user tuning. However, a critical drawback was evident: the MIO models tend to be computationally intractable for realistic clinical cases. Solving a full-scale IMPT MIO can require thousands of binary variables (one per voxel or per dose increment) in addition to the continuous beamlet intensity variables, leading to prohibitively long solve times (hours or more) or even solver failure due to memory limits. Indeed, Liu et al. had to restrict their study to a small number of beams and employed simplifying assumptions to make the problem manageable, and even then the solution times were significant. In general, due to these computational challenges, despite the theoretical appeal of MIO-based planning, very few treatment planning systems or clinics use it in practice. The research community recognizes that new optimization techniques are needed to bridge this gap between ideal plan quality and practical solvability in a clinical time frame.

In summary, there exists a research gap between the high-quality treatment plans that advanced optimization models, especially mixed-integer formulations, can produce and the efficiency demanded by clinical workflows. On one hand, purely continuous optimization approaches are fast but rely on iterative, manual tuning to approximately meet DVH goals. On the other hand, full mixed-integer approaches can meet DVH constraints exactly and automatically, but are often too slow for routine use. Attempts to mitigate this, such as pre-processing the problem or tightening the formulation, have only partially alleviated the difficulty. For example, our recent work explored problem size reduction techniques, i.e., cropping irrelevant regions of the patient scan, sparsifying negligible dose contributions, and aggregating voxels in

low-gradient regions, to speed up proton plan optimization, see Mohammadisiahroudi et al., 2024.

While these techniques reduced computation in the continuous optimization setting, a full MIO with thousands of DVH-related constraints was still far from real-time solvable. Thus, a new algorithmic approach is needed to retain the benefits of MIO (guaranteed constraint satisfaction and optimal trade-offs) while dramatically cutting down the computation time. In this paper, we propose an adaptive relaxation technique to achieve this goal. The key idea is to start with a relaxed mixed-integer model that does not include all DVH constraints explicitly, solve a series of simpler problems, and based on solution feedback iteratively introduce or tighten constraints only as needed. In essence, our method automatically identifies the most critical subset of voxels or constraints that govern DVH violations and focuses the combinatorial search on those, while relaxing less critical constraints to keep the problem solvable. This is somewhat analogous to a cutting-plane or column-generation method, but applied to dose-volume constraints: we add back constraints (or binary variables) adaptively rather than include the full set from the outset. A sensitivity analysis approach, inspired by methods like interior point sensitivity analysis, is used to determine which constraint relaxations cause the most deviation from DVH targets. By progressively refining the solution in a series of iterations, the algorithm converges to a plan that satisfies all DVH criteria with high precision, yet each iteration remains fast.

We demonstrate that, for typical IMPT cases, our approach can produce clinically equivalent, or superior, plans in a matter of minutes, whereas a naive MIO approach would be computationally prohibitive. The result is a planning framework that bridges the gap between automation and optimality: it requires minimal human intervention (no trial-and-error tuning) and yields high-quality plans that strictly meet DVH-based prescriptions, all within a runtime suitable for clinical practice.

The rest of the paper is structured as follows. In Section 2, the MIO model is presented. Section 3 outlines the proposed adaptive relaxation technique. The results of the numerical experiments are reviewed in Section 4. Finally, Section 5 concludes the paper.

2 Proton Therapy Treatment Optimization Models

The general approach in our treatment optimization models is to minimize the deviation of dose on the target from the prescribed dose, while ensuring that the dose received by OARs does not exceed clinical constraints specified via DVH limits. Let V denote the set of voxels in the patient domain, and O the set of organs-at-risk (OARs).

Let $V_T \subseteq V$ denote the set of voxels in the target volume, and for each $o \in O$, let $V_o \subseteq V$ be the set of voxels belonging to OAR o. Let B denote the set of beamlets available in the treatment planning. For each pair (i, j) where $i \in V$ and $j \in B$, let d_{ij} be the dose delivered to voxel i by unit intensity

of beamlet j. The matrix $D \in \mathbb{R}^{|V|\times|B|}$, known as the dose-influence matrix, stores these values.

Our goal is to determine optimal beamlet intensities x_j for all $j \in B$. The accumulated dose delivered to voxel i is then defined by:

$$\mathcal{D}_i = \sum_{j\in B} d_{ij}x_j.$$

We use MatRad (Wieser et al., 2017), an open-source radiotherapy planning platform, to compute the dose-influence matrix efficiently. We now introduce two optimization formulations with increasing model complexity.

2.1 Linear Optimization (LO) Model

The linear optimization model seeks to minimize total deviation from target dose bounds using auxiliary variables, while maintaining per-voxel upper bounds for OARs. Let $t_{\max}$ and $t_{\min}$ be the upper and lower bounds for dose in the target volume, and m_o the upper bound for OAR $o \in O$. The model is given as:

$$\min \quad \sum_{i\in V_T} s_i + \sum_{i\in V_T} s'_i \tag{1}$$

$$\text{s.t.} \quad \sum_{j\in B} d_{ij}x_j - s_i \le t_{\max} \quad \forall i \in V_T \tag{2}$$

$$\sum_{j\in B} d_{ij}x_j + s'_i \ge t_{\min} \quad \forall i \in V_T \tag{3}$$

$$\sum_{j\in B} d_{ij}x_j \le m_o \quad \forall i \in V_o,\ \forall o \in O \tag{4}$$

$$x_j, s_i, s'_i \ge 0 \quad \forall j \in B, i \in V_T. \tag{5}$$

Here, s_i and s'_i are auxiliary slack variables capturing violations above the upper target threshold and below the lower target threshold, respectively. The objective (1) minimizes the total deviation from the prescribed dose range in the target. Constraint (2) ensures the dose per target voxel does not exceed $t_{\max}$ unless compensated by slack. Constraint (3) ensures coverage above $t_{\min}$. Constraint (4) provides pointwise upper bounds for each voxel in OARs. This model can be solved efficiently using off-the-shelf LP solvers (*Gurobi* v11).

2.2 Mixed-Integer Linear Optimization (MIO) Model

While efficient, the LO model fails to guarantee volumetric control, which is clinically required through DVH specifications. For instance, an OAR constraint may state that at most w_o fraction of its voxels receive dose above

threshold u_o. To model such volumetric constraints, we formulate a MIO model that introduces binary variables.

Let $n_o = |V_o|$ be the number of voxels in OAR o. For each $i \in V_o$, define a binary variable y_i indicating whether voxel i is allowed to exceed the dose threshold u_o. Let M denote a sufficiently large upper bound on dose.

$$\min \quad \sum_{i \in V_T} s_i + \sum_{i \in V_T} s'_i \tag{6}$$

$$\text{s.t.} \quad \sum_{j \in B} d_{ij} x_j - s_i \leq t_{\max} \quad \forall i \in V_T \tag{7}$$

$$\sum_{j \in B} d_{ij} x_j + s'_i \geq t_{\min} \quad \forall i \in V_T \tag{8}$$

$$\sum_{j \in B} d_{ij} x_j \leq u_o + (M - u_o) y_i \quad \forall i \in V_o,\ \forall o \in O \tag{9}$$

$$\sum_{i \in V_o} y_i \leq w_o n_o \quad \forall o \in O \tag{10}$$

$$x_j, s_i, s'_i \geq 0,\ y_i \in \{0,1\} \quad \forall j \in B, i \in V_T, i \in V_o. \tag{11}$$

Constraints (7)–(8) handle target dose consistency, identical to the LO model. Constraints (9) and (10) model DVH bounds. Specifically, (9) enforces that voxel $i \in V_o$ receives dose at most u_o if $y_i = 0$, or at most M if $y_i = 1$, allowing a controlled violation. Constraint (10) ensures that the number of violations does not exceed the allowed fraction w_o.

This structure ensures that DVH constraints are satisfied exactly. If multiple DVH thresholds are specified (e.g., for multiple dose-volume points), we replicate constraints (9) and (10) with separate binary variables and dose thresholds.

Despite its clinical appeal, the MIO model introduces a large number of binary variables, making it in reasonable time computationally intractable for full-resolution clinical instances. Solving such models using branch-and-bound or branch-and-cut methods can be time-consuming or infeasible for large-scale problems. The next section proposes an adaptive relaxation technique to reduce this computational burden while maintaining plan quality.

3 Adaptive Relaxation Technique

In this section, we present an adaptive relaxation technique designed to solve the MIO problem introduced in Section 2 more efficiently. The central idea is to avoid solving the full MIO model with all binary variables by identifying and activating only the most critical ones. The activation procedure utilized dual sensitivity analysis information. This is done iteratively, refining the set of voxels that violate DVH constraints, until the solution stabilizes.

Let $m_o = \lfloor u_o n_o \rfloor$ for each $o \in O$ be the maximum allowable number of voxels in OAR o that can exceed the dose threshold u_o.

3.1 Algorithm Description

Algorithm 1 Adaptive Relaxation for DVH-Constrained Optimization

1: Solve initial linear problem with softened OAR constraints:

$$\begin{aligned}
\min \quad & \sum_{i \in V_T} s_i + \sum_{i \in V_T} s_i' \\
\text{s.t.} \quad & \sum_{j \in B} d_{ij} x_j - s_i \leq t_{\max}, \quad \forall i \in V_T \\
& \sum_{j \in B} d_{ij} x_j + s_i' \geq t_{\min}, \quad \forall i \in V_T \\
& \sum_{j \in B} d_{ij} x_j \leq u_o, \quad \forall i \in V_o, \ \forall o \in O \\
& x_j, s_i, s_i' \geq 0
\end{aligned}$$

2: Compute shadow prices (dual values) sp_i of the OAR constraints for all $i \in V_o$, $o \in O$
3: For each $o \in O$, select m_o voxels with highest sp_i into OV_o (overdose candidates), and set $LV_o = V_o \setminus OV_o$
4: **repeat**
5: Solve the following relaxed MIO problem:

$$\begin{aligned}
\min \quad & \sum_{i \in V_T} s_i + \sum_{i \in V_T} s_i' \\
\text{s.t.} \quad & \sum_{j \in B} d_{ij} x_j - s_i \leq t_{\max}, \quad \forall i \in V_T \\
& \sum_{j \in B} d_{ij} x_j + s_i' \geq t_{\min}, \quad \forall i \in V_T \\
& \sum_{j \in B} d_{ij} x_j \leq u_o, \quad \forall i \in LV_o, \ o \in O \\
& \sum_{j \in B} d_{ij} x_j \leq M, \quad \forall i \in OV_o, \ o \in O \\
& x_j, s_i, s_i' \geq 0
\end{aligned}$$

6: Update shadow prices sp_i for $i \in V_o$ based on dual values
7: Recompute OV_o and LV_o based on updated sp_i
8: **until** convergence (no change in OV_o)

3.2 Convergence and Optimality

We now provide theoretical guarantees for monotonicity and optimality of the proposed method.

Theorem 1 (Monotonic Improvement) *Let z^k denote the optimal objective value at iteration k. Then the objective sequence $\{z^k\}$ is monotonically non-increasing.*

Proof Assume that the algorithm progresses to iteration $k+1$ after iteration k. Then there exists at least one voxel $i \in LV_o^k$ and one voxel $j \in OV_o^k$ such that their roles are swapped in LV_o^{k+1} and OV_o^{k+1}. Since voxel i had a larger shadow price than j ($sp_i > sp_j$), replacing constraint $\sum_j d_{ij}x_j \leq u_o$ with a relaxed upper bound M for voxel i results in potential decrease in objective cost, while tightening the constraint for voxel j, whose sensitivity is lower, does not impact optimality significantly.

The change in objective is lower bounded by

$$z^k - z^{k+1} \geq (M - u_o)(sp_i - sp_j) \geq 0,$$

ensuring monotonicity.

Theorem 2 (Global Optimality) *The final solution of the adaptive relaxation algorithm is an optimal solution to the MIO formulation presented in Section 2.*

Proof Let LV_o^*, OV_o^* be the final sets identified by the algorithm for each OAR o. Suppose, for contradiction, that there exists a feasible MIO solution with sets LV_o', OV_o' that achieves a strictly better objective $z' < z^*$.

Since the MIO model explicitly limits the number of overdosed voxels to m_o, both partitions satisfy $|OV_o^*| = |OV_o'| = m_o$. Therefore, there must exist voxels $i \in OV_o^*, j \in LV_o^*$ such that $i \in LV_o', j \in OV_o'$. In other words, the optimal MIO solution would swap voxels i and j.

But since the algorithm terminated with $sp_i < sp_j$, switching them would contradict the greedy choice made in the final iteration. Moreover, this implies that such a swap cannot improve the objective, contradicting $z' < z^*$. Hence, the final solution is optimal for the MIO problem.

4 Numerical Experiments

We evaluate the performance of our proposed adaptive relaxation technique (ART) using a series of realistic clinical cases. Our implementation is carried out in both MATLAB and Python, with code available at `https://github.com/psm-optimizes/IMPT-SuccRelax` for public access and reproducibility.

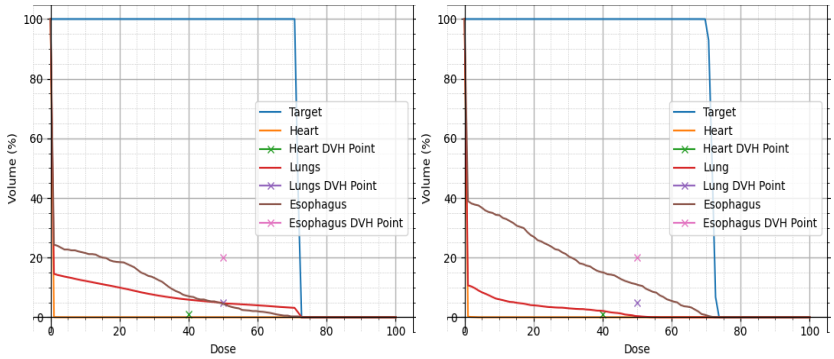
All numerical experiments were conducted on a workstation equipped with dual Intel Xeon® E5-2630 CPUs (2.20 GHz, 20 cores total) and 64 GB of RAM. The dataset includes six lung cancer patients and three brain tumor patients. Each case involves its corresponding clinical prescription and imaging data, anonymized for this IRB-approved retrospective study.

The lung patient plans were created based on average CT images reconstructed from 4DCT scans. To mitigate tumor motion, patients were treated with abdominal compression, restricting motion to under 8 mm. Lung patients were prescribed 2 Gy per fraction for a total of 35 fractions, while brain patients received 2 Gy per fraction over 30 fractions. The clinical contours and dosimetric goals for all OARs were applied consistently across all experiments.

To reduce problem dimensionality and improve computational tractability, we employed a set of preprocessing techniques from Mohammadisiahroudi et al., 2024. These include CT image truncation (before dose calculation), followed by sparsification and inner voxel aggregation (after dose calculation). The order of these preprocessing steps may be adjusted to optimize runtime without impacting dosimetric accuracy.

Dose-influence matrices were calculated using MatRad's generic dose engine on high-resolution CT volumes of size $(512, 512, 370)$ with voxel spacing $(1\text{ mm}, 1\text{ mm}, 3\text{ mm})$. After preprocessing, the resulting sparse dose matrices were used as input to the ART optimization procedure. The optimized spot intensities were then re-applied to the original full-resolution CT images to reconstruct the final dose distribution. The total time for dose calculation and data preparation ranged from 2 to 5 minutes per patient and was not the dominant component of runtime.

In our ART implementation, we utilized Gurobi's interior point (barrier) method to solve the initial linear optimization (LO) model. For subsequent iterations of the adaptive relaxation loop, Gurobi's dual simplex solver was employed, as it efficiently warm-starts from the previous iteration's solution. This hybrid strategy significantly reduces the total computational time.

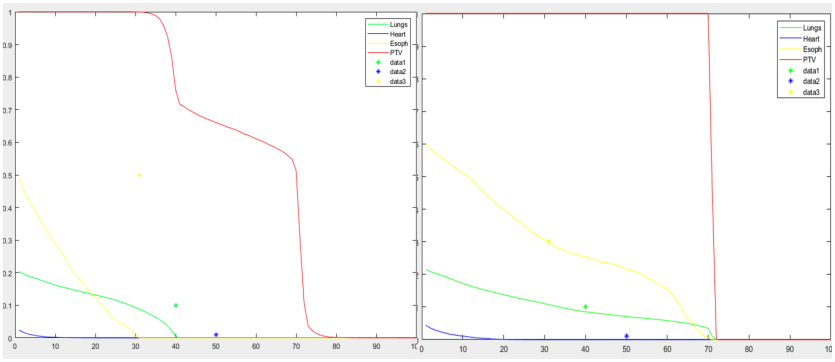


(a) Total time: 10.61 minutes (b) Total time: 11.63 minutes

Fig. 1: DVH curves for two lung patients obtained by ART.

Figure 1 shows dose-volume histogram (DVH) curves for two lung patients optimized using ART. Despite the challenging nature of IMPT planning for thoracic tumors, owing to large motion uncertainties and anatomical complexity, our method efficiently produced highly conformal treatment plans in under 12 minutes.

In Figure 2, we illustrate the progressive refinement of the dose distribution by ART. The first iteration is conservative, strictly limiting OAR doses, which leads to insufficient target coverage. In the second iteration, ART selectively relaxes a small subset of OAR constraints, informed by dual sensitivity val-



(a) First Iteration (b) Second Iteration

Fig. 2: DVH curves across two ART iterations for a lung patient.

ues, enabling the algorithm to recover target coverage while maintaining OAR constraints. The final DVH curves confirm that all clinical objectives are met.

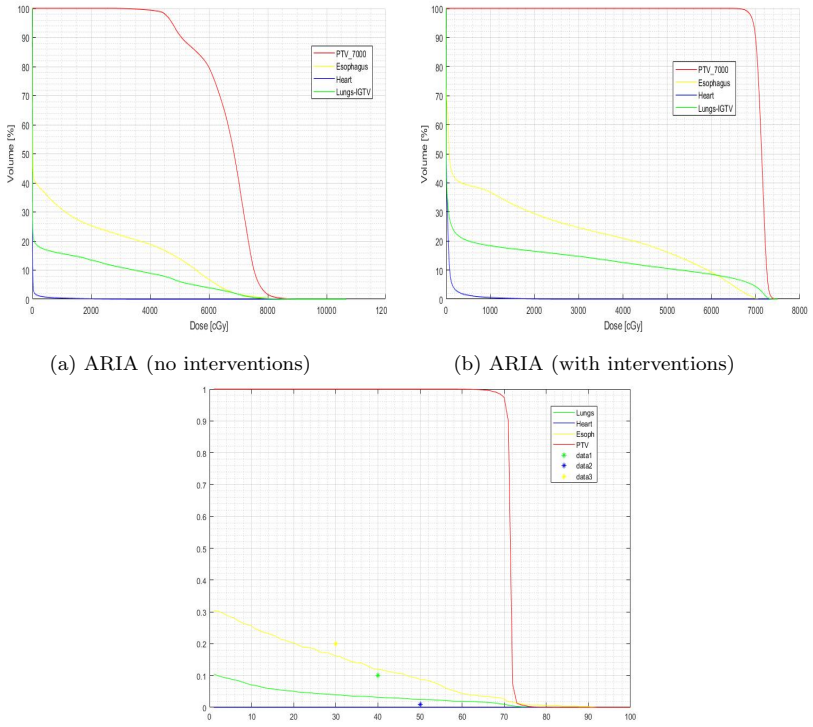
As seen in Figure 3, with no human intervention ART is able to achieve dosimetric outcomes comparable to, or better than, manually tuned clinical plans. Figure 3c shows a final DVH curve from ART that meets all prescription criteria with a total runtime under 10 minutes, suggesting that our method could streamline clinical planning workflows.

Finally, Figure 4 compares the dose distributions from three optimization strategies: linear optimization (LO), second-order cone optimization (SOCO), and mixed-integer optimization (MIO). Both SOCO and MIO generate highly conformal plans that satisfy clinical criteria. In contrast, LO fails to enforce volumetric constraints adequately, particularly for OARs, leading to suboptimal target coverage and excess OAR irradiation.

Together, these results demonstrate that our adaptive relaxation technique enables fast, automated, and high-quality treatment planning. The ART generated plans are comparable to state-of-the-art methods, and so suitable for real-world clinical adaptation.

5 Conclusion

This paper presents a novel adaptive relaxation technique (ART) for solving large-scale mixed-integer optimization (MIO) models in intensity-modulated proton therapy (IMPT) treatment planning. We began by reviewing the challenges in modern proton therapy planning, particularly those related to the exact enforcement of dose-volume histogram (DVH) constraints. Traditional linear optimization approaches struggle to guarantee such volumetric constraints, while MIO models, though accurate, are often computationally intractable for realistic clinical cases.



(a) ARIA (no interventions)

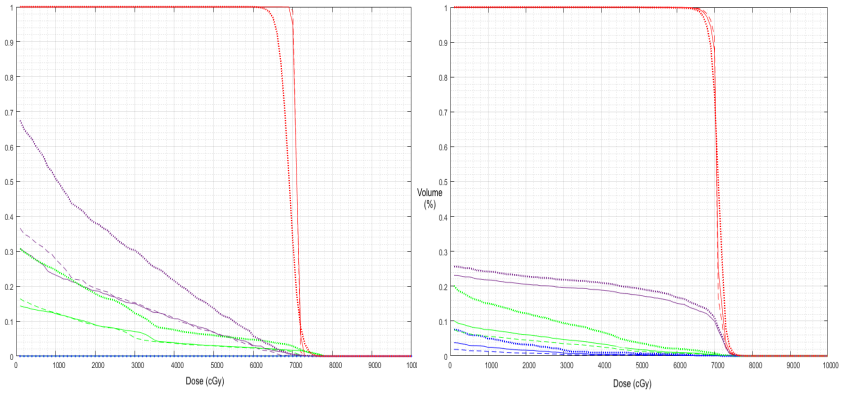
(b) ARIA (with interventions)

(c) Proposed ART approach

Fig. 3: Comparison of DVH curves for one lung patient: (a) clinical ARIA plan without planner tuning, (b) ARIA plan with manual interventions, and (c) ART-generated plan.

To address this gap, we developed ART, a sensitivity-guided algorithm that iteratively refines the feasible set by identifying and enforcing only the most critical DVH constraints. By leveraging dual information from linear relaxations, ART adaptively determines which voxels require strict constraint enforcement, thereby significantly reducing the number of binary variables active in any given iteration. Theoretical guarantees of monotonic improvement and optimality were established, and the method was implemented using MATLAB and Python.

Through comprehensive numerical experiments using lung and brain cancer cases, ART demonstrated strong performance in generating high-quality treatment plans in minutes. Compared to existing clinical systems and optimization strategies, ART achieved comparable or superior dosimetric outcomes with minimal human intervention. In particular, the method delivered



(a) Lung Patient 1 (b) Lung Patient 2

Fig. 4: DVH comparisons across three models: LO (dotted), SOCO (solid), and MIO (dashed). Color legend: Target, Esophagus, Lungs, Heart.

plans that satisfied all DVH constraints while maintaining efficient runtimes, indicating that the method is suitable for integration into clinical workflows.

Although this study focused on deterministic planning, many practical IMPT scenarios involve uncertainty, such as organ motion and setup errors. Classical robust optimization formulations that account for such uncertainties typically result in even larger MIO models. Our proposed ART framework can be naturally extended to these robust settings, offering a tractable alternative to solving full robust MIO models.

Another promising direction for future work is the integration of beam angle optimization (BAO) into the IMPT planning process. BAO introduces an additional combinatorial layer to the problem, further increasing the size and complexity of the optimization model. Adapting ART to handle BAO alongside DVH constraints could provide a scalable and automated solution for holistic treatment planning.

Overall, ART opens new avenues for efficient, accurate, and fully automated treatment planning in proton therapy, and potentially other forms of radiation therapy that are formulated as large scale optimization problems with complex constraints. Future research will focus on extending the algorithm to robust and multi-criteria formulations and validating its clinical impact through prospective studies.

6 Aknowledgement

This work is supported by Varian Medical Systems, Inc. Project: *Fully robust integrated fluence map and beam-angle selection optimization for IMPT.*

References

Breedveld, Sebastiaan, Pascal RM Storchi, Peter WJ Voet, and Ben JM Heijmen (2012). "iCycle: Integrated, multicriterial beam angle, and profile optimization for generation of coplanar and noncoplanar IMRT plans". In: *Medical Physics* 39.2, pp. 951–963.

Buti, Gregory, Kevin Souris, Ana M Barragán Montero, Marie Cohilis, John A Lee, and Edmond Sterpin (2020). "Accelerated robust optimization algorithm for proton therapy treatment planning". In: *Medical Physics* 47.7, pp. 2746–2754.

Ferris, Michael C, Robert R Meyer, and Warren D'Souza (2006). "Radiation treatment planning: Mixed integer programming formulations and approaches". In: *Handbook on Modelling for Discrete Optimization*. Springer, pp. 317–340.

Gurobi (v11). Version v11. Gurobi Optimization, LLC. URL: `https://www.gurobi.com`.

Hussein, Mohammad, Ben JM Heijmen, Dirk Verellen, and Andrew Nisbet (2018). "Automation in intensity modulated radiotherapy treatment planning—a review of recent innovations". In: *The British Journal of Radiology* 91.1092, p. 20180270.

McIntosh, Chris, Mattea Welch, Andrea McNiven, David A Jaffray, and Thomas G Purdie (2017). "Fully automated treatment planning for head and neck radiotherapy using a voxel-based dose prediction and dose mimicking method". In: *Physics in Medicine & Biology* 62.15, pp. 5926–3944.

Merchant, Thomas E, Chia-ho Hua, Hemant Shukla, Xiaofei Ying, Simeon Nill, and Uwe Oelfke (2008). "Proton versus photon radiotherapy for common pediatric brain tumors: comparison of models of dose characteristics and their relationship to cognitive function". In: *Pediatric Blood & Cancer* 51.1, pp. 110–117.

Mohammadisiahroudi, Mohammadhossein, Jennifer Wei Zou, Lei Dong, Yuriy Zinchenko, and Tamás Terlaky (2024). "On Optimization Challenges in Proton Therapy Treatment Planning". In: *AAPM 66th Annual Meeting & Exhibition*. AAPM.

Nystrom, Hakan, Maria Fuglsang Jensen, and Petra Witt Nystrom (2020). "Treatment planning for proton therapy: what is needed in the next 10 years?" In: *The British Journal of Radiology* 93.1107, p. 20190304.

Rocha, Humberto, Joana Matos Dias, Brigida Costa Ferreira, and M d C Lopes (2012). "Discretization of optimal beamlet intensities in IMRT: a binary integer programming approach". In: *Mathematical and Computer Modelling* 55.7-8, pp. 1969–1980.

Romeijn, H Edwin, Ravindra K Ahuja, James F Dempsey, and Arvind Kumar (2006). "A new linear programming approach to radiation therapy treatment planning problems". In: *Operations Research* 54.2, pp. 201–216.

Taasti, Vicki T, Linda Hong, Joseph O Deasy, and Masoud Zarepisheh (2020). "Automated proton treatment planning with robust optimization using

constrained hierarchical optimization". In: *Medical Physics* 47.7, pp. 2779–2790.

Tuncel, Ali T, Felisa Preciado, Ronald L Rardin, Mark Langer, and Jean-Philippe P Richard (2012). "Strong valid inequalities for fluence map optimization problem under dose-volume restrictions". In: *Annals of Operations Research* 196.1, pp. 819–840.

Wieser, Hans-Peter, Eduardo Cisternas, Niklas Wahl, Silke Ulrich, Alexander Stadler, Henning Mescher, Lucas-Raphael Müller, Thomas Klinge, Hubert Gabrys, Lucas Burigo, et al. (2017). "Development of the open-source dose calculation and optimization toolkit MatRad". In: *Medical Physics* 44.6, pp. 2556–2568.

Wu, Qiuwen and Radhe Mohan (2000). "Algorithms and functionality of an intensity modulated radiotherapy optimization system". In: *Medical Physics* 27.4, pp. 701–711.

Zaghian, Maryam, Gino Lim, Wei Liu, and Radhe Mohan (2014). "An automatic approach for satisfying dose-volume constraints in linear fluence map optimization for IMPT". In: *Journal of Cancer Therapy* 5.2, p. 198.