THE RADIATION THERAPY PLANNING PROBLEM

CHRISTOPH BÖRGERS*

1. Introduction. The purpose of this paper is to describe mathematical aspects of radiation therapy planning to readers with a background in applied mathematics.

The use of X-rays for cancer therapy began a few days after their discovery. Wilhelm Röntgen announced the discovery of X-rays on December 28, 1895, and Emil Grubbe used them for cancer therapy on January 12, 1896 [40]. X-rays are still the most common form of radiation used for cancer therapy, but beams of electrons, protons, neutrons, and other particles are used as well. The planning of the radiation treatment of a tumor begins with the creation of a three-dimensional image of the tumor and surrounding healthy tissue, using techniques such as computed tomography or MRI. The treatment planning discussed in this article occurs after the imaging is completed. It involves substantial use of computational algorithms.

Radiation therapy planning requires the study of radiation penetrating a background (a portion of a patient's body and the surrounding air, for instance). Both the radiation and the background are, of course, made up of particles. We shall distinguish between the two by referring to *radiation particles* and *background particles*. Background particles can be set in rapid motion as a result of interactions with radiation particles, thereby becoming radiation particles themselves. The transport of the radiation particles through the background is described by a system of coupled Boltzmann transport equations; see for instance Ref. [15], and also Sec. 2 of this article. A solution of this system is a vector of *phase space number densities*, that is, numbers of radiation particles per unit volume in *phase space*, i.e. position-direction-energy space. Different components of this vector correspond to different particle types. Even if the beams aimed at the tumor consist of one particle type only (for instance photons, as in X-rays), interactions between radiation particles and the background will set in motion other types of particles. Careful calculations therefore require consideration of several types of radiation particles in any case.

Interactions of radiation particles with each other are negligible in this context. The relevant transport equations are therefore linear. The speed of the radiation particles is the speed of light (for photons) or a significant fraction of the speed of light. As a result, a steady state is reached in a time that is extremely short in comparison with the times for which the beams are typically turned on during treatment, which are on the order of seconds or minutes. The relevant transport equations therefore contain no time

*Department of Mathematics, Tufts University, Medford, MA 02155, E-mail: borgers@math.tufts.edu

derivatives. Information obtained through imaging, such as the locations of soft tissue, bone, or air gaps, yields coefficients in these equations.

Radiation therapy is *fractionated*, that is, delivered in multiple sessions. Furthermore, during a single session, several beam configurations may be used. A radiation therapy plan specifies beam positions, directions, energies, etc., as well as when and how long the specified beams are to be turned on. This can be viewed as specifying a sequence of inflow boundary value problems for a system of steady linear Boltzmann equations.

The full solutions to these boundary value problems are never considered in radiation therapy planning. Of greatest interest is the total *dose*, that is, the amount of energy per unit background mass deposited, during the entire course of the treatment, as a result of excitation and ionization events. In the language of kinetic theory, dose is a *macroscopic* quantity, whereas the solution to a linear Boltzmann equation is a *mesoscopic* quantity. Dose depends on position; to emphasize this dependence, it is often called the *dose distribution.*

Strictly speaking, the dose distribution is not all that matters. Biological effects also depend on the type and energy of radiation used, the fractionation schedule, etc.; see Chapter 17 of Ref. [23] for a discussion of these factors. However, in practice it is usually assumed that for a given type of treatment (for instance, treatment with X-rays of a given energy range, and using a standard fractionation schedule), the effectiveness of a given treatment plan can be predicted from the dose distribution alone.

Computing dose distributions is a matter of computational physics, based on well-understood physical principles. In order to devise a good treatment plan, one must also be able to evaluate the *desirability* of a given dose distribution. This is most typically done by a physician based on experience and intuition, and is not a matter of rigorous science. However, one approach to evaluating the desirability of a given dose distribution is to first estimate, based on clinical data or even radiobiological models, the probabilities $p_1, ..., p_n$ of certain events, such as eradication of the tumor, damage to or destruction of healthy organs, pain relief as a result of tumor size reduction, etc. One can then use a function $\varphi = \varphi(p_1, ..., p_n)$ as the measure of desirability. Obviously φ depends on subjective preferences. Refs. [32] and [38] are basic articles on this sort of approach.

2. Dose calculation. As discussed in the introduction, dose calculation means the computation of macroscopic information related to the solution of an inflow boundary value problem for a system of steady linear Boltzmann equations; see Ref. [28] for a recent survey of this aspect of the problem, and an extensive literature list. To make this more concrete for readers not familiar with the linear Boltzmann equation, we shall outline the derivation of the equation for the special case of a single species of particles moving through a homogeneous, scattering, non-absorbing background; see Sec. 1.4 of Ref. [39].

Consider a particle moving through a background. Assume that the particle experiences collisions with the background at random times, and that the times between collisions are independent of each other, exponentially distributed, with an expected value depending on the pre-collision kinetic energy of the particle. This expected value is called the *mean free time.* Assume further that the collisions cause random direction and energy changes. As a result, the *phase space coordinates* of the particle, i.e. its position $\mathbf{x} \in \mathbb{R}^3$, direction $\omega \in S^2$, and energy $E > 0$, at time t are random. Denote their probability density by $f(\mathbf{x}, \omega, E, t)$. When a particle with pre-collision direction $\omega' \in S^2$ and pre-collision energy $E' > 0$ undergoes a collision, its post-collision direction $\omega \in S^2$ and energy $E > 0$ are random, with probability density

$$\frac{1}{2\pi}\, p(\omega \cdot \omega', E, E') \ . \tag{2.1}$$

This expression depends on the dot product $\omega \cdot \omega'$, but not on ω and ω' individually, reflecting isotropy of scattering. If we define $\mu = \omega{\cdot}\omega' \in [-1, 1]$ to be the cosine of the angle between the pre- and post-collision directions, the probability density of the pair (μ, E), for a given E', is $p(\mu, E, E')$, without the factor of $1/2\pi$. A particle with kinetic energy $E > 0$ has velocity $v(E)$ and mean free time $\overline{\tau}(E)$. The probability density $p(\mu, E, E')$ is close to zero unless μ is close to one. That is, the deflection experienced by a particle in a single collision is likely to be small; see for instance Ref. [45]. One expresses this by saying that the scattering is *strongly forward-peaked.*

With the notation introduced above,

$$f_t + v(\omega \cdot \nabla) f = Q_0 f \ , \tag{2.2}$$

with

$$(Q_0 f)(\omega, E) = \frac{1}{2\pi} \int_0^\infty \int_{\omega' \in S^2} p(\omega \cdot \omega', E, E') \frac{f(\omega', E')}{\overline{\tau}(E')} d\omega' dE' - \frac{f(\omega, E)}{\overline{\tau}(E)}. \tag{2.3}$$

In Eq. (2.3), we have omitted the dependence on $\mathbf{x}$ and t for notational simplicity. The left-hand side in Eq. (2.2) corresponds to the streaming of the particle between collisions. On the right-hand side of Eq. (2.3), the term with the minus sign corresponds to the particle being "lost" by entering a collision, and the integral corresponds to the particle "re-emerging" from a collision with altered direction and energy. Up to now, we have thought of a single particle, and of f as the probability density function of its phase space location. Alternatively, we can think of a very large number of particles, independent of each other, and of f as their phase space number density. This is how we shall think from now on.

It is customary to introduce the independent variable

$$\psi = vf \ , \tag{2.4}$$

called the flux, and the quantity

$$\sigma_s(E) = \frac{1}{v(E)\overline{\tau}(E)}\ , \tag{2.5}$$

called the scattering cross-section. Using this notation, and dropping the time derivative, Eq. (2.2) becomes

$$(\omega \cdot \nabla)\psi = Q\psi\ , \tag{2.6}$$

where

$$\begin{aligned} Q\psi(\omega, E) = \frac{1}{2\pi}\int_0^\infty \int_{\omega' \in S^2} & p(\omega \cdot \omega', E, E')\sigma_s(E')\psi(\omega', E')d\omega' dE' \\ & -\sigma_s(E)\psi(\omega, E)\ . \end{aligned} \tag{2.7}$$

Background inhomogeneity makes σ_s and p functions of $\mathbf{x}$.

Let $\Omega \subseteq \mathbb{R}^3$ be a bounded region with a smooth boundary $\partial\Omega$. Let $\mathbf{n} = \mathbf{n}(\mathbf{x})$, $\mathbf{x} \in \partial\Omega$, denote the exterior unit normal vector field on $\partial\Omega$. A well-posed boundary value problem for $\psi = \psi(\mathbf{x}, \omega, E)$, $(\mathbf{x}, \omega, E) \in \Omega \times S^2 \times \mathbb{R}_+$, is obtained by supplementing Eq. (2.6) with the *inflow boundary condition*

$$\psi(\mathbf{x}, \omega, E) = g(\mathbf{x}, \omega, E) \quad \text{for } \mathbf{x} \in \partial\Omega,\ \omega \in S^2,\ \omega \cdot \mathbf{n}(\mathbf{x}) < 0,\ E > 0\ . \tag{2.8}$$

For the mathematical theory of inflow boundary value problems for linear Boltzmann equations, see for instance Chapter 21 of Ref. [13]

To illustrate how dose distributions can be obtained from the solutions to boundary value problems for linear Boltzmann equations, let us compute an expression for the time rate at which the particles deposit energy in the background in our simplified setting. The expected amount of energy lost by a particle with pre-collision direction ω' and pre-collision energy E' in a collision is

$$\overline{\Delta E}(\mathbf{x}, \omega', E') = \frac{1}{2\pi}\int_0^\infty \int_{\omega \in S^2} (E' - E)\, p(\mathbf{x}, \omega \cdot \omega', E, E')\, d\omega\, dE\ . \tag{2.9}$$

The time rate of energy deposition is

$$d(\mathbf{x}) = \int_0^\infty \int_{\omega' \in S^2} \overline{\Delta E}(\mathbf{x}, \omega', E')\, \sigma_s(\mathbf{x}, E')\, \psi(\mathbf{x}, \omega', E')\, d\omega' dE'\ , \tag{2.10}$$

and the energy deposited during a time interval of duration T is

$$D(\mathbf{x}) = T\, d(\mathbf{x})\ . \tag{2.11}$$

As explained earlier, the true equations are a little more complicated, and in particular are coupled systems of linear Boltzmann equations.

In discussions of dose calculation in the Medical Physics literature, the underlying system of linear Boltzmann equations is not usually mentioned. With the codes used in current clinical practice, one typically obtains the dose directly, that is, without first computing the solution of the system of linear Boltzmann equations. There is a wide variety of different algorithms. However, they share the following basic ideas. The incoming radiation is thought of as composed of a finite number of pencil beams, that is, infinitesimally thin, mono-directional, mono-energetic beams. Mathematically, this means approximation of the boundary data by a finite sum of δ-functions. Approximations to the dose distributions due to pencil beams are obtained by laboratory experiments, numerical experiments using Monte Carlo simulation, mathematical analysis, or a combination of these approaches. The overall dose distribution is then obtained by summing such approximations. For discussions of dose calculation methods of this kind, see Refs. [24] and (for electron beams) [22]. There is an extensive literature on the mathematical analysis of pencil beams, starting with work due to Fermi [16]; see Ref. [22] for a survey and references. We studied this subject in Refs. [5]–[7].

In the past, Monte Carlo methods have been too slow for routine clinical use. However, the combination of gains in computer speed and development of faster Monte Carlo methods makes their future widespread clinical use increasingly likely; see for instance Refs. [1], [2], and [30] for Monte Carlo methods for particle transport calculations in general, and [35] for a Monte Carlo method specifically for radiation therapy planning.

Grid-based methods for the linear Boltzmann equation, using finite difference or finite element discretizations of spatial derivatives and, for example, discrete ordinates for the collision operator, are rarely mentioned in the Medical Physics literature. The *phase space evolution methods* (see Refs. [20] and [21]) come close to being such schemes. In general, the use of grid-based deterministic methods requires the development of efficient solvers for linear Boltzmann boundary value problems. This subject has been studied extensively in the Nuclear Engineering literature; see for instance Ref. [27] and references given there. However, most of this work does not apply to the case of strongly forward-peaked scattering. It appears that this is a gap that needs to be filled if grid-based deterministic methods are to become practical for dose calculations; see Refs. [25], [34], and [3] for methods for simplified (one- and two-dimensional) problems.

One might think that deterministic methods, such as finite difference or finite element methods, are not likely to compete well with Monte Carlo methods because of the large number of phase space dimensions (three space and three velocity dimensions). I discussed this argument in detail in Ref. [4], coming to the conclusion that it is not convincing. Therefore the question which of the two families of methods is preferable remains, at least in my view, unsettled.

We conclude this section by mentioning that the unit of dose commonly used in radiation therapy planning is the Gray, abbreviated Gy:

$$1\text{Gy} = 1\text{J}/\text{kg} \; . \tag{2.12}$$

For realistic values, see for instance Sec. 6 of Ref. [29]. One of the cases discussed there is a brain tumor, for which a dose of 90Gy was prescribed, with limits on the doses to brainstem and optic nerve of 20Gy and 10Gy, respectively.

3. Realizable dose distributions. We call the mapping from beam intensity distributions to dose distributions the *dose operator.* We call a dose distribution *realizable* if there is a realizable beam intensity distribution generating it. Which beam intensity distributions are realizable depends, of course, on the hardware used to deliver the radiation therapy. The most obvious constraint is that beam intensities must be non-negative. Typically there also is an upper bound on the number of beams that can be used.

Let $\mathcal{R}$ denote the set of realizable dose distributions. The question whether $\mathcal{R}$ is convex will be of interest to us in later sections. If the non-negativity of beam intensities is the only constraint on the treatment plan, then the set of permitted beam intensity distributions is convex, and therefore $\mathcal{R}$, being the image of a convex set under the (linear) dose operator, is convex as well. On the other hand, if there is a bound on the number of beams, but freedom in choosing beam positions and directions, then the set of permitted beam intensity distributions is non-convex, and so is $\mathcal{R}$ in general. We briefly refer to the problem of choosing beam positions and directions in the presence of a bound on the number of beams as the *beam selection problem.* So inclusion of the beam selection problem in the optimization problem makes $\mathcal{R}$ non-convex. The beam selection problem is discussed extensively in Ref. [29].

4. Biological response models. Models attempting to predict the probabilities of certain events, desirable or undesirable, for a given dose distribution, are called *biological response models.* For an introduction to this aspect of the problem, see for instance Sec. 1.1 of Ref. [42] and Refs. [37] and [41]. To illustrate the flavor of these models, we shall consider some simple examples. They are found in the references given above, although our notation is a little non-standard here.

We denote the region occupied by the tumor by Ω_t, the region occupied by healthy tissue by Ω_h, and the region of interest by $\Omega = \Omega_t \cup \Omega_h$. There may be ambiguity about the extent of a tumor; one can model that by not requiring that the intersection of Ω_t and Ω_h be empty.

We first discuss the *tumor control probability* (TCP). Assume that the tumor contains a very large number of small units called clonogens, and that the tumor is eradicated if and only if each clonogen is eradicated. Denote by ρ the number density of clonogens. Further assume that the deaths of clonogens are independent random events, and that for a given

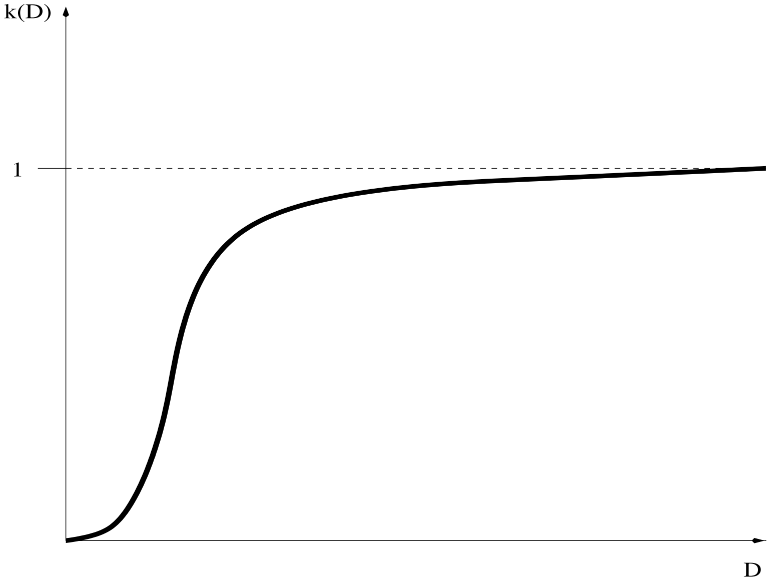

FIG. 1. *Probability of killing a single clonogen with dose D.*

clonogen, the probability of its death only depends on the dose D received by it. Denote this probability by $k(D)$. Suppose now that the tumor region Ω_t is divided into a large number of subregions $\Omega_{t,k}$ of volume V_k, $k = 1, ..., n$. Assume that these subregions are so small that the dose and the clonogen number density in $\Omega_{t,k}$ can be approximated by constants D_k and ρ_k, but so large that $\Omega_{t,k}$ contains many clonogens. Then

$$TCP \approx \prod_{k=1}^{n} k(D_k)^{\rho_k V_k} = \exp \sum_{k=1}^{n} \rho_k V_k \ln k(D_k) \ . \tag{4.1}$$

A continuous analog of (4.1) is

$$TCP = \exp \int_{\Omega_t} \rho(\mathbf{x}) \ln k(D(\mathbf{x}))\, d\mathbf{x} \ . \tag{4.2}$$

In the special case of constant D and ρ, this reduces to the obvious formula

$$TCP = k(D)^N \ , \tag{4.3}$$

where N denotes the total number of clonogens. So Eq. (4.2) gives the right way of modifying Eq. (4.3) for non-constant D and ρ. To complete the model of the TCP, one has to specify the function $k(D)$. It is always taken to be sigmoidal, as sketched in Fig. 1; compare for instance Fig. 1.18 on p. 37 of Ref. [42].

Lyman [31] proposed a simple formula for *normal tissue complication probabilities* (*NTCP*s). It is not a fundamental model based on radiobiology, but a data fitting scheme. Lyman's model applies to cases when a fraction v, $0 \leq v \leq 1$, of an organ at risk receives a constant dose D, and the rest of the organ receives no dose at all. Several ways of extending this model to the general case of a spatially varying dose have been proposed. The one due to Kutcher and Burman [26] can be shown, after a small amount of algebra, to be equivalent to

$$NTCP = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(\langle D \rangle_{L^p} - D_{50})/\sigma} \exp(-t^2/2)\, dt \ , \tag{4.4}$$

where $\langle D \rangle_{L^p}$ denotes the L^p-average of the dose over the organ at risk, that is:

$$\langle D \rangle_{L^p} = \frac{\|D\|_{L^p}}{V^{1/p}} \ , \tag{4.5}$$

where V is the volume of the organ at risk, and the parameters $p > 0$, $D_{50} > 0$, and $\sigma > 0$ are adjusted to fit experimental data. (The denominator of $V^{1/p}$ in Eq. (4.5) is needed to ensure that $\langle D \rangle_{L^p} = C$ if $D(\mathbf{x}) = C$ for all $\mathbf{x}$.) Eq. (4.4) predicts that irradiation at a dose with L^p-average D_{50} leads to a complication with probability 50%; this explains the notation. Table 1 of Ref. [9] suggests values of the parameters p, σ, and D_{50} for various different tissues and organs. The three parameters in Ref. [9] are called n, TD_{50}, and m. These parameters are related to ours as follows: $p = 1/n$, $D_{50} = TD_{50}$, and $\sigma = mTD_{50}$. For example, for the spinal chord, $p = 20$, $D_{50} = 66.5$Gy, and $\sigma = 11.6$Gy, and for the lungs, $p = 1.15$, $D_{50} = 24.5$Gy, and $\sigma = 4.4$Gy. The difference in the values of p reflects that for the spinal chord, even small regions of large dose must be avoided, whereas for the lungs, the average dose is essentially all that matters.

A different approach is described in Ref. [44]. Our presentation of it is close to that of Ref. [41]. Consider an organ at risk occupying a region Ω_o in space (or, more generally, any $\Omega_o \subseteq \Omega_h$). Assume that the organ at risk is made up of a large number of small units called *functional subunits*.

For some organs, such as the spinal chord, it may be appropriate to assume that significant damage to the organ occurs as soon as one of the subunits is destroyed. Organs of this kind are said to have a *serial* structure [42]. The probability of *no* normal tissue complication is then the probability that all subunits survive. This is analogous to the TCP model discussed earlier, where the probability of tumor control is the probability that all clonogens are killed. Following the same arguments that lead to Eq. (4.2), denoting by ρ the number density of subunits, and by $s(D)$ the probability that a single subunit survives irradiation at dose D, we are lead to the formula

$$NTCP = 1 - \exp \int_{\Omega_o} \rho(\mathbf{x}) \, \ln s(D(\mathbf{x}))\, d\mathbf{x} \ . \tag{4.6}$$

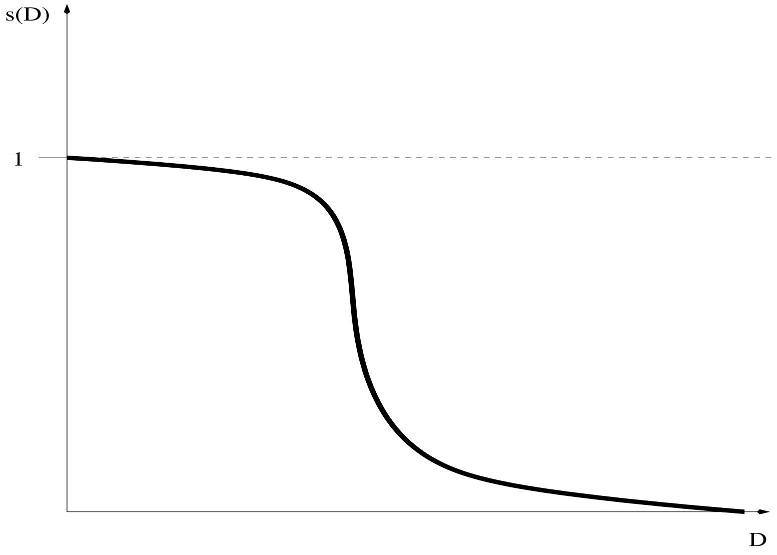

FIG. 2. *Survival probability of a single subunit receiving dose D.*

It is natural to assume that s is sigmoidal, as shown in Fig. 2.

For other organs, such as the lungs, it may be appropriate to assume that significant damage to the organ occurs only when a certain percentage of the subunits is destroyed. Organs of this kind are said to have a *parallel* structure [42]. The $NTCP$ might, more generally, be a decreasing function of the number of surviving subunits, that is:

$$NTCP = G\left(\int_{\Omega_o} \rho(\mathbf{x})\, s(D(\mathbf{x}))\, d\mathbf{x}\right) \ , \tag{4.7}$$

where G is a decreasing function.

5. Dose-volume constraints. Biological response models offer one way of formulating requirements on the dose distribution. Dose-volume constraints are an alternative approach. For a given dose distribution, and a given region in space, for instance the region occupied by the tumor or by a healthy organ, define $v(d)$ to be the volume fraction that receives a dose $\geq d$. It is clear that v is a decreasing function of d with $v(0) = 1$ and $v(d) = 0$ for sufficiently large d. In the Medical Physics literature, v is called a (differential) *dose-volume histogram*; see for instance Sec. 1.1.9 of [42]. An *upper dose-volume constraint* is an inequality of the form

$$v(d) \leq v_{max}(d) \quad \text{for all } d \ . \tag{5.1}$$

This is appropriate for an organ at risk. A *lower dose-volume constraint* is an inequality of the form

$$v(d) \geq v_{min}(d) \quad \text{for all } d \ . \tag{5.2}$$

This is appropriate for a tumor. Notice that a constraint of the form $v(d_0) \leq v_0$ for one particular d_0 is a special upper dose-volume constraint, with $v_{max}(d) = v_0$ for $d \geq d_0$, and $v_{max}(d) = 1$ for $d < d_0$. Similarly, $v(d_0) \geq v_0$ is a special lower dose-volume constraint.

6. Minimizing the distance from an ideal dose distribution. Ideally, one would, of course, like to achieve the dose distribution

$$\hat{D}(\mathbf{x}) = \begin{cases} D_0 & \text{in } \Omega_{\mathrm{t}} \ , \\ 0 & \text{in } \Omega_{\mathrm{h}} \ , \end{cases} \tag{6.1}$$

where D_0 is as large as needed to kill the tumor with certainty, but not very much larger; see for instance Ref. [17]. It is therefore natural to use, as a measure of desirability of a dose distribution D, the quantity

$$\varphi(D) = -||D - \hat{D}|| \tag{6.2}$$

for some function norm $|| \cdot ||$; see for instance Ref. [19]. The minus sign ensures that larger φ means greater desirability. Of course $\hat{D}$ is not realizable in general, since radiation must pass through healthy tissue to reach a tumor that does not lie at the surface of the patient's body, so in general the maximum of φ is negative.

Assume now that $|| \cdot ||$ is an L^p-norm:

$$||D - \hat{D}|| = ||D - \hat{D}||_{L^p(\Omega)} = \left[\int_\Omega |D - \hat{D}|^p \, d\mathbf{x} \right]^{1/p} \tag{6.3}$$

for some $p \geq 1$. Denote the region occupied by the tumor by Ω_t, and the region occupied by the healthy tissue by Ω_h. Then

$$||D - \hat{D}||^p_{L^p(\Omega)} = ||D - D_0||^p_{L^p(\Omega_t)} + ||D||^p_{L^p(\Omega_h)} \ . \tag{6.4}$$

Therefore minimizing $||D - \hat{D}||_{L^p(\Omega)}$ is equivalent to assuming that $1 - TCP$ is proportional to $||D - D_0||^p_{L^p(\Omega_t)}$, $NTCP$ is proportional to $||D||^p_{L^p(\Omega_h)}$, and maximizing an expression of the form $\alpha TCP - \beta NTCP$, where α and β are positive weights. For $p > 1$, D approximately constant in Ω_t, and $||D - D_0||^p_{L^p(\Omega_t)}$ and small $||D||^p_{L^p(\Omega_h)}$, this can, depending on the parameter values, closely resemble the use of the biological response models given by Eqs. (4.2), (4.4), and (4.5); for the TCP, this is illustrated in Fig. 3.

It seems to me that the approach of Ref. [11] (see also Ref. [10]) can essentially be viewed as an improvement on minimizing the distance from an ideal dose distribution. To see the similarity, compare for instance

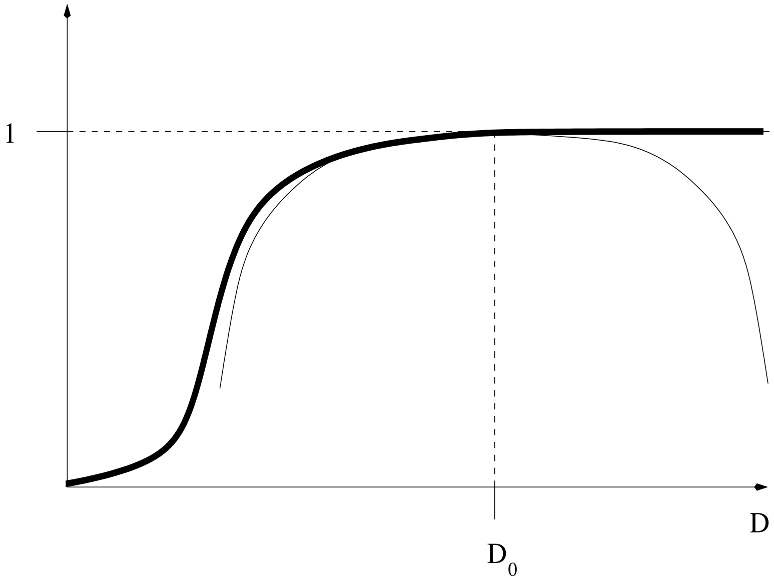

FIG. 3. *TCP (bold) and* $1 - c|D - D_0|^p$.

Eq. (6.2) with the displayed equation following Eq. (40) in [11]. Notice, however, that exceeding $\hat{D}$ in the tumor is penalized when minimizing the distance from $\hat{D}$, whereas exceeding the lower dose bound in the tumor is not penalized in the approach of [11], unless there is an upper dose bound in the tumor, and that bound is exceeded as well. Similarly, there is no penalty for a positive dose in healthy tissue in the approach of Ref. [11], as long as the upper dose bound in the healthy tissue is not exceeded.

7. Convexity of the set of "acceptably safe" dose distributions. Using biological response models as in Sec. 4, we might call a dose distribution $D \in \mathcal{R}$ *acceptably safe* if it satisfies inequalities of the form

$$NTCP_i \leq \epsilon_i \ , \tag{7.1}$$

where the index i labels possible normal tissue complications, and the $\epsilon_i \in (0,1)$ are prescribed bounds. Alternatively, based on Sec. 5, $D \in \mathcal{R}$ could be called acceptably safe if it satisfies dose-volume constraints

$$v_i(d) \leq v_{max,i}(d) \ , \tag{7.2}$$

where the index i labels organs at risk, and the $v_{max,i}$ are prescribed decreasing functions with values between 0 and 1. Based on Sec. 6, we might call $D \in \mathcal{R}$ acceptably safe simply if it satisfies a constraint of the form

$$||D|_{\Omega_h}|| \leq \epsilon \ , \tag{7.3}$$

where $\|\cdot\|$ denotes a function norm, $D|_{\Omega_h}$ denotes the restriction of D to Ω_h, and $\epsilon > 0$ is a prescribed bound.

Let us denote by $\mathcal{A}$ the set of acceptably safe dose distributions. Is $\mathcal{A}$ convex? That is, is a convex combination of acceptably safe dose distributions acceptably safe? There are two reasons for asking this question. First, the answer will be useful in Sec. 8. Second, the answer may reveal a flaw in the formulation used. It is *desirable* that the definition of *acceptably safe* permit $\mathcal{A}$ to be non-convex in some cases, even when $\mathcal{R}$ is convex. Namely, it can be acceptable to sacrifice a small volume of healthy tissue, but not a large one; lung tissue is an example. In such a case, two treatment plans giving large doses to small volumes of healthy tissue may each be considered acceptably safe, but their average, giving a moderate dose to a larger volume, may not be. This point is made, explicitly or implicitly, in several places in the literature on radiation therapy planning; see for instance the discussion of the difference between the prostate and lung cases in Ref. [33], or p. 1296 of Ref. [41].

For the remainder of the section, we assume that $\mathcal{R}$ is convex, in particular that the beam selection problem is not included in the optimization problem. If the definition of $\mathcal{A}$ is then based on $NTCP$ models as defined by Eqs. (4.4) and (4.5), $\mathcal{A}$ is assured to be convex, regardless of the parameter choices. The same holds if its definition is based on (7.3), regardless of the choice of norm, or on an improved inequality along the lines of Ref. [11] (compare the discussion at the end of Sec. 6). On the other hand, if the definition is based on (7.2), then $\mathcal{A}$ is assured *not* to be convex, except in the trivial case when all $v_{max,i}$ are constant functions. If an $NTCP$ model of the form (4.6) underlies the definition of $\mathcal{A}$, convexity of $\mathcal{A}$ is assured provided that $\ln s(D)$ is a concave (that is, concave-down) function of D. Whether or not this is the case cannot be deduced from the general qualitative shape in Fig. 2. Finally, if an $NTCP$ model of the form (4.7) underlies the definition of $\mathcal{A}$, then convexity of $\mathcal{A}$ is certainly not assured, since the function s is not everywhere concave.

8. Uniqueness of solutions to the optimization problem. We conclude with a discussion of conditions that imply uniqueness of the solution to the radiation therapy optimization problem. From a practical point of view, this is important because the choice of optimization algorithm depends on it. For instance, algorithms such as simulated annealing have been proposed for radiation therapy optimization because of the possibility of multiple local optima [43]. As suggested by Niemierko in Ref. [36] and shown by Deasy in Ref. [14], insight into the issue of uniqueness of solutions can be gained from elementary convexity considerations. We shall present a variation on Deasy's argument, and point out an issue arising in this context that seems important and not yet well-understood [8].

The issue of uniqueness of the optimal treatment plan can be divided into two parts as follows. The first question is when, and in which sense, the optimal realizable dose distribution is unique. The second question is to which extent, for a given optimal realizable dose distribution, the treatment plan generating it is unique. This amounts to studying the nullspace of the dose operator.

To discuss the first question, let us assume that our goal is to determine a dose distribution $D \in \mathcal{A}$ (see Sec. 7) with maximal TCP. This corresponds to the choice $\varphi = TCP$ if $D \in \mathcal{A}$, and $\varphi = 0$ otherwise. Other choices of φ could be discussed similarly. Of course, maximizing TCP means the same as maximizing $q(TCP)$ if q is a strictly increasing function. If $\mathcal{A}$ is convex, then uniqueness of the optimal dose distribution in $\mathcal{A}$ depends on concavity properties of $q(TCP)$ as a functional of the dose distribution D. Strict concavity rules out multiple local maxima. Non-strict concavity rules out multiple local maxima with different TCP values.

Two sources of non-convexity of $\mathcal{A}$ have already been discussed. First, $\mathcal{R}$ is non-convex if beam selection is included in the optimization problem, as discussed in Sec. 3. But even if $\mathcal{R}$ is convex, $\mathcal{A}$ can be non-convex, and ought to be non-convex at least in some cases because of dose-volume constraints; see Sec. 7.

Let us assume now that $\mathcal{A}$ is convex for our problem. Using Eq. (4.2),

$$\ln TCP = \int_{\Omega_t} \rho(\mathbf{x}) \, \ln k(D(\mathbf{x})) \, d\mathbf{x} \ . \tag{8.1}$$

If k were a concave function, then (8.1) would be a concave functional of D. Although k is not concave everywhere (compare Fig. 1), it is concave where k is close to 1. This implies uniqueness of the locally optimal TCP value at least among dose distributions for which the minimum tumor dose is not too low.[1] Things are simpler if we maximize the minimum tumor dose instead of the TCP over $\mathcal{A}$:

$$\min_{\mathbf{x} \in \Omega_t} D(\mathbf{x}) \tag{8.2}$$

is a concave functional of D.

We are currently studying the second question [8]. Making greatly simplifying assumptions, including the absence of scattering, the dose operator can be modeled as the dual exponential X-ray transform; see Refs. [12] and [29]. The question raised then reduces to studying the nullspace of

[1] Deasy [14] proposed using the notion of *quasi-concavity* instead of concavity. A function $g = g(\mathbf{z})$, $\mathbf{z} \in \mathbb{R}^l$, is called quasi-concave if $g(\theta\mathbf{z}_1 + (1-\theta)\mathbf{z}_2) \geq \min(g(\mathbf{z}_1), g(\mathbf{z}_2))$ whenever $\mathbf{z}_1 \neq \mathbf{z}_2$ and $0 < \theta < 1$, and strictly quasi-concave if strict inequality holds. Strict quasi-concavity rules out multiple local maxima. Deasy argued that the TCP, as a function of beam weights, should be strictly quasi-concave. His reasoning, however, seems based on the incorrect assumption that strict monotonicity in each coordinate direction implies strict quasi-concavity ([14], p. 1159). A counterexample is $g(u, w) = (1 + u^2)(1 + w^2)$, $u \geq 0$ and $w \geq 0$. This function is strictly increasing in u and w, but not quasi-concave since $g(1/2, 1/2) = 25/16 < \min(g(1,0), g(0,1)) = \min(2,2) = 2$.

this transform. However, the question should also be posed with discrete sets of permitted beam positions and directions, and with more realistic dose operators. Even when the nullspace of the dose operator is trivial, one may ask whether the dose operator has singular values that are *nearly* zero. If the answer is yes, then there may be beam intensity distributions that are significantly different from the optimal one(s), but generate nearly optimal dose distributions. Among these beam intensity distributions, one could then try to find a particularly simple one.

9. Acknowledgments. While writing this article, I had helpful discussions with Martin Altschuler, Yair Censor, Todd Quinto, and Jim Satterthwaite. My work is supported in part by NSF grant DMS-9626696.

REFERENCES

[1] A. F. Bielajew, *Photon Monte Carlo Simulation*, Lecture Notes, National Research Council of Canada, Report PIRS-0393, available on the internet at http://ehssun.lbl.gov/egs/epub/course.html, 1993.

[2] A. F. Bielajew and D. W. O. Rogers, *Electron Monte Carlo Simulation*, Lecture Notes, National Research Council of Canada, Report PIRS-0394, available on the internet at http://ehssun.lbl.gov/egs/epub/course.html, 1993.

[3] C. Börgers, *A fast iterative method for computing particle beams penetrating matter*, J. Comp. Phys., **133**, 323–339, 1997.

[4] C. Börgers, *Complexity of Monte Carlo and deterministic dose-calculation methods*, Phys. Med. Biol., **43**, 517-528, 1998.

[5] C. Börgers and E. W. Larsen, *The transversely integrated scalar flux of a narrowly focused particle beam*, SIAM J. Appl. Math, **50**, No. 1, 1–22, 1995.

[6] C. Börgers and E. W. Larsen, *Asymptotic derivation of the Fermi pencil beam approximation*, Nucl. Sci. Eng., **123**, No. 3, 343–357, 1996.

[7] C. Börgers and E. W. Larsen, *On the accuracy of the Fokker-Planck and Fermi pencil beam equations for charged particle transport*, Med. Phys., **23**, 1749–1759, 1996.

[8] C. Börgers and E. T. Quinto, *Nullspace and conditioning of the mapping from beam weights to dose distributions in radiation therapy planning*, in preparation.

[9] C. Burman, G. J. Kutcher, B. Emami, and M. Goitein, *Fitting of normal tissue tolerance data to analytic functions*, Int. J. Rad. Onc. Biol. Phys., **21**, 123–135, 1991.

[10] Y. Censor, *Mathematical aspects of radiation therapy treatment planning: Continuous inversion versus full discretization and optimization versus feasibility*, in this volume.

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

[12] A. M. Cormack and E. T. Quinto, *The mathematics and physics of radiation dose planning using X-rays*, Contemporary Mathematics, **113**, 41–55, 1990.

[13] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, vol. 6, Springer-Verlag, 1993.

[14] J. O. Deasy, *Multiple local minima in radiotherapy optimization problems with dose-volume constraints*, Med. Phys., **24**, No. 7, 1157–1161, 1997.

[15] J. J. Duderstadt and W. R. Martin, *Transport Theory*, John Wiley & Sons, 1979.

[16] E. FERMI, result reported in B. Rossi and K. Greisen, *Cosmic ray theory*, Rev. Mod. Phys., **13**, 240, 1941.

[17] M. GOITEIN, *Causes and consequences of inhomogeneous dose distributions in radiation therapy*, Int. J. Rad. Onc. Biol. Phys., **12**, 701–704, 1986.

[18] M. GOITEIN AND A. NIEMIERKO, *Biologically based models for scoring treatment plans*, presentation to the Joint U.S./Scandinavian Symposium on Future Directions of Computer-Aided Radiotherapy, San Antonio, 1988.

[19] T. HOLMES AND T. R. MACKIE, *A comparison of three inverse treatment planning algorithms*, Phys. Med. Biol., **39**, 91–106, 1994.

[20] J. J. JANSSEN, D. E. J. RIEDEMAN, M. MORAWSKA-KACZYŃSKA, P. R. M. STORCHI, AND H. HUIZENGA, *Numerical calculation of energy deposition by high-energy electron beams: III. Three-dimensional heterogeneous media*, Phys. Med. Biol., **39**, 1351–1366, 1994.

[21] J. J. JANSSEN, E. W. KOREVAAR, R. M. STORCHI, AND H. HUIZENGA, *Numerical calculation of energy deposition by high-energy electron beams: III-B. Improvements to the 6D phase space evolution model*, Phys. Med. Biol., **42**, 1441–1449, 1997.

[22] D. JETTE, *Electron beam dose calculations*, in Radiation Therapy Physics, A. R. Smith (ed.), 95–121, Springer-Verlag, Berlin, 1995.

[23] H. E. JOHNS AND J. R. CUNNINGHAM, *The Physics of Radiology*, fourth edition, Charles C. Thomas, 1983.

[24] F. M. KHAN, *The Physics of Radiation Therapy*, second edition, Williams & Wilkins, 1994.

[25] K. M. KHATTAB AND E. W. LARSEN, *Synthetic acceleration methods for linear transport problems with highly anisotropic scattering*, Nucl. Sci. Eng., **107**, 217–227, 1991.

[26] G. J. KUTCHER AND C. BURMAN, *Calculation of complication probability factors for non-uniform normal tissue irradiation: the effective volume method*, Int. J. Rad. Onc. Biol. Phys., **16**, 1623–1630, 1989.

[27] E. W. LARSEN, *Diffusion-synthetic acceleration methods for discrete-ordinate problems*, Transport Theory Stat. Phys., **13**, 107–126, 1984.

[28] E. W. LARSEN, *Tutorial: The nature of transport calculations used in radiation oncology*, Transport Theory Stat. Phys., **26**, No. 7, 739, 1997.

[29] R. Y. LEVINE, E. A. GREGERSON, AND M. M. URIE, *The application of the X-ray transform to 3D conformal radiotherapy*, in this volume.

[30] I. LUX AND L. KOBLINGER, *Monte Carlo Particle Transport Methods: Neutron and Photon Calculations*, CRC Press, 1991.

[31] J. T. LYMAN, *Complication probability as assessed from dose-volume histograms*, Rad. Res., **104**, S-13–S-19, 1985.

[32] R. MOHAN, G. S. MAGERAS, B. BALDWIN, L. J. BREWSTER, G. J. KUTCHER, S. LEIBEL, C. M. BURMAN, C. C. LING, AND Z. FUKS, *Clinically relevant optimization of 3-D conformal treatments*, Med. Phys., **19**, 933–944, 1992.

[33] R. MOHAN, X. WANG, A. JACKSON, T. BORTFELD, A. L. BOYER, G. J. KUTCHER, S. A. LEIBEL, Z. FUKS, AND C. C. LING, *The potential and limitations of the inverse radiotherapy technique*, Radiother. Oncol., **32**, 232–248, 1994.

[34] J. E. MOREL AND T. A. MANTEUFFEL, *An angular multigrid acceleration technique for S_n equations with highly forward-peaked scattering*, Nucl. Sci. Eng., **107**, 330–342, 1991.

[35] H. NEUENSCHWANDER, T. R. MACKIE, AND P. J. RECKWERDT, *MMC – A high-performance Monte Carlo code for electron beam treatment planning*, Phys. Med. Biol., **40**, 543, 1995.

[36] A. NIEMIERKO, *Optimization of intensity modulated beams: Local or global optimum?*, Med. Phys., **23**, 1072, 1996.

[37] A. NIEMIERKO AND M. GOITEIN, *Calculation of normal tissue complication probability and dose-volume histogram reduction schemes for tissues with a critical element structure*, Radiother. Oncol., **20**, 166–176, 1991.

[38] A. Niemierko, M. Urie, and M. Goitein, *Optimization of 3D radiation therapy with both physical and biological end points and constraints*, Int. J. Rad. Onc. Biol. Phys., **23**, 99–108, 1992.

[39] G. C. Pomraning, *Linear Kinetic Theory and Particle Transport in Stochastic Mixtures*, Series on Advances in Mathematics for Applied Sciences Vol. 7, World Scientific, Singapore, 1991.

[40] *Radiology Centennial, Inc.*, A Century of Radiology, http://www.xray.hmc.psu.edu/rci/centennial.html, 1993.

[41] C. Raphael, *Mathematical modelling of objectives in radiation therapy treatment planning*, Phys. Med. Biol., **37**, No. 6, 1293–1311, 1992.

[42] S. Webb, *The Physics of Three-Dimensional Radiation Therapy*, IOP Publishing, Bristol and Philadelphia, 1993.

[43] S. Webb, *Optimization by simulated annealing of three-dimensional conformal treatment planning for radiation fields defined by multi-leaf collimators: II. Inclusion of two-dimensional modulation of X-ray intensities*, Phys. Med. Biol., **37**, 1689–1704, 1992.

[44] H. R. Withers, J. M. G. Taylor, and B. Maciejewski, *Treatment volume and tissue tolerance*, Int. J. Rad. Onc. Biol. Phys., **14**, 751–759, 1988.

[45] C. D. Zerby and F. L. Keller, *Electron transport theory, calculations, and experiments*, Nucl. Sci. Eng., **27**, 190–218, 1967.