SIAG/OPT Views-and-News

A Forum for the SIAM Activity Group on Optimization

Volume 17 Number 2 October 2006

Contents

Optimization in Medicine

Introduction by the Editors

Optimization is increasingly becoming a vital component in medical and biological advances. In this issue, we present summaries of three areas (imaging, medical diagnosis, and treatment design) for which optimization has played a key role in advancing the state-of-the-art. First, Herman and Kuba describe discrete tomography and provide an overview of the most frequently used optimization methods in this imaging application. Next, Chaovalitwongse and Pardalos present optimization approaches in medical diagnosis. In particular, they report on their research, using statistical analysis and quadratic programming, involving the prediction of seizures and localization in brain epilepsy. This is followed by the article by Lee and Deasy, who describe treatment planning optimization in intensity-modulated radiation therapy, and present their experience using mixed integer programming approaches. Their article also briefly summarizes other optimization techniques applied to this area.

Eva. K. Lee and Ariela Sofer, September 2006.

Optimization in the Medical Applications of Discrete Tomography

Gabor T. Herman
Department of Computer Science,
Graduate Center, City University of New York,
New York, NY 10016, USA
(gabortherman@yahoo.com).

Attila Kuba
Department of Image Processing and Computer Graphics,
University of Szeged,
H-6701 Szeged, Hungary (kuba@inf.u-szeged.hu).

1. Introduction

Discrete Tomography (DT), as we perceive the field, has to do with determining a function (perhaps only partially, perhaps only approximately) from its projections, when the function has a known discrete range. The knowledge of the discrete range, possibly together with some prior information, can significantly reduce the number of projections required for a high-quality reconstruction. The reconstruction methods used in DT applications are usually based on some optimization problem. In this survey paper we are going to give an overview of the most frequently used optimization methods used in DT, also mentioning a few examples of its possible applications. Further survey papers [15, 16] about the medical applications of DT have been published discussing its general, not only its optimization, aspects.

Why is optimization so generally used in the applications of DT? The answer comes from the fact that, in most applications, the limited number of projections determines not only one but many solutions. This is usually the case even when some prior information about the function f to be reconstructed is available (prior information can be, for example, that f is the characteristic function of a 3D convex body) and included into the reconstruction process. Using optimization terminology, we can say that the projection data and the prior information together determine the set of feasible solutions and in order to single out one of them we have to select an objective function to be minimized over the set of feasible solutions. Then the function f optimal in this sense is to be considered as the solution of the DT reconstruction problem.

Since the different imaging techniques of tomography (*e.g.*, CT, SPECT, PET, MR, and US) have been so successfully applied in medicine, it is natural that the early versions of DT have been tried in human diagnostics (*e.g.*, angiography) as early as in the 1970s. However, there is a very clear limitation of DT in medicine: the human body cannot be (or can be only roughly) represented by a function with a discrete range. Such rough representation is possible, for example, when the absorption of the different tissues can be approximated by a function having only three possible values corresponding to the absorption coefficients of bone, lung, and the so-called soft tissues. Even such an approximate function is useful for absorption correction in Single-Photon Emission Computed Tomography (SPECT) in order to improve the quality of imaging. Another example for DT application in medicine is angiography, in which contrast material with high absorption value is injected into the blood vessel to be imaged and in this way a two-valued function (absorption coefficients of the contrast material and the background) can be used for representation.

The structure of the paper is the following. The reconstruction problem of DT with the necessary definitions and notation is described in Section 2. The next section presents the ways the DT reconstruction problem can be reformulated as some optimization problem. This section contains also the different optimization methods applied for image reconstruction in medicine. Section 4 discusses DT results in different medical imaging techniques. A brief discussion section concludes the paper.

2. Definitions and notation

Let $f : X \to D$ be the function to be reconstructed. The domain of f, X can be continuous or discrete, however the range of f is a set of known real numbers $D = \{d_1, d_2, \ldots, d_c\}$ $(c > 1)$.

In the applications we can usually give some prior information about the function to be reconstructed. For example, we may assume a class of functions having constant values on closed 3D regions with triangulated boundary surfaces. The set of functions satisfying the given prior property is denoted by $\mathcal{F}$.

The projections of the functions f are defined as

integrals on certain subsets $\mathcal{S}$ of X. (These subsets typically consist of straight lines, strips, or tubes.) Let us suppose that all elements f of $\mathcal{F}$ are integrable on each element of $\mathcal{S}$. Then the *projection* of $f \in \mathcal{F}$ for an $S \in \mathcal{S}$ is defined with the help of the *projection transformation* $\mathcal{P}$ as

$$[\mathcal{P}f](S) = \int_S f(x)\, dx. \tag{1}$$

In most applications the projections can be classified as parallel or fan-beam. In the case of *parallel projections*, the elements of $\mathcal{S}$ can be partitioned such that each partition class contains all the lines which are parallel to one direction in X. *Fan-beam projections* mean that the partition classes contain all the lines that diverge from a single point of X (in 3D and in higher dimensions, this is also called *cone beam* projections).

In the following we need to keep a clear distinction between the projection of the function f, denoted by $g = \mathcal{P}f$, and the *projection data*, y, available from the measurements. The projection data y has the same domain and range as g, but in practice it is only an approximation to g.

The *reconstruction problem* in DT can be expressed as follows. Let $\mathcal{F}$ be a class of functions $f : X \to D$ and let $\mathcal{S}$ be a finite set of subsets of X over each all elements of $\mathcal{F}$ are integrable.

RECONSTRUCTION($\mathcal{F}, \mathcal{S}$).

Given: *The projection data $y(S) \in \mathbb{R}$, for all $S \in \mathcal{S}$.*

Task: *Find a function $f \in F$ such that*

$$[\mathcal{P}f](S) \approx y(S), \tag{2}$$

for all $S \in \mathcal{S}$.

(The symbol $\approx$ denotes approximate equality.)

It is clear that posing the reconstruction problem in this way is more realistic than demanding exact equations instead of (2). Due to noisy projections or modeling errors it is quite probable that there is no solution if we replace $\approx$ with $=$ in (2).

There are several approaches different from the optimization one to solve the reconstruction problem for special sets $\mathcal{F}$ and $\mathcal{S}$ [14]. In the rest of this paper we consider only reconstruction methods that use optimization.

3. Reconstruction as an optimization problem

The general formulation of the reconstruction problem in the medical application of DT in the form of an optimization problem is:

MINIMIZATION OF A COST FUNCTION($\mathcal{F}, \mathcal{S}$).

Constraint: $f \in \mathcal{F}$.

Task: *Find the minimum of a given real valued cost function $C(f)$.*

For example, a popular form in the literature for the cost function is

$$C(f) \;=\; \|\mathcal{P}f - y\|^2 + \Phi(f), \tag{3}$$

where $\|\cdot\|$ denotes a two-norm and Φ is a real valued function. In the cost function (3), $\Phi(f)$ indicates how undesirable a solution f is from the viewpoint of our application. More generally, variants of (3) may be applied; the specific form of the resulting $C(f)$ depends on the representation of f, the selected norm $\|\cdot\|$, and the function Φ.

A most frequently used representation of f is when the domain X is a finite set of I elements (called points, pixels, or voxels). In this case the projection transformation $\mathcal{P}$ is usually replaced by a linear equation system $Af = g$,

$$\sum_i a_{ij} f_i \;=\; g_j, \qquad j = 1, \ldots, J, \tag{4}$$

where f_i denotes the ith pixel value ($i = 1, 2, \ldots, I$), $A = (a_{ij})_{I \times J}$, and a_{ij} describes the contribution of the ith pixel to the jth element of $\mathcal{S}$.

Another way of representing the function to be reconstructed in DT is to consider f as the characteristic function of some 3D set F. One possibility is to suppose that F is a finite polyhedron having a surface of triangles [4].

Another class of cost functions is based on some *probabilistic model*. For example, let us suppose that the function f to be reconstructed is a typical member of a class of images having a known probabilistic distribution $\Pi(f)$ (*e.g.*, a Gibbs distribution with

given parameters). Furthermore, suppose that we know the conditional probability $L(y|f)$ of measuring projection data y if the function is f.

Optimization of probabilistic models $(\mathcal{F}, \Pi, L, \mathcal{S})$.

Constraint: $f \in \mathcal{F}$.

Given: *The probabilistic distributions* $\Pi(f)$ *and* $L(y|f)$.

Task: *Find the optimum of a given real valued cost function* $C(f)$ *depending on* $\Pi(f)$ *and* $L(y|f)$.

As examples, we can take the maximum likelihood, the maximum *a posteriori*, or the minimum mean square error solution (see, *e.g.*, [8, 9, 10]).

Another way to reformulate the DT reconstruction problem is to consider it as a *linear integer programming* problem [1, 11, 13]. Take again the representation and problem given by the linear equation system (4). Instead of looking for the binary solution directly, first let us solve the problem with the constraint $0 \le f_i \le 1$, $i = 1, 2, \ldots, I$. Accordingly we have

Linear programming (A).

Constraints: $0 \le f_i \le 1, \quad i = 1, 2, \ldots, I,$
$Af \le y.$

Task: *Find the minimum of a given linear cost function* $C(f)$.

In most cases we are interested in a binary solution. The usual optimization method is to apply LP-relaxation to the range $[0, 1]$ and round the fractional solution. Further specialization of this approach is:

Smoothness (A, B).

Constraint: $0 \le f_i \le 1, \quad i = 1, 2, \ldots, I,$
$Af \le y,\ z \ge Bf,\ z \ge -Bf.$

Task: *Find the minimum of* $C(f) = -\sum_i f_i + \gamma \cdot \sum_i z_i$.

Here B is an $I \times I$ real matrix describing some property of the solution to be found. For example,

$$B = \begin{pmatrix} -1 & 1 & 0 & 0 & \ldots & \ldots & \ldots & 0 \\ 0 & -1 & 1 & 0 & \ldots & \ldots & \ldots & 0 \\ \vdots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \vdots \\ \vdots & \ldots & \ldots & \ldots & 0 & -1 & 1 & 0 \\ \vdots & \ldots & \ldots & \ldots & 0 & 0 & -1 & 1 \end{pmatrix}$$

means that the first order differences of f are taken into account in the constraints and we are looking for an f that is smooth in this sense. Similar idea can be applied for higher order differences and even more complex properties (see, *e.g.*, [23]).

4. Medical applications

One of the first medical applications of optimization in discrete tomography was published by Slump and Gerbrands [22]. They reconstructed the left ventricle of a dog from two projections. Because of the noise in the projection data they selected the constraints as

$$\max\{0, y_j - \sqrt{y_j}\} \le \sum_{i=1}^{I} a_{ij} f_i \le \min\{y_j + \sqrt{y_j}, n\},$$

where n denotes the number of columns and rows in the $n \times n$ binary matrix representing f. The cost function is defined as

$$C(f) \ = \ \sum_{i=1}^{I} c_i f_i,$$

where the values of c_i, $i = 1, 2, \ldots, I$, are determined on the base of a binary model (*i.e.*, a binary matrix) as follows. The element $c_i = 0$ if the ith pixel in the model is 1, otherwise c_i is a positive integer reflecting the distance of the pixel from the nearest 1-pixel of the model. That is, the *a priori* knowledge about f is that the positions of the 1s in the model constitute a subset of the positions of 1s in the cross-section to be determined. The size of the reconstructed sections was small, 46×46.

This method was modified by Reiber and coworkers [19] in order to be applicable to coronary artery reconstruction from two X-ray projections. The constraints were changed to

$$y_j - \gamma \cdot \sqrt{y_j - b} \ \le$$

$$\leq \sum_i a_{ij} f_i \leq y_j + \gamma \cdot \sqrt{y_j + b} \quad j = 1, 2, \ldots, J,$$

$$\sum_i f_i = 1/2 \cdot (\sum_j y_j), \tag{5}$$

where the constants b and γ are related to the background thickness (*i.e.*, the thickness of the tissues different from the artery) and the X-ray source current, respectively. Equation (5) expresses that we look for a solution whose total sum is the average of the total sums of the two projections (which can differ because of the noise). Reiber and co-workers performed tests using perspex phantoms with circular cross-sections. The method reconstructed the cross-sections with 18% relative mean error [19].

In 1985 Gerbrands and Slump published another reconstruction method [12], which is an extension of [22], that takes into account the stochastic nature of the X-ray imaging process. The reconstruction problem is formulated as the minimization of

$$C(f) = \omega \cdot \sum_j \frac{(\sum_i a_{ij} f_i - y_j)^2}{\sigma_j^2} + \sum_i c_i f_i$$

under the constraint (5). When the coefficient ω has a higher value, then the solution is more consistent with the projection data and is further from the model forced by the coefficients c_i. This method was applied to reconstruct a segment of a coronary artery from the same data as in [19].

Pellot and co-workers reconstructed vascular structures from two X-ray projections [17]. Let us suppose that the previously reconstructed adjacent cross-section is $f^{(p)}$. The cost function was defined by (3) with

$$\Phi(f) = \lambda_1 \cdot \sum_i \Phi_i(f) + \lambda_2 \cdot \sum_i |f_i - f_i^{(p)}|,$$

where λ_1 and λ_2 are coefficients (which are reduced during the iterative process of optimization) and $\Phi_i(f)$ is the number of pixels in the 8-neighborhood of pixel i whose value is different from f_i. The first term forces the reconstructed vessel cross-sections to be as compact as possible, while the second term encourages solutions which have similar neighboring cross-sections. For each cross-section, the optimization procedure starts with an initial f, which is the characteristic function of the ellipse that best fits the projection data in the least squares sense. Then simulated annealing (SA) is applied to find the optimal 3D shape, but in such a way that only peripheral pixels are changed in the iterative steps. Reconstructions from simulated projections of known shapes were used to set the coefficients λ_1 and λ_2 and the parameters of the SA procedure. Experiments were also performed on real angiograms. The iliac bifurcation of a patient was reconstructed from two projections and, according to a subjective comparison of the projections of the reconstructed shape with the real radiological projections, the conformity was judged to be correct.

Robert, Peyrin, and Yaffe reconstructed simulated vascular cross-sections from a few (2 to 9) cone-beam projections [20]. The cost function to be minimized was (3) with $\Phi(f)$ defined as a continuity term that encourages a voxel to have the same value as the majority of its 3D neighbors. To find the optimum, an iterative procedure based on SA was applied. Experiments showed that the value of MV error, defined as

$$MV = \frac{\sum_i |f_i - f_i^{(0)}|}{2 \sum_i f_i^{(0)}} \cdot 100\%,$$

is reduced from 9% without the continuity term in the cost function to 4% with the continuity term when reconstructing from three cone-beam projections a simulated branched vessel exhibiting a stenosis.

Chan and co-workers tested discrete tomography methods for phantom studies in Positron Emission Tomography (PET) [8, 9]. They applied the following two-stage reconstruction procedure:

1. Perform a reconstruction using some classical (nondiscrete) algorithm, *e.g.*, filtered backprojection [18], to produce an initial estimate image f'.

2. Perform a Bayesian restoration of f' to produce f.

The restoration is done by minimization of the function

$$C = \Pi(f) \cdot L'(f' \,|\, f).$$

Simulated annealing was applied for minimizing this function C.

Cunningham, Hanson, and Battle studied the reconstruction of a physical emission heart phantom from Single-Photon Emission Computed Tomography (SPECT) data [4, 10]. Altogether 24 cone-beam projections were available for the reconstruction. Assuming that the emitting radio-tracer is homogeneously distributed throughout the volume, we have again a discrete tomography reconstruction problem: the volume to be reconstructed contains two values, the absorption coefficients of the heart phantom and the background. The authors used the *maximum a posteriori* (*MAP*) estimate, *i.e.*, the f that maximizes

$$\Pi(f) \cdot L(y \,|\, f),$$

where $\Pi(f)$ and $L(y \,|\, f)$ denote the known distribution of the possible objects $f \in \mathcal{F}$ and the known probability of the measured projection y given that the image is f, respectively. The heart phantom was represented by a function having constant value within the 3D region of the heart with a triangulated boundary surface. The cost function was

$$C \;=\; \sum_j (g_j - y_j \cdot \log g_j) + \Phi(f),$$

where the function Φ enforces smoothness on the surface of the reconstructed object. A gradient-based method was proposed. The results showed the expected forms in most regions of the phantom.

Battle and co-workers reconstructed also *free-form deformation* (FFD) models to create 3D attenuation maps of the torso for attenuation correction of SPECT studies [2, 5, 6]. They considered the object to be reconstructed to be a set of closed regions: soft tissues, lungs and the spine. The regions were embedded one into another and each region was assumed to have a uniform attenuation coefficient. Altogether 37 parallel projections were collected. FFDs were used to describe continuous deformations of the space embedding the surfaces of the regions. The FFDs were given by the displacements of control points. Thus reconstruction consisted of estimation of the deformation of the initial set of control points and of estimation of the attenuation coefficient of each region. The cost function was a log likelihood function:

$$C \;=\; \log \sum_j \frac{(g_j - y_j)^2}{g_j}.$$

The optimization method to minimize the cost function was a quasi-Newton algorithm. Simulation studies were performed.

Battle and co-workers showed [3] that a similar FFD technique can be used for the lung by SPECT. In that case the distribution of the radioactive gas, and so the radioactivity, can be considered to be uniform in the regions of the lungs. The closed surfaces of the regions are represented by sets of triangles. The reconstruction starts with an initial 3D object consisting of two distinct regions of the lungs, and then goes on finding the displacements of the control points, and so the corresponding deformed object, that best match the given projection data. They reported on using ML and MAP solutions. Experiments were performed on simulated data sets using 36 parallel projections. The FFD was quicker and gave superior results than a direct deformation method.

Another optimization method was tested by Carvalho and co-workers [7]. They supposed that the image f is a random sample from a known Gibbs distribution. The cost function to be minimized was defined as

$$C(f) \;=\; -\sum_i I_i(f) + \gamma \cdot \sum_j |y_j - g_j|,$$

where $I_i(f)$ is the so-called *local energy function* for the pixel i depending on the binary value of f_i and those of its eight neighbors, *i.e.*, *configuration* in the 8-neighborhood of the pixel i. The local energy function determines which are the preferred and less-preferred configurations in the reconstructed image. The software phantoms used for testing consisted of mathematically-described approximations of the left and right ventricles of the heart and the left atrium, the image sizes were 63×63. Three views were generated (from the horizontal, the vertical, and a diagonal direction). The average MV error where $f^{(0)}$ denotes the phantom, was between 1.3% and 3.1% depending on the simulated noise level.

Senasli *et al.* [21] published a reconstruction method using (cubic) B-spline functions to describe the vessel contours in each cross-section. In this case, reconstruction consisted of finding the optimal control points of B-splines. The cost function was written as (3) with

$$\Phi(f) \;=\; \lambda_1 \cdot U_{reg} + \lambda_2 \cdot U_{cont},$$

where the contour regularity term U_{reg} and U_{cont} measure the irregularity of the actual contour and the continuity between the previous and the current cross-sections, respectively. Simulated annealing was used for positioning the control points. In one iterative step a randomly selected control point could move to one of its 8-neighbors. The initial contour was an ellipse and the initial contour points were uniformly distributed over it. Experiments were performed on software and physical phantoms simulating both concentric and eccentric stenoses, and the error of reconstruction was measured as the relative mean error $\overline{R}$ between f and the real object $f^{(0)}$, defined by

$$\overline{R} \;=\; \frac{\sum_i |f_i - f_i^{(0)}|}{\sum_i f_i^{(0)}} \cdot 100\%.$$

5. Conclusion

Optimization is an appropriate tool in DT because it selects a particular solution among the many that would be available if only constraint satisfaction were required. In this paper we have given an overview of the specific choices for constraints and optimizing functionals that were suggested in literature and their reported performance in medical applications.

Acknowledgments

This work was supported by the National Science Foundation grant DMS 0306215 "Aspects of Discrete Tomography", and by the Hungarian Research Foundation grant T 048476 "New Problems of Discrete Tomography and Its Application in Neutron Radiography."

REFERENCES

[1] R. Aharoni, G. T. Herman, and A. Kuba, *Binary vectors partially determined by linear equation systems*, Discrete Math., 171 (1997), pp. 1–16.

[2] X. L. Battle and Y. Bizais, *Binary 3D attenuation map reconstruction using geometrical models and free form deformations*, in *Proc. Fully Three- Dimensional Image Reconstruction in Radiology and Nuclear Medicine*, (1999), pp. 181–184.

[3] X. L. Battle, Y. Bizais, C. Le Rest, and A. Turzó, *Tomographic reconstruction using free-form deformation models*, in *Medical Imaging: Image Processing*, K. M. Hanson *et al.*, editors, Proc. SPIE. 3661 (1999), pp. 356–367.

[4] X. L. Battle, K. M. Hanson, and G. S. Cunningham, *Tomographic reconstruction using 3D deformable models*, Phys. Med. Biol., 43 (1998), pp. 983–990.

[5] X. L. Battle, C. Le Rest, A. Turzó, and Y. Bizais, *Free-form deformation in tomographic reconstruction. Application to attenuation map reconstruction*, IEEE Trans. Nucl. Sci., 47 (2000), pp. 1065–1071.

[6] X. L. Battle, C. Le Rest, A. Turzó, and Y. Bizais, *3D attenuation map reconstruction using geometrical models and free-form deformations*, IEEE Trans. Medical Imaging, 19 (2000), pp. 404–411.

[7] B. M. Carvalho, G. T. Herman, S. Matej, C. Salzberg, and E. Vardi, *Binary tomography for triplane cardiography*, in *Information Processing in Medical Imaging*, A. Kuba, M. Samal, and A. Todd-Pokropek, editors, LNCS-1613, Springer-Verlag, Berlin, (1999), pp. 29–41.

[8] M. T. Chan, G. T. Herman, and E. Levitan, *Bayesian image reconstruction using image-modeling Gibbs priors*, Int. J. Imaging Systems Techn., 9 (1998), pp. 85–98.

[9] M. T. Chan, G. T. Herman, and E. Levitan, *Probabilistic modeling of discrete images*, in *Discrete Tomography. Foundations, Algorithms, and Applications*, G. T. Herman, A. Kuba, editors, Birkhäuser, Boston, (1999), pp. 213–235.

[10] G. S. Cunningham, X. L. Battle, and K. M. Hanson, *Three-dimensional reconstructions from low-count SPECT data using deformable models*, Opt. Express, 2 (1998), pp. 227–236.

[11] P. C. Fishburn, P. Schwander, L. A. Shepp, and R. J. Vanderbei, *The discrete Radon transform and its approximate inversion via linear programming*, Discrete Appl. Math., 75 (1997), pp. 39-61.

[12] J. J. Gerbrands and C. H. Slump, *3-D reconstruction of homogeneous objects from two Poisson-distributed projections*, Pattern Recognition Letters, 3 (1985), pp. 137–145.

[13] P. Gritzmann, S. de Vries, and M. Wiegelmann, *Approximating binary images from discrete X-rays*, SIAM J. Optim., 11 (2000), pp. 522-546.

[14] G. T. Herman and A. Kuba (eds.), *Discrete Tomography: Foundations, Algorithms and Applications*, Birkhäuser, Boston, 1999.

[15] G. T. Herman and A. Kuba, *Discrete tomography in medical imaging*, Proc. of IEEE, 91 (2003), pp. 1612–1626.

[16] A. Kuba, G. T. Herman, S. Matej, and A. Todd-Pokropek, *Medical applications of discrete tomography*, in *Discrete Mathematical Problems with Medical Applications, DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, vol. 55, D. Z. Du, P. M. Pardalos, and J. Wang, editors, American Mathematical Society, Rhode Island, (2000), pp. 195–208.

[17] C. Pellot, A. Herment, M. Sigelle, P. Horain, H. Maitre, and P. Peronneau, *A 3D reconstruction of vascular structures from two X-ray angiograms using an adapted simulated annealing algorithm*, IEEE Trans. Medical Imaging, 13 (1994) pp. 48–60.

[18] G. N. Ramachandran and A. V. Lakshminarayanan, *Three-dimensional reconstruction from radiographs and electron micrographs: Application of convolutions instead of Fourier transforms*, Proc. Natl. Acad. Sci. USA, 68 (1971), pp. 2236–2240.

[19] J. H. C. Reiber, J. J. Gerbrands, G. J. Troost, C. J. Kooijman, and C. H. Slump, *3-D reconstruction of coronary arterial segments from two projections*, in *Digital Imaging in Cardiovascular Radiology*, P. H. Heintzen, R. Brennecke, editors, Georg Thieme Verlag, Stuttgart, Germany, (1983), pp. 151–163.

[20] N. Robert, F. Peyrin, and M. J. Yaffe, *Binary vascular reconstruction from a limited number of cone beam projections*, Med. Phys, 21 (1994), pp. 1839–1851.

[21] M. Sensali, L. Garnero, A. Herment, and E. Mousseaux, *3D reconstruction of vessel lumen from very few angiograms by dynamic contours using a stochastic approach*, Graphical Models, 62 (2000), pp. 105–127.

[22] C. H. Slump and J. J. Gerbrands, *A network flow approach to reconstruction of the left ventricle from two projections*, Comp. Graphics Image Proc., 18 (1982), pp. 18–36.

[23] S. Weber, C. Schnörr, and J. Hornegger, *A linear programming relaxation for binary tomography with smoothness priors*, in *Proceedings of International Workshop on Combinatorial Image Analysis (IWCIA)*, A. Del Lungo, V. Di Gesu, and A. Kuba, editors, Palermo, Italy, (2003).

Optimization Approaches to Characterize the Hidden Dynamics of the Epileptic Brain: Seizure Prediction and Localization

Wanpracha Art Chaovalitwongse
Department of Industrial and Systems Engineering,
Rutgers, the State University of New Jersey,
Piscataway, NJ 08854, USA
(wchaoval@rci.rutgers.edu).

Oleg A. Prokopyev
Department of Industrial Engineering,
University of Pittsburgh,
Pittsburgh, PA 15261, USA
(prokopyev@engr.pitt.edu).

Panos M. Pardalos
Departments of Industrial and Systems Engineering and
Biomedical Engineering,
Brain Institute, University of Florida,
Gainesville, FL 32611-6595, USA
(pardalos@ufl.edu).

1. Introduction

At least 40 million people worldwide (or 1% of the population) currently suffer from epilepsy, which is the second most common serious brain disorder after stroke. Epilepsy is a chronic condition of diverse etiologies with the common symptom of spontaneous recurrent seizures, which is characterized by intermittent paroxysmal and highly organized rhythmic neuronal discharges in the cerebral cortex. In some types of epilepsy (*e.g.*, focal or partial epilepsy), there is a localized structural change in neuronal circuitry within the cerebrum which produces organized quasi-rhythmic discharges, which spread from the region of origin (epileptogenic zone) to activate other areas of the cerebral hemisphere [29]. The transitional development of the epileptic state can be considered as changes in network circuitry of neurons in the brain that produce changes in voltage potential, which can be captured by an electroencephalogram (EEG), the most common tool for evaluating the physiological state of the brain. These changes are reflected by wriggling lines along the time axis in a typical EEG recording.

Approximately 25 to 30% of epileptic patients remain unresponsive to the treatment with antiepileptic drugs (AEDs), which is the mainstay of epilepsy treatment, and continue to have seizures and still have inadequate seizure control. Epilepsy surgery is another alternative treatment for medically refractory patients with the aim of excising the portion of brain tissue supposed to be responsible for seizure initiation. However, at least 50% of pre-surgical candidates eventually will not undergo respective surgery because a single epileptogenic zone could not be identified or was located in functional brain tissue. Besides, only 60 to 85% of epilepsy surgery cases result in seizure free. In the recent years, the vagus nerve stimulator Neurocybernetic Prosthesis has been available as an alternative epilepsy treatment that reduces seizure frequency; however, the parameters of this device (amplitude and duration of stimulation) continue to be arbitrarily adjusted by physicians. Moreover, more than a minority of patients have minor side effects and can benefit from this treatment. Due to the shortcomings and side effects of current epilepsy treatment, there has been an urgency for new development of novel therapeutic treatments for epilepsy. During the last few years, there has been a great deal of research interest in epilepsy research shifted from the research in curing epilepsy to the ability to anticipate/predict the onset of seizures. Although spontaneous epileptic seizures seem to occur randomly and unpredictably and begin intermittently as a result of complex dynamical interactions among many regions of the brain, neurologists still believe that seizures occur in a predictable fashion. Seizure prediction is a very promising option for the effective and safe treatment of people with epilepsy by avoiding both the side effects of drugs and cutting out pieces of brain. The most realizable application of seizure prediction development is its potential for use in therapeutic epilepsy devices to either warn about an impending seizure or trigger intervention to prevent seizures before they begin.

Work on seizure prediction started in the 1970s [37] and early 1980s [30] to show the seizure's predictability. Most of the work was focused on visible features in the EEG (*e.g.*, epileptic spiking) to extract seizure precursors. More advanced quantitative analyses (*e.g.*, spectral analysis) in the EEG are applied to discover the abnormal activity and demonstrate the predictability in seizure patterns. Since the complexity and variability of the seizure development cannot be captured by traditional methods used to process physiological signals,

Iasemidis and co-workers were the first group to attempt to apply the theory of nonlinear dynamics to the EEG for predicting seizures [17]. The results of this work indicates that the EEG becomes progressively less chaotic as seizures advance, with respect to the estimation of short-term maximum Lyapunov exponents (STL_{max}), which is a measure of the order or disorder (chaos) of signals. During the past decade, Iasemidis and his group have demonstrated dynamical properties and the large-scale patterns of EEG that emerge when neurons interact all together, which demonstrate that the convulsive firing of neurons in epileptic seizures offers such a clear case of collective dynamics. For example, evidence for nonlinear time dependencies in the inter-seizure intervals observed from patients with frequent partial seizures is reported in [11]. This observation suggests that the occurrence of seizures, though displaying a complex time structure, is not a random process and may be driven by deterministic mechanisms. Later attempts to apply measures in nonlinear dynamics were followed by other investigations [20, 21, 24, 31, 26, 19]. The application of the correlation dimension, another nonlinear dynamics approach, is employed to measure the neuron complexity of the EEG and correlation density and dynamical similarity were employed to show evidence of seizure anticipation in pre-seizure segments [8, 20, 21]. In these studies, reductions in the effective correlation dimension (D_2^{eff}, a measure of the complexity of the EEG signals) are shown to be more prominent in pre-ictal EEG samples than at times more distant from a seizure. Elger and co-workers estimate that a detectable change in dynamics can be observed at least 2 minutes before a seizure in most cases [8]. Because their datasets were only of 10 to 30 minutes in duration, the exact duration of the pre-ictal state cannot be determined. These studies were followed by the measure of phase synchronization in the pre-seizure EEG signals [24, 31]. Martinerie and co-workers also report significant differences between dimension measures obtained in pre-ictal versus inter-ictal EEG samples [24]. They find an abrupt decrease in dimension during the pre-ictal transition. This study also employs relatively brief (40 minutes) samples of pre-ictal and inter-ictal data. More recently, this group has reported changes in brain dynamics obtained from scalp electrode recordings of the EEG. By comparing pre-ictal EEG samples to a reference sample selected from inter-ictal data, they find evidence of dynamical changes that anticipate temporal lobe seizures by periods of up to 15 minutes [31]. In that study, they employ a method, inspired by Manuca and Savit [23], which measures the degree of stationarity of EEG signals. The changes or sustained bursts in long-term energy profiles of the EEG are reported to be increasing in volume that leads to seizure onset [22]. In the most recent study, the application of the correlation dimension, correlation integral, and autocorrelation is studied to demonstrate the fluctuations of seizure dynamics [26, 19].

Although the aforementioned studies have successfully demonstrated that there exist temporal changes in the brain dynamics reflected from seizure development, the collective physiological dynamics of billions of interconnected neurons in brain are not well studied or understood. Since temporal properties of the brain dynamics can only capture the interaction of some groups of locally-connected neurons, they are not sufficient to demonstrate either the mechanism or the propagation of seizure development, which involves billions of interconnected neurons throughout the brain. For example, extensive investigations indicate that the quantification of only temporal properties of the brain dynamics (*e.g.*, STL_{max}) fail to demonstrate the capability and sufficiency to predict seizures [6].

For this reason, a study that considers both temporal and spatial properties of the brain dynamics is proposed to demonstrate that the spatiotemporal dynamical properties of EEG's can reveal patterns that correspond to specific clinical states [28, 14, 27]. These studies lead to the development of an Automated Seizure Warning System (ASWS) [33, 35, 5], which not surprisingly demonstrates that the inter-ictal, ictal, and immediate post-ictal states are distinguishable with respect to the spatiotemporal dynamical patterns/properties of intracranial EEG recordings. These patterns are considered to be seizure precursors reflected from the convergence of STL_{max} profiles from a group of electrode sites during the hour preceding seizures. The transition from a seizure precursor to a seizure onset has been defined as a "pre-ictal transition". In essence, the ASWS algorithm is developed to study the real

seizure prediction, which is proposed as a natural extension of previous investigations based on an analysis of spatio-temporal properties of the brain dynamics [17, 11, 14, 6, 27]. The experiments in these studies were undertaken to determine if it is possible to predict an impending seizure automatically by a robust method employing the ASWS algorithm, which is an online computer-based algorithm.

The spirit of the ASWS algorithm involves (1) Quantitative Approach to Characterize the Dynamics of EEG time series: an estimation of STL_{max} to quantify temporal properties of the brain dynamics; (2) Statistical Measure to Quantify Similarity Patterns of the Brain Dynamics: a statistical estimate of the degree of similarity of patterns/properties of the brain dynamics; (3) Quadratic Programming Approach to Select Critical Electrode Sites: an optimization technique to identify critical spatial features (the most similar statistical properties) of the brain dynamics.

2. Quantitative approach to characterize the dynamics of EEG time series

Since the brain is a nonstationary system, algorithms used to estimate measures of the brain dynamics should be capable of automatically identifying and appropriately weighing existing transients in the data. In the ASWS algorithm, EEG signals are divided into sequential epochs (non-overlapping windows) to properly account for possible nonstationarities in the epileptic EEG. For each epoch of each channel of EEG signals, we quantify the brain dynamics by applying measures of chaos. An estimation of STL_{max} is employed as a measure of chaos, quantification of the chaoticity of the attractor. In other words, it measures the average uncertainty along the local eigenvectors of an attractor in the phase space. In fact, the rate of divergence is an important aspect of the system dynamics and is reflected in the value of Lyapunov exponents. Next, a short overview of mathematical models used in the estimation of STL_{max} will be discussed.

To characterize the brain dynamics from multidimensional EEG time series, the initial step in analyzing the dynamical properties of EEG signals is to embed it in a higher dimensional space of dimension p, which enables us to capture the behavior in time of the p variables that are primarily responsible for the dynamics of the EEG. We can now construct p-dimensional vectors $X(t)$, whose components consist of values of the recorded EEG signal $x(t)$ at p points in time separated by a time delay. Construction of the embedding phase space from a data segment $x(t)$ of duration T is made with the method of delays. The vectors X_i in the phase space are constructed as:

$$X_i = (x(t_i), x(t_i + \tau) \dots x(t_i + (p-1) * \tau)) \quad (1)$$

where τ is the selected time lag between the components of each vector in the phase space, p is the selected dimension of the embedding phase space, and $t_i \in [1, T - (p-1)\tau]$.

The method for estimation of the Short Term Maximum Lyapunov Exponent (STL_{max}) for nonstationary data (*e.g.*, EEG time series) is previously explained in [10, 13, 38]. In this article, only a short description and basic notation of our mathematical models used to estimate STL_{max} will be discussed. First, let us define the following notation.

- $X(t_i)$ is the point of the fiducial trajectory $\phi_t(X(t_0))$ with $t = t_i$, $X(t_0) = (x(t_0), \dots, x(t_0 + (p-1) * \tau))$, and $X(t_j)$ is a properly chosen vector adjacent to $X(t_i)$ in the phase space.

- $\delta X_{i,j}(0) = X(t_i) - X(t_j)$ is the displacement vector at t_i, that is, a perturbation of the fiducial orbit at t_i, and $\delta X_{i,j}(\Delta t) = X(t_i + \Delta t) - X(t_j + \Delta t)$ is the evolution of this perturbation after time Δt.

- $t_i = t_0 + (i-1) * \Delta t$ and $t_j = t_0 + (j-1) * \Delta t$, where $i \in [1, N_a]$ and $j \in [1, N]$ with $j \neq i$.

- Δt is the evolution time for $\delta X_{i,j}$, that is, the time one allows $\delta X_{i,j}$ to evolve in the phase space. If the evolution time Δt is given in seconds, then L is in bits per second.

- t_0 is the initial time point of the fiducial trajectory and coincides with the time point of the first data in the data segment of analysis. In the estimation of L, for a complete scan of the attractor, t_0 should move within $[0, \Delta t]$.

- N_a is the number of local L_{max}'s that will be estimated within a duration T data segment. Therefore, if D_t is the sampling period of the time domain data, $T = (N-1)D_t = N_a\Delta t + (p-1)\tau$.

Let L be an estimate of the short term maximum Lyapunov exponent, defined as the average of local Lyapunov exponents in the state space. L can be calculated as follows.

$$L = \frac{1}{N_a \Delta t} \sum_{i=1}^{N_a} \log_2 \frac{|\delta X_{i,j}(\Delta t)|}{|\delta X_{i,j}(0)|} \tag{2}$$

with

$$\delta X_{i,j}(0) = X(t_i) - X(t_j) \tag{3}$$
$$\delta X_{i,j}(\Delta t) = X(t_i + \Delta t) - X(t_j + \Delta t). \tag{4}$$

Per electrode, we computed the STL_{max} profile using the method proposed by Iasemedis *et al.* [10], which is a modification of the method by Wolf *et al.* [38]. Modification of the Wolf's algorithm is necessary to better estimate of STL_{max} in small epochs that include transients, such as inter-ictal spikes. The modification of the STL_{max} algorithm is primarily in the searching procedure for a replacement vector at each point of a fiducial trajectory. In the previous study of EEG analysis, the crucial parameter in the L_{max} estimation is found to be an adaptive estimation (in time and phase space) of the magnitude bounds of the candidate displacement vector to avoid catastrophic replacements. This parameter plays a very important role in distinguishing the pre-ictal, the ictal, and the post-ictal stages.

3. Statistical measure to quantify similarity patterns of the brain dynamics

A similarity measure is proposed to estimate the difference of the dynamics of EEG time series between different groups of the brain states. In other words, the T-index is employed as a measure of statistical distance between two epochs of STL_{max} profiles. The T-index at time t between electrode sites i and j is defined as:

$$T_{i,j}(t) = \sqrt{N} \times |E\{STL_{max,i} - STL_{max,j}\}| / \sigma_{i,j}(t) \tag{5}$$

where $E\{\cdot\}$ is the sample average difference for the $STL_{max,i} - STL_{max,j}$ estimated over a moving window $w_t(\lambda)$ defined as:

$$w_t(\lambda) = \begin{cases} 1 & \text{if } \lambda \in [t-N-1, t] \\ 0 & \text{if } \lambda \notin [t-N-1, t], \end{cases}$$

where N is the length of the moving window. Then, $\sigma_{i,j}(t)$ is the sample standard deviation of the STL_{max} differences between electrode sites i and j within the moving window $w_t(\lambda)$. The thus defined T-index follows a t-distribution with $N-1$ degrees of freedom. In this study, STL_{max} profiles are divided into overlapping 10-minute epochs ($N = 60$ points).

4. Quadratic programming approach to select critical electrode sites

Motivated by the Sherrington-Kirkpatric Hamiltonian, one of the most interesting problems about this model is the determination of the minimal-energy states (GROUND STATE problem) [2, 3, 4]. For this reason, quadratic 0-1 programming techniques have been extensively used to study Ising spin glass models [1, 9, 25, 4]. In this research, quadratic 0-1 programming techniques are employed to select the critical cortical sites, where each electrode has only two states, and to determine the minimal-average T-index state (brain areas with the most similar dynamical states). This problem is formulated as a multi-quadratic 0-1 knapsack problem with objective function to minimize the average T-index (a measure of statistical distance between the mean values of STL_{max}) among electrode sites, the knapsack constraint to identify the number of critical cortical sites [18, 16], and an additional quadratic constraint to ensure that the optimal group of critical sites shows the divergence in STL_{max} profiles after a seizure. In essence, we basically aim to select electrode sites such that they are most similar (minimum T-index value) prior to the seizure, conditional on the divergence of STL_{max} profiles after the seizure onset.

The optimization problem is formulated as the followings.

1. A T-matrix corresponding to the 10-minute epoch prior to the seizure onset was generated

and put into the objective function, which needs to be minimized.

2. A T-matrix corresponding to the 10-minute epoch after the seizure onset was generated and put into the quadratic constraint, which ensures that the selected electrode sites (solution to the optimization problem) show the divergence in STL_{max} after the seizure onset.

3. A linear constraint of the number of critical electrode sites (k) was added in the optimization problem.

4.1 Notation

Let A be $n \times n$ matrix, whose each element $a_{i,j}$ represents the T-index between electrode i and j within 10-minute window before the onset of a seizure. Define $x = (x_1, ..., x_n)$, where each x_i represents the cortical electrode site i. If the cortical site i is selected to be one of the critical electrode sites, then $x_i = 1$; otherwise, $x_i = 0$. k denotes the number of selected critical electrode sites. Let B be $n \times n$ matrix, whose each element $b_{i,j}$ represents the T-index between electrode i and j within 10-minute window after the onset of a seizure.

4.2 Formulation

The electrode selection problem can be formulated as the following multi-quadratic 0-1 programming problem given by:

$$\begin{aligned} \min \quad & x^T A x & (6)\\ \text{s.t.} \quad & \textstyle\sum_{i=1}^{n} x_i = k & (7)\\ & x^T B x \geq T_\alpha k(k-1) & (8)\\ & x \in \{0,1\}^n. & (9) \end{aligned}$$

Eq. (8) is added to ensure that the optimal group of critical sites shows this divergence by adding one more quadratic constraint. The constant T_α is the critical value of T-index, as previously defined, to reject H_o: "two brain sites acquire identical STL_{max} values within time window $w_t(\lambda)$".

Note that the problem in Eqs. (6)-(9) is a special case of multi-quadratic 0–1 programming problems. In this case, for the matrices A and B, $\forall i, j\ a_{ij} \geq 0,\ b_{ij} \geq 0$ and $\forall i\ a_{ii} = 0,\ b_{ii} = 0$.

Consider the following problem

$$\min_{x \in \{0,1\}^n,\ e^T x = k} x^T Q x, \tag{10}$$

where $\forall i, j\ q_{ij} \geq 0$ and $\forall i\ q_{ii} = 0$.

Problem (10) can be shown to be *NP*-hard as follows. In [9] it is shown that the maximum clique problem (which is known to be *NP*-hard) in a graph $G = (V, E)$ with vertex set $V = \{1, \ldots, n\}$ and edge set E is polynomially equivalent to

$$\begin{aligned} \min f(x) \quad &= -\sum_{i=1}^{n} x_i + 2 \sum_{\substack{(i,j) \notin E \\ i > j}} x_i x_j \\ &= -e^T x + 2 \sum_{\substack{(i,j) \notin E \\ i > j}} x_i x_j \qquad (11)\\ \text{s.t.} \quad & x \in \{0,1\}^n. \end{aligned}$$

Obviously, the problem (11) can be solved by solving $n + 1$ problems of the form

$$\begin{aligned} \min \quad & f_k(x) = \sum_{\substack{(i,j) \notin E \\ i > j}} x_i x_j \\ \text{s.t.} \quad & e^T x = k,\ x \in \{0,1\}^n. \end{aligned} \tag{12}$$

for each $k \in [0, n]$. Note that problem (12) is a restricted version of problem (10). The solution of the problem (11) will be the one that yields the value of minimal $2f_k(x) - k$. Therefore, we can solve the maximum clique problem by solving $n+1$ problems (12). Hence, problem (10) is *NP*-hard. As the problem (10) with additional quadratic constraint is a generalization of the problem (10), the problem in Eqs. (6)-(9) is also *NP*-hard. To solve the *NP*-hard problem in Eqs. (6)-(9), two computational approaches have been proposed in [7, 27].

5. Performance of an automated seizure warning system (ASWS) algorithm

The development of an ASWS algorithm was extended from the results of our previous studies demonstrating that if one knows which critical electrode sites will participate in the next pre-ictal transition, it may be possible to detect the seizure precursors in time to warn about an impending seizure [6]. The main components that constitute the ASWS algorithm are as follows.

1. The estimation of STL_{max} profiles is used to measure the degree of order or disorder (chaos) of the EEG signals.

2. The critical electrode selection was accomplished by an automated optimization technique based upon the behavior of STL_{max} profiles before and after each preceding seizure.

3. Such a warning will be triggered when the similarity degree of the brain dynamics from critical electrode sites crosses the threshold. In practice, this warning will activate a therapeutic intervention to abort an impending seizure.

The prospective analysis of the ASWS algorithm in the continuous long-term intracranial EEG recordings constitutes, for the first time to our knowledge, an automated seizure warning device. Cases with continuous recordings of several days in duration are selected for this initial evaluation of the method. To evaluate the performance of the ASWS algorithm, we calculate the sensitivity and false positive rate of the algorithm tested on continuous long-term intracranial EEG recordings, which have previously been obtained for clinical purposes. In the algorithm, there are ranges of different parameter settings, which need to be adjusted and optimized. In order to find the optimal parameter setting, Receiver Operating Characteristics (ROC) curve analysis is employed to indicate an appropriate trade-off that one can achieve between the false positive rate (1-Specificity, plotted on X-axis) and the true positive rate (Sensitivity, plotted on Y-axis). To test the ASWS algorithm on-line, we first trained the algorithm by dividing the data set into training data set and testing data set. In each of the 10 test patients, the first half of seizures are used to train for the optimal parameter setting. With the optimal parameter setting obtained from the training phase, the algorithm is tested prospectively on the testing data set. During the training step, in order to find the most appropriate trade-off, the optimal parameter setting is defined as the one closest to the ideal point in ROC curve (100% sensitivity and 0 false positive rate). A "prediction score" is employed to measure the closeness to the ideal point, which represents the "goodness" of a prediction algorithm. The lower the prediction score, the better the prediction algorithm. In fact, the prediction score is actually a distance from the performance point (sensitivity and false positive rate) of a predictor on the ROC curve to the ideal point (100% sensitivity and 0 false positive rate). The prediction score can be calculated from $\sqrt{(1 - Sensitivity)^2 + FPR^2)}$.

To demonstrate the importance of optimization techniques in the electrode selection process that it can capture the critical spatial connections of the brain dynamics, a statistical testing experiment is proposed to verify if the pre-ictal transitions detected by the ASWS algorithm are truly the physiological changes in the seizure development by comparing the prediction scores of the ASWS algorithm in the cases with and without optimizing the electrode selection process. To validate that the optimization process in the ASWS algorithm is the key component capable of identifying vital spatial dependence of the brain dynamics in the seizure development, the ASWS algorithm without optimization in electrode site selection process will be tested. In other words, the non-optimized ASWS (NASWS) will follow the same experimental procedure except that the groups of electrode sites used in the monitoring process are randomly selected. The non-optimized ASWS (NASWS) algorithm is tested on the real dataset for 100 iterations. To demonstrate that the optimization process is a vital ingredient in the ASWS algorithm to capture the spatial properties of the seizure development, the performance characteristics and average prediction score of the NASWS algorithm are calculated.

The ASWS algorithm was tested on the dataset of 10 patients, whose characteristics are described in Table 1. The performance characteristics of the ASWS algorithm in the first phase (training phase) and the second phase (testing phase) are described as follows. In the training phase for the optimal parameter setting, the optimal parameter setting is trained by testing the algorithm on the first half of seizures for individual patient. With the optimal parameter, the ASWS algorithm achieves a sensitivity (an average probability of a seizure to be predicted) of 76.12% with an average false prediction rate of 0.17 per hour, which is equivalent to a prediction score of 0.295. The average true warning time in the training phase is 72.18 minutes while the average ratio of warning times to inter-seizure in-

Table 1: Data Characteristics: (Onset region: LH, Left Hippacampal; RH, Right Hippacampal; RF, Right Orbi-frontal. Seizure types: CP, Complex Partial; SC, Subclinical; GTC, Generalized Tonic/Clonic)

Patient ID	Gender	Onset region	Age	Seizure types	Duration of EEG	Number of seizures	Average inter-seizure interval
1	F	RH	45	CP, SC	3.63	9	8.69
2	M	RH, RF	60	CP, GTC, SC	11.98	7	20.32
3	F	RH	41	CP	9.06	24	3.62
4	M	RH	19	CP, SC	13.45	17	17.10
5	M	RH	33	CP, SC	12.24	18	15.30
6	M	RH	38	CP, SC	3.18	9	7.59
7	M	LH, RH	44	CP, SC	6.24	23	7.77
8	M	RH	29	CP, SC	6.07	19	6.61
9	F	LH, RH	37	CP, SC	11.80	20	12.54
10	M	LH, RH	37	CP, GTC	9.88	12	19.49
Total					87.53 days	158	10.86 hours

tervals is 0.101. In testing phase, the algorithm is tested on the other half of seizures from individual patient using the optimal parameter from the training phase. The algorithm achieves a sensitivity of 68.75% with an average false prediction rate of 0.15 per hour, which is equivalent to a prediction score of 0.322. This false prediction rate corresponds to false warning every 6.7 hours. On average the algorithm generates a true warning approximately 72 minutes before each seizure) while the average ratio of warning times to inter-seizure intervals is 0.317. These results demonstrate the reliability of the ASWS algorithm, which is an indication of the possibility to develop automated seizure warning devices for diagnostic and therapeutic purposes.

The NASWS algorithm is tested on the real dataset by randomly selecting groups of electrode sites used in the monitoring process for 100 iterations. For individual patient, we use the same optimal parameter setting as in the previous experiment in both training and testing phases. The performance characteristics and average prediction score of the NASWS algorithm in the training and testing phases for individual patient and overall are summarized as follows. In the training phase, the algorithm achieved an average sensitivity of 50.90% with an average false prediction rate of 0.287 per hour, which is equivalent to a prediction score of 0.509. In the testing phase, the algorithm achieves a sensitivity of 59.10% with an average false prediction rate of 0.433 per hour, which is equivalent to a prediction score of 0.471. Note that the prediction score of the ASWS algorithm is significantly lower than the average prediction score of the NASWS algorithm. Examples of the distribution of prediction scores in 100 iterations of the NASWS algorithm tested on patient 3 compared to the prediction score of the ASWS in the training and testing phases are illustrated in Figures 1 (A) and (B), respectively. The prediction score of the ASWS algorithm is significantly lower, and is statistically better, than the prediction score of the NASWS with p-value's of 0.09 and 0.02 in the training and testing phases, respectively. These results indicate that the optimization process is a vital ingredient in the ASWS algorithm to capture dynamical interactions in the spatial properties of brain dynamics as seizures advance. The results of this study suggest that the optimal electrode sites

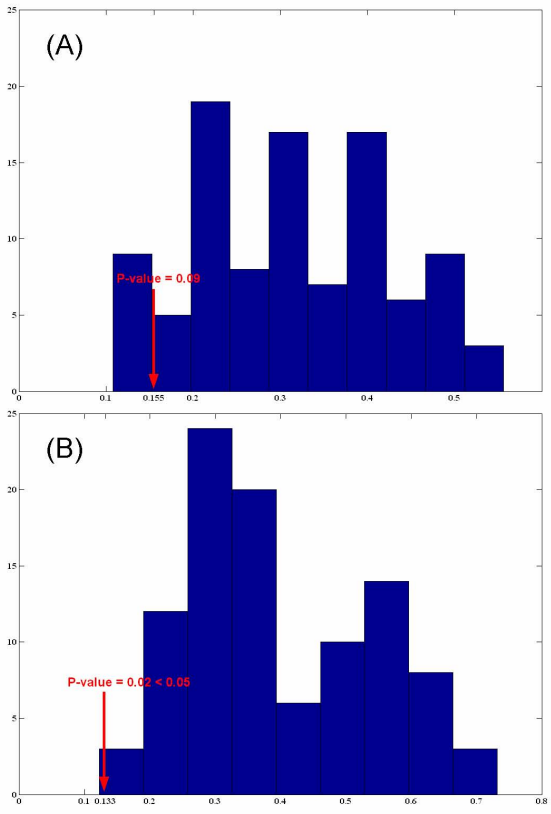

Figure 1: Example of the the prediction score histogram of the NASWA algorithm versus the prediction score of the ASWS algorithm on the training (A) and testing (B) sets of patient 1. The p-value's of real seizure points in training and testing sets are 0.09 and 0.02 respectively, which is significantly less than that of the NASWA algorithm.

selected by the ASWS algorithm demonstrate evidence that they can portray the unique physiological changes in the seizure development with sufficient lead-time (the prediction horizon).

6. Concluding remarks

The results of this study indicate that the ASWS algorithm designed to detect dynamical patterns of critical electrode sites is capable of providing a seizure warning well in advance of a seizure. In the cases analyzed for this study, the average seizure warning time ranges from 22.4 to 135.0 minutes. This time interval is sufficient to allow a wide range of therapeutic interventions. However, the performance (sensitivity and false prediction rate) of the ASWS algorithm are still considerably inferior to the results reported in the previous seizure predictability studies [12, 15]. One of the reasons is that the electrode selection in those retrospective studies was done in advance during the next seizure (using the future information). On the other hand, in this prospective study, the algorithm is tested online without using any future information. In addition, the electrode selection process only uses the information from the previous seizure. However, the ASWS algorithm may be improved since we use the same parameter settings for every patient in the procedures to quantify the brain dynamics, optimize electrode selection, and detection of pre-ictal transition. Those parameters remain to be further investigated. Nevertheless, the temporal and spatial properties of the brain dynamics captured by the proposed method have been proven capable of reflecting the real physiological changes in the brain as they correspond specifically to the real seizure precursors.

These results are considered to be the groundwork of seizure prediction research. Potential diagnostic applications include a seizure warning system used during long-term EEG recordings performed in a diagnostic epilepsy-monitoring unit. This type of system could potentially be used to warn professional staff of an impending seizure or to trigger functional imaging devices in order to measure regional cerebral blood flow during seizure onset. This type of seizure warning algorithm could also be incorporated into digital signal processing chips for use in implantable devices. Such devices could be utilized to activate pharmacological or physiological interventions designed to abort an impending seizure. Future studies, employing novel experimental designs are required to investigate the therapeutic potential for implantable seizure warning devices.

In addition, Iasemidis and co-workers have also explored the possibility of applying the ASWS algorithm to non-invasive scalp EEG recordings [32, 34, 36]. The implementation is complicated by the fact that the parameter settings (embedding dimension and time delay) in the estimation of STL_{max} is optimized based on the ictal EEG depth recordings in human subject with respect to minimization of the transient and reduction of the nonstationarity of EEG. Therefore the ASWS algorithm cannot gain the maximum prediction power with non-optimal

parameter setting, which remains to be further investigated. The clinical utility of a seizure warning system depends upon the false positive rate as well as the sensitivity of the system. The system utilized in the present study produces false warnings at an average of 6.5 hours, depending upon the parameter settings. Further investigations are required to determine the cause of these false warnings. Several explanations are plausible. The value of STL_{max} is only one dynamical feature of the EEG signals. In theory, there is one Lyapunov exponent for each dimension of a system. STL_{max} is an estimate of only the largest Lyapunov exponent in a multidimensional system. Knowledge of additional Lyapunov exponents may make it possible to distinguish between a true pre-ictal transition and other conditions in which there is convergence of the largest Lyapunov exponent. Other potential measures to characterize different aspects of the dynamics of a system also exist, such as the correlation dimension, Kolmorogov-Sanai entropy, or other estimates of complexity. Further investigation is required to determine whether other measures, or some combination of these measures, may provide a means to distinguish between true and false detections of the pre-ictal state. It is also possible that the false warnings correctly detect a pre-ictal or seizure susceptibility state, but normal physiological resetting mechanisms intervene returning the brain to a more normal dynamical state. Finally, it is possible that the dynamics of the pre-ictal transition are not unique and may be found in other physiological states.

This pre-clinical research forms a bridge between seizure prediction research and the implementation of seizure prediction/warning devices, which is a revolutionary approach for handling epileptic seizures, very similar to the brain-pacemaker. It may also lead to clinical investigations of the effects of medical diagnosis, drug effects, or therapeutic intervention during invasive EEG monitoring of epileptic patients. Future research towards the treatment of human epilepsy and therapeutic intervention of epileptic activities, as well as the development of seizure feedback control devices, may be feasible. Thus, it represents a necessary first step in the development of implantable biofeedback devices to directly regulate therapeutic pharmacological or physiological intervention to prevent impending seizures or other brain disorders. For example, such an intervention might be achieved by electrical or magnetic stimulation (*e.g.*, vagal nerve stimulation) or by a timely release of an anticonvulsant drug. Another practical application of the proposed approach would be to help neurosurgeons quickly identify the epileptogenic zone without having patients stay in the hospital for the invasive long-time (10-14 days in duration) EEG monitoring. This research has the potential to revolutionize the protocol to identify the epileptogenic zone, which could drastically reduce the healthcare cost during the hospital stay for these patients. In addition, this protocol will help physicians identify epileptogenic zones without the necessity to risk patient safety by implanting depth electrodes in the brain. In addition, the results from this study could also contribute to the understanding of the intermittency of other dynamical neurophysiological disorders of the brain (*e.g.*, migraines, panic attacks, sleep disorders, and Parkinsonian tremors). This research could also contribute to the localization of defects (flaws), classification and prediction of spatiotemporal transitions in other high dimensional biological systems such as heart fibrillation and heart attacks.

7. Acknowledgements

Thanks are due to Prof. J.C. Sackellares, Prof. L.D. Iasemidis, Prof. P.R. Carney, D.-S. Shiau, who have been very helpful in sharing their expert knowledge on the brain dynamics and physiology and their fruitful comments and discussion. Research was partially supported by the Medical Research Service of the Department of Veterans Affairs, grants from the Department of Veterans Affairs Research, the NSF grants DBI-980821, EIA-9872509, and NIH grant R01-NS-39687-01A1.

REFERENCES

[1] G. G. Athanasiou, C. P. Bachas, and W. F. Wolf, *Invariant geometry of spin-glass states*, Phys. Rev. B, 35 (1987), pp. 1965–1968.

[2] F. Barahona, *On the computational complexity of spin glass models*, J. Phys. A, 15 (1982), pp. 3241–3253.

[3] F. Barahona, *On the exact ground states of three-dimensional ising spin glasses*, J. Phys. A, 15 (1982), pp. 611–615.

[4] F. Barahona, M. Grötschel, M. Jüger, and G. Reinelt, *An application of combinatorial optimization to statistical physics and circuit layout design*, Oper. Res., 36 (1988), pp. 493–513.

[5] W. Chaovalitwongse, P. M. Pardalos, L. D. Iasemidis, J. C. Sackellares, and D.-S. Shiau, *Optimization of spatio-temporal pattern processing for seizure warning and prediction*, U.S. Patent application filed August 2004, Attorney Docket No. 028724–150, 2004.

[6] W. Chaovalitwongse, P. M. Pardalos, L. D. Iasemidis, D.-S. Shiau, and J. C. Sackellares, *Applications of global optimization and dynamical systems to prediction of epileptic seizures*, P. M. Pardalos, J. C. Sackellares, L. D. Iasemidis, and P. R. Carney, editors, Quantitative Neuroscience, Kluwer Academic Publishers, Dordrecht, (2003) pp. 1–36.

[7] W. Chaovalitwongse, P. M. Pardalos, and O. A. Prokoyev, *A new linearization technique for multi-quadratic 0-1 programming problems*, Oper. Res. Lett., 32 (2004), pp. 517–522.

[8] C. E. Elger and K. Lehnertz, *Seizure prediction by nonlinear time series analysis of brain electrical activity*, European Journal of Neuroscience, 10 (1998), pp. 786–789.

[9] R. Horst, P. M. Pardalos, and N. V. Thoai, *Introduction to Global Optimization*, Kluwer Academic Publishers, Dordrecht, 1995.

[10] L. D. Iasemidis, *On the Dynamics of the Human Brain in Temporal Lobe Epilepsy*, Ph.D. Thesis, University of Michigan, Ann Arbor, 1991.

[11] L. D. Iasemidis, L. D. Olson, J. C. Sackellares, and R. S. Savit, *Time dependencies in the occurrences of epileptic seizures: a nonlinear approach*, Epilepsy Research, 17 (1994), pp. 81–94.

[12] L. D. Iasemidis, P. M. Pardalos, J. C. Sackellares, and D.-S. Shiau, *Quadratic binary programming and dynamical system approach to determine the predictability of epileptic seizures*, J. Comb. Optim., 5 (2001), pp. 9–26.

[13] L. D. Iasemidis, J. C. Principe, and J. C. Sackellares, *Measurement and quantification of spatiotemporal dynamics of human epileptic seizures*, M. Akay, editor, *Nonlinear Biomedical Signal Processing*, Wiley–IEEE Press, Vol. II (2000), pp. 294–318.

[14] L. D. Iasemidis, D.-S. Shiau, W. Chaovalitwongse, J. C. Sackellares, P. M. Pardalos, P. R. Carney, J. C. Principe, A. Prasad, B. Veeramani, and K. Tsakalis, *Adaptive epileptic seizure prediction system*, IEEE Transactions on Biomedical Engineering, 5 (2003), pp. 616–627.

[15] L. D. Iasemidis, D.-S. Shiau, P. M. Pardalos, and J. C. Sackellares, *Phase entrainment and predictability of epileptic seizures*, P. M. Pardalos and J. C. Principe, editors, Biocomputing, Kluwer Academic Publishers, Dordrecht, (2001) pp. 59–84.

[16] L. D. Iasemidis, D.-S. Shiau, J. C. Sackellares, and P. M. Pardalos, *Transition to epileptic seizures: Optimization*, D. Z. Du, P. M. Pardalos, and J. Wang, editors, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, American Mathematical Society, (1999) pp. 55–74.

[17] L. D. Iasemidis, H. P. Zaveri, J. C. Sackellares, and W. J. Williams, *Phase space analysis of eeg in temporal lobe epilepsy*, IEEE Eng. in Medicine and Biology Society, 10th Ann. Int. Conf., (1998) pp. 1201–1203.

[18] L. D. Iasemidis, H. P. Zaveri, J. C. Sackellares, and W. J. Williams, *Phase space topography of the electrocorticogram and the lyapunov exponent in partial seizures*, Brain Topography, 2 (1990), pp. 187–201.

[19] Y. C. Lai, I. Osorio, M. A. F. Harrison, and M.G. Frei, *Correlation-dimension and autocorrelation fluctuations in seizure dynamics*, Phys. Rev., 65 (2002), 3 Pt 1:031921.

[20] K. Lehnertz and C. E. Elger, *Spatio-temporal dynamics of the primary epileptogenic area in temporal lobe epilepsy characterized by neuronal complexity loss*, Electroencephalogr. Clin. Neurophysiol., 95 (1995), pp. 108–117.

[21] K. Lehnertz and C. E. Elger, *Can epileptic seizures be predicted? evidence from nonlinear time series analysis of brain electrical activity*, Phys. Rev. Lett., 80 (1998), pp. 5019–5022.

[22] B. Litt, R. Esteller, J. Echauz, D. A. Maryann, R. Shor, T. Henry, P. Pennell, C. Epstein, R. Bakay, M. Dichter, and G. Vachtservanos, *Epileptic seizures may begin hours in advance of clinical onset: A report of five patients*, Neuron, 30 (2001), pp. 51–64.

[23] R. Manuca and R. Savit, *Stationary and nonstationary in time series analysis*, Phys. D, 99 (1999), pp. 134–161.

[24] J. Martinerie, C. Van Adam, and M. Le Van Quyen, *Epileptic seizures can be anticipated by non-linear analysis*, Nature Medicine, 4 (1998), pp. 1173–1176.

[25] M. Mezard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond*, World Scientific, 1987.

[26] I. Osorio, M. A. F. Harrison, M. G. Frei, and Y. C. Lai. *Observations on the application of the correlation dimension and correlation integral to the prediction of seizures*, J. Clin. Neurophysiol., 18 (2001), pp. 269–274.

[27] P. M. Pardalos, W. Chaovalitwongse, L. D. Iasemidis, J. C. Sackellares, D.-S. Shiau, P. R. Carney, O. A. Prokopyev, and V. A. Yatsenko. *Seizure warning algorithm based on optimization and nonlinear dynamics*, Math. Program., 101 (2004), pp. 365–385.

[28] P. M. Pardalos and J. C. Principe, editors, *Biocomputing*, Kluwer Academic Publishers, Dordrecht, 2003.

[29] P. M. Pardalos, J. C. Sackellares, P. R. Carney, and L. D. Iasemidis, *Quantitative Neuroscience*, editors Kluwer Academic Publishers, Dordrecht, 2004.

[30] V. Piccone, J. Piccone, L. Piccone, R. LeVeen, and E. L. Veen, *Implantable epilepsy monitor apparatus*, US Patent 4, 566,464, 1981.

[31] M. Le Van Quyen, J. Martinerie, M. Baulac, and F. Varela, *Anticipating epileptic seizures in real time by non-linear analysis of similarity between eeg recordings*, NeuroReport, 10 (1999), pp. 2149–2155.

[32] J. C. Sackellares, L. D. Iasemidis, and D.-S. Shiau. *Detection of the preictal transition in scalp eeg*, Epilepsia, 40 (1999) pp. 176.

[33] J. C. Sackellares, L. D. Iasemidis, D.-S. Shiau, L. K. Dance, P. M. Pardalos, and W. Chaovalitwongse, *Optimization of multi-dimensional time series processing for seizure warning and prediction*, International Patent Application filed August 2003, Attorney Docket No. 028724–142, 2003.

[34] J. C. Sackellares, L. D. Iasemidis, D.-S. Shiau, W. Suharitdamrong, L. K. Dance, W. Chaovalitwongse, P. M. Pardalos, and P. R. Carney, *An automated seizure warning algorithm for scalp eeg*, Epilepsia, 44 (2003), pp. (S9):228.

[35] J. C. Sackellares, L. D. Iasemidis, V. A. Yatsenko, D.-S. Shiau, P. M. Pardalos, and W. Chaovalitwongse, *Multi-dimensional multi-parameter time series processing for seizure warning and prediction*, International Patent Application filed Septempber 2003, Attorney Docket No. 028724–143, 2003.

[36] D.-S. Shiau, L. D. Iasemidis, W. Suharitdamrong, L. K. Dance, W. Chaovalitwongse, P. M. Pardalos, and J. C. Sackellares, *Detection of the preictal period by dynamical analysis of scalp eeg*, Epilepsia, 44 (2003), pp. 233–234.

[37] S. Viglione, V. Ordon, W. Martin, and C. Kesler, *Epileptic seizure warning system*, US Patent 3, 863,625, 1975.

[38] A. Wolf, J. B. Swift, H. L. Swinney, and J. A. Vastano, *Determining lyapunov exponents from a time series*, Phys. D, 16 (1985), pp. 285–317.

Optimization in Intensity Modulated Radiation Therapy

Eva K. Lee
Center for Operations Research in Medicine,
School of Industrial and Systems Engineering,
Georgia Institute of Technology,
Atlanta, GA 30332-0205, USA
Winship Cancer Institute and Dept. of Radiation Oncology,
Emory University School of Medicine
(evakylee@isye.gatech.edu).

Joseph O. Deasy
Division of Bioinformatics and Outcomes Research,
Department of Radiation Oncology,
Washington University School of Medicine,
St. Louis, MO 63110, USA
(deasy@radonc.wustl.edu).

Abstract: An overview and some computational challenges in intensity modulated radiation therapy are presented. Experience with a mixed-integer programming treatment planning model is described. The MIP model allows simultaneous optimization over the space of beamlet intensity weights and beam and couch angles. The model uses two classes of decision variables to capture the beam configuration and intensities simultaneously. Binary (0/1) variables are used to capture "on" or "off" or "yes" or "no" decisions for each field, and nonnegative continuous variables are used to represent intensities of beamlets. Binary and continuous variables are also used for each voxel to capture dose level and dose deviation from target bounds. The treatment planning model was designed to explicitly incorporate the following planning constraints: (a) upper/lower/mean dose-based constraints, (b) dose-volume and equivalent-uniform-dose constraints for critical structures, (c) homogeneity constraints (underdose/overdose) for the planning target volume (PTV), (d) coverage constraints for PTV, and (e) maximum number of beams allowed. Results of applying the MIP Model to a patient case are presented. Brief discussions of recent linear programming and nonlinear programming treatment planning models are also described, as is an MIP approach for direct aperture optimization.

1. Introduction

Every year over 1.4 million new cancer cases are diagnosed [1] in the United States, and over half of the patients receive radiation treatment at some point during the course of their disease. The key to the effectiveness of radiation therapy for the treatment of cancer lies both in the fact that the repair mechanisms for cancerous cells are less efficient than that of normal cells, and the ability to deliver higher doses to the target volume using "cross-fired" radiation beam. Thus, a dose of radiation sufficient to kill cancerous cells may not be lethal for nearby healthy tissue. Nevertheless, avoiding or minimizing radiation exposure to healthy tissue is extremely important.

Using multiple beams of radiation from multiple directions to cross-fire at the tumor volume provides a method to keep radiation exposure to normal tissue at relatively low levels, while dose to tumor cells is elevated. The crux of the treatment planning process involves designing beam profiles (*i.e.*, a collection of beams) that delivers a sterilizing dose of radiation to the tumor volume, while dose levels to critical normal tissues are kept below established tolerance levels. Often, one attempts to design a plan for which the prescription dose isodose surface conforms to the geometric shape of the specified tumor volume [28, 67]. (The term *prescription dose* typically refers to the minimum dose desired to be delivered to the tumor volume; it is generally physician specified.)

Linear accelerators (LINAC) are common beam delivery units used for external beam radiotherapy. The table on which the patient lies and the beam delivery mechanism for the LINAC rotate about separate orthogonal axes, providing the ability to adjust the angle and entry point of radiation fields used during treatments. Each field is further defined by a bank of multileaf collimators (MLC), small metallic leaves located inside the LINAC treatment unit. These leaves can be opened or closed, and used to shape the radiation beam as it exits the machine. Figure 1 shows a linear accelerator.

Intensity-modulated radiation therapy (IMRT) is an important recent advance in radiation therapy [68]. In conventional radiotherapy treatment, the planning process consists of determining a set of external beams that meet, as best as possible, the clinical dose distribution criteria. In many cases, significant compromises to critical structure function have to be made to enable a tumoricidal dose to

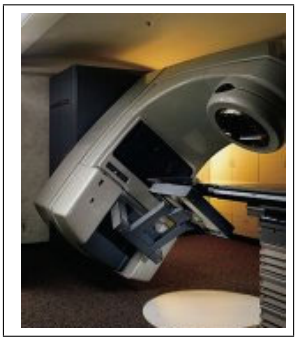

Figure 1: A linear accelerator used for external beam radiotherapy treatment

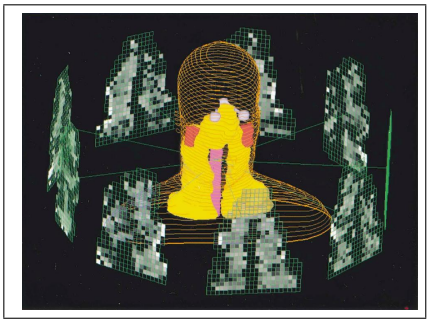

Figure 2: The treatment of a head-and-neck case via IMRT. Shown is a 3D view of the patient, the planning tumor volume (PTV), yellow; the spinal cord, pink; and the parotid glands, red. The 9 beams, shown with gray levels, reflect the modulated radiation intensity. (Use with permission from [2])

be delivered to the targets. In IMRT, the radiation fluence is varied across the beam, which allows a higher degree of conformation to the tumor than previously possible and allows concave isodose profiles to be generated, which may block, for example, dose to critical structure anterior or posterior to the target from that view. Specifically, not only is the shape of the beam controlled, but combinations of open and closed multileaf collimators modulate the intensity as well. For this reason, IMRT provides improved delivery control over conventional treatment. Indeed, it provides an unprecedented capability to dynamically vary the dose to accommodate the shape of the tumor from different angles, and to spare normal tissues and organs-at-risk (OAR) that may be potentially harmed during treatment.

Due to the complexity of the beam intensity profile associated with IMRT, there has been a tremendous research effort among medical physicists and radiation oncologists related to IMRT treatment planning and delivery, and there remain many opportunities for computational advances, particularly in treatment design. A computer-driven optimization algorithm must be used to determine the beam fluences (intensity maps) that provide the best compromise between target underdosing, target overdosing and critical structure overdosing. The textbook by Webb [68] has a good list of references for IMRT optimization.

In Sections 2.1 and 2.2, we describe the treatment planning problem for IMRT, and discuss relevant input data and the dose matrix. In Section 2.3, we discuss our experience of a mixed integer programming treatment planning model. The mixed integer programming model allows one to simultaneously incorporate dose coverage, underdose, overdose, homogeneity and conformity criteria on the tumor volume; dose volume restrictions on the critical structures (how much volume can receive more than a specified dose); and physical constraints on the total number of beams. Section 2.4 describes briefly the associated clinical results, and Section 2.5 provides a very brief discussion of current mathematical programming approaches. Summary and discussion is presented in Section 3.

2. Intensity-modulated radiation therapy treatment planning

Treatment planning in intensity modulated radiation therapy consists of a sequence of steps:

- Acquiring a 3D image of the affected region
- Delineating target volumes and healthy anatomical structures
- Selecting the appropriate radiation source and energy
- Selecting a set of beam angles for use in treatment
- Computing dose from each beam
- Performing intensity map optimization for the selected beams
- Developing optimal collimator sequences for actual delivery

These steps can be performed sequentially, or some can be combined together, resulting in complex numerical problems. In the sections presented herein, much of the description follows our recent work on mixed integer programming in this area [35, 36]. Very brief discussions on linear programming and nonlinear programming approaches are included.

2.1 Input data and dose calculation

Image Acquisition and Segmentation. The planning process begins when the patient is diagnosed with a tumor mass and radiation is selected as part of the treatment regime. A 3D image, or volumetric studyset, of the affected region, which contains the tumor mass and the surrounding areas, is acquired via computed tomography (CT) scans. These CT data are used for treatment planning, and electron density information derived from them are used in the photon dose calculations. Additionally, magnetic resonance imaging (MRI) scans may be acquired, fused with the CT volumetric studyset, and used to more accurately identify the location and extent of some tumors — especially those in the brain. Based on these scans, the physician outlines the tumor, and also outlines anatomic structures that need to be held to a low dose during treatment.

It is common practice to identify three "volumes" associated with the tumor. The gross tumor volume (GTV) represents the volume that encompasses the imageable or palpable macroscopic disease; that is, the disease that can be detected and localized by the oncologist. The clinical target volume (CTV) expands the GTV to include regions of suspected microscopic disease. The delineation of the CTV depends heavily on *a priori* knowledge of the behavior of a given tumor type. For a given GTV, tumor histologic features, and patient type, a set of probabilities exist (imperfectly known) that the tumor will, or will not, extend microscopically into a given regional organ or lymph node. However, accurate specific data are usually not available to the radiation oncologist, only general principles are known. A more quantitative and consistent definition of the CTV is an important need. The planning target volume (PTV) includes additional margins for anatomical and patient setup uncertainties related to organ and patient movement over time. All volumetric data is discretized into voxels (point representations of volume properties) at a granularity that is conducive both to generating a realistic model and to ensuring that the resulting treatment planning instances are tractable (*i.e.*, capable of being solved in a reasonable amount of computational time for practical clinical usage).

Dose Calculation Radiation dose, measured in Gray (Gy), is energy (Joules) deposited locally per unit mass (kg). Fluence for external beam photon radiation is defined mathematically by the number of photon crossings per surface area. Dose tends to be proportional to fluence, but is also influenced by photons and electrons scattered in the patient's tissues as well as the incident energy and media involved.

The calculation of the dose distribution associated with IMRT delivery is a critical aspect of the IMRT optimization and delivery processes. The calculated dose distribution from each candidate set of plan parameters is evaluated at each iteration or at the end of the optimization process, and the objective function values (costs or scores) for the iterative optimization are typically obtained by analysis of the dose distribution. For most systems, after the fluence-optimized plan is obtained, another dose calculation/optimization procedure, called leaf sequencing, is performed which first breaks the beams up into machine-deliverable multileaf sequencing steps, and then includes a final dose calculation step based on the details of the multileaf field shapes.

One of the most commonly used IMRT dose calculation algorithms involves a simple pencil beam method and is usually part of a broader class of correction-based dose-calculation algorithms [40, 4]. While these models offer significant speed advantages for use in the optimization code, they have varying limitations in accuracy.

In contrast, convolution/superposition, energy deposition kernel-based approaches can take into account beam energy, geometry, beam modifiers, patient contour, and electron density distribution [41, 10, 3, 6, 43]. Both the convolution method and the Monte Carlo method compute the dose per unit energy fluence (or fluence) incident on the patient.

Although it is clear that improved dose-calculation accuracy afforded by the convolution-type calcula-

tions may be important for IMRT, the long calculation times make this difficult.

Recently, significant progress has been made in the development of Monte Carlo calculation algorithms for photon beams, which simulate particle tracks individually, that are fast enough to compete with other current methods [24, 39, 70, 57]. In several situations, the Monte Carlo method is likely to be even more accurate than the convolution method [71]. For example, multiple scatter (second and higher order scatter) may be perturbed near the surface of a patient and the Monte Carlo method may be able to account for this as long as the number of simulated particles is sufficient. Direct Monte Carlo simulation may be the only option for achieving accurate dose computations in these complex situations. However, the application of Monte Carlo methods to optimization for IMRT is an area that requires much more work before relevant results will be available.

Access and usage of realistic radiotherapy data can be facilitated by using an open-source toolbox, developed by Deasy *et al.* [21], which enables users to import clinical plan data into Matlab for viewing and manipulation, and furthermore includes tools to generate the dose influence matrices.

2.2 Treatment planning strategies

In a strategy known as forward treatment planning, the beam geometry (beam orientation, shape, modifier, beam weights, etc.) is first defined, followed by calculation of the 3D dose distribution. After qualitative review of the dose distribution by the treatment planner and/or radiation oncologist, plan improvement is often attempted by modifying the initial geometry (*e.g.*, changing the beam weights and/or modifiers, adding another beam), to improve the target dose coverage and/or decrease the dose in the organs at risk. This forward planning process is repeated until a satisfactory plan is generated. As one can imagine, this is a time consuming approach to treatment planning.

In newer inverse treatment planning, the focus is on the desired outcome (*e.g.*, a specified dose distribution or tumor control probability (TCP) and normal tissue complication probability (NTCP)) rather than on how the outcome is achieved. The user of the system specifies the goals; the computer (optimization system) then adjusts the beam parameters (mainly the intensities) iteratively in an attempt to achieve the desired outcome. After review of the computer optimized dose distribution, some modification of the desired outcome and adjustment of the relative importance of each end point might be needed if the physician is not satisfied with the dose to the target volume or organs-at-risk (OARs).

Clearly, optimization is a classical inverse planning approach: constraints and an objective function are utilized to guide the optimization solver to select a plan with pre-specified clinical properties. Beginning with the work of Bahr *et al.* [5] in the late 60's, a number of research articles, authored primarily by medical researchers, discussed the use of mathematical programming and other optimization techniques in conventional external beam radiation treatment planning [18, 30, 31, 32, 34, 56, 61, 65, 73].

Much of IMRT treatment planning research has focused on the determination of the fluence map [67, 2, 7, 8, 11, 13, 44, 45, 72, 26, 12, 15, 19, 25, 27, 23]; that is, the radiation intensity or beam weights associated to each of the small beamlets of a selected radiation field/beam. However, the determination of beam angles, shapes, modifiers, couch positions and radiation energy to be used are best modeled using discrete variables.

At present, most IMRT optimization systems use dose-based and/or dose-volume-based criteria. One method commonly used to create dose-based and dose-volume objective functions involves minimizing the variance of the dose relative to the prescribed dose for the target volumes or dose limits for the organs at risk. This type of objective function has been used for traditional radiation therapy treatment optimization for the past several decades [62]. Variance is defined as the sum of the squares of the differences between the calculated dose and the prescribed dose or dose limit. Thus, a typical dose-based or dose-volume-based objective function is the sum of the variance terms representing each anatomic structure multiplied by appropriate penalty factors (*i.e.*, importance factors). Just as in conventional radiation therapy [14], the resulting unconstrained quadratic programming problem is often solved via the gradient method [61, 72], although the inclusion of dose-volume constraints makes the problem non-convex [20].

Within the optimization community, linear programming and nonlinear programming have been used to determine the optimal intensity map [55, 58], while mixed integer programming has been introduced to simultaneously determine the optimal beam angles and beam intensities [35, 36], and in finding optimal apertures for radiation delivery [52]. Below, we describe the MIP models formulated for simultaneous beam angle and intensity map optimization, closely following the presentation in Lee *et al.* [35, 36]. Results from a patient case will be briefly summarized. We then briefly describe linear and nonlinear programming approaches by others. Besides mathematical programming approaches, heuristic approaches — such as simulated annealing and genetic algorithms — have been commonly used for radiation therapy treatment optimization.

2.3 Mixed integer programming treatment planning models

The treatment planning models in [35, 36] use two classes of decision variables to capture the beam configuration and intensities simultaneously: Binary (0/1) variables are used to capture "on" or "off" or "yes" or "no" decisions for each field, and nonnegative continuous variables are used to represent intensities of beamlets. Binary and continuous variables are also used for each voxel to capture dose level and dose deviation from target bounds. Below, we provide the mathematical description of the treatment planning models.

Let $\mathcal{B}$ denote the set of candidate beams (each with an associated beam angle), and let $\mathcal{N}_i$ denote the set of beamlets (discretized sub-beams — usually rectangular in cross-section — which comprise the beam) associated with beam $i \in \mathcal{B}$. Beamlets associated with a beam can only be used when the beam is chosen to be "on". If a beam is on, the beamlets with positive dose intensity will contribute a certain amount of radiation dosage to each voxel in the target volume and other anatomical structures. Once the set of potential beamlet intensities is specified, the total radiation dose received at each voxel can be modelled. Let $w_{ij} \geq 0$ denote the intensity of beamlet j from beam i (in calibrated monitor units). Then the total radiation dose at a voxel P is given by

$$D_P(w) = \sum_{i \in \mathcal{B}} \sum_{j \in \mathcal{N}_i} D_{P,ij}\ w_{ij}, \tag{1}$$

where $D_{P,ij}$ denotes the dose per monitor unit intensity contribution to voxel P from beamlet j in beam i. Various dose constraints are involved in the design of treatment plans. Clinically prescribed lower and upper bounds, say L_P and U_P, for dose at voxel P are incorporated with (1) to form the basic dosimetric constraints:

$$\sum_{i \in \mathcal{B}} \sum_{j \in \mathcal{N}_i} D_{P,ij}\ w_{ij} \geq L_P, \quad \sum_{i \in \mathcal{B}} \sum_{j \in \mathcal{N}_i} D_{P,ij}\ w_{ij} \leq U_P. \tag{2}$$

Our model also allows selection of optimal beam angles out of a collection of candidate beams. Thus, the resulting plan returns the optimal beam geometry as well as beam intensities.

Let x_i be a binary variable denoting the use or non-use of beam i. The following constraints limit the total number of beams used in the final plan and ensure that beamlet intensities are zero for beams not chosen:

$$\sum_{i \in \mathcal{B}} x_i \leq B_{\max} \qquad \text{and} \qquad w_{ij} \leq M_i x_i. \tag{3}$$

Here, $B_{\max}$ is the maximum number of beams desired in an optimal plan, and M_i is a positive constant that can be chosen as the largest possible intensity emitted from beam i.

For each voxel in each anatomical structure, we associate one binary variable and one continuous variable to capture whether or not desired dose level is achieved and the deviation of received dose from desired dose. We also impose additional constraints into our treatment plan design, as discussed below.

Clinically, it may be desirable to incorporate coverage constraints within the model. For example, the clinicians may consider that it is acceptable if, say, 95% of the PTV receives the prescription dose, *PrDose*. Such a coverage requirement can be modelled as follows.

$$\sum_{i \in \mathcal{B}} \sum_{j \in \mathcal{N}_i} D_{P,ij} w_{ij} - r_P = PrDose,\ P \in PTV \tag{4}$$

$$r_P \leq D^{OD}_{PTV} v_P \quad (5)$$

$$r_P \geq D^{UD}_{PTV}(v_P - 1) \quad (6)$$

$$\sum_{P \in PTV} v_P \geq \alpha |PTV|. \quad (7)$$

Here, r_P is a real-valued variable that measures the discrepancy between prescription dose and actual dose; v_P is a 0/1 variable that captures whether voxel P is above or below the prescription dose bounds or not; α corresponds to the minimum percentage of coverage required (*e.g.*, $\alpha = 0.95$); D^{OD}_{PTV} and D^{UD}_{PTV} are the maximum overdose and maximum underdose levels tolerated for tumor cells; and $|PTV|$ represents the total number of voxels used to represent the planning target volume. The values D^{OD}_{PTV} and D^{UD}_{PTV} can be chosen according to the homogeneity level desired by the clinician for the resulting plan. If $r_P > 0$, then voxel P receives sufficient radiation dose to cover the prescribed dose. In this case, $v_P = 1$ and the amount of radiation for voxel P above the prescribed dose is controlled by the maximum-allowed-overdose constant, D^{OD}_{PTV}. Similarly, when $r_P < 0$, voxel P is underdosed, and the amount of underdose is limited by D^{UD}_{PTV}. In this case, $v_P = 0$.

By design, constraints (5) and (6) serve two purposes: 1) they capture the number of PTV voxels satisfying the prescription dose, and 2) they provide a means of controlling underdose, overdose, and dose homogeneity in the tumor. For the latter, the ratio $(PrDose + D^{OD}_{PTV})/(PrDose - D^{UD}_{PTV})$ can be viewed as an implied PTV homogeneity constraint associated with the model. Using a model with a smaller homogeneity constraint can be expected to result in a more homogeneous plan. Constraint (7) corresponds to the coverage level desired by the clinician.

Recently Equation (4) has been used to capture dose gradient fall-off when 100% tumor coverage is demanded. This was achieved by minimizing the dose surrounding the tumor region [34]. For IMRT planning optimization, it alone was used to model the deviation from prescribed dose for the PTV [13, 72, 17]. In these studies, a nonlinear objective function was formulated to steer the gradient-based optimization engine towards achieving the prescribed dose for the target volume; specifically, the objective was to minimize the sum of dose deviation across the target volume: $\|r\|_q = (\sum_P |r_P|^q)^{1/q}$ (with no imposed constraints). When $q = 2$, this is a least-squares problem.

It is desirable that dose received by radiation sensitive organs/tissues other than the tumor volume should be controlled to reduce the risk of injury. Thus, for other anatomical structures involved in the planning process, along with the basic dose constraints given in (2), additional binary variables are employed for modeling the dose-volume-tolerance relationships. To incorporate this concept into the model, let $\alpha_k,\ \beta_k \in (0, 1]$ for k in some index set K. (In our implementations, the cardinality of the index set K is typically between 3 and 10 but could be larger.) The following set of constraints ensures that at least $100\beta_k\%$ of the voxels in an organ-at-risk, OAR, receive dose less than or equal to $\alpha_k\, PrDose$. The symbols $y^{\alpha_k}_P$ and z^{OAR}_P denote binary variables.

$$\sum_{i \in \mathcal{B}} \sum_{j \in \mathcal{N}_i} D_{P,ij} w_{ij} \leq [\alpha_k\, PrDose] y^{\alpha_k}_P + \mathcal{D}_{max} z^{OAR}_P,\ P \in OAR \quad (8)$$

$$\sum_{P \in OAR} y^{\alpha_k}_P \geq \beta_k |\, OAR\,| \quad (9)$$

$$y^{\alpha_k}_P + z^{OAR}_P = 1 \quad (10)$$

$$y^{\alpha_{k_1}}_P \leq y^{\alpha_{k_2}}_P \text{ for } \alpha_{k_1} \leq \alpha_{k_2}. \quad (11)$$

Here, $\mathcal{D}_{max}$ is the maximum dose allowed for OAR (often determined by the maximum dose thought to be well-tolerated), and $\alpha_k,\ \beta_k$ combinations are patient and tumor specific. When the total dose received by a voxel P is less than $\alpha_k\, PrDose$, $y^{\alpha_k}_P = 1$, and this contributes to a voxel count in Constraint (9). When it does not satisfy the dose bound $\alpha_k\, PrDose$, then $y^{\alpha_k}_P = 0$, and in this case the dose will be forced to be lower than the maximum dose tolerance allowed, $\mathcal{D}_{max}$, and $z^{OAR}_P = 1$. Note that by using discrete variables to represent each voxel and controlling the number of points satisfying a certain dose level, we can impose strict dose-volume criteria within the solution space. This is in contrast to the common approach of incorporating "soft" dose-volume criteria into a composite objective function [72]. Langer[31] was the first to apply MIP ideas to model dose-volume relationships in conventional radiation therapy. For IMRT, the challenge is that the resulting problem instances are large-scale (involving hundreds of thousands or even millions of in-

equalities), and are computationally taxing and difficult to solve without the development of specialized algorithms [35].

Besides the commonly used least-squares dose deviation objective function, other objective functions have been used, including: minimizing the squared radiation dose to OARs, maximizing the minimum dose to tumor target, maximize/minimize weighted sum of doses to target and OARs. Other more complex biological objective functions — involving equivalent uniform dose (the p-norm or generalized mean value), tumor control probability, and normal tissue complication probability — have also been proposed [37, 38, 60, 59, 47, 50, 48, 49, 46].

The MIP treatment planning models for real patient cases involve tens to hundreds of thousands of binary variables and constraints. Our experience is that the resulting MIP instances are intractable via commercial MIP solvers. However, we have observed that, by using specialized algorithms [35], clinically superior treatment plans can be obtained [36].

2.4 Computational results for a real patient case

We briefly describe a patient study. Input data includes 3D images of tissue to be treated. On these images, the planning target volume (PTV) is delineated, and contours of organs-at-risk (OAR) and normal tissue are outlined. In addition to these structures, a tissue ring of 5 mm thickness is drawn around the PTV. We call this ring the *critical-normal-tissue-ring*. In [31] it was demonstrated that this normal tissue construct can assist in obtaining conformal plans for radiosurgery. In [35, 36], we have shown its usefulness in designing conformal IMRT plans. For the results herein, depending on the volume of the anatomical structure, a 3–5 mm voxel size (for dose computation) is used for setting up the MIP model instances.

For each beamlet, the dose per monitor unit intensity to a voxel is calculated. The total dose per unit intensity deposited to a voxel is equal to the sum of dose per intensity deposited from each beamlet. For the results described here, 16-24 coplanar fields of size 10×10 cm^2 to 15×15 cm^2 are generated as candidate fields, each of which consists of 400-900 0.5×0.5 cm^2 beamlets. This results in a large set of candidate beamlets used for instantiating the treatment planning models.

In [35], we study the effect of maximum beam angles allowed on plan quality. In [36], five objective functions are considered and contrasted on three different tumor sites to compare plan quality and to gain understanding of the steering effects of clinical objectives. Below, we illustrate the results for a head-and-neck case obtained via multiple objectives.

Some common metrics for reporting quality of treatment plans include:

- Coverage — Coverage is computed as the ratio of the target volume enclosed by the prescription isodose surface to the total target volume. Coverage is always less than or equal to 1.
- Conformity — Conformity is a measure of how well the prescription isodose surface conforms to the target volume; it is computed as the ratio of the total volume enclosed by the prescription isodose surface to the target volume enclosed by this same surface. Conformity is always greater than or equal to 1.
- Homogeneity — The homogeneity index is defined as the ratio of the maximum dose to the minimum dose received by the tumor volume.
- Mean dose and maximum dose for each critical structure.
- Dose-volume histograms (volume receiving more than each given dose level) and isodose curves.

Observe that these metrics are not entirely independent. For example, while it is desirable to obtain a prescription isodose surface big enough to cover the target volume in order to ensure good coverage, it is also desirable to have this surface "small" in order to conform to the target volume. In addition, variations in conformity and coverage affect the amount of irradiation to nearby organs at risk, thus affecting dose distribution levels of these organs.

Head-and-neck tonsil cancer. We focus on a tonsil cancer case where the PTV is adjacent to the left submandibular salivary gland. The following structures with their respective clinical dose limits are considered. PTV should receive 68 Gy; left parotid: $30\% \leq 27$ Gy and $100\% \leq 68$ Gy; right parotid: $100\% \leq 15$ Gy; right submandibular gland:

100% $\leq$ 30 Gy; left submandibular gland: 10% $\leq$ 27 Gy and 100% $\leq$ 68 Gy; larynx: 80% $\leq$ 30 Gy and 100% $\leq$ 55 Gy; spinal cord and brainstem: 100% $\leq$ 45 Gy.

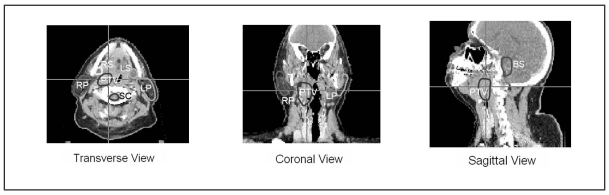

Figure 3: Anatomical structures for the head-and-neck. Notation: right parotid (RP), left parotid (LP), right submandibular gland (RS), left submandibular gland (LS), spinal cord (SP), brain stem (BS).

A total of 1501 PTV voxels, 406 critical-normal-tissue-ring voxels, 3247 voxels for the OARs and 6416 normal tissue voxels were used to instantiate the MIP treatment model.

Here, we report the results for a plan with a maximum of 7 beams in which the objectives include minimizing the total dose to the critical structures and optimizing the PTV conformity. The results are based on the utilization of a specialized branch-and-bound MIP solver for large-scale external beam radiation [35] that is built on top of a general-purpose mixed integer research code (MIPSOL) [33]. Figure 4 shows the dose volume histograms, and Figure 5 shows the isodose curves. Compared to the clinical plan, we observe the following:

a. For all critical structures, the mean dose and max dose received are drastically less than the clinical plan.

b. For OARs that are close to the tumor volume, namely the left parotid ($< 10mm$) and the left submandibular gland ($< 10mm$), the mean dose received is significantly reduced (70% and 50%, respectively). The spinal cord enjoys moderate dose reduction (33%).

c. The coverage constraint and the objective helped in achieving 98% coverage. Underdose and overdose constraints kept minimum and maximum dose to the tumor relatively uniform, with a homogeneity index of 1.24. And the conformity objective helped to achieve a superior conformity value of 1.34. These all improve over the clinical plan, which had 97% coverage, and scores of 1.4 for homogeneity, and 1.6 for conformity.

d. The total overall monitor units of radiation from the MIP optimized plan is less than that from the clinical plan, indicating that the plan uses less radiation but yet can still deliver the required prescription dose to the tumor, thus sparing excessive radiation dose to the critical structures and normal tissue.

It is noteworthy that in contrast to the typically equispaced beams chosen when beam configurations are pre-selected, the optimal 7-beam plans obtained herein (that are considered clinically acceptable) do not have equispaced beams. Indeed, the optimal beam angles returned appear to be non-intuitive, and to depend on PTV size and geometry and the spatial relationship between the tumor and the critical structures.

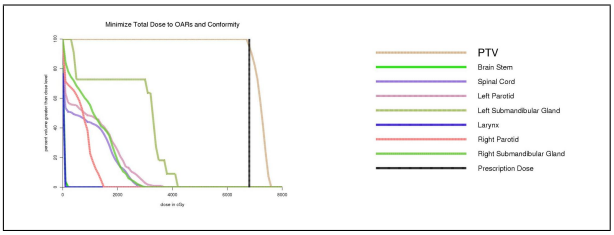

Figure 4: Dose-volume histogram for the head-and-neck for the MIP model with objective of minimizing the OARs dose and optimizing prescription dose conformity to tumor.

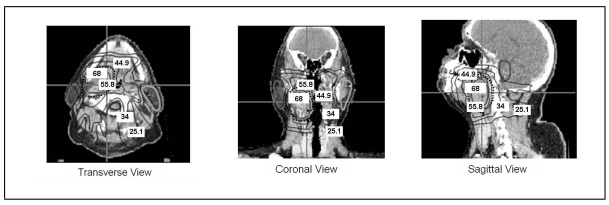

Figure 5: Isodose curves for the head-and-neck case. The critical-normal-tissue-ring is represented by the dotted curve.

2.5 Other mathematical programming approaches

As previously mentioned, linear and nonlinear programming have long been used for radiation therapy treatment optimization [5, 56, 55]. Lacking discrete variables, LP and NLP models typically use a preselected beam configuration, and focus on determining beam intensities. Below, we briefly outline some recent approaches in this area.

Simplified least-squares objective function and dose-volume constraints: In [55], using pre-selected beam angles, linear programming approaches were used to determine the associated optimal intensity map. The authors approximated the least-squares objective function measuring deviation of tumor voxel dose from prescribed dose via a piecewise linear function. They also utilized conditional value-at-risk (CVaR) constraints to control the mean dose received by subsets of voxels receiving the highest or lowest doses among all voxels in a given structure. Two forms of such constraints were used:
(i) lower α-CVaR: The average dose received by the subset of a target of relative volume 1-α receiving the lowest doses must be at least equal to L^α.
(ii) upper α-CVaR: The average dose received by the subset of a structure of relative volume 1-α receiving the highest doses may be no more than U^α.
CVaR constraints were originally proposed by Rockafellar and Uryasev [54] to formulate risk management constraints in terms of the tail means of distributions of financial risk. Mathematically, the upper α-CVaR constraint on a structure S is defined as

$$\bar{\zeta}_S^\alpha(w)+\frac{1}{(1-\alpha)|S|}\sum_{j\in S}\max\{0, D_P(w)-\bar{\zeta}_S^\alpha(w)\} \le U_S^\alpha, \tag{12}$$

where U_S^α is an upper bound target, $D_P(w)$ is the total dose from intensity vector w for voxel P, and $\bar{\zeta}_S^\alpha(w)$ denotes the smallest dose level with the property that no more than $100(1-\alpha)$ percent of the structure S receives a larger dose. The authors showed that including such partial-volume constraints to bound the tail averages of the differential dose-volume histograms of structures helps to improve dose homogeneity to the target and to spare dose to critical structures.

Nonlinear programming approach: Sheperd *et al.* [58] summarized several LP and NLP models for determining optimal intensity maps. To model dose-volume constraints, they applied a nonlinear error function approach. Their problem involved minimizing the standard objective of sum of the square differences between the prescribed and the actual doses over all of the voxels in the tumor, subject to two partial volume constraints — to OARs and to normal tissue. For an OAR S, the partial volume constraint defined on S was:

$$\sum_{P\in S} erf(D_P(w)-\Lambda_P) \le \alpha|S| \tag{13}$$

where Λ denotes a selected dose limit and α denotes the fraction of the volume allowed to exceed this limit. The error function *erf(x)* realizing the partial volume constraints is a nonlinear function. (See fig. 4.4 in [58].)

The authors also compared this with an MIP approach to model partial volume constraints on OARs, involving a simplified version of constraints (8)–(11) described above.

Direct aperture approaches via mixed integer programming: Preciado-Walters *et al.* [52] formulated the treatment planning problem as a mixed integer program over a coupled pair of column generation processes: the first designed to produce intensity maps for the IMRT beamlet grid, followed by the second to specify protected area choices aiding in reducing the computational burden of enforcing the dose-volume restrictions on tissues.

Instead of determining the beamlet intensity for each beam, and then applying leaf-sequencing to determine delivery patterns, the planning involved first selecting a fixed set of deliverable beams. For each of these beams, the authors pre-determined heuristically a set of delivery patterns. They then introduced continuous nonnegative decision variables x_{jq} to represent the assigned intensity to whole pattern q of beam j. Then the dose at any voxel P, is calculated by

$$D_P = \sum_{j=1}^{n}\sum_{q\in\mathcal{Q}_j} a_{Pjq}x_{jq} \tag{14}$$

where $x_{jq} \ge 0$, $\mathcal{Q}_j$ is the set of patterns for beam j, and a_{Pjq} is the implied dose coefficient of pattern q from beam j at voxel P when the pattern q is constructed.

The resulting MIP model for treatment planning employed the objective function of maximizing the minimum tumor dose. Similar to the above MIP models, the constraints include upper and lower dose bounds on tumor voxels, and upper dose bounds for healthy tissues. Dose-volume constraints are formulated just as constraints (8)–(11) above.

3. Summary and discussion

This article provides a brief overview of optimization issues in intensity-modulated radiation therapy, and summarizes our experience with an integer programming approach. The MIP model described allows simultaneous optimization over the space of beamlet intensity weights and beam angles. Based on experiments with clinical data, this approach can return good plans that are clinically acceptable and practical. This work is distinguished from recent IMRT research in several ways. First, in previous methods beam angles are selected prior to intensity map optimization. Herein, we employ 0/1 variables to model the set of candidate beams, and thereby allow the optimization process itself to select optimal beams. Second, instead of incorporating dose-volume criteria within the objective function as in previous work, herein, a combination of discrete and continuous variables associated with each voxel provides a mechanism to strictly enforce dose-volume criteria within the constraints. The challenge of using MIP modeling for IMRT is that the resulting instances are very large-scale, and since general MIP is NP-hard, specialized algorithms designed to solve IMRT instances are required. Third, incorporating the critical-normal-tissue-ring can improve conformity in general tumor sites, without addition of other dose-shaping structures. In general, our MIP approach uses constraints to control a variety of clinical criteria (coverage, homogeneity, underdose to PTV, overdose to PTV, dose-volume limits on organs-at-risk and normal tissue), while assigning an objective to help with the solution search. The model can also be expanded to incorporate energy selection, couch angles and other treatment parameters.

Patient studies indicate that using the MIP approach, one can produce good clinical plans that aggressively lower OAR dose below pre-imposed levels without compromising local tumor control [36]. This is appealing since lower OAR dose should translate to lower normal tissue complication probability.

Computationally, the specialized optimization engine returns good feasible solutions within 30 minutes. We have performed standard leaf-sequencing techniques on the resulting optimal intensity map, and showed that returned plans are deliverable. The results provide evidence that the MIP approach is viable in producing good treatment plans that can potentially lead to significant improvement in local tumor control and reduction in normal tissue complication.

With pre-selected beam angles, other approaches such as linear programming [55] and nonlinear programming [72, 58] can be used for intensity map optimization. Comparisons are needed to gauge the quality of these plans versus those from MIP approaches. Direct aperture optimization [52] is appealing, since resulting segments are implementable directly. Again, comparisons are needed to determine the effectiveness, advantages and tradeoffs among different planning optimization methods.

Other computational challenges actively pursued by medical physics experts include image segmentation, planning under uncertainties, biological modeling, leaf-sequencing and treatment outcome analysis.

4. Acknowledgement

This research was partially supported by grants from the National Science Foundation, the National Institute of Health, and the Charles Edison Foundation.

REFERENCES

[1] American Cancer Society, *Source: Cancer Facts and Figures – 2006*, Atlanta, Georgia 2006.

[2] *Intensity-Modulated Radiotherapy: Current Status and Issues of Interest*, Intensity Modulated Radiation Therapy Collaborative, Working Group, Int. J. Radiat. Oncol. Biol. Phys., 51 (2001), pp. 880–914.

[3] A. Ahnesjo, *Collapsed cone convolution of radiant energy for photon dose calculation in heterogeneous media*, Med. Phys., 16 (1989), pp. 577–592.

[4] A. Ahnesjo and M. M. Aspradakis, *Dose calculations for external photon beams in radiotherapy*, Phys. Med. Biol., 44 (1999), pp. 99–155.

[5] G. K. Bahr, J. G. Kereiakes, H. Horwitz, R. Finney, J. Galvin, and K. Goode, *The method of linear programming applied to radiation treatment planning*, Radiology, 91 (1968), pp. 686–693.

[6] J. J. Battista and M. B. Sharpe, *True three-dimensional dose computations for megavoltage x-ray therapy: A role for the superposition principle*, Aust. Phys. Eng. Sci. Med., 15 (1992), pp. 159–178.

[7] T. Bortfeld, *Optimized planning using physical objectives and constraints*, Semin. Radiat. Oncol., 9 (1999), pp. 20–34.

[8] T. Bortfeld, K. Jokivarsi, M. Goitein, J. Kung, and S. B. Jiang, *Effects of intra-fraction motion on IMRT dose delivery: statistical analysis and simulation*, Phys. Med. Biol., 47 (2002), pp. 2203–2220.

[9] J. D. Bourland and E. L. Chaney, *A finite-size pencil beam model for photon dose calculations in three dimensions*, Med. Phys., 19 (1992), pp. 1401–1412.

[10] A. Boyer and E. Mok, *A photon dose distribution model employing convolution methods*, Med. Phys., 12 (1985), pp. 169–177.

[11] A. Brahme, *Development of radiation therapy optimization*, Acta Oncol., 39 (2000), pp. 579–595.

[12] M. P. Carol, *Integrated 3-D conformal multivane intensity modulation delivery system for radiotherapy*, Proceedings of the 11th International Conference on the Use of Computers in Radiation Therapy, Madison, WI, 1994.

[13] Y. Chen, D. Michalski, C. Houser, and J. M. Galvin, *A deterministic iterative least-squares algorithm for beam weight optimization in conformal radio therapy*, Phys. Med. Biol., 47 (2002), pp. 1647–1658.

[14] R. E. Cooper, *A gradient method of optimizing external-beam radiotherapy treatment plans*, Radiology, 128 (1978), pp. 235–243.

[15] C. Cotrutz and L. Xing, *Using voxel-dependent importance factors for interactive DVH-based dose optimization*, Phys. Med. Biol., 47 (2002), pp. 1659–1669.

[16] C. Cotrutz and L. Xing, *Segment-based dose optimization using a genetic algorithm*, Phys. Med. Biol., 48 (2003), pp. 2987–2998.

[17] S. M. Crooks and L. Xing, *Linear algebraic methods applied to intensity modulated radiation therapy*, Phys. Med. Biol., 46 (2001), pp. 2587–2606.

[18] S. Das and L. Marks, *Selection of coplanar and non-coplanar beams using three-dimensional optimization based on maximum beam separation and minimized nontarget irradiation*, Int. J. Radiat. Oncol. Biol. Phys., 38 (1997), pp. 643–655.

[19] C. De Wagter, C. O. Colle, L. G. Fortan, B. B. Van Duyse, D. L. Van den Berge, and W. J. De Neve, *3D conformal intensity-modulated radiotherapy planning: interactive optimization by constrained matrix inversion*, Radiot. Oncol., 47 (1998), pp. 69–76.

[20] J. O. Deasy, *Multiple local minima in radiotherapy optimization problems with dose-volume constraints*, Med. Phys., 24 (1997), pp. 1157–1161.

[21] J. O. Deasy, E. K. Lee, T. Bortfeld, M. Langer, K. Zakarian, J. Alaly, Y. Zhang, H. Liu, R. Mohan, R. Ahuja, A. Pollack, J. Purdy, and R. Rardin, *A collaboratory for radiation therapy treatment planning optimization research*, Ann. Oper. Res., to appear.

[22] D. Djajaputra, Q. Wu, Y. Wu, and R. Mohan, *Algorithm and performance of a clinical IMRT beam-angle optimization system*, Phys. Med. Biol., 48 (2003), pp. 3191–3212.

[23] B. A. Fraass, M. L. Kessler, D. L. McShan, L. H. Marsh, B. A. Watson, W. J. Dusseau, A. Eisbruch, H. M. Sandler, and A. S. Lichter, *Optimization and clinical use of multisegment intensity-modulated radiation therapy for high-dose conformal therapy*, Semin. Radiather. Oncol., 9 (1999), pp. 60–77.

[24] C. L. Hartmann Siantar, P. M. Bergstrom, W. P. Chandler, *et al.*, *Lawrence Livermore National Laboratory's PEREGRINE Project*, Proceedings of the XII International Conference on the Use of Computers in Radiation Therapy, Salt Lake City, Utah, 1997.

[25] M. Hilbig, R. Hanne, P. Kneschaurek, F. Zimmermann, and A. Schweikard, *Design of an inverse planning system for radiotherapy using linear optimization*, Med. Phys., 12 (2002), pp. 89–96.

[26] T. Holmes and T. R. Mackie, *A comparison of three inverse treatment planning algorithms*, Phys. Med. Biol., 39 (1994), pp. 91–106.

[27] D. H. Hristov and B. G. Fallone, *An active set algorithm for treatment planning optimization*, Med. Phys., 24 (1997), pp. 1455–1464.

[28] F. Khan, *The Physics of Radiation Therapy*, Williams and Wilkins, second edition, Baltimore, 1992.

[29] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, *Optimization by simulated annealing*, Science, 220 (1983), pp. 671–680.

[30] H. Kooy, L. Nedzi, J. Loeffler, E. Alexander, C. Cheng, E. Mannarino, E. Holupka, and R. Siddon, *Treatment planning for stereotactic radiosurgery of intracranial lesions*, Int. J. Radiat. Oncol. Biol. Phys., 21 (1991), pp. 683–693.

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

[32] M. Langer, S. Morrill, R. Brown, O. Lee, and R. Lane, *A comparison of mixed integer programming and fast simulated annealing for optimizing beam weights in radiation therapy*, Med. Phys., 23 (1996), pp. 957–964.

[33] E. K. Lee, *Computational Experience with a General Purpose Mixed 0/1 Integer Programming Solver (MIPSOL)*, Software Report, School of Industrial and Systems Engineering, Georgia Institute of Technology, 1997.

[34] E. K. Lee, T. Fox, and I. Crocker, *Optimization of radiosurgery treatment planning via mixed integer programming*, Med. Phys., 27 (2000), pp. 995–1004.

[35] E. K. Lee, T. Fox, and I. Crocker, *Integer programming applied to intensity-modulated radiation therapy treatment planning*, Ann. Oper. Res., 119 (2003), pp. 165–181.

[36] E. K. Lee, T. Fox, and I. Crocker, *Simultaneous beam geometry and intensity map optimization in intensity-modulated radiation therapy*, Int. J. Radiat. Oncol. Biol. Phys., 64 (2006), pp. 301–320.

[37] J. T. Lyman, *Complication probability as assessed from dose volume histograms*, Radiat. Res., 104 (1985), pp. 13–19.

[38] J. T. Lyman and A. B. Wolbarst, *Optimization of radiation therapy. III. A method of assessing complication probabilities from dose-volume histograms*, Int. J. Radiat. Oncol. Biol. Phys., 13 (1987), pp. 103–109.

[39] C.-M. Ma, E. Mok, A. Kapur, *et al.*, *Clinical implementation of a Monte Carlo treatment planning system*, Med. Phys., 26 (1999), pp. 2133–2143.

[40] T. R. Mackie, P. Reckwerdt, T. McNutt T, *et al.*, *Photon beam dose computations*, J. Palta and T. R. Mackie, editors, Teletherapy: Present and Future, Advanced Medical Publishing, College Park, MD (1996) pp. 103–136.

[41] T. R. Mackie, J. W. Scrimger, and J. J. Battista, *A convolution method of calculating dose for 15-MV x-rays*, Med. Phys., 12 (1985), pp. 188–196.

[42] G. S. Mageras and R. Mohan, *Application of fast simulated annealing to optimization of conformal radiation treatments*, Med. Phys., 20 (1993), pp. 639–647.

[43] R. Mohan, C. Chui, and L. Lidofsky, *Differential pencil beam dose computation model for photons*, Med. Phys., 13 (1986), pp. 64–73.

[44] R. Mohan, G. S. Mageras, B. Baldwin, *et al.*, *Clinically relevant optimization of 3-D conformal treatments*, Med. Phys., 19 (1992), pp. 933–944.

[45] R. Mohan, X. Wang, A. Jackson, *et al.*, *The potential and limitations of the inverse radiotherapy technique*, Radiot. Oncol., 32 (1994), pp. 232–248.

[46] A. Niemierko, *Reporting and analyzing dose distributions: A concept of equivalent uniform dose*, Med. Phys., 24 (1997), pp. 103–110.

[47] A. Niemierko and M. Goitein, *Calculation of normal tissue complication probability and dose-volume histogram reduction schemes for tissues with a critical element architecture*, Radiot. Oncol., 20 (1991), pp. 166–176.

[48] A. Niemierko and M. Goitein, *Modeling of normal tissue response to radiation: The critical volume model*, Int. J. Radiat. Oncol. Biol. Phys., 25 (1992), pp. 135–145.

[49] A. Niemierko and M. Goitein, *Implementation of a model for estimating tumor control probability for an inhomogeneously irradiated tumor*, Radiot. Oncol., 29 (1993), pp. 140–147.

[50] A. Niemierko, M. Urie, and M. Goitein, *Optimization of 3D radiation therapy with both physical and biological end points and constraints*, Int. J. Radiat. Oncol. Biol. Phys., 23 (1992), pp. 99–108.

[51] O. Z. Ostapiak, Y. Zhu, and J. Van Dyk, *Refinements of the finite size pencil beam model of three-dimensional photon dose calculation*, Med. Phys., 24 (1997), pp. 743–750.

[52] F. Preciado-Walters, R. Rardin, M. Langer, and V. Thai, *A coupled column generation, mixed integer approach to optimal planning of intensity modulated radiation therapy for cancer*, Math. Program., 101 (2004), pp. 319–338.

[53] A. B. Pugachev, A. L. Boyer, and L. Xing, *Beam orientation optimization in intensity-modulated radiation treatment planning*, Med. Phys., 27 (2000), pp. 1238–1245.

[54] R. Rockafellar and S. Uryasev, *Conditional value-at-risk for general loss distributions*, J. Banking Finance, 26 (2002), pp. 1443–1471.

[55] H. E. Romeijn, R. K. Ahuja, J. F. Dempsey, A. Kumar, and J. G. Li, *A novel linear programming approach to fluence map optimization for intensity modulated radiation therapy treatment planning*, Phys. Med. Biol., 48 (2003), pp. 3521–3542.

[56] I. I. Rosen, R. G. Lane, S. Morrill, and J. A. Belli, *Treatment plan optimization using linear programming*, Phys. Med. Biol., 18 (1991), pp. 141–152.

[57] J. Sempau, S. J. Wilderman, and A. F. Bielajew, *DPM, a fast, accurate Monte Carlo code optimized for photon and electron radiotherapy treatment planning dose calculations*, Phys. Med. Biol., 45 (2000), pp. 2263–2291.

[58] D. M. Shepard, M. C. Ferris, G. H. Olivera, and T. Rockwell Mackie, *Optimizing the delivery of radiation therapy to cancer patients*, SIAM Rev., 41 (1999), pp. 721–744.

[59] A. R. Smith and C. C. Ling, editors, *Implementation of three dimensional conformal radiotherapy*, Int. J. Radiat. Oncol. Biol. Phys., 33 (1995), pp. 779–976.

[60] A. R. Smith, J. A. Purdy, editors, *Three-dimensional photon treatment planning*, Report of the Collaborative Working Group on the Evaluation of Treatment Planning for External Photon Beam Radiotherapy, Int. J. Radiat. Oncol. Biol. Phys., 21 (1991), pp. 3–268.

[61] S. V. Spirou and C. S. Chui, *A gradient inverse planning algorithm with dose-volume constraints*, Med. Phys., 25 (1998), pp. 321–333.

[62] G. Starkschall, *A constrained least-squares optimization method for external beam radiation therapy treatment planning*, Med. Phys., 11 (1984), pp. 659–665.

[63] J. Stein, R. Mohan, X. Wang, T. Bortfeld, Q. Wu, K. Preiser, C. C. Ling, and W. Schlegel, *Number and orientations of beams in intensity-modulated radiation treatments*, Med. Phys., 24 (1997), pp. 149–160.

[64] H. Szu and R. Hartley, *Fast simulated annealing*, Phys. Lett. A., 122 (1987), pp. 157–162.

[65] U. Treuer, H. Treuer, M. Hoevels, P. Muller, and V. Sturm, *Computerized optimization of multiple isocenters in stereotactic convergent beam irradiation*, Phys. Med. Biol., 43 (1998), pp. 49–64.

[66] S. Webb, *Optimisation by simulated annealing of three-dimensional conformal treatment planning for radiation fields defined by a multileaf collimator*, Phys. Med. Biol., 36 (1991), pp. 1201–1226.

[67] S. Webb, *The Physics of Three-Dimensional Radiation Therapy: Conformal Radiotherapy, Radiosurgery and Treatment Planning*, Institute of Physics Publishing, Philadelphia, 1993.

[68] S. Webb, *Intensity-Modulated Radiation Therapy*, Institute of Physics Publishing, Bristol, 2000.

[69] S. Webb, D. J. Convery, P. M. Evans, *Inverse planning with constraints to generate smoothed intensity-modulated beams*, Phys. Med. Biol., 43 (1998), pp. 2785–2794.

[70] A. E. S. von Wittenau, L. J. Cox, P. M. Bergstrom, *et al.*, *Correlated histogram representation of Monte Carlo derived medical accelerator photon-output phase space*, Med. Phys., 26 (1999), pp. 1196–1211.

[71] M. K. Woo, J. R. Cunningham, and J. J. Jerioranski, *Extending the concept of primary and scatter separation to the condition of electronic disequilibrium*, Med. Phys., 17 (1990), pp. 588–595.

[72] Q. Wu and R. Mohan, *Algorithms and functionality of an intensity modulated radiotherapy optimization system*, Med Phys., 27 (2000), pp. 701–711.

[73] Y. Yu, M. Schell, and J. B. Zhang, *Decision theoretic steering and genetic algorithm optimization: Application to stereotactic radiosurgery treatment planning*, Med. Phys., 24 (1997), pp. 1742–1750.

Bulletin

1. Event Announcements

IPCO 2007
The Twelfth Conference on Integer Programming and Combinatorial Optimization
June 25–27, 2007
Cornell University, Ithaca, New York, USA
`http://ipco2007.orie.cornell.edu`

The IPCO conference is held every year, except for those years in which the 'Symposium on Mathematical Programming' takes place. The conference is meant to be a forum for researchers and practitioners working on various aspects of integer programming and combinatorial optimization. The aim is to present recent developments in theory, computation, and applications in that area. The scope of IPCO includes algorithmic and structural results in topics such as approximation algorithms algorithmic game theory, branch and bound algorithms, branch and cut algorithms, computational biology, computational complexity, computational geometry, cutting plane algorithms, diophantine equations, geometry of numbers, graph and network algorithms, integer programming, matroids and submodular functions, on-line algorithms and competitive analysis, polyhedral combinatorics, randomized algorithms, random graphs, scheduling theory and scheduling algorithms, semidefinite programs.

IPCO is not restricted to theory. Computational and practical work, implementations, novel applications of these techniques to practical problems, and revealing computational studies, are most welcome.

During the conference, approximately 30-35 papers will be presented in a series of non-parallel sessions. Each lecture will be 30 minutes long. The program committee will select the papers to be presented on the basis of extended abstracts to be submitted as described at the conference webpage (see above). The proceedings will contain full texts of all presented papers. Each participant will receive a copy at the conference.

Program Committee:

- Dimitris Bertsimas (MIT)
- Dan Bienstock (Columbia)
- Alberto Caprara (Bologna)
- Bill Cook (Georgia Tech)
- Gerard Cornuejols (CMU)
- Matteo Fischetti, Chair, Program Committee (Padova)
- Bertrand Guenin (Waterloo)
- Christoph Helmberg (TU Chemnitz)
- Tibor Jordn (ELTE Budapest)
- Tom McCormick (UBC)
- David Williamson, Chair, Local Arrangements (Cornell)
- Gerhard Woeginger (Eindhoven)

For further information, please contact us via the email address **ipco2007@orie.cornell.edu**.

Optimization 2007
July 22-25, 2007
University of Oporto, Oporto, Portugal
`http://www.fep.up.pt/opti2007`

Optimization 2007 is the sixth international conference on optimization organized in Portugal since 1991. We are proud to announce that six world-renowned scientists have accepted invitations to give plenary lectures. We hope to provide a friendly atmosphere and a lively social program.

The meeting will be held at the Faculty of Economics of the University of Porto, and it is supported by APDIO and SPM (the portuguese operations research and mathematical societies).

List of plenary speakers:

- Charles Audet
  École Polytechnique Montréal, Canada
  Direct Search Methods for Non-Smooth Optimization
- Egon Balas
  Carnegie Mellon University, USA
  Lift-and-Project and Its Impact on the State of the Art in Integer Programming

- Adam N. Letchford
  Lancaster University, UK
  The Max-Cut Problem: Applications and Algorithms
- Sven Leyffer
  Argonne National Laboratory, USA
  Recent Progress in Mixed Integer Nonlinear Programming
- Michael J. Todd
  Cornell University, USA
  Conic Optimization: Interior-Point Methods and Beyond
- Xin Yao
  University of Birmingham, UK
  Evolutionary Optimization and Constraint Handling

José Fernando Gonçalves (Conference Chair) **opti2007@fep.up.pt**.

ICCOPT II
Second Mathematical Programming Society International Conference on Continuous Optimization
August 12–16, 2007
McMaster University, Hamilton, Ontario, Canada
`http://iccopt-mopta.mcmaster.ca`

ICCOPT is held every three years and is a forum for researchers interested in all aspects of continuous optimization. ICCOPT-II will be held together with MOPTA-07. The conference will be preceded by graduate-level tutorials on nonlinear programming, modeling languages, and applications. More information on the invited speakers, stream topics, *etc.* will be available soon on the official webpage.

We are looking forward to welcoming you at ICCOPT-II / MOPTA-07 in Hamilton at McMaster University in August 2007.

Henry Wolkowicz (**hwolkowicz@uwaterloo.ca**), Támas Terlaky (**terlaky@mcmaster.ca**).

Chairman's Column

The triennial International Symposium on Mathematical Programming (ISMP) and SIAM Optimization meetings have become the major international conferences in the field of optimization. Having recently returned from ISMP 2006 in Rio de Janeiro, I would like to give a brief "report from Rio" for SIAG/OPT members who were unable to attend. The ISMP 2006 meeting had approximately 750 registrants with sessions held over 5 days. In addition to an opening session at the spectacular Teatro Municipal that featured a program of Brazilian music, and a "rodizio" conference dinner, the meeting had an outstanding slate of Plenary and Semi-Plenary sessions on a wide variety of topics including special sessions dedicated to George Dantzig and Leonid Khachiyan. The location of ISMP 2006, along with that of ISMP 2003 (Copenhagen) and the 2005 SIAM Optimization meeting (Stockholm) dramatically illustrates how truly international the field of optimization has become. The 2006 meeting was the first ISMP held in South America, and the large number of South American participants shows how vibrant the optimization research community is in countries such as Argentina, Brazil, Chile and Venezuela. Although the 2008 SIAM Optimization meeting in Boston will offer less-expensive travel for US participants, co-organizer Sven Leyffer and I will be hard-pressed to match the Brazilian dance demonstration at the ISMP 2006 conference dinner for spectator interest. We welcome your suggestions on any and all aspects of the 2008 meeting as we begin planning for it.

Kurt M. Anstreicher, SIAG/OPT Chair
Department of Management Sciences
University of Iowa
S210 PBB Iowa City, IA 52242,
USA
kurt-anstreicher@uiowa.edu
`http://www.biz.uiowa.edu/faculty/anstreicher`

Comments from the Editor

In this issue of SIAM/Optimization Views-and-News (Volume 17, Number 2) we publish a set of articles on Optimization in Medicine. Continuous and discrete optimization techniques are being applied to several important problems in Medicine with great success. This is due in part to the current dissemination of technology in Medicine, but is also related to the wide range of applicability of Optimization as a fundamental discipline in Applied Mathematics.

SIAM/Optimization Views-and-News is grateful to our guest editors, Eva K. Lee and Ariela Sofer, for having accepted to edited this special issue on Optimization in Medicine, and to the authors who have contributed with interesting and relevant papers. I would also like to thank Pedro Martins (IPC, Coimbra, Portugal) for helping me revising the manuscripts.

Luís N. Vicente, Editor
Department of Mathematics
University of Coimbra
3001-454 Coimbra
Portugal
lnv@mat.uc.pt
`http://www.mat.uc.pt/~lnv`