An orthogonal matching pursuit optimization method for solving minimum-monitor-unit problems: applications to proton IMPT, ARC and FLASH

Ya-Nan Zhu[1], **Xiaoqun Zhang**[1], **Yuting Lin**[2], **Chris Lominska**[2], **Hao Gao**[2]

[1]Institute of Natural Sciences and School of Mathematics, Shanghai Jiao Tong University, China

[2]Department of Radiation Oncology, University of Kansas Medical Center, USA

Abstract

Background: The intensities (i.e., number of protons in monitor unit (MU)) of deliverable proton spots need to be either zero or meet a minimum-MU (MMU) threshold, which is a nonconvex problem. Since the dose rate is proportionally associated with the MMU threshold, higher-dose-rate proton radiation therapy (RT) (e.g., efficient intensity modulated proton therapy (IMPT) and ARC proton therapy, and high-dose-rate-induced FLASH effect needs to solve the MMU problem with larger MMU threshold, which however makes the nonconvex problem more difficult to solve.

Purpose: This work will develop a more effective optimization method based on orthogonal matching pursuit (OMP) for solving the MMU problem with large MMU thresholds, compared to state-of-the-art methods, such as alternating direction method of multipliers (ADMM), proximal gradient descent method (PGD), or stochastic coordinate descent method (SCD).

Methods: The new method consists of two essential components. First, the iterative convex relaxation method is used to determine the active sets for dose-volume planning constraints and decouple the MMU constraint from the rest. Second, a modified OMP optimization algorithm is used to handle the MMU constraint: the non-zero spots are greedily selected via OMP to form the solution set to be optimized, and then a convex constrained subproblem is formed and can be conveniently solved to optimize the spot weights restricted to this solution set via OMP. During this iterative process, the new non-zero spots localized via OMP will be adaptively added to or removed from the optimization objective.

Results: The new method via OMP is validated in comparison with ADMM, PGD and SCD for high-dose-rate IMPT, ARC, and FLASH problems of large MMU thresholds, and the results suggest that OMP substantially improved the plan quality from PGD, ADMM and SCD in terms of both target dose conformality (e.g., quantified by max target dose and conformity index) and normal tissue sparing (e.g., mean and max dose). For example, in the brain case, the max target dose for IMPT/ARC/FLASH was 368.0%/358.3%/283.4% respectively for PGD, 154.4%/179.8%/150.0% for ADMM, 134.5%/130.4%/123.0% for SCD, while it was <120% in all scenarios for

xqzhang@sjtu.edu.cn .

OMP; compared to PGD/ADMM/SCD, OMP improved the conformity index from 0.42/0.52/0.33 to 0.65 for IMPT and 0.46/0.60/0.61 to 0.83 for ARC.

Conclusions: A new OMP-based optimization algorithm is developed to solve the MMU problems with large MMU thresholds, and validated using examples of IMPT, ARC, and FLASH with substantially improved plan quality from ADMM, PGD, and SCD.

Keywords

orthogonal matching pursuit; proton therapy; IMPT; ARC; FLASH

---

1. Introduction

The radiation therapy (RT) is to deliver the tumoricidal dose for tumor controls while at the same time minimize radiation-induced normal tissue toxicities, for which pencil beam scanning (PBS) proton RT has become increasingly popular due to its conformal dose to tumor targets and less integral dose to normal tissues with the introduction of intensity modulated proton therapy (IMPT) [2].

For the presentation purpose, the treatment plan optimization problem (e.g., IMPT) considered in this work is formulated mathematically as to solve the following optimization problem

$$\min_{x \in \mathbb{R}^{N_x}, x \geq 0} f(x),$$

(1)

where x denotes the spot weight with the non-negative constraint, N_x is the number of spots, and $f(x)$ is the sum of planning objectives. The unit of x is the so-called monitor unit (MU), i.e., the number of protons.

For proton spots to be deliverable, there exists a machine-specific minimum-MU (MMU) threshold G_{min}, such that the number of protons delivered per spot must be at least G_{min}. On the other hand, it is desirable to have a treatment plan with large MMU threshold g ($g \; G_{min}$) to be satisfied by all proton spots, since larger value of g corresponds to higher dose rate. This will be useful to improve the delivery efficiency of IMPT [2] and proton ARC therapy [13], such as the recent push for lung cancer treatment with a single-breath-hold delivery of IMPT [41]. This will also be needed for FLASH [25], which demands the ultra-high dose rate in order to induce the biological FLASH effect to realize the full biological efficacy of FLASH for normal tissue protection. The core optimization problem for achieving these applications is the so-called the MMU optimization problem

$$\min_{x \in \mathbb{R}^{N_x}} f(x)$$
$$s.t. \; x_j \in \{0\} \cup [g, +\infty), \; j \leq N_x$$

(2)

The MMU constraint in Eq. (2) is nonconvex and states that the spot weight x is nonnegative and at least g if x is nonzero. Note that g is the planning MMU threshold, which has to be larger than or equal to the machine MMU threshold G_{min}, i.e., $g \geq G_{min}$.

Note that the MMU problem can also be formulated as a mixed integer problem

$$\min_{x \in \mathbb{R}^{N_x}, s \in \{0, 1\}^{N_x}} f(x \cdot s)$$
$$s.t. \, x_j \in [g, +\infty), j \leq N_x$$

(3)

where s is a binary vector. This formulation again shows the nonconvexity of the problem, and the high dimensionality of the problem makes it difficult to solve, especially for large g.

Various methods have been developed to solve the MMU problem, including postprocessing methods [9,16,21] and optimization methods [11,19,22,24]. For example, a post-processing method [9] with simple thresholding of proton spots was used by Varian Eclipse treatment planning system; the spot-reduction method [42,43] was proposed to solve the MMU problem, which heuristically handles the MMU constraint by iteratively removing small-weight spots, i.e., the smallest m% spot weights are set to 0 every n iterations during a total of N iterations; a mathematically rigorous optimization approach was developed to solve the MMU problem, using alternating direction method of multipliers (ADMM) [22,24,37]; very recently, a so-called stochastic coordinate descent (SCD) method was proposed to solve the MMU problem with relatively large g [31].

However, the MMU problem with large g remains challenging, as it becomes increasingly difficult to maintain the plan quality (e.g., the conformal dose coverage for tumor target and the low dose for normal tissue) for large value of g. Note that although some methods can have better plan quality than others, this observation of deteriorated plan quality for the MMU problem with large g is intrinsic to the MMU problem, and not specific to an optimization method. To demonstrate this point, we will also compare with another popular optimization algorithm called proximal gradient descent (PGD) method (a generalization of proximal forward-backward splitting or FISTA method) [38–40] that can be used to solve the MMU problem.

The contribution of this work is to develop a new optimization method based on orthogonal matching pursuit (OMP) [5] that can substantially improve the solution to the MMU problem with large g in terms of plan quality.

2. Methods

2.1. Notations

Before introducing the new optimization method (Section 2.2), we will provide the notations and planning objectives based on dose-volume-histogram (DVH) constraints.

The dose d is calculated based on the spot x via dose influence matrix D, i.e.,

$$d_i = D_i x = \sum_{j=1}^{N_x} D_{ij} x_j, i \leq N_d,$$

(4)

where x is the spot weight vector of N_x spots indexed by j, d is the dose vector of N_d voxels indexed by i, and D_i is the ith row of D with the jth entry D_{ij} as the dose from the jth spot of the unit weight to the ith voxel.

The planning objective f (Eq. (2)) is the sum of various optimization objective terms corresponding to DVH constraints [1,3], including DVH-max constraint (e.g., for organs-at-risk (OAR))

$$D_{p\%} \leq c : \leq p\% \text{ of ROI receives } \geq c \text{ dose,}$$

(5)

and DVH-min constraint (e.g., for planning target volume (CTV))

$$D_{p\%} \geq c : \geq p\% \text{ of ROI receives } \geq c \text{ dose,}$$

(6)

where ROI denotes the region of interest for a DVH constraint under consideration.

The DVH planning constraints Eq. (5) and (6) are nonconvex with respect to d and therefore nonconvex with respect to x. This is because, for specific DVH constraints, not all voxels of d need to be included and the locations (i.e., the active set Ω of indices) of d where these planning constraints are enforced (i.e., as included in corresponding optimization objectives) also depend globally on d and this DVH constraint (e.g., c and p).

The active set Ω for the optimization objective corresponding to a given DVH constraint can be rigorously determined as a set of indices where the DVH constraint is least-violated [3,26,31]. Mathematically, Ω for DVH-max constraint can be determined by

$$\Omega = \left\{ i \mid p \leq i' \leq p*, d'_{p*} = c \right\}, \text{ if } d'_p > c,$$

(7)

for DVH-min constraint by

$$\Omega = \left\{ i \mid p* \leq i' \leq p, d'_{p*} = c \right\}, \text{ if } d'_p < c,$$

(8)

where d' (indexed by i') is the sort of d (indexed by i) in descending order.

The total planning objective f is

$$f(x) = \sum_{m=1}^{M} \frac{W_m}{2N_m} \sum_{i \in \Omega_m} \|D_i x - c_m\|_2^2,$$

(9)

where c_m is the dose objective, Ω_m the active set, N_m the number of voxels in Ω_m, w_m the objective weight for the mth DVH constraint, and M is the number of DVH constraints.

To simplify the notation, Eq. (9) is rewritten as

$$f(x) = \frac{1}{2}\|A(\Omega)x - b(\Omega)\|_2^2,$$

(10)

where the matrix A consists of D, ω_m, N_m, and the vector b consists of c_m, ω_m, N_m. Both A and b depend on Ω via Eq. (7)–(9).

Note that although the total planning objective Eq. (10) is nonconvex and nonlinear with respect to x (i.e., A and b depend on Ω, which depends on x), it is a convex least-square summation in x once Ω is fixed. This motivates the use of iterative convex relaxation (ICR) method [22–24,26]. ICR for solving Eq. (2) consists of two following iterative steps.

In Step 1, we fix Ω^k and solve the MMU problem with a convex optimization objective

$$x^{k+1} = \underset{x}{\text{argmin}} \frac{1}{2}\left\|A\left(\Omega^k\right)x - b\left(\Omega^k\right)\right\|_2^2$$
$$s.t. \quad x \in \{0\} \cup [g, +\infty)$$

(11)

In Step 2, we update the active set Ω^{k+1} based on x^{k+1}

$$\Omega^{k+1} = H\left(Dx^{k+1}\right),$$

(12)

where H denotes the determination of Ω based on $d=Dx$ formally via Eq. (7) and (8).

2.2. Orthogonal Matching Pursuit

However, although the optimization objective in Eq. (11) is convex, Eq. (11) is still a nonconvex problem, because the MMU constraint is nonconvex, which poses the major optimization challenge. The nonconvexity of the MMU constraint is also apparent from the alternative formulation of the problem as the mixed-integer programming Eq. (3).

Here we propose to use a modified version of orthogonal matching pursuit (OMP) method [5] to handle the MMU constraint, i.e., an iterative process consisting of first the selection of nonzero spots and then convex optimization problem restricted to this nonzero set.

To start with, we will provide the OMP method (Algorithm 1) for solving Eq. (11) without the MMU constraint, i.e.,

$$\min_{x \in \mathbb{R}^{N_x}} \frac{1}{2} \|Ax - b\|_2^2.$$

(13)

OMP belongs to the category of greedy optimization methods: starting from a null set, OMP iteratively adds to the current support one element index at a time, which has the maximal residual among all indexes. For clarity, the OMP algorithm for solving Eq. (13) is provided as follows.

Algorithm 1: OMP for solving Eq. (13)

Input: A, b. $S^0 = \varnothing$, $x^0 = 0$.
1.   **For** $n = 1, 2, \ldots N_x$
2. $j^{n+1} = \underset{j}{\text{argmax}} \left\{ [A^*(b - Ax^n)]_j \right\}$
3. $S^{n+1} = S^n \cup \left\{ j^{n+1} \right\}$
4. $x^{n+1} = \underset{\text{suppt}(z) \in S^{n+1}}{\min} \|b - Az\|_2^2$
5.   Set $n := n+1$
6.   **End**
Output: x^{N_x}.

It can be rigorously proven that the objective value monotonically decreases with respect to the number of iterations for OMP. In another word, the newly added spot guarantees the maximal reduction of optimization objective. This nice property of OMP motivates us to develop an OMP method for solving the MMU problem Eq. (11) by OMP.

However, the OMP method does not directly apply to the MMU problem Eq. (11), due to the MMU constraint, for which we propose the following modified OMP method (Algorithm 2).

Algorithm 2: **Modified OMP for solving Eq. (11)**

Input: $A(\Omega), b(\Omega), S_1{}^0 = S_2{}^0 = \varnothing, \quad x^0 = y^0 = 0, \quad N_{Inner}$ and g.

1.　　　　**For** $n = 0, 1, \quad 2, \quad \cdots, \quad N_{Inner} - 1$

2.　　　　　　　$j^{n+1} = \underset{j}{\operatorname{argmax}}\left\{\left[A^T(b - Ay^n)\right]_j\right\}$

3.　　　　　　　$S_1^{n+1} = S_1^n \cup \left\{j^{n+1}\right\}$

4.　　　　　　　$y^{n+1} = \underset{Supp(z) \subset S_1^{n+1}}{\operatorname{argmin}}\left\|b - Az\right\|_2^2$

5.　　　　　　　$S_2^{n+1} = S_2^n \cup \left\{j^{n+1}\right\}$

6.　　　　　　　$\left(f^{n+1}, x^{n+1}\right) = \begin{cases} \underset{z}{\operatorname{argmin}}\frac{1}{2}\|Az - b\|_2^2 \\ s.t. \quad z \geq g, Supp(z) = S_2^{n+1} \end{cases}$

7.　　　　　　**If** $f^{n+1} \geq f^n$

8.　　　　　　　　$S_2^{n+1} = S_2^n$

9.　　　　　　　　$x^{n+1} = x^n$

10.　　　　**endif**

11.　　**End**

Output: $x^{N_{Inner}}$.

The key components of modified OMP (Algorithm 2) include the followings.

- Two support sets S_1 and S_2 are defined in modified OMP instead of just one in OMP: S_1 is the support set from original OMP to iteratively add the index with the most contribution to the planning objective (without considering MMU constraint), and S_2 is a subset of S_1 that has decreasing total planning objective (when the MMU constraint is accounted for) and is also feasible in the sense that all spot weights satisfy the MMU constraint.

- Steps 3, and 4 correspond to the steps in the original OMP to update S_1, for identifying the spot j^{n+1} that makes the most contribution to the planning objective residual based on the most recent spot weight y^n, and then reoptimizing for y^{n+1} based on the new support S_1^{n+1} that includes j^{n+1}.

- Steps 5 and 6 are modified OMP steps to update S_2 and x subject to the MMU constraint. Note that in Step 6, the MMU constraint becomes a convex constraint when restricted to S_2 that consists of nonzero spots, and thus the optimization of x is a convex constrained problem which can be effectively solved, e.g., by ADMM or PGD.

- Steps 7–11 are to ensure that the updated x has a decreasing total planning objective f; otherwise, x and thus f are not updated.

The overall solution algorithm for solving the MMU problem Eq. (2) via ICR and modified OMP is summarized in Algorithm 3, where Algorithm 2 is denoted by $OMP(\Omega, N_{Inner})$ with N_{Inner} as the number of OMP iterations for the inner loop.

Algorithm 3: **Solution algorithm to Eq.(2)**

Input: $x^0 = 0$, N_{Inner}, N_{outer}.

1.   **For** $k = 0, 1, 2, \cdots, N_{outer} - 1$

2. $\Omega^{k+1} = H(Dx^k)$

3. $x^{k+1} = OMP(\Omega^{k+1}, N_{Inner})$

4.   **End**

Output: $x^{N_{Outer}}$.

2.3. Materials

The new MMU method via OMP was validated in comparison with methods (ADMM and PGD) and a recently developed MMU method SCD, for solving the MMU problems with large MMU threshold g for three clinical cases: brain (fraction prescription dose of 2Gy, 2Gy and 20Gy for IMPT, ARC and FLASH respectively), lung (fraction prescription dose of 2Gy, 2Gy and 12Gy for IMPT, ARC and FLASH respectively) and head-and-neck (HN) (fraction prescription dose of 2.12Gy, 2.12Gy and 15Gy for IMPT, ARC and FLASH respectively). For each case, we considered three scenarios of the MMU constraints for high dose rates: IMPT (3–4 beam angles), ARC (24 equally sampled beam angles), and FLASH (3–4 transmission beams) and the MMU threshold g is chosen as the largest value that reasonable plan quality (e.g., D_{max}<120% for the tumor target) can be achieved by OMP. MatRad [17] was used to generate dose influence matrices D with 5mm spot lateral spacing on 3 mm^3 dose grid. The planning details of IMPT, ARC, and FLASH can be found in [22,24], [36,48], and [32] respectively. All plans were normalized to have D95 = 100% to CTV. The FLASH coverage was quantified using FLASH modifying factor based on a sigmoid model of FLASH dose and dose rate [32]. The conformal index (CI) was defined as *V2 100/(V×V' 100)* (V_{100}: CTV volume receiving at least 100% of prescription dose; V: CTV volume; $V' 100$: total volume receiving at least 100% of prescription dose). The plan robustness is quantitively evaluated in the robustness variance (RV): RVp defined at $p\%$ dose is the averaged variation between maximum and minimum percentage volume at *(p-1)%, p%* and *(p+1)%*.

3. Results

ADMM, PGD, SCD, and OMP are comparatively evaluated in terms of target coverage and OAR sparing for IMPT, ARC and FLASH respectively.

3.1. IMPT

In terms of the target coverage, Table 1–3 show that OMP was quantitatively better than PGD, ADMM and SCD: the CI improved from 0.42, 0.52 and 0.33 respectively to 0.65 for brain (Table 1), from 0.50, 0.54 and 0.59 respectively to 0.79 for lung (Table 2), and from 0.33, 0.52 and 0.44 respectively to 0.81 for HN (Table 3); under the same target dose normalization, the max target dose decreased from 368.0%, 154.4% and 134.5% to 119.5% for brain, from 322.9%, 151.8% and 125.6% to 117.5% for lung, and from 574.5%, 174.4% and 130.6% to 119.1% for HN. The improvement of target coverage via OMP was also

illustrated by comparing dose plots in (a), (e), (i), (m) and DVH plots in (q) of Figure 1–3, e.g., the DVH curve of the target (red solid line) was generally under that from PGD (magenta solid line), ADMM (blue solid line), and SCD (green solid line).

In terms of the OAR sparing, Table 1–3 demonstrate that OMP generally had a lower OAR dose than PGD, ADMM and SCD. For example, compared to PGD, ADMM and SCD, OMP decreased the max dose of the brainstem from 49.2%, 25.4% and 22.9% to 20.0% (Table 1), mean heart dose from 7.6%, 3.6% and 7.2% to 1.4% (Table 2), and mean cavity oral dose from 15.4%, 16.7% and 21.0% to 12.4% (Table 3).

3.2. ARC

In terms of the target coverage, Table 1–3 show that OMP was substantially better than PGD, ADMM and SCD: CI improved from 0.46, 0.60 and 0.61 respectively to 0.83 for brain (Table 1), from 0.64, 0.72 and 0.50 respectively to 0.87 for lung (Table 2), and from 0.35, 0.51 and 0.62 respectively to 0.83 for HN (Table 3). PGD and ADMM failed to generate an acceptable plan, e.g., the max target dose of PGD was 358.3%, 419.6%, and 311.2% respectively for brain, lung, and HN; the max target dose of ADMM for three cases was 179.8%, 163.3%, and 212.1% respectively; Although SCD had lower max dose, e.g., 130.4% for brain, 129.3% for lung and 118.2% for HN, OMP still outperformed SCD, e.g., with max target dose 119.8%, 119.3%, and 116.8% for brain, lung and HN respectively. The improvement of target coverage and organ at risk sparing via OMP was also evident by comparing dose and DVH plots in (b), (f), (j), (n) and (r) of Figure 1–3.

In terms of the OAR sparing, Table 1–3 suggest that OMP also had a lower OAR dose than PGD, ADMM and SCD. For example, compared to PGD, ADMM and SCD, OMP decreased max dose of the brainstem from 41.8%, 26.6% and 22.6% to 20.8% (Table 1), mean esophagus dose from 7.3%, 4.3% and 7.9% to 3.8% (Table 2), and mean oral cavity dose from 21.5%, 5.2% and 15.3% to 3.9% (Table 3).

3.3. FLASH

In terms of the target coverage, Table 1–3 show that PGD did not provide acceptable plan, and OMP had comparable results with ADMM and SCD. However, OMP had smaller max target dose: a decrease from 283.4%, 150.0% and 123.0% to 119.6% for brain (Table 1), from 286.1%, 125.1% and 125.3% to 118.7% for lung (Table 2), and from 120.3%, 128.7% and 126.0% to 118.5% for HN (Table 3). The improved target dose and DVH plots via OMP were also evident in in (c), (d), (g), (h), (k), (l), (o), (p) and (s) of Figure 1–3.

In terms of the OAR sparing, Table 1–3 demonstrate that OMP generally had lower OAR dose than PGD, ADMM and SCD. For example, compared to PGD, ADMM and SCD, OMP generally decreased max dose of the brainstem from 35.7%, 15.2%, and 12.1% to 14.3% (Table 1) (here SCD had lower max dose than OMP), mean esophagus dose from 8.1%, 4.9% and 5.1% to 2.4% (Table 2), and mean oral cavity dose from 15.2%, 7.3% and 11.1% to 2.0% (Table 3). On the other hand, OMP provided comparable dose rate and FLASH coverage for the OAR CTV10mm (10mm isotropic expansion of CTV excluding CTV) as summarized in Table 4.

3.4. Robust Analysis

The robust analysis for the brain case is presented in Table 5 and Fig. 4, using 3mm setup uncertainty and 3.5% range uncertainty. As shown in Fig. 4, since PGD and ADMM had severely deteriorated target dose conformality due to large MMU threshold g, the robustness comparison is primarily between SCD and OMP, which still relatively preserved target dose conformality. Fig. 4 suggests that OMP had tighter DVH's from various uncertainty scenarios to each other than SCD. This was also quantitatively true as summarized in Table 5: compared to OMP, SCD had smaller RV_{115} for CTV and smaller RV_{30} for brainstem, e.g., the reduction of RV_{115} from 12.94% to 2.63% for IMPT, from 6.09% to 3.39% for ARC, and from 19.23% to 2.67% for FLASH.

4. Discussion

It is interesting to note the plan difference between PGD, ADMM, SCD and OMP for solving the proton FLASH problem as presented in Fig. 1–3 (c), (d), (g), (h), (k), (l), (o) and (p): for OMP, the proton spots are added for optimization one by one in a greed fashion that minimizes the total objective residual (Step 3 of Algorithm 2), while in PGD, ADMM, all proton spots are included in optimization from the beginning. As a result, the beam angles for each axial dose slice were utilized differently in OMP: for the brain case (Fig. 1), the FLASH dose was mostly from one beam (90º) for S1 (Fig. 1(o)) and three beams (0º, 45º, 90º) for S2 (Fig. 1(p)); for the lung case (Fig. 2), the FLASH dose was mostly from two beam (0º, 60º) for S1 (Fig. 2(o)) and all three beams (0º, 60º, 120º) for S2 (Fig. 2(p)); for the HN case (Fig. 3), the FLASH dose was mostly from one beam (90º) for S1 (Fig. 3(o)) and two beams (45º, 90º) for S2 (Fig. 3(p)). In comparison, all the beam angles were always utilized for all dose slices in PGD, ADMM and SCD: for the brain case (Fig. 1), the FLASH dose was from all four beams (0º, 45º, 90º, 135º) (Fig. 1(c), (d) (g), (h), (k) and (l)); for the lung case (Fig. 2), the FLASH dose was from all three beams (0º, 60º, 120º) (Fig. 2 (c), (d), (g), (h), (k) and (l)) for the HN case (Fig. 3), the FLASH dose was from all four beams (0º, 45º, 90º, 135º) (Fig. 3 (c), (d), (g), (h), (k) and (l)). This observation of does distribution difference between OMP and PGD, ADMM, SCD further suggests OMP is more efficient to solve the MMU problem with large g, in the sense that it selects the most relevant spots for optimization to meet the planning objective. On the other hand, OMP used fewer spots than ADMM with generally more weights per spot, which helped to meet the large MMU threshold.

In current clinical practice, stereotactic body radiation therapy (SBRT) is not commonly used for IMPT. This is partially due to the combination of the long treatment time and the lack of effective motion management to monitor patient movement during the treatment. With this respect to this, the proposed MMU method may enable the SBRT plan of preserved plan quality (which cannot be generated by IMPT optimizer such as ADMM) and at the same time high-dose-rate IMPT delivery per beam to within a single breathing-hold time, which we plan to investigate in a future work, by exploring the combination of this new MMU optimization method together with other techniques for high-dose-rate IMPT, such as the use of range modulator [18], energy layer reduction method [22], and variable MMU threshold method [24].

It is noted that conventional optimization methods such as sequential quadratic programing (SQP) [44] and interior point (IP) method [45] have been widely used to handle the convex constraints in RT, such as lower bounds. However, it is not straightforward for these conventional methods to handle the nonconvex MMU constraint (i.e., allowing spots to be turned on or off). The main challenge of the MMU problem is to localize non-zeros spots. After the non-zero spots are localized, the optimization problem degenerates to a convex low-bounded problem that can be handled easily by many optimization methods, including conventional methods. For example, SQP or IP can replace ADMM or PGD for solving Step 6 in OMP, which however is likely to make little or no difference.

In contrast, ADMM or PGD can be directly used to solve the nonconvex MMU problem, noting that the projection onto the nonconvex MMU bounded set admits an analytical solution, although the performance of ADMM or PGD is not satisfactory for large MMU threshold [31]. That is, the MMU problem can be reformulated as an unconstrained problem by lifting the MMU constraint into the objective using the indicator function. Then the splitable structure of the objective of this reformulation, and the non-smoothness of the objective (the indicator function) motivates the use of first-order proximal splitting algorithms such as ADMM or PGD.

Note that there are some convergence analysis results of using first-order proximal splitting algorithms for non-convex problems (e.g., [37] for ADMM, [38–40] for PGD and [46]). For example, the theoretical convergence relies on the Kurdyka-Łojasiewicz (KL) property [47]. It can be shown that the MMU constraint set is a semi-algebraic set and the indicator function of the semi-algebraic set is called the semi-algebraic function which satisfies the KL property. Regarding the optimization objective, which is the summation of least square function (a semi-algebraic function) and an indicator function of a semi-algebraic set when the active set Ω is fixed, it is also semi-algebraic [38], and therefore the optimization problem Eq. (2) also satisfies the KL property. However, since the active set Ω also varies iteratively, there is no theoretical guarantee of solution convergence for Eq. (2). Plus, even if the solution converges, it converges to a local minimum, which explains why different solution algorithms can have different solutions.

This OMP based MMU method should be also needed for spot-scanning ARC therapy [13], as the delivery time optimization is important to ARC. For example, frequent low-to-high energy layer switching (so called switch-up (SU)) can significantly degrade delivery efficiency, because the SU takes much longer than the usual high-to-low energy layer switching [27,34,36]. It will be highly interesting to develop a full optimization method for efficient and accurate ARC therapy based on OMP to handle the MMU constraint that also explicitly minimizes the delivery time while optimizing the plan quality, by incorporating accurate machine delivery model [35] and the SU regularization method [27,34].

5. Conclusion

We have developed an OMP based optimization algorithm for solving MMU problems with large MMU threshold. The applications to IMPT, ARC, and FLASH have demonstrated

the effectiveness of the OMP method, with substantially improved plan quality from the methods (PGD and ADMM) and a recently proposed method SCD.

Acknowledgment

References

[1]. Bortfeld T, Stein J, and Preiser K, 1997. Clinically relevant intensity modulation optimization using physical criteria. 12th Int. Conf. on the Use of Computers in Radiation Therapy. 1–4.

[2]. Lomax A, 1999. Intensity modulation methods for proton radiotherapy. Physics in Medicine & Biology, 44(1), 185–205. [PubMed: 10071883]

[3]. Wu Q, and Mohan R, 2000. Algorithms and functionality of an intensity modulated radiotherapy optimization system. Medical Physics, 27(4), 701–711. [PubMed: 10798692]

[4]. Combettes PL, and Wajs VR, 2005. Signal recovery by proximal forward-backward splitting. Multiscale Modeling & Simulation, 4(4), 1168–1200.

[5]. Tropp JA, and Gilbert AC, 2007. Signal recovery from random measurements via orthogonal matching pursuit. IEEE Transactions on Information Theory, 53(12), 4655–4666.

[6]. Beck A, and Teboulle M, 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences, 2(1), 183–202.

[7]. Goldstein T, and Osher S, 2009. The split Bregman algorithm for l_1 regularized problems. SIAM Journal on Imaging Sciences, 2(2), 323–343.

[8]. Albertini F, Gaignat S, Bosshardt M, and Lomax AJ, 2009. Planning and optimizing treatment plans for actively scanned proton therapy Biomedical Mathematics: Promising Directions in Imaging, Therapy Planning, and Inverse Problems (ed 1). Madison: Medical Physics Pub Corp, 1–18.

[9]. Zhu XR, Sahoo N, Zhang X, et al. , 2010. Intensity modulated proton therapy treatment planning using single-field optimization: the impact of monitor unit constraints on plan quality. Medical Physics, 37(3), 1210–1219. [PubMed: 20384258]

[10]. Boyd S, Parikh N, Chu E, et al. , 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine learning 3(1), 1–122.

[11]. Cao W, Lim G, Li X, et al. , 2013. Incorporating deliverable monitor unit constraints into spot intensity optimization in intensity-modulated proton therapy treatment planning. Physics in Medicine & Biology, 58(15), 5113. [PubMed: 23835656]

[12]. Howard M, Beltran C, Mayo CS, et al. , 2014. Effects of minimum monitor unit threshold on spot scanning proton plan quality. Medical Physics, 41(9), 091703. [PubMed: 25186378]

[13]. Ding X, Li X, Zhang JM, et al. , 2016. Spot-scanning proton arc (SPArc) therapy: the first robust and delivery-efficient spot-scanning proton arc therapy. International Journal of Radiation Oncology Biology Physics, 96(5), 1107–1116. [PubMed: 27869083]

[14]. Gao H, 2016. Fused analytical and iterative reconstruction (AIR) via modified proximal forward–backward splitting: a FDK-based iterative image reconstruction example for CBCT. Physics in Medicine & Biology, 61(19), 7187. [PubMed: 27649259]

[15]. Gao H, 2016. Robust fluence map optimization via alternating direction method of multipliers with empirical parameter optimization. Physics in Medicine & Biology, 61(7), 2838. [PubMed: 26987680]

[16]. Lin Y, Kooy H, Craft D, et al. , 2016. A Greedy reassignment algorithm for the PBS minimum monitor unit constraint. Physics in Medicine & Biology, 61(12), 4665–4678. [PubMed: 27245098]

[17]. Wieser HP, Cisternas E, Wahl N, et al. , 2017. Development of the open-source dose calculation and optimization toolkit matRad. Medical Physics, 44(6), 2556–2568. [PubMed: 28370020]

[18]. Simeonov Y, Weber U, Penchev P, et al. , 2017. 3D range-modulator for scanned particle therapy: development, Monte Carlo simulations and experimental evaluation. Physics in Medicine & Biology, 62(17), 7075. [PubMed: 28741595]

[19]. Shan J, An Y, Bues M, et al. , 2018. Robust optimization in IMPT using quadratic objective functions to account for the minimum MU constraint. Medical Physics, 45(1), 460–469. [PubMed: 29148570]

[20]. Gu W, O'Connor D, Nguyen D, et al. , 2018. Integrated beam orientation and scanning spot optimization in intensity modulated proton therapy for brain and unilateral head and neck tumors. Medical Physics, 45(4), 1338–1350. [PubMed: 29394454]

[21]. Gao H, Clasie B, Liu T, et al. , 2019. Minimum MU optimization (MMO): an inverse optimization approach for the PBS minimum MU constraint. Physics in Medicine & Biology, 64(12), 125022. [PubMed: 31082813]

[22]. Lin Y, Clasie B, Liu T, et al. , 2019. Minimum-MU and sparse-energy-level (MMSEL) constrained inverse optimization method for efficiently deliverable PBS plans. Physics in Medicine & Biology, 64(20), 205001. [PubMed: 31530746]

[23]. Gao H, 2019. Hybrid proton-photon inverse optimization with uniformity-regularized proton and photon target dose. Physics in Medicine & Biology, 64(10), 105003. [PubMed: 30978714]

[24]. Gao H, Clasie B, McDonald M, et al. , 2020. Plan-delivery-time constrained inverse optimization method with minimum-MU-per-energy-layer (MMPEL) for efficient pencil beam scanning proton therapy. Medical Physics, 47(9), 3892–3897. [PubMed: 32614472]

[25]. Wilson JD, Hammond EM, Higgins GS, et al. , 2020. Ultra-high dose rate (FLASH) radiotherapy: Silver bullet or fool's gold?. Frontiers in Oncology, 9, 1563. [PubMed: 32010633]

[26]. Gao H, Lin B, Lin Y, et al. , 2020. Simultaneous dose and dose rate optimization (SDDRO) for FLASH proton therapy. Medical Physics, 47(12), 6388–6395. [PubMed: 33068294]

[27]. Gu W, Ruan D, Lyu Q, et al. , 2020. A novel energy layer optimization framework for spot scanning proton arc therapy. Medical Physics, 47(5), 2072–2084. [PubMed: 32040214]

[28]. Chen G, Hong X, Ding Q, et al. , 2020. AirNet: fused analytical and iterative reconstruction with deep neural network regularization for sparse data CT. Medical Physics, 47(7), 2916–2930. [PubMed: 32274793]

[29]. Chen G, Zhao Y, Huang Q, et al. , 2020. 4D-AirNet: a temporally-resolved CBCT slice reconstruction method synergizing analytical and iterative method with deep learning. Physics in Medicine & Biology, 65(17), 175020. [PubMed: 32575088]

[30]. Lin Y, Lin B, Fu S, et al. , 2021. SDDRO-Joint: simultaneous dose and dose rate optimization with the joint use of transmission beams and Bragg peaks for FLASH proton therapy. Physics in Medicine & Biology, 66(12), 125011.

[31]. Cai JF, Chen RC, Fan J, et al. , 2022. Minimum-monitor-unit optimization via a stochastic coordinate descent method. Physics in Medicine & Biology, 67(1), 015009.

[32]. Gao H, Liu J, Lin Y, et al. , 2022. Simultaneous dose and dose rate optimization (SDDRO) of the FLASH effect for pencil-beam-scanning proton therapy. Medical Physics, 49(3), 2014–2025. [PubMed: 34800301]

[33]. Shen H, Zhang G, Lin Y, et al. , 2023. Beam angle optimization for proton therapy via group-sparsity based angle generation method. Medical Physics. 10.1002/mp.16392

[34]. Wuyckens S, Saint-Guillain M, Janssens G, et al. , 2022. Treatment planning in arc proton therapy: Comparison of several optimization problem statements and their corresponding solvers. Computers in Biology and Medicine, 105609. [PubMed: 35803749]

[35]. Zhao L, Liu G, Chen S, et al. , 2022. Developing an accurate model of spot scanning treatment delivery time and sequence for a compact superconducting synchrocyclotron proton therapy system. Radiation Oncology, 17(1), 1–19. [PubMed: 34980178]

[36]. Zhang G, Shen H, Lin Y, et al. , 2022. Energy layer optimization via energy matrix regularization for proton spot-scanning arc therapy. Medical Physics. 49(9), 5752–5762. [PubMed: 35848227]

[37]. Wang Y, Yin W, and Zeng J, 2019. Global convergence of ADMM in nonconvex nonsmooth optimization. Journal of Scientific Computing,78(1), 29–63.

[38]. Bolte J, Sabach S, and Teboulle M, 2014. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Mathematical Programming, 146(1), 459–494.

[39]. Attouch H, Bolte J, Redont P, et al. , 2010. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. Mathematics of operations research, 35(2), 438–457.

[40]. Pock T, and Sabach S (2016). Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems. SIAM Journal on Imaging Sciences, 9(4), 1756–1787.

[41]. Maradia V, van de Water S, Meer D, et al. , 2022. Ultra-fast pencil beam scanning proton therapy for locally advanced non-small-cell lung cancers: field delivery within a single breath-hold. Radiotherapy and oncology, 174, 23–29. [PubMed: 35788354]

[42]. Albertini F, Gaignat S, Bosshard M. et al. , 2009. Planning and optimizing treatment plans for actively scanned proton therapy. Biomedical Mathematics Promising Directions in Imaging, Therapy Planning, and Inverse Problems. 1–18.

[43]. Van De Water S, Belosi MF, Albertini F, et al. , 2020. Shortening delivery times for intensity-modulated proton therapy by reducing the number of proton spots: An experimental verification. Physics in Medicine & Biology, 65(9), 095008. [PubMed: 32155594]

[44]. Wieser HP, Cisternas E, Wahl N, et al. , 2017. Development of the open-source dose calculation and optimization toolkit matRad. Medical physics, 44(6), 2556–2568. [PubMed: 28370020]

[45]. Breedveld S, van den Berg B, Heijmen B, 2017. An interior-point implementation developed and tuned for radiation therapy treatment planning. Computational Optimization and Applications, 68(2), 209–242.

[46]. Attouch H, Bolte J, Svaiter BF, 2013. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss–Seidel methods. Mathematical Programming, 137, 91–129.

[47]. Bian F, Liang J, and Zhang X, 2021. A stochastic alternating direction method of multipliers for non-smooth and non-convex optimization. Inverse Problems, 37(7), p.075009.

[48]. Zhang G, Long Y, Lin Y, et al. , 2023. A treatment plan optimization method with direct minimization of number of energy jumps for proton arc therapy. Physics in Medicine & Biology. 68, 085001.
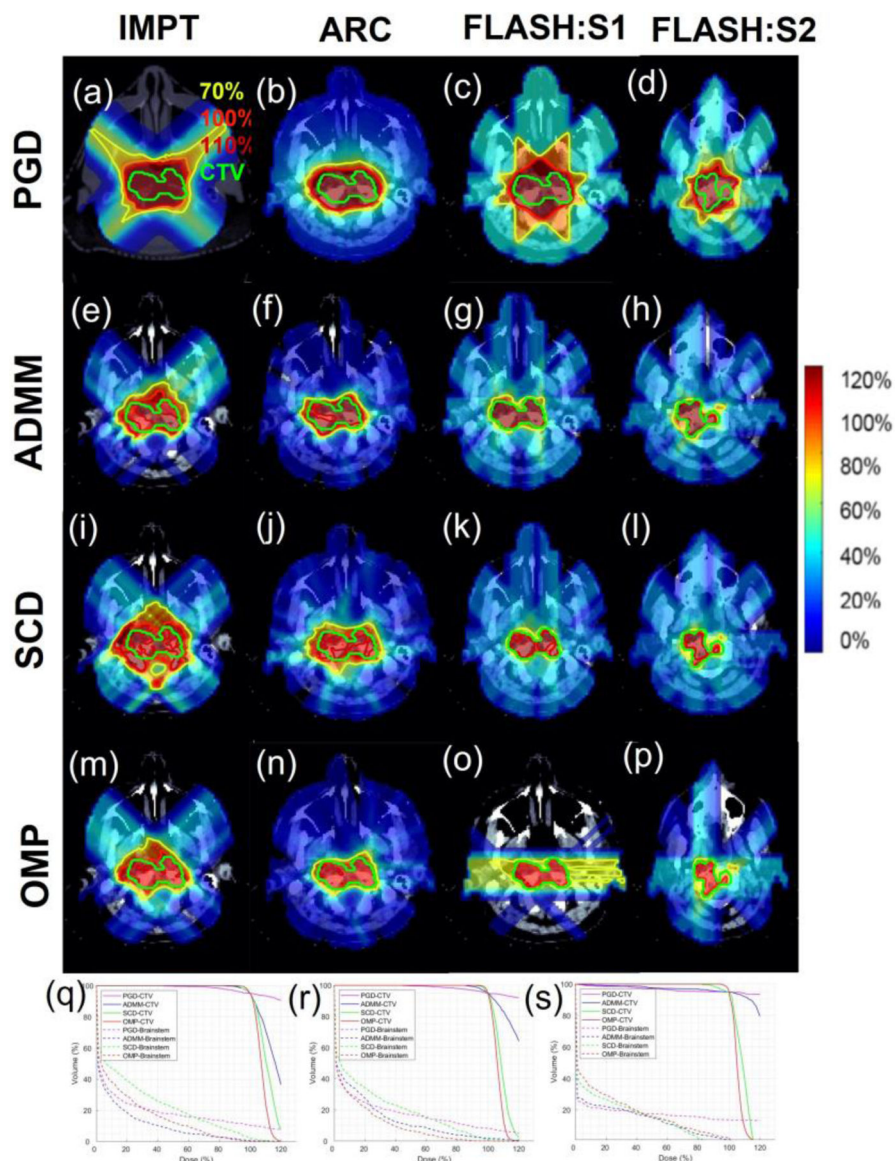
Figure 1. Brain.
(a)-(d): PGD dose plots from IMPT, ARC, FLASH (slice 1), FLASH (slice 2) respectively; (e)-(h): ADMM dose plots from IMPT, ARC, FLASH (slice 1), FLASH (slice 2) respectively; (i)-(l): SCD dose plots from IMPT, ARC, FLASH (slice 1), FLASH (slice 2) respectively; (m)-(p): OMP dose plots from IMPT, ARC, FLASH (slice 1), FLASH (slice 2) respectively; (q)-(s): DVH plots for IMPT, ARC, FLASH respectively, where solid lines are for CTV and dotted lines are for Brainstem. The dose plot window is [0%, 120%]. 110%, 100% 70% isodose lines and CTV are highlighted in dose plots.
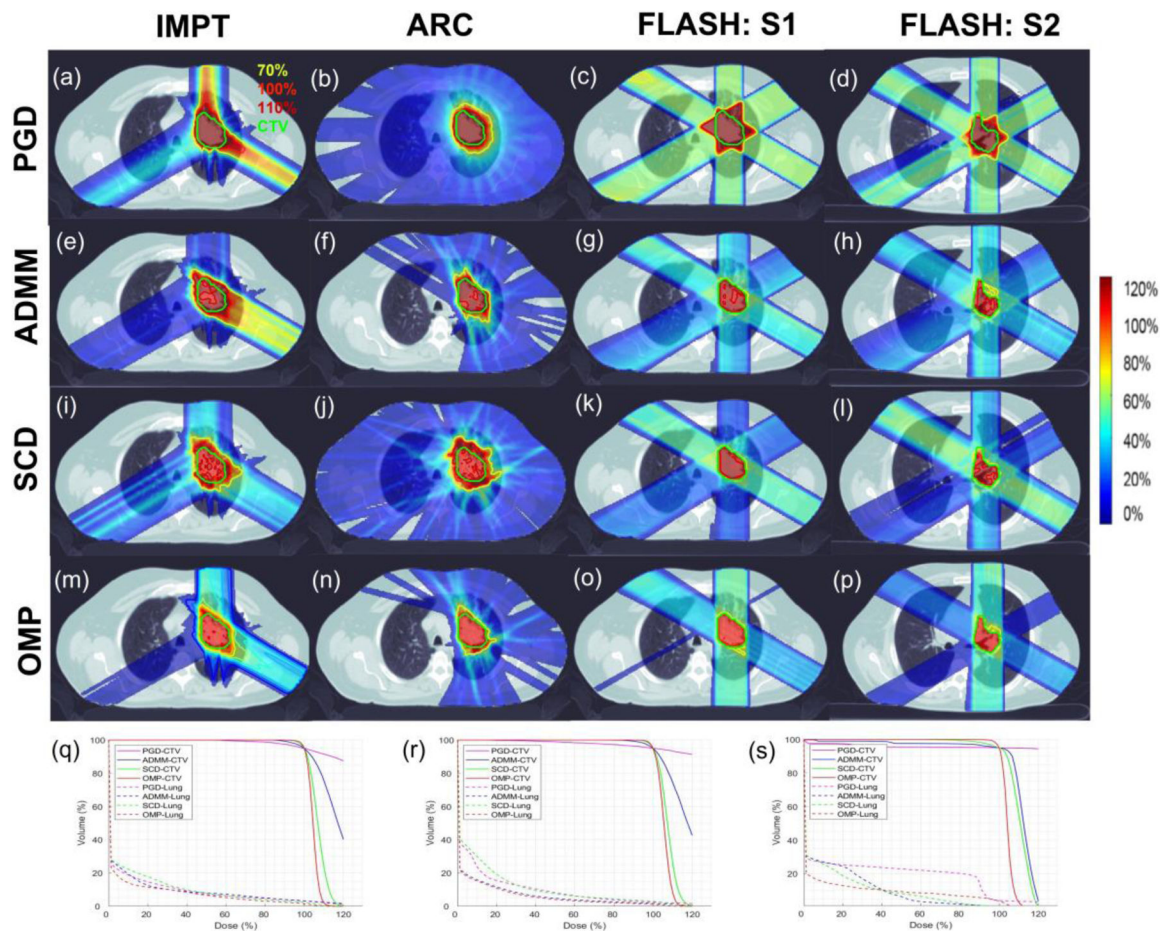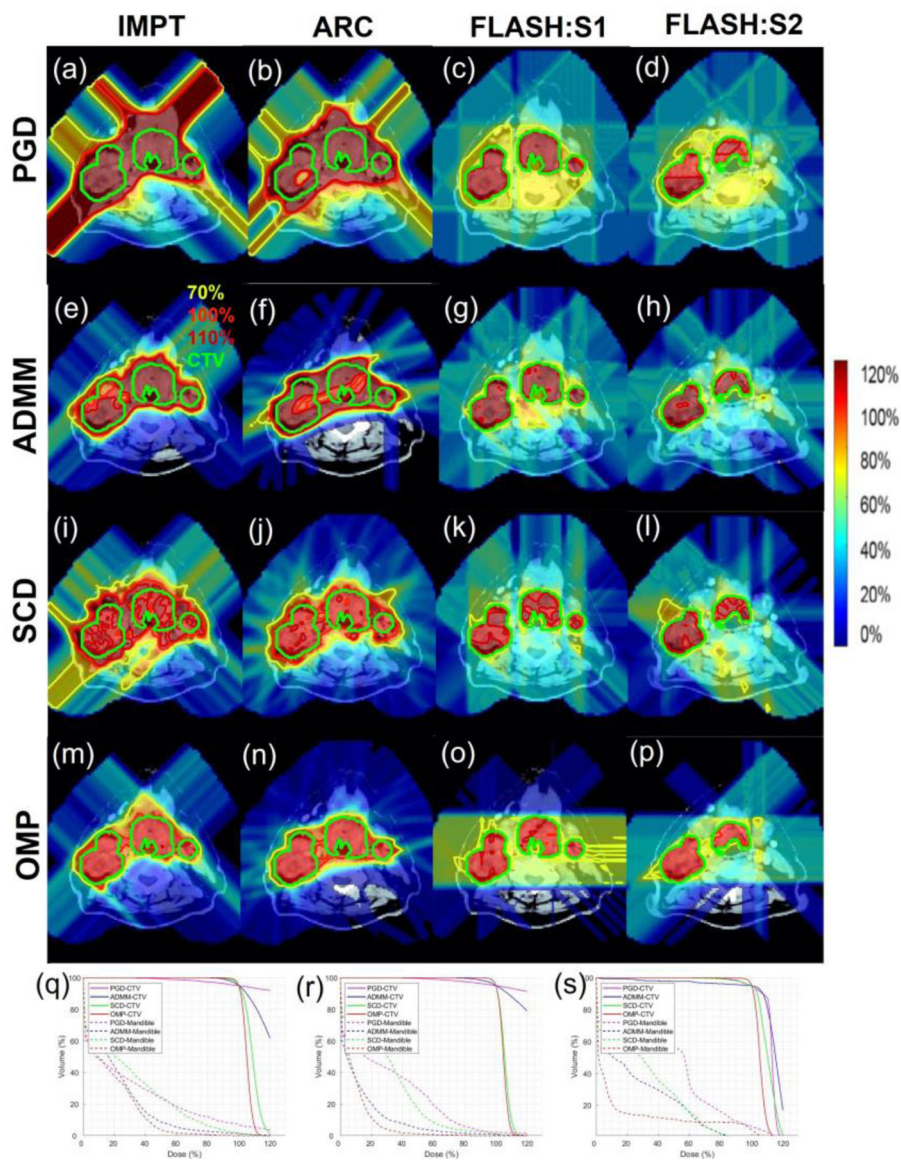
Figure 2. Lung.
(a)-(d): PGD dose plots from IMPT, ARC, FLASH (slice 1), FLASH (slice 2) respectively; (e)-(h): ADMM dose plots from IMPT, ARC, FLASH (slice 1), FLASH (slice 2) respectively; (i)-(l): SCD dose plots from IMPT, ARC, FLASH (slice 1), FLASH (slice 2) respectively; (m)-(p): SCD dose plots from IMPT, ARC, FLASH (slice 1), FLASH (slice 2) respectively; (q)-(s): DVH plots for IMPT, ARC, FLASH respectively, where solid lines are for CTV and dotted lines are for Lung. The dose plot window is [0%, 120%]. 110%, 100% 70% isodose lines and CTV are highlighted in dose plots.

Figure 3. HN.
(a)-(d): PGD dose plots from IMPT, ARC, FLASH (slice 1), FLASH (slice 2) respectively; (e)-(h): ADMM dose plots from IMPT, ARC, FLASH (slice 1), FLASH (slice 2) respectively; (i)-(l): SCD dose plots from IMPT, ARC, FLASH (slice 1), FLASH (slice 2) respectively; (m)-(p): OMP dose plots from IMPT, ARC, FLASH (slice 1), FLASH (slice 2) respectively; (q)-(s): DVH plots for IMPT, ARC, FLASH respectively, where solid lines are for CTV and dotted lines are for Mandible. The dose plot window is [0%, 120%]. 110%, 100% 70% isodose lines and CTV are highlighted in dose plots.
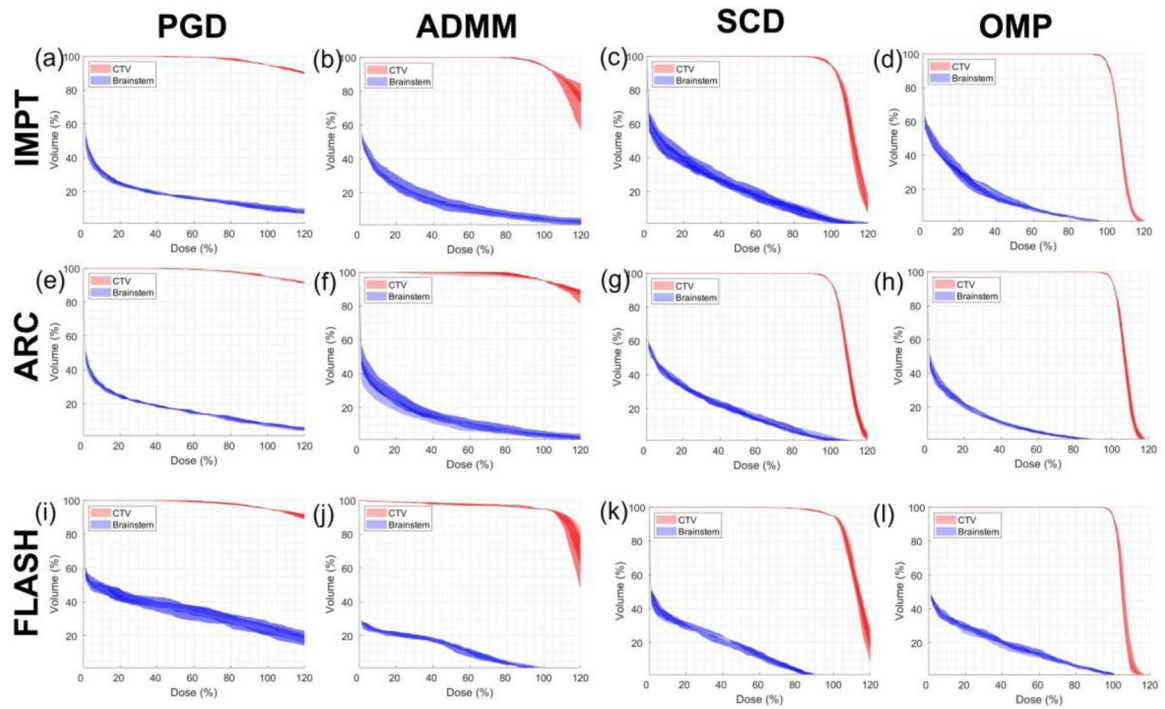
Figure 4. Robustness analysis for the brain case.
(a)-(d): IMPT DVH plots for PGD, ADMM, SCD, OMP respectively; (e)-(h): ARC DVH plots for PGD, ADMM, SCD, OMP respectively; (i)-(l): FLASH DVH plots for PGD, ADMM, SCD, OMP respectively.

Table 1.

Brain.

The dosimetric quantities from left to right: MMU threshold (unit: 10^6 protons), optimization objective value, max dose of CTV, conformal index, max dose of brainstem (BS), and mean dose of body. All quantities are in percentage except for g, with dose values in percentage to the prescription dose.

Unit：%		g	f	D_{max}	CI	BS	Body
IMPT	PGD	80	223.8	368.0	42	49.2	17.3
	ADMM		4.8	154.4	52	25.4	5.2
	SCD		2.1	134.5	33	22.9	17.3
	OMP		0.9	119.5	65	20.0	5.1
ARC	PGD	60	360.4	358.3	46	41.8	5.0
	ADMM		21.1	179.8	60	26.6	4.6
	SCD		1.5	130.4	61	22.6	6.9
	OMP		1.5	119.8	83	20.8	3.2
FLASH	PGD	1300	875.7	283.4	43	35.7	1.8
	ADMM		18.3	150.0	95	15.2	1.1
	SCD		1.2	123.0	95	12.1	1.2
	OMP		1.1	119.6	95	14.3	0.9

Table 2.

Lung.

The dosimetric quantities from left to right: MMU threshold (unit: 10^6 protons), optimization objective value, max dose of CTV, conformal index, mean dose of lung, heart, esophagus (ESO), and body. All quantities are in percentage except for g, with dose values in percentage to the prescription dose.

Unit：%		g	f	D_{max}	CI	Lung	Heart	ESO	Body
IMPT	PGD	45	133.5	322.9	50	1.4	7.6	6.4	3.3
	ADMM		5.2	151.8	54	1.8	3.6	6.1	2.9
	SCD		0.8	125.6	59	2.0	7.2	6.3	3.2
	OMP		0.3	117.5	79	1.3	1.4	4.6	2.2
ARC	PGD	50	356.8	419.3	64	2.1	5.1	7.3	3.9
	ADMM		6.1	163.3	72	1.3	0.8	4.3	2.0
	SCD		0.9	129.3	50	2.3	12.6	7.4	3.8
	OMP		0.5	119.3	87	1.1	0.4	3.8	1.7
FLASH	PGD	550	310.8	286.1	46	6.0	1.8	8.1	3.8
	ADMM		0.30	125.1	95	3.7	1.3	4.9	2.3
	SCD		0.21	125.3	95	3.3	1.5	5.1	2.4
	OMP		0.04	118.7	93	2.7	0.9	2.4	1.8

Table 3.

HN.

The dosimetric quantities from left to right: MMU threshold (unit: 10^6 protons), optimization objective value, max dose of CTV, conformity index, mean dose of mandible (MA), oral cavity (OC), and body All quantities are in percentage except for g, with dose values in percentage to the prescription dose.

Unit：%		g	f	D_{max}	CI	MA	OC	Body
IMPT	PGD	60	617.6	574.5	33	19.9	15.4	4.2
	ADMM		8.1	174.4	52	14.2	16.7	2.4
	SCD		1.0	130.6	44	19.8	21.0	3.3
	OMP		0.4	119.1	81	11.4	12.4	1.9
ARC	PGD	45	190.6	311.2	35	20.5	21.5	4.4
	ADMM		32.7	212.1	51	8.2	5.2	2.2
	SCD		0.8	118.2	62	20.5	15.3	2.9
	OMP		0.6	116.8	83	6.1	3.9	1.7
FLASH	PGD	550	0.6	120.3	95	19.4	15.2	3.3
	ADMM		1.1	128.7	95	11.2	7.3	2.4
	SCD		0.4	126.0	95	14.5	11.1	2.6
	OMP		0.05	118.5	91	6.4	2.0	2.0

Table 4.

Dose-rate and FLASH coverage.

Dose-rate coverage P_γ: the percentage of an OAR (CTV10mm: 10mm isotropic expansion of CTV excluding CTV) volume receiving $\gamma \geq 40$ Gy/s; P_d the percentage of the OAR volume receiving $d \geq 8$ Gy; FLASH coverage $P_{\gamma,d}$: the percentage of the OAR volume receiving ≥ 40 Gy/s and $d \geq 8$ Gy.

Case	Method	P_γ	P_d	$P_{\gamma,d}$
Brain	PGD	81.5	68.9	68.9
	ADMM	89.1	64.7	64.7
	SCD	96.2	70.1	70.1
	OMP	94.8	63.5	63.5
Lung	PGD	82.0	76.6	76.6
	ADMM	88.8	48.5	48.5
	SCD	95.3	53.3	53.3
	OMP	88.7	60.9	60.9
HN	PGD	99.7	95.5	95.5
	ADMM	90.3	71.7	71.7
	SCD	96.7	71.6	71.6
	OMP	91.9	72.1	72.1

Table 5.

Robust analysis for the brain case.

The dosimetric quantities include RV_{115} for the CTV and RV_{30} for the brainstem.

		PGD	ADMM	SCD	OMP
IMPT	RV_{115}	1.02%	10.99%	12.94%	2.63%
	RV_{30}	1.82%	8.38%	8.68%	7.61%
ARC	RV_{115}	0.83%	2.97%	6.09%	3.39%
	RV_{30}	1.66%	9.07%	3.32%	3.08%
FLASH	RV_{115}	2.45%	21.93%	19.23%	2.67%
	RV_{30}	6.56%	3.12%	5.55%	5.51%