AD_____

Award Number:
W81XWH-12-1-0122

TITLE:
Superior Volumetic Modulated Arc Therapy Planning Solution for Prostate Patients

PRINCIPAL INVESTIGATOR:
Ran Davidi

CONTRACTING ORGANIZATION:
Stanford University Stanford, CA 94305-5847

REPORT DATE:
July 2014

TYPE OF REPORT:
Annual Summary (Final) Report

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
July 2014	Annual Summary (FINAL)	1 April 2012 - 31 March 2014

4. TITLE AND SUBTITLE
Superior Volumetic Modulated Arc Therapy Planning Solution for Prostate Patients

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-12-1-0122

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
RAN DAVIDI

E-Mail: r_davidi@yahoo.com

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Stanford University
Stanford, CA 94305-5847

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Inverse planning is at the heart of prostate Volumetric Modulated Arc Therapy (VMAT) treatment procedure and critically determines its level of success. As practiced now, the capacity of VMAT is greatly underutilized because of inferior computing performance of existing optimization methods. An alternative mathematical approach that improves both the efficiency and the efficacy is needed and is the center of this research. We propose to develop a new innovative inverse planning tool, based on the novel idea of superiorization, to replace the classical constrained optimization approaches employed in clinics today for prostate VMAT cases.

The training award focused on formulating the VMAT problem as a constrained superiorization problem and on the development of a framework of fast converging inverse planning algorithms. Empty solution sets and infeasibility constraints that often exist in real-world applications were incorporated into the model. The new framework was proven mathematically and its efficacy was demonstrated when it was compared to a classical optimization method. The superiorization methodology was Implemented, tested and evaluated on a previously treated prostate case where good results were obtained.

15. SUBJECT TERMS
PROSTATE CANCER, VMAT, SUPERIORIZATION, INVERSE-PLANNING, PROJECTION-METHODS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	103	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Introduction

Inverse planning is at the heart of prostate Volumetric Modulated Arc Therapy (VMAT) treatment procedure and critically determines its level of success. As practiced now, the capacity of VMAT is greatly underutilized because of inferior computing performance of existing optimization methods. An alternative mathematical approach that improves both the efficiency and the efficacy is needed and is the center of this research. We propose to develop a new innovative inverse planning tool, based on the novel idea of *superiorization*, to replace the classical constrained optimization approaches employed in clinics today for prostate VMAT cases.

Towards this goal, year 1 of the training award focused on formulating the VMAT problem as a constrained superiorization problem and on the development of a framework of fast converging inverse planning algorithms. The new approach was then implemented, tested and evaluated on a previously treated prostate cancer case where initial results were obtained. In year 2, the work concentrated on developing further the modality assumed for superiorization when applied to the inverse planning in radiation therapy. Further, the work was implemented and compared with the previous developed method. Towards the overarching goal of the award we expanded the superiorization framework to other applications, such as proton imaging and therapy, to help with the same kind of computational relief that is needed in these types of applications.

Body

1. *Research Accomplishments*

SOW Aim 1: Develop algorithms for inverse planning using superiorization techniques for prostate VMAT

Meeting the goal outlined in the SOW aim 1, we have studied the problem of inverse planning for prostate VMAT and developed a framework of algorithms using the superiorization methodology that is specifically tailored to this application. We first defined the problem mathematically by reformulating it as a linear feasibility problem (instead of a minimization problem) and suggested a solution to solve it using the superiorization methodology. In developing the tools, we have also generalized the approach to include other medical physics applications, and provided conditions that are simple to meet both in theory and in practice. Our claims were proved mathematically and the results were submitted to three journal (archival) publications [2, 4, 11].

Task 1: Formulating the VMAT treatment planning as a constrained superiorization problem

Our approach to a VMAT treatment planning started by studying the current mathematical models used for this application. Since the superiorization methodology requires a different mathematical formulation, the first step was to model the problem accordingly.

Consider the system of equations

$$Ax = d, \tag{1}$$

where A is the $J \times I$ dose matrix that maps any intensity of beamlets vector $x = (x_i)_{i=1}^{I} \in R^I$ onto a dose in voxels vector $d = (d_j)_{j=1}^{J} \in R^J$. Here I is the total number of beamlets and J is the total number of voxels.

The minimization problem can then be formulated as

$$\begin{array}{ll} \text{minimize} & \sum_{s=1}^{S} \lambda_s \left\| A_s x - d_{(s)} \right\|^2 \\ \text{subject to} & x \geq 0, \end{array} \tag{2}$$

where the index s stands for different structures, A_s is the submatrix of A related to structure s and $d_{(s)}$ is the subvector of d related to structure s, respectively, and λ_s is the importance factor associated with the sth structure which is decided by the planner. There is an assumption that x is achievable using apertures (aperture constraints).

Assume that we have S structures, for $s = 1, 2, \ldots, S$, (including the complement of all identified structures), and denote by O_s the set of indices of voxels that belong to the sth structure, such that

$$O_s = \{j_{s,1}, j_{s,2,} \ldots j_{s,m(s)}\} \tag{3}$$

where $m(s)$ is the number of voxels in this structure. Then the system matrix A can be partitioned into blocks

$$A = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_S \end{bmatrix} \tag{4}$$

so that a submatrix A_s will contain the rows from A whose indices appear in O_s, (similarly, let $d_{(s)}$ be the subvector of d whose component indices appear in O_s) and then the system becomes

$$\begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_S \end{bmatrix} x = \begin{pmatrix} d_{(1)} \\ d_{(2)} \\ \vdots \\ d_{(S)} \end{pmatrix}. \tag{5}$$

An optimization method aims at satisfying the system (1) (equivalently (5)) while minimizing a given objective function.

Reformulating the problem as a constrained superiorization problem: We suggest the following modifications to the above modality. Replace the prescription method that gives rise to the system $Ax = d$ in (1) by a more flexible one in which we ask the planner to provide lower- and upper- dose bounds vectors, $\underline{d}$ and $\overline{d}$, respectively, on all voxels in all structures, and instead of (2) we aim at solving the following linear feasibility problem

$$\underline{d} \leq Ax \leq \overline{d}. \tag{6}$$

By transforming the problem of (1) into a linear feasibility problem of the form (6), we allow many iterative projection method to derive a solution. This enables a formulation for the superiorization methodology to be applied to VMAT inverse planning problem since many of these algorithms are also perturbation resilient. Specifically, methods that belong to the two classes of projection methods, String Averaging Projection (SAP) and Block-Iterative Projection (BIP) methods, can be applied towards solving this formulation and achieve finding a superior solution in addition to satisfying the feasibility constraints (see [1, 3]). That is, an x obtainable by a projection method alone will be an intensity of beamlets vector trying to solve (6), while using a projection method that is also perturbation resilient allows for obtaining an x that solves (6) but also provides a solution that is *superior* with respect to an objective function.

The solution vector x of the beamlet intensities that results from the superiorization approach will then be evaluated. Tools such as dose volume histograms (DVHs) will help assess conformality to the prostate (the target) and to the organs at risk (OAR). These will be compared against what is recommended by a physician in the clinic and governed by the specifications of the Radiation Treatment Oncology Group (RTOG) protocol for prostate cancer patients [5].

The adaptation to our model based on the RTOG protocol is as follows: Given a structure s that is an OAR, we define $\overline{d_{(s)}}$ to be the upper-bound subvector of the prescribed dose

$$\overline{d_{(s)}} \equiv d_{(s)}, \tag{7}$$

and define $\underline{d_{(s)}}$ to be a lower-bound subvector for any target structure s

$$\underline{d_{(s)}} \equiv d_{(s)}. \tag{8}$$

This allows the constraints in (6) to be written as

$$0 \leq A_s x \leq d_{(s)}, \tag{9}$$

for an OAR structure s and as

$$d_{(s)} \leq A_s x \leq e_{(s)}, \tag{10}$$

for a target structure s, where $e_{(s)}$ is a clinically-specified upper-bound subvector for the target.

In assessing the solution provided by the superiorization method, if the acceptance criteria is not met, then a refined selection of $\underline{d}$ and $\overline{d}$ will be provided and the process will repeat until a superior feasible solution is found (this step is identical to how it is done in the clinic today).

Task 2: Development of a framework for fast converging inverse planning superiorization techniques
And
Task 7: Investigate the underlying principles and put their concept on a firm mathematical ground

In developing a framework for fast converging inverse planning superiorization techniques we first identified several problems that currently exist in optimization methods. In classical optimization it is assumed that there is a constraints set C and the task is to find an $x \in C$ for which $\phi(x)$ is minimal. Problems with this approach are the following: (1) The constraints may not be consistent and so C could be empty and the optimization task as stated would not have a solution. (2) Iterative methods of classical constrained optimization typically converge to a solution only in the limit. In practice some stopping rule is applied to terminate the process and the actual output at that time may not be in C and, even if it is in C, it is most unlikely to be a minimizer of ϕ over C.

Both problems were addressed in the newly developed superiorization framework. Mathematical definitions and conditions were introduced and were theoretically proven. The new foundations include two new notions of *constraints-compatibility* and *strong perturbation resiliency*. The new concepts allow to take into the modality the infeasibility and practical convergence problems that exist in optimization methods. More specifically, in the superiorization model we suggested to replace the constraints set C by a nonnegative real-valued function Pr that serves as an indicator of how incompatible a given x is with the constraints. Then the merit of an actual output of an algorithm is given by the smallness of the two numbers $Pr(x)$ and $\phi(x)$. Roughly, if an iterative algorithm produces an output x, then its superiorized version will produce an output x' for which $Pr(x')$ is not larger then $Pr(x)$, but (in general) $\phi(x')$ is much smaller than $\phi(x)$.

In addition to the theoretical developments of superiorization, a practical and systematic way was developed to turn any iterative algorithm that solves a feasibility problem into an algorithm that does superiorization. For an iterative algorithm **P** and for any optimization criterion ϕ for which we know how to produce nonascending vectors (see definition p. 5536 in [4]), the following pseudocode automatically takes **P** and produces a version of **P** that is superiorized for ϕ (exact details of the procedure can be found on page 5537 in [4]):

Superiorized Version of the Basic Algorithm

set $k = 0$
set $y^k = y^0$
set $\ell = -1$
repeat
    set $n = 0$
    set $y^{k,n} = y^k$
    while $n<N$
        set $v^{k,n}$ to be a nonascending vector for ϕ at $y^{k,n}$
        set *loop=true*
        while *loop*
            set $\ell = \ell + 1$
            set $\beta_{k,n} = \eta_\ell$
            set $z = y^{k,n} + \beta_{k,n} v^{k,n}$
            if $\phi(z) \leq \phi\left(y^k\right)$ **then**
                set $n=n + 1$
                set $y^{k,n}=z$
                set *loop = false*
    set $y^{k+1}=\boldsymbol{A}_C\left(y^{k,N}\right)$
    set $k = k + 1$

By bridging the gap that typically exist between theory and practice in the new model, superiorization was made more general. That is, the framework fit many other medical physics application, not just VMAT or radiation therapy inverse planning type applications. All the results mentioned briefly here have been published in an archival journal publication in the journal of Medical Physics [4], see the Appendix Section for the full manuscript.

Another accomplishment related to this task touches on an additional aspect of superiorization. Constrained optimization problems that arise in real-life applications are often huge (such an example is the

	$Total\,Variation$ value	Time (seconds)
projected subgradient method	919	2217
superiorization method	873	102

Table 1: Performance comparison of the projected subgradient method and the superiorization method with Total Variation as the objective function.

VMAT problem). It can then happen that the traditional algorithms for constrained optimization require computational resources that are not easily available and, even if they are, the length of time needed to produce an acceptable output is too long to be practicable. As part of our goal to show that superiorization can handle large size problems efficiently, we have illustrated that the computational requirements of a superiorized algorithm can be significantly less than that of a traditional optimization algorithm, by reporting on a comparison of superiorization with the projected subgradient method (PSM), which is a standard method of classical optimization. Table 1 summarizes the comparison we have performed between the PSM method and the superiorization method. In our experiment, we set the the stopping rule to guarantee that the output of the superiorization method is at least as constraints-compatible as the output of the PSM. The superiorization method showed clearly superior efficacy to the PSM: it obtained a result with a lower objective function value (TV) at less than one twentieth of the computational cost.

The complete report that summarizes this work was published in the Journal of Optimization Theory and Applications [2]. It is attached to this report in the Appendix Section.

Task 3: Implementation and testing of the developed algorithms
And
Task 5: Early-stage algorithm testing on a prostate cases
And
Task 6: Testing on clinical data

In these tasks we wanted to assess our proposed approach to using superiorization on a realistically yet simple test case. The goal set here is two-fold: the first is to show that the developed method can produce good results and the second is to obtain a clear indication if the nonacsending-type superiorization techniques should be replaced with alternative derivative-free approaches (see, SOW Task 4: the development of alternative derivative-free techniques to superiorization).

We proposed to use as a test case in this task a previously treated intensity modulated radiation therapy (IMRT) prostate patient case. As was explained in the research proposal, the VMAT technique delivers an IMRT type treatment in a single arc. Getting good results on a previously treated IMRT case would establish a level of confidence that the superiorization method can deliver superior results by referencing a previously treated clinical case. The modality that was given above (in Task 1) is identical for these two radiation therapy techniques (i.e., IMRT and VMAT); the difference lies in the size of the problem and its level of complexity. Since superiorization was never tried with any type of radiation therapy treatment planning it is important to provide such evidence on an actual clinical case. The mathematical model which we have developed along with the theories and proofs of the superiorization methodology (in Task 2) fit both problems. Satisfactory results will encourage us to continue develop the method as it is proposed in Tasks 1 and 2 and tailor it further more to the VMAT approach.

Algorithmic operator and objective function The framework that was developed is quite general for many medical physics applications. With the modality of the superiorization approach in (6), a choice for

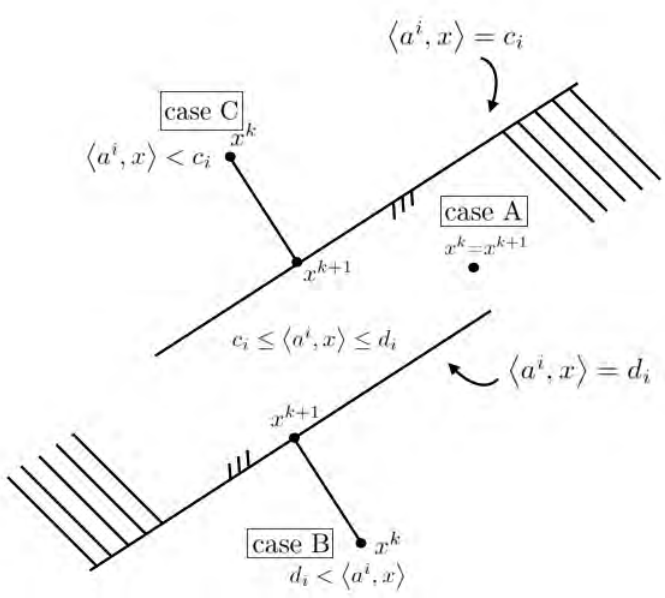


Figure 1: Geometrical description of ART with inequalities constraints

a projection operator that is perturbation resilient is needed as well as a choice of an objective function. The algorithmic operator that was chosen for our implementation was the Algebraic Reconstruction Technique (ART) for inequalities constraints. This operator was proven to be perturbation-resilient in [1]. The constraints of the system in (6) can be thought of as hyperslabs. The algorithm projects the current point according to its location in relation to the two hyperplanes that form a hyperslab. A geometrical description to this feasibility problem is provided in Figure 1. The analytical formulation associated with it is the following:

$$x^{k+1} = \begin{cases} x^k, & \text{if } c_i \le \langle a^i, x^k\rangle \le d_i \text{ (case A)}, \\ x^k + \lambda_k \frac{d_i - \langle a^i, x^k\rangle}{\|a^i\|^2} a^i, & \text{if } d_i < \langle a^i, x^k\rangle \text{ (case B)}, \\ x^k + \lambda_k \frac{c_i - \langle a^i, x^k\rangle}{\|a^i\|^2} a^i, & \text{if } \langle a^i, x^k\rangle < c_i \text{ (case C)}. \end{cases} \tag{11}$$

The objective function used in our implementation was the total variation (TV) functional of the beamlet intensity vector x, see Eq. (12) in [4] and the discussion in the research proposal under Specific Aim 1 regarding this choice. We denote herein the superiorization algorithm that uses TV as the objective function by TV-Superiorization.

Prostate patient data and planning The data for testing the approach were of a previously treated prostate cancer patient. A seven field radiation treatment IMRT plan was created. The organs that were included in the plan were the prostate (target), rectum, bladder, small bowel (OARs) and the full body. Figure 2 shows the CT and the contours of these structures. Using RTOG 0815 [5] we set in Table 2 the acceptance criteria for the implemented TV-Superiorization algorithm.

Figure 2: CT of the prostate patient case used in the experiment.

Organ	Target?	Acceptance criteria
I. Prostate	Yes	1. Dose will be normalized s.t. 98% of the PTV receives the prescription dose. (Prescribed dose to PTV is 79.2 Gy.) 2. The maximum allowable dose within the PTV is 107% of the prescribed dose (i.e., maximum allowed dose is 84.744 Gy). 3. The minimum allowable dose within the PTV should be $\geq$95% of the prescribed dose (i.e., 100% of the dose should be $\geq$75.24 Gy.
II. Rectum	No	1. No more than 15% volume receives dose that exceeds 75 Gy 2. No more than 25% volume receives dose that exceeds 70 Gy 3. No more than 35% volume receives dose that exceeds 65 Gy 4. No more than 50% volume receives dose that exceeds 60 Gy
III. Bladder	No	1. No more than 15% volume receives dose that exceeds 80 Gy 2. No more than 25% volume receives dose that exceeds 75 Gy 3. No more than 35% volume receives dose that exceeds 70 Gy 4. No more than 50% volume receives dose that exceeds 65 Gy
IV. Small Bowel	No	1. Upper bound is set to 52 Gy.

Table 2: Acceptance criteria for prostate patients according to RTOG 0815 [5].

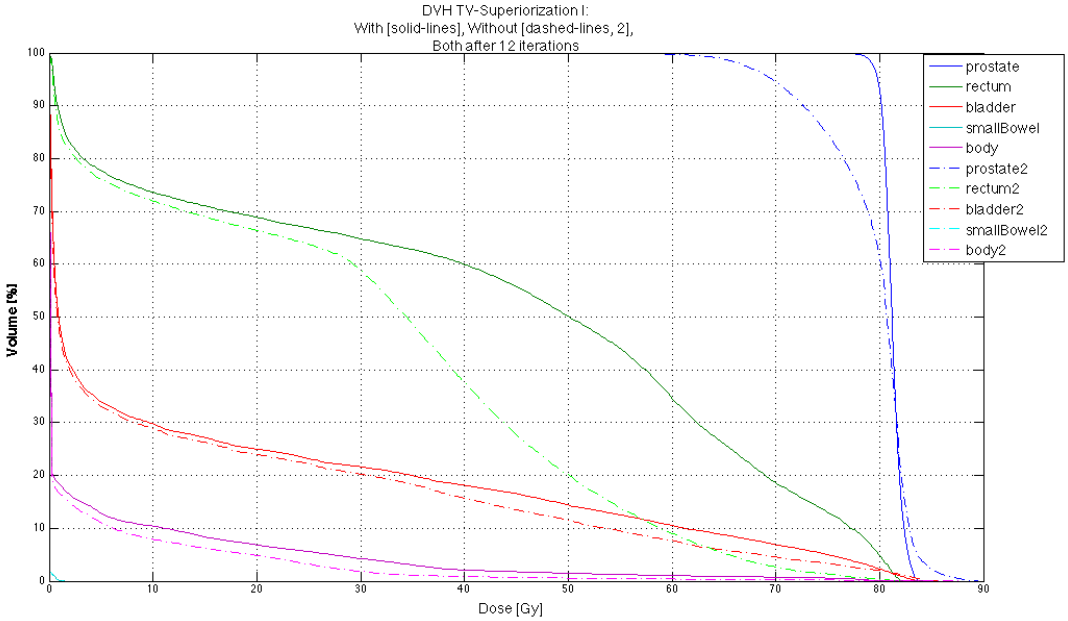


Figure 3: DVH plots for a prostate case experiment with and without TV-Superiorization.

Results We compared the results when superiorization was present versus when it was not. The TV-Superiorization algorithm was able to meet the RTOG acceptance criteria while the one without TV-Superiorization was not. Moreover, the TV-Superiorization algorithm was able to achieve this in a relatively short amount time of only 12 iterations. Figure 3 shows the DVH curves of the two algorithms side-by-side. The solid lines represent the TV-Superiorization algorithm and the dashed lines represent the algorithm without Superiorization. The corresponding numbers for assessing the acceptance criteria are specified in Table 3. As can be seen, the criteria that is based on the RTOG protocol [5] was fully met by the superiorization method for this prostate case.

Task 4: Alternative approach

The goal of this task was modified to reflect the success of tasks 3 and 5. Instead, we developed a different modality to be tested against the original one proposed in Task 1, but also one that is directly inherited from the mathematical foundation laid out in task 7. In the new proposed approach we aimed at removing the linear two-sided inequalities feasibility model in favor of a least-squares model approach.

Quadratic Programming Superiorization (QPS)

Consider the system of equations as in (1) above, i.e., $Ax = d$, where A is the dose matrix mapping the intensity of beamlets vector x to the dose in voxels vector d, where the total number of beamlets is I the total number of voxels is J. Further, in this new work we assume the notation and fundamentals that were used in equation (1)-(5) (not copied here). In the new formulation we propose the following changes: We suggest to use the famous least-squares model of (2) and not our previously suggested model of linear

Organ	Target?	Criterion	TV-Superiorization
I. Prostate	Yes	%vol $>$ 79.2 Gy $=$ 98	%vol $>$ 79.2 Gy $=$ 98
		%vol $>$ 84.744 Gy $=$ 0	%vol $>$ 84.6 Gy $=$ 0
		%vol $>$ 75.24 Gy $=$ 100	%vol $>$ 75.24 Gy $=$ 100
II. Rectum	No	%vol $>$ 75 Gy $\leq$ 15	%vol $>$ 75 Gy $\leq$ 12.7
		%vol $>$ 70 Gy $\leq$ 25	%vol $>$ 70 Gy $\leq$ 18.6
		%vol $>$ 65 Gy $\leq$ 35	%vol $>$ 65 Gy $\leq$ 25.8
		%vol $>$ 60 Gy $\leq$ 50	%vol $>$ 60 Gy $\leq$ 34.5
III. Bladder	No	%vol $>$ 80 Gy $\leq$ 15	%vol $>$ 80 Gy $\leq$ 2.2
		%vol $>$ 75 Gy $\leq$ 25	%vol $>$ 75 Gy $\leq$ 4.9
		%vol $>$ 70 Gy $\leq$ 35	%vol $>$ 70 Gy $\leq$ 6.8
		%vol $>$ 65 Gy $\leq$ 50	%vol $>$ 65 Gy $\leq$ 8.7
IV. Small Bowel	No	%vol $>$ 52 Gy $\leq$ 0	%vol $>$ 1.4 Gy $\leq$ 0

Table 3: Results of the criteria for the TV-Superiorization algorithm.

interval inequalities above. It is also typical to include a second term that will be minimized with the original objective function. Such a term (called a *regularization term*) carry the means for incorporating total variation, i.e.,

$$\begin{array}{ll} \text{minimize} & \sum_{s=1}^{S} \lambda_s \left\| A_s x - d_{(s)} \right\|^2 + \beta TV(x) \\ \text{subject to} & x \geq 0, \end{array} \tag{12}$$

In this work we will not regularize the objective function but follow the superiorization framework as laid out in the pseudocode of the Superiorized Version of the Basic Algorithm (SVoBA). Instead we propose to perform TV superiorization on top of a quadratic programming (QP) algorithm that is intended for the least-squares model. In this way ϕ will be the TV function. Further, the "Basic Algorithm" in line 18 of the SVoBA will be the QP algorithm and not the feasibility-seeking algorithm that was used in the previous section. We next describe in detail the QP algorithm.

The QP algorithm

The QP algorithm was originally designed to solve (2) iteratively. In our framework, whenever line 18 of the SVoBA is called for the algorithmic operator A_C we design it such that it will be an iteration of the QP algorithm. There are many QP algorithms for solving (2). We decided to implement the Projected Landweber method [6, 7] since it fits the superiorization framework (more details below). Let us denote the quadratic objective function of (2) by

$$F(x) = \frac{1}{2} \sum_{s=1}^{S} \lambda_s \left\| A_s x - d_{(s)} \right\|^2. \tag{13}$$

The gradient of F can then be calculated

$$\nabla F(x) = \sum_{s=1}^{S} \lambda_s A_s^T \left(A_s x - d_{(s)} \right) \tag{14}$$

where A_s^T is the transpose matrix of the submatrix A_s. In our pseudocode of SVoBA, when line 18 is reached, in the algorithm $y^{k,N}$ should treat it as the x^k in the Landweber Algorithm and calculate from it the next

The projected Landweber method
Initialization: $x^0 \in R^I$ is arbitrary,
Iterative step: Given the current iterate x^k calculate the next iterate x^{k+1} by

$$x^{k+1} = P_\Omega(x^k - \tau_k \nabla F(x^k)) \tag{15}$$

where $\Omega = \{x \in R^I \mid x_i \geq 0 \text{ for all } i = 1, 2, \ldots, I\}$ is the nonnegative orthant of R^I and P_Ω is the projection onto it, namely, for any point $z \in R^I$

$$(P_\Omega(z))_i = \max(z_i, 0) = \begin{cases} z_i, & \text{if } z_i \geq 0, \\ 0 & \text{if } z_i < 0. \end{cases} \tag{16}$$

The stepsizes: The stepsizes τ_k should be chosen to be either diminishing steps $\tau_k = \frac{0.1}{\sqrt{k}}$ or square summable steps $\tau_k = \frac{1}{k}$.

Organ	Target?	Criterion	QP-Superiorization
I. Prostate	Yes	%vol > 78 Gy = 95	%vol > 78 Gy = 95
		%vol > 84.744 Gy = 0	%vol > 84.70 Gy = 0
		%vol > 75.24 Gy = 100	%vol > 76.02 Gy = 100
II. Rectum	No	%vol > 75 Gy ≤ 15	%vol > 75 Gy ≤ 2.4
		%vol > 70 Gy ≤ 25	%vol > 70 Gy ≤ 3.1
		%vol > 65 Gy ≤ 35	%vol > 65 Gy ≤ 3.6
		%vol > 60 Gy ≤ 50	%vol > 60 Gy ≤ 6.4
III. Bladder	No	%vol > 80 Gy ≤ 15	%vol > 80 Gy ≤ 0
		%vol > 75 Gy ≤ 25	%vol > 75 Gy ≤ 0.8
		%vol > 70 Gy ≤ 35	%vol > 70 Gy ≤ 2.3
		%vol > 65 Gy ≤ 50	%vol > 65 Gy ≤ 4.7
IV. Small Bowel	No	%vol > 52 Gy ≤ 0	%vol > 1.2 Gy ≤ 0

Table 4: Results of the criteria for the QP-Superiorization algorithm.

iterate according to (15), which, in turn, will be the y^{k+1} of line 18 of the SVoBA. The Projected Landweber Method was proven to be perturbation resilient in [8].

Results:

We report here on the results obtained for the same data of a prostate patient (which we reported earlier using the TV-Superiorization method). As can be seen from the Dose Volume Histogram and from the Table, the algorithm was able to produce acceptable results and meet all the criteria. The number of iterations that it needed to reach this result was 20.

Additional tasks that were completed by the PI during the duration of the award (not included in the SOW):

- Implementation of the algorithms and code and its availability to the community: The mathematical foundation behind superiorization will be available to the community in the package framework of SNARK09 in the context of image reconstruction from projections. A paper that summarizes the work was published in the journal of Computer Methods and programs in Biomedicine, see [9] for further

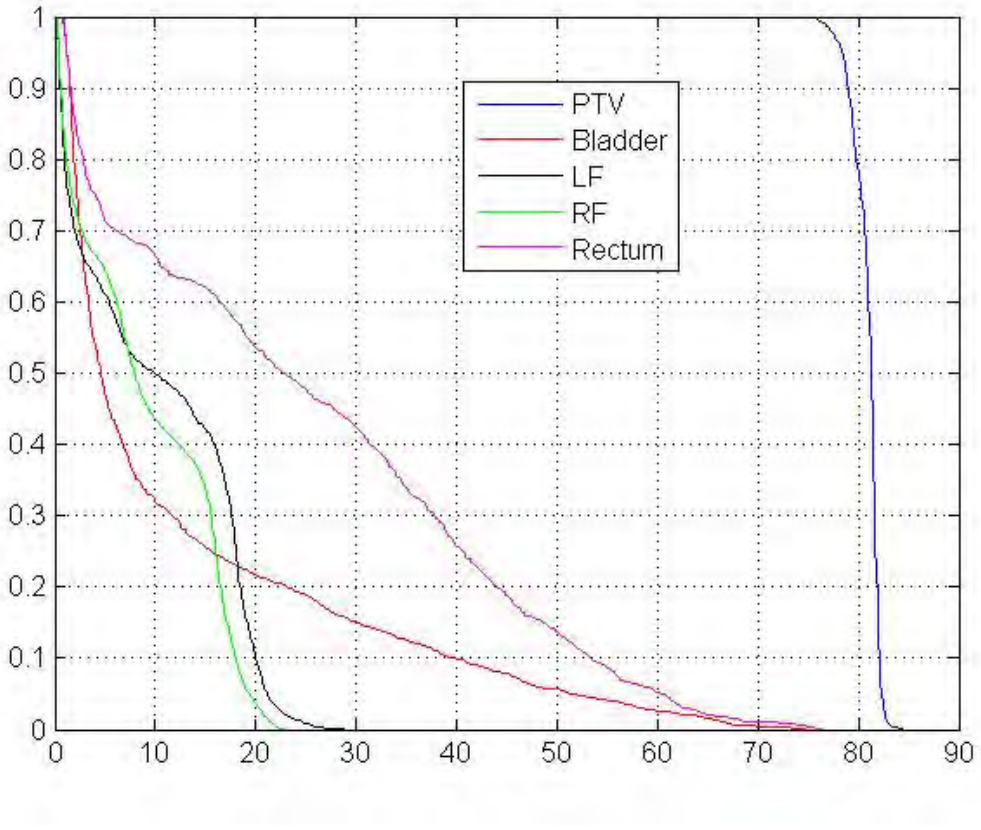


Figure 4: DVH plots for the newly proposed Quadratic Programming Superiorization (QPS).

details (also attached to the appendix).

- The PI participated in a collaborative effort to incorporate the method of superiorization into prostate proton therapy and proton CT imaging techniques. The paper entitled: "200 MeV Proton Radiography Studies with a Hand Phantom Using a Prototype Proton CT Scanner, IEEE Trans Med Imaging 2013", provides the reasoning behind proton CT and its usefulness for using protons to treat cancer as oppose to photons. The acknowledgment section specifies the contribution of the PI and acknowledgment of the grant, [10].
- The PI participated in the development and planning of an Intensity Modulated Proton Therapy (IMpRT) inverse planning method which incorporates principles of the work developed during this award. The attached IMpRT proposal summarizes future aspect of this collaboration and continuation effort among multiple institutes and how this may enrich the career path of the PI.

Training Accomplishments

Task 8: Seminar, lectures and meetings

Task 9: Research training

Task 10: Clinical training

During the duration of the training award the PI had attended regular meetings, seminars and journal clubs with presentations on topics related to radiation therapy treatment planning. Other presentations of visiting scholars and professionals were also available throughout the year and had enriched his knowledge on the topic. The PI was trained in the clinic to operate the Eclipse and Aria system stations for treatment planning available at Stanford Cancer Center (Eclipse and Aria are commercial tools for treatment planning developed by Varian Medical Systems). He collaborated with radiation oncologists, radiation therapists, physicists and dosimetrists and obtained first-hand the knowledge and experience of the process of prostate radiation treatment planning.

Key Research Accomplishments

- Formulated the VMAT treatment planning as a constrained superiorization problem.
- Developed a framework for fast converging inverse planning superiorization techniques.
- Derived the necessary conditions of the superiorization framework for VMAT treatment planning
- Placed the newly developed concepts on a firm mathematical ground.
- Implemented and tested the new superiorization framework and showed good initial results.
- Developed implemented and tested a new modality based on Quadratic Programing and incorporated it into the superiorization framework.
- Participated in additional collaborations for using the developed superiorization framework in other applications including: image reconstruction, proton CT and proton radiation therapy.
- Trained in treating prostate cancer as it is done in the clinic today.
- Invited speaker to international meetings and workshops.

Reportable Outcomes

- Four journal publications were submitted. Three appeared and the fourth was accepted:

  1. G.T. Herman, E. Garduño, R. Davidi and Y. Censor, Superiorization: An optimization heuristic for medical physics, *Medical Physics* **39** (2012), 5532–5546.
  2. Y. Censor, R. Davidi, G.T. Herman, R.W. Schulte and L. Tetruashvili, Projected subgradient minimization versus superiorization, Journal of Optimization Theory and Applications, DOI 10.1007/s10957-013-0408-3, 2013.
  3. J. Klukowska, R. Davidi, and G.T. Herman: SNARK09 - A software package for the reconstruction of 2D images from 1D projections, Computer Methods and Programs in Biomedicine, 110:424-440, 2013.
  4. R. Davidi, Y. Censor, R.W. Schulte, S. Geneser, and L. Xing. Feasibility-Seeking and Superiorization Algorithms Applied to Inverse Treatment Planning in Radiation Therapy, Contemporary Mathematics, (to appear), 2014.

- The above work has been accepted for presentation at the joint workshop sponsored by the American Society for Therapeutic Radiology and Oncology (ASTRO), the National Cancer Institute (NCI) and the American Association of Physicists in Medicine (AAPM) , June 13-14, 2013, National Institutes of Health, Bethesda, MD, USA.

- The above work has been accepted for presentation at the workshop on Projection Methods: Theory and Practice, June 19-21, 2013, Fraunhofer Institute for Industrial Mathematics ITWM, Kaiserslautern, Germany.

- The above work has been accepted for presentation at the meeting on Projection Methods in Feasibility, Superiorization and Optimization, December 19, 2013, Center for Mathematics and Scientific Computation (CMSC) and the Caesarea Rothschild Institute (CRI) for Interdisciplinary Applications of Computer Science at the University of Haifa, Mt. Carmel, Israel.

Conclusion

We were able to extend the superiorization methodology into a larger framework, one that is more realistic from the point of view of the application at hand. By taking into account the discrepancy that exist between theory and practice and incorporate it into our model, we minimized potential issues that typically appear when a theory is applied to a real-life application.

Superiorization was developed to be a general tool for medical physics applications. It is capable of turning any iterative algorithm that tries to satisfy a set of constraints into one that is also capable of superiorizing an objective function. The work that came out of this research can help other applications that use optimization methods as the main tool. Examples include X-ray CT using image reconstruction from projections, proton CT that uses superiorization as the iterative engine to superiorize an objective function, and utilizing its eficacy for implementing fast converging techniques in proton radiation therapy.

Using the developed methodology, we tailored it specifically to solve the problem of IMRT and VMAT in radiation therapy inverse planning. The initial results obtained on a realistic prostate case were satisfactory and show good indication that superiorization works and can be applied to a radiation treatment planning

problems. We further extended the framework to include quadratic programming and provided the means to use superiorization using this approach.

Finally, we implemented, tested and evaluated our new framework on prostate cases. We performed a thorough investigation as detailed in the SOW and reported in the literature and in meetings on our results.

References

[1] D. Butnariu, R. Davidi, G.T. Herman and I.G. Kazantsev, Stable convergence behavior under summable perturbations of a class of projection methods for convex feasibility and optimization problems, *IEEE Journal of Selected Topics in Signal Processing* **1** (2007), 540–547.

[2] Y. Censor, R. Davidi, G.T. Herman, R.W. Schulte and L. Tetruashvili, Projected subgradient minimization versus superiorization, Journal of Optimization Theory and Applications, **160** (2014), 730–747.

[3] R. Davidi, G.T. Herman and Y. Censor, Perturbation-resilient block-iterative projection methods with application to image reconstruction from projections, *International Transactions in Operational Research* **16** (2009), 505–524.

[4] G.T. Herman, E. Garduño, R. Davidi and Y. Censor, Superiorization: An optimization heuristic for medical physics, *Medical Physics* **39** (2012), 5532–5546.

[5] A. Martinez *et al.* A phase III prospective randomized trial of dose-escalated radiotherapy with or without short-term androgen deprivation therapy for patients with intermediate-risk prostate cancer. Last accessed April 19, 2013. http://www.rtog.org/clinicaltrials/protocoltable/studydetails.aspx?action=openFile&FileID=4662

[6] M. Bertero and P. Boccacci, *Introduction to Inverse Problems in Imaging*, Institute of Physics Publishing, Bristol, UK, 1998.

[7] B. Johansson, T. Elfving, V. Kozlov, Y. Censor, P.-E. Forssén and G. Granlund, The application of an oblique-projected Landweber method to a model of supervised learning, *Mathematical and Computer Modelling* **43** (2006), 892–909. DOI:10.1016/j.mcm.2005.12.010.

[8] W. Jin, Y. Censor and M. Jiang, Convergence of projected scaled gradient methods with summable perturbations, *Technical report*, December 2013.

[9] J. Klukowska, R. Davidi, G.T. Herman, SNARK09 - A software package for the reconstruction of 2D images from 1D projections, Computer Methods and Programs in Biomedicine, **110** (2013), 424–440.

[10] T. Plautz, V. Bashkirov, V. Feng, F. Hurley, R.P. Johnson, C. Leary, S. Macafee, A. Plumb, V. Rykalin, H.F. Sadrozinski, K. Schubert, R. Schulte, B. Schultze, D. Steinberg, M. Witt, A. Zatserklyaniy, 200 MeV proton radiography studies with a hand phantom using a prototype proton CT scanner, IEEE Trans Med Imaging, **33** (2014), 875–881.

[11] R. Davidi, Y. Censor, R.W. Schulte, S. Geneser, and L. Xing. Feasibility-Seeking and Superiorization Algorithms Applied to Inverse Treatment Planning in Radiation Therapy, Contemporary Mathematics, (2014), to appear.

Appendices

Superiorization: An optimization heuristic for medical physics

Gabor T. Herman[a)]
Department of Computer Science, The Graduate Center, City University of New York, New York, New York 10016

Edgar Garduño
Departamento de Ciencias de la Computación, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Cd. Universitaria, Mexico City C.P. 04510, Mexico

Ran Davidi
Department of Radiation Oncology, Stanford University, Stanford, California 94305

Yair Censor
Department of Mathematics, University of Haifa, Mt. Carmel, 31905 Haifa, Israel

(Received 12 January 2012; revised 28 July 2012; accepted for publication 29 July 2012; published 22 August 2012)

Purpose: To describe and mathematically validate the superiorization methodology, which is a recently developed heuristic approach to optimization, and to discuss its applicability to medical physics problem formulations that specify the desired solution (of physically given or otherwise obtained constraints) by an optimization criterion.
Methods: The superiorization methodology is presented as a heuristic solver for a large class of constrained optimization problems. The constraints come from the desire to produce a solution that is constraints-compatible, in the sense of meeting requirements provided by physically or otherwise obtained constraints. The underlying idea is that many iterative algorithms for finding such a solution are perturbation resilient in the sense that, even if certain kinds of changes are made at the end of each iterative step, the algorithm still produces a constraints-compatible solution. This property is exploited by using permitted changes to steer the algorithm to a solution that is not only constraints-compatible, but is also desirable according to a specified optimization criterion. The approach is very general, it is applicable to many iterative procedures and optimization criteria used in medical physics.
Results: The main practical contribution is a procedure for automatically producing from any given iterative algorithm its superiorized version, which will supply solutions that are superior according to a given optimization criterion. It is shown that if the original iterative algorithm satisfies certain mathematical conditions, then the output of its superiorized version is guaranteed to be as constraints-compatible as the output of the original algorithm, but it is superior to the latter according to the optimization criterion. This intuitive description is made precise in the paper and the stated claims are rigorously proved. Superiorization is illustrated on simulated computerized tomography data of a head cross section and, in spite of its generality, superiorization is shown to be competitive to an optimization algorithm that is specifically designed to minimize total variation.
Conclusions: The range of applicability of superiorization to constrained optimization problems is very large. Its major utility is in the automatic nature of producing a superiorization algorithm from an algorithm aimed at only constraints-compatibility; while nonheuristic (exact) approaches need to be redesigned for a new optimization criterion. Thus superiorization provides a quick route to algorithms for the practical solution of constrained optimization problems. © *2012 American Association of Physicists in Medicine*. [http://dx.doi.org/10.1118/1.4745566]

Key words: superiorization, constrained optimization, heuristic optimization, tomography, total variation

I. INTRODUCTION

Optimization is a tool that is used in many areas of Medical Physics. Prime examples are radiation therapy treatment planning and tomographic reconstruction, but there are others such as image registration. Some well-cited classical publications on the topic are Refs. 1–12 and some recent articles are Refs. 13–26.

In a typical medical physics application, one uses *constrained optimization*, where the constraints come from the desire to produce a solution that is *constraints-compatible*, in the sense of meeting the requirements provided by physically or otherwise obtained constraints. In radiation therapy treatment planning, the requirements are usually in the form of constraints prescribed by the treatment planner on the doses to be delivered at specific locations in the body. These doses in turn depend on information provided by an imaging instrument, typically a magnetic resonance imaging (MRI) or a computerized tomography (CT) scanner. In tomography, the constraints come from the detector readings of the instrument.

In such applications, it is typically the case that a large number of solutions would be considered good enough from the point of view of being constraints-compatible; to a large extent, but not entirely, due to the fact that there is uncertainty as to the exact nature of the constraints (for example, due to noise in the data collection). In such a case, an optimization criterion is introduced that helps us to distinguish the "better" constraints-compatible solutions (for example, this criterion could be the total dose to be delivered to the body, which may vary quite a bit between radiation therapy treatment plans that are compatible with the constraints on the doses delivered to individual locations).

The superiorization methodology (see, for example, Refs. 22 and 27–32) is a recently developed heuristic approach to optimization. The word *heuristic* is used here in the sense that the process is not guaranteed to lead to an optimum according to the given criterion; approaches aimed at processes that are guaranteed in that sense are usually referred to as *exact*. Heuristic approaches have been found useful in practical applications of optimization, mainly because they are often computationally much less expensive than their exact counterparts, but nevertheless provide solutions that are appropriate for the application at hand.[33–35]

The underlying idea of the superiorization approach is the following. In many applications there exists a computationally efficient iterative algorithm that produces a constraints-compatible solution for the given constraints. (An example of this for radiation therapy treatment planning is reported in Ref. 36, its clinical use is discussed in Ref. 15.) Furthermore, often the algorithm is *perturbation resilient* in the sense that, even if certain kinds of changes are made at the end of each iterative step, the algorithm still produces a constraints-compatible solution.[27–30] This property is exploited in the superiorization approach by using such perturbations to steer the algorithm to a solution that is not only constraints-compatible, but is also desirable according to a specified optimization criterion. The approach is very general, it is applicable to many iterative procedures and optimization criteria.

The current paper presents a major advance in the practice and theory of superiorization. The previous publications[22,27–32] used the intuitive idea to present some superiorization algorithms, in this paper the reader will find a totally automatic procedure that turns an iterative algorithm into its superiorized version. This version will produce an output that is as constraints-compatible as the output of the original algorithm, but it is superior to that according to an optimization criterion. This claim is mathematically shown to be true for a very large class of iterative algorithms and for optimization criteria in general, typical restrictions (such as convexity) on the optimization criterion are not essential for the material presented below. In order to make precise and validate this broad claim, we present here a new theoretical framework. The framework of Ref. 29 is a precursor of what we present here, but it is a restricted one, since it assumes that the constraints can be all satisfied simultaneously, which is often false in medical physics applications. There is no such restriction in the presentation below.

The idea of designing algorithms that use interlacing steps of two different kinds (in our case, one kind of steps aim at constraints-compatibility and the other kind of steps aim at improvement of the optimization criterion) is well-established and, in fact, is made use of in many approaches that have been proposed with exact constrained optimization in mind; see, for example, the works of Helou Neto and De Pierro,[37,38] Nurminski,[39] Combettes and co-workers,[40,41] Sidky and co-workers,[23,42,43] and Defrise and co-workers.[44] However, none of these approaches can do what can be done by the superiorization approach as presented below, namely, the automatic production of a heuristic constrained optimization algorithm from an iterative algorithm for constraints-compatibility. For example, in Ref. 37 it is assumed (just as in the theory presented in our Ref. 29) that all the constraints can be satisfied simultaneously.

A major motivator for the additional theory presented in the current paper is to get rid of this assumption, which is not reasonable when handling real problems of medical physics. Motivated by similar considerations, Helou Neto and De Pierro[38] present an alternative approach that does not require this unreasonable assumption. However, in order to solve such a problem, they end up with iterative algorithms of a particular form rather than having the generality of being able to turn any constraints-compatibility seeking algorithm into a superiorized one capable of handling constrained optimization. Also, the assumptions they have to make in order to prove their convergence result (their Theorem 15) indicate that their approach is applicable to a smaller class of constrained optimization problems than the superiorization approach whose applicability seems to be more general. However, for the mathematical purist, we point out that they present an exact constrained optimization algorithm, while superiorization is a heuristic approach. Whether this is relevant to medical physics practice is not clear: exact algorithms are not run forever, but are stopped according to some stopping-rule, the relevant questions in comparing two algorithms are the quality of the actual output and the computation time needed to obtain it.

Ultimately, the quality of the outputs should be evaluated by some figures of merit relevant to the medical task at hand. An example of a careful study of this kind that involves superiorization is in Sec. 4.3 of Ref. 30, which reports on comparing in CT the efficacy of constrained optimization reconstruction algorithms for the detection of low-contrast brain tumors by using the method of statistical hypothesis testing (which provides a P-value that indicates the significance by which we can reject the null hypothesis that the two algorithms are equally efficacious in favor of the alternative that one is preferable). Such studies bundle together two things: (i) the formulation of the constrained optimization task and (ii) the performance of the algorithm in performing that task. The first of these requires a translation of the medical aim into a mathematical model, it is important that this model should be appropriately chosen.

The superiorization approach is not about choosing this model, it kicks in once the model is chosen and aims at producing an output that is "good" according to the

mathematical specifications of the constraints and of the optimization criterion. Thus superiorization has been used to compare the effects on the quality of the output in CT when the optimization criterion is specified by total variation (TV) versus by entropy[28] or versus by the ℓ_1-norm of the Haar transform.[32] However, the current paper is not about discussing how to translate the underlying medical physics task into a constrained optimization problem. For our purposes here, we are assuming that the mathematical model has been worked out and concentrate on the algorithmic approach for solving the resulting constrained optimization problem. We claim that the evaluation of such algorithms should not be based on the medical figures of merit mentioned at the beginning of the previous paragraph, but rather on their performance in solving the mathematical problem. If "good" solutions to the constrained optimization problem are not medically efficacious, that indicates that something is wrong with the mathematical model and not that something is wrong with the algorithmic approach. For this reason, in this paper we will not carry out a careful investigation of the medical efficacy of any algorithm in the manner that we have done in Sec. 4.3 of Ref. 30, but will restrict ourselves to a simple illustration of the performance of the superiorization approach as compared to the previously published algorithm of Ref. 42 that is aimed at performing exact minimization.

Examples of such studies already exist. Superiorization was compared in Ref. 27 with Algorithm 6 of Ref. 40 and in Ref. 45 with the algorithm of Goldstein and Osher that they refer to as TwIST (Ref. 46) with split Bregman[47] as the substep. In both cases the implementation was done by the proposers of the algorithms. In these reported instances superiorization did well: the constraints-compatibility and the value of the function to be minimized were very similar for the outputs produced by the algorithms being compared, but the superiorization algorithm produced its output four times faster than the alternative. It would be unjustified to draw any general conclusions on the mathematical performance and speed of superiorization based on just a few experiments, but the reported results are encouraging.

However, the main reason why we advocate superiorization is different from what is discussed above. The reason why we claim it to be helpful in medical physics research is that it has the potential of saving a lot of time and effort for the researcher. Let us consider a historical example. Likelihood optimization using the iterative process of expectation maximization (EM) (Ref. 48) gained immediate and wide acceptance in the emission tomography community. It was observed that irregular high amplitude patterns occurred in the image with a large number of iterations, but it was not until five years later that this problem was corrected[49] by the use of a maximum a posteriority probability (MAP) algorithm with a multivariate Gaussian prior. Had we had at our disposal the superiorization approach, then the introduction of an optimization criterion (Gaussian or other) into the iterative EM process would have been a simple matter and we would have saved the time and effort spent on designing a special purpose algorithm for the MAP formulation. A TV-superiorization of the EM algorithm is presented in Ref. 50.

Even though our major claim for superiorization is that it provides a quick route to algorithms for the practical solution of constrained optimization problems, before leaving this introduction let us bring up a question that has to do with the performance of the resulting algorithms: Will superiorization produce superior results to those produced by contemporary MAP methods or is it faster than the better of such methods? At this stage we have not yet developed the mathematical notation to discuss this question in a rigorous manner, we return to it in Subsection II.F.

In Sec. III, we present in detail the superiorization methodology. In Sec. III, we provide an illustrative example by reporting on reconstructions produced by algorithms applied to simulated computerized tomography data of a head cross section. In the final section, we discuss our results and present our conclusions.

II. THE SUPERIORIZATION METHODOLOGY

II.A. Problem sets, proximity functions, and ε-compatibility

Although optimization is often studied in a more general context (such as in Hilbert or Banach spaces), in medical physics we usually deal with a special case, where optimization is performed in a *Euclidean space* $\mathbb{R}^J$ (the space of J-dimensional vectors of real numbers, where J is a positive integer). As often appropriate in practice, we further restrict the domain of optimization to a nonempty subset Ω of $\mathbb{R}^J$ (such as the *non-negative orthant* $\mathbb{R}^J_+$ that consists of vectors all of whose components are non-negative).

We now turn to formalizing the notion of being compatible with given constraints, a notion that we have used informally in Sec. I. In any application, there is a *problem set* $\mathbb{T}$; each *problem* $T \in \mathbb{T}$ is essentially a description of the constraints in that particular case. For example, for a tomographic scanner, the problem of reconstruction for a particular patient at a particular time is determined by the measurements taken by the scanner for that patient at that time. The intuitive notion of constraints-compatibility is formalized by the use of a *proximity function* $\mathcal{P}r$ on $\mathbb{T}$ such that, for every $T \in \mathbb{T}$, $\mathcal{P}r_T$ maps Ω into $\mathbb{R}_+$, the set of non-negative real numbers; i.e., $\mathcal{P}r_T : \Omega \to \mathbb{R}_+$. Intuitively, we think of $\mathcal{P}r_T(\boldsymbol{x})$ as an indicator of how incompatible $\boldsymbol{x}$ is with the constraints of T. For example, in tomography, $\mathcal{P}r_T(\boldsymbol{x})$ should indicate by how much a proposed reconstruction that is described by an $\boldsymbol{x}$ in Ω violates the constraints of the problem T that are provided by the measurements taken by the scanner. For example, if we use $\boldsymbol{b}$ to denote the vector of estimated line integrals based on the measurements obtained by the scanner and by $\boldsymbol{A}$ the system matrix of the scanner, then a possible choice for the proximity function is the norm-distance $\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|$, which we will use as an example in the discussions that follow. An alternative legitimate choice for the proximity function is the Kullback-Leibler distance $KL(\boldsymbol{b}, \boldsymbol{A}\boldsymbol{x})$, which is the negative log-likelihood of a statistical model in tomography. The

special case $\mathcal{P}r_T(\boldsymbol{x}) = 0$ is interpreted by saying that $\boldsymbol{x}$ is perfectly compatible with the constraints; due to the presence of noise in practical applications, it is quite conceivable that there is no $\boldsymbol{x}$ that is perfectly compatible with the constraints, and we accept an $\boldsymbol{x}$ as constraints-compatible as long as the value of $\mathcal{P}r_T(\boldsymbol{x})$ is considered to be small enough to justify that decision. Combining these two concepts leads to the notion of a *problem structure*, which is a pair $\langle \mathbb{T}, \mathcal{P}r \rangle$, where $\mathbb{T}$ is a nonempty problem set and $\mathcal{P}r$ is a proximity function on $\mathbb{T}$. For a problem structure $\langle \mathbb{T}, \mathcal{P}r \rangle$, a problem $T \in \mathbb{T}$, a non-negative ε, and an $\boldsymbol{x} \in \Omega$, we say that $\boldsymbol{x}$ is *ε-compatible* with T provided that $\mathcal{P}r_T(\boldsymbol{x}) \leq \varepsilon$.

As an example (whose applicability to tomographic reconstruction is illustrated in Sec. III), consider the problem structure that arises from the desire to find non-negative solutions of sequences of blocks of linear equations. Then the appropriate choices are $\Omega = \mathbb{R}^J_+$ and the problem structure is $\langle \mathbb{S}, Res \rangle$, where the problem set $\mathbb{S}$ is

$$\begin{aligned}\mathbb{S} = \{&(\{(\boldsymbol{a}^1, b_1), \ldots, (\boldsymbol{a}^{\ell_1}, b_{\ell_1})\}, \ldots,\\ &\{(\boldsymbol{a}^{\ell_1+\ldots+\ell_{W-1}+1}, b_{\ell_1+\ldots+\ell_{W-1}+1}), \ldots, (\boldsymbol{a}^{\ell_1+\ldots+\ell_W}, b_{\ell_1+\ldots+\ell_W})\}) |\\ &W \text{ is a positive integer and,}\\ &\text{for } 1 \leq w \leq W,\ \ell_w \text{ is a positive integer and,}\\ &\text{for } 1 \leq i \leq \ell_1 + \ldots + \ell_W,\ \boldsymbol{a}^i \in \mathbb{R}^J \text{ and } b_i \in \mathbb{R}\} \end{aligned} \tag{1}$$

and the proximity function *Res* on $\mathbb{S}$ is defined, for any problem $S = (\{(\boldsymbol{a}^1, b_1), \ldots, (\boldsymbol{a}^{\ell_1}, b_{\ell_1})\}, \ldots, \{(\boldsymbol{a}^{\ell_1+\ldots+\ell_{W-1}+1}, b_{\ell_1+\ldots+\ell_{W-1}+1}), \ldots, (\boldsymbol{a}^{\ell_1+\ldots+\ell_W}, b_{\ell_1+\ldots+\ell_W})\})$ in $\mathbb{S}$ and for any $\boldsymbol{x} \in \Omega$, by

$$Res_S(\boldsymbol{x}) = \sqrt{\sum_{i=1}^{\ell_1+\ldots+\ell_W} (b_i - \langle \boldsymbol{a}^i, \boldsymbol{x} \rangle)^2}. \tag{2}$$

Note that each element of this problem set $\mathbb{S}$ specifies an ordered sequence of W blocks of linear equations of the form $\langle \boldsymbol{a}^i, \boldsymbol{x} \rangle = b_i$ where $\langle *, * \rangle$ denotes the inner product in $\mathbb{R}^J$ (and thus $\mathbb{S}$ is an appropriate representation of the so-called "ordered subsets" approach to tomographic reconstruction,[51] as well as of other earlier-published block-iterative methods that proposed essentially the same idea[52–54]). The proximity function *Res* on $\mathbb{S}$ is the *residual* that we get when a particular $\boldsymbol{x}$ is substituted into all the equations of a particular problem S.

II.B. Algorithms and outputs

We now define the concept of an algorithm in the general context of problem structures. For technical reasons that will become clear as we proceed with our development, we introduce an additional set Δ, such that $\Omega \subseteq \Delta \subseteq \mathbb{R}^J$. (Both Ω and Δ are assumed to be known and fixed for any particular problem structure $\langle \mathbb{T}, \mathcal{P}r \rangle$.) An *algorithm* $\mathbf{P}$ for a problem structure $\langle \mathbb{T}, \mathcal{P}r \rangle$ assigns to each problem $T \in \mathbb{T}$ an operator $\mathbf{P}_T : \Delta \to \Omega$. This definition is used to define iterative processes that, for any *initial point* $\boldsymbol{x} \in \Omega$, produce the (potentially) infinite sequence $((\mathbf{P}_T)^k \boldsymbol{x})_{k=0}^{\infty}$ (that is, the sequence $\boldsymbol{x}, \mathbf{P}_T\boldsymbol{x}, \mathbf{P}_T(\mathbf{P}_T\boldsymbol{x}), \cdots$) of points in Ω. We discuss below how such a potentially infinite process is terminated in practice.

Selecting $\Omega = \mathbb{R}^J_+$ and $\Delta = \mathbb{R}^J$ for the problem structure $\langle \mathbb{S}, Res \rangle$ of Subsection II.A, an example of an algorithm $\mathbf{R}$ is specified by

$$\mathbf{R}_S\boldsymbol{x} = \mathbf{Q}\mathbf{B}_{S_W} \cdots \mathbf{B}_{S_1}\boldsymbol{x}, \tag{3}$$

where S is the problem specified above in Eq. (2) and, for $1 \leq w \leq W$, $\mathbf{B}_{S_w} : \Delta \to \Delta$ is defined by

$$\mathbf{B}_{S_w}\boldsymbol{x} = \boldsymbol{x} + \frac{1}{\ell_w} \sum_{i=\ell_1+\ldots+\ell_{w-1}+1}^{\ell_1+\ldots+\ell_w} \frac{b_i - \langle \boldsymbol{a}^i, \boldsymbol{x} \rangle}{\|\boldsymbol{a}^i\|^2} \boldsymbol{a}^i, \tag{4}$$

where $\|\boldsymbol{a}\|$ denotes the norm of the vector $\boldsymbol{a}$ in $\mathbb{R}^J$, and $\mathbf{Q} : \Delta \to \Omega$ is defined by

$$(\mathbf{Q}\boldsymbol{x})_j = \max\{0, \boldsymbol{x}_j\}, \text{ for } 1 \leq j \leq J. \tag{5}$$

Note that $\mathbf{R}_S : \Delta \to \Omega$. This specific algorithm $\mathbf{R}$ is a typical example of the so-called block-iterative methods mentioned above. Except for the presence of $\mathbf{Q}$ in Eq. (3), which enforces non-negativity of the components, it is identical to an algorithm used and illustrated in Ref. 31. With the $\mathbf{Q}$ absent from the definition of the algorithm, Ω has to be the whole of $\mathbb{R}^J$; the practical consequence of the presence versus the absence of $\mathbf{Q}$ in the tomographic application is illustrated in Subsection III.D. We also note that special cases of the presented algorithm include the classical reconstruction methods such as algebraic reconstruction technique (ART) (if $\ell_w = 1$, for $1 \leq w \leq W$) and SIRT (if $W = 1$); see, for example, Chaps. 11 and 12 of Ref. 55.

For a problem structure $\langle \mathbb{T}, \mathcal{P}r \rangle$, a $T \in \mathbb{T}$, an $\varepsilon \in \mathbb{R}_+$, and a sequence $R = (\boldsymbol{x}^k)_{k=0}^{\infty}$ of points in Ω, we use $O(T, \varepsilon, R)$ to denote the $\boldsymbol{x} \in \Omega$ that has the following properties: $\mathcal{P}r_T(\boldsymbol{x}) \leq \varepsilon$ and there is a non-negative integer K such that $\boldsymbol{x}^K = \boldsymbol{x}$ and, for all non-negative integers $k < K \mathcal{P}r_T(\boldsymbol{x}^k) > \varepsilon$. Clearly, if there is such an $\boldsymbol{x}$, then it is unique. If there is no such $\boldsymbol{x}$, then we say that $O(T, \varepsilon, R)$ is *undefined*, otherwise we say that it is *defined*. The intuition behind this definition is the following: if we think of R as the (infinite) sequence of points that is produced by an algorithm (intended for the problem T) without a termination criterion, then $O(T, \varepsilon, R)$ is the *output* produced by that algorithm when we add to it instructions that make it terminate as soon as it reaches a point that is ε-compatible with T.

II.C. Bounded perturbation resilience

The notion of a *bounded perturbations resilient* algorithm $\mathbf{P}$ for a problem structure $\langle \mathbb{T}, \mathcal{P}r \rangle$ has been defined in a mathematically precise manner.[29] However, that definition is not satisfactory from the point of view of applications in medical physics (or indeed in any area involving noisy data), because it is useful only for problems T for which there is a perfectly compatible solution (that is, an $\boldsymbol{x}$ such that $\mathcal{P}r_T(\boldsymbol{x}) = 0$). We therefore extend here that notion as follows. An algorithm $\mathbf{P}$ for a problem structure $\langle \mathbb{T}, \mathcal{P}r \rangle$ is said to be *strongly perturbation resilient* if, for all $T \in \mathbb{T}$,

(i) there exists an $\varepsilon \in \mathbb{R}_+$ such that $O(T, \varepsilon, ((\mathbf{P}_T)^k \boldsymbol{x})_{k=0}^{\infty})$ is defined for every $\boldsymbol{x} \in \Omega$;

(ii) for all $\varepsilon \in \mathbb{R}_+$ such that $O(T, \varepsilon, ((\mathbf{P}_T)^k \boldsymbol{x})_{k=0}^{\infty})$ is defined for every $\boldsymbol{x} \in \Omega$, we also have that $O(T, \varepsilon', R)$ is defined for every $\varepsilon' > \varepsilon$ and for every sequence $R = (\boldsymbol{x}^k)_{k=0}^{\infty}$ of points in Ω generated by

$$\boldsymbol{x}^{k+1} = \mathbf{P}_T(\boldsymbol{x}^k + \beta_k \boldsymbol{v}^k), \text{ for all } k \geq 0, \tag{6}$$

where $\beta_k \boldsymbol{v}^k$ are *bounded perturbations*, meaning that the sequence $(\beta_k)_{k=0}^{\infty}$ of non-negative real numbers is *summable* (that is, $\sum_{k=0}^{\infty} \beta_k < \infty$), the sequence $(\boldsymbol{v}^k)_{k=0}^{\infty}$ of vectors in $\mathbb{R}^J$ is bounded and, for all $k \geq 0$, $\boldsymbol{x}^k + \beta_k \boldsymbol{v}^k \in \Delta$.

In less formal terms, the second of these properties says that for a strongly perturbation resilient algorithm we have that, for every problem and any non-negative real number ε, if it is the case that for all initial points from Ω the infinite sequence produced by the algorithm contains an ε-compatible point, then it will also be the case that all perturbed sequences satisfying Eq. (6) contain an ε'-compatible point, for any $\varepsilon' > \varepsilon$.

Having defined the notion of a strongly perturbation resilient algorithm, we next show that this notion is of relevance to problems in medical physics. We illustrate the use of this in tomography in Sec. III. We first need to introduce some mathematical concepts.

Given an algorithm $\mathbf{P}$ for a problem structure $\langle \mathbb{T}, \mathcal{P}r \rangle$ and a $T \in \mathbb{T}$, we say that $\mathbf{P}$ is *convergent for* T if, for every $\boldsymbol{x} \in \Omega$, there exists a unique $\boldsymbol{y}(\boldsymbol{x}) \in \Omega$ such that, $\lim_{k\to\infty} (\mathbf{P}_T)^k \boldsymbol{x} = \boldsymbol{y}(\boldsymbol{x})$, meaning that for every positive real number δ, there exists a non-negative integer K, such that $\|(\mathbf{P}_T)^k \boldsymbol{x} - \boldsymbol{y}(\boldsymbol{x})\| \leq \delta$, for all non-negative integers $k \geq K$. If, in addition, there exists a $\gamma \in \mathbb{R}_+$ such that $\mathcal{P}r_T(\boldsymbol{y}(\boldsymbol{x})) \leq \gamma$, for every $\boldsymbol{x} \in \Omega$, then we say that $\mathbf{P}$ is *boundedly convergent for* T.

A function $f : \Omega \to \mathbb{R}$ is *uniformly continuous* if, for every $\varepsilon > 0$ there exists a $\delta > 0$, such that, for all $\boldsymbol{x}, \boldsymbol{y} \in \Omega$, $|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq \varepsilon$ provided that $\|\boldsymbol{x} - \boldsymbol{y}\| \leq \delta$. An example of a uniformly continuous function is Res_S of Eq. (2), for any $S \in \mathbb{S}$. This can be proved by observing that the right-hand side of Eq. (2) can be rewritten in vector/matrix form as $\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|$ and then selecting, for any given $\varepsilon > 0$, δ to be $\varepsilon / \|\boldsymbol{A}\|$, where $\|\boldsymbol{A}\|$ denotes the matrix norm of $\boldsymbol{A}$.

An operator $\mathbf{O}$: $\Delta \to \Omega$, is *nonexpansive* if $\|\mathbf{O}\boldsymbol{x} - \mathbf{O}\boldsymbol{y}\| \leq \|\boldsymbol{x} - \boldsymbol{y}\|$, for all $\boldsymbol{x}, \boldsymbol{y} \in \Delta$. An example of a nonexpansive operator is the $\mathbf{R}_S$ of Eq. (3). The proof of this is also simple. It follows from discussions regarding similar claims in Ref. 27 that the $\mathbf{B}_{S_w} : \mathbb{R}^J \to \mathbb{R}^J$ of Eq. (4) is a nonexpansive operator, for $1 \leq w \leq W$, and that the operator $\mathbf{Q}$ of Eq. (5) is also nonexpansive. Obviously, a sequential application of nonexpansive operators results in a nonexpansive operator and thus $\mathbf{R}_S$ is nonexpansive.

Now we state an important new result that gives sufficient conditions for strong perturbation resilience: If $\mathbf{P}$ is an algorithm for a problem structure $\langle \mathbb{T}, \mathcal{P}r \rangle$ such that, for all $T \in \mathbb{T}$, $\mathbf{P}$ is boundedly convergent for T, $\mathcal{P}r_T : \Omega \to \mathbb{R}$ is uniformly continuous, and $\mathbf{P}_T : \Delta \to \Omega$ is nonexpansive, then $\mathbf{P}$ is strongly perturbation resilient. The importance of this result lies in the fact that the rather ordinary condition of uniform continuity for the proximity function and the reasonable conditions of bounded convergence and nonexpansiveness of the algorithmic operators guarantee that we end up with a strongly perturbation resilient algorithm. The proof of this new result involves some mathematical technicalities and is therefore presented in the Appendix as Theorem 1.

II.D. Optimization criterion and nonascending vector

Now suppose, as is indeed the case for the constrained optimization problems discussed in Sec. I, that in addition to a problem structure $\langle \mathbb{T}, \mathcal{P}r \rangle$ we are also provided with an optimization criterion, which is specified by a function $\phi : \Delta \to \mathbb{R}$, with the convention that a point in Δ for which the value of ϕ is smaller is considered *superior* (from the point of view of our application) to a point in Δ for which the value of ϕ is larger. In the tomography context, any of the functions of $\boldsymbol{x}$ that are listed as a "secondary optimization criterion" (an alternative name is a "regularizer") in Sec. 6.4 of Ref. 55 is an acceptable choice for the optimization criterion ϕ. These include weighted norms, the negative of Shannon's entropy and total variation. It is the last of these that we discuss in detail in the illustrative example below. The essential idea of the *superiorization methodology* presented in this paper is to make use of the perturbations of Eq. (6) to transform a strongly perturbation resilient algorithm that seeks a constraints-compatible solution into one whose outputs are equally good from the point of view of constraints-compatibility, but are superior according to the optimization criterion. We do this by producing from the algorithm another one, called its *superiorized* version, by making sure not only that the $\beta_k \boldsymbol{v}^k$ are bounded perturbations, but also that $\phi(\boldsymbol{x}^k + \beta_k \boldsymbol{v}^k) \leq \phi(\boldsymbol{x}^k)$, for all $k \geq 0$.

In order to ensure this we introduce a new concept (closely related to the concept of a "descent direction" that is widely used in optimization). Given a function $\phi : \Delta \to \mathbb{R}$ and a point $\boldsymbol{x} \in \Delta$, we say that a vector $\boldsymbol{d} \in \mathbb{R}^J$ is *nonascending* for ϕ at $\boldsymbol{x}$ if $\|\boldsymbol{d}\| \leq 1$ and

$$\begin{aligned} &\text{there is a } \delta > 0 \text{ such that for all } \lambda \in [0, \delta], \\ &\quad (\boldsymbol{x} + \lambda \boldsymbol{d}) \in \Delta \text{ and } \phi(\boldsymbol{x} + \lambda \boldsymbol{d}) \leq \phi(\boldsymbol{x}). \end{aligned} \tag{7}$$

Note that irrespective of the choices of ϕ and $\boldsymbol{x}$, there is always at least one nonascending vector $\boldsymbol{d}$ for ϕ at $\boldsymbol{x}$, namely, the zero-vector, all of whose components are zero. This is a useful fact for proving results concerning the guaranteed behavior of our proposed procedures. However, in order to steer our algorithms towards a point at which the value of ϕ is small, we need to find a $\boldsymbol{d}$ such that $\phi(\boldsymbol{x} + \lambda \boldsymbol{d}) < \phi(\boldsymbol{x})$ rather than just $\phi(\boldsymbol{x} + \lambda \boldsymbol{d}) \leq \phi(\boldsymbol{x})$ as in Eq. (7). In some earlier papers on superiorization[27–31] it was assumed that $\Delta = \mathbb{R}^J$ and that ϕ is a convex function. This implied that, for any point $\boldsymbol{x} \in \Delta$, ϕ had a subgradient $\boldsymbol{g} \in \mathbb{R}^J$ at the point $\boldsymbol{x}$. It was suggested that if there is such a $\boldsymbol{g}$ with a positive norm, then $\boldsymbol{d}$ should be chosen to be $-\boldsymbol{g}/\|\boldsymbol{g}\|$, otherwise $\boldsymbol{d}$ should be chosen to be the zero vector. However, there are approaches (not involving subgradients) to selecting an appropriate $\boldsymbol{d}$; an example can be found in Ref. 32 in which $\boldsymbol{d}$ is found without using subgradients for the case when ϕ is the ℓ_1-norm of the Haar transform.

The method we used for selecting a nonascending vector in the experiments reported in this paper is specified at the end of Subsection III.A.

II.E. Superiorized version of an algorithm

We now make precise the ingredients needed for transforming an algorithm into its superiorized version. Let Ω and Δ be the underlying sets for a problem structure $\langle \mathbb{T}, \mathcal{P}r \rangle$ ($\Omega \subseteq \Delta \subseteq \mathbb{R}^J$, as discussed at the beginning of Subsection II.B), **P** be an algorithm for $\langle \mathbb{T}, \mathcal{P}r \rangle$ and $\phi : \Delta \to \mathbb{R}$. The following description of the Superiorized Version of Algorithm **P** produces, for any problem $T \in \mathbb{T}$, a sequence $R_T = (\boldsymbol{x}^k)_{k=0}^{\infty}$ of points in Ω for which, for all $k \geq 0$, Eq. (6) is satisfied. We show this to be true, for any algorithm **P**, after the description of the Superiorized Version of Algorithm **P**. Furthermore, since the sequence R_T is steered by Superiorized Version of Algorithm **P** towards a reduced value of ϕ, there is an intuitive expectation that the output of the superiorized version is likely to be superior (from the point of view of the optimization criterion ϕ) to the output of the original unperturbed algorithm. This last statement is not precise and so it cannot be proved in a mathematical sense for an arbitrary algorithm **P**; however, that should not stop us from applying the easy procedure given below for automatically producing the superiorized version of **P** and experimentally checking whether it indeed provides us with outputs superior to those of the original algorithm. The well-demonstrated nature of heuristic optimization approaches is that they often work in practice even when their performance cannot be guaranteed to be optimal.[33–35]

Nevertheless, we can push our theory further than the hope expressed in the last paragraph, by considering superiorized versions of algorithms that satisfy some condition. In this paper, the condition that we discuss is strong perturbation resilience. We show below that if **P** is strongly perturbation resilient, then, for any problem $T \in \mathbb{T}$, a sequence R_T produced by its superiorized version has the following desirable property: For all $\varepsilon \in \mathbb{R}_+$, if $O(T, \varepsilon, ((\mathbf{P}_T)^k \boldsymbol{x})_{k=0}^{\infty})$ is defined for every $\boldsymbol{x} \in \Omega$, then $O(T, \varepsilon', R_T)$ is also defined for every $\varepsilon' > \varepsilon$; in other words, the Superiorized Version of Algorithm **P** provides an ε'-compatible output. As stated above, the advantage of the superiorized version is that its output is likely to be superior to the output of the original unperturbed algorithm. We point out that strong perturbation resilience is a sufficient, but not necessary, condition for guaranteeing such desirable behavior of the superiorized version, finding additional sufficient conditions and proving that algorithms that we wish to superiorize satisfy such conditions is part of our ongoing research.

The superiorized version assumes that we have available a summable sequence $(\gamma_\ell)_{\ell=0}^{\infty}$ of positive real numbers (for example, $\gamma_\ell = a^\ell$, where $0 < a < 1$) and it generates, simultaneously with the sequence $(\boldsymbol{x}^k)_{k=0}^{\infty}$, sequences $(\boldsymbol{v}^k)_{k=0}^{\infty}$, and $(\beta_k)_{k=0}^{\infty}$. The latter is generated as a subsequence of $(\gamma_\ell)_{\ell=0}^{\infty}$, resulting in a summable sequence $(\beta_k)_{k=0}^{\infty}$. The algorithm further depends on a specified initial point $\bar{\boldsymbol{x}} \in \Omega$ and on a positive integer N. It makes use of a logical variable called *loop*.

Superiorized Version of Algorithm P

(i) **set** $k = 0$
(ii) **set** $\boldsymbol{x}^k = \bar{\boldsymbol{x}}$
(iii) **set** $\ell = -1$
(iv) **repeat**
(v) **set** $n = 0$
(vi) **set** $\boldsymbol{x}^{k,n} = \boldsymbol{x}^k$
(vii) **while** $n < N$
(viii) **set** $\boldsymbol{v}^{k,n}$ to be a nonascending vector for ϕ at $\boldsymbol{x}^{k,n}$
(ix) **set** *loop* = *true*
(x) **while** *loop*
(xi) **set** $\ell = \ell + 1$
(xii) **set** $\beta_{k,n} = \gamma_\ell$
(xiii) **set** $\boldsymbol{z} = \boldsymbol{x}^{k,n} + \beta_{k,n} \boldsymbol{v}^{k,n}$
(xiv) **if** $\boldsymbol{z} \in \Delta$ **and** $\phi(\boldsymbol{z}) \leq \phi(\boldsymbol{x}^k)$, **then**
(xv) **set** $n = n + 1$
(xvi) **set** $\boldsymbol{x}^{k,n} = \boldsymbol{z}$
(xvii) **set** *loop* = *false*
(xviii) **set** $\boldsymbol{x}^{k+1} = \mathbf{P}_T \boldsymbol{x}^{k,N}$
(xix) **set** $k = k + 1$.

Next we analyze the behavior of the Superiorized Version of Algorithm **P**.

The iteration number k is set to 0 in (i) and $\boldsymbol{x}^k = \boldsymbol{x}^0$ is set to its initial value $\bar{\boldsymbol{x}}$ in (ii). The integer index ℓ for picking the next element from the sequence $(\gamma_\ell)_{\ell=0}^{\infty}$ is initialized to -1 by line (iii), it is repeatedly increased by line (xi). The lines (v)–(xix) that follow the **repeat** in (iv) perform a complete iterative step from $\boldsymbol{x}^k$ to $\boldsymbol{x}^{k+1}$, infinite repetitions of such steps provide the sequence $R_T = (\boldsymbol{x}^k)_{k=0}^{\infty}$. During one iterative step, there is one application of the operator $\mathbf{P}_T$, in line (xviii), but there are N steering steps aimed at reducing the value of ϕ; the latter are done by lines (v)–(xvii). These lines produce a sequence of points $\boldsymbol{x}^{k,n}$, where $0 \leq n \leq N$ with $\boldsymbol{x}^{k,0} = \boldsymbol{x}^k$, $\boldsymbol{x}^{k,n} \in \Delta$, and $\phi(\boldsymbol{x}^{k,n}) \leq \phi(\boldsymbol{x}^k)$.

We prove the truth of the last sentence by induction on the non-negative integers. For $n = 0$, we have by lines (v) and (vi) that $\boldsymbol{x}^{k,0} = \boldsymbol{x}^k$. But $\boldsymbol{x}^k \in \Omega$, since it is either $\bar{\boldsymbol{x}}$ that is assumed to be in Ω due to lines (i) and (ii) or it is in the range Ω of $\mathbf{P}_T$ due to lines (xviii) and (xix). Now we assume, for any $0 \leq n < N$, that $\boldsymbol{x}^{k,n} \in \Delta$ and $\phi(\boldsymbol{x}^{k,n}) \leq \phi(\boldsymbol{x}^k)$ and show that lines (viii)–(xvii) perform a computation that leads from $\boldsymbol{x}^{k,n}$ to an $\boldsymbol{x}^{k,n+1} \in \Delta$ that satisfies $\phi(\boldsymbol{x}^{k,n+1}) \leq \phi(\boldsymbol{x}^k)$. To see this, observe that line (viii) sets $\boldsymbol{v}^{k,n}$ to be a nonascending vector for ϕ at $\boldsymbol{x}^{k,n}$, which implies that Eq. (7) is satisfied with $\boldsymbol{x} = \boldsymbol{x}^{k,n}$ and $\boldsymbol{d} = \boldsymbol{v}^{k,n}$. Line (ix) sets *loop* to *true*, and it remains *true* while searching for the desired $\boldsymbol{x}^{k,n+1}$, by repeatedly executing the loop sequence that follows line (x). In this sequence, line (xi) increases ℓ by 1 and line (xii) sets $\beta_{k,n}$ to γ_ℓ. Thus for the vector $\boldsymbol{z}$ defined by line (xiii), $\boldsymbol{z} \in \Delta$ and $\phi(\boldsymbol{z}) \leq \phi(\boldsymbol{x}^{k,n})$, provided that $\beta_{k,n}$ is not greater than the δ in Eq. (7). Since $(\gamma_\ell)_{\ell=0}^{\infty}$ is a summable sequence of positive real numbers, there must be a positive integer L such that $\gamma_\ell \leq \delta$, for all $\ell \geq L$. This implies that if we applied lines (xi)–(xiii) often enough, we would reach a vector $\boldsymbol{z}$ that satisfies $\boldsymbol{z} \in \Delta$ and $\phi(\boldsymbol{z}) \leq \phi(\boldsymbol{x}^{k,n})$. If the condition in line (xiv) is not satisfied when the process gets to it, then lines

(xi)–(xiii) are again executed and eventually we get a vector $\boldsymbol{z}$ for which the condition in line (xiv) is satisfied due to the induction hypothesis that $\phi(\boldsymbol{x}^{k,n}) \leq \phi(\boldsymbol{x}^k)$. By lines (xv) and (xvi) we see that at that time $\boldsymbol{x}^{k,n+1}$ is set to $\boldsymbol{z}$ and so we obtain that $\boldsymbol{x}^{k,n+1} \in \Delta$ and $\phi(\boldsymbol{x}^{k,n+1}) \leq \phi(\boldsymbol{x}^k)$, as desired. Line (xvii) sets *loop* to *false* and so control is returned to line (vii). When this happens for the Nth time, it will be the case that $n = N$ and, therefore, line (xviii) is used to produce $\boldsymbol{x}^{k+1} \in \Omega$ and the increasing of k by line (xix) allows us then to move on to the next iterative step. Infinite repetition of such steps produces the sequence $R_T = (\boldsymbol{x}^k)_{k=0}^{\infty}$ of points in Ω.

We now show that if $O(T, \varepsilon, ((\mathbf{P}_T)^k \boldsymbol{x})_{k=0}^{\infty})$ is defined for every $\boldsymbol{x} \in \Omega$, then, for any $\varepsilon' > \varepsilon$, the Superiorized Version of Algorithm $\mathbf{P}$ produces an ε'-compatible output. Since $\mathbf{P}$ is assumed to be strongly perturbation resilient, this desired result follows if we can show that there exists a summable sequence $(\beta_k)_{k=0}^{\infty}$ of non-negative real numbers and a bounded sequence $(\boldsymbol{v}^k)_{k=0}^{\infty}$ of vectors in $\mathbb{R}^J$ such that Eq. (6) is satisfied for all $k \geq 0$. In view of line (xviii), this is achieved if we can define the β_k and the $\boldsymbol{v}^k$ so that $\boldsymbol{x}^{k,N} = \boldsymbol{x}^k + \beta_k \boldsymbol{v}^k$. This is done by setting

$$\beta_k = \max\{\beta_{k,n} \,|\, 0 \leq n < N\}, \tag{8}$$

$$\boldsymbol{v}^k = \sum_{n=0}^{N-1} \frac{\beta_{k,n}}{\beta_k} \boldsymbol{v}^{k,n}. \tag{9}$$

That these assignments result in $\boldsymbol{x}^{k,N} = \boldsymbol{x}^k + \beta_k \boldsymbol{v}^k$ follows from lines (v)–(xvii). From line (xii) follows that $(\beta_k)_{k=0}^{\infty}$ is a subsequence of $(\gamma_\ell)_{\ell=0}^{\infty}$ and, hence, it is a summable sequence of non-negative real numbers. Since each $\|\boldsymbol{v}^{k,n}\| \leq 1$ by the definition of a nonascending vector, it follows from Eqs. (8) and (9) that $\|\boldsymbol{v}^k\| \leq N$ and so $(\boldsymbol{v}^k)_{k=0}^{\infty}$ is bounded. Part of the condition expressed in Eq. (6) is that, for all $k \geq 0$, $\boldsymbol{x}^k + \beta_k \boldsymbol{v}^k \in \Delta$. This follows from the fact that $\boldsymbol{x}^{k,N} = \boldsymbol{x}^k + \beta_k \boldsymbol{v}^k$ is assigned its value by line (xvi), but only if the condition expressed in line (xiv) is satisfied.

In conclusion, we have shown that the superiorized version of a strongly perturbation resilient algorithm produces outputs that are essentially as constraints-compatible as those produced by the original version of the algorithm. However, due to the repeated steering of the process by lines (vii)–(xvii) towards reducing the value of the optimization criterion ϕ, we can expect that the output of the superiorized version will be superior (from the point of view of ϕ) to the output of the original algorithm.

II.F. Information on performance comparison with MAP methods

Using our notation, the constrained minimization formulation that we are considering is as follows: Given an $\varepsilon \in \mathbb{R}_+$,

$$\text{minimize } \phi(\boldsymbol{x}), \text{ subject to } \mathcal{P}r_T(\boldsymbol{x}) \leq \varepsilon. \tag{10}$$

The aim of superiorization is not identical with the aim of constrained minimization in Eq. (10). One difference is that ε is not "given" in the superiorization context. The superiorization of an algorithm produces a sequence and, for any ε, the associated output of the algorithm is considered to be the first $\boldsymbol{x}$ in the sequence for which $\mathcal{P}r_T(\boldsymbol{x}) \leq \varepsilon$. The other difference is that we do not claim that this output is a minimizer of ϕ among all points that satisfy the constraint, but hope only that it is usually an $\boldsymbol{x}$ for which $\phi(\boldsymbol{x})$ is at the small end of its range of values over the set of constraint-satisfying points. This latter difference is generally shared by comparisons of a heuristic approach with an exact approach to solving a constrained minimization problem.

The MAP (or regularized) formulation of a physical problem that leads to the constrained minimization problem (10) is the unconstrained minimization problem of the form: Given a $\beta \in \mathbb{R}_+$,

$$\text{minimize } [\phi(\boldsymbol{x}) + \beta \mathcal{P}r_T(\boldsymbol{x})]. \tag{11}$$

Formulations of both kinds [i.e., the ones of Eqs. (10) and (11)] are widely used for solving medical physics problems and the question "Which of these two formulations leads to faster or better solutions of the underlying physical problem?" is open. Examples of both formulations with various choices for $\mathcal{P}r_T$ and ϕ are listed in the beginning parts of the paper of Goldstein and Osher.[47]

We now return to the question raised near the end of Sec. I: Will superiorization produce superior results to those produced by contemporary MAP methods or is it faster than the better of such methods? As yet, there is very little information available regarding this general question; in fact, we are aware of only one published study.[45] That study compared a superiorization algorithm with the algorithm of Goldstein and Osher that they refer to as TwIST (Ref. 46) with split Bregman[47] as the substep, which is indeed a contemporary method that uses the MAP formulation. (For example, see the discussion of the split Bregman method in Ref. 56.) The problem S to which the two algorithms were applied was one from the tomographic problem set $\mathbb{S}$ defined in Eq. (1). Res_S as defined in Eq. (2) was used as the proximity function and total variation, TV as defined below in Eq. (12), was the choice for ϕ. It is reported in Ref. 45 that for the outputs of the two algorithms that were being compared, the values of Res_S and TV were very similar, but the superiorization algorithm produced its output four times faster than the MAP method.

III. AN ILLUSTRATIVE EXAMPLE

III.A. Application to tomography

We use *tomography* to refer to the process of reconstructing a function over a Euclidean space from estimated values of its integrals along lines (that are usually, but not necessarily, straight). The particular reconstruction processes to which our discussion applies are the *series expansion methods*, see Sec. 6.3 of Ref. 55, in which it is assumed that the function to be reconstructed can be approximated by a linear combination of a finite number (say J) of basis functions and the

reconstruction task becomes one of estimating the coefficients of the basis functions in the expansion. Sometimes, prior knowledge about the nature of the function to be reconstructed allows us to confine the sought-after vector $\boldsymbol{x}$ of coefficients to a subset Ω of $\mathbb{R}^J$ (such as the non-negative orthant $\mathbb{R}^J_+$). We use i to index the lines along which we integrate, $\boldsymbol{a}^i \in \mathbb{R}^J$ to denote the vector whose jth component is the integral of the jth basis function along the ith line, and b_i to denote the measured integral of the function to be reconstructed along the ith line. Under these circumstances the constraints come from the desire that, for each of the lines, $\langle \boldsymbol{a}^i, \boldsymbol{x} \rangle$ should be close (in some sense) to b_i.

To make this concrete, consider Eq. (1). Such a description of the constraints arises in tomography by grouping the lines of integration into W blocks, with ℓ_w lines in the wth block. Such groupings often (but not always) are done according to some geometrical condition on the lines (for example, in case of straight lines, we may decide that all the lines that are parallel to each other form one block). In this framework, the proximity function *Res* defined by Eq. (2) provides a reasonable measure of the incompatibility of a vector $\boldsymbol{x}$ with the constraints. The algorithm **R** described by Eqs. (3)–(5) is applicable to this concrete formulation.

There are many optimization criteria that have been used in tomography, see Sec. 6.4 of Ref. 55, here we discuss the one called TV, whose use has been popular in medical physics recently, see as examples Refs. 20, 22, 23, and 41–44. The definition of TV that we use here requires a certain way of selecting the basis functions. It is assumed that the function to be reconstructed is defined in the plane $\mathbb{R}^2$ and is zero-valued outside a square-shaped region in the plane. This region is subdivided into J smaller equal-sized squares (*pixels*) and the J basis functions are defined by having value one in exactly one pixel and value zero everywhere else. We index the pixels by j and we let C denote the set of all indices of pixels that are not in the rightmost column or the bottom row of the pixel array. For any pixel with index j in C, let $r(j)$ and $b(j)$ be the index of the pixel to its right and below it, respectively. We define $TV : \mathbb{R}^J \to \mathbb{R}$ by

$$TV(\boldsymbol{x}) = \sum_{j \in C} \sqrt{(x_j - x_{r(j)})^2 + (x_j - x_{b(j)})^2}. \tag{12}$$

The method we adopted to generate a nonascending vector for the TV function at an $\boldsymbol{x} \in \mathbb{R}^J$ is based on Theorem 2 of the Appendix. It is applicable since $TV : \mathbb{R}^J \to \mathbb{R}$ is a convex function; see, for example, the end of the Proof of Proposition 1 of Ref. 41. Now consider an integer j' such that $1 \le j' \le J$. Looking at the sum in Eq. (12), we see that $x_{j'}$ appears in at most three terms, in which j' must be either j, or $r(j)$, or $b(j)$ for some $j \in C$. By taking the formal partial derivatives of these three terms, we see that $\frac{\partial TV}{\partial x_{j'}}(\boldsymbol{x})$ is well defined if the denominator in the formal derivative of each of the three terms is not zero for $\boldsymbol{x}$. In view of this, we define the $\boldsymbol{g}$ in Theorem 2 as follows. If the denominator in any of the three formal partial derivatives with respect to $x_{j'}$ has an absolute value less than a very small positive number (we used 10^{-20}), then we set $g_{j'}$ to zero, otherwise we set it to $\frac{\partial TV}{\partial x_{j'}}(\boldsymbol{x})$. Clearly, the resulting $\boldsymbol{g} \in \mathbb{R}^J$ satisfies the condition in Theorem 2 and hence provides a $\boldsymbol{d}$ that is a nonascending vector for TV at $\boldsymbol{x}$.

Previously reported reconstructions using TV-superiorization selected the $\boldsymbol{d}$ using subgradients as discussed in the paragraph following Eq. (7); such a $\boldsymbol{d}$ is not guaranteed to be a nonascending vector for the TV function. What we are proposing here is not only mathematically rigorous (in the sense that it is guaranteed to produce a nonascending vector for the TV function), but it can also lead to a better reconstructions, as illustrated in Subsection III.D.

III.B. The data generation for the experiments

The datasets used in the experiments reported in this paper were generated in such a way that they share the noise-characteristics of CT scanners when used for scanning the human head and brain; as discussed, for example, in Chap. 5 of Ref. 55. They were generated using the software SNARK09.[57]

The head phantom that was used for data generation is based on an actual cross section of the human head. It is described as a collection of geometrical objects (such as ellipses, triangles, and segments of circles) whose combination accurately resembles the anatomical features of the actual head cross section. In addition, the basic phantom contains a large tumor. The actual phantom used was obtained by a random variation of the basic phantom, by incorporating into it local inhomogeneities and small low-contrast tumors at random locations. This phantom is represented by the image in Fig. 1(a). That image comprises 485×485 pixels each of size 0.376 mm by 0.376 mm. The values assigned to the pixels are obtained by an 11×11 subsampling of the pixels and averaging the values assigned to the subsamples by the geometrical objects that are used to describe the anatomical features and the tumors. Those values are approximate linear attenuation coefficients per cm at 60 keV (0.416 for bone, 0.210 for brain, 0.207 for cerebrospinal fluid). The contrast of the small tumors with their background is 0.003 cm^{-1}. In order to clearly see the low-contrast details in the interior of the skull, we use zero (black) to represent the value 0.204 (or anything less) and 255 (white) to represent 0.21675 or anything more). The same is true for all the images in the rest of this paper.

For the selected head phantom we generated *parallel projection data*, in which one *view* comprises estimates of integrals through the phantom for a set of 693 equally spaced parallel lines with a spacing of 0.0376 cm between them. (We chose to simulate parallel rather than divergent projection data, since the reconstruction by the method of Ref. 42 with which we wish to compare the superiorization approach was performed for us by the authors of Ref. 42 on parallel data. Even though contemporary CT scanners use divergent projection data, results obtained by the use of parallel projection data are relevant to them, since it is known that the quality of reconstructions from these two modes of data collection are very similar as long as the data generations use similar frequencies of sampling of lines and similar noise characteristics

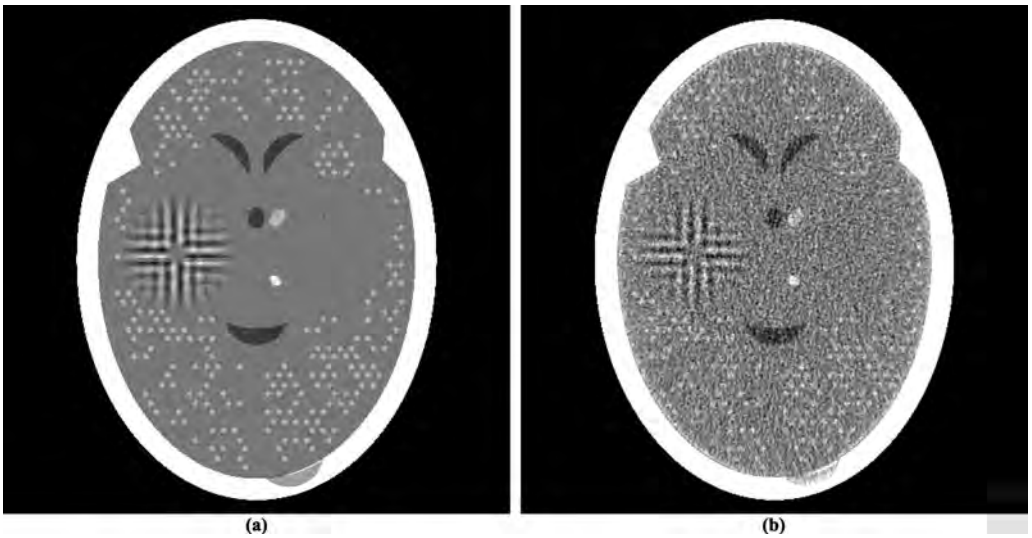


FIG. 1. (a) A head phantom. (b) Reconstruction of the head phantom from realistically simulated projection data for 360 views using ART with blob basis functions.

in the estimated integrals for those lines; see, for example, the reconstructions from divergent and parallel projection data in Fig. 5.15 of Ref. 55.) In calculating these estimates. we take into consideration the effects of photon statistics, detector width, and scatter. Details of how we do this exactly can be found in Secs. 5.5 and 5.9 of Ref. 55. Briefly, quantum noise is calculated based on the assumption that approximately 2 000 000 photons enter the head along each ray, detector width is simulated by using 11 subrays along each of which the attenuation is calculated independently and then combined at the detector, and 5% of the photons get counted not by the detector for the ray in question but detectors for the neighboring rays. For the experiments in this paper, we did not simulate the polyenergetic nature of the x-ray source. To indicate what can be achieved in clinical CT, we show in Fig. 1(b) a reconstruction that was made from data comprising of 360 such views with the reconstruction algorithm known as ART with blob basis functions; see Chap. 11 of Ref. 55.

III.C. Superiorization reconstruction from a few views

The main reason in the literature for advocating the use of TV as the optimization criterion is that by doing so one can achieve efficacious reconstructions even from sparsely sampled data. In our own work[31] with realistically simulated CT data, we found that this is not always the case and this will be demonstrated again by the experiments reported in the current paper.

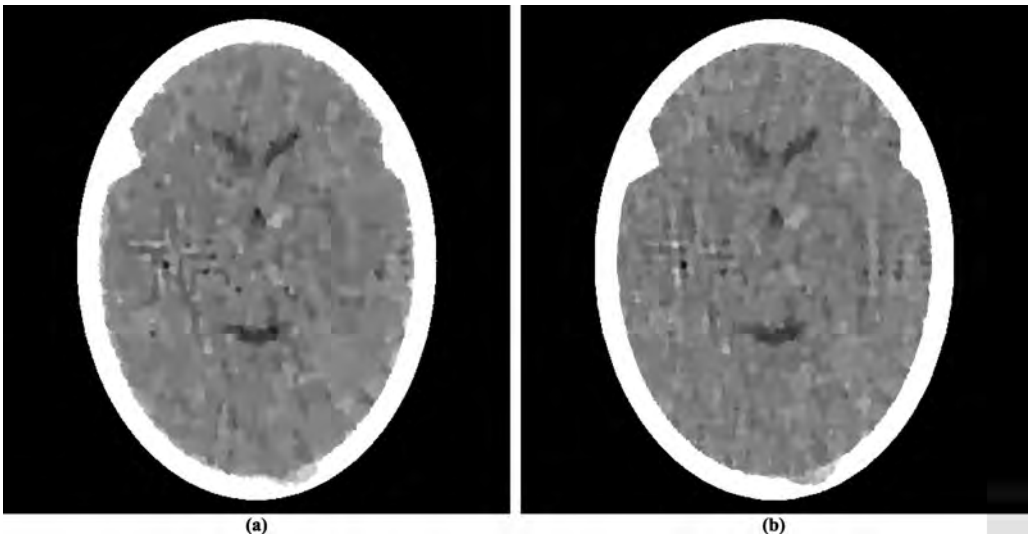


FIG. 2. Reconstructions using TV as the optimization criterion from realistically simulated projection data for 60 views using (a) ASD-POCS and (b) superiorization. As compared to Fig. 1(b), these reconstructions fail in two ways: they do not show some of the fine details in the phantom and they present some artifactual variations. The former of these is a consequence of reconstructing from a much smaller dataset than used for Fig. 1(b). The latter is due to using a very narrow window (13.5 HU) in these displays. Were we to use a wider display window (e.g., from –429 HU to 429 HU) for the reconstructions in this figure and in Fig. 1(b), the visual appearance of the resulting images would be nearly indistinguishable.

There have appeared in the literature some approaches to TV minimization that seem to indicate a more efficacious performance for CT than the one reported in Ref. 31. One of these is the adaptive steepest descent projections onto convex sets (ASD-POCS) algorithm, which is described in detail in the much-cited paper of Sidky and Pan[42] and whose use has been since reported in a number of subsequent publications, for example, in Refs. 23 and 43. We note that ASD-POCS was designed with the aim of producing an exact minimization algorithm, in contrast to our heuristic superiorization approach. Translating Eqs. (6)–(8) of Ref. 42 into our terminology, the aim of ASD-POCS is the following: Given an $\varepsilon \in \mathbb{R}_+$, find an ε-compatible $\boldsymbol{x} \in \Omega = \mathbb{R}_+^J$ for which $TV(\boldsymbol{x})$ is minimal. [Note that this aim is a special case of the constrained optimization formulation presented in Eq. (10).] In order to test ASD-POCS, we generated realistic projection data as described in Subsection III.B but for only 60 views at 3° increments with the spacing between the lines for which integrals are estimated set at 0.752 mm. Thus the number of rays (and hence the number photons put into the head) in this dataset is a 12th of what it is in the dataset used to produce the reconstruction in Fig. 1(b). A reconstruction from these data was produced for us using ASD-POCS by the authors of Ref. 42 (this ensured that it does not suffer due to our misinterpretation of the algorithm or from our inappropriate choices of the free parameters), it is shown in Fig. 2(a).

Since the image quality of Fig. 2(a) is not anywhere near to that of Fig. 1(b), we present here a brief discussion as to why we are showing such images. Many publications in the recent medical imaging literature have claimed that medically efficacious reconstructions can be obtained by the use of TV-minimization from data as sparse as what was used to produce Fig. 2(a). (In fact, ASD-POCS was motivated and used with such an aim in mind.[23,42,43]) Such publications usually show reconstructions from sparse data as evidence for the validity of their claims. They can do this because in their presented illustrations the features that are observable in the reconstructions are usually much larger and/or of much higher contrast against their backgrounds than the small "tumors" in Fig. 1(a), which are perfectly visible in the reconstruction in Fig. 1(b), but are not detectable in the reconstruction from sparse data in Fig. 2(a). The reason why that reconstruction appears to be unacceptably bad is that the display window (from 0.204 cm^{-1} linear attenuation coefficient to 0.21675 cm^{-1} linear attenuation coefficient) is very narrow; it was selected to enhance the visibility of the small low-contrast tumors. The width of this window corresponds to about 13.5 Hounsfield units (HU). As compared to this, in their evaluation of sparse-view reconstruction from flat-panel-detector cone-beam CT, Bian *et al.*[43] use what they call a "soft-tissue grayscale window" (also a "narrow window") from –429 HU to 429 HU to display head phantom reconstructions. Using such a window for our reconstructions shown Figs. 2(a) and 1(b) would result in images that are nearly indistinguishable from each other. Thus reporting the images using such a display window is consistent with the claim that a *TV*-minimizing reconstruction from a few views is similar in quality to a more traditional reconstruction from many views. However, our much narrower display window reveals that this is not really so. We therefore continue using our much narrower window in what follows, since it clearly reveals the nature of the reconstructions being compared, warts and all.

While this ASD-POCS reconstruction is not as good as it should be for diagnostic CT of the brain (due to the sparsity of the data), it is visually better than the reconstruction using superiorization from similar data as reported in Ref. 31. We discuss the reasons for this in Subsection III.D. Here, we concentrate on examining whether one can achieve a reconstruction using superiorization that is as good as that produced by ASD-POCS from the same data.

For this we first need to examine the numerical properties of the ASD-POCS reconstruction. This reconstruction uses 485×485 pixels each of size 0.376 mm by 0.376 mm. This implies that $J = 235{,}225$ and it also determines the components of the vectors $\boldsymbol{a}^i \in \mathbb{R}^J$ in the precise specification of the problem S. The Res_S, as defined by Eq. (2), of the ASD-POCS reconstruction is 0.33 and the TV, as defined by Eq. (12), is 835.

We applied to the same problem S a superiorized version of the algorithm **R** defined by Eq. (3). To complete the specification of **R**, we point out that for the ordering of views we chose the "efficient" one that was introduced in Ref. 58 and is also discussed on p. 209 of Ref. 55. The choices we made for the superiorization are the following: $\gamma_\ell = 0.99995^\ell$, $\overline{\boldsymbol{x}}$ is the zero vector, and $N = 20$. The nonascending vector was computed by the method described in the paragraph below [Eq. (12)]. Denoting by R_S the infinite sequence of points in Ω that is produced by the superiorized version of the algorithm **R** when applied to the problem S, we chose as our reconstruction $\boldsymbol{x}^* = O(S, 0.33, R_S)$. For such a reconstruction we have, by the definition of O, that $Res_S(\boldsymbol{x}^*) \leq 0.33$; in other words, the output of the superiorization algorithm is at least as constraints-compatible with S as the output of ASD-POCS. From the point of view of TV-minimization, our $\boldsymbol{x}^*$ is slightly better: $TV(\boldsymbol{x}^*) = 826$.

The superiorization reconstruction is displayed in Fig. 2(b). Visually, it is similar to the reconstruction produced by ASD-POCS. From the optimization point of view it achieves the desired aim better than ASD-POCS does, since it results in smaller values for both Res_S and for TV, even though only slightly.

That the two reconstructions in Fig. 2 are very similar is not surprising because a comparison of the pseudocodes reveals that the ASD-POCS algorithm in Ref. 42 is essentially a special case of the Superiorized Version of Algorithm **P**, even though it has been derived from rather different principles. To obtain the ASD-POCS algorithm from our methodology described here, we would have to choose ART (see Chap. 11 of Ref. 55) as the algorithm that we are superiorizing. Such a superiorization of ART was reported in the earliest paper on superiorization.[27] For the illustration in our current paper, we decided to superiorize the block-iterative algorithm **R** defined by Eq. (3). This illustrates the generality of the superiorization approach: it is applicable not only to a large class of constrained optimization problems, but also enables the use of any of a large class of iterative algorithms designed to

produce a constraints-compatible solutions. A recent publication aimed at producing an exact TV-minimizing algorithm based on the block-iterative approach is Ref. 44.

III.D. Effects of variations in the reconstruction approach

The reconstruction in Fig. 2(a) produced by ASD-POCS definitely "looks better" than a reconstruction in Ref. 31, which was obtained using superiorization from similar data. Since, as discussed in the last paragraph of Subsection III.C, the ASD-POCS algorithm in Ref. 42 can be obtained as a special case of superiorization, it must be that some of the choices made in the details of the implementations are responsible for the visual differences. An analysis of the implementational details adopted by the two approaches revealed several differences. After removing these differences, the superiorization approach produced the image in Fig. 2(b), which is very similar to the reconstruction produced by ASD-POCS. We now list the implementational choices that were made for superiorization to make its performance match that of the reported implementation of ASD-POCS.

One implementational difference is in the stopping-rule of the iterative algorithm; that is, the choice of ε in determining the output $O(S, \varepsilon, R_S)$. Since the data are noisy, the phantom itself does not match the data exactly. In previously reported implementations of superiorization it was assumed that the iterative process should terminate when an image is obtained that is approximately as constraints-compatible as the phantom; in the case of the phantom and the projections data on which we report here the value of Res_S for the phantom is approximately 0.91, which is larger than its value (0.33) for the reconstruction produced by ASD-POCS. The output $O(S, 0.91, R_S)$ is shown in Fig. 3(a). This is a wonderfully smooth reconstruction, its TV value is only 771. However, this smoothness comes at a price: we lose not only the ability to detect the large tumor, but we cannot even see anatomic features (such as the ventricular cavities) inside the brain. So it appears that, in order to see medically relevant features in the brain, *overfitting* (in the sense of producing a reconstruction from noisy data that is more constraints-compatible than the phantom) is desirable.

In the implementations that produced previously reported reconstructions by superiorization, the number N in the Superiorized Version of Algorithm **P** was always chosen to be 1. It is possible that this is the wrong choice, making only this change to what lead to the reconstruction in Fig. 2(b) results in the reconstruction shown in Fig. 3(b). That image appears similar to the image in Fig. 2(b), but it has a higher TV value, namely, 832, which is still very slightly lower than that of the ASD-POCS reconstruction. The choice $N = 20$ was based on the desire to maintain consistency with what has been practiced using ASD-POCS, see p. 4790 of Ref. 42. It appears that in the context of our paper the additional computing cost due to choosing N to be 20 rather than 1 is not really justified. (We note that if $\boldsymbol{d}$ is selected using subgradients as discussed in the paragraph following Eq. (7) and thus $\boldsymbol{d}$ is not guaranteed to be a nonascending vector for the TV function, then the choice of 20 rather than 1 for N results in a considerable improvement. However, an even greater improvement is achieved even with $N = 1$ by selecting $\boldsymbol{d}$ as recommended in this paper.)

Another important difference between the ASD-POCS implementation and the previous implementations of the superiorization approach is the size of the pixels in the reconstructions. For the ASD-POCS reconstruction this was selected to be 0.376 mm by 0.376 mm. In previously reported reconstructions by superiorization it was assumed that the edge of a pixel should be the same as the distance between the parallel lines along which the data are collected; that is, 0.752 mm for our problem S. This assumption proved to be false. TV-minimization takes care of undesirable artifacts that may otherwise arise due to the smaller pixels and this leads to a visual improvement. A superiorizing reconstruction with the larger pixels, using $\varepsilon = 0.33$ and $N = 20$, is shown in Fig. 3(c). (We note that the use of smaller pixels during iterative x-ray CT reconstructions was also suggested in Ref. 59. However, that approach is quite different from what is presented here: its final result uses larger pixels whose values are obtained by averaging assemblies of values provided by the iterative process to the smaller pixels. There is no such downsampling in our approach, our final result is presented using the smaller pixels. Its smoothness is due to reduction of TV by the superiorization approach rather than to averaging pixel values in a denser digitization.)

Combining the use of the larger pixels with $\varepsilon = 0.91$ and $N = 1$ results in the reconstruction shown in Fig. 3(d). This reconstruction, for which the superiorization options were selected according to what was done in Ref. 31, is visually inferior to those shown in our Fig. 2. The reconstructions displayed in Fig. 3 also illustrate another important point, namely, that even though the mathematical results discussed in this paper are valid for a large range of choices of the parameters in the superiorization algorithms, for medical efficacy of the reconstructions attention has to be paid to these choices since they can have a drastic effect on the quality of the reconstruction.

It has been mentioned in Subsection II.B that except for the presence of **Q** in Eq. (3), which enforces non-negativity of the components, **R** is identical to the algorithm used and illustrated in Ref. 31. It is known that CT reconstruction of the brain from many views does not suffer from ignoring the fact that the components of the $\boldsymbol{x}$, which represent linear attenuation coefficients, should be non-negative; as is illustrated in Fig. 1(b). This remains so when reconstructing from a few views using the method and data that we have been discussing: if we do everything in exactly the same way as was done to obtain the reconstruction with TV value 826 that is shown in our Fig. 2(b) but remove **Q** from Eq. (3), then we obtain a reconstruction in Fig. 4(a) whose TV value is 829.

Another variation that deserves discussion, because it has been suggested in the literature,[22] is one that does not come about by making choices for the general approach of the Superiorized Version of Algorithm **P** but rather by changing the nature of the approach. The variation in question is not applicable in general, but can be applied to the special case when the algorithm to be superiorized is the **R** defined by Eq. (3). It

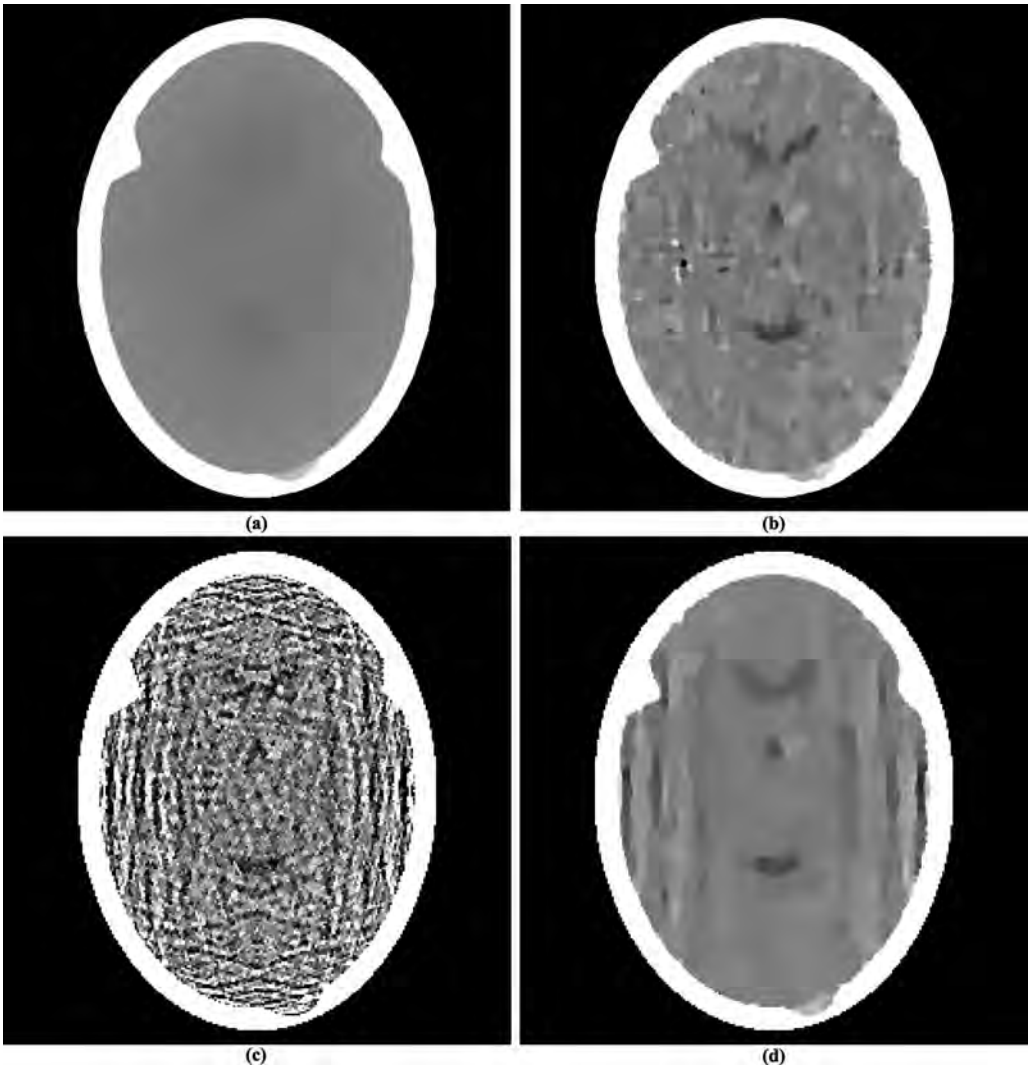


FIG. 3. Reconstructions produced by varying some of the parameters in the algorithm that produced Fig. 2(b). (a) Changing the termination criterion form $\varepsilon = 0.33$ to $\varepsilon = 0.91$. (b) Changing the value of N from 20 to 1. (c) Reconstructing with pixel size 0.752 mm by 0.752 mm instead of 0.376 mm by 0.376 mm. (d) Reconstructing with all the three changes of (a)–(c).

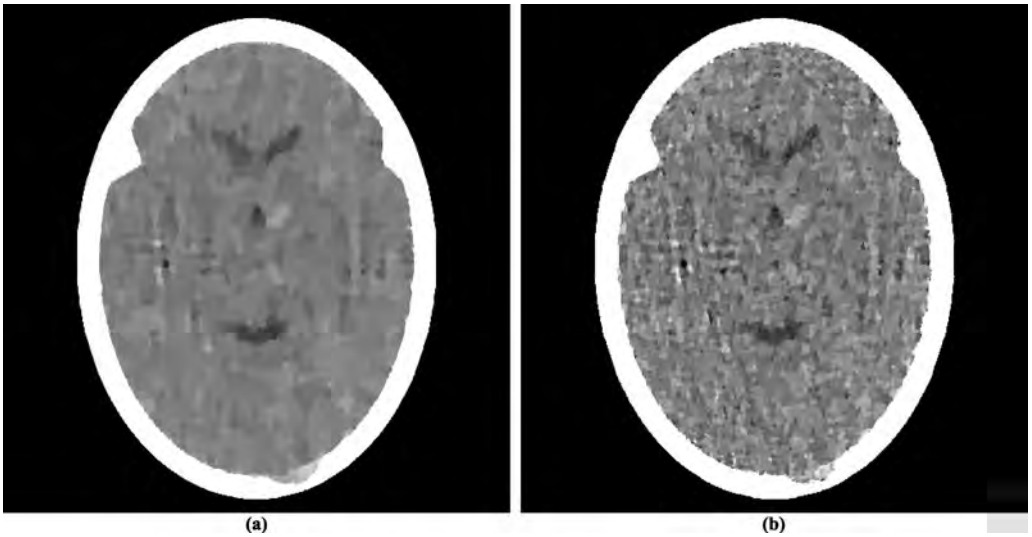


FIG. 4. Reconstructions by variations that do not fit into the framework within which the previously shown reconstructions were produced. (a) Not using non-negativity in the algorithm. (b) Interleaving perturbations with blocks.

was suggested as an improvement to the approach presented above with the choice $N = 1$. The idea was based on recognizing the block-iterative nature of the algorithmic operator $\mathbf{R}_S$ in Eq. (3) and intermingling the perturbation steps of lines (vii)–(xvii) of the Superiorized Version of Algorithm **R** with the projection steps $\mathbf{B}_{S_1}, \ldots, \mathbf{B}_{S_W}$ of Eq. (3). It was reported in Ref. 22 that doing this is advantageous to using the Superiorized Version of Algorithm **R**. However, when we applied the variation of the Superiorized Version of Algorithm **R** that is proposed in Ref. 22 to the problem S that we have been using in this section, we ended up with the reconstruction in Fig. 4(b) whose TV value is 920. This is not as good as what was obtained using the version of the algorithm that produced the reconstruction in Fig. 2(b). We conclude that the variation suggested by Ref. 22, which does not fit into the theory of our paper, does not have an advantage over what we are proposing here, at least for the problem S that we have been discussing in this section. We conjecture that the improvement reported in Ref. 22 is due to selecting $\boldsymbol{d}$ using subgradients as discussed in the paragraph following Eq. (7) and, as discussed earlier, such an improvement is not obtained if $\boldsymbol{d}$ is selected by the more appropriate method recommended in this paper.

IV. DISCUSSION AND CONCLUSIONS

Constrained optimization is an often-used tool in medical physics. The methodology of superiorization is a heuristic (as opposed to exact) approach to constrained optimization.

Although the idea of superiorization was introduced in 2007 and its practical use has been demonstrated in several publications since, this paper is the first to provide a solid mathematical foundation to superiorization as applied to the noisy problems of the real world. These foundations include a precise definition of constraints-compatibility, the concept of a strongly perturbation resilient algorithm, simple conditions that ensure that an algorithm is strongly perturbation resilient, the superiorized version of an algorithm and the showing that the superiorized version of a strongly perturbation resilient algorithm produces outputs that are essentially as constraints-compatible as those produced by the original version but are likely to have a smaller value of the chosen optimization criterion.

The approach is very general. For any iterative algorithm **P** and for any optimization criterion ϕ for which we know how to produce nonascending vectors, the pseudocode given in Subsection II.E automatically provides the version of **P** that is superiorized for ϕ.

We demonstrated superiorization for tomography when total variation is used as the optimization criterion. In particular, we illustrated on a particular tomography problem that, in spite of its generality, superiorization produced a reconstruction that is as good as (from the points of view of constraints-compatibility and TV-minimization) what was obtained by the ASD-POCS algorithm that was specially designed for TV-minimization in tomography.

ACKNOWLEDGMENTS

The detailed and penetrating comments of three reviewers and the editors helped us to improve this paper in a significant way. The authors thank Professor Xiaochuan Pan and his co-workers from the University of Chicago for providing them with the reconstruction from their data using their implementation of their ASD-POCS algorithm. Our work is supported by the National Science Foundation Award No. DMS-1114901, the United States-Israel Binational Science Foundation (BSF) Grant No. 200912, the U.S. Department of Defense Prostate Cancer Research Program Award No. W81XWH-12-1-0122, and the U.S. Department of Army Award No. W81XWH-10-1-0170.

APPENDIX: MATHEMATICAL PROOFS

1. Conditions for strong perturbation resilience

Theorem 1. Let **P** be an algorithm for a problem structure $\langle \mathbb{T}, \mathcal{P}r \rangle$ such that, for all $T \in \mathbb{T}$, **P** is boundedly convergent for T, $\mathcal{P}r_T : \Omega \to \mathbb{R}$ is uniformly continuous, and $\mathbf{P}_T : \Delta \to \Omega$ is nonexpansive. Then **P** is strongly perturbation resilient.

Proof. We first show that there exists an $\varepsilon \in \mathbb{R}_+$ such that $O(T, \varepsilon, ((\mathbf{P}_T)^k \boldsymbol{x})_{k=0}^{\infty})$ is defined for every $\boldsymbol{x} \in \Omega$. Under the assumptions of the theorem, let $\gamma \in \mathbb{R}_+$ be such that $\mathcal{P}r_T(\boldsymbol{y}(\boldsymbol{x})) \leq \gamma$, for every $\boldsymbol{x} \in \Omega$. We prove that $O(T, 2\gamma, ((\mathbf{P}_T)^k \boldsymbol{x})_{k=0}^{\infty})$ is defined for every $\boldsymbol{x} \in \Omega$ as follows. Select a particular $\boldsymbol{x} \in \Omega$. By uniform continuity of $\mathcal{P}r_T$, there exists a $\delta > 0$, such that $|\mathcal{P}r_T(\boldsymbol{z}) - \mathcal{P}r_T(\boldsymbol{y}(\boldsymbol{x}))| \leq \gamma$, for any $\boldsymbol{z} \in \Omega$ for which $\|\boldsymbol{z} - \boldsymbol{y}(\boldsymbol{x})\| \leq \delta$. Since **P** is convergent for T, there exists a non-negative integer K, such that $\|(\mathbf{P}_T)^K \boldsymbol{x} - \boldsymbol{y}(\boldsymbol{x})\| \leq \delta$. It follows that

$$|\mathcal{P}r_T((\mathbf{P}_T)^K \boldsymbol{x})| \leq |\mathcal{P}r_T((\mathbf{P}_T)^K \boldsymbol{x}) - \mathcal{P}r_T(\boldsymbol{y}(\boldsymbol{x}))| + |\mathcal{P}r_T(\boldsymbol{y}(\boldsymbol{x}))| \leq 2\gamma. \tag{A1}$$

Now let $T \in \mathbb{T}$ and $\varepsilon \in \mathbb{R}_+$ be such that $O(T, \varepsilon, ((\mathbf{P}_T)^k \boldsymbol{x})_{k=0}^{\infty})$ is defined for every $\boldsymbol{x} \in \Omega$. To prove the theorem, we need to show that $O(T, \varepsilon', R)$ is defined for every $\varepsilon' > \varepsilon$ and for every sequence $R = (\boldsymbol{x}^k)_{k=0}^{\infty}$ of points in Ω for which, for all $k \geq 0$, Eq. (6) is satisfied for bounded perturbations $\beta_k \boldsymbol{v}^k$. Let ε' and R satisfy the conditions of the previous sentence.

For $k \geq 0$, we have, due to the nonexpansiveness of $\mathbf{P}_T$, that

$$\|\boldsymbol{x}^{k+1} - \mathbf{P}_T \boldsymbol{x}^k\| = \|\mathbf{P}_T(\boldsymbol{x}^k + \beta_k \boldsymbol{v}^k) - \mathbf{P}_T \boldsymbol{x}^k\| \leq \|\beta_k \boldsymbol{v}^k\|. \tag{A2}$$

Denote $\|\beta_k \boldsymbol{v}^k\|$ by r_k. Clearly, $r_k \in \mathbb{R}_+$ and it follows from the definition of bounded perturbations that $\sum_{k=0}^{\infty} r_k < \infty$.

We next prove by induction that, for every pair of non-negative integers k and i,

$$\|\boldsymbol{x}^{k+i} - (\mathbf{P}_T)^i \boldsymbol{x}^k\| \leq \sum_{j=k}^{k+i-1} r_j. \tag{A3}$$

Let k be an arbitrary non-negative integer. If $i = 0$, then the value is zero on both sides of the inequality and hence Eq. (A3) holds. Now assume that Eq. (A3) holds for an integer $i \geq 0$. Then, by Eq. (A2) and the nonexpansiveness of $\mathbf{P}_T$,

$$\begin{aligned}\|\boldsymbol{x}^{k+i+1} - (\mathbf{P}_T)^{i+1}\boldsymbol{x}^k\| &\leq \|\boldsymbol{x}^{k+i+1} - \mathbf{P}_T\boldsymbol{x}^{k+i}\| \\ &\quad + \|\mathbf{P}_T\boldsymbol{x}^{k+i} - (\mathbf{P}_T)^{i+1}\boldsymbol{x}^k\| \\ &\leq r_{k+i} + \|\boldsymbol{x}^{k+i} - (\mathbf{P}_T)^{i}\boldsymbol{x}^k\| \\ &\leq r_{k+i} + \sum_{j=k}^{k+i-1} r_j \\ &= \sum_{j=k}^{k+i} r_j, \end{aligned} \tag{A4}$$

which completes our inductive proof. A consequence of Eq. (A3) is that, for every pair of non-negative integers k and i,

$$\|\boldsymbol{x}^{k+i} - (\mathbf{P}_T)^{i}\boldsymbol{x}^k\| \leq \sum_{j=k}^{\infty} r_j. \tag{A5}$$

Due to the summability of the non-negative sequence $(r_k)_{k=0}^{\infty}$, the right-hand side (and hence the left-hand side) of this inequality gets arbitrarily close to zero as k increases.

Since $\mathcal{P}r_T$ is uniformly continuous, there exists a δ such that, for all $\boldsymbol{x}, \boldsymbol{y} \in \Omega$, $|\mathcal{P}r_T(\boldsymbol{x}) - \mathcal{P}r_T(\boldsymbol{y})| \leq \varepsilon' - \varepsilon$ provided that $\|\boldsymbol{x} - \boldsymbol{y}\| \leq \delta$. Select a k so that $\sum_{j=k}^{\infty} r_j \leq \delta$. By the assumption that $O(T, \varepsilon, ((\mathbf{P}_T)^k\boldsymbol{x})_{k=0}^{\infty})$ is defined for every $\boldsymbol{x} \in \Omega$, there exists a non-negative integer i for which $\mathcal{P}r((\mathbf{P}_T)^i\boldsymbol{x}^k) \leq \varepsilon$. From Eq. (A5) we have, for this k and i, that $\|\boldsymbol{x}^{k+i} - (\mathbf{P}_T)^i\boldsymbol{x}^k\| \leq \delta$ and, hence,

$$\begin{aligned}|\mathcal{P}r_T(\boldsymbol{x}^{k+i})| &\leq |\mathcal{P}r_T(\boldsymbol{x}^{k+i}) - \mathcal{P}r_T((\mathbf{P}_T)^i\boldsymbol{x}^k)| \\ &\quad + |\mathcal{P}r_T((\mathbf{P}_T)^i\boldsymbol{x}^k)| \\ &\leq (\varepsilon' - \varepsilon) + \varepsilon = \varepsilon', \end{aligned} \tag{A6}$$

proving that $O(T, \varepsilon', R)$ is defined. □

2. Nonascending vectors for convex functions

Theorem 2: Let $\phi : \mathbb{R}^J \to \mathbb{R}$ be a convex function and let $\boldsymbol{x} \in \mathbb{R}^J$. Let $\boldsymbol{g} \in \mathbb{R}^J$ satisfy the property: For $1 \leq j \leq J$, if the jth component g_j of $\boldsymbol{g}$ is not zero, then the partial derivative $\frac{\partial\phi}{\partial x_j}(\boldsymbol{x})$ of ϕ at $\boldsymbol{x}$ exists and its value is g_j. Define $\boldsymbol{d}$ to be the zero vector if $\|\boldsymbol{g}\| = 0$ and to be $-\boldsymbol{g}/\|\boldsymbol{g}\|$ otherwise. Then $\boldsymbol{d}$ is a nonascending vector for ϕ at $\boldsymbol{x}$.

Proof: The theorem is trivially true if $\|\boldsymbol{g}\| = 0$, so we assume that this is not the case. We denote by I the nonempty set of those indices j for which $g_j \neq 0$.

For $1 \leq j \leq J$, let s_j be $g_j/|g_j|$ for $j \in I$ and be 0 otherwise, and let $\boldsymbol{e}^j \in \mathbb{R}^J$ be the vector all of whose components are zero except for the jth, which is one. Then, for $1 \leq j \leq J$, there exists a $\delta_j > 0$ such that, for $0 \leq \lambda_j \leq \delta_j$,

$$\phi(\boldsymbol{x} - \lambda_j s_j \boldsymbol{e}^j) \leq \phi(\boldsymbol{x}). \tag{A7}$$

This is obvious if $s_j = 0$. Otherwise, $\frac{\partial\phi}{\partial x_j}(\boldsymbol{x})$ exists and indicates ϕ increases at $\boldsymbol{x}$ if $s_j = 1$ or that ϕ decreases at $\boldsymbol{x}$ if $s_j = -1$. The existence of the desired δ_j can be derived from the standard definition of the partial derivative as a limit.

We define $\delta > 0$ by

$$\delta = \frac{\|\boldsymbol{g}\|}{J} \min_{j \in I}\left\{\frac{\delta_j}{|g_j|}\right\}. \tag{A8}$$

Then we have that, for $0 \leq \lambda \leq \delta$,

$$\begin{aligned}\phi(\boldsymbol{x} + \lambda\boldsymbol{d}) &= \phi\left(\boldsymbol{x} - \lambda\sum_{j=1}^{J}\frac{|g_j|}{\|\boldsymbol{g}\|}s_j\boldsymbol{e}^j\right) \\ &= \phi\left(\sum_{j=1}^{J}\frac{1}{J}\left(\boldsymbol{x} - \lambda J\frac{|g_j|}{\|\boldsymbol{g}\|}s_j\boldsymbol{e}^j\right)\right) \\ &\leq \frac{1}{J}\sum_{j=1}^{J}\phi\left(\boldsymbol{x} - \lambda J\frac{|g_j|}{\|\boldsymbol{g}\|}s_j\boldsymbol{e}^j\right) \\ &\leq \frac{1}{J}\sum_{j=1}^{J}\phi(\boldsymbol{x}) \\ &= \phi(\boldsymbol{x}). \end{aligned} \tag{A9}$$

The first inequality above follows from the convexity of ϕ and the second one follows from Eq. (A7), with λ_j defined to be $\lambda J\frac{|g_j|}{\|\boldsymbol{g}\|}$, combined with Eq. (A8). Thus $\boldsymbol{d}$ is a nonascending vector for ϕ at $\boldsymbol{x}$. □

[a]Author to whom correspondence should be addressed. Electronic mail: gaborherman@yahoo.com; URL: http://www.dig.cs.gc.cuny.edu/gabor/index.html.

[1]J. O. Deasy, "Multiple local minima in radiotherapy optimization problems with dose-volume constraints," Med. Phys. **24**, 1157–1161 (1997).
[2]G. A. Ezzell, "Genetic and geometric optimization of three-dimensional radiation therapy treatment planning," Med. Phys. **23**, 293–305 (1996).
[3]A. Gustafsson, B. K. Lind, and A. Brahme, "A generalized pencil beam algorithm for optimization of radiation-therapy," Med. Phys. **21**, 343–357 (1994).
[4]A. Gustafsson, B. K. Lind, R. Svensson, and A. Brahme, "Simultaneous-optimization of dynamic multileaf collimation and scanning patterns or compensation filters using a generalized pencil beam algorithm," Med. Phys. **22**, 1141–1156 (1995).
[5]E. Lessard and J. Pouliot, "Inverse planning anatomy-based dose optimization for hdr-brachytherapy of the prostate using fast simulated annealing algorithm and dedicated objective function," Med. Phys. **28**, 773–779 (2001).
[6]R. Manzke, M. Grass, T. Nielsen, G. Shechter, and D. Hawkes, "Adaptive temporal resolution optimization in helical cardiac cone beam CT reconstruction," Med. Phys. **30**, 3072–3080 (2003).
[7]A. B. Pugachev, A. L. Boyer, and L. Xing, "Beam orientation optimization in intensity-modulated radiation treatment planning," Med. Phys. **27**, 1238–1245 (2000).
[8]D. M. Shepard, M. A. Earl, X. A. Li, S. Naqvi, and C. Yu, "Direct aperture optimization: A turnkey solution for step-and-shoot IMRT," Med. Phys. **29**, 1007–1018 (2002).
[9]C. Studholme, D. L. G. Hill, and D. J. Hawkes, "Automated three-dimensional registration of magnetic resonance and positron emission tomography brain images by multiresolution optimization of voxel similarity measures," Med. Phys. **24**, 25–35 (1997).
[10]Q. W. Wu and R. Mohan, "Algorithms and functionality of an intensity modulated radiotherapy optimization system," Med. Phys. **27**, 701–711 (2000).
[11]Y. Yu and M. C. Schell, "A genetic algorithm for the optimization of prostate implants," Med. Phys. **23**, 2085–2091 (1996).

[12]T. Z. Zhang, R. Jeraj, H. Keller, W. G. Lu, G. H. Olivera, T. R. McNutt, T. R. Mackie, and B. Paliwal, "Treatment plan optimization incorporating respiratory motion," Med. Phys. **31**, 1576–1586 (2004).
[13]M. Abdoli, M. R. Ay, A. Ahmadian, R. A. Dierckx, and H. Zaidi, "Reduction of dental filling metallic artifacts in CT-based attenuation correction of PET data using weighted virtual sinograms optimized by a genetic algorithm," Med. Phys. **37**, 6166–6177 (2010).
[14]S. Bartolac, S. Graham, J. Siewerdsen, and D. Jaffray, "Fluence field optimization for noise and dose objectives in CT," Med. Phys. **38**, S2–S17 (2011).
[15]W. Chen, D. Craft, T. M. Madden, K. Zhang, H. M. Kooy, and G. T. Herman, "A fast optimization algorithm for multicriteria intensity modulated proton therapy planning," Med. Phys. **37**, 4938–4945 (2010).
[16]J. Fiege, B. McCurdy, P. Potrebko, H. Champion, and A. Cull, "PARETO: A novel evolutionary optimization approach to multiobjective IMRT planning," Med. Phys. **38**, 5217–5229 (2011).
[17]A. Fredriksson, A. Forsgren, and B. Hardemark, "Minimax optimization for handling range and setup uncertainties in proton therapy," Med. Phys. **38**, 1672–1684 (2011).
[18]C. Holdsworth, M. Kim, J. Liao, and M. H. Phillips, "A hierarchical evolutionary algorithm for multiobjective optimization in IMRT," Med. Phys. **37**, 4986–4997 (2010).
[19]C. Holdsworth, R. D. Stewart, M. Kim, J. Liao, and M. H. Phillips, "Investigation of effective decision criteria for multiobjective optimization in IMRT," Med. Phys. **38**, 2964–2974 (2011).
[20]T. Kim, L. Zhu, T.-S. Suh, S. Geneser, B. Meng, and L. Xing, "Inverse planning for IMRT with nonuniform beam profiles using total-variation regularization (TVR)," Med. Phys. **38**, 57–66 (2011).
[21]C. Men, H. E. Romeijn, X. Jia, and S. B. Jiang, "Ultrafast treatment plan optimization for volumetric modulated arc therapy (VMAT)," Med. Phys. **37**, 5787–5791 (2010).
[22]S. N. Penfold, R. W. Schulte, Y. Censor, and A. B. Rosenfeld, "Total variation superiorization schemes in proton computed tomography image reconstruction," Med. Phys. **37**, 5887–5895 (2010).
[23]E. Y. Sidky, Y. Duchin, X. Pan, and C. Ullberg, "A constrained, total-variation minimization algorithm for low-intensity x-ray CT," Med. Phys. **38**, S117–S125 (2011).
[24]H. Stabenau, L. Rivera, E. Yorke, J. Yang, R. Lu, R. J. Radke, and A. Jackson, "Reduced order constrained optimization (ROCO): Clinical application to lung IMRT," Med. Phys. **38**, 2731–2741 (2011).
[25]Y. Yang and M. J. Rivard, "Dosimetric optimization of a conical breast brachytherapy applicator for improved skin dose sparing," Med. Phys. **37**, 5665–5671 (2010).
[26]X. Zhang, J. Wang, and L. Xing, "Metal artifact reduction in x-ray computed tomography (CT) by constrained optimization," Med. Phys. **38**, 701–711 (2011).
[27]D. Butnariu, R. Davidi, G. T. Herman, and I. G. Kazantsev, "Stable convergence behavior under summable perturbations of a class of projection methods for convex feasibility and optimization problems," IEEE J. Sel. Top. Signal Process. **1**, 540–547 (2007).
[28]R. Davidi, G. T. Herman, and Y. Censor, "Perturbation-resilient block-iterative projection methods with application to image reconstruction from projections," Int. Trans. Oper. Res. **16**, 505–524 (2009).
[29]Y. Censor, R. Davidi, and G. T. Herman, "Perturbation resilience and superiorization of iterative algorithms," Inverse Probl. **26**, 065008 (2010).
[30]T. Nikazad, R. Davidi, and G. T. Herman, "Accelerated perturbation-resilient block-iterative projection methods with application to image reconstruction," Inverse Probl. **28**, 035005 (2012).
[31]G. T. Herman and R. Davidi, "Image reconstruction from a small number of projections," Inverse Probl. **24**, 045011 (2008).
[32]E. Garduño, R. Davidi, and G. T. Herman, "Reconstruction from a few projections by ℓ_1-minimization of the Haar transform," Inverse Probl. **27**, 055006 (2011).
[33]R. L. Rardin and R. Uzsoy, "Experimental evaluation of heuristic optimization algorithms: A tutorial," J. Heuristics **7**, 261–304 (2001).
[34]L. Wernisch, S. Hery, and S. J. Wodak, "Automatic protein design with all atom force-fields by exact and heuristic optimization," J. Mol. Biol. **301**, 713–736 (2000).
[35]S. H. Zanakis and J. R. Evans, "Heuristic optimization: Why, when, and how to use it," Interfaces **11**, 84–91 (1981).
[36]G. T. Herman and W. Chen, "A fast algorithm for solving a linear feasibility problem with application to intensity-modulated radiation therapy," Linear Algebra Appl. **428**, 1207–1217 (2008).
[37]E. S. Helou Neto and Á. R. De Pierro, "Incremental subgradients for constrained convex optimization: A unified framework and new methods," SIAM J. Optim. **20**, 1547–1572 (2009).
[38]E. S. Helou Neto and Á. R. De Pierro, "On perturbed steepest descent methods with inexact line search for bilevel convex optimization," Optim. **60**, 991–1008 (2011).
[39]E. A. Nurminski, "Envelope stepsize control for iterative algorithms based on Fejer processes with attractants," Optim. Methods Software **25**, 97–108 (2010).
[40]P. L. Combettes and J. Luo, "An adaptive level set method for nondifferentiable constrained image recovery," IEEE Trans. Image Process. **11**, 1295–1304 (2002).
[41]P. L. Combettes and J.-C. Pesquet, "Image restoration subject to a total variation constraint," IEEE Trans. Image Process. **13**, 1213–1222 (2004).
[42]E. Y. Sidky and X. Pan, "Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization," Phys. Med. Biol. **53**, 4777–4807 (2008).
[43]J. Bian, J. H. Siewerdsen, X. Han, E. Y. Sidky, J. L. Prince, C. A. Pelizzari, and X. Pan, "Evaluation of sparse-view reconstruction from flat-panel-detector cone-beam CT," Phys. Med. Biol. **55**, 6575–6599 (2010).
[44]M. Defrise, C. Vanhove, and X. Liu, "An algorithm for total variation regularization in high-dimensional linear problems," Inverse Probl. **27**, 065002 (2011).
[45]Y. Censor, W. Chen, P. L. Combettes, R. Davidi, and G. T. Herman, "On the effectiveness of projection methods for convex feasibility problems with linear inequality constraints," Comput. Optim. Appl. **51**, 1065–1088 (2012).
[46]J. Bioucas-Dias and M. Figueiredo, "A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration," IEEE Trans. Image Process. **16**, 2992–3004 (2007).
[47]T. Goldstein and S. Osher, "The split Bregman method for L1 regularized problems," SIAM J. Imaging Sci. **2**, 323–343 (2009).
[48]L. A. Shepp and Y. Vardi, "Maximum likelihood reconstruction for emission tomography," IEEE Trans. Med. Imaging **1**, 113–122 (1982).
[49]E. Levitan and G. T. Herman, "A maximum *a posteriori* probability expectation maximization algorithm for image reconstruction in emission tomography," IEEE Trans. Med. Imaging **6**, 185–192 (1987).
[50]W. Jin, Y. Censor, and M. Jiang, "A heuristic superiorization-like approach to bioluminescence tomography," in *Proceedings of the International Federation for Medical and Biological Engineering (IFMBE)* (Springer-Verlag, Berlin, 2012), Vol. 39, pp. 1026–1029.
[51]H. M. Hudson and R. S. Larkin, "Accelerated image reconstruction using ordered subsets of projection data," IEEE Trans. Med. Imaging **13**, 601–609 (1994).
[52]T. Elfving, "Block-iterative methods for consistent and inconsistent linear equations," Numer. Math. **35**, 1–12 (1980).
[53]P. P. B. Eggermont, G. T. Herman, and A. Lent, "Iterative algorithms for large partitioned linear systems, with applications to image reconstruction," Linear Algebra Appl. **40**, 37–67 (1981).
[54]R. Aharoni and Y. Censor, "Block-iterative projection methods for parallel computation of solutions to convex feasibility problems," Linear Algebra Appl. **120**, 165–175 (1989).
[55]G. T. Herman, *Fundamentals of Computerized Tomography: Image Reconstruction from Projections*, 2nd ed. (Springer, New York, 2009).
[56]J. F. P. J. Abascal, J. Chamorro-Servent, J. Aguirre, S. Arridge, T. Correia, J. Ripoli, J. J. Vaquero, and M. Desco, "Fluorescence diffuse optical tomography using the split Bregman method," Med. Phys. **38**, 6275–6284 (2011).
[57]R. Davidi, G. T. Herman, and J. Klukowska, SNARK09: A programming system for the reconstruction of 2D images from 1D projections, 2009 (available URL: http://www.snark09.com).
[58]G. T. Herman and L. B. Meyer, "Algebraic reconstruction techniques can be made computationally efficient," IEEE Trans. Med. Imaging **12**, 600–609 (1993).
[59]W. Zbijewski and F. J. Beekman, "Characterization and suppression of edge and aliasing artefacts in iterative x-ray CT reconstruction," Phys. Med. Biol. **49**, 145–157 (2004).

J Optim Theory Appl
DOI 10.1007/s10957-013-0408-3

Projected Subgradient Minimization Versus Superiorization

Yair Censor · Ran Davidi · Gabor T. Herman · Reinhard W. Schulte · Luba Tetruashvili

Received: 5 February 2013 / Accepted: 17 August 2013

Abstract The projected subgradient method for constrained minimization repeatedly interlaces subgradient steps for the objective function with projections onto the feasible region, which is the intersection of closed and convex constraints sets, to regain feasibility. The latter poses a computational difficulty, and, therefore, the projected subgradient method is applicable only when the feasible region is "simple to project onto." In contrast to this, in the superiorization methodology a feasibility-seeking algorithm leads the overall process, and objective function steps are interlaced into it. This makes a difference because the feasibility-seeking algorithm employs projections onto the individual constraints sets and not onto the entire feasible region.

We present the two approaches side-by-side and demonstrate their performance on a problem of computerized tomography image reconstruction, posed as a constrained minimization problem aiming at finding a constraint-compatible solution that has a reduced value of the total variation of the reconstructed image.

Communicated by Masao Fukushima.

Y. Censor (✉) · L. Tetruashvili
Department of Mathematics, University of Haifa, Mt. Carmel, 3190501 Haifa, Israel
e-mail: yair@math.haifa.ac.il

R. Davidi
Department of Radiation Oncology, Stanford University, Stanford, CA 94305, USA

G.T. Herman
Department of Computer Science, The Graduate Center, City University of New York, New York, NY 10016, USA

R.W. Schulte
Department of Radiation Medicine, Loma Linda University Medical Center, Loma Linda, CA 92354, USA

Published online: 10 September 2013

Keywords Constrained minimization · Feasibility-seeking · Bounded convergence · Superiorization · Projected subgradient method · Proximity function · Strong perturbation resilience · Image reconstruction · Computerized tomography

1 Introduction

Our aim in this paper is to expose the recently developed superiorization methodology and its ideas to the optimization community by "confronting" it with the projected subgradient method. We juxtapose the *projected subgradient method* (PSM) with the *superiorization methodology* (SM) and demonstrate their performance on a large-size real-world application that is modeled, and needs to be solved, as a constrained minimization problem. The PSM for constrained minimization has been extensively investigated, see, e.g., [1, Sect. 7.1.2], [2, Sect. 3.2.3]. Its roots are in the work of Shor [3] for the unconstrained case and in the work of Polyak [4, 5] for the constrained case. More recent work can be found in, e.g., [6]. The superiorization methodology was first proposed in [7], although without using the term superiorization. In that work, perturbation resilience (without using this term) was proved for the general class of *string-averaging projection* (SAP) methods, see [8–12], that use orthogonal projections and relate to consistent constraints. Subsequent investigations and developments of the SM were done in [13–17]. More information on superiorization-related work is given in Sect. 3.

It is not claimed that the PSM is the best optimization method for solving constrained minimization problems and there are many different alternative methods with which SM could be compared. So, why did we chose to confront the PSM with our SM? In a nutshell, our answer is that both methods interlace steps related to the objective function with steps oriented toward feasibility, but they differ in how they restore or preserve feasibility. A major difficulty with the PSM is the need to perform, within each iterative step, an orthogonal projection onto the feasible set of the constrained minimization problem. If the feasible set is not "simple to project onto," then the projection requires an independent inner-loop calculation to minimize the distance from a point to the feasible set, which can be costly and hamper the overall effectiveness of the PSM.

In the SM, we replace the notion of a fixed feasible set by that of a nonnegative real-valued proximity function. This function serves as an indicator of how incompatible a vector is with the constraints. In such a formulation, the merit of an actual output vector of any algorithm is indicated by the smallness of the two numbers, i.e., the values of the proximity function and the objective function. The underlying idea of SM is that many iterative algorithms that produce outputs for which the proximity function is small are *strongly perturbation resilient* in the sense that, even if certain kinds of changes are made at the end of each iterative step, the algorithm still produces an output for which the proximity function is not larger. This property is exploited by using permitted changes to steer the algorithm to an output that has not only a small proximity function value, but has also a small objective function value.

The PSM requires that feasibility is regained after each subgradient step by performing a projection onto the entire feasible set, whereas in the SM the feasibility-seeking projection method proceeds by projecting (in a well-defined algorithmically

structured regime dictated by the specific projection method) onto the individual sets, whose intersection is the entire feasible set, and not onto the whole feasible set itself. This has a potentially great computational advantage.

We elaborate on the motivation for this work in Sect. 2. In Sect. 3 we discuss some superiorization-related work, in Sect. 4 the SM is presented, and in Sect. 5 we demonstrate the approaches of the SM and the PSM on a realistically-large-size problem with data that arise from the significant problem of x-ray computed tomography (CT) with total variation (TV) minimization, followed by some conclusions in Sect. 6.

2 Motivation and Basic Notions

Throughout this paper, we assume that Ω is a nonempty subset of the J-dimensional Euclidean space $\mathbb{R}^J$. We consider constrained minimization problems of the form

$$\text{minimize}\big\{\phi(x) \mid x \in C\big\}, \tag{1}$$

where $\phi : \mathbb{R}^J \to \mathbb{R}$ is an objective function, and $C \subseteq \Omega$ is a given feasible set.

Since we juxtapose the *projected subgradient method* (PSM) with the *superiorization methodology* (SM) and demonstrate their performance on a large-size real-world application that is modeled, and needs to be solved, as a constrained minimization problem, we now outline these two methods and explain our choice in detail.

In order to apply the PSM to solving (1), we need to assume that C is a nonempty closed convex set and that ϕ is a convex function. The PSM generates a sequence of iterates $\{x^k\}_{k=0}^{\infty}$ according to the recursion formula

$$x^{k+1} = P_C\big(x^k - t_k \phi'\big(x^k\big)\big), \tag{2}$$

where $t_k > 0$ is a step-size, $\phi'(x^k) \in \partial\phi(x^k)$ is a subgradient of ϕ at x^k, and P_C stands for the orthogonal (least Euclidean norm) projection onto the set C.

A major difficulty with (2) is the need to perform, within each iterative step, the orthogonal projection. If the feasible set C is not "simple to project onto," then the projection requires an independent inner-loop calculation to minimize the distance from the point $x^k - t_k\phi'(x^k)$ to the set C, which can be costly and hamper the overall effectiveness of an algorithm that uses (2). Also, if the inner loop converges to the projection onto C only in the limit, then, in practical implementations, it will have to be stopped after a finite number of steps, and so x^{k+1} will be only an approximation to the projection onto C, and it could even happen that it is not in C.

Even if we set aside our worries about projecting onto C in (2), there are still two concerns when applying the PSM to real-world problems. One is that the iterative process usually converges to the desired solution only in the limit. In practice, some stopping rule is applied to terminate the process, and the output at that time may not even be in C, and, even if it is in C, it is most unlikely to be the minimizer of ϕ over C. The second problem in real-world applications comes from the fact that the constraints, derived from the real-world problem, may not be consistent (e.g., because they come from noisy measurements), and so C is empty.

Similar criticism applies actually to many constrained-minimization-seeking algorithms for which asymptotic convergence results are available. In the SM, both of these objections can be handled by replacing the notion of a fixed feasible set C by that of a nonnegative real-valued proximity function $\mathrm{Prox}_C : \Omega \to \mathbb{R}_+$. This function serves as an indicator of how incompatible a vector x is with the constraints. In such a formulation, the merit of the actual output x of any algorithm is indicated by the smallness of the two numbers $\mathrm{Prox}_C(x)$ and $\phi(x)$. For the formulation of (1), we would define Prox_C so that its range is the ray of nonnegative real numbers with $\mathrm{Prox}_C(x) = 0$ if, and only if, $x \in C$, and then the constrained minimization problem (1) is precisely that of finding an x that is a minimizer of $\phi(x)$ over $\{x \mid \mathrm{Prox}_C(x) = 0\}$. The above discussion allows us to do away with the nonemptiness assumption and also to compare the merits of actual outputs of algorithms that only approximate the aim of the constrained minimization problem.

The recently invented SM incorporates the ideas of the previous paragraph in its very foundation and formulates the problem with the function Prox_C instead of the set C. The underlying idea of SM is that many iterative algorithms that produce outputs x for which $\mathrm{Prox}_C(x)$ is small are *strongly perturbation resilient* in the sense that, even if certain kinds of changes are made at the end of each iterative step, the algorithm still produces an output x' for which $\mathrm{Prox}_C(x')$ is not larger. This property is exploited by using permitted changes to steer the algorithm to an output that has not only a small Prox_C value, but has also a small ϕ value. The algorithm that incorporates such a steering process is referred to as the *superiorized version* of the original iterative algorithm. The main practical contribution of SM is the automatic creation of the superiorized version, according to a given objective function ϕ, of just about any iterative algorithm that aims at producing an x for which $\mathrm{Prox}_C(x)$ is small.

Nevertheless, in order to carry out our comparative study, we restrict our attention here to a subset of all possible problems to which not only the SM but also the PSM is applicable. We assume that we are given a family of constraints $\{C_\ell\}_{\ell=1}^{L}$, where each set C_ℓ is a nonempty closed convex subset of $\mathbb{R}^J$ such that

$$C = \bigcap_{\ell=1}^{L} C_\ell \tag{3}$$

is a nonempty subset of Ω and that it is the feasible set C of (1). Under these assumptions, we illustrate the application of the SM by the superiorization of feasibility-seeking *projection methods*, see, e.g., [18–22] and the recent monograph [23]. Such methods use projections onto the individual sets C_ℓ in order to generate a sequence $\{x^k\}_{k=0}^{\infty}$ that converges to a point $x^* \in C$. Therefore, contrary to the PSM, one does not need to assume that C is a "simple to project onto" set, but rather that the individual sets C_ℓ have this property. The latter is indeed often the case, such as, for example, where the sets C_ℓ are hyperplanes or half-spaces onto which we can project easily, but their intersection is not "simple to project onto."

The SM is accurately presented in Sect. 4 below. However, the discussion above is sufficient to explain why we chose the PSM and the SM for our comparative study. Namely, both methods interlace objective-function-reduction steps with steps

oriented toward feasibility. But exactly here lies a big difference between the two approaches. The PSM requires that feasibility is regained after subgradient nonascent steps by performing a projection onto C, whereas in the SM the feasibility-seeking projection method proceeds by projecting (in a well-defined algorithmically structured regime dictated by the specific projection method) onto the individual sets C_ℓ and not onto the whole feasible set C. This has a potentially great computational advantage.

3 Superiorization-Related Previous Work

The superiorization methodology was first proposed in [7], although without using the term superiorization. In that work, perturbation resilience (without using this term) was proved for the general class of *string-averaging projection* (SAP) methods, see [8–12], that use orthogonal projections and relate to consistent constraints. Subsequent investigations and developments were done in [13–17]. In [13], the methodology was formulated over general *problem structures* that enabled rigorous analysis and revealed that the approach is not limited to feasibility and optimization. In [14], perturbation resilience was analyzed for the class of *block-iterative projection* (BIP) methods, see [18–22], and applied in this manner. In [15], the advantages of superiorization for image reconstruction from a small number of projections was studied, and in [16] two acceleration schemes based on (symmetric and nonsymmetric) BIP methods were proposed and experimented with. In [17], total variation superiorization schemes in proton computed tomography (pCT) image reconstruction were investigated.

In [24], we introduced the notion of ε-compatibility into the superiorization approach in order to handle inconsistent constraints. This enabled us to close the logical discrepancy between the assumption of consistency of constraints and the actual experimental work done previously. We also introduced there the new notion of strong perturbation resilience, which generalizes the previously used notion of perturbation resilience. Algorithmically, the new superiorized algorithm introduced there (and used here) is different from all previous ones in that it uses the notion of nonascending direction and in that it allows several perturbation steps for each feasibility-seeking step, an aspect that has practical advantages.

In [25], superiorization was applied to the *expectation maximization* (EM) algorithm instead of the feasibility-seeking projection methods that were used in superiorization previously. The approach was implemented there to solve an inverse problem of *bioluminescence tomography* (BLT) image reconstruction. Such EM superiorization was investigated further and applied to a problem of *Single Photon Emission Computed Tomography* (SPECT) in [26]. Most recently, in [27], the SM was further investigated numerically, along with many projection methods for the feasibility problem and for the best approximation problem.

Our superiorization methodology should be distinguished from the works of Helou Neto and De Pierro [28, 29], of Nedić [30], Ram, Nedić, and Veeravalli [31], and of Nurminski [32–35]. The lack of cross-referencing between some of these papers shows that, in spite of the similarities between their approaches, their results were apparently reached independently.

There are various differences among the works mentioned in the previous paragraph, differences in overall setup of the problems, differences in the assumptions used for the various convergence results, etc. This is not the place for a full review of all these differences. But we wish to clarify the fundamental difference between them and the SM. The point is that when two activities are interlaced, here, feasibility steps and objective function reduction steps, then once the process is running all such methods look alike. From looking at the iterative formulas, one cannot tell if (a) "feasibility steps are interlaced into an iterative gradient scheme for objective function minimization" or if (b) "objective function reduction steps are interlaced into an iterative projections scheme for feasibility-seeking." The common thread of all works mentioned in the previous paragraph is that they fall into the category (a), while the SM is of the kind (b). In all methods of category (a) the condition that is needed to guarantee convergence to a constrained minimum point is that the diminishing step-sizes $\alpha_k \to 0$ as $k \to \infty$ must be such that $\sum_{k=0}^{\infty} \alpha_k = +\infty$. In contrast, since the feasibility-seeking projection method is the "leader" of the overall process in the SM, we must have that the perturbations (that do the objective function reduction) will use diminishing step-sizes $\beta_k \to 0$ as $k \to \infty$ but such that $\sum_{k=0}^{\infty} \beta_k < \infty$. The latter condition guarantees the perturbation resilience of the original feasibility-seeking projection method so that, regardless of the interlaced objective function reduction steps, the overall process converges to a feasible, or ε-compatible, point of the constraints.

Yet another fundamental difference between the superiorization methodology and the algorithms of category (a) mentioned above is that those algorithms perform the interlaced objective function descent and feasibility steps alternatingly according to a rigid predetermined scheme, whereas in the superiorization methodology the activation of these steps and the decisions whether to keep an iterate or discard it are done inside the superiorized algorithm in a controlled and automatically supervised manner. Thus, the superiorization methodology has the following features not present in the algorithms of category (a) mentioned above: (i) it conducts iterations of a feasibility-seeking projection method which is strongly perturbation resilient (as defined below), (ii) it interlaces objective function nonascent steps into the process in a controlled and automatically supervised manner, (iii) it is not known to guarantee convergence to a solution of the constrained minimization problem, and it might (we do not know if this is so or not) instead only be shown to lead to a feasible point whose objective function value is less than that of a feasible point that would have been reached by the same feasibility-seeking projection method without the perturbations exercised by the superiorized algorithm.

The *adaptive steepest descent projections onto convex sets* (ASD-POCS) algorithm described in [36] has some similarities to the SM. However, it is not as general as the SM; see [24] for a comparison.

4 The Superiorization Methodology

In this section we present a restricted version of the SM of [24] adapted to our problem (1). As discussed in Sect. 2, we associate with the feasible set C in (1) a proximity

function $\text{Prox}_C : \Omega \to \mathbb{R}_+$ that is an indicator of how incompatible an $x \in \Omega$ is with the constraints. For any given $\varepsilon > 0$, a point $x \in \Omega$ for which $\text{Prox}_C(x) \leq \varepsilon$ is called an *ε-compatible solution* for C. We further assume that we have, for the C in (1), a feasibility-seeking *algorithmic operator* $\boldsymbol{A}_C : \mathbb{R}^J \to \Omega$, with which we define the following basic algorithm.

The Basic Algorithm

(B1) **Initialization**: Choose an arbitrary $x^0 \in \Omega$,
(B2) **Iterative Step**: Given the current iterate x^k, calculate the next iterate x^{k+1} by

$$x^{k+1} = \boldsymbol{A}_C\big(x^k\big). \tag{4}$$

The following definition helps to evaluate the output of the Basic Algorithm upon termination by a stopping rule.

Definition 4.1 (The ε-output of a sequence) Given $C \subseteq \mathbb{R}^J$, a proximity function $\text{Prox}_C : \Omega \to \mathbb{R}_+$, a sequence $\{x^k\}_{k=0}^{\infty} \subset \Omega$ and an $\varepsilon > 0$, then an element x^K of the sequence which has the properties: (i) $\text{Prox}_C(x^K) \leq \varepsilon$, and (ii) $\text{Prox}_C(x^k) > \varepsilon$ for all $0 \leq k < K$, is called an ε-`output of the sequence` $\{x^k\}_{k=0}^{\infty}$ `with respect to the pair` (C, Prox_C). We denote it by $O(C, \varepsilon, \{x^k\}_{k=0}^{\infty}) = x^K$.

Clearly, an ε-output $O(C, \varepsilon, \{x^k\}_{k=0}^{\infty})$ of a sequence $\{x^k\}_{k=0}^{\infty}$ might or might not exist, but if it does, then it is unique. If $\{x^k\}_{k=0}^{\infty}$ is produced by an algorithm intended for the feasible set C, such as the Basic Algorithm, without a termination criterion, then $O(C, \varepsilon, \{x^k\}_{k=0}^{\infty})$ is the *output* produced by that algorithm when it includes the termination rule to stop when an ε-compatible solution for C is reached.

Definition 4.2 (Strong perturbation resilience) Assume that we are given a $C \subseteq \Omega$, a proximity function Prox_C, an algorithmic operator $\boldsymbol{A}_C$ and an $x^0 \in \Omega$. We use $\{x^k\}_{k=0}^{\infty}$ to denote the sequence generated by the Basic Algorithm when it is initialized by x^0. The Basic Algorithm is said to be `strongly perturbation resilient` iff the following hold:

(i) there exists an $\varepsilon > 0$ such that the ε-output $O(C, \varepsilon, \{x^k\}_{k=0}^{\infty})$ exists for every $x^0 \in \Omega$;
(ii) for every $\varepsilon > 0$, for which the ε-output $O(C, \varepsilon, \{x^k\}_{k=0}^{\infty})$ exists for every $x^0 \in \Omega$, we have also that the ε'-output $O(C, \varepsilon', \{y^k\}_{k=0}^{\infty})$ exists for every $\varepsilon' > \varepsilon$ and for every sequence $\{y^k\}_{k=0}^{\infty}$ generated by

$$y^{k+1} = \boldsymbol{A}_C\big(y^k + \beta_k v^k\big) \quad \text{for all } k \geq 0, \tag{5}$$

where the vector sequence $\{v^k\}_{k=0}^{\infty}$ is bounded, and the scalars $\{\beta_k\}_{k=0}^{\infty}$ are such that $\beta_k \geq 0$ for all $k \geq 0$ and $\sum_{k=0}^{\infty} \beta_k < \infty$.

Definition 4.3 (Bounded convergence) Assume that we are given a $C \subseteq \mathbb{R}^J$, a proximity function Prox_C, and an algorithmic operator $\boldsymbol{A}_C : \mathbb{R}^J \to \Omega$. Then the Basic Algorithm is said to be `convergent over` Ω iff for every $x^0 \in \Omega$, there exists

the limit $\lim_{k\to\infty} x^k = y(x^0)$ and $y(x^0) \in \Omega$. It is said to be `boundedly convergent over` Ω iff, in addition, there exists a $\gamma \geq 0$ such that $\text{Prox}_C(y(x^0)) \leq \gamma$ for every $x^0 \in \Omega$.

Next theorem, which gives sufficient conditions for strong perturbation resilience of the Basic Algorithm, has been proved in [24, Theorem 1] (in different wording).

Theorem 4.1 *Assume that we are given a* $C \subseteq \mathbb{R}^J$, *a proximity function* Prox_C, *and an algorithmic operator* $\boldsymbol{A}_C : \mathbb{R}^J \to \Omega$. *If* $\boldsymbol{A}_C$ *is nonexpansive and is such that it defines a boundedly convergent Basic Algorithm and if the proximity function* Prox_C *is uniformly continuous, then the Basic Algorithm defined by* $\boldsymbol{A}_C$ *is strongly perturbation resilient.*

Along with the $C \subseteq \mathbb{R}^J$, we look at the objective function $\phi : \mathbb{R}^J \to \mathbb{R}$, with the convention that a point in $\mathbb{R}^J$ for which the value of ϕ is smaller is considered *superior* to a point in $\mathbb{R}^J$ for which the value of ϕ is larger. The essential idea of the SM is to make use of the perturbations of (5) to transform a strongly perturbation resilient algorithm that seeks a constraints-compatible solution for C into one whose outputs are equally good from the point of view of constraints-compatibility, but are superior (not necessarily optimal) according to the objective function ϕ.

This is done by producing from the Basic Algorithm another algorithm, called its *superiorized* version, that makes sure not only that the $\beta_k v^k$ are bounded perturbations, but also that $\phi(y^k + \beta_k v^k) \leq \phi(y^k)$ for all k. To do so, we use the next concept, closely related to the concept of "descent direction."

Definition 4.4 Given a function $\phi : \mathbb{R}^J \to \mathbb{R}$ and a point $y \in \mathbb{R}^J$, we say that a vector $d \in \mathbb{R}^J$ is `nonascending for` ϕ `at` y iff $\|d\| \leq 1$ and there is a $\delta > 0$ such that

$$\text{for all } \lambda \in [0, \delta], \quad \text{we have } \phi(y + \lambda d) \leq \phi(y). \tag{6}$$

Obviously, the zero vector is always such a vector, but for superiorization to work, we need a sharp inequality to occur in (6) frequently enough.

The Superiorized Version of the Basic Algorithm assumes that we have available a summable sequence $\{\eta_\ell\}_{\ell=0}^{\infty}$ of positive real numbers (for example, $\eta_\ell = a^\ell$, where $0 < a < 1$) and it generates, simultaneously with the sequence $\{y^k\}_{k=0}^{\infty}$ in Ω, sequences $\{v^k\}_{k=0}^{\infty}$ and $\{\beta_k\}_{k=0}^{\infty}$. The latter is generated as a subsequence of $\{\eta_\ell\}_{\ell=0}^{\infty}$, resulting in a nonnegative summable sequence $\{\beta_k\}_{k=0}^{\infty}$. The algorithm further depends on a specified initial point $y^0 \in \Omega$ and on a positive integer N. It makes use of a logical variable called *loop*. The superiorized algorithm is presented next by its pseudo-code.

Superiorized Version of the Basic Algorithm

1. **set** $k = 0$
2. **set** $y^k = y^0$
3. **set** $\ell = -1$
4. **repeat**

5. **set** $n = 0$
6. **set** $y^{k,n} = y^k$
7. **while** $n < N$
8. **set** $v^{k,n}$ to be a nonascending vector for ϕ at $y^{k,n}$
9. **set** *loop* = *true*
10. **while** *loop*
11. **set** $\ell = \ell + 1$
12. **set** $\beta_{k,n} = \eta_\ell$
13. **set** $z = y^{k,n} + \beta_{k,n} v^{k,n}$
14. **if** $\phi(z) \leq \phi(y^k)$ **then**
15. **set** $n = n + 1$
16. **set** $y^{k,n} = z$
17. **set** *loop* = *false*
18. **set** $y^{k+1} = \boldsymbol{A}_C(y^{k,N})$
19. **set** $k = k + 1$

Theorem 4.2 *Any sequence* $\{y^k\}_{k=0}^{\infty}$, *generated by the Superiorized Version of the Basic Algorithm, satisfies* (5). *Further, if, for a given* $\varepsilon > 0$, *the* ε*-output* $O(C, \varepsilon, \{x^k\}_{k=0}^{\infty})$ *of the Basic Algorithm exists for every* $x^0 \in \Omega$, *then every sequence* $\{y^k\}_{k=0}^{\infty}$, *generated by the Superiorized Version of the Basic Algorithm, has an* ε'*-output* $O(C, \varepsilon', \{y^k\}_{k=0}^{\infty})$ *for every* $\varepsilon' > \varepsilon$.

This theorem follows from the analysis of the behavior of the Superiorized Version of the Basic Algorithm in [24]. In other words, the Superiorized Version produces outputs that are essentially as constraints-compatible as those produced by the original not superiorized algorithm. However, due to the repeated steering of the process by lines 7 to 17 toward reducing the value of the objective function ϕ, we can expect that the output of the Superiorized Version will be superior (from the point of view of ϕ) to the output of the original algorithm.

5 A Computational Demonstration

5.1 The x-Ray CT Problem

The fully discretized model in the series expansion approach to the image reconstruction problem of x-ray computerized tomography (CT) is formulated in the following manner. A Cartesian grid of square picture elements, called *pixels*, is introduced into the region of interest so that it covers the whole picture that has to be reconstructed. The pixels are numbered in some agreed manner, say from 1 (top left corner pixel) to J (bottom right corner pixel).

The x-ray attenuation function is assumed to take a constant value x_j throughout the jth pixel for $j = 1, 2, \ldots, J$. Sources and detectors are assumed to be points, and the rays between them are assumed to be lines. Further, assume that the length of intersection of the ith ray with the jth pixel, denoted by a_j^i, for $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$, represents the weight of the contribution of the jth pixel to the total attenuation along the ith ray.

The physical measurement of the total attenuation along the ith ray, denoted by b_i, represents the line integral of the unknown attenuation function along the path of the ray. Therefore, in this fully discretized model, the line integral turns out to be a finite sum, and the model is described by a system of linear equations

$$\sum_{j=1}^{J} x_j a_j^i = b_i, \quad i = 1, 2, \ldots, I. \tag{7}$$

In matrix notation we rewrite (7) as

$$Ax = b, \tag{8}$$

where $b \in \mathbb{R}^I$ is the *measurement vector*, $x \in \mathbb{R}^J$ is the *image vector*, and the $I \times J$ matrix $A = (a_j^i)$ is the *projection matrix*. See [37], especially Sect. 6.3, for a complete treatment of this subject.

5.2 The Algorithms that We Use

In this section we describe the PSM and SM algorithms specifically used in our demonstration. We applied both algorithms to solve the fully discretized model in the series expansion approach to the image reconstruction problem of x-ray CT, formulated in the previous section and represented by the optimization problem

$$\text{minimize}\big\{\phi(x) \mid Ax = b \text{ and } 0 \le x \le 1\big\}. \tag{9}$$

The box constraints are natural for this problem: If x_j represents the linear attenuation coefficient, measured in cm^{-1}, at a medically used x-ray energy spectrum in the jth pixel, then the box constraints $0 \le x \le 1$ are reasonable for tissues in the human body; see Table 4.1 of [37]. Hence, for the image reconstruction problem of x-ray CT, we define Ω by

$$\Omega = \big\{x \in \mathbb{R}^J \mid 0 \le x \le 1\big\}. \tag{10}$$

We note that this Ω is bounded.

The choice of C in (1) is of the type specified in (3), with $L = I + 1$, $C_i = \{x \in \mathbb{R}^J \mid \langle a^i, x\rangle = b_i\}$ for $i = 1, 2, \ldots, I$ and $C_{I+1} = \Omega$. Furthermore, since in the experiment reported below, we start with a specific image vector $x \in \Omega$ and calculate from it the measurement vector $b \in \mathbb{R}^I$ using (7), we know that C is a nonempty subset of Ω, which is the requirement stated below (3).

For any such C, we define $\text{Prox}_C : \Omega \to \mathbb{R}_+$ by

$$\text{Prox}_C(x) = \sqrt{\sum_{i=1}^{I} \big(b_i - \langle a^i, x\rangle\big)^2}. \tag{11}$$

Note that this proximity function Prox_C is uniformly continuous and thus satisfies the condition stated for it in Theorem 4.1.

Our choice for the objective function ϕ is the total variation (TV) of the image vector x. Denoting the $G \times H$ image array X ($GH = J$) obtained from the image vector x by $X_{g,h} = x_{(g-1)H+h}$ for $1 \le g \le G$ and $1 \le h \le H$, we use

$$\phi(x) = \mathrm{TV}(X) = \sum_{g=1}^{G-1} \sum_{h=1}^{H-1} \sqrt{(X_{g+1,h} - X_{g,h})^2 + (X_{g,h+1} - X_{g,h})^2}. \tag{12}$$

5.2.1 The Projected Subgradient Method

We implemented the PSM with the choice of C and the objective function ϕ described above. We used the PSM recursion formula (2) and adopted a nonsummable diminishing step-length rule of the form $t_k = \gamma_k / \|\phi'(x^k)\|$, where $\gamma_k \ge 0$, $\lim_{k\to\infty} \gamma_k = 0$, and $\sum_{k=0}^{\infty} \gamma_k = \infty$.

The PSM Algorithm

(P1) **Initialization**: Select a point $x^0 \in \mathbb{R}^J$, select integers K and M, use two real number variables $\boldsymbol{curr}$ and $\boldsymbol{prev}$, and set $\boldsymbol{curr} = \phi(x^0)$ and $\boldsymbol{prev} = \boldsymbol{curr}$.

(P2) **Iterative step**: Given the current iterate x^k, calculate the next one as follows:

(P2.1) Calculate a subgradient of ϕ at x^k, i.e., $\phi'(x^k) \in \partial\phi(x^k)$, a step-size $t_k = k^{-1/4}/\|\phi'(x^k)\|_2$, and the vector

$$q^k = x^k - t_k \phi'\big(x^k\big). \tag{13}$$

(P2.2) Calculate the next iterate as the projection of q^k onto C by solving

$$x^{k+1} = \arg\min_x \left\{ \frac{1}{2} \big\| x - q^k \big\|^2 \mid Ax = b \text{ and } 0 \le x \le 1 \right\}. \tag{14}$$

(P2.3) If $\phi(x^{k+1}) \le \boldsymbol{curr}$, then $\boldsymbol{curr} = \phi(x^{k+1})$.

(P3) **Stopping rule**: If $k \bmod K = 0$ (i.e., k is divisible by K), then: If $\boldsymbol{prev} - \boldsymbol{curr} < \boldsymbol{prev}/M$ then stop. Otherwise, $\boldsymbol{prev} = \boldsymbol{curr}$ and go to (P2).

That the PSM algorithm converges to a solution of (1) follows from [2, Sect. 3.2.3], in particular, from Theorem 3.2.2 therein, provided that ϕ is convex and locally Lipschitz continuous and C is closed and convex. The latter is indeed the case for the C in (9). The convexity of the ϕ of (12) follows from the end of the proof of Proposition 1 in [38]. Its Lipschitz continuity on the whole space $\mathbb{R}^J$ follows from the fact that the TV function can be rewritten as

$$\mathrm{TV}(X) = \sum_{g=1}^{G-1} \sum_{h=1}^{H-1} \|A_{g,h} X\|_2, \tag{15}$$

where $A_{g,h}$ is a square matrix having only two nonzero rows, with the first nonzero row containing only two nonzero elements 1 and -1 that correspond to the variables $X_{g+1,h}$ and $X_{g,h}$, respectively, and the second nonzero row containing only two nonzero elements 1 and -1 that correspond to the variables $X_{g,h+1}$ $X_{g,h}$, respectively.

In our implementation we solved problem (14), in step (P2.2) above, by considering its dual

$$\text{maximize}\{f(\lambda) \mid \lambda \in \mathbb{R}^I\}, \tag{16}$$

where

$$f(\lambda) = \frac{1}{2}\|q^k - A^T\lambda - P_{C_{I+1}}(q^k - A^T\lambda)\|^2 - \frac{1}{2}\|q^k - A^T\lambda\|^2 - \langle\lambda, b\rangle + \frac{1}{2}\|q^k\|^2. \tag{17}$$

The optimal point x^{*k} of (14) is then

$$x^{*k} = P_{C_{I+1}}(q^k - A^T\lambda^{*k}), \tag{18}$$

where λ^{*k} is the optimal solution of (16). To find λ^{*k}, we minimized $-f(\lambda)$ using the Optimal Method of Nesterov [39], as generalized by Güler [40, p. 188], whose generic description for unconstrained minimization of a convex function $\theta(\lambda)$, which is continuously differentiable with Lipschitz continuous gradient, is as follows.

(N1) **Initialization:** Select a $\mu^0 \in \mathbb{R}^J$ and a positive α_{-1} and put $\lambda^{-1} = \mu^0$, $\beta_0 = 1$, and $k = 0$.

(N2) **Iterative Step:** Given λ^{k-1}, μ^k, α_{k-1}, and β_k:

(N2.1) Calculate the smallest index $s \geq 0$ for which the following inequality holds:

$$\theta(\mu^k) - \theta(\mu^k - 2^{-s}\alpha_{k-1}\nabla\theta(\mu^k)) \geq 2^{-s-1}\alpha_{k-1}\|\nabla\theta(\mu^k)\|^2. \tag{19}$$

(N2.2) Calculate the next iterate by

$$\alpha_k = 2^{-s}\alpha_{k-1} \quad \text{and} \quad \lambda^k = \mu^k - \alpha_k\nabla\theta(\mu^k), \tag{20}$$

and update

$$\beta_{k+1} = \left(\frac{1}{2} + \frac{1}{2}\sqrt{4\beta_k^2 + 1}\right) \tag{21}$$

and

$$\mu^{k+1} = \lambda^k + \frac{\beta_k - 1}{\beta_{k+1}}(\lambda^k - \lambda^{k-1}). \tag{22}$$

When a stopping rule applies, then the point λ^k is the output of the method.

In the reported experiments, we used the starting points x^0 in the PSM Algorithm and $\lambda^{-1} = \mu^0$ in (N1) above to be zero vectors. In the initialization step of the PSM Algorithm, we selected $K = 10$ and $M = 5000$. In (N1), we chose $\alpha_{-1} = 10$.

5.2.2 The Superiorization Method

Our selected choice for the operator $\boldsymbol{A}_C$ in the Basic Algorithm as well as in the Superiorized Version of the Basic Algorithm, as described in Sect. 4, is based on

an algebraic reconstruction technique (ART), see [37, Chap. 11]. Specifically, for $i = 1, 2, \ldots, I$, we define the operators $U_i : \mathbb{R}^J \to \mathbb{R}^J$ by

$$U_i(x) = x + \frac{b_i - \langle a^i, x\rangle}{\|a^i\|^2} a^i. \tag{23}$$

Defining the projection operator onto the unit box Ω by $Q : \mathbb{R}^J \to \Omega$

$$\big(Q(x)\big)_j = \begin{cases} x_j & \text{if } 0 \le x_j \le 1, \\ 0 & \text{if } x_j < 0, \\ 1 & \text{if } 1 < x_j, \end{cases} \tag{24}$$

for $j = 1, 2, \ldots, J$, we specify the algorithmic operator $\boldsymbol{A}_C : \Omega \to \Omega$ by

$$\boldsymbol{A}_C(x) = QU_I \cdots U_2 U_1(x). \tag{25}$$

Since the individual U_is as well as the Q are clearly nonexpansive operators, the same is true for $\boldsymbol{A}_C$.

By well-known properties of ART (see, for example, Sects. 11.2 and 15.8 of [37]), the Basic Algorithm with this algorithmic operator is convergent over Ω, and, in fact, for every $x^0 \in \Omega$, the limit $y(x^0)$ is in C. It follows that, for every $x^0 \in \Omega$, $\mathrm{Prox}_C(y(x^0)) = 0$, and so the Basic Algorithm is boundedly convergent. According to Theorem 4.1, this, combined with the facts that $\boldsymbol{A}_C$ is nonexpansive and the proximity function Prox_C is uniformly continuous, implies that the Basic Algorithm defined by $\boldsymbol{A}_C$ is strongly perturbation resilient.

The following uses the convergence of the Basic Algorithm to an element of C and Theorem 2. Since for all $\varepsilon > 0$, the ε-output $O(C, \varepsilon, \{x^k\}_{k=0}^{\infty})$ of the Basic Algorithm is defined for every $x^0 \in \Omega$, we also have that every sequence $\{y^k\}_{k=0}^{\infty}$ generated by the Superiorized Version of the Basic Algorithm has an ε'-output $O(C, \varepsilon', \{y^k\}_{k=0}^{\infty})$ for every $\varepsilon' > 0$. This means that for the specific type of C that is used in our comparative study, the Superiorized Version of the Basic Algorithm is guaranteed to produce an ε'-compatible output for any $\varepsilon' > 0$ and any initial point $y^0 \in \Omega$.

The specific choices made when running the Superiorized Version of the Basic Algorithm for our comparative study were the following. We selected $\eta_\ell = 0.999^\ell$, y^0 to be the zero vector, and $N = 9$. All these choices we made are based on auxiliary experiments (not included in this paper) that helped determine optimal parameters for the data-set discussed in Sect. 5.3. In addition, we need to specify how the nonascending vector $v^{k,n}$ is selected in line 8 of the Superiorized Version of the Basic Algorithm. We use the method specified in [24] (especially Sect. II.D, the paragraph following Eq. (12) and Theorem 2 in the Appendix). Specifically, we define another vector w and set $v^{k,n}$ to be the zero vector if $\|w\| = 0$ and $-\frac{w}{\|w\|}$ otherwise. The components of w are computed by $w_j = \frac{\partial \phi}{\partial x_j}(y^{k,n})$ if the partial derivative can be calculated without a numerical difficulty and $w_j = 0$ otherwise, for $1 \le j \le J$. Looking at (12), we see that formally the partial derivative $w_j = \frac{\partial \phi}{\partial x_j}(y^{k,n})$ is the sum of at most three fractions; the phrase "numerical difficulty" in the previous sentence refers to the situation where in one of these fractions the denominator has an absolute value less than 10^{-20}.

5.3 The Computational Result

The computational work reported here was done on a single machine using a single CPU, an Intel i5-3570K 3.4 GHz with 16 GB RAM using the SNARK09 software package [41, 42]; the phantom, the data, the reconstructions and displays were all generated within this same framework. In particular, this implies that differences in the reported reconstruction times are not due to the different algorithms being implemented in different environments.

Figure 1 shows the phantom used in our study, which is a 485×485 digitized image whose TV is 984. The phantom corresponds to a cross-section of a human head (based on [37, Fig. 4.6]). It is represented by a vector with $235,225$ components, each standing for the average x-ray attenuation coefficient within a pixel. Each pixel is of size 0.376×0.376 mm^2. The values of the components are in the range of $[0, 0.6241749]$, however, the display range used here was much smaller, namely $[0.204, 0.21675]$. The mapping between the two ranges is such that any value below 0.204 is shown as black and any value above 0.21675 is shown as white with a linear mapping in-between. We used this display window for all images presented here.

Data were collected by calculating line integrals through the digitized head phantom in Fig. 1 using 60 sets of equally rotated (in 3 degrees increments) parallel lines, with lines in each set spaced at 0.752 mm from each other. Each line integral gives rise to a linear equation and represents a hyperplane in $\mathbb{R}^J$. The phantom itself lies in the intersection of all the hyperplanes that are associated with these lines, and it also satisfies the box constraints in (10). The total number of linear equations is $18,524$,

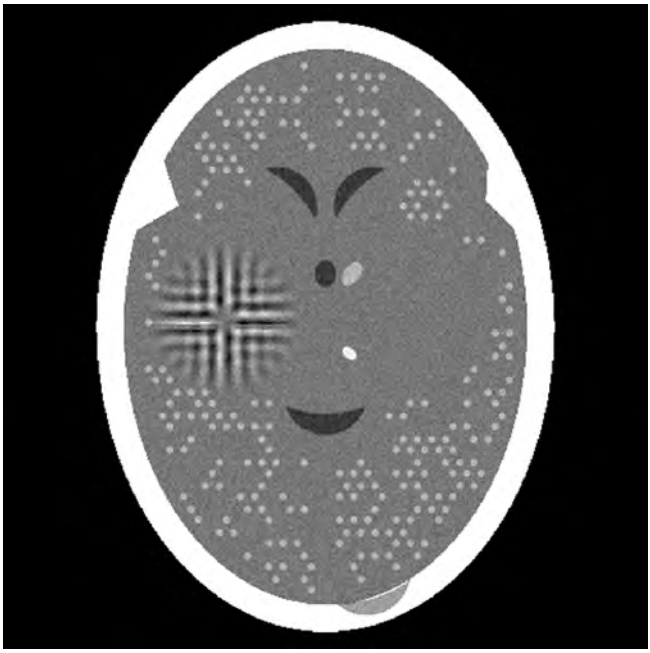
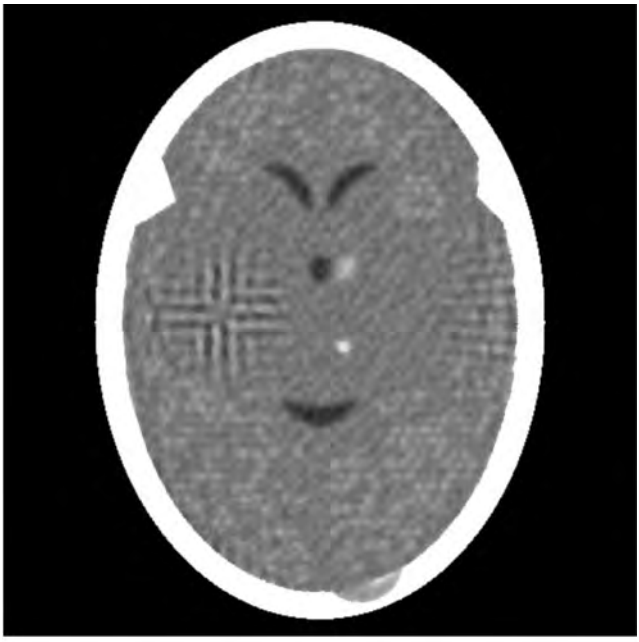


Fig. 1 The head phantom. The value of its TV is 984. Its tomographic data was obtained for 60 views



(a)

(b)

Fig. 2 Reconstructions of the head phantom of Fig. 1. (**a**) The image reconstructed by the PSM has TV = 919 and was obtained after 2217 seconds. **(b)** The image reconstructed by the SM has TV = 873 and was obtained after 102 seconds

Table 1 Performance comparison of the PSM and the SM when producing the reconstructions in Fig. 2

	TV value	Time (seconds)
PSM	919	2217
SM	873	102

making our problem underdetermined with 235,225 unknowns (the intersection of all the hyperplanes is in an at least 216,701-dimensional subspace of $R^{235,225}$). In the comparative study, we first applied the PSM and then the SM to these data as follows.

The PSM was implemented as described in Sect. 5.2.1. In particular, it started with the zero vector, for which $\text{Prox}_C(x^0) = 326$. It was stopped according to the Stopping Rule (P3), the iteration number at that time was 815, and the value of the proximity function was $\text{Prox}_C(x^{815}) = 0.0422$, which is very much smaller than the value at the initial point. The computer time required was 2217 seconds. The TV of the output was 919, which is less than that of the phantom, indicating that the PSM is performing its task of producing a constraints-compatible output with a low TV. This output is shown in Fig. 2(a).

We used the Superiorized Version of the Basic Algorithm, as described in Sect. 5.2.2 to generate a sequence $\{y^k\}_{k=0}^{\infty}$ until it reached $O(C, 0.0422, \{y^k\}_{k=0}^{\infty})$ and considered that to be the output of the SM. We know that this output must exist for our problem and that its constraints-compatibility will not be greater than that of the output of the PSM. The computer time required to obtain this output was 102 seconds, which is over twenty times shorter than what was needed by the PSM to get its output. The TV of the SM output was 876, which is also less than that of the output of PSM. The SM output is shown in Fig. 2(b).

As summarized in Table 1, with the stopping rule that guarantees that the output of the SM is at least as constraints-compatible as the output of the PSM, the SM showed superior efficacy compared to the PSM: it obtained a result with a lower TV value at less than one twentieth of the computational cost.

6 Conclusions

The superiorization methodology (SM) allows the conversion of a feasibility-seeking algorithm, designed to find an ε-compatible solution of the constraints, into a superiorized algorithm that inserts, into the feasibility-seeking algorithm, objective function reduction steps while preserving the guaranteed feasibility-seeking nature of the algorithm. The superiorized algorithm interlaces objective function nonascent steps into the original process in an automatic manner. In case of strong perturbation resilience of the original feasibility-seeking algorithm, mathematical results indicate why the superiorized algorithm will be efficacious for producing an ε-compatible solution output with a low value of the objective function.

We have presented an example for which the SM finds a better solution to a constrained minimization problem than the projected subgradient method (PSM), and in significantly less computation time. This finding is understandable in view of the nature of how the methods interlace feasibility-oriented activities with optimization

activities. While the PSM requires a projection onto the feasible region of the constrained minimization problem, the SM needs to do only projections onto the individual constraints whose intersection is the feasible region. We demonstrated this experimentally on a large-sized application that is modeled, and needs to be solved, as a constrained minimization problem.

Acknowledgements We thank the editor and reviewer for their constructive comments. We would like to acknowledge the generous support by Dr. Ernesto Gomez and Dr. Keith Schubert in allowing us to use the GPU cluster at the Department of Computer Science and Engineering at California State University San Bernardino. We are also grateful to Joanna Klukowska for her advice on using optimized compilation for speeding up SNARK09. This work was supported by the United States–Israel Binational Science Foundation (BSF) Grant No. 200912, the U.S. Department of Defense Prostate Cancer Research Program Award No. W81XWH-12-1-0122, the National Science Foundation Award No. DMS-1114901, the U.S. Department of Army Award No. W81XWH-10-1-0170, and by Grant No. R01EB013118 from the National Institute of Biomedical Imaging and Bioengineering and the National Science Foundation. The contents of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Biomedical Imaging and Bioengineering or the National Institutes of Health.

References

1. Ruszczyński, A.: Nonlinear Optimization. Princeton University Press, Princeton (2006)
2. Nesterov, Y.: Introductory Lectures on Convex Optimization. Kluwer Academic, Doredrecht (2004)
3. Shor, N.Z.: Minimization Methods for Non-Differentiable Functions. Springer, Berlin (1985)
4. Poljak, B.T.: A general method of solving extremum problems. Sov. Math. Dokl. **8**, 593–597 (1967)
5. Polyak, B.T.: Minimization of unsmooth functionals. USSR Comput. Math. Math. Phys. **9**, 14–29 (1969)
6. Beck, A., Teboulle, M.: Mirror descent and nonlinear projected subgradient methods for convex optimization. Oper. Res. Lett. **31**, 167–175 (2003)
7. Butnariu, D., Davidi, R., Herman, G.T., Kazantsev, I.G.: Stable convergence behavior under summable perturbations of a class of projection methods for convex feasibility and optimization problems. IEEE J. Sel. Top. Signal Process. **1**, 540–547 (2007)
8. Censor, Y., Elfving, T., Herman, G.T.: Averaging strings of sequential iterations for convex feasibility problems. In: Butnariu, D., Censor, Y., Reich, S. (eds.) Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications, pp. 101–114. Elsevier, Amsterdam (2001)
9. Censor, Y., Segal, A.: On the string averaging method for sparse common fixed point problems. Int. Trans. Oper. Res. **16**, 481–494 (2009)
10. Censor, Y., Segal, A.: On string-averaging for sparse problems and on the split common fixed point problem. Contemp. Math. **513**, 125–142 (2010)
11. Censor, Y., Tom, E.: Convergence of string-averaging projection schemes for inconsistent convex feasibility problems. Optim. Methods Softw. **18**, 543–554 (2003)
12. Penfold, S.N., Schulte, R.W., Censor, Y., Bashkirov, V., McAllister, S., Schubert, K.E., Rosenfeld, A.B.: Block-iterative and string-averaging projection algorithms in proton computed tomography image reconstruction. In: Censor, Y., Jiang, M., Wang, G. (eds.) Biomedical Mathematics: Promising Directions in Imaging, Therapy Planning and Inverse Problems, pp. 347–367. Medical Physics, Madison (2010)
13. Censor, Y., Davidi, R., Herman, G.T.: Perturbation resilience and superiorization of iterative algorithms. Inverse Probl. **26**, 065008 (2010)
14. Davidi, R., Herman, G.T., Censor, Y.: Perturbation-resilient block-iterative projection methods with application to image reconstruction from projections. Int. Trans. Oper. Res. **16**, 505–524 (2009)
15. Herman, G.T., Davidi, R.: Image reconstruction from a small number of projections. Inverse Probl. **24**, 045011 (2008)
16. Nikazad, T., Davidi, R., Herman, G.: Accelerated perturbation-resilient block-iterative projection methods with application to image reconstruction. Inverse Probl. **28**, 035005 (2012)

17. Penfold, S.N., Schulte, R.W., Censor, Y., Rosenfeld, A.B.: Total variation superiorization schemes in proton computed tomography image reconstruction. Med. Phys. **37**, 5887–5895 (2010)
18. Aharoni, R., Censor, Y.: Block-iterative projection methods for parallel computation of solutions to convex feasibility problems. Linear Algebra Appl. **120**, 165–175 (1989)
19. Bauschke, H.H., Borwein, J.M.: On projection algorithms for solving convex feasibility problems. SIAM Rev. **38**, 367–426 (1996)
20. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, New York (2011)
21. Censor, Y., Chen, W., Combettes, P.L., Davidi, R., Herman, G.T.: On the effectiveness of projection methods for convex feasibility problems with linear inequality constraints. Comput. Optim. Appl. **51**, 1065–1088 (2012)
22. Censor, Y., Zenios, S.A.: Parallel Optimization: Theory, Algorithms, and Applications. Oxford University Press, New York (1997)
23. Cegielski, A.: Iterative Methods for Fixed Point Problems in Hilbert Spaces. Lecture Notes in Mathematics, vol. 2057. Springer, Berlin (2012)
24. Herman, G.T., Garduño, E., Davidi, R., Censor, Y.: Superiorization: an optimization heuristic for medical physics. Med. Phys. **39**, 5532–5546 (2012)
25. Jin, W., Censor, Y., Jiang, M.: A heuristic superiorization-like approach to bioluminescence tomography. In: International Federation for Medical and Biological Engineering (IFMBE) Proceedings, vol. 39, pp. 1026–1029 (2012)
26. Luo, S., Zhou, T.: Superiorization of EM algorithm and its application in single-photon emission computed tomography (SPECT). Inverse Problems and Imaging. Accepted for publication
27. Bauschke, H.H., Koch, V.R.: Projection methods: swiss army knives for solving feasibility and best approximation problems with halfspaces. In: Reich, S., Zaslavski, A. (eds.) Proceedings of the Workshop "Infinite Products of Operators and Their Applications", Haifa, 2012 (2013). Accepted for publication https://people.ok.ubc.ca/bauschke/Research/c16.pdf
28. Helou Neto, E.S., De Pierro, Á.R.: Incremental subgradients for constrained convex optimization: a unified framework and new methods. SIAM J. Optim. **20**, 1547–1572 (2009)
29. Helou Neto, E.S., De Pierro, Á.R.: On perturbed steepest descent methods with inexact line search for bilevel convex optimization. Optimization **60**, 991–1008 (2011)
30. Nedić, A.: Random algorithms for convex minimization problems. Math. Program., Ser. B **129**, 225–253 (2011)
31. Ram, S.S., Nedić, A., Veeravalli, V.: Incremental stochastic subgradient algorithms for convex optimization. SIAM J. Optim. **20**, 691–717 (2009)
32. Nurminski, E.A.: The use of additional diminishing disturbances in Fejer models of iterative algorithms. Comput. Math. Math. Phys. **48**, 2154–2161 (2008). Original Russian text: Nurminski, E.A.: Ž. Vyčisl. Mat. Mat. Fiz. **48**, 2121–2128 (2008)
33. Nurminski, E.A.: Fejer processes with diminishing disturbances. Dokl. Math. **78**, 755–758 (2008). Original Russian text: Nurminski, E.A.: Dokl. Akad. Nauk SSSR **422**, 601–604 (2008)
34. Nurminski, E.A.: Envelope stepsize control for iterative algorithms based on Fejer processes with attractants. Optim. Methods Softw. **25**, 97–108 (2010)
35. Nurminski, E.A.: Fejer algorithms with an adaptive step. Comput. Math. Math. Phys. **51**, 741–750 (2011). Original Russian text: Nurminski, E.A.: Ž. Vyčisl. Mat. Mat. Fiz. **51**, 791–801 (2011)
36. Sidky, E.Y., Kao, C., Pan, X.: Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. Phys. Med. Biol. **53**, 4777–4807 (2008)
37. Herman, G.T.: Fundamentals of Computerized Tomography: Image Reconstruction from Projections, 2nd edn. Springer, Berlin (2009)
38. Combettes, P.L., Pesquet, J.C.: Image restoration subject to a total variation constraint. IEEE Trans. Image Process. **13**, 1213–1222 (2004)
39. Nesterov, Y.E.: A method for solving the convex programming problem with convergence rate $O(1/k^2)$. Sov. Math. Dokl. **27**, 372–376 (1983)
40. Güler, O.: Complexity of smooth convex programming and its applications. In: Pardalos, P.M. (ed.) Complexity of Numerical Optimization, pp. 180–202. World Scientific, Singapore (1993)
41. Davidi, R., Herman, G.T., Klukowska, J.: SNARK09: A programming system for the reconstruction of 2D images from 1D projections (2009). http://www.dig.cs.gc.cuny.edu/software/snark09/
42. Klukowska, J., Davidi, R., Herman, G.T.: SNARK09—A software package for reconstruction of 2D images from 1D projections. Comput. Methods Programs Biomed. **110**, 424–440 (2013)

Contemporary Mathematics

Feasibility-Seeking and Superiorization Algorithms Applied to Inverse Treatment Planning in Radiation Therapy

Ran Davidi, Yair Censor, Reinhard W. Schulte, Sarah Geneser, and Lei Xing

Abstract. We apply the recently proposed superiorization methodology (SM) to the inverse planning problem in radiation therapy. The inverse planning problem is represented here as a constrained minimization problem of the total variation (TV) of the intensity vector over a large system of linear two-sided inequalities. The SM can be viewed conceptually as lying between feasibility-seeking for the constraints and full-fledged constrained minimization of the objective function subject to these constraints. It is based on the discovery that many feasibility-seeking algorithms (of the projection methods variety) are perturbation-resilient, and can be proactively steered toward a feasible solution of the constraints with a reduced, thus superiorized, but not necessarily minimal, objective function value.

December 3, 2013

1. Introduction

Computationally demanding numerical minimization techniques are often used in optimizing the treatment plan of different types of intensity modulated radiation therapy (IMRT), for example, in volumetric-modulated arc therapy (VMAT). However, some commonly employed objective functions and corresponding minimization techniques are not necessarily the most appropriate for achieving the desired radiation dose distribution behavior in the patient. This disconnect occurs because minimal solutions to current minimization formulations are not guaranteed to provide the desired dose coverage, conformality, or homogeneity. Therefore, the considerable computational cost associated with some of these minimization techniques may not be justified.

We propose to apply the recently developed novel superiorization method (SM) that improves computational tractability by aiming at a solution that is guaranteed to satisfy the IMRT planning constraints and results in a reduced, but not necessarily minimal, value of the objective function.

The first author was supported by the U.S. Department of Defense Prostate Cancer Research Program Award No. W81XWH-12-1-0122.

The SM can be viewed conceptually as lying between feasibility-seeking for the constraints and full-fledged constrained minimization of the objective function subject to these constraints. It is based on the discovery that many feasibility-seeking algorithms (of the projection methods variety) are perturbation-resilient, and can be proactively steered toward a feasible solution of the constraints with a reduced, but not necessarily minimal, objective function value.

The SM is, thus, capable of producing "superior feasible solutions" by employing less-demanding feasibility-seeking projection methods. Therefore, it may replace current computationally demanding constrained minimization methods, and potentially lead to shorter computational times and improved dose distributions.

The paper is laid out as follows. In Section 2 we briefly acquaint the reader with the inverse problem of radiation therapy treatment planning and the mathematical model that we use. In Section 3 a short review of the SM is given, and, in Section 4, we present an illustrative example how SM can be applied to planning a prostate cancer IMRT case. Finally, in Section 5 we provide our conclusions.

2. The inverse problem of radiation therapy treatment planning

Inverse planning is at the heart of intensity modulated treatment procedures and critically determines the quality of the resulting treatment plan. Usually, the radiation oncologist in charge defines the boundaries of the clinical and gross tumor volumes and organs at risk (OAR) for radiation late effects and prescribes the minimum and maximum target doses, threshold doses and/or volumes not to be exceeded in OAR and gives importance factors for each. These constraints give rise to a mathematical model that requires the solution of an inverse problem. A solution method is run to find a treatment plan consisting of intensities and timing of different beam segments which best matches all the input criteria.

However, as practiced now, the therapeutic capacity of these applications is underutilized because of the computing performance of some of the currently used minimization methods. In this work, we suggest to use the SM to reach an acceptable treatment plan. Let us first briefly describe the inverse problem at hand; for more technical details related to different types of IMRT, the reader may consult review articles, such as, [**A, B, C**], to name but a few.

IMRT-type techniques are currently the most advanced form of external radiation therapy. Different from its predecessor, 3D conformal radiation therapy (3DCRT), the physician needs to clearly define the objective of the treatment plan by specifying dose and/or volume constraints for the planning target volume (PTV) and OAR that aims at maximum tumor cell killing and minimum harm to the patient's normal tissues. The treatment plan resulting from solving a corresponding mathematical problem defines multiple field directions and the movement of computer controlled pairs of multileaf collimator (MLC) leaves for each direction.

The pairs of MLC leaf positions dynamically change during treatment and are physically controlled plates that move during treatment and help modulate the beam to achieve the objectives of the physician-defined treatment plan. The beam, therefore, can be conceptually subdivided to a two-dimensional grid of beam subunits called beamlets. Finding a deliverable treatment plan comprised of beam apertures and weights for the multiple directions and possible locations of the MLCs is the goal of the inverse treatment planning problem. In the next paragraph we

discuss a typical model for the inverse treatment planning problem that leads to a constrained minimization problem, which in turn, fits the SM framework.

Denote the physician's prescribed dose distribution to the patient by a dose vector $d = (d_j)_{j=1}^{J} \in R^J$ where d_j is the dose in voxel j of the fully-discretized patient's cross-section. The dose distribution d is known to have a linear relationship with the intensities of the beamlets, denoted by an intensity vector $x = (x_i)_{i=1}^{I} \in R^I$, such that x_i is the intensity of the beamlet i. The inversion problem can, therefore, be formulated as a linear system of equations

$$d = Ax, \tag{2.1}$$

where A is the $J \times I$ dose matrix that maps any intensity of beamlets vector x onto a dose in voxels vector d. Here I is the total number of beamlets and J is the total number of voxels.

Further assume that there are S structures in the patient's cross-section, for $s = 1, 2, \ldots, S$, and let O_s be the set of voxel indices that belong to a structure s

$$O_s = \{j_{s,1}, j_{s,2,} \ldots j_{s,m(s)}\}, \tag{2.2}$$

where $m(s)$ is the number of voxels in the s structure. Then the system matrix A can be partitioned into blocks

$$A = \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_S \end{pmatrix}, \tag{2.3}$$

so that a submatrix A_s will contain the rows of A whose indices appear in O_s, and $d_{(s)}$ will be the subvector of d whose component indices appear in O_s, and the system (2.1) becomes

$$\begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_S \end{pmatrix} x = \begin{pmatrix} d_{(1)} \\ d_{(2)} \\ \vdots \\ d_{(S)} \end{pmatrix}. \tag{2.4}$$

Following a well-trodden path in this area, with roots in [**E**] and [**F**], we replace the system (2.1) by a more flexible model in which we ask the physician to specify lower- and upper-dose bounds vectors, $\underline{d}$ and $\overline{d}$, respectively, on all voxels in the respective structures. For a structure s that is an OAR we define

$$\overline{d_{(s)}} \equiv d_{(s)}, \tag{2.5}$$

and for any target structures s such as the PTV we define

$$\underline{d_{(s)}} \equiv d_{(s)}. \tag{2.6}$$

Hence, for an s that is an OAR we obtain

$$0 \leq A_s x \leq d_{(s)}, \tag{2.7}$$

and for a target structure s

$$d_{(s)} \leq A_s x \leq e_{(s)}, \tag{2.8}$$

where $e_{(s)}$ is an additional clinically-specified upper-bound subvector on the target. Denoting by a^t the tth row of the matrix A, the inequalities of (2.7) are, componentwise,

$$0 \leq \left\langle a^{j_{s,\ell}}, x \right\rangle \leq d_{(s)}, \text{ for all } \ell = 1, 2, \ldots, m(s), \tag{2.9}$$

where $j_{s,\ell} \in O_s$, for a structure s, and the inequalities of (2.8) are,

$$d_{(s)} \leq \left\langle a^{j_{s,\ell}}, x \right\rangle \leq e_{(s)}, \text{ for all } \ell = 1, 2, \ldots, m(s), \tag{2.10}$$

where $\langle \cdot, \cdot \rangle$ stands for the inner product.

This leads to a system of linear inequalities

$$\begin{pmatrix} \underline{d}_{(1)} \\ \underline{d}_{(2)} \\ \vdots \\ \underline{d}_{(S)} \end{pmatrix} \leq \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_S \end{pmatrix} x \leq \begin{pmatrix} \overline{d}_{(1)} \\ \overline{d}_{(2)} \\ \vdots \\ \overline{d}_{(S)} \end{pmatrix} \tag{2.11}$$

which serves as the constraints set for the inverse problem modeled as a minimization problem. For the objective function ϕ we use the *total variation* (TV) of the intensity vector x, given by

$$\phi(X) = TV(X) = \sum_{u=1}^{U-1} \sum_{u=1}^{V-1} \sqrt{(x_{u+1,v} - x_{u,v})^2 + (x_{u,v+1} - x_{u,v})^2}, \tag{2.12}$$

where the two-dimensional array is obtained from the intensity vector x by $X = \{x_{u,v}\}_{u=1,\ v=1}^{U,\ V}$ where u and v are integers (and $uv = J$). The use of TV minimization in radiation therapy treatment planning was suggested by Zhu *et al.* in [**G**] but they used there a different modeling approach that led them to a minimization problem, rather than a feasibility problem like ours in (2.11). They handled the TV minimization by using it to regularize their objective function and applied an exact constrained minimization algorithm, which resulted in a large computational burden.

Our approach leads us to the constrained minimization problem (3.1) with (2.12) as the objective and (2.11) as the constraints.

3. A short review of the SM

The superiorization methodology (SM) of [**H, I**] is intended for nonlinear constrained minimization (CM) problems of the form

$$\text{minimize} \left\{ \phi(x) \mid x \in C \right\}, \tag{3.1}$$

where $\phi : R^J \to R$ is an objective function and $C \subseteq \Theta \subseteq R^J$ is a given feasible set defined by a family of constraints $\{C_i\}_{i=1}^I$, where each set C_i is a nonempty closed convex subset of R^J, so that $C = \cap_{i=1}^I C_i \neq \emptyset$.

In a nutshell, the new paradigm of superiorization lies between feasibility-seeking and CM. It is not quite trying to solve the full fledged CM; rather, the task is to find a feasible point that is superior (with respect to the objective function value) to one returned by a feasibility-seeking only algorithm.

The SM could be beneficial for a problem for which an exact CM algorithm has not yet been discovered, or when existing exact optimization algorithms are very time consuming or require too much computer resources for realistic large problems.

If, in such cases, there exist (space- and time-) efficient iterative feasibility-seeking projection methods that provide non-optimal but constraints-compatible solutions, then they can be turned by the SM into methods that will be practically useful from the point of view of the function to be optimized. Examples of such situations are given in [**H, I**].

We associate with the feasible set C a proximity function $Prox_C : \Theta \to R_+$, which is an indicator of how incompatible a vector $x \in \Theta$ is with the constraints. For any given $\varepsilon > 0$, a point $x \in \Theta$ for which $Prox_C(x) \leq \varepsilon$ is called an *ε-compatible solution* for C. We assume that we have a feasibility-seeking *algorithmic operator* $\boldsymbol{A}_C : R^J \to \Theta$, that defines a Basic Algorithm whose iterative step, given the current iterate vector x^k, calculates the next iterate x^{k+1} by

$$x^{k+1} = \boldsymbol{A}_C\left(x^k\right). \tag{3.2}$$

Given $C \subseteq R^J$, a proximity function $Prox_C$, a sequence $\left\{x^k\right\}_{k=0}^{\infty} \subset \Theta$ and an $\varepsilon > 0$, then an element x^K of the sequence which has the properties: (i) $Prox_C\left(x^K\right) \leq \varepsilon$, and (ii) $Prox_C\left(x^k\right) > \varepsilon$ for all $0 \leq k < K$, is called an *ε-output of the sequence* $\left\{x^k\right\}_{k=0}^{\infty}$ *with respect to the pair* $(C,\ Prox_C)$. We denote it by $O\left(C, \varepsilon, \left\{x^k\right\}_{k=0}^{\infty}\right) = x^K$, O standing for output.

Clearly, an ε-output $O\left(C, \varepsilon, \left\{x^k\right\}_{k=0}^{\infty}\right)$ of a sequence $\left\{x^k\right\}_{k=0}^{\infty}$ might or might not exist, but if it does, then it is unique. If $\left\{x^k\right\}_{k=0}^{\infty}$ is produced by an algorithm intended for the feasible set C, such as the Basic Algorithm (3.2), without a termination criterion, then $O\left(C, \varepsilon, \left\{x^k\right\}_{k=0}^{\infty}\right)$ is the *output* produced by that algorithm when it includes the termination rule to stop when an ε-compatible solution for C is reached.

In order to "superiorize" such an algorithm we need it to be *strong perturbation resilience* in the sense that for every $\varepsilon > 0$, for which an ε-output is defined for a sequence generated by the Basic Algorithm, for every $x^0 \in \Theta$, we have also that the ε'-output is defined for every $\varepsilon' > \varepsilon$ and for every sequence $\left\{y^k\right\}_{k=0}^{\infty}$ generated by $y^{k+1} = \boldsymbol{A}_C\left(y^k + \beta_k v^k\right)$, for all $k \geq 0$, where the vector sequence $\left\{v^k\right\}_{k=0}^{\infty}$ is bounded and the scalars $\{\beta_k\}_{k=0}^{\infty}$ are such that $\beta_k \geq 0$, for all $k \geq 0$, and $\sum_{k=0}^{\infty} \beta_k < \infty$. See our recent [**H**] for details.

Along with the constraints set $C \subseteq R^J$, we look at the objective function $\phi : R^J \to R$, with the convention that a point in R^J for which the value of ϕ is smaller is considered *superior* to a point in R^J for which the value of ϕ is larger.

The essential idea of the SM is to make use of the perturbations in order to transform a strongly perturbation resilient algorithm that seeks a constraints-compatible solution for C (i.e., is seeking feasibility) into one whose outputs are equally good from the point of view of constraints-compatibility, but are superior (not necessarily optimal) according to the objective function ϕ.

This is done by producing from the Basic Algorithm another algorithm, called its *superiorized* version, that makes sure not only that the $\beta_k v^k$ are bounded perturbations, but also that $\phi\left(y^k + \beta_k v^k\right) \leq \phi\left(y^k\right)$, for $k \geq L$ for some integer $L \geq 0$. The Superiorized Version of the Basic Algorithm assumes that we have available a summable sequence $\{\eta_\ell\}_{\ell=0}^{\infty}$ of positive real numbers (for example, $\eta_\ell = a^\ell$, where $0 < a < 1$) and it generates, simultaneously with the sequence $\left\{y^k\right\}_{k=0}^{\infty}$ in Θ, sequences $\left\{v^k\right\}_{k=0}^{\infty}$ and $\{\beta_k\}_{k=0}^{\infty}$. The latter is generated as a subsequence of $\{\eta_\ell\}_{\ell=0}^{\infty}$,

resulting in a nonnegative summable sequence $\{\beta_k\}_{k=0}^{\infty}$. The algorithm further depends on a specified initial point $y^0 \in \Theta$ and on a positive integer N. It makes use of a logical variable called *loop*. The superiorized algorithm is presented next by its pseudo-code.

The Superiorized Version of the Basic Algorithm
set $k = 0$
set $y^k = y^0$
set $\ell = -1$
repeat
  set $n = 0$
  set $y^{k,n} = y^k$
  while $n<N$
    set $v^{k,n}$ to be a nonascending vector for ϕ at $y^{k,n}$
    set *loop=true*
    while *loop*
      set $\ell = \ell + 1$
      set $\beta_{k,n} = \eta_\ell$
      set $z = y^{k,n} + \beta_{k,n} v^{k,n}$
      if $\phi(z) \leq \phi\left(y^k\right)$ **then**
        set $n=n+1$
        set $y^{k,n}=z$
        set *loop = false*
  set $y^{k+1}=\boldsymbol{A}_C\left(y^{k,N}\right)$
  set $k = k + 1$

Analysis of the Superiorized Version of the Basic Algorithm [**H, I**], shows that it produces outputs that are essentially as constraints-compatible as those produced by the original (not superiorized) Basic Algorithm. However, due to the repeated steering of the process toward reducing the value of the objective function ϕ, we can expect that the output of the Superiorized Version will be superior (from the point of view of ϕ) to the output of the original algorithm. A recent work that includes results about the SM appears in this volume [**D**].

4. Demonstrative examples

The anonymized pelvic planning CT (computed tomography) of a prostate cancer patient was employed for the IMRT treatment planning using the proposed method. Seven equispaced fields were used for targeting the PTV. The dose constraints were set using the RTOG 0815 randomized trial protocol [**J**].

Our preliminary testing of the approach was done by comparing the outputs of a TV-superiorization algorithm with an, otherwise identical, algorithm that aimed at only satisfying the dose constraints, without applying the SM. Here $\boldsymbol{A}_C$ was chosen to be ART for inequalities [**K**]. It was proven to be perturbation resilient in [**L**].

From a radiation delivery stand point, a solution that is easy to deliver is one that has a piecewise constant intensety-beamlet map. The reason has to do with the physical constraints coming from the MLCs, they require that the beamlets have a small number of signal levels. It was, therefore, suggested in the literature to use total-variation (TV) to force the solution to be piecewise constant [**M, N**].

We performed two experiments with different starting conditions. For the first experiment, we initiated the algorithm with the zero vector of dose weights and for the second experiment all dose weights were given the value 10. Tables 1 and 2 summarize the results for the two experiments and in Figure 1 we present the associated DVH (dose-volume histogram) curves.

For the first experiment, the TV-superiorization algorithm produced a solution that met the acceptance criteria after 12 iterations whereas the feasibility-seeking algorithm was not able to reach an acceptable solution after this number of iterations. For the second experiment, the TV-superiorization algorithm reached an acceptable solution even faster, i.e., after 7 iterations, and the feasibility-seeking algorithm again failed some of the acceptance criteria after this number of iterations.

TABLE 1. RTOG 0815 acceptance criteria and results of experiment 1 described in Section 4 (TVS stands for TV-superiorization)

Acceptance criteria	Exp 1 with TVS	Exp 1 without TVS
PTV: Min Allowed Dose: 75.24 Gy	75.24 Gy	56.13 Gy
PTV: Max Allowed Dose: 84.74 Gy	84.69 Gy	89.42 Gy
Rectum: No more than 50% of the volume should exceed 60.00 Gy	34.50 %	8.50 %
Rectum: Max Dose	82.64 Gy	82.71 Gy

TABLE 2. RTOG 0815 acceptance criteria and results of experiment 2 described in Section 4 (TVS stands for TV-superiorization)

Acceptance criteria	Exp 2 with TVS	Exp 2 without TVS
PTV: Min Allowed Dose: 75.24 Gy	77.80 Gy	76.15 Gy
PTV: Max Allowed Dose: 84.74 Gy	84.71 Gy	87.63 Gy
Rectum: No more than 50% of the volume should exceed 60.00 Gy	36.90 %	40.50 %
Rectum: Max Dose	84.09 Gy	87.25 Gy

5. Conclusions

Our proposed method successfully produced conformal solutions that met the acceptance criteria while that an otherwise identical algorithm without superiorization failed to do so with the same number of iterations. Future work will assess the computational gain of the superiorization method compared to a conventional one and investigate the utility of it for a computationally more complex problems that can be found in modulated techniques for arc therapy.

FIGURE 1. Dose volume histograms (DVH) of the two experiments. Solid lines represent the algorithm with TV-superiorization (broken lines represent no superiorization). The first (top) took 12 iterations and the second (bottom) took 7 iterations. Exact numbers are given in Tables 1 and 2.

Acknowledgements. This work is supported by the U.S. Department of Defense Prostate Cancer Research Program Award No. W81XWH-12-1-0122, by grant number 2009012 from the United States–Israel Binational Science Foundation (BSF) and by the U.S. Department of the Army Award No. W81XWH-10-1-0170. Some of the material was presented as a poster at the Technology for Innovation in Radiation Oncology, Joint Workshop of the American Society for Radiation Oncology (ASTRO), the National Cancer Institute (NCI) and the American Association of Physicists in Medicine (AAPM), National Institutes of Health (NIH), Bethesda, MD, USA, June 13–14, 2013.

References

[A] K. Otto, *Volumetric modulated arc therapy: IMRT in a single gantry arc*, Medical Physics **35** (2008), 310–317.

[B] S. Webb, *The physical basis of IMRT and inverse planning*, British Journal of Radiology **76** (2003), 678–689.

[C] T. Bortfeld, *IMRT: a review and preview*, Physics in Medicine and Biology **51** (2006), R363–R379.

[D] H.H. Bauschke and V.R. Koch, *Projection methods: Swiss army knives for solving feasibility and best approximation problems with halfspaces*, The Conference on Infinite Products and Their Applications (Technion, Haifa, Israel, 2012), Contemporary Mathematics, accepted for publication. http://arxiv.org/abs/1301.4506v1.

[E] M.D. Altschuler and Y. Censor, *Feasibility solutions in radiation therapy treatment planning*, in: Proceedings of the Eighth International Conference on the Use of Computers in Radiation Therapy, IEEE Computer Society Press, Silver Spring, MD, USA (1984), 220–224.

[F] Y. Censor, M.D. Altschuler and W.D. Powlis, *A computational solution of the inverse problem in radiation-therapy treatment planning*, Applied Mathematics and Computation **25** (1988), 57–87.

[G] L. Zhu, L. Lee, Y. Ma, Y. Ye, R. Mazzeo and L. Xing, *Using total-variation regularization for intensity modulated radiation therapy inverse planning with field-specific numbers of segments*, Physics in Medicine and Biology **53** (2008), 6653–6672.

[H] Y. Censor, R. Davidi, G.T. Herman, R.W. Schulte and L. Tetruashvili, *Projected subgradient minimization versus superiorization*, Journal of Optimization Theory and Applications, accepted for publication. DOI:10.1007/s10957-013-0408-3.

[I] G.T. Herman, E. Garduño, R. Davidi and Y. Censor *Superiorization: An optimization heuristic for medical physics*, Medical Physics **39** (2012), 5532–5546.

[J] Radiation Therapy Oncology Group: *RTOG 0815 Protocol Information*. http://www.rtog.org/ClinicalTrials/ProtocolTable/StudyDetails.aspx?study=0815. Updated: 11/24/2013.

[K] G.T. Herman and A. Lent *A family of iterative quadratic optimization algorithms for pairs of inequalities with application in diagnostic radiology*, Mathematical Programming Studies **9** (1978), 15–29.

[L] D. Butnariu, R. Davidi, G.T. Herman and I. G. Kazantsev *Stable Convergence Behavior Under Summable Perturbations of a Class of Projection Methods for Convex Feasibility and Optimization Problems*, IEEE Journal of Selected Topics In Signal Processing **1** (2007), 540–547.

[M] K.T. Block , M. Uecker and J. Frahm *Undersampled radial MRI with multiple coils. Iterative image reconstruction using a total variation constraint*, Magnetic Resonance in Medicine **57** (2007), 1086–1098.

[N] V. Kolehmainen, A. Vanne, S. Siltanen, S. Jrvenp, J.P. Kaipio, M. Lassas and M. Kalke *Parallelized Bayesian inversion for three-dimensional dental x-ray imaging*, IEEE Transactions on Medical Imaging **25** (2006), 218–228.

Department of Radiation Oncology, Stanford University School of Medicine, Stanford, CA 94305, USA

E-mail address: rdavidi@stanford.edu

Department of Mathematics, University of Haifa, Mount Carmel, Haifa 3190501, Israel

E-mail address: yair@math.haifa.ac.il

Department of Radiation Medicine, Loma Linda University Medical Center, Loma Linda, CA 92354, USA

E-mail address: rschulte@llu.edu

Department of Radiation Oncology, University of California San Francisco, San Francisco, CA 94143, USA

E-mail address: genesers@radonc.ucsf.edu

Department of Radiation Oncology, Stanford University School of Medicine, Stanford, CA 94305, USA

E-mail address: lei@stanford.edu

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

SNARK09 – A software package for reconstruction of 2D images from 1D projections

Joanna Klukowska[a], *Ran Davidi*[b], *Gabor T. Herman*[a,*]

[a] *Department of Computer Science, Graduate Center, City University of New York, New York, NY 10016, USA*
[b] *Department of Radiation Oncology, Stanford University School of Medicine, Stanford, CA 94305, USA*

ARTICLE INFO

Article history:
Received 15 August 2012
Received in revised form
3 December 2012
Accepted 7 January 2013

Keywords:
SNARK09
Reconstruction from projections
Software package
Evaluation
Simulation
Computerized tomography

ABSTRACT

The problem of reconstruction of slices and volumes from 1D and 2D projections has arisen in a large number of scientific fields (including computerized tomography, electron microscopy, X-ray microscopy, radiology, radio astronomy and holography). Many different methods (algorithms) have been suggested for its solution.

In this paper we present a software package, SNARK09, for reconstruction of 2D images from their 1D projections. In the area of image reconstruction, researchers often desire to compare two or more reconstruction techniques and assess their relative merits. SNARK09 provides a uniform framework to implement algorithms and evaluate their performance. It has been designed to treat both parallel and divergent projection geometries and can either create test data (with or without noise) for use by reconstruction algorithms or use data collected by another software or a physical device. A number of frequently-used classical reconstruction algorithms are incorporated. The package provides a means for easy incorporation of new algorithms for their testing, comparison and evaluation. It comes with tools for statistical analysis of the results and ten worked examples.

1. Introduction

The need for reconstruction from projections occurs in many biomedical areas. Practical applications, such as computerized tomography (CT), positron emission tomography (PET) and X-ray microscopy, use physically collected projection data to reconstruct real objects. Simulation packages (SNARK09 [1,2] is an example) allow for thorough testing of effects of various factors on the projection data and on the outputs of reconstruction algorithms. For example, in SNARK09, various sources of noise that occurs during X-ray data collection can be simulated separately so that their effects can be studied and understood. (The name SNARK09 originates from the Lewis Carroll nonsense poem "The Hunting of the Snark.")

SNARK09 provides a total framework for reconstruction from projections for both simulated and real data, as well as statistical evaluation of the results. Mathematical phantoms can be generated either as piecewise constant objects, appropriate for materials science, or as objects containing inhomogeneities to better simulate biological materials. Projection datasets can be obtained based on mathematically described phantoms. The user has options of investigating various scanner modes, including noise models comparable to actual imaging devices. There is also an option of using projection data obtained from external sources (a medical scanner or data generated by other software). The package comes with several built-in reconstruction algorithms. It provides either pixels or blobs as basis functions. Users have a means of implementing their own reconstruction algorithms.

* *Corresponding author*. Tel.: +1 212 817 8193; fax: +1 212 817 1510.
E-mail addresses: jklukowska.gc@gmail.com (J. Klukowska), rdavidi@stanford.edu (R. Davidi), gaborherman@yahoo.com (G.T. Herman).

0169-2607/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cmpb.2013.01.003

The results of the reconstructions can be evaluated using a statistically sound methodology built into the package.

1.1. Features of the package

Below we give a short summary of some features included in SNARK09. This is not intended to be a complete list, but rather a representative sample of what SNARK09 has to offer.

Polychromatic and monochromatic X-ray simulation. The package provides a means of simulating X-rays with either a monochromatic or polychromatic spectrum at energy levels specified by the user. X-rays generated by most medical imaging devices are polychromatic in nature.

Beam hardening correction. Due to the polychromatic nature of X-rays, beam hardening correction is needed to compensate for different levels of X-ray attenuation. Such a correction is available in SNARK09.

Projection computation. The projection data through the phantom can be computed either based on a mathematically defined phantom (line integrals through geometrical features), or a digitized phantom (using lengths of intersections of a ray with each pixel of the phantom).

Digital difference analyzer (DDA). Projections through digitized phantoms can be computed very fast using a DDA. This method, originally developed for drawing lines using a digital plotter [3], is used for computation of pixel intersections by a single line.

Basis functions. Both pixels and blobs [4,5] can be used as the basis functions for mathematical representation of the reconstructions. Blobs have been shown to be superior for representation of biological structures due to their smoothness.

Reconstruction algorithms. The package provides a large selection of reconstruction algorithms based on transform methods and series expansion methods with parameters selected by the user. User-defined reconstruction algorithms can also be easily implemented and used in SNARK09. The code for these algorithms has to be written in C++. The programmers have at their disposal a selection of functions and objects already implemented in SNARK09.

Routines and classes for use in user-defined algorithms. Researchers who need to implement their own reconstruction algorithms have a large range of routines and classes available for their use. One example is a function that, for a given projection angle and ray number, computes the pixels intersected by that ray. This greatly simplifies implementation of additional algorithms.

Statistical comparison of algorithms. The built-in and user-defined algorithms can be evaluated for their superiority for a given task. SNARK09 uses an ensemble of phantoms from which a particular phantom is (randomly) chosen and is reconstructed from its (randomly generated) projection data by the algorithms to be compared. Statistical evaluation is performed based on this multiplicity of reconstructions, which allow us to assign a statistical significance by which we can reject the null hypothesis that two algorithms are equally good in favor of the alternative hypothesis that one of them is better than the other.

Figure of merit (FOM). A meaningful statistical evaluation of reconstruction methods has to be done in terms of a specific task at hand. The package provides several built-in FOMs, as well as a means for the user to provide their own definitions. One of the built-in FOMs is imagewise region of interest, which has been shown to correlate well with the performance of humans for detection of small tumors in lung tissue [6].

Use of simulated or real data. The package can either simulate the data generation based on a mathematically defined phantom, or use data obtained by another simulation package or an actual device.

Graphical user interface. SNARK09 runs from the command line. There are two other programs, SNARK09Input and SNARK09Display, that provide assistance in creation of input files and visualization of sinograms and of reconstructions. SNARK09Display has also the capability of displaying profile lines of the reconstructed images as well as plotting several built-in evaluation parameters.

1.2. Related work

There are many packages available for reconstruction from projections. Some of them, like SNARK09, are designed to work with 2D images and their 1D projections; some are designed to work with 3D objects and their 2D projections; some can do both. The packages provide a varied selection of simulation capabilities, choices of reconstruction algorithms and evaluation techniques. Some are developed for the purpose of reconstruction from real data and for obtaining results used in practical studies in medicine or biology. Others, like SNARK09, provide means of evaluation and examination of different effects that occur during the imaging process and are testbeds for new and existing reconstruction methodologies. Below, we mention a few of these packages that we are most familiar with as examples; it is by no means a complete review of what is available.

jSNARK [7] incorporates a large subset of the capabilities of SNARK09 for reconstructing 2D objects from their 1D projections, but it also extends those capabilities to reconstructing 3D objects from their 2D projections. It is written in Java allowing for greater platform flexibility and for user-defined routines written as plugins.

There are many software packages designed primarily for reconstruction from transmission electron microscopy data. Their goal is to produce reconstructions that can be used by biologists, rather than to allow experimentation with new algorithms. Three examples of such packages are Xmipp [8], SPIDER [9] and IMOD [10].

MATLAB® provides some very basic tomography routines and phantoms. Many researchers write their own routines using MATLAB® environment. Two examples of such code available for free download are the Image Reconstruction Toolbox [11] and AIR Tools [12] that implement iterative algebraic reconstruction methods.

STIR [13] is open-source software for use in tomographic imaging. Its aim is to provide a multi-platform object-oriented framework for all data manipulations in tomographic imaging. Currently, the emphasis is on iterative image reconstruction in positron emission tomography.

SNARK09 offers much more than just the reconstruction routines. It provides a basic platform that can be used by researchers (with no or not much extra coding) to study,

simulate and perform data collection, reconstruction and statistical analysis.

1.3. *Outline of the rest of this paper*

In this paper we review the functionality of the SNARK09 software package. We discuss phantom creation, data collection and reconstruction methods in Section 2. The system description follows in Section 3. We go over an example of the use of the package in Section 4. Finally, mode of availability and system requirements are covered in Section 5 and future work in Section 6.

2. Computational methods and theory

The *reconstruction problem* may be stated roughly as follows: given approximations (based on physical measurements) of the real ray sums of a picture for a number of rays, estimate the $N \times N$ digitization of the picture. The SNARK09 package provides a means of simulating each step of this process. The details of this are discussed in this section.

SNARK09 deals with *pictures* defined over the 2D plane of points (x, y) in some assumed fixed coordinate system. To be exact, a *picture* has two components:

(1) the *picture region*, which is a square whose center is at the origin of the coordinate system and whose sides are parallel to its axes;
(2) a *function* of two variables whose value is zero outside the picture region.

Identical functions may give rise to different pictures if the picture regions are different.

We often refer to the value $f(x, y)$ of the picture f at the point (x, y) as the *density* of f at (x, y). Within SNARK09, $f(x, y)$ is approximated by a *grid* G (which is a finite set $\{(g_1, h_1), \ldots, (g_J, h_J)\}$ of points in the plane), a *basic basis function* b (which is just a function of two variables), and *coefficients* c_j associated with each grid point (g_j, h_j) as follows. For $1 \le j \le J$, each of the functions

$$b_j(x, y) = b(x - g_j,\ y - h_j) \tag{1}$$

is called a *basis function* and f is defined by

$$f(x, y) = \sum_{j=1}^{J} c_j b_j(x, y); \tag{2}$$

i.e., as an expansion over the basis functions b_j with coefficients c_j. SNARK09 allows the use of two different kinds of basis functions: pixels and blobs [4,5], each with its own type of grid.

The exact definition of a *pixel basis function* depends on a variable called PIXSIZ that is specified by the input to SNARK09. The *pixel basis function* is then defined to have the value 1 at points strictly inside the square that is centered at the origin and that has edges of length PIXSIZ parallel to the coordinate axes, and to have the value 0 at points strictly outside this square. The associated grid G is defined, by an additional input-specified variable called NELEM, to be the set

$$\{(m \times \mathrm{PIXSIZ}, n \times \mathrm{PIXSIZ}) \mid m \text{ and } n \text{ integers},\ \max\{|m|, |n|\} \le \mathrm{NELEM}/2\}, \tag{3}$$

where $|\cdot|$ denotes the absolute value. This approach subdivides the picture region into NELEM^2 equal squares. Each of these smaller squares is called a *pixel* (short for picture element). In the interior of a pixel, the density of the function, as defined by (2), is uniform. An arbitrary picture can be approximated by such an expansion by simply assigning to each c_j the average density of the picture in the corresponding pixel; such an approximation is referred to as the NELEM^2 *digitization* of the picture. In SNARK09, the picture region (which is sometimes referred to as the *reconstruction region*) is determined by the program (based on the input-specified variables PIXSIZ and NELEM) as the square whose corners have coordinates (c, c), $(-c, c)$, $(-c, -c)$, $(c, -c)$, where

$$c = \frac{\mathrm{PIXSIZ} \times \mathrm{NELEM}}{2}. \tag{4}$$

A *blob basis function* also depends on some input-specified variables; however, SNARK09 will automatically calculate reasonable values for these parameters, relieving the user from the need of having to study the mathematical definitions that now follow. *Blob basis functions* are generalizations of the well-known window functions in digital signal processing called *Kaiser–Bessel* [4,5]; they are circularly symmetric, have nonzero values only in a circular disk around the origin, and smoothly decrease from a positive value at the origin to zero at the edge of the disk. The exact definition depends on the three variables SUPPORT, DELTA and SHAPE as follows:

$$k(x, y) = \begin{cases} \dfrac{\sqrt{3} \times \mathrm{DELTA}^2 \times \mathrm{SHAPE} \times (1 - r^2)}{4\pi \times \mathrm{SUPPORT}^2 \times I_3(\mathrm{SHAPE})} \times I_2(\mathrm{SHAPE} \times \sqrt{1 - r^2}), & \text{if } 0 \le r \le 1, \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where I_i denotes the modified Bessel function of the first kind of order i and $r = \sqrt{x^2 + y^2}/\mathrm{SUPPORT}$. The grid G that determines the blob basis functions using (5) is hexagonal [14]; it is defined as the set of all points in the set

$$\left\{ \left(\frac{m}{2} \times \mathrm{DELTA}, \frac{\sqrt{3}n}{2} \times \mathrm{DELTA} \right) \middle| \, m \text{ and } n \text{ are integers, and } m + n \text{ is even} \right\} \tag{6}$$

that are also inside the reconstruction region specified by (4).

2.1. *Creation of a phantom*

In practical applications one wishes to reconstruct a real object from its projections. During the development of reconstruction methods, though, it is preferable to work with mathematically described objects, called *phantoms*. The reason for this is that with real objects evaluation of the accuracy of a reconstruction method is practically impossible. The purpose of imaging and reconstruction is to visualize an object that cannot be seen otherwise (due to its size or because the internal structure is desired, for example, the internal structures of a cell or a cross section of a human head). Use of mathematically defined phantoms allows for evaluation of quality of reconstruction because these objects are known. Computer simulations with phantoms also allow for investigation of various phenomena occurring during both imaging and reconstruction separately from any other phenomena.

A phantom is a picture on which we wish to test reconstruction algorithms or data collection methods. In SNARK09 the phantom is put together by superimposing a number of *elemental objects*. There are five different types of elemental objects available: rectangles, ellipses, isosceles triangles, segments of circles and sectors of circles. The elemental objects are illustrated in Fig. 1. Each elemental object is described by its position in the plane (denoted by CX, CY), size along the two perpendicular directions (denoted by U, V), orientation (denoted by ANG) and density (which, for example, in case of computerized tomography represents the linear attenuation coefficient). They are allowed to overlap, in which case the densities of overlap areas are added together. Examples

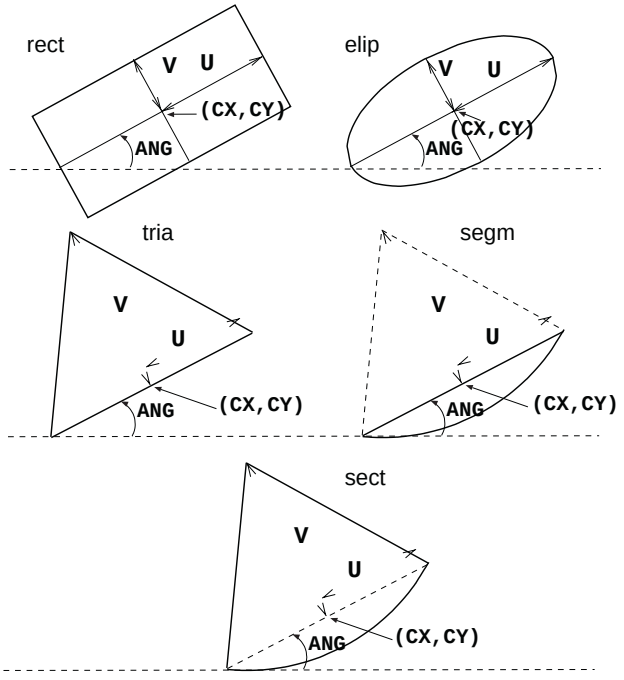


Fig. 1 – Elemental objects used to construct phantoms in SNARK09: rectangle (rect), ellipse (elip), triangle (tria), segment of a circle (segm) and sector of a circle (sect).

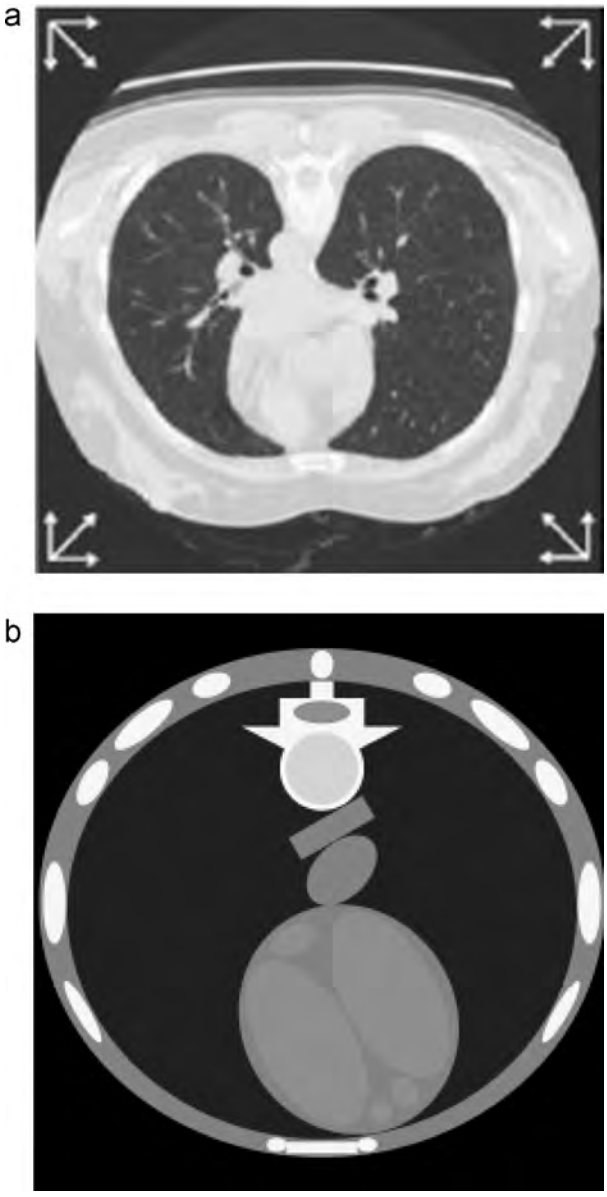


Fig. 2 – (a) CT image of a cross section of the human thorax (reproduced with permission from [15]). (b) 255 × 255 digitization of a thorax phantom created based on real cross sections such as the one shown in (a).

of phantoms that can be created from these simple elemental objects are shown in Fig. 2(b) (this is a phantom based on images, such as the one shown in Fig. 2(a) from [15], of an actual cross section of the human thorax; its exact specification is given in Table 1), Fig. 3(a) (which is based on a phantom in [16] designed for testing algorithms with data available only from a limited angular range) and Fig. 3(b) (which is a head phantom from [17]). The widely used Shepp–Logan phantom [18] can also be described using the elemental objects provided by SNARK09.

The $N \times N$ digitization of the phantom is an $N \times N$ array of pixels, where N is a user-specified integer. Each pixel's density is determined as the average of the densities at $K \times K$ uniformly-spaced points within a pixel. K is a user-specified

Table 1 – Basic thorax phantom description used in SNARK09, see Fig. 2(b). The linear attenuation coefficients (LAC) in this table are specified for a single energy level of 60 keV. To obtain the actual LAC at a point, the LACs provided in the table should be added together for all objects that contain that point. For example, the lungs, which are inside the smaller ellipse specified for the thorax (second row), have LAC value 0.196 − 0.147 = 0.049; this is in units of cm^{-1}.

Organ	Elemental object	CX	CY	U	V	ANG	LAC[a]
Thorax	elip	0.000	0.000	200.000	180.000	0.00	0.196
	elip	0.000	−5.000	180.000	162.000	0.00	−0.147
Heart	elip	19.022	−82.533	76.000	83.000	30.00	0.147
	elip	−10.778	−99.267	33.000	65.000	30.00	0.018
	elip	47.867	−64.956	33.000	65.000	30.00	0.018
	elip	−20.578	−26.356	17.000	10.000	30.00	0.018
	elip	58.711	−134.090	9.000	12.000	30.00	0.018
Sternum	elip	41.244	−147.533	8.000	8.000	30.00	0.018
	elip	32.000	−171.000	7.000	5.000	0.00	0.175
	rect	0.000	−173.000	25.000	4.000	0.00	0.175
Spinal column	elip	0.000	95.000	29.750	29.750	0.00	0.322
	elip	0.000	95.000	25.500	25.500	0.00	−0.053
	sect	0.000	124.750	57.000	29.750	180.00	0.322
	segm	0.000	124.750	57.000	29.750	180.00	−0.322
	sect	0.000	108.760	26.374	13.765	180.00	−0.322
	rect	0.000	134.750	30.000	10.000	0.00	0.322
	elip	0.000	134.750	20.000	8.000	0.00	−0.157
	rect	0.000	150.750	8.000	6.000	0.00	0.322
Ribs	elip	−169.000	−77.000	5.000	25.000	30.00	0.175
	elip	169.000	−77.000	5.000	25.000	−30.00	0.175
	elip	−189.000	0.000	8.000	29.000	0.00	0.175
	elip	189.000	0.000	8.000	29.000	0.00	0.175
	elip	−162.000	85.000	8.000	18.000	−30.00	0.175
	elip	162.000	85.000	8.000	18.000	30.00	0.175
	elip	−126.000	126.000	8.000	26.000	−50.00	0.175
	elip	126.000	126.000	8.000	26.000	50.00	0.175
	elip	−78.000	154.000	8.000	14.000	−70.00	0.175
	elip	78.000	154.000	8.000	14.000	70.00	0.175
Other	elip	14.000	23.000	19.000	29.500	−45.00	0.147
	rect	7.000	53.000	30.000	9.000	30.00	0.147

[a] Linear attenuation coefficient.

integer. As the value of K increases, the digitized phantom resembles the mathematical phantom more closely. The density assigned to a pixel can be expressed as a sum

$$\frac{1}{K^2}\sum_{k=1}^{K^2}\sum_{j=1}^{J}\delta_{k,j}d_j, \tag{7}$$

where J is the number of elemental objects in the phantom, d_j is the density of the jth elemental object, and $\delta_{k,j} = 1$ if the kth of the K^2 points in the pixel is in the jth elemental object and it is zero otherwise. Two possible digitizations, with $K = 1$ and $K = 13$, of a disk phantom are shown in Fig. 4.

In order to obtain phantoms that resemble actual biological objects more closely, SNARK09 has an option of adding local inhomogeneities to phantoms. Using locally piecewise phantoms may lead to misleading conclusions about the efficacy of an algorithm in practice, because biological structures are far from being piecewise constant [19]. Examples of a piecewise constant phantom and one with local inhomogeneities are presented in Fig. 3.

Phantoms in SNARK09 can carry information about attenuation at different energy levels of polychromatic X-ray radiation. The way such information is made use of is discussed in Section 2.2.3.

SNARK09 is not designed to work with "real" images, like the one in Fig. 2(a). One could create a phantom that consists of as many small squares as there are pixels in such image with the densities corresponding to the grayscale values in the image. This is not efficient and not advisable, and produces data biased toward reconstructions based on the pixelized images.

2.2. *Data collection*

SNARK09 is capable of simulating several modes of data collection used in various applications such as computerized tomography (CT) and positron emission tomography (PET). We first describe some general ideas and then specify how SNARK09 simulates the CT and PET modes of data collection. These simulations are based closely on the actual behavior of the instruments (see, for example, [17, Chapter 4] for CT and [20] for PET) and thus provide us with a tool capable of predicting the performance of such instruments in practice. As explained below (in particular in Section 3.1.1), once a projection data set is generated, it is stored internally (together

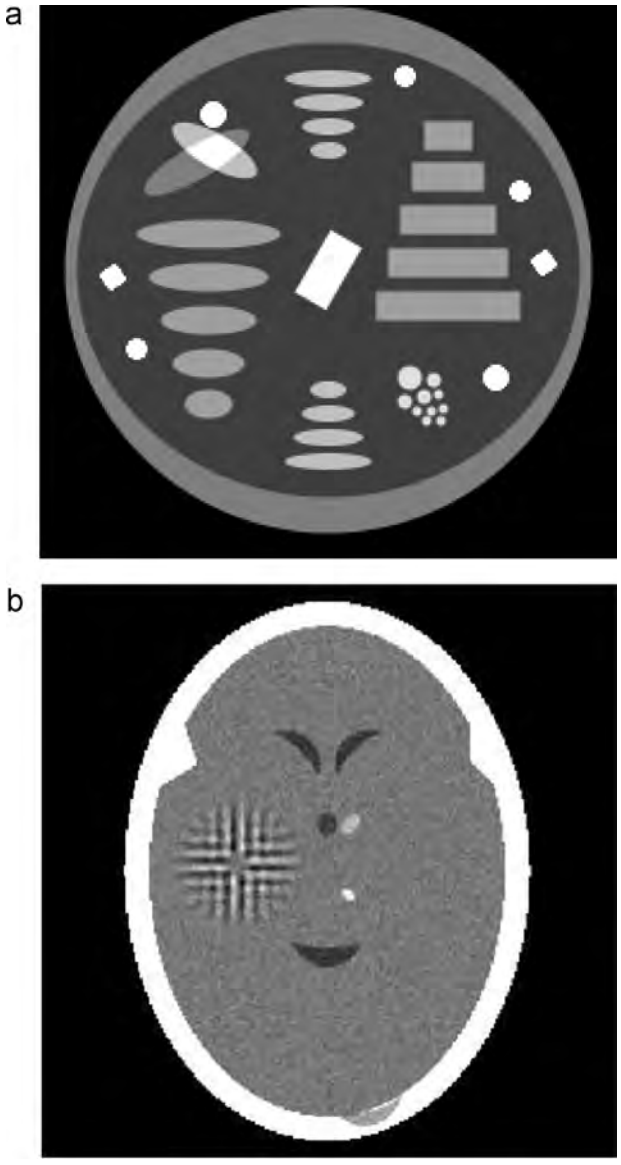


Fig. 3 – Examples of phantoms: (a) 241 × 241 digitization of a phantom based on [16]. It was designed for testing algorithms with data available only from a limited angular range. (b) 245 × 245 digitization of a head phantom from [17], with local inhomogeneities present.

with all the information that was utilized for its generation) to be used repeatedly by other processes such as reconstruction algorithms.

2.2.1. *Computation of ray sums*

Projection data collection is simulated by computation of approximate line integrals through the image according to the options indicated in the input file. The actual line integrals are approximated by summations of products of densities and lengths of intersection of all the elemental objects intersected by the given ray. There are two kinds of *rays* available:

(1) a *line ray*, which is a straight line, and
(2) a *strip ray*, which is a region of the plane between a pair of parallel straight lines.

Given a picture and a ray, the *real ray sum* is the integral of the picture along the ray (either a line or a strip). This is computed using the geometric description of the elemental objects, not the digitized version of the phantom.

SNARK09 also deals with *pseudo ray sums*; these are defined only for expansions of the form of Eq. (2). In the line ray case, the pseudo ray sum is the real ray sum of the picture defined by the function f of Eq. (2) (it uses the intersections of the ray with the basis function, either pixel or blob) and a picture region large enough to contain all points at which the value of f is not zero. (Since the grid G is finite, see paragraph above Eq. (1), it is always possible to find such a picture region.) In the case of strip rays, the pseudo ray sum is defined as

$$\left(\sum_{j\in S} c_j\right) \times \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} b(x, y)dxdy, \tag{8}$$

where S contains exactly those j ($1 \le j \le J$) for which (g_j, h_j) is in the strip. Note that the integral in the above equation depends only on the basic basis function; it is equal to the area of the pixel in the pixel case and area under the curve of the blob in the blob case.

Pseudo ray sums are used in iterative reconstruction algorithms in which the objective of the algorithm is the estimation of the coefficients c_j in Eq. (2). When SNARK09 is used for simulating how a physical device produces projection data, real ray sums should be used to obtain high accuracy. Depending on the parameters of the basis functions, pseudo ray sums may be very inaccurate.

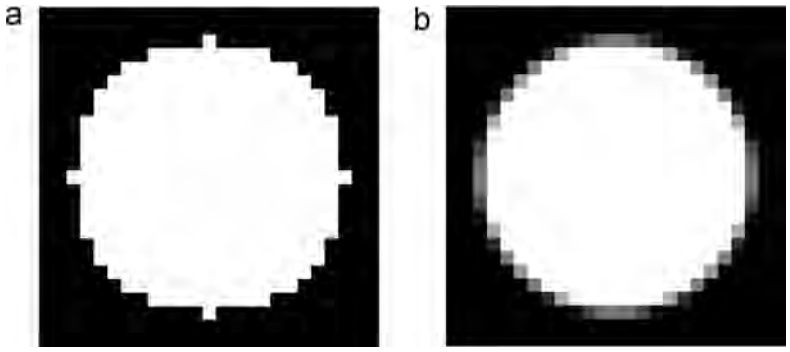


Fig. 4 – Two 25 × 25 digitizations of a mathematically described disk phantom, using (a) $K = 1$ and (b) $K = 13$.

2.2.2. *Geometry of data collection*

SNARK09 is not capable of handling an arbitrary arrangement of rays, but it can handle a number of arrangements of rays that are typical of what one might come across in practice.

The set of all rays along which data are collected is divided into a number of subsets, called *projections*. The number of projections and number of rays per projection are user-defined values, although SNARK09 can compute the number of rays sufficient to cover the entire area of the phantom. There are two basically different modes of data collection: divergent and parallel.

In *divergent geometry* (Fig. 5(a)), a projection consists of a set of line rays that go through a common point (the *source position*). In all the projections, the source is at a fixed distance from the origin. The angle between the line from origin to source and the *x*-axis (marked THETA in Fig. 5(a)) is called the *projection angle*. The rays in one projection connect the source to points (*detectors*) that lie either on an arc of a circle whose center is at the source or on a straight line tangent to that circle. In either case, one of the detectors (marked *C* in Fig. 5(a)) lies on the line connecting the source to the origin, at a distance STOD (for Source TO Detector) from the source. The other detectors are spaced symmetrically at equal intervals on the two sides of *C*, either on the arc whose center is the source or on the tangent line to this arc at *C*. The spacing between detectors (the length of the arc or that of the tangent line between two neighboring detectors) is denoted by PINC (specified by the user in the input file to SNARK09).

In *parallel geometry* (Fig. 5(b)), a projection consists of a set of parallel line or strip rays. The angle these rays make with the *x*-axis (denoted by THETA in Fig. 5(b)) is called the projection angle. In the line case one of the rays goes through the origin, in the strip case the origin is equidistant to the two lines bounding one of the strips. The other rays are spaced symmetrically at equal intervals on the two sides of this ray. In the strip case the rays are abutting. Let d denote the distance between the rays in the line case or the width of the rays in the strip case; it is determined as follows. The input specifies the variable PINC and also whether the *ray spacing* is to be uniform or variable. In the uniform case, $d = \text{PINC}$, for all projections. In the variable case it depends on the projection angle THETA: $d = \text{PINC} \times \max\{|\sin \text{THETA}|, |\cos \text{THETA}|\}$. (A consequence of this definition is that the distance between two consecutive intercepts with either the *x*- or the *y*-axis is PINC.)

2.2.3. *Simulating CT*

Computerized tomography (CT) is a method of imaging the interior of an object (frequently a human body) based on the measurements of X-ray radiation that passes through that object. The density values assigned to elemental objects in the phantom are interpreted as attenuation coefficients of the elemental object. The imaging of three-dimensional (3D) objects can be done in thin sections that are in practice considered to be two-dimensional (2D) images. Each measurement is related to the X-ray source and detector positions lying in the plane of such section. For each pair of source and detector positions two measurements are taken: a *calibration measurement* and an *actual measurement*. A calibration measurement is taken without the object, only through the background material. An actual measurement is taken through the object. Ideally, the

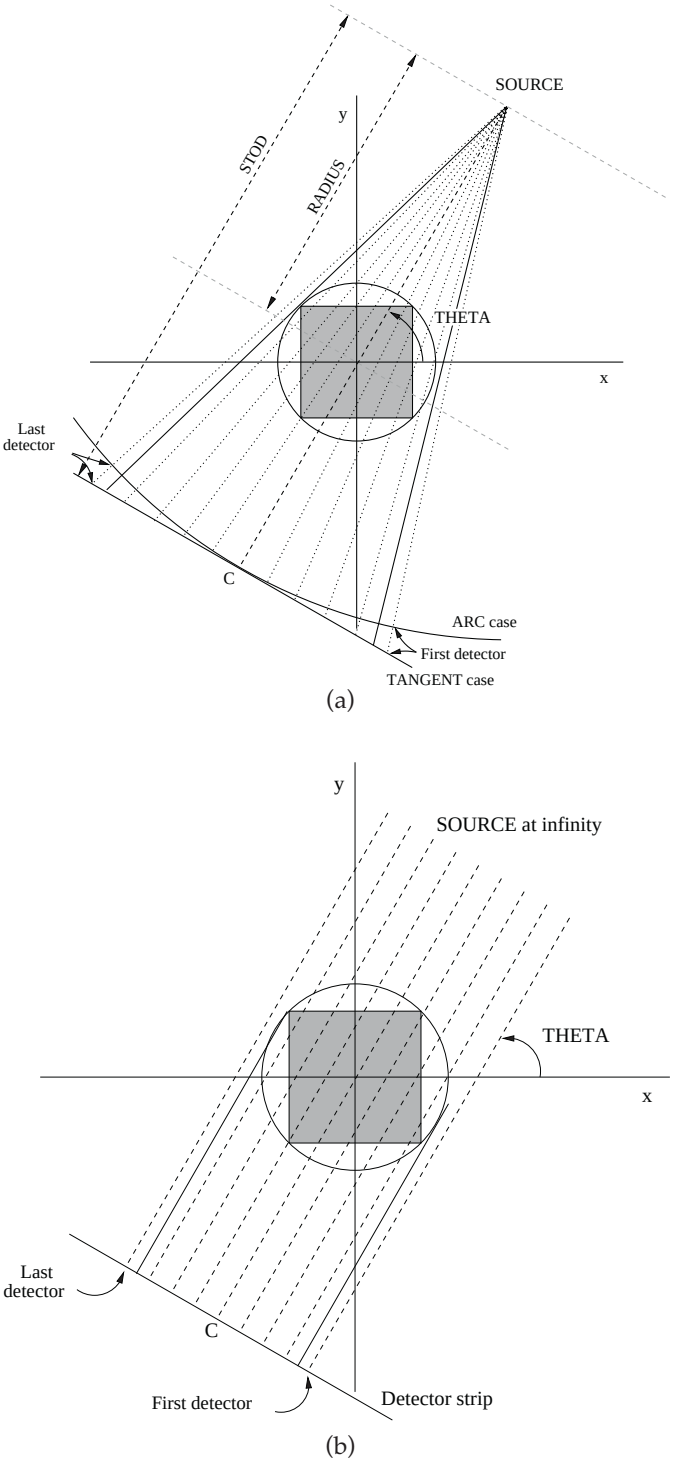


Fig. 5 – Schematic of geometry of data collection in CT: (a) divergent and (b) parallel.

X-ray spectrum would have a fixed energy level (monochromatic X-rays), which means that each point has a uniquely assigned attenuation coefficient. In practice, X-rays are made up of a continuous energy spectrum and this can be simulated in SNARK09. When this is done, attenuation of the X-ray beam at a point depends on the material traversed through by the beam prior to reaching that point, because more lower energy

photons get absorbed by that material than higher energy photons (this is referred to as *beam hardening*).

SNARK09 is capable of simulating both monochromatic and polychromatic X-rays. The polychromaticity of the X-ray is represented by up to seven discrete energy levels. The ray sum for a fixed pair of source-detector positions is defined by $p = -\ln(A/C)$, where A and C are actual and calibration measurements, respectively. The set of p values for all source-detector pairs is called the projection data. For polychromatic simulation the phantom description contains a list of linear attenuation coefficients corresponding to each discrete energy level. For the display purposes, the image of the phantom is generated based on linear attenuation coefficients of only one energy level.

When CT data collection is simulated in SNARK09, the values of ray sums are used in computations of A. According to options, specified in the input file, the simulated data reflects effects of beam hardening, detector width and scatter, quantum noise and various scanning modes (for a detailed discussion of these effects see, for example, [17]).

2.2.4. *Simulating PET*

In positron emission tomography (PET) we are interested in the uptake of positron-emitting isotopes by various parts of the human body. When a positron is emitted it is annihilated with a nearby electron and produces two γ-ray photons of identical energy traveling in approximately opposite directions [20, Fig. 1]. The two photons are detected in near coincidence by a pair of opposite detectors. The annihilation, and thus the positron emission, is known to take place somewhere along the line joining the detector pair [20, Fig. 2]. We count such coincidences for a number of detector pairs around the body. From these measured counts our aim is to estimate the concentration of the positron emitter at various points in the body cross-section.

Fig. 6(a) shows a simplification of PET geometry [20, Fig. 2] consisting of a ring of eight detectors. For simplicity of the illustration, we assume that each detector is coupled with three opposite detectors to detect (near) coincidence arrivals of photons. Thus the lines sampled by each detector form a divergent pattern. By analogy with X-ray CT, we refer to the collections of such (divergent) lines as a *projection* and the lines themselves as *rays* in the projection. Thus in Fig. 6(a) we have eight projections with three rays per projection and twelve rays in total in all the projections.

To simulate measurements by a PET system, SNARK09 utilizes many of the routines written to simulate X-ray CT measurements. Consider Fig. 6(b) to see how the X-ray CT organization is used to simulate data collection for the simplified PET geometry. It shows a schematic of a divergent geometry with the detectors located on an arc. The ray sums are measured along lines joining the "source" and three opposite detectors located on an arc of a circle whose center is at the source. Fig. 6(b) illustrates the situation when the "source" is at location D_2 and the detectors are on the arc $D'_5D'_6D'_7$. By simple geometrical considerations we see that, for this arc, the PINC of Section 2.2.2 can be selected so that the line connecting the "source" to a detector D' in the CT geometry goes through the corresponding detector D in the PET geometry and so the rays sums are calculated for the rays whose locations are the correct ones for the PET geometry of Fig. 6(a). In Fig. 6(b) a full scan is made by measuring the ray sums as the source rotates through locations D_1–D_8. The PET data simulation is completed by generating, for each ray sum, a Poisson random variable whose mean is given by the value of the ray sum. When simulating PET data collection, SNARK09 ignores attenuation effects.

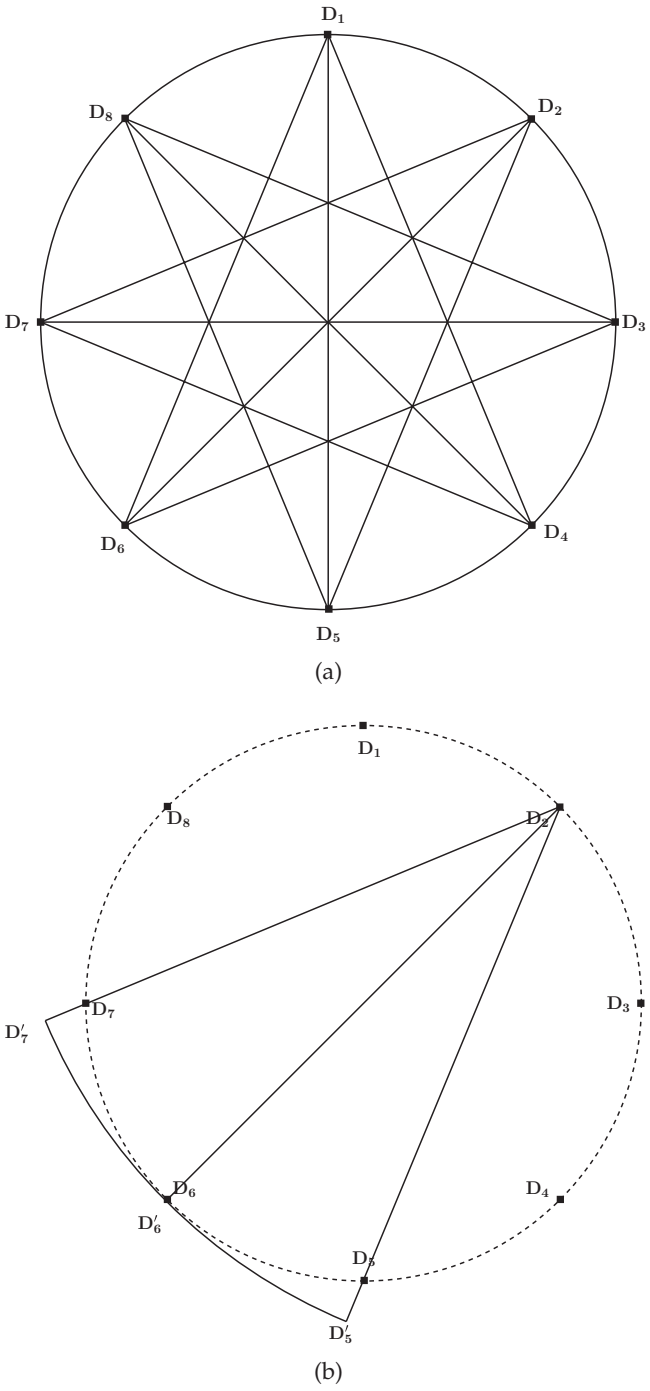


Fig. 6 – Schematic of geometry of data collection in PET: (a) Simplified geometry of eight-detector PET system, and (b) SNARK09 divergent geometry used to simulate the PET geometry shown in (a).

2.3. Built-in reconstruction algorithms

The SNARK09 package comes with several built-in reconstruction algorithms. It also provides the option for users to define their own reconstruction algorithms and termination tests. In this section we provide a very brief listing of the reconstruction algorithms that are available in SNARK09 together with references to works that further describe those algorithms (a more detailed description of all the reconstruction algorithms is beyond the scope of this publication). The references are not necessarily to the sources that originated the method, but rather to the ones that provide comprehensive descriptions.

The reconstruction methods are often categorized into two groups: transform methods and series expansion methods. SNARK09 provides several algorithms in both categories. The transform methods provided are filtered backprojection (FBP), rho-filtered layergram, Fourier method, and linograms. The series expansion methods provided are algebraic reconstruction techniques (ART), both additive and multiplicative, simultaneous iterative reconstruction technique (SIRT), quadratic optimization methods, and a maximum *a posteriori* probability (MAP) algorithm for PET based on a modified expectation–maximization (EM) algorithm (referred to as EMAP). The series expansion methods can be used with either pixels or blobs; the transform methods are limited to only pixel reconstructions.

FBP can be used for reconstruction from either parallel (see, e.g., [17,21]) or divergent rays (see, e.g., [22,17]). The standard backprojection works by estimating the density at a point by adding all the ray sums of the lines through that point. FBP filters projection data before it is used in the backprojection. Several different types of filters can be specified for this method.

Rho-filtered layergram (see, e.g., [17,21,23]) is a reconstruction method that attempts to deblur the picture that is obtained by backprojection alone. SNARK09 provides various deblurring methods that can be used with this algorithm.

The *Fourier method* (see, e.g., [17,24]) is based on the projection theorem. Roughly speaking, the projection data are first transformed using the one-dimensional Fourier transform. This provides values of the two-dimensional Fourier transform of the picture on radial lines. From these values the Fourier transform of the picture is estimated at the centers of the pixels of a grid, and the discrete inverse two-dimensional Fourier transform is used to get the reconstructed picture.

Linogram is another method based on the projection theorem, (see, e.g., [17,25,26]). Provided that the projection data are collected in a way that matches certain assumptions, the linogram algorithm produces reconstructions faster than FBP and the quality of the reconstructions tends to be better.

ART (see, e.g., [27,17]) is a family of iterative algorithms that, starting from an initial estimate of the picture to be reconstructed, update the estimate through a sequence of steps. A single step is influenced by exactly one ray for which we have an estimate of the ray sum. Only those basis function (pixel or blob) densities that contribute to the associated pseudo ray sum are updated. The updating is done by the addition of a correction term (additive ART) or multiplication by a correction term (multiplicative ART) to the density in each such basis function, so that after the correction the pseudo ray sum for the ray in question will be nearer to the ray sum in the projection data.

SIRT (see, e.g., [28,29]) is an iterative procedure that, starting from an initial estimate of the picture to be reconstructed, updates the estimate through a sequence of steps. Roughly speaking, the correction at each update is the discrete backprojection of a set of "projection error data" that consists of all the differences between the given ray sums and corresponding pseudo ray sums from the current estimate of the image.

Quadratic optimization techniques are a family of algorithms (see, e.g., [17,30,31]) that minimize a quadratic function of the vector of basis function densities using an iterative process. There are several choices available for the quadratic function to be minimized and the minimization method to be used.

EMAP is a maximum *a posteriori* probability (MAP) algorithm for PET based on a modified expectation–maximization (EM) algorithm (see [32,33]).

In addition to the built-in algorithms, the users can add up to ten user-defined reconstruction algorithms to SNARK09.

2.4. Evaluation

For single reconstructions, SNARK09 provides means for the evaluation of some quantitative measures of the overall difference between a digitized test phantom and its reconstruction. Such an evaluation can be performed either over the entire region of the image or over selected areas, and can be also restricted to pixels whose densities fall within a user-selected range.

2.5. Experimenter

It is often desirable to evaluate the relative efficacy of two or more reconstruction methods for a specific medical task in a manner that is statistically sound [34–37]. Such an evaluation must be done using a sample set that is large enough to provide a statistically significant result. Performing this evaluation on mathematical phantoms requires a means of running the competing algorithms on projection data obtained from a large number of randomly generated phantoms. Thereafter, various numerical measures of agreement between the reconstructed images and the original phantoms may be used to reach a conclusion that has some statistical substance. A straightforward way of achieving this goal is to provide a front-end or driver program that contains all the requisite commands that may be fed to SNARK09 to generate as many phantoms as needed together with their projection data, to implement the desired reconstruction algorithms on these data, and to evaluate the reconstructed images. Such a driver program is provided by SNARK09 in form of the Experimenter module. The method used in the comparative evaluation of the algorithms consists of the following:

- generation of random samples from a statistically described ensemble of phantoms and their projection data;
- reconstruction from the projection data by each of the algorithms to be compared;

- assignment to each reconstructed image a figure of merit (FOM), which measures the appropriateness of the image for solving the specified task;
- calculation of the statistical significance, based on the FOMs for all reconstructions, at which we can reject the null hypothesis that the methods are equally helpful for solving the task in favor of the alternative hypothesis that the one with the higher average FOM is more helpful.

The ensemble of phantoms available for multiple runs of SNARK09 within the Experimenter module has several possible sources of randomness:

(1) Users may specify a list of multiple phantom descriptions that are chosen at random during the experiment.
(2) For a fixed pair of distinct density values, paired structures, which are elemental objects that appear symmetrically with respect to the vertical line through the center of a phantom, can be assigned densities in such a way that in each pair exactly one structure has one of the two fixed density values. This assignment is done in a random manner at the time of phantom generation. Given s paired structures, there are 2^s possible phantoms, assuming that the paired structures are the only source of variability in the ensemble.
(3) Random inhomogeneity can be added to the pixel densities each time a new phantom is generated.
(4) Noise in the projection data may be generated at random each time a projection dataset is generated.

SNARK09 provides several built-in FOMs. It also allows users to create new FOMs that are more appropriate for a task at hand. Below we provide a brief description of FOMs that are built into SNARK09 together with references to works that further describe them.

The *structural accuracy* FOM [35,37] is computed as follows. Consider a phantom that contains a total of N structures. For a reconstruction, let α_k^r be the average pixel value for those pixels whose centers are within the structure k. Let α_k^p be the average pixel value of the corresponding structure in the phantom. The structural accuracy of a reconstruction is defined as

$$-\frac{1}{N}\sum_{k=1}^{N}|\alpha_k^r-\alpha_k^p|. \tag{9}$$

The *pointwise accuracy* FOM [35,37] is defined as the negative of the normalized root mean square distance between a reconstruction and the phantom. It is sometimes desirable to compute the pointwise accuracy when both the phantom and the reconstruction are clipped to a specified density range.

The *hit-ratio* FOM [35,37] is calculated only for those phantoms containing paired structures (such paired structures have unequal densities). For such pairs a hit occurs if the structure in the pair with the higher average density in the phantom is also the structure in the pair with the higher average density in the reconstruction. The hit-ratio for a reconstruction is the number of hits divided by the total number of pairs.

The *imagewise region of interest (IROI)* FOM [6] is calculated only for phantoms that contain paired structures. Such paired structures must have unequal densities with one of the structures having non-zero density (we refer to it as the *tumor*) and the other having density zero. The pairs of structures are numbered from 1 to B. For $1 \le b \le B$, let $\alpha_t^p(b)$ (respectively, $\alpha_n^p(b)$) denote the average density in the phantom of the structure of the bth pair that is (respectively, is not) the tumor. We specify similarly $\alpha_t^r(b)$ (respectively, $\alpha_n^r(b)$), for the reconstruction. The imagewise region of interest FOM is defined by

$$\mathrm{IROI} = \frac{\left[\sum_{b=1}^{B}\left(\alpha_t^r(b)-\alpha_n^r(b)\right)\right] \Big/ \left[\sqrt{\sum_{b=1}^{B}\left(\alpha_n^r(b)-(1/B)\sum_{b'=1}^{B}\alpha_n^r(b')\right)^2}\right]}{\left[\sum_{b=1}^{B}\left(\alpha_t^p(b)-\alpha_n^p(b)\right)\right] \Big/ \left[\sqrt{\sum_{b=1}^{B}\left(\alpha_n^p(b)-(1/B)\sum_{b'=1}^{B}\alpha_n^p(b')\right)^2}\right]}. \tag{10}$$

The first thing to note about this formula is that the numerator and the denominator in the big fraction are exactly the same except that the numerator refers to the reconstruction and the denominator refers to the phantom. Thus, if the reconstruction is perfect (in the sense of being identical to the phantom) then IROI = 1. Analyzing the contents of the numerator and the denominator, we see that they are (except for constants that cancel out) the mean difference between the average values at the tumor site and the corresponding non-tumor site divided by the standard deviation of the average values at the non-tumor sites. It has been found by experiments with human observers that this FOM correlates well with the performance of people [6].

3. System description

In this section we discuss the structure of the SNARK09 package, its various modules and the graphical user interfaces.

3.1. Application framework

A SNARK09 run can be subdivided into three phases: (1) data generation, (2) initialization and reconstruction, and (3) analysis.

Each of these phases requires some input data and produces some output data. Some of the output is used as the input of a later (or even the same) phase. Provided that the appropriate input data are available, a single SNARK09 run may consist of one, two or all three of these phases.

We now proceed with a description of each of the three phases. The reader should consult Fig. 7 for an overview. The specific details of mandatory and optional parts of the input are discussed in the online manual [2]. The reader should be aware that the word “input” is used for both the stream of commands that drive a whole SNARK09 run, but also to what

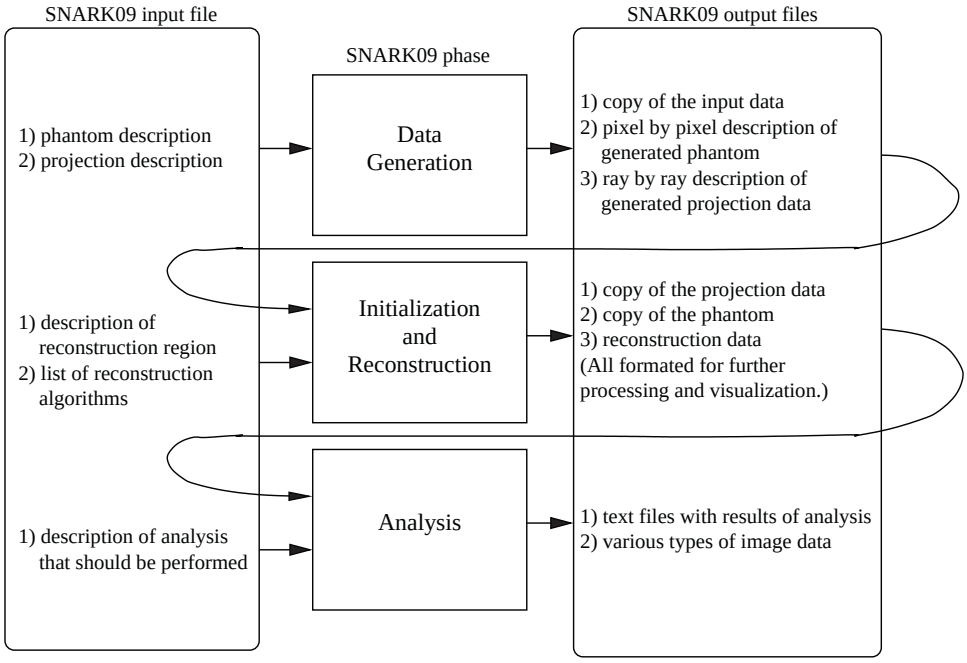


Fig. 7 – Data flow in SNARK09.

is considered to be the input data to any of the phases of such a run.

3.1.1. *Data generation phase*

During this phase SNARK09 generates a phantom and projection data of it. The projection data consist of real ray sums of the phantom, possibly contaminated by the types of noise that one may come across in a device used for collecting data for reconstruction.

Input: The input for this phase consists of (1) the geometrical description of the phantom, and (2) description of the projections including geometry of data collection, number and distribution of projections and noise present during the data collection process.

Output: The output of this phase consists of (1) a copy of the geometrical description of the phantom, (2) a pixel by pixel description of the phantom for one or more energy levels (according to specifications provided in the input), (3) a copy of the geometry of data collection, number and distribution of projections and noise present during the data collection process, and (4) a ray by ray description of each projection. The output is self contained in the sense that it contains all the information provided by the input together with the newly generated data. This is used as input for the subsequent phases either in the same run of SNARK09 or in separate runs, in which the presence of the original input data is not required.

3.1.2. *Initialization and reconstruction phase*

The main goal of this phase is to perform the reconstruction based on the available projection data. This phase can be further divided into two sub-phases: (1) initialization and (2) reconstruction. The latter cannot be performed without the former having been performed in the same run.

During the initialization sub-phase, the nature of the grid for the reconstruction region as well as the assumed geometry of data collection are determined. The phantom and projection data are written as output files in the XML (extensible markup language) format, which is suitable for easy access in subsequent processing steps and for visualization. During the

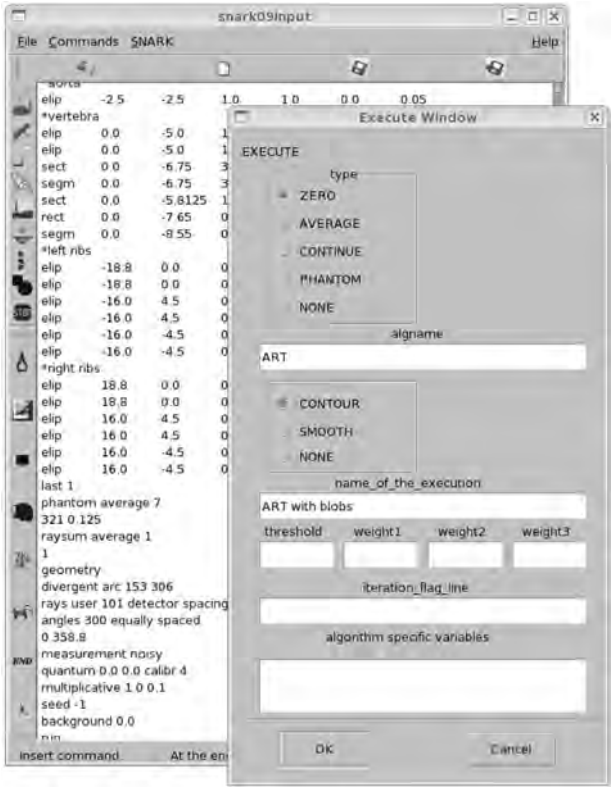


Fig. 8 – SNARK09Input is a graphical user interface used for creation of SNARK09 input files.

reconstruction sub-phase the reconstruction algorithms are carried out and the results are saved.

Input: The first source of input data is the main input file for the SNARK09 run. It contains (1) the description of the reconstruction region, and (2) the list of algorithms to be used for the reconstruction. The second input file is simply the output file produced in the previous phase. This file does not have to be created in the same run of SNARK09. In fact, the second file can be manually produced and filled by the data obtained by a real imaging device.

Output: There are two XML output files produced by this phase. The first one contains just a copy of the projection data. It is written in the format used in the subsequent phases and for visualization. The second file contains a copy of the phantom (if it is available) and of all the reconstructions in the current run of SNARK09. If iterative reconstruction methods are used, then the reconstructions produced by each of the iterative steps are saved.

3.1.3. *Analysis phase*

In the final phase, the data obtained by the reconstruction algorithms can be further processed and analyzed. Depending on the commands in the input file there are several things that may be achieved here: (1) statistical analysis of the results, (2) comparison of reconstruction(s) with a phantom, (3) storing of a reconstruction in a format that allows for its later use as a starting point for another reconstruction algorithm, and (4) saving of the reconstructions in a standard image format.

Input: The first source of input data comes from the main input file for the SNARK09 run. It contains commands that indicate what needs to be computed and written in what format. The second input file is the data file produced by the previous phase (in the same or a separate run) that contains all the reconstructions and the original phantom (if there is one).

Output: The output files depend on what the user specified in the original input file. They can be text files with results of statistical analysis and image files with requested graphics.

3.2. *DIG libraries*

The DIG libraries are used in the creation of the projection and reconstruction data files produced in SNARK09 runs. They provide routines that can be used easily to access, extract and modify the data stored in those data files. Programmers who need to process further the projection and reconstruction data should use these libraries to obtain easy and safe access to them.

3.3. *Graphical user interfaces*

To ease the use of SNARK09, two interactive graphical user interfaces have been designed for creating input files and for visualization of projection data, reconstructions and the analysis of the results. SNARK09Input assists users in the creation of input files used in SNARK09; see Fig. 8. SNARK09Display allows users to display the outputs of a SNARK09 run; see

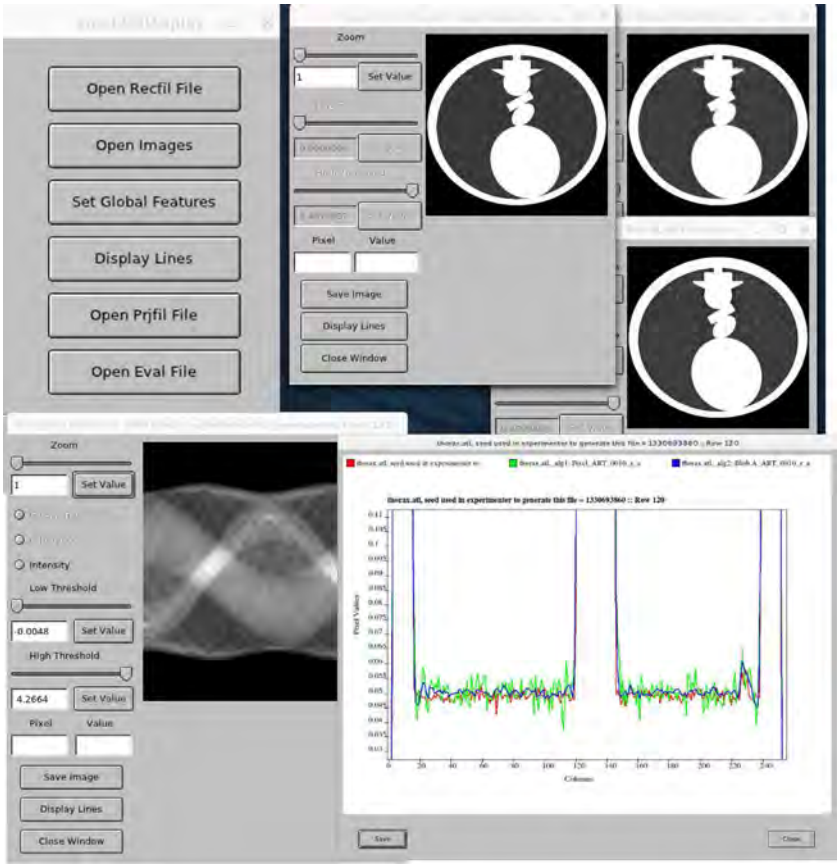


Fig. 9 – SNARK09Display is a graphical user interface used for display of results obtained by a SNARK09 run.

Fig. 9. It can display 2D images of projection data and of reconstructions at user-defined gray-level intensity values and plot their row/column profiles. It also can display graphically data analysis results. The images presented in the next section for an example run of SNARK09 have all been generated using SNARK09Display.

4. Example of use

In this section we illustrate in detail an example of how SNARK09 can be used in practice. We present multiple features of the package, but it is impossible to make use of all features in a single example. The reader is referred to the SNARK09 manual [2] for the detailed listing of all the available features and for many more examples of its use. There are ten worked out examples in the manual with input files and output generated by SNARK09; these examples are provided also when SNARK09 package is downloaded from its website. The book [17] used SNARK09 and the phantom from Fig. 3b for demonstration of many concepts related to computerized tomography.

In the example reported here, we evaluate the usefulness of two reconstruction algorithms for recovery of low-contrast

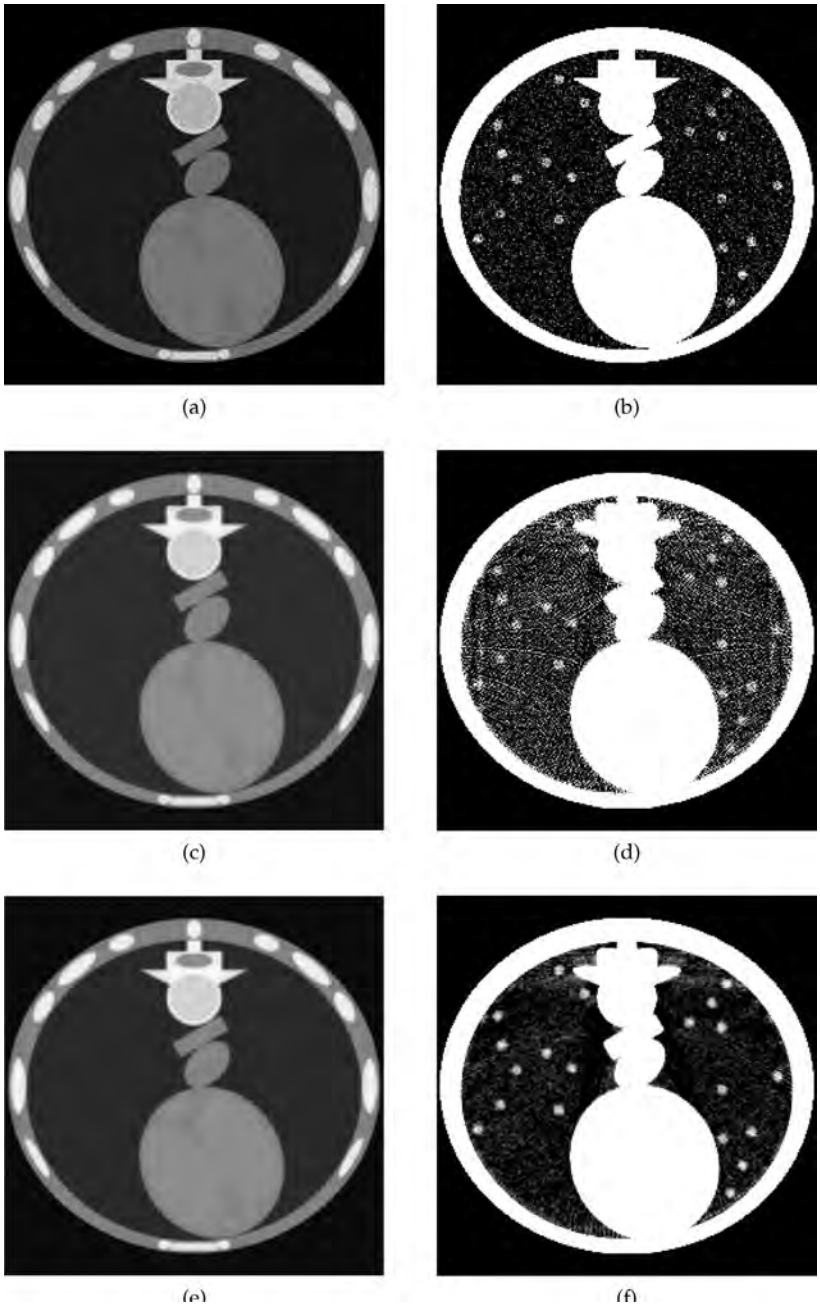


Fig. 10 – (a and b) Thorax phantom with randomly generated small low-contrast tumors in the lungs and tissue inhomogeneity, (c and d) ART reconstruction using pixel basis functions, (e and f) ART reconstruction using blob basis functions. In all cases, the image on the left uses the full display window of grayscale values while the image on the right uses a narrow display window for better visualization of the low-contrast tumors.

tumors in lung tissue. The Experimenter part of SNARK09 allows us to do such an evaluation. We need to choose an ensemble of phantoms, one or more figures of merit (FOMs), and two or more algorithms whose performance is being compared.

Consider the thorax phantom shown in Fig. 2(b). We modified the phantom by adding a list of 20 pairs of possible tumor sites in the lung. The tumors are represented by small circles with linear attenuation coefficients 15% higher than the underlying lung tissue. Furthermore, we added inhomogeneity to the phantom to represent the biological tissue more accurately. This is done in SNARK09 by adding to each pixel a random value from a zero mean Gaussian distribution with a specified standard deviation. In this case, we used a standard deviation of 6% of the underlying density. The inhomogeneity of the tissue lowers significantly the difference in density between the lung tissue and the tumors, making it more challenging for the algorithms to recover the tumors correctly. During the experiment, for each pair of possible tumor sites, the tumor is randomly placed either in the left or right lung, giving us 2^{20} possible phantoms even before the inhomogeneity is added. The contrast between the tumors and lung tissue is so low, that when the phantom is displayed using the full range of attenuation coefficients mapped to grayscale values, the tumors are practically invisible, see Fig. 10(a). This is due to much higher attenuation coefficients for bone and muscle tissue as compared to the lung. The tumor sites become visible when the display window of gray values is narrowed, see Fig. 10(b).

We simulated polychromatic X-rays. To do so, we used five discrete energy levels. The attenuation coefficients for five different energy levels are listed in Table 2. The attenuation coefficients for energy of 60 keV correspond to the ones in the phantom description in Table 1. The projection data were obtained from 360 angles equally spaced in the range $[0--360^{\circ})$ using divergent rays. Each projection contained 363 rays. The collected data were corrupted by quantum and scatter noise. The data were corrected for beam hardening (due to polychromaticity of the X-ray beam) before it was used for reconstruction. The projections for one of the randomly generated phantoms are shown as columns of the image in Fig. 11.

We compared the two variants of the built-in ART algorithm: one using pixels, the other using blobs. We used a built-in FOM: imagewise region of interest (IROI). The IROI FOM has been confirmed to correlate well with human observers for detectability of small, low density features [6]. The number of FOMs computed for each experiment is up to the user.

Using SNARK09 Experimenter, both versions of ART were automatically run thirty times, each time generating a new phantom and a new set of projection data based on which reconstructions were computed. The reconstructions computed in one such run are shown in Fig. 10(c)–(f). When viewed in the full window of grayscale values (see Fig. 10(c) and (e)) the two reconstructions are almost indistinguishable. The differences appear only when the images are viewed using a much narrower display window (see Fig. 10(d) and (f)). After all runs completed, statistical significance was computed using the FOM values for each reconstruction algorithm. The results of this statistical analysis are presented in Table 3. The table shows average values of the figure of merit computed for different iterations of ART. For ART using pixels the highest average FOM was obtained at the eighth iteration. For ART using blobs the highest average FOM was obtained at the fifth iteration. The last column in Table 3 shows that the differences are statistically significant. Thus, according to the values of the IROI FOM, we can reject the null hypothesis that the two variants of ART perform equally well for detection of small low-contrast tumors in favor of the alternative hypothesis that ART using blobs performs better.

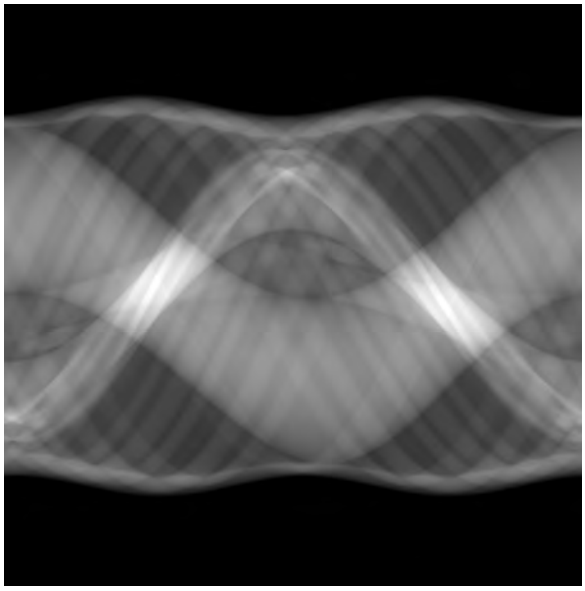


Fig. 11 – Projection dataset obtained based on the thorax phantom shown in Fig. 10(a). Each column of pixels in the image corresponds to a single projection.

The reconstruction obtained using ART with blobs smoothes the inhomogeneities in the lungs, resulting in an increased contrast between the tumors and their background. This can be seen in the plots of the density values along column 191 of the reconstructions and the phantom shown in Fig. 12. In fact, the lung in the reconstruction using blobs appears to be smoother than in the phantom; which results in the IROI FOM having a value greater than one. The claims of superiority of one algorithm over another can be made only for the FOM that measures the task at hand. From Fig. 12 it is clear that the variability in the lung tissue was not recovered well by the ART with blobs. But recovery of the background variability was not the task that was evaluated; the task was to detect small low-contrast tumors and, by smoothing the reconstruction, ART with blobs generated reconstructions with small tumors that are visible more clearly than in the reconstructions produced by the ART with pixels.

5. Availability and system requirements

SNARK09, SNARK09Input and SNARK09Display are all open source. They are available for download on the SNARK09 website [1] free of charge.

Table 2 – Linear attenuation coefficients (in cm^{-1}) as a function of photon energy for tissues that occur in the thorax phantom.

Energy (keV)	Muscle	Blood	Fat	Lung	Compact bone	Soft bone	Tumors
40	0.249	0.278	0.224	0.062	0.642	0.520	0.071
50	0.214	0.234	0.198	0.055	0.455	0.382	0.063
60	0.196	0.214	0.184	0.049	0.371	0.318	0.056
80	0.178	0.189	0.170	0.047	0.298	0.261	0.054
100	0.167	0.176	0.161	0.042	0.265	0.235	0.048

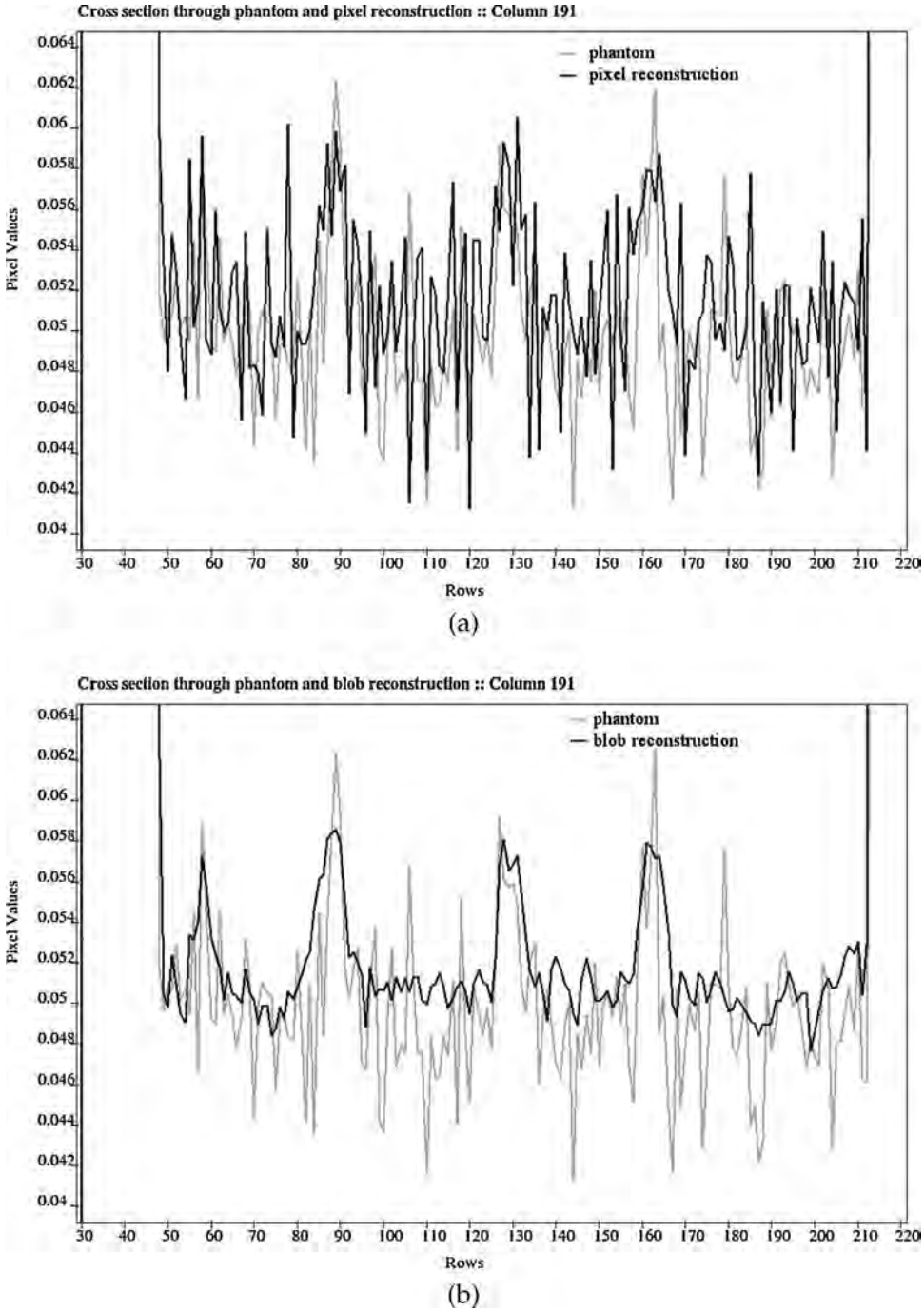


Fig. 12 – Comparison of the plots of the density values along column 191 through the two reconstructions (in black) and the phantom (in gray): (a) ART with pixels and (b) ART with blobs.

Table 3 – Statistical analysis results computed by SNARK09 experimenter for the example in Section 4. The statistical significance of the observed differences between the performance of the pixel and blob algorithms (as measured by the IROI FOM) was calculated and is given in the last row.

FOM: Imagewise-ROI				
ART with pixels		ART with blobs		Significance level
Iteration	Mean	Iteration	Mean	
1	0.2438	1	0.8386	0.00000005
3	0.7339	3	1.0320	0.00000906
5	0.8557	5	1.1144	0.00002006
8	0.8659	8	1.1043	0.00001781
10	0.8497	10	1.0957	0.00001194
8	**0.8659**	**5**	**1.1144**	**0.00001688**

The package is a Linux/Unix based system. It runs on a typical modern PC and has no specific hardware requirements. The software libraries used by SNARK09 are provided in repositories of all the major Linux distributions. SNARK09 is implemented in C/C++, which are available on a wide variety of hardware and operating system platforms and are currently among the most popular programming languages used by computer scientists. The standard development packages that come with a typical Linux distribution are sufficient to compile and build the package.

6. Future work

SNARK09 is a package that is the result of more than three decades of continuous development. It evolves as the field of tomographic reconstruction changes. There are many aspects of it that can be modified and expanded. We plan to rewrite some of the existing code to make it computationally more efficient and to take advantage of some of the multiprocessing hardware (multicore processors and/or graphics processing units) that have become, in recent years, standard on a typical desktop computer. We also plan to incorporate into the standard SNARK09 code new reconstruction algorithms that we are currently using as user-defined routines. An example of this is the recently developed superiorization methodology for image reconstruction, that has been implemented and thoroughly investigated within SNARK09 via user-defined routines; see, e.g., [38,39].

Acknowledgements

The early versions of this package go back to the 1970s. Since then many researchers and programmers have been involved in its development. Their names, along with names of funding institutions are listed in SNARK09 manual [2]. The work of all these people is greatly appreciated, they contributed a great deal to this feature-rich package.

The work of Joanna Klukowska and Gabor T. Herman is currently supported by the National Science Foundation award number DMS-1114901. The work of Ran Davidi is supported by the Department of Defense Prostate Cancer Research Program award number W81XWH-12-1-0122.

The authors would also like to thank the reviewers of the paper for helpful comments and suggestions that were gratefully followed.

REFERENCES

[1] SNARK09, http://www.dig.cs.gc.cuny.edu/software/snark09 (accessed November 2012).
[2] R. Davidi, G.T. Herman, J. Klukowska, SNARK09: A Programming System for the Reconstruction of 2D Images from 1D Projections, http://www.dig.cs.gc.cuny.edu/software/snark09/SNARK09.pdf, 2012.
[3] J.E. Bresenham, Algorithm for computer control of a digital plotter, IBM Systems Journal 4 (1965) 25–30.
[4] R.M. Lewitt, Multidimensional digital image representations using generalized Kaiser–Bessel window functions, Journal of the Optical Society of America A: Optics and Image Science 7 (1990) 1834–1846.
[5] S. Matej, R.M. Lewitt, Practical considerations for 3-D image reconstruction using spherically symmetric volume elements, IEEE Transactions on Medical Imaging 15 (1996) 68–78.
[6] T.K. Narayan, G.T. Herman, Prediction of human observer performance by numerical observers: an experimental study, Journal of the Optical Society of America A: Optics and Image Science 16 (1999) 679–693.
[7] S.W. Rowland, jSNARK 1.0.5, http://jsnark.sourceforge.net/ (accessed November 2012).
[8] S.H.W. Scheres, R. Nú nez-Ramírez, C.O.S. Sorzano, J.M. Carazo, R. Marabini, Image processing for electron microscopy single-particle analysis using XMIPP, Nature Protocols 3 (2008) 977–990.
[9] T.R. Shaikh, H. Gao, W.T. Baxter, F.J. Asturias, N. Boisset, A. Leith, J. Frank, SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs, Nature Protocols 3 (2008) 1941–1974.
[10] IMOD, http://bio3d.colorado.edu/imod/ (accessed November 2012).
[11] Image Reconstruction Toolbox, http://web.eecs.umich.edu/~fessler/code/index.html (accessed November 2012).
[12] P.C. Hansen, M. Saxild-Hansen, AIR tools – a MATLAB package of algebraic iterative reconstruction methods, Journal of Computational and Applied Mathematics 236 (2012) 2167–2178.
[13] STIR, http://stir.sourceforge.net/ (accessed November 2012).
[14] G.T. Herman, Geometry of Digital Spaces, Birkhäuser, Boston, MA, 1998.
[15] Y. Yim, H. Hong, Correction of segmented lung boundary for inclusion of pleural nodules and pulmonary vessels in chest

CT images, Computers in Biology and Medicine 38 (2008) 845–857.
[16] A.H. Delaney, Y. Bresler, Globally convergent edge-preserving regularization: an application to limited-angle tomography, IEEE Transactions on Image Processing 7 (1998) 204–221.
[17] G.T. Herman, Fundamentals of Computerized Tomography: Image Reconstruction from Projections, 2nd ed., Springer, London, 2009.
[18] L.A. Shepp, B. Logan, Reconstructing interior head tissue from X-ray transmissions, IEEE Transactions on Nuclear Science NS21 (1974) 228–236.
[19] G.T. Herman, R. Davidi, Image reconstruction from a small number of projections, Inverse Problems 24 (2008) 045011.
[20] T.T. Turkington, Introduction to PET instrumentation, Journal of Nuclear Medicine Technology 29 (2001) 4–11.
[21] S.W. Rowland, Computer implementation of image reconstruction formulas, in: G.T. Herman (Ed.), Image Reconstruction from Projections: Implementation and Applications, Springer-Verlag, Berlin, 1979, pp. 9–79.
[22] T. Chang, G.T. Herman, A scientific study of filter selection for a fan-beam convolution reconstruction algorithm, SIAM Journal on Applied Mathematics 39 (1980) 83–105.
[23] P.R. Smith, T.M. Peters, R.H.T. Bates, Image reconstruction from finite numbers of projections, Journal of Physics A: Mathematical, Nuclear and General 6 (1973) 361–382.
[24] R.M. Mersereau, Direct Fourier transform techniques in 3-D image reconstruction, Computers in Biology and Medicine 6 (1976) 247–258.
[25] P. Edholm, G.T. Herman, Linograms in image reconstruction from projections, IEEE Transactions on Medical Imaging 6 (1987) 301–307.
[26] P. Edholm, G.T. Herman, D.A. Roberts, Image reconstruction from linograms: implementation and evaluation, IEEE Transactions on Medical Imaging 7 (1988) 239–246.
[27] R. Gordon, R. Bender, G.T. Herman, Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography, Journal of Theoretical Biology 29 (1970) 471–482.
[28] P. Gilbert, Iterative methods for three-dimensional reconstruction of an object from projections, Journal of Theoretical Biology 36 (1972) 105–117.
[29] A.V. Laksminarayanan, A. Lent, Methods of least squares and SIRT in reconstruction, Journal of Theoretical Biology 76 (1979) 267–295.
[30] G.T. Herman, A. Lent, Quadratic optimization for image reconstruction I, Computer Graphics and Image Processing 5 (1976) 319–332.
[31] E. Artzy, T. Elfving, G.T. Herman, Quadratic optimization for image reconstruction II, Computer Graphics and Image Processing 11 (1979) 242–261.
[32] L.A. Shepp, Y. Vardi, Maximum likelihood reconstruction for emission tomography, IEEE Transactions on Medical Imaging 1 (1982) 113–122.
[33] G.T. Herman, A.R. De Pierro, N. Gai, On methods for maximum a posteriori image reconstruction with a normal prior, Journal of Visual Communication and Image Representation 3 (1992) 316–324.
[34] G.T. Herman, K.T.D. Yeung, Evaluators of image reconstruction algorithms, International Journal of Imaging Systems and Technology 1 (1989) 187–195.
[35] G.T. Herman, D. Odhner, Performance evaluation of an iterative image reconstruction algorithm for positron emission tomography, IEEE Transactions on Medical Imaging 10 (1991) 336–346.
[36] S.S. Furuie, G.T. Herman, T.K. Narayan, P. Kinahan, J.S. Karp, R.M. Lewitt, S. Matej, A methodology for testing for statistically significant differences between fully 3-D PET reconstruction algorithms, Physics in Medicine & Biology 39 (1994) 341–354.
[37] J.A. Browne, G.T. Herman, Computerized evaluation of image reconstruction algorithms, International Journal of Imaging Systems and Technology 7 (1998) 256–267.
[38] T. Nikazad, R. Davidi, G.T. Herman, Accelerated perturbation-resilient block-iterative projection methods with application to image reconstruction, Inverse Problems 28 (2012) 035005.
[39] G.T. Herman, E. Garduño, R. Davidi, Y. Censor, Superiorization: an optimization heuristic for medical physics, Medical Physics 39 (2012) 5532–5546.

200 MeV Proton Radiography Studies with a Hand Phantom Using a Prototype Proton CT Scanner

Tia Plautz, V. Bashkirov, V. Feng, F. Hurley, R.P. Johnson, C. Leary, S. Macafee, A. Plumb, H.F.W. Sadrozinski, K. Schubert, R. Schulte, B. Schultze, D. Steinberg, M. Witt, A. Zatserklyaniy

***Abstract*—Proton radiography generates two-dimensional projection images of an object and has applications in patient alignment and verification procedures for proton beam radiation therapy. The quality of the image, both contrast and spatial resolution, is affected by the energy of the protons used in the creation of the radiograph, as well as by multiple Coulomb scattering and energy-loss straggling. Here we report an experiment which used 200 MeV protons to generate proton energy-loss and scattering radiographs of a hand phantom. It was found that while both radiographs displayed anatomical details of the hand phantom, the energy-loss radiograph has a noticeably higher spatial resolution. The scattering radiograph may yield sharper edges between soft and bone tissue than energy loss radiograph, but this requires further study. These radiographs demonstrate the new promise of proton imaging (proton radiography and CT) now within reach of becoming a new, potentially low-dose medical imaging modality. The experiment used the current first-generation proton CT scanner prototype, which is installed on the research beam line of the clinical proton synchrotron at Loma Linda University Medical Center. This study contributes to the optimization of the performance of a clinical proton CT scanner.**

***Index Terms*—proton imaging, tomographic reconstruction of material properties, spatial resolution, data reduction**

I. Introduction

With increasing use of proton radiation therapy for cancer patients, research into new imaging methods that can improve the accuracy of proton range estimates in radiation therapy planning have become a high priority. Protons are particularly desirable for treating cancerous tissue in close proximity to radiosensitive normal tissues, such as at the base of skull and near the spinal cord. Protons are preferable to photons because their energies are easily tuned, the unhealthy area can be isolated, and the dose can be localized reducing the threat of damaging otherwise healthy tissue. Most importantly, the greatest radiation dose occurs only in the last 2% of the proton's range, at the Bragg peak, so a maximum amount of healthy tissue can be spared when the position of the Bragg peak is controlled.

Tia Plautz, V. Feng, R. P. Johnson, C. Leary, S. Macafee, A. Plumb, H. F. W. Sadrozinski, D. Steinberg and A. Zatserklyaniy are with the Santa Cruz Institute for Particle Physics, University of California Santa Cruz, Santa Cruz, CA 95064 USA (e-mail: tiaplautz@gmail.com).

V. Bashkirov, R. F. Hurley, R. Schulte are with Loma Linda University Medical Center, Loma Linda, CA 92354 USA.

K. Schubert, B. Schultze, M. Witt are with CSU San Bernardino, San Bernardino, CA 92407 USA.

Manuscript received 11 December, 2012.

This work was supported in part by Grant No. 1R01EB013118-01 from the National Institute of Health.

In order to obtain relative stopping power (RSP), Hounsfield units (i.e. units of x-ray attenuation used in x-ray CT) are transformed using a calibration curve. However, there is no unique relationship between Hounsfield units and RSP, especially in the regime of RSP=1 (i.e. water, human tissue). This means that during conversion, errors in proton range are consistently 3-4% of the nominal proton range or even higher in regions containing bone [1]. A recent survey by the American Association of Physicists in Medicine (AAPM) showed that 33% of attendees polled said that range uncertainties are the main obstacle to making proton therapy mainstream [2]. Simulations and first experimental results have shown that using a proton CT imaging system one may be able to reduce this range uncertainty to about 1% or less without increasing the dose to the patient.

Proton CT differs in several key aspects from x-ray CT. While unscattered photons travel in straight line paths, protons do not and rather undergo many multiple Coulomb scattering (MCS) events, which limits the usefulness of the standard filtered back projection (FBP) approach to reconstruction. In fact, proton CT images reconstructed with the classical FBP algorithm suffer from loss of spatial resolution since the proton path deviates from the assumed straight lines by up to several millimeters in anatomical objects encountered in medical proton CT imaging. The accuracy of those path estimates is critical for achieving a high spatial resolution in proton CT.

A. Current Prototype Design

A low intensity, high energy (100-200 MeV) cone beam of protons traverses a phantom. Silicon strip detectors (228 μm pitch) record the proton path in 4 planes (each 400 um thick) so entry and exit vectors can be easily determined. Detectors interface through a high speed field programmable gate array (FPGA)-based data acquisition system. A calorimeter composed of an array of 18 CsI crystals is used to detect the residual energies of incident protons at a rate of up to 100k protons/sec.

B. Reconstruction Software

Mathematical algorithms and computer software are used to reconstruct the phantom from raw data [3]. Raw data contain the proton tracker coordinates and the calorimeter's response for each proton. The software bins the exit tracker data into spatial bins (pixels) and determines cuts in relative angle, defined as the difference between entry and exit angle, at 3σ

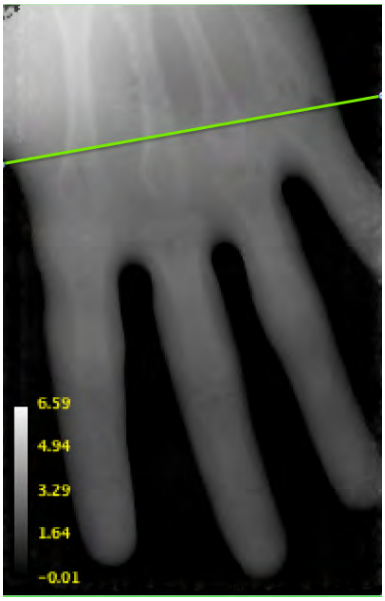


Fig. 1 – First radiograph of a hand phantom with 0.5 mm pixels (scale in cm of WEPL). The RSP of bone is only about 50% greater than that of water, resulting in the low contrast between the bones and soft tissue. The line traversing the image corresponds to the image profile analyzed in Fig. 4.

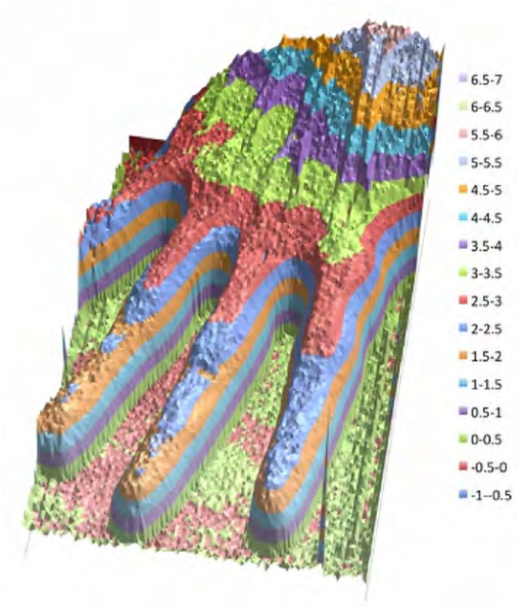


Fig. 2 – Radiograph of a hand phantom (Fig. 1) in terms of water equivalent thickness (WET) calculated from the summed-up stopping power of the phantom. The image shows the varying thickness of the hand and clear structural details. The scale on the right hand side is in cm.

from each pixel's mean relative scattering angle. These cuts are made to exclude events that have very large scattering angles, caused by inelastic nuclear interactions or elastic large angle scattering events inside the phantom. The software also makes cuts in water equivalent path length (WEPL) given by:

$$L = \int_{\ell} \varrho d\ell, \tag{1}$$

where ϱ is the ratio of the stopping power of the material to the stopping power of water (i.e. the RSP) and ℓ defines the path of the proton. These cuts are also made at 3σ from the mean pixel value, and are necessary to insure that erroneously large energy measurements, caused by the coincidence of two or more particles in the calorimeter, are excluded.

II. Energy-Loss Radiography and Water Equivalent Path Length (WEPL)

The quantity of importance for proton treatment planning is relative stopping power (RSP) of protons with respect to water. RSP, or ϱ in Eq. 1, is practically energy independent and is determined mostly by the electron density of the material or tissue.

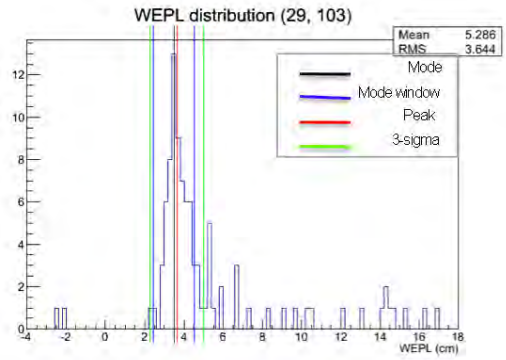
We calibrate the calorimeter response to the integral of the RSP directly. For each pixel, we define a mode window of WEPL that accepts protons within $\pm$ 30% of the mode, or ± 1 cm if 30% is less than 1 cm, and make the appropriate cuts during reconstruction. Fig. 1 is a radiograph of a hand phantom using this energy-loss technique and data reduction process.

The WEPL distribution of protons in each pixel is roughly gaussian, as seen in Fig. 3(a). The distribution is usually skewed to the right (high WEPL) which corresponds to the left-skewed (low-energy) distributions in energy. The protons in the tails are protons that underwent nuclear scattering events. These are the events that we wish to reduce by appropriate cuts.

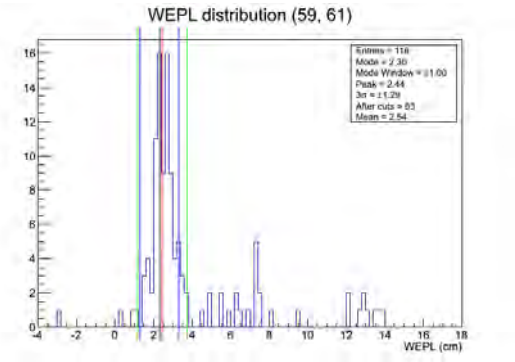
We did find that a significant percentage of pixels contained non-gaussian, or anomolous WEPL distributions. These distributions, as in 3(b), are bimodal and correspond to pixels that lie on the boundary between two materials of different RSP. Currently, the reconstruction algorithm selects the mode that is closest to the mean, and the appropriate cuts are determined based on that value. This, however, ignores valuable information and leads to lower spatial resolution. Methods such as averaging the two modes, or "splitting" pixels have been proposed and have yet to be explored.

An image of the radiographic hand phantom in terms of WEPL (Fig. 1 and 2) was created by plotting values of WEPL for each pixel (in cm). The image clearly depicts the varying thickness of the hand in different places, and shows clear structural details. The agreement between this image and the phantom shows that there is great promise in our technique.

As a further exploration of WEPL, we investigated radiographs of various pixel sizes: 1-mm, 0.5-mm, 0.25-mm. The plots in Fig. 4 illustrate the image profile along the line indicated in Fig. 1 for the various pixel sizes. Fig. 4 shows that as pixel size is systematically decreased, the steepness of the slope of the image profile increases from a relatively shallow incline in the 1-mm (pixel size) plot to a steep rise from 0 to 1 cm of WEPL in the 0.25-mm plot, due to the improved spatial resolution with smaller pixel size. However, decreasing the size of the pixel also increases the amount of spatial noise added to the profile, due to the lower statistics (fewer protons in each pixel). While some regions of the 0.5 mm and the 0.25 mm plots are relatively sharp, other regions are entirely washed out with almost no way to tell what the signal actually is. One can increase the number of protons, but this will increase the dose to the patient, which should be

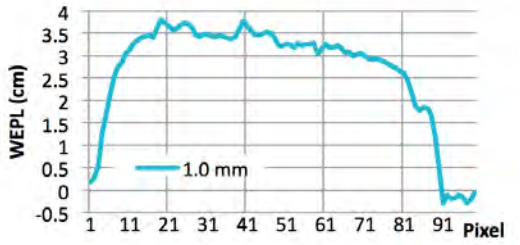


(a)

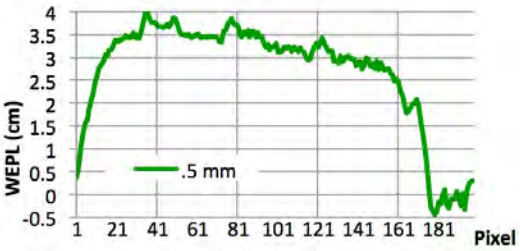


(b)

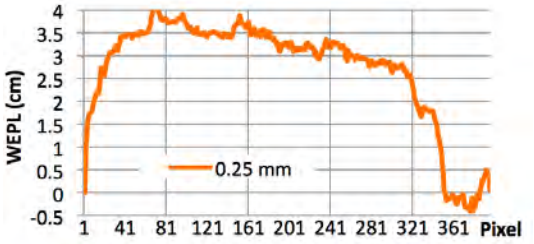
Fig. 3 – Distribution in WEPL for pixels described by the coordinates 3(a) ($v = 29, t = 103$) and 3(b) ($v = 61, t = 59$) before cuts are made. The black line defines the mode of the distribution and the red line defines the mean or "peak" of the distribution. The blue lines indicate the mode window which contains the particles within $\pm 30\%$ of the mode, and provides the distribution on which the 3σ cuts are based. The green lines indicate the cuts made on this specific pixel. Notice the straggling in the large WEPL range. These values correspond to particles that underwent nuclear interactions. Fig 3(a) illustrates an example of a roughly gaussian WEPL distribution. Fig. 3(b) is that for a boundary pixel with a bimodal WEPL distribution.



(a) 1-mm pixels



(b) 0.5-mm pixels



(c) 0.25-mm pixels

Fig. 4 – Image profiles for 1-mm, 0.5-mm and 0.25-mm pixels. Profiles show that as pixel size is decreased from 1-mm (Fig. 4(a)) to 0.5-mm (Fig. 4(b)), the spatial resolution increases (i.e. the details become more clear). Further reducing the pixel size seems only to increase statistical noise in the image (Fig. 4(c)). An ideal pixel size must be found that maximizes spatial resolution while minimizing dose delivered to the patient.

kept as small as possible due to the small risk of secondary cancer. This analysis suggests that, for a given dose, there is an ideal pixel size which will provide a balance between spatial resolution and dose. We have found that at least 20 protons/pixel are required for reasonable statistics.

III. Multiple Coulomb Scattering and Proton Scattering Radiography

The amount that a proton is scattered between its entry and exit from a phantom is proportional to the inverse of its energy and can be described by the Lynch-Dahl approximation for multiple scattering events [4]:

$$\theta = \frac{13.6eV}{\beta cp} z \sqrt{\frac{x}{X_o}} [1 + 0.038 \log \frac{x}{X_o}] \qquad (2)$$

where θ is the width of the Gaussian approximation for angular deflection in a plane, β, p are the velocity and momentum of the proton, respectively, z is the charge of the proton and x/X_o is the thickness of the material traversed in radiation lengths, where we calculate X_o of the material using:

$$\frac{1}{X_o} = \sum \frac{w_j}{X_j} \qquad (3)$$

where the w_j's are the fractions by weight of each element in a given material. The second term in Eq. 2 tends to be small and can thus be ignored for purposes of estimation. Note that this approximation is good only for relatively thin objects (i.e. $10^{-3} < x/X_o < 100$) where the energy and momentum are assumed to be approximately constant. For a thicker phantom, we must account for energy-loss by introducing an integral over x (see Ref. [5] for details).

A scattering radiograph (scale in mrad) is given in Fig. 5. A gaussian distribution of scattering angles in each of the t (vertical) and v (horizontal) planes in each pixel was obtained. The mean v and t angles were determined in each

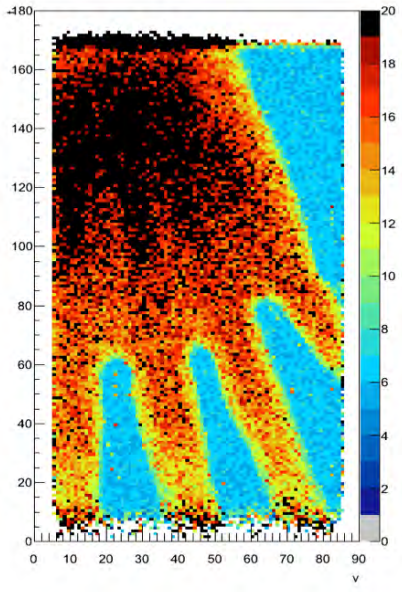


Fig. 5 – This scattering radiograph shows a strong agreement between predicted thickness given by Eq. 2 and the thickness of real materials. Variation in the thickness of the hand is clearly visible. Regions of dark orange and black are those corresponding to thick regions of bone. Blue region in the background corresponds to the scattering due to SSD's alone. Scale is in mrad.

pixel from these distributions. These mean angles were added in quadrature in order to obtain the mean spatial scattering angle, defined as the angle of scattering from the beam axis. Areas of high scattering power, such as bone, were expected to yield greater scattering angles, while protons scattered only by SSDs were expected to have the smallest scattering angle. The scattering angle value was then compared with the expected scattering estimated using Eq. 2.

TABLE I – Densities and radiation lengths of materials commonly encountered in pCT. Data for bone: [6]. Data for tissue, water and silicon: [7]

Material	Density (g/cm^3)	Radiation Length, X_o (g/cm^2)
bone	1.45	16.6
tissue	1.00	38.2
water	1.00	36.1
silicon	2.33	21.8

Table I provides radiation length values for material that we typically deal with in medical proton imaging. For a 200 MeV proton, $\beta = .566$ and $p = 644$ MeV/c, and therefore, by Eq. 2, the scattering due to the four silicon tracker plates (1.6 mm total thickness) is expected to be approximately 5.2 mrad. Comparing this estimate with the background (blue) region in Fig. 5, we find that this estimate agrees well with the image, which depicts scattering of 5-6 mrad due to the SSD's alone.

While the spatial resolution of the scattering radiograph is not as good as with the energy-loss radiograph, one can still observe regions of varying thickness around the edges of the fingers, where the protons traversed only skin and soft tissue (yellow and green region), and in the hand, where the thickest bone exists (black region). The scattering angles correspond to realistic proton path lengths through the hand.

A remarkable aspect of scattering radiography is that the contrast between bone and soft tissue for proton scattering power is, in principle, higher than that of proton stopping

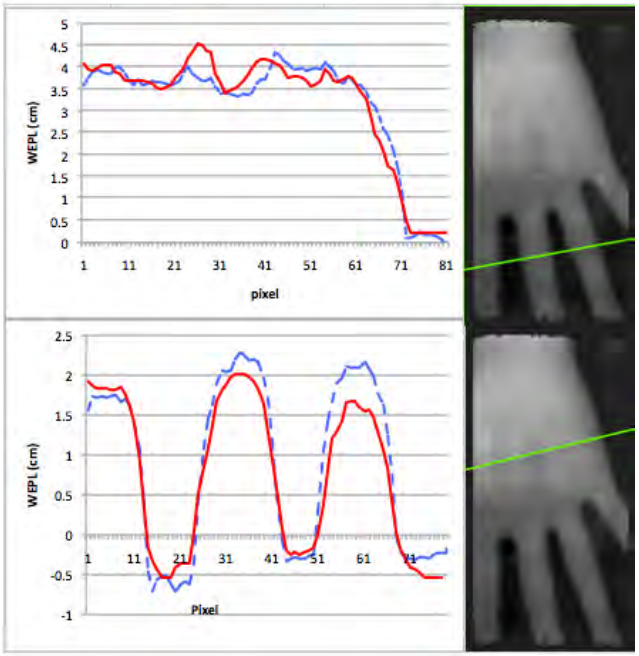


Fig. 6 – Normalizing the scattering radiograph (solid curve) to the energy-loss radiograph (dashed curve), we see roughly the same shape and even some subtle features, however these are quite a bit washed out. The profile slopes of the scattering radiograph in the bottom plots are shallower, indicating reduced spatial resolution.

power. The stopping power of bone is 50% - 80% greater than that of water, but the scattering power of bone is about 2.5 times that of water. Fig. 6 compares two image profiles for the energy-loss radiograph (dashed curve) and the scattering radiograph (solid curve). When the scattering curve is normalized to the energy-loss curve, we find that the general shapes of the two curves of each plot are almost identical, which shows that in this case, regions of greater stopping power are also regions of higher scattering power. The energy-loss curve clearly provides higher spatial resolution, but more importantly, it provides the RSP information required for treatment planning. The scattering radiograph, however, may provide us with higher contrast resolution, since contrast depends upon the difference in material properties of those materials being imaged. Information about the radiation length of the material, X_o can be gleaned from the scattering radiograph and can provide us with the the effective atomic number of the material, Z (which is inversely proportional to the radiation length). The quality and usefulness of this information, however, requires further investigation.

IV. Conclusion

Our proton radiographs demonstrate the new promise of proton imaging (proton radiography and CT) now within reach of becoming a new, potentially low-dose medical imaging modality. This work indicates that choosing an optimal pixel size is important for balanced image quality in terms of low-contrast and spatial resolution. The image profile comparison suggests that scattering radiography may yield sharper edges

(greater contrast) between soft and bone tissue than energy loss radiography, alone. However, this requires further study. Scattering radiography (like x-ray radiography) does provide information about the radiation length of materials which is inversely proportional to the effective atomic number distribution in the tissue. Energy-loss radiography cannot provide this information since stopping power depends only on Z/A which is practically identical for most soft tissues and water, leading to very low contrast. Therefore, scattering radiography will likely have useful applications in proton treatment planning.

ACKNOWLEDGMENT

We acknowledge contributions from Y. Censor (The University of Haifa (Israel)), S. Penfold (University of Wollongong (Australia)), and R. Davidi (Stanford University).

This research in proton CT is supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and the National Science Foundation (NSF), award Number R01EB013118, the U.S. Department of Defense Prostate Cancer Research Program, award No. W81XWH-12-1-0122, and the United States - Israel Binational Science Foundation (BSF). The content of this poster is solely the responsibility of the authors and does not necessarily represent the official views of NIBIB, NIH, NSF and DOD.

The proton imaging detectors were built at UCSC and Northern Illinois University with support from the U.S. Department of Defense Prostate Cancer Research Program, award No. W81XWH-12-1-0122 and the Department of Radiation Medicine at LLU.

REFERENCES

[1] R. Schulte. A Status Update on Proton Imaging for Applications in Medicine. IEEE Nuclear Science Symposium and Medical Imaging Conference. Anaheim. 30 Oct. 2012.

[2] Tami Freeman. Will protons gradually replace photons? Medical Physics Web. 22 August, 2012. http://medicalphysicsweb.org/cws/article/research/50584.

[3] S. Penfold, Image Reconstruction and Monte Carlo Simulations in the Development of Proton Computed Tomography for Applications in Proton Radiation Therapy PhD thesis Univ. of Wollongong, 2010.

[4] Particle Data Group. *Review of Particle Physics 2008*. Section 27.3 "Multiple scattering through small angles." p. 271. 2008.

[5] Schulte R.W., Penfold S. N., Tafas J. T., and Schubert K. E., A maximum likelihood proton path formalism for application in proton computed tomography, Med. Phys. 35: 4849-4856, 2008.

[6] D. C. Williams, Phys. Med. Biol. 49 (2004) 2899-2911.

[7] Particle Data Group. Atomic and Nuclear Properties of Materials for more than 300 Materials. Accessed 15 November, 2012. http://hepdata.cedar.ac.uk/lbl/2011/AtomicNuclearProperties/index.html

[8] R. F. Hurley, V. A. Bashkirov, R. W. Schulte, A. J. Wroe, A. Ghebremedhin, P. Koss, B. Patyal, H. Sadrozinski, V. Rykalin, G. Coutrakon. Water-equivalent path length calibration of a prototype proton CT scanner. Med Phys. 2012 May;39(5):2438-46.

Parallel Algorithms for Intensity Modulated Proton Radiation Therapy

PIs: Reinhard Schulte, Yair Censor, Ran Davidi, John DeMarco, Keith E. Schubert
Graduate Students: Aarohi Padhye, Tai Dou

1 Scope

The purpose of this document is to describes the scientific background of a multi-institutional project on intensity modulated proton therapy, including mathematical formulations pertinent references relevant to this project.

2 Background

2.1 Principles of IMpRT

Intensity modulated proton radiation therapy and radiosurgery, short IMpRT and IMpRS, are evolving techniques for highly conformal dose delivery to tumor or other targets in close proximity to sensitive and critical organs at risk. IMpRT is delivered in several dose fractions, while IMpRS is delivered in as a single dose or a few (up to 5) dose fractions applying stereotactic techniques. The underlying principle of these techniques is to aim at the target from many different directions (either in 2D or 3D) with multiple narrow proton beams, or *pencil beams*, and to modulate the intensity (or fluence) of each beam, taking into account whether they pass through critical organs at risk or not. The most important characteristic of a proton beam is that it delivers a low dose in the initial part of the beam followed by a rapid increase of dose, leading to a dose peak (the Bragg peak) and a rapid distal dose fall-off to zero dose behind the Bragg peak. The Bragg peak is placed inside the target at a given *beam aiming point*. Note that several pencil beams sharing the same central axis can be "stacked" in beam direction, and this arrangement may be called a *beamlet*.

The starting point of each IMpRT/RS calculation is a digital model of the patient volume of interest, e.g., the patient's head, usually provided by a computed tomography (CT) scan. A head CT scan consists of about 200 slices

of 1-2 mm thickness and each slice is organized into a matrix of 512×512 image pixels. In 3D, this creates a digital space comprised of the order of 50 million voxels. Each voxel has material properties that are needed to calculate the proton dose delivered by the different proton pencil beams.

In practical applications, one generates a generic pencil beam dose model for a unit-intensity proton beam in water and scales the distance between the entry point of a proton beam into the object and the beam aiming point by multiplying the intersection length of each voxel with the so-called relative stopping power (RSP) with respect to water. This information is provided by converting the numbers provided by the CT scan (Hounsfield units) to RSP, using a HU-to-RSP calibration curve. In the future, the RSP of voxels will be directly reconstructed from a proton CT (pCT) scan. Knowing the central beam axis dose as a function of depth in water, one can then assign the correct dose of the unit-intensity proton pencil beam to each voxel on the central beam axis. Similarly, knowing the lateral dose fall-off at each depth, one can calculate the correct dose for each off-axis voxel based on its orthogonal distance from the beam axis.

Given a distribution of the intensities of in the limit, continuously spaced proton pencil beams directed at the target, one can calculate the resulting dose distribution in the voxels of the object using a proton dose operator $\mathfrak{D}$ that mathematically connects the two quantities. Often times, the chosen intensities do not result in a satisfactory dose distribution, i.e., one that meets the dose constraints dictated by the radiosensitivity of the tumor and the organs at risk. In general, one wants the target dose to exceed some minimum value and the dose in organs at risk not to exceed a maximum value that can lead to serious complications. Therefore, it is better to "prescribe" a dose distribution selected from a subset in a continuum of possible dose distributions that meet the clinical requirements and then to find a fluence distribution that that will lead to a dose distribution that is a member of this "solution" subset. As we will see below, the solution of such an "inverse" treatment planning problem can be found mathematically by formulating a discrete mathematical model of IMpRT that can be solved, in principle.

2.2 The discrete model of IMpRT

In the absence of a closed-form analytic representation of the proton dose operator $\mathfrak{D}$ that calculates the dose distribution given a the fluence of an continuum of proton pencil beams, and, therefore, the absence of such a

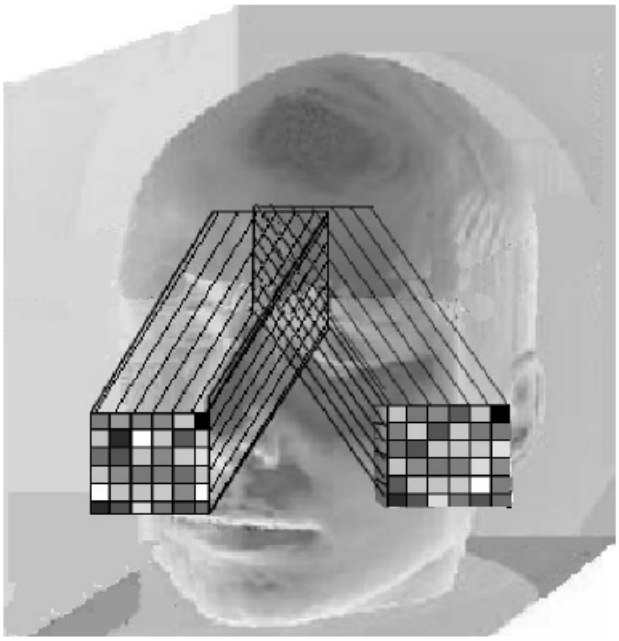


Figure 1: Two IMpRT beams from different directions. Variable shades of gray correspond to different fluences (number of protons per area). Note that each square in the beam cross section can be occupied by more than one proton pencil beam, making up a *beamlet*, each with a different Bragg peak depth and intensity.

presentation of its inverse operator $\mathfrak{D}^{-1}$, one must resort to a fully-discretized model of the problem. The term *full* in "fully-discretized model" refers to the fact that both the external proton radiation field and the patient volume are discretized, leading to a problem formulated in a finite-dimensional vector space. To do this we divide the beam's cross-section into a finite rectangular grid of squares and the beam angles into discrete angular steps separated by a constant interval, which may be chosen differently for each IMpRT treatment plan (see Figure 1). Further, we discretize the proton energy into steps, such that the proton Bragg peaks, i.e., the dose maximum of a proton pencil beam, are located at well-defined discrete aiming points within the patient volume. Each proton pencil beam is thus assigned a discrete direction and a discrete energy.

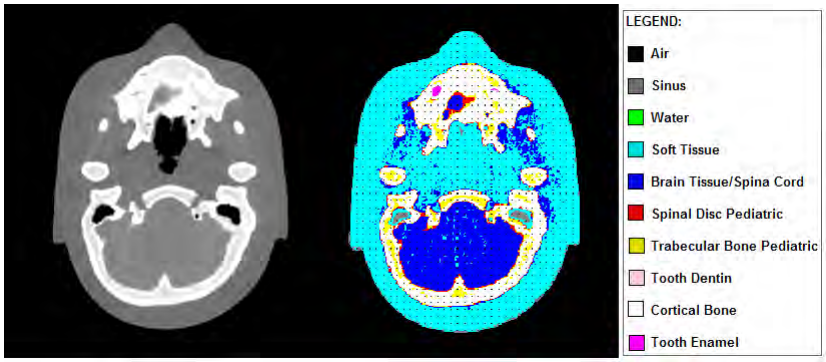


Figure 2: Example of a CT head section before (left) and after conversion to a color-coded image that gives each voxel a tissue assignment (right).

Figure 2 (left) shows a representative two-dimensional (2D) cross-section through the object. In a contiguous set of cross-sections, the treatment planner defines a set of voxels that belong to the target. Other voxels sets may be defined that are assigned to an organ at risk, e.g., the brainstem, or other normal tissue regions, such as brain and skull bone. In order to simplify the image segmentation process and to calculate the dose of unit-intensity beams, each image of the CT data set needs to be processed in order to assign a given tissue type to each voxel based on the CT (HU or RSP) value. This is shown in Figure 2 (right).

2.3 Mathematical formulation of the discrete IMpRT model

The patient volume Ω is divided into a discrete grid of voxels the centers of which are the desired dose calculation points. These are represented by the family of triplets of 3D coordinates $\{(r_j) \mid j = 1, 2, \ldots, J\}$. Further, we define a discrete number of proton pencil beams by their entry direction unit vectors $\{v_i \mid i = 1, 2, \ldots, I\}$. and aiming point $\{(\hat{r}_i) \mid i = 1, 2, \ldots, I\}$.

Let a_{ij} be the dose deposited at the jth grid point (r_j) in the patient volume Ω due to the ith pencil beam $(\hat{r}_i, v_i)$ of unit proton fluence and define the I-dimensional vector $a^j = (a_{ij})_{i=1}^I$ for $j = 1, 2, \ldots, J$. Let x_i denote the actual (yet unknown) fluence of the ith pencil beam $(\hat{r}_i, v_i)$ and define the I-dimensional vector $x = (x_i)_{i=1}^I$ which is unknown vector of all pencil

beams' fluences that should deliver the required dose to the patient volume Ω. Finally, let $\overline{d}_j$ and $\underline{d}_j$ be an upper-bound and a lower-bound, on the permitted or required, respectively, dose in the jth grid point (r_j) in the patient volume Ω.

With these notions we can define discrete forward and inverse problems of IMpRT as follows.

The discrete forward problem of IMpRT: Given a patient volume Ω, whose physical properties are known, and a discretized (into I proton pencil beams) external proton radiation field $\{(\hat{r}_i, v_i) \mid i = 1, 2, \ldots, I\}$, along with a proton pencil beams intensity vector x, find the discretized proton dose distribution function $D(r_j)$ for all $(r_j) \in \Omega$.

This discrete forward problem can be solved if all I-dimensional vectors $a^j = (a_{ij})_{i=1}^I$ for $j = 1, 2, \ldots, J$, are known to us, e.g., by having been pre-calculated by a forward problem solver computer package. In that case, denoting $d_j = D(r_j, \theta_j)$ for all $j = 1, 2, \ldots, J$, we just need to calculate

$$\sum_{i=1}^{I} a_{ij} x_i = d_j, \;\; j = 1, 2, \ldots, J. \tag{1}$$

The J-dimensional vector $d = (d_j)_{j=1}^J$, whose components are the discretized proton dose distribution function $D(r_j)$ values, is called a dose vector.

The discrete inverse problem of IMpRT: Given are a patient volume Ω, whose physical properties are known, an upper-bound dose vector $\overline{d} = (\overline{d}_j)_{j=1}^J$ and a lower-bound dose vector $\underline{d} = (\underline{d}_j)_{j=1}^J$, on the permitted and required, respectively, doses at the grid points $\{(r_j, \theta_j) \mid j = 1, 2, \ldots, J\}$ in the patient volume Ω. Find a proton pencil beams fluence vector x such that

$$\underline{d}_j \leq \sum_{i=1}^{I} a_{ij} x_i \leq \overline{d}_j, \text{ for all } j = 1, 2, \ldots, J \text{ and } x_i \geq 0 \text{ for all } i = 1, 2, \ldots, I. \tag{2}$$

This formulation of the discrete inverse problem of IMpRT does not aim at a proton pencil beams fluence vector x that will deposit a fixed prescribed dose in each voxel but rather calls for a solution of that is called in optimization theory the solution of a *linear feasibility problem*. The term "feasibility" refers here to the fact that no exogeneous objective function is set up for optimization but rather any point in the feasible set $\{x \in R^I \mid \underline{d}_j \leq \sum_{i=1}^I a_{ij} x_i \leq \overline{d}_j$, for all $j = 1, 2, \ldots, J\}$ will be "acceptable" by the treatment planner.

This feasibility approach to setting up the discrete inverse problem has its roots in some early papers on radiation therapy treatment planning where the term IMRT was even not used, see [1, 5, 6, 7].

The J individual linear feasibility constraints of (2) can be grouped according to volumes of interest in the patient volume Ω.

3 Scientific Tasks

The graduate students will support the development of the Geant4 beam libraries (Tai) of a GPU-based platform for testing new algorithms (Aarohi) that solve the discrete inverse problem of IMpRT. A brief summary and motivation of each task is provided below.

3.1 Identification and Storage of Volumes of Interest

The starting point for IMpRT calculations is a CT image set, as described in the background section. The images are in DICOM format, which is a standardized medical imaging format. Within this image set, the physician defines the boundaries of volumes of interest (VOIs) in pertinent slices. This task is usually performed with a commercial computer treatment planning program. The program provides the tools to draw the VOI regions in individual slices and to display them as overlay on the original CT images. The program also outputs a standardized DICOM RT structure set that contains the geometrical information of the VOI boundaries.

The students will import the DICOM image data as well as the DICOM RT structure set file into a Matlab program. Matlab interprets the image set as a hypermatrix of 512×512 matrices that contain the numerical voxel values (in HU) as elements. The students need to develop software that stores the information of which voxel indices belong to each VOI in condensed sparse row format. This information will later be needed to assign the individual linear feasibility constraints to the correct voxels according to their assignment to VOIs.

3.2 CT Image Segmentation

For the forward dose calculation, it is necessary to assign different regions in the CT images to different materials, in this case to different human tissues.

The simplest way to do this is to define HU intervals and assign them to a specific tissue, as shown in Table 1, which is the conversion table for a pediatric head phantom with 9 different tissue types. However, as can be seen in Figure 2, this assignment is not always perfect due to the presence of noise and artifacts in the CT images.

The students will develop a program that finds the boundaries between different tissue regions and will assign voxels inside these boundaries to the correct materials. The voxel volumes are generally small enough to ignore partial volume effects, i.e., individual voxels will be assigned only one material type.

Table 1: *Tissue categorization according to HU value.*

HU Interval	Tissue
$[-1000, -800)$	air
$[-800, -700)$	sinus
$[-700, 40)$	soft tissue
$[40, 90)$	brain
$[90, 150)$	spinal disc
$[150, 200)$	trabecular bone
$[200, 1000)$	cortical bone
$[1000, 2000)$	tooth dentin
≥ 2000	tooth enamel

3.3 Interface to Geant4 Program Output

Geant4 is a toolkit written in C++ code that performs radiation transport calculations. The students will obtain a source model for the Geant4 forward dose calculations. Geant4 will provide a dose model for a standard library of proton pencil beams in water with energies between 60 MeV and 160 MeV in 10 MeV steps. The students will also develop a program that creates an array of beaming aiming points for each of a set of beam directions. The program will then calculate the water equivalent depth of each point by multiplying beam axis intersection lengths by the assigned relative stopping power (RSP)

for each voxel on the central beam path. In addition the water-equivalent distance of voxels lateral to the central beam axis will need to be calculated.

4 Potential for Publications

The development of the beam libraries and computing platform will be reported by the students at scientific meetings in computer science and medical physics fields. This will typically lead to abstracts and conference papers with the students being the first author (depending on the type of conference). The aim is to also publish a series of original papers with on solution algorithms developed by Ran and Yair with postdoc Ran Davidi as first author and students as co-authors. There could well be other original papers written by students on spin-off projects resulting from the main project.

References

[1] M.D. Altschuler and Y. Censor, Feasibility solutions in radiation therapy treatment planning, in: *Proceedings of the Eighth International Conference on the Use of Computers in Radiation Therapy*, IEEE Computer Society Press, Silver Spring, MD, USA, 1984, pp. 220–224.

[2] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall International, Englwood Cliffs, NJ, USA, 1989.

[3] D. Butnariu, R. Davidi, G.T. Herman, and I.G. Kazantsev, Stable convergence behavior under summable perturbations of a class of projection methods for convex feasibility and optimization problems, *IEEE Journal on Special Topics in Signal Processing* **1** (2007), 540–547.

[4] A. Cegielski, *Iterative Methods for Fixed Point Problems in Hilbert Spaces*, Lecture Notes in Mathematics 2057, Springer-Verlag, Berlin, Heidelberg, Germany, 2012.

[5] Y. Censor, W.D. Powlis and M.D. Altschuler, On the fully discretized model for the inverse problem of radiation therapy treatment planning,

in: *Proceedings of the Thirteenth Annual Northeast Bioengineering Conference*, (K.R. Foster, Editor), Institute of Electrical and Electronics Engineers (IEEE), New York, NY, USA, Vol. **1** (1987), pp. 211–214.

[6] Y. Censor, M.D. Altschuler and W.D. Powlis, On the use of Cimmino's simultaneous projections method for computing a solution of the inverse problem in radiation therapy treatment planning, *Inverse Problems* **4** (1988), 607–623.

[7] Y. Censor, M.D. Altschuler and W.D. Powlis, A computational solution of the inverse problem in radiation-therapy treatment planning, *Applied Mathematics and Computation* **25** (1988), 57–87.

[8] Y. Censor and S.A. Zenios, *Parallel Optimization: Theory, Algorithms, and Applications*, Oxford University Press, New York, NY, USA, 1997.

[9] Y. Censor, W. Chen, P.L. Combettes, R. Davidi and G.T. Herman, On the effectiveness of projection methods for convex feasibility problems with linear inequality constraints, *Computational Optimization and Applications* **51** (2012), 1065–1088.

[10] Y. Censor and J. Unkelbach, From analytic inversion to contemporary IMRT optimization: Radiation therapy planning revisited from a mathematical perspective, *Physica Medica: European Journal of Medical Physics* **28** (2012), 109–118.

[11] Y. Censor, T. Elfving, N. Kopf and T. Bortfeld, The multiple-sets split feasibility problem and its applications for inverse problems, *Inverse Problems* **21** (2005), 2071–2084.

[12] Y. Censor, R. Davidi and G.T. Herman, Perturbation resilience and superiorization of iterative algorithms, *Inverse Problems* **26** (2010), 065008.

[13] Y. Censor and T. Elfving , A multiprojection algorithm using Bregman projections in product space, *Numerical Algorithms* **8** (1994), 221–239.

[14] Y. Censor, T. Bortfeld, B. Martin and A. Trofimov, A unified approach for inversion problems in intensity-modulated radiation therapy, *Physics in Medicine and Biology* **51** (2006), 2353–2365.

[15] Y. Censor, R. Davidi, and G.T. Herman, Perturbation resilience and superiorization of iterative algorithms, *Inverse Problems* **26**, 065008, (2010).

[16] R. Davidi, G.T. Herman, and Y. Censor, Perturbation-resilient block-iterative projection methods with application to image reconstruction from projections, *International Transactions on Operational Research* **16** (2009), 505–524.

[17] R. Davidi, R.W. Schulte, Y. Censor and L. Xing, Fast superiorization using a dual perturbation scheme for proton computed tomography, *Transactions of the American Nuclear Society* **106** (2012), 73–76.

[18] G.T. Herman, E. Garduño, R. Davidi and Y. Censor, Superiorization: An optimization heuristic for medical physics, *Medical Physics* **39** (2012), 5532–5546.

[19] A. Lomax, Intensity modulated methods for proton therapy, *Physics in Medicine and Biology* **44** (1999), 185–205.

[20] T. Nikazad, R. Davidi, and G.T. Herman, Accelerated perturbation-resilient block-iterative projection methods with application to image reconstruction, *Inverse Problems* **28**, 035005, (2012).

Ran Davidi[1], Yair Censor[2], Sarah Geneser[3], Reinhard Schulte[4] and Lei Xing[1]

[1]Department of Radiation Oncology, Stanford University School of Medicine, Stanford, CA 94305
[2]Department of Mathematics, University of Haifa, Mount Carmel, Haifa 3190501, Israel
[3]Department of Radiation Oncology, University of California San Francisco, San Francisco, CA 94143
[4]Department of Radiation Medicine, Loma Linda University Medical Center, Loma Linda, CA 92354

Introduction and Objective

Computationally demanding numerical minimization techniques are often used in IMRT treatment planning, but the commonly employed cost functions and corresponding solution approaches are not necessarily the most appropriate for achieving the desired dose behavior. This disconnect occurs because minimal solutions to current cost function formulations are not guaranteed to provide the necessary dose coverage, conformality, or homogeneity. Therefore, the considerable computational cost associated with some of these minimization techniques may not be justified.

We propose a novel superiorization approach that substantially improves computational tractability by producing a solution with reduced, but not necessarily minimal, value of the defined cost function that is guaranteed to satisfy the given IMRT planning constrains. Superiorization is a new paradigm that can be viewed as lying in-between feasibility-seeking for the dose constraints and full-fledged constrained minimization of the cost function subject to these constraints. This method is based on the discovery that many feasibility-seeking algorithms are perturbation-resilient, and superiorization proactively steers the feasibility-seeking projection method towards a feasible solution of the dose constraints with a reduced, but not necessarily minimal, cost function value.

The superiorization method produces "superior feasible solutions" and can replace current IMRT constrained minimization methods, potentially leading to shorter computational times and improved dose distributions.

Materials and Methods

We model a given IMRT problem as a linear feasibility one, by formulating the constraints into upper- and lower-bounds vectors. The bounds are set and depend whether the constrained volume is a target or an organ at risk (OAR). The bounds reflect the dose acceptance criteria, which are determined by the treating physician and reflect generally accepted dose guidelines. A projection method that is perturbation-resilent aims at solving this linear feasibility system of hyperslabs constraints. This feasibility-seeking algorithm uses the resiliency to perturbations to steer the iterates to a superior feasible point with respect to an objective function. Here we use ART for inequality constraints and total variation (TV) [2] of the beam intensity space as the objective function.

The complete superiorization algorithm is provided in the pseudocode (Fig. 1) and is based on [2].

How superiorization works: The algorithm starts from an arbitrary point. In lines 7-17 it perturbs the current point N times. A nonascending vectoris computed in line 8 and the perturbation is performed in line 13 with some step size β. The value of the objective function Φ is assessed in line 14 to make sure that the perturbation superiorized (obtained a lower value) the objective function compared to the previous point. At the end of the N perturbation steps, the projection method is applied and a new point is obtained. The process repeats until the acceptance dose criteria is met.

IMRT plan: The anonymized pelvic planning CT of a prostate cancer patient was employed for the IMRT treatment planning using the proposed method. Seven equispaced fields were used for targeting the PTV. The dose constraints were set using the RTOG 0815 randomized trial protocol [3].

Results

We have initially tested this new approach by comparing the TV-superiorization algorithm with an otherwise identical algorithm that aimed at only satisfying the dose constraints without applying superiorization. We performed two experiments with different starting conditions. For the first experiment, we started the algorithm with the zero vector of dose weights and for the second experiment all dose weights were given the value 10. Table 1 summarizes the results for the two experiments and in Fig. 2 we present the associated DVH curves. For the first experiment, the TV-superiorization produced a solution that met the acceptance criteria after 12 iterations whereas the conventional algorithm was not able to reach an acceptable solution after this number of iterations. For the second experiment, the superiorization algorithm reached an acceptable solution even faster, i.e., after 7 iterations, and the conventional algorithm again failed some of the acceptance criteria after this number of iterations.

Table 1: RTOG 0815 acceptance criteria and results of the two experiments described in the Results section

Acceptance criteria	Exp 1 with superiorization	Exp 1 without superiorization	Exp 2 with superiorization	Exp 2 without superiorization
PTV min allowed dose (95% of prescribed dose) is 75.24 Gy	75.24 Gy	56.13 Gy	77.80 Gy	76.15 Gy
PTV max allowed dose: 84.74 Gy	84.69 Gy	89.42 Gy	84.71 Gy	87.63 Gy
Rectum – No more than 50% volume receives dose that exceeds 60.00 Gy	34.50 %	8.50 %	36.90 %	40.50 %
Rectum – max dose	82.64 Gy	82.71 Gy	84.09 Gy	87.25 Gy

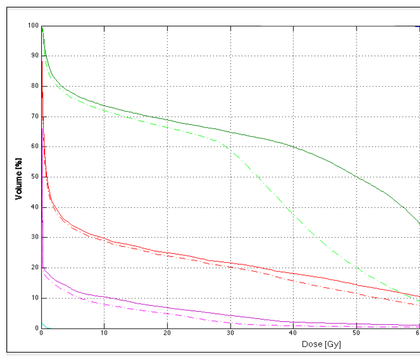
Conclusions

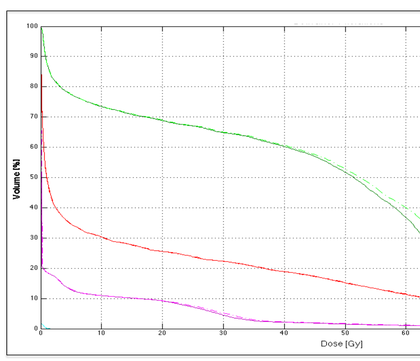
Our proposed method successfully produced conformal solutions that met the acceptance criteria while that an otherwise identical algorithm without superiorization failed to do so with the same number of iterations. Future work will assess the computational gain of the superiorization method compared to a conventional one and investigate the utility of it for a computationally more complex problems such as Volumetric Modulated Arc Therapy (VMAT).

Fig. 1: Pseudocode of the Superiorization Algo

1. **set** $k = 0$
2. **set** $y^k = y^0$
3. **set** $\ell = -1$
4. **repeat**
5. **set** $n = 0$
6. **set** $y^{k,n} = y^k$
7. **while** $n<N$
8. **set** $v^{k,n}$ to be a nonascendin
9. **set** *loop=true*
10. **while** *loop*
11. **set** $\ell = \ell + 1$
12. **set** $\beta_{k,n} = \eta_\ell$
13. **set** $z = y^{k,n} + \beta_{k,n} v^{k,n}$
14. **if** $\phi(z) \leq \phi(y^k)$ **then**
15. **set** $n = n + 1$
16. **set** $y^{k,n} = z$
17. **set** *loop* = *false*
18. **set** $y^{k+1} = \boldsymbol{A}_C(y^{k,N})$
19. **set** $k = k + 1$

Fig. 2: Dose Volume Histograms (DVH) of the lines represent the algorithm with TV-superi represent no superiorization). The first (top) to second (bottom) took 7 iterations. Exact numbe





References

[1] G.T. Herman and A. Lent, A family of iterative quadratic optimization algorithms for pairs of inequalities with application in diagnostic radiology, Mathematical Programming

[2] G.T. Herman, E. Garduño, R. Davidi , Y. Censor, Superiorization: An optimization heuristic for medical physics, Med. Phys. 39, 5532–46, 2012.

[3] Radiation Therapy Oncology Group: RTOG 0815 Protocol Information. http://www.rtog.org/ClinicalTrials/ProtocolTable/StudyDetails.aspx?study=0815 Updated: 5/24/2013

Acknowledgement: This work is supported by the U.S. Department of Defense Prostate Cancer Research Program Award No. W81XWH-12-1-0122, by grant number 2009012 from the United State Binational Science Foundation (BSF) and by the U.S. Department of the Army Award No. W81XWH-10-1-0170.

INFORMATION AND CONTACT

Organisation

There will be plenary talks on projection methods by invited researchers and presentations on industry problems and current solutions by ITWM researchers. The aim of the workshop is to discuss applications of state-of-the-art projection methods to real-world problems. The workshop is organized by the Department of Optimization, Fraunhofer ITWM.

Participation

Open to all interested researchers, faculty and students. Participation is free.

Venue

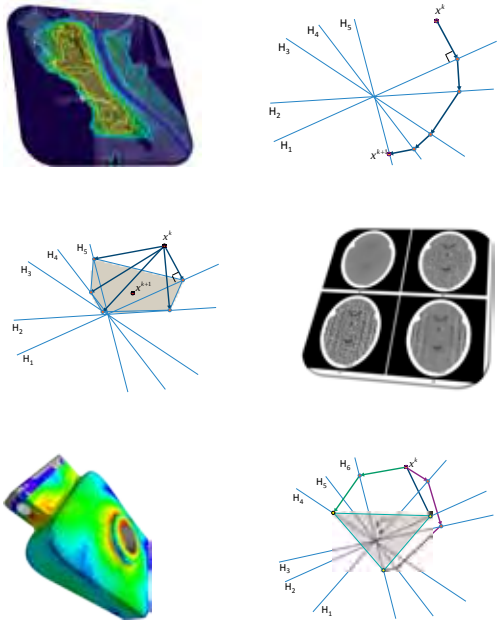
Fraunhofer Institute for Industrial
Mathematics ITWM
Fraunhofer-Platz 1
67663 Kaiserslautern

Directions: www.itwm.fraunhofer.de

For further information please contact

Aviv Gibali
Phone +49 631 31600-4707
Email aviv.gibali@itwm.fraunhofer.de
or
Philipp Süss
Phone +49 631 31600-4295
Emai philipp.suess@itwm.fraunhofer.de

WORKSHOP



Projection Methods – Theory and Practice

June 19 – 20, 2013
Fraunhofer Institute for Industrial Mathematics ITWM
Kaiserslautern, Germany

FELIX KLEIN
ZENTRUM FÜR
MATHEMATIK

Projection Methods

Projection methods for solving large-scale systems of equations are very attractive for various practical applications. Aside from having desirable convergence properties, they exhibit robustness against perturbations (data, algorithm parameters, etc.). Moreover, they are typically easy to understand and implement and most often lend themselves to parallelization.

Applications

Projection methods have been developed for solving convex feasibility problems, in particular, linear and nonlinear system of equations and inequalities. They have found applications in Material Science, Radiation Therapy Treatment Planning, Computed Tomography and Image Reconstruction, and more.

Invited Speakers

- Adi Ben-Israel, Rutgers Business School, USA
- Andrzej Cegielski, University of Zielona Góra, Poland
- Charles L. Byrne, University of Massachusetts Lowell, USA
- Gabor T. Herman, City University of New York, USA
- Ran Davidi, Stanford University, USA
- Tommy Elfving, University of Linköping, Sweden
- Yair Censor, University of Haifa, Israel

June 19, 2013

08.30 – 09.00	registration and coffee	
09.00 – 12.00	45 min. invited	**Yair Censor**
	25 min. break	
	45 min. invited	**Andrzej Cegielski**
	20 min. break	
	45 min. Fraunhofer	**Michael Bortz**
12.00 – 14.00	lunch break and discussions	
14.00 – 17.00	45 min. invited	**Gabor T. Herman**
	25 min. break	
	45 min. invited	**Ran Davidi**
	20 min. break	
	45 min. Fraunhofer	**Philipp Süss**

June 20, 2013

08.30 – 09.00	coffee	
09.00 – 12.00	45 min. invited	**Charles L. Byrne**
	25 min. break	
	45 min. invited	**Tommy Elfving**
	20 min. break	
	45 min. Fraunhofer	**Jan Schwientek**
12.00 – 14.00	lunch break and discussions	
14.00 – 15.00	45 min. invited	**Adi Ben-Israel**
	15 min. break	
15.00 – 17.00	open forum and discussions	

Fraunhofer
ITWM

FRAUNHOFER INSTITUTE FOR INDUSTRIAL MATHEMATICS ITWM

PROJECTION METHODS THEORY & PRACTICE

JUNE 19-20, 2013

FRAUNHOFER ITWM, KAISERSLAUTERN, GERMANY

PROJECTION METHODS

Projection methods for solving large-scale systems of equations are very attractive for various practical applications. Aside from having desirable convergence properties, they exhibit robustness against perturbations (data, algorithm parameters, etc.). Moreover, they are typically easy to understand and implement and most often lend themselves to parallelization.

APPLICATIONS

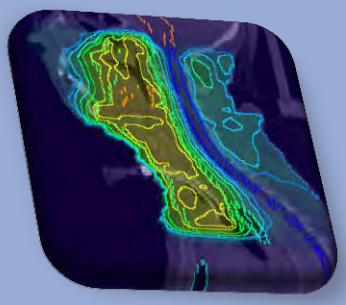
Projection methods have been developed for solving convex feasibility problems, in particular, linear and nonlinear system of equations and inequalities. They have found applications in Material Science, Radiation Therapy Treatment Planning, Computed Tomography & Image Reconstruction, and more.

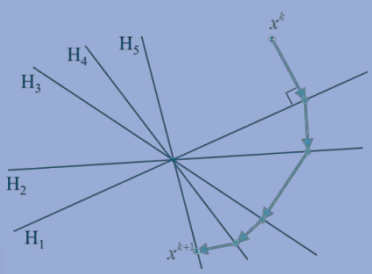
ORGANIZATION

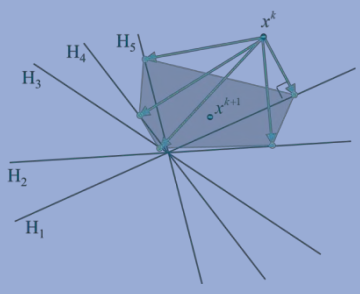
The workshop will be held at the Fraunhofer Institute in Kaiserslautern, Germany. There will be plenary talks on projection methods by invited researchers, presentations on industry problems and current solutions by ITWM researchers, as well as open workshops and discussions on applications of state-of-the-art projection methods to real-world problems.

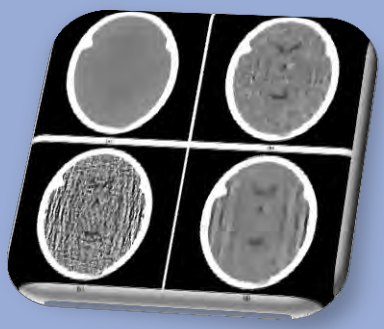
INVITED SPEAKERS

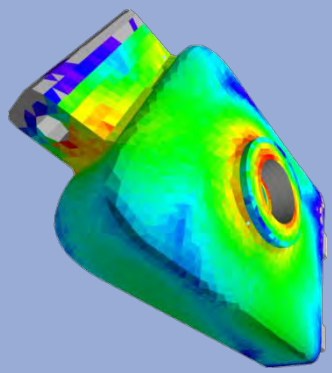
- Charles L. Byrne, USA
- Andrzej Cegielski, Poland
- Yair Censor, Israel
- Ran Davidi, USA
- Tommy Elfving, Sweden
- Valentin R. Koch, Canada
- Akhtar A. Khan, USA

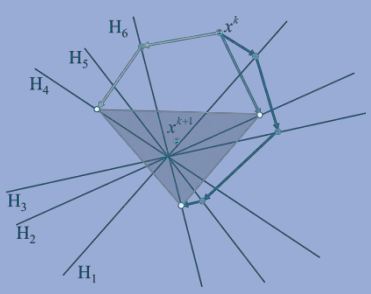












Superiorization of Projection Methods Applied to Radiation Therapy Treatment Planning

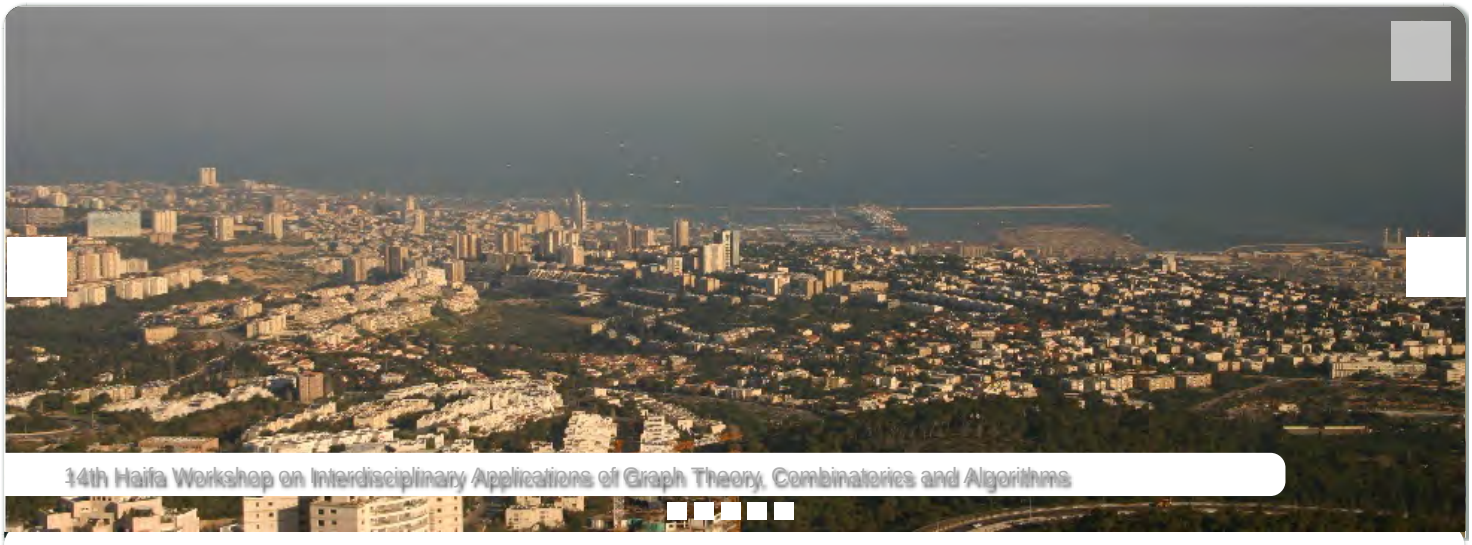
Ran Davidi

Department of Radiation Oncology, Stanford University, Stanford, CA, USA

In Radiation Therapy Treatment Planning applications, such as IMRT and VMAT, the main objective is to deliver precise radiation doses to a malignant tumor while sparing the surrounding normal tissues. Minimization techniques are often used as the main tool for the inverse treatment planning. They commonly employ the minimization of an objective function subject to a set of dose constraints. These methods often produce a solution that is not guaranteed to provide the necessary dose coverage and conformality that is required for a successful treatment and are executed with a high computational demand. Satisfying only the dose constraints on the other hand, can be implemented quite efficiently using projection methods, however, these methods are lacking the machinery of a reduction of an objective function value.

Superiorization is a new paradigm that aims to bridge the gap between these two approaches of optimization and feasibility-seeking methods. It utilizes the fact that many projection methods are perturbation resilient and steers the process in the direction of a lower objective function value while satisfying the dose constraints of the problem. We present how superiorization can be applied to the inverse planning problem in radiation therapy and demonstrate its efficacy on prostate patient data.

Acknowledgement: This work is supported by the U.S. Department of Defense Prostate Cancer Research Program Award No. W81XWH-12-1-0122.



Main Menu

About Us
Events
2014
2013
2012
2011
2010
2009
2008
2007
2006
2005
2004
Video
Call for Papers, Visitors, Post Doc, Workshops, Proposals
Scholarships
Photo Gallery
People
Research Laboratories
Partners
The Hecht Project
Publications
Annual Reports

Related Links

Etgar Program (Hebrew)
The Israeli AAI
CRI Visitors FAQ
CS Home Page

Projection Methods in Feasibility, Superiorization and Optimization

Written by Daniel

Details: *| Published: 15 December 2013*

The Center for Mathematics and Scientific Computation (CMSC)

and

The Caesarea Rothschild Institute (CRI) for Interdisciplinary Applications of Computer Science

cordially invite you to

A one-day marathon on

"Projection Methods in Feasibility, Superiorization and Optimization"

Thursday, December 19, 2013

The event is under the joint auspices of the **Center for Mathematics and Scientific Computation (CMSC)** and the **Caesarea Rothschild Institute (CRI) for Interdisciplinary Applications of Computer Science** at the University of Haifa, and is organized by Yair Censor.

The talks will be given at the University of Haifa, in the Abraham and Rachel Kluger Education and Sciences Building,

6th Floor, Room number 665,

Participation is free but let us please know in advance if you intend to come via the following contact information.

For entrance and parking permit to the campus please write to: Ms. Danielle Friedlander, Administrative Coordinator, the CRI, Phones: 04-8288337 (office), 050-9777907 (cell), e-mail: dfridl1@univ.haifa.ac.il

For all other questions please contact: Prof. Yair Censor, Phones: 04-8240837 (office), 050-8816144 (cell), e-mail: yair@math.haifa.ac.il

Program

08:45 -- 9:00 Getting together

09:00 -- 09:05 Yair Censor: Opening comments

09:05 -- 10:00 Andrzej Cegielski: Methods for the split common fixed point problem

10:00 -- 10:30 Coffee break

10:30 -- 11:30 Simeon Reich: Porosity and the bounded linear regularity property

11:30 -- 12:30 Ran Davidi: Superiorization of projection methods and their use in medical applications

12:30 -- 14:00 Lunch break

14:00 -- 15:00 Aviv Gibali, Projection-based scheme for solving convex constrained optimization problems

15:00 -- 16:00 Rafiq Mansour: The cyclic Douglas-Rachford algorithm

16:00 -- 16:30 Open forum and closing of the meeting

Titles, Speakers (in alphabetical order) and Abstracts:

Title: Methods for the split common fixed point problem

Speaker: Andrzej Cegielski, University of Zielona Gora, Poland

Abstract: We present a general method for solving a split common fixed point problem in Hilbert spaces, with an application of strongly quasi-nonexpansive operators satisfying the demi closedness principle. We present a general convergence theorem and show that the known methods satisfy the assumptions of this theorem.

Title: Superiorization of projection methods and their use in medical applications

Speaker: Ran Davidi, Stanford University, California, USA

Abstract: Computationally demanding numerical minimization techniques are often used in medical applications such as radiation therapy treatment planning and computerized tomography. They often employ cost functions and corresponding solution approaches that are not necessarily most appropriate for achieving the desired solutions. This disconnect occurs because minimal solutions to current cost function formulations are not guaranteed to provide the optimal solution from the point of view of the application. Therefore, the considerable computational cost associated with some of these minimization techniques may not be justified. Superiorization is a new paradigm that substantially improves computational tractability by producing a solution with reduced, but not necessarily minimal, value of a defined cost function that is guaranteed to satisfy the constraints of the problem. The ability to do so stems from the fact that many feasibility-seeking projection methods are perturbation-resilient which enables to steer the process to a solution with a reduced (i.e., superior) cost function value. In this talk we present how superiorization can be applied to real-world applications and demonstrate its usefulness with a few examples taken from the medical field.

Title: projection-based scheme for solving convex constrained optimization problems

Speaker: Aviv Gibali, Fraunhofer Institute for Industrial Mathematics (ITWM), Kaiserslautern, Germany

Abstract: In this talk we present a new projection-based scheme for general convex constrained optimization problem. The general idea is to transform the original optimization problem to a sequence of feasibility problems by iteratively constraining the objective function from above until the feasibility problem is inconsistent. Then, for each of the feasibility problems one may apply any of the existing projection methods for solving it, which are known to be very efficient and practical. Some numerical experiments to illustrate the performance of the suggested scheme.

Title: The cyclic Douglas-Rachford algorithm

Speaker: Rafiq Mansour, University of Haifa, Israel

Abstract: The Douglas-Rachford (DR) algorithm is a projection method for finding the projection of a point onto the nonempty intersection of two sets. It draws great attention in the literature recently. We review recent results on the cyclic Douglas-Rachford algorithm which extends the DR algorithm to handle a family of n sets. Our presentation is based on a recent paper on this topic by J.M. Borwein and M.K. Tam.

Title: Porosity and the bounded linear regularity property

Speaker: Simeon Reich, The Technion, Israel

Abstract: H. H. Bauschke and J. M. Borwein showed that in the space of all tuples of bounded, closed and convex subsets of a Hilbert space with a nonempty intersection, a typical tuple has the bounded linear regularity property. This property is important because it leads to the convergence of infinite products of the corresponding nearest point projections to a point in the intersection. We show that the subset of all tuples possessing the bounded linear regularity property has a porous complement. Moreover, our result is established in all normed spaces and for tuples of closed and convex sets which are not necessarily bounded.

This is joint work with A. J. Zaslavski.

List of personnel receiving pay from the research effort:

- Ran Davidi (PI)