Dissertation
submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by

MSc Physics:	**Daniel Pflugfelder**
born in:	**Stuttgart, Germany**

Oral examination: 6th February 2008

Risk-adapted Optimization in Intensity Modulated Proton Therapy (IMPT)

Referees: **Prof. Dr. Uwe Oelfke**
Prof. Dr. Wolfgang Schlegel

Zusammenfassung

Risikoadaptierte Optimierung in der intensitätsmodulierten Protonentherapie (IMPT)

Die ausgeprägten Dosisgradienten eines Protonenstahls können in der Protonentherapie zu Bestrahlungsplänen führen, die auf Unsicherheiten in der Bestrahlungsplanung und -applizierung sehr anfällig sind. Allerdings bietet die IMPT viele Lösungen des inversen Problems an, die vergleichbare Dosisverteilungen aufweisen. Diese Arbeit beschäftigt sich mit Möglichkeiten, diese Entartung der Lösungen zur Generierung robuster Bestrahlungspläne auszunutzen. Eine Überprüfung des Optimierungsalgorithmus der verwendeten IMPT Software KonRad ergab, daß der Standardalgorithmus den optimalen Bestrahlungsplan nicht in angemessener Zeit ermitteln kann. Deshalb wurden zusätzlich mehrere Optimierungsalgorithmen in KonRad implementiert und getestet. Die besten Ergebnisse erzielte der L-BFGS Algorithmus. Zur Beurteilung der Empfindlichkeit der Dosisverteilung einzelner Beamlets im Hinblick auf Unsicherheiten, wurde das Konzept der Heterogenitätszahl H_i entwickelt. Es wurde gezeigt, daß H_i sowohl mit dem Dosisberechnungsfehler, der durch den üblicherweise verwendeten Pencilbeam Algorithmus entsteht, als auch mit der Empfindlichkeit der einzelnen Beamlets im Bezug auf Fehllagerungen korreliert. Schließlich wurde die "worst case Optimierung" entwickelt um Unsicherheiten in die inverse Bestrahlungsplannung mit einzubeziehen. Diese Technik wurde auf Reichweitenunsicherheiten, Fehllagerungen des Patienten sowie deren Kombination angewandt. Die Bestrahlungspläne, die mit dieser neuen Methode erzeugt wurden, weisen im Vergleich zu konventioneller IMPT und sogar zu konventionellen Ein-Feld Bestrahlungsplänen eine deutlich größere Robustheit gegen die jeweiligen Unsicherheiten auf.

Abstract

Risk-adapted Optimization in Intensity Modulated Proton Therapy (IMPT)

Due to the pronounced dose gradients generated by proton beams, proton treatment plans can be very sensitive to treatment uncertainties. However in IMPT many different solutions of the inverse problem exist which result in dose distributions of comparable quality. This thesis investigates methods to exploit this degeneracy of solutions to generate treatment plans which are robust to uncertainties. An investigation of the optimization algorithm in the used IMPT software KonRad revealed that the standard optimization algorithm is not capable to find the optimal treatment plan in a reasonable time. Thus several additional optimization algorithms were implemented and tested in KonRad. The best results were achieved using the L-BFGS algorithm. To rate the sensitivity to uncertainties of individual beamlet dose distributions the heterogeneity number H_i was developed. It was shown that H_i correlates with the dose calculation error introduced by the commonly employed pencil beam algorithm as well as with the sensitivity to setup errors of individual beamlets. Finally, the "worst case optimization" was developed to account for uncertainties during the inverse treatment planning. This technique was applied to account for range uncertainties,

setup errors and a combination of both uncertainties. The treatment plans generated with this new method are much more robust to the respective uncertainties as conventional IMPT and even as conventional single-field proton plans.

Contents

Chapter 1

Introduction

Cancer is among the leading causes for death in Germany. In 2002, more than 400.000 newly diagnosed cases and more than 200.000 deaths due to cancer were reported [1]. The three main therapies against cancer are surgery, chemotherapy and radiation therapy. Often a combined therapy is employed. The *Deutsche Gesellschaft für Radioonkologie (DEGRO)* states that almost 60% of all cancer patients receive radiation at least as a part of their therapy [2]. This results in approximately 240.000 radiation treatments in Germany per year. The vast majority of these treatments employ photon radiation. However accelerated charged particles such as proton radiation offer the possibility to generate dose distributions which are not achievable with photons. Tumors which are so close to critical structures that they cannot be irradiated with photons might be cured by proton therapy. Nevertheless proton treatment plans cannot only deliver dose distributions with very steep dose gradients. Precisely because of these sharp dose gradients proton treatment plans can be very sensitive to uncertainties in the treatment planning and the dose delivery process. There are multiple sources for such uncertainties. The treatment planning CT which serves as the patient model for the treatment planning process is usually taken a few days before the beginning of the actual treatment course. Furthermore due to radiobiological effects radiation therapy is usually delivered in multiple fractions. At each day of the treatment the patient has to be positioned relative to the treatment beam. A positioning error of a few *mm* compared to the treatment planning CT can already lead to a deteriorated dose distribution [3]. Furthermore, a different patient anatomy such as a different filling of the bladder and rectum or weight -gain or -loss of the patient can lead to a different range of the proton beam and thus degenerate the dose distribution [4]. Besides these uncertainties in the treatment delivery process the planning CT already introduces uncertainties into the treatment planning process. The CT provides electron densities quantified in so-called

Hounsfield units (HU) for each volume element in space. To calculate the proton range and thus the dose distribution in the medium the relative stopping power of each volume element is needed. However there is no one-to-one transition between relative stopping power values and Hounsfield units. Two different materials with different relative stopping power values can result in the same Hounsfield unit [5]. Furthermore CT artefacts can have a large impact on the dose calculation. Metal implant lead to pronounced CT artefacts. Such implants are often present in patients which received surgery prior to radiotherapy. Although large efforts are made to minimize these uncertainties there still remain unavoidable uncertainties which can have a large impact on the delivered dose distributions. The aim of this thesis is to investigate methods for a risk-adapted proton treatment planning. For photon intensity modulated radiation therapy (IMRT) it has been shown that there are multiple possibilities to deliver dose distributions of similar quality, leading to a degenerated space of solutions for IMRT (see e.g. Alber et al. [6]). The intensity of photon radiation can only be modulated across the two-dimensional lateral beam profile. The proton beam can additionally be varied in the beam energy, leading to a different range of the beam. Thus the intensity of the proton radiation can be modulated in three dimensions. Due to this additional flexibility in the dose delivery process for proton therapy the degeneracy of solutions for IMPT is even expected to increase beyond the degeneracy for IMRT. In risk-adapted proton treatment planning, this degeneracy of creating treatment plans is utilized to search for high quality treatment plans that show a reduced sensitivity to treatment uncertainties.

The thesis is subdivided into six chapters. Chapter 2 will give a short introduction into proton therapy. The research done for this thesis will be presented in the following chapters. An investigation of the optimization algorithms used in the inverse planning process is shown in chapter 3. A method to quantify the risks resulting from lateral tissue heterogeneities of an individual beamlet is presented in chapter 4. In chapter 5, a method is developed to account for uncertainties in the inverse treatment planning process. Finally a summary and an outlook for further research is given in chapter 6. Parts of this thesis have been published or submitted for publication. References [7], [8] and [3] each cover parts of chapters 3, 4 and 5, respectively. Parts of the thesis have also been presented on international conferences. Abstracts to these talks can be found in references [9, 10, 11, 12, 13, 14].

Chapter 2

Proton Therapy

2.1 Introduction

The absorption mechanisms of charged particle radiation in matter are different compared to those of photon radiation. Already in 1946 R. Wilson [15] recognized the physical advantages of accelerated charged particles for radiation therapy. Photons are absorbed in one (photoelectric effect, pair production) or a few (Compton effect) interactions. This results in an exponential decay of the number of absorbed photons as the photon beam transverses a medium. Assuming that the photon energy is deposed locally at the site of the interaction this results in an exponential decay of the deposed dose in depth. (This assumption is not exactly fulfilled, leading to a buildup effect as it can be see in figure 2.1.) Compared to this charged particles release their energy in multiple interactions, predominately Coulomb interactions with the atomic electrons of the target material. The energy loss of charged particles as they traverse matter can be described by the Bethe-Bloch equation (see e.g.[16]). For therapeutical energies the energy loss increases in this equation as the particles slow down. This results in an increase of the absorbed energy in depth up to a point where all particles are absorbed, leading to a distinct peak in the depth-dose distribution. This peak was first measured by Sir William Henry Bragg who published his work "On the Ionization Curves of Radium" together with Richard Kleeman in 1904 [17]. To his honor this peak is called "Bragg peak". Often this behavior of charged particles is also referred to as inverse depth-dose profile. An analytical derivation of this curve can be found in [18].The penetration depth of the charged particles depends on their initial kinetic energy thus the range can be adjusted. A comparison between the absorbed dose in water of a photon beam and proton beams of different initial energies is shown in figure 2.1. The physical advantage of charged particles for radiation therapy compared to photons lies in

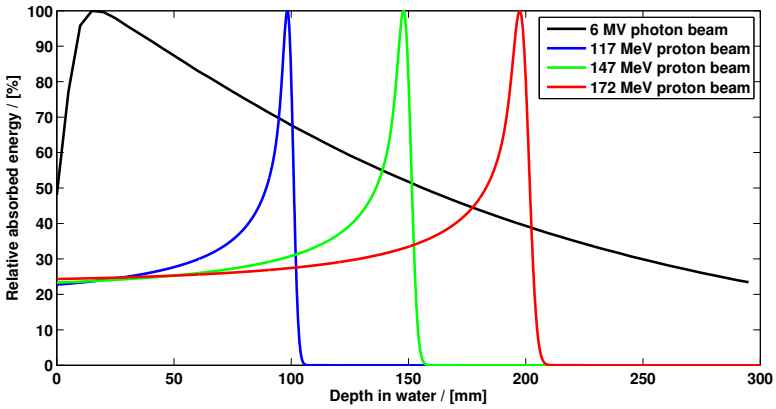

Figure 2.1: *The absorbed dose in water of broad proton beams of different energies and a 6 MV photon beam. The curves are normalized to the maximal value. The photon beam was measured, the proton beams are calculated using the analytical form of the Bragg peak as described in [18].*

the reduced dose in the entrance region as well as in the vanishing dose behind the Bragg peak. All heavy charged particles beams show this characteristic Bragg peak. Several particles were utilized for radiation therapy, besides protons, helium ions, carbon ions and pions were employed. However protons are the major modality for charged particles radiation therapy. Of all patients treated with charged particles up to 2005, 88% received a proton treatment and 9% received a carbon ion therapy [19]. This thesis concentrates on proton therapy, although the methods derived here can also be applied to other charged particle beams.

In principle all tumors treated with photons can also be treated with protons. The main argument against proton therapy is the costs associated with this modality. Proton therapy is estimated to be 1.7 to 3 times more expensive than the most advanced photon therapy (IMRT) [20].

2.2 Delivering techniques for proton therapy

Radiation therapy aims to deliver a prescribed dose homogeneously to the tumor while minimizing the dose to the surrounding healthy tissue. To realize this, several delivering techniques were utilized in proton therapy, starting from passive techniques to active techniques and finally to intensity modulated proton therapy (IMPT). This order represents the chronological order as well as the order of technical complexity as well as the order for the improvement of the sparing of healthy tissue. The three techniques will be explained in the following sections.

2.2.1 Passive techniques

A proton beam is generated by a particle accelerator, resulting in an almost monoenergetic, narrow beam of protons. To make such a beam useful for therapy it has to be broadened laterally as well as in depth to enclose the whole tumor with a homogeneous dose. The lateral broadening of the beam can be achieved by scattering foils as a pure passive technique or using wobbling magnets. To realize a homogeneous dose in depth, several initial proton energies are needed as shown in figure 2.2. The resulting beam shown there is called a spread-out Bragg peak (SOBP) [16]. For passive techniques, this SOBP can be generated

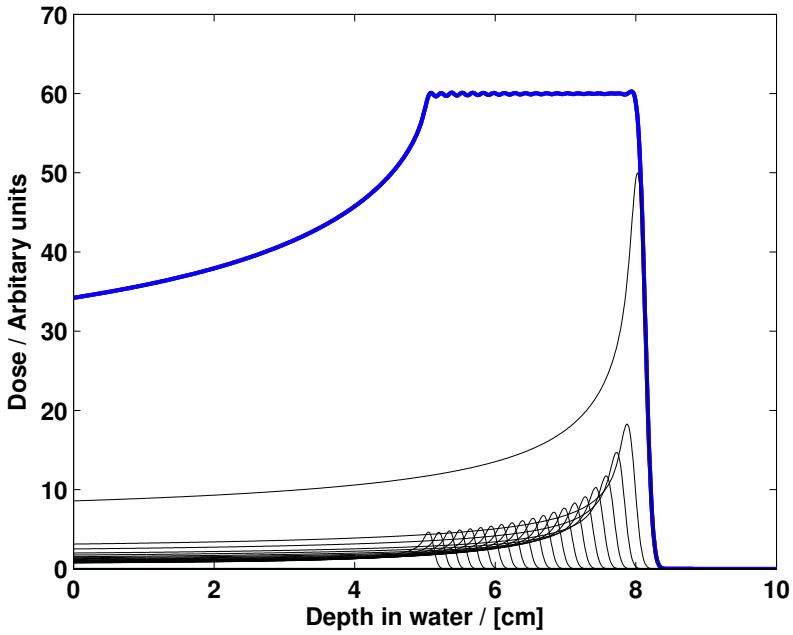

Figure 2.2: *To generate a homogeneous dose in depth, Bragg peaks of different energies with different weights are superimposed. This is called a Spread-out Bragg Peak (SOBP).*

using a modulator wheel which is spinning in the beam line. This wheel consists of sectors of varying thickness of material. It resembles a spiral staircase, where the height of each stair determines the individual energy and the width of each stair its weight. Two patient specific devices are additionally used to deliver proton therapy with passive techniques. A collimator is applied to laterally shape the beam to the tumor. Finally a compensator is used to align the distal edge of the tumor with the distal edge of the SOBP. The devices are schematically presented in figure 2.3, where it is indicated that the width of the SOBP is determined by the largest extension in depth of the tumor. Since the width of the SOBP cannot be changed laterally using a passive technique, a large part of the surrounding healthy tissue is receiving the full therapeutic dose. Additionally, the material placed into

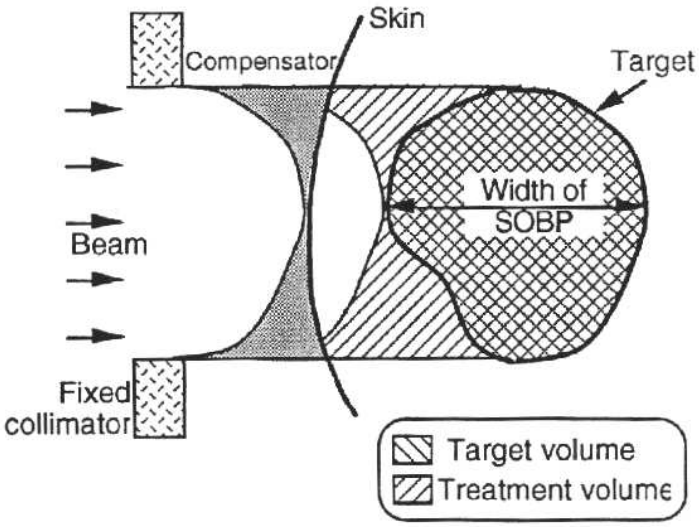

Figure 2.3: *Picture taken from [16]. Patient specific devices and the resulting dose distribution is sketched for passive proton therapy.*

the proton beam like the scattering foil, the collimator and the compensator results in the production of neutrons which can be harmful to the patient (see e.g. [21]). These two problems are the main disadvantages of passive techniques, the main advantage being its simplicity.

2.2.2 Active techniques

The active delivering techniques were pioneered in the 1990s at the Paul Scherrer Institute (PSI), Switzerland, for proton therapy and at the Gesellschaft für Schwerionenforschung (GSI), Darmstadt, Germany, for carbon ion therapy. The principle of active scanning is sketched in figure 2.4. Instead of broadening the beam by scattering foils, a narrow proton beam with a fixed energy is scanned over the tumor using scanning magnets. After the tumor has been scanned the energy of the proton beam is changed, either passively by inserting a plate of material, a so-called range shifter, into the beam line or actively by directly changing the accelerator settings. This process is repeated until the whole tumor has been irradiated. The active delivering technique is technically more challenging than the passive technique. While for the passive technique the SOBP is shaped using one universal modulator wheel per SOBP width, the number of protons applied to each scanning position has to be calculated for each patient individually. As for intensity modulated therapy this is done using inverse treatment planning. This is described in the following section. To control

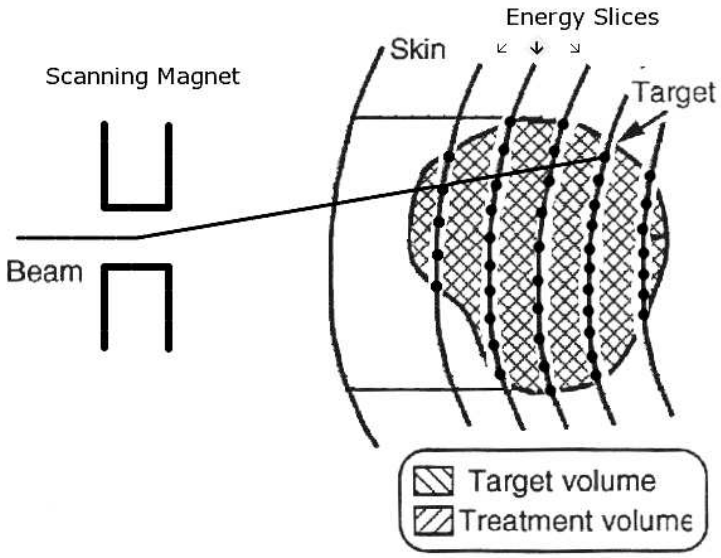

Figure 2.4: *Picture modified from [16]. Using the active delivering technique, the narrow proton beam is scanned over the tumor. By changing the beam energy the range of the Bragg peak can be adjusted to the corresponding energy slices. For simplicity, only one scanning magnet is shown. A second scanning magnet is needed to deflect the proton beam in the third dimension.*

the beam scanning within the required accuracy is another challenge. This is especially true if the tumor is able to move during the treatment as it is the case for e.g. lung tumors. There the interplay between the tumor motion and the beam scanning can deteriorate the dose distribution. These interplay effects are one reason why the active scanning technique has mainly been applied to immobile tumors such as brain tumors and lesions at the spinal cord. However compared to the passive technique the active technique offers the possibility to spare healthy tissue proximal to the tumor since the width of the SOBP can be adjusted for each lateral position. Furthermore much less material is placed in the proton beam using the active technique, thus much less neutrons are generated.

2.2.3 Intensity Modulated Proton Therapy (IMPT)

Both the active and the passive technique can be applied with only one treatment beam. Even if multiple beams are used in these techniques every beam is applied independently, that is every beam results in a homogeneous dose distribution in the target. In intensity modulated proton therapy (IMPT), multiple treatment beams are used. The superposition of all treatment beams result in a homogeneous dose distribution to the target, whereas each treatment beam alone can deliver an inhomogeneous dose to the target. To generate

IMPT plans, the treatment beams are subdivided into multiple beamlets. The treatment beam can now be modulated by assigning a weight to each individual beamlet. The optimal beamlet weights are usually identified by an iterative process, the so-called optimization loop. For this an initial weight is assigned to each beamlet. The dose distribution due to these beamlet weights is then calculated and compared with the desired dose distribution. The difference between these two dose distributions is quantified by the so-called objective (or cost) function. The beamlet weights are then updated such that the value of the objective function is reduced. This loop is then repeated until a suitable treatment plan is identified. This process is called inverse treatment planning. In the DKFZ, the inverse planning is done using the research version of the in-house treatment planning software KonRad [24]. The following sections describe the elements of the the inverse treatment planning process for IMPT in more detail. The optimization process to update the beamlet weights will be the topic of chapter 3 and is excluded here.
Inverse treatment planning is also used to determine the weights of the individual Bragg peaks for the active delivery technique. Thus the active delivery technique could also be seen as intensity modulated proton therapy. However it has been agreed on that only techniques where each individual treatment beam may result in an inhomogeneous dose distribution are called IMPT [22]. This is motivated by the fact that for the active technique there is only one modulation of the treatment beams that leads to the desired dose distribution. In contrast to this there are usually multiple sets of beamlet weights which result in comparable dose distributions for the IMPT techniques. The inverse problem for IMPT is thus degenerated.

IMPT techniques

Lomax described four techniques to deliver IMPT [23]. The methods differ in the way they subdivide the treatment beam into individual beamlets. They are illustrated in figure 2.5. The beamlets for the 2D method are narrow spread-out Bragg peaks (SOBPs). As for the passive technique, the extension of the SOBP in depth is identical for each lateral position. The beamlets for the 2.5D method are SOBPs with a variable plateaulength such that the size of the SOBP is matched with the target size at each lateral position.
The distal edge tracking (DET) technique proposed by Deasy et al. [25] uses single Bragg peaks as beamlets. However only one Bragg peak is used for each lateral position, namely the Bragg peak which stops at the distal edge of the target. With this technique a homogeneous dose in the target can only be achieved if multiple beam directions are employed. Finally the 3D method proposed by Brahme et al. [26] uses individual Bragg peaks as

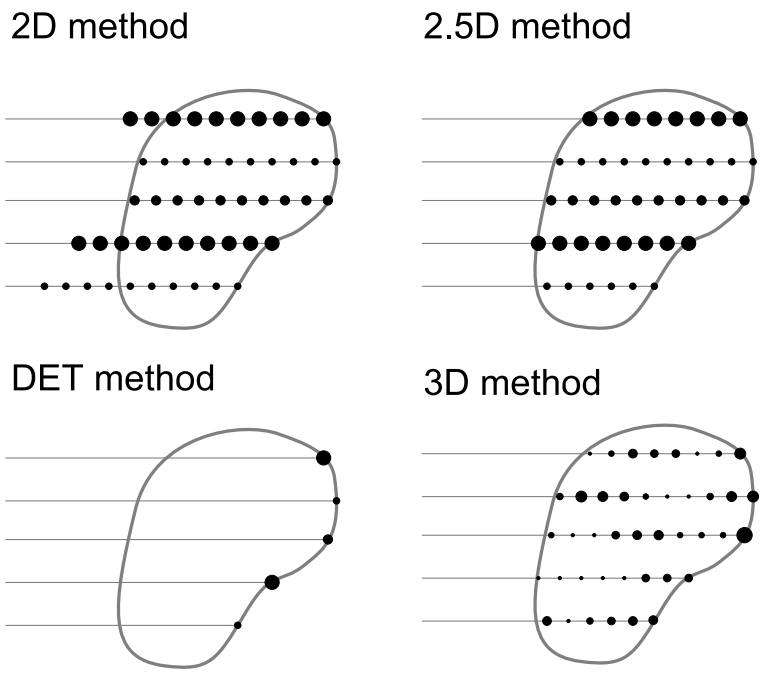

Figure 2.5: *The four methods to deliver IMPT as described by Lomax [23].*

beamlets. Contrary to the DET method Bragg peaks of multiple energies are used at each lateral position, covering the target from the distal edge up to the proximal edge.
The full three dimensional modulation is the most flexible IMPT technique, the other methods are merely constrained versions of it. All solutions obtained with the other three methods can also be obtained using the 3D technique. The rationale for the other three methods is to reduce the number of beamlets which reduces the complexity of the inverse problem as well as the complexity of the treatment delivery. Depending on the actual geometry the number of beamlets can be reduced by approximately a factor of ten when changing from the 3D to e.g. the DET method. Additionally, the use of SOBPs in the 2D and 2.5D method can decrease the sensitivity to range uncertainties of the resulting treatment plan.
Although the DET method only uses a very limited number of Bragg peaks it can, depending on the tumor size and the number of used treatment beams, result in dose distributions which are comparable to the 3D method (see e.g. [27]). Figure 2.2 shows that for an SOBP dose delivery the Bragg peak at the distal edge has by far the largest weight to generate a homogeneous dose in the target. Furthermore this Bragg peak has the best dose to target / dose to normal tissue ratio. Thus the DET method concentrates on the most important Bragg peaks.

Objective function

The optimal dose distribution in radiation therapy would be a prescribed dose applied homogeneously to the target with no dose to the surrounding healthy tissue. Obviously this goal cannot be fulfilled. The objective function F quantifies the difference between the actual and the desired dose distribution. Since the actual dose distribution depends on the beamlet weights this function can be expressed in terms of the beamlet weight vector, $\vec{w}$. Thus $F(\vec{w})$ is a N-dimensional function, N being the number of beamlets. Several different objective functions have been proposed. In this thesis the following quadratic objective function is used:

$$\begin{aligned} F(\vec{w}) \quad &= \frac{1}{N_{Target}} \sum_{i}^{N_{Target}} \left(p_{u_i} \left[D_{pres_i} - D_i(\vec{w}) \right]_+^2 + p_{o_i} \left[D_i(\vec{w}) - D_{pres_i} \right]_+^2 \right) \\ &+ \sum_{k}^{OAR} \frac{1}{N_k} \sum_{i}^{N_k} p_{o_i} \left[D_i(\vec{w}) - D_{tol_i}^k \right]_+^2 \qquad (2.1) \\ &\text{with} \quad D_i(\vec{w}) = \sum_{j}^{N} w_j D_{ij} \qquad (2.2) \\ &\text{and} \quad [x]_+ = 0 \ \text{ for } \ x < 0, \quad [x]_+ = x \ \text{ for } \ x \geq 0. \end{aligned}$$

Here N_m is the number of voxels in the respective volume m. $D_i(\vec{w})$ is the actual dose in voxel i. D_{pres_i} is the dose prescribed to the target, $D_{tol_i}^k$ the dose tolerated by the k-th organ at risk (OAR). p_{u_i/o_i} are penalty factors for the under/over dosage of voxel i. D_{ij} is the dose in voxel i due to unit fluence of beamlet j. This D_{ij} (or influence) matrix can be calculated once prior to the optimization. During the optimization, where multiple dose calculations are required, this D_{ij} matrix can be used as a quick lookup table for a fast dose calculation. With this approach demanding dose calculation algorithms as e.g. Monte Carlo algorithms can be used in the inverse treatment planning process.
Due to the positivity operator $[\cdot]_+$ this objective function is strictly speaking not a quadratic function. Nevertheless it is commonly referred to as quadratic objective function thus this terminology is kept in this thesis.

2.3 Proton dose calculation

Several dose calculation algorithms are available for proton therapy. An overview can be found e.g. in [27]. Two common approaches are described in this section, namely the pencil beam algorithm (PB) and the Monte Carlo approach (MC). The description of the pencil

beam algorithm follows reference [27]. While pencil beam algorithms are very fast and reasonably accurate, Monte Carlo algorithms provide a gold standard at the cost of prolonged calculation times.

2.3.1 Initial phase space

Often the modeling of the particle accelerator is decoupled from the dose calculation using the concept of the initial phase space (IPS) [27]. The initial phase space $IPS(\vec{r}, \vec{p}, E)$ is a complete description of the beam in a plane perpendicular to the beam axis, including the direction of flight ($\vec{p}$) distribution and the energy (E) distribution of the incoming protons at each position $\vec{r}$ of the plane. The energy distribution of the protons is usually assumed to be independent of the lateral position, which reduces the complexity of the initial phase space. This thesis investigates IMPT which is delivered using a scanned narrow beam. Such a beam is controlled by the beam optics, which usually yields a Gaussian lateral distribution and an almost parallel momentum distribution of the protons. Assuming a parallel initial phase space with all protons moving in the direction of the beam axis further reduces the complexity of the IPS.

2.3.2 Pencil beam algorithm

The pencil beam algorithm described here makes use of the two simplification described in the previous section, thus it assumes a parallel initial phase space with a constant energy spectrum. The initial phase space is thus reduced from $IPS(\vec{r}, \vec{p}, E)$ to $IPS(\vec{r})$. The pencil beam algorithm calculates the dose due to an infinitesimally narrow proton beam with the given energy spectrum starting at the plane on which the IPS is defined. By convolving the dose of this pencil beam, $P(\vec{r}, d, E_0)$, with the IPS the complete dose distribution D is obtained:

$$D(\vec{r}, d) = \int d^2\vec{r'}\ IPS(\vec{r'}) \cdot P(\vec{r} - \vec{r'}, d, E_0) \tag{2.3}$$

Here $\vec{r}$ are the two coordinates perpendicular to the beam axis while d represents the coordinate along the beam axis.

The pencil beam algorithm is now described for the dose calculation in a homogeneous target such as a water phantom. Two physical effects dominate the behavior of a proton beam transversing material. The energy loss leads to the characteristic depth-dose curve including the Bragg peak. The energy loss of the accelerated protons is mainly due to Coulomb interaction with the atomic electrons of the target material. Additionally, the

multiple Coulomb scattering (MCS) of the accelerated protons with the target atoms results in a broadening of the beam. Since the energy loss and the scattering are almost independent a common approximation of all pencil beam algorithms is to decouple these two effects. This leads to a factorization of the dose of the pencil beam into a central axis depth-dose curve $D_{cax}(d, E_0)$ and a lateral distribution $L(r, d, E_0)$:

$$P(\vec{r}, d, E_0) = D_{cax}(d, E_0) \times L(\vec{r}, d, E_0) \tag{2.4}$$

Both terms can either be measured or calculated using analytical models. An analytical approximation for the depth-dose curve was given by Bortfeld [18]. The multiple Coulomb scattering which leads to the lateral distribution $L(\vec{r}, d, E_0)$ was investigated by Molière [28]. He found that the lateral distribution can be described to first order by a Gaussian distribution, thus $L(\vec{r}, d, E_0) = (2\pi\sigma^2(d, E_0))^{-1} \cdot exp\left(-\frac{(\vec{r}-\vec{r_0})^2}{2\sigma^2(d,E_0)}\right)$. Here $\vec{r_0}$ is the position of the central beamlet axis and $\sigma^2(d, E_0)$ is the width of the Gaussian distribution which is dependent of the depth d in the target and the initial energy E_0 of the proton beam. With this approximation the lateral distribution can be parameterized simply by $\sigma^2(d, E_0)$. Furthermore a Gaussian initial phase space as that of a scanned narrow beam is now easily included into the dose calculation. The convolution of two Gaussian distributions a and b is another Gaussian distribution c with $\sigma_c^2 = \sigma_a^2 + \sigma_b^2$. Thus the convolution with the dose of an infinitesimally narrow pencil beam and the initial phase space can be achieved analytically by adding the two sigmas in quadrature. This approach is called finite size pencil beam algorithm. This algorithm is implemented in our in-house treatment planning system KonRad which is used throughout this thesis. For a fast dose calculation, the depth-dose curve $D_{cax}(d, E_0)$ and the depth-dependent sigma $\sigma^2(d, E_0)$ in water are stored in KonRad for each available energy as lookup tables. Thus measured data as well as base data derived from analytical models can be employed for the dose calculation.
Up to now, a homogeneous phantom was assumed. To include inhomogeneities such as air cavities or bones the depth axis is scaled. The geometrical depth d is replaced by the water equivalent (or radiological) depth WED. The WED takes into account the different slowing down properties of the different materials. The ability of a material to slow down accelerated particles can be quantified by the mass stopping power S_{Mat}. It is defined as the energy loss per unit length in that material. Often, the relative stopping power compared to water is given, that is $S_{H_2O}^{Mat} = \frac{S_{Mat}}{S_{H_2O}}$. The water equivalent depth is now defined as $WED = \sum_i \Delta d_i \cdot S_{H_2O}^{Mat}$. With this definition, a particle that traveled a distance d in the inhomogeneous target lost the same energy as a particle which traveled a distance WED in a water phantom. For the dose calculation in the inhomogeneous target, equation 2.4 is

thus changed to:

$$P(\vec{r}, d, E_0) = D_{cax}(WED, E_0) \times L(\vec{r}, WED, E_0) \tag{2.5}$$

While this one dimensional scaling is a good approximation for $D_{cax}(d, E_0)$ which describes the energy loss of the particle it can lead to large errors in the lateral distribution. Szymanowski et al. [5] presented a two dimensional scaling algorithm which improves the calculation of the lateral dose distribution in the presence of inhomogeneities. Nevertheless the commonly used 1D scaling is employed in this thesis.
Both the 1D and the 2D algorithm calculate the water equivalent depth only on the central axis of the pencil beam, thus both algorithms assume a slab geometry of the phantom. Lateral heterogeneities cannot be included into the dose calculation using a pencil beam algorithm. This is its major limitation. One approach to solve this problem known as fine sampling is to split the finite pencil beam into multiple subpencil beams [29]. However, since every subpencil beam then still needs a slab geometry for the dose calculation. But even an infinitesimally narrow proton beam of e.g. 200 MeV is broadened due to multiple Coulomb scattering to approximately $\sigma = 0.7\ cm$ in a depth in water of 25 cm [27], thus this approach is limited. In cases with large lateral inhomogeneities more accurate dose calculation algorithms as e.g. Monte Carlo methods are needed.

2.3.3 Monte Carlo algorithm

Monte Carlo (MC) algorithms trace individual particles through the medium. Models to predict the probability of specific interactions and their outcome are employed. At each interaction of the particle with the medium random numbers are used to decide the further path of the particle according to these models. This generates individual particle tracks in the medium. To calculate the dose of a proton beam, multiple primary protons are generated according to the initial phase space. The dose of all individual primary proton tracks is then averaged to yield the dose of the whole proton beam. Since the Monte Carlo algorithm is a statistical method the statistical uncertainty of the resulting dose distribution can be calculated. This can be used to estimate the number of individual protons necessary to reduce the statistical uncertainty in the dose calculation below a required level.
In this thesis the Monte Carlo code GEANT4.7.0 [30] is used. In GEANT4 the utilized models for the physical interactions are defined in the so-called physics list. A physics list suitable for proton therapy was provided by Hanitra Szymanowski. To be able to trace the protons through patients a patient model is needed which assigns a specific material to each volume element in space. The available CT of the patients only offers Hounsfield

Units (HU), thus a correlation between HUs and the corresponding materials is needed. This is done as described by Schneider et al. [31].

2.4 Challenges in proton therapy: Towards risk adapted optimization

The physical properties of a proton beam such as the distinct Bragg peak and the adjustable range of such a beam offer the possibility to create dose distributions which are superior to the ones obtained by photon radiation. However this new modality also poses new challenges. The dose gradient behind the Bragg peak is very steep. Thus small errors in the range calculation can lead to large errors in the dose distributions. Multiple sources can lead to errors in the calculated range. The calculation of the proton range is based on a CT scan of the patient. This scan is obtained using photon radiation. There is no unique transformation from the Hounsfield units (HU) resulting from a CT scan to the stopping powers required for the calculation of the proton range. As shown e.g. in [32] two materials with different stopping powers can result in the same value of the Hounsfield unit. Schaffner et al. [33] stated that solely the conversion from HU to stopping powers leads to "a range precision of about $1-3\ mm$ in typical treatment situations." Additionally to this uncertainty CT scans can exhibit large artefacts especially if a material with a high atomic number such as metal is present in the CT. These artefacts also contribute to the range uncertainty of the proton beam. Many patients receive surgery prior to the radiation therapy. Often metal implants are used to e.g. stabilize the spine after this surgery thus the metal artefacts in treatment planning CT scans are a common problem. Changes of the patient anatomy such as weight gain/loss or (hopefully) tumor shrinkage during the approximately 30 days of the treatment can change the calculated proton range.
Another risk factor are beamlets which encounter large lateral inhomogeneities as e.g. a scanned narrow beam which passes parallel to bone/air interface. Small changes such as setup errors or shifts caused by e.g. breathing motion can shift the beam to the other side of the interface, again resulting in a large change of the calculated proton range. Additionally the commonly used pencil beam algorithm cannot account for lateral inhomogeneities thus large dose calculation errors occur using this algorithm for such geometries.
In the active scanning technique a narrow proton beam is scanned over the target. If the tumor is also moveable as it can be the case e.g. for lung tumors these two motions lead to interplay effects which can deteriorate the resulting dose distribution.
As stated in section 2.2.3, the inverse problem of intensity modulated proton therapy is

degenerated, thus there are multiple sets of beamlet weights which result in a comparable dose distribution in the patient. This degeneracy can be exploited to create treatment plans with a reduced sensitivity to uncertainties. The aim of this thesis is to develop methods for such a risk adapted optimization.

Chapter 3

Optimization in IMPT

In intensity modulated radiation therapy, the optimal beamlet weights are determined by minimizing an objective function. Several optimization algorithms have been proposed for radiation therapy, see e.g. [23, 34, 35, 36]. Most optimization algorithms are local algorithms, meaning that if the function which is to be minimized has more than one minima then the optimization algorithm will terminate in an arbitrary one. This is not guaranteed to be the minima with the lowest possible function value. Usually objective functions with only one minima are used in radiation therapy to prevent the optimization algorithm to get stuck in a local minimum. Such functions are e.g. convex functions as the quadratic objective function presented in section 2.2.3. Instead of using convex functions global optimization algorithms such as simulated annealing could be used. In practice these algorithms are usually too slow [16]. The objective function is a function of the beamlet weights. These beamlet weights are constraint to positive values as negative fluences cannot be delivered. Moreover, if we could deliver negative fluences the resulting dose distributions would certainly be improved. In fact, using negative beamlet weights the optimal dose distribution could be realized. This means that the minimum of the unconstrained objective function will be in the forbidden region. Thus this positivity constraint will be important during the optimization.

In this chapter several optimization algorithms for IMPT are investigated. A summary of all optimization algorithms implemented into KonRad is presented in table 3.2 on page 29. The standard optimization algorithm (std) in our inverse planning tool KonRad is a quasi-Newton algorithm introduced by Bortfeld et al. [34] for intensity modulated photon therapy in 1990. This routine yields good optimization results for clinical cases, however it has several drawbacks: It has a very crude approach to fulfill the positivity constraint of the

beamlet weights. Often the algorithm is only slowly converging, especially towards the end of the optimization. Finally this routine uses a free parameter, the so-called damping factor α, that has to be chosen appropriately for every new optimization. The standard algorithm does contain an algorithm to determine a value for α. However it is almost impossible to find the optimal value for a parameter that changes with every new optimization. Therefore, an improved optimization algorithm has been implemented into KonRad. The improved algorithm is based on the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm [37] which is also a quasi-Newton algorithm. The implementation follows the description of Nocedal and Wright [38].
Another widely used optimization algorithm is the conjugate gradient algorithm (CG) (see e.g. [39, 36] and others). To show the results in a larger context the CG was integrated into this comparison.

3.1 Optimization algorithms

Most optimization algorithms are iterative. At each iteration, the algorithm chooses a direction and a step length in the variable space. This is repeated until a termination criteria is fulfilled. A common class of optimization algorithms are the Newton methods: At each iteration, the function is approximated at the current position by a quadratic function. The minimum of this quadratic function determines the next position. The method is illustrated in one dimension in figure 3.1. In 1D, the step p_k from position x_k to x_{k+1} is given

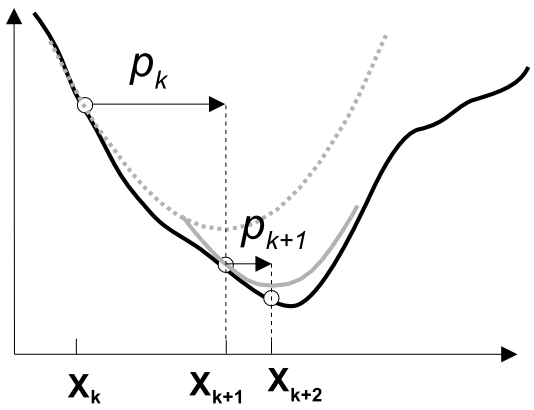

Figure 3.1: *Illustration of the Newton method in one dimension.*

by $p_k = \frac{-f'(x)}{f''(x)}$. For n dimensions, this generalizes to $\vec{p_k} = -H(\vec{x}_k)^{-1}\vec{\nabla} f(\vec{x}_k)$. The Hessian matrix H of the second derivatives is a $n \times n$ matrix. For large n it is very demanding to calculate and invert it, thus H^{-1} is often approximated. Algorithms doing so are called

quasi-Newton algorithms. The Newton method as described up to now is not guaranteed to converge. To guarantee convergence, the length of the vector $\vec{p_k}$ has to fulfill a convergence criteria. Frequently used convergence criteria are e.g. the so-called strong Wolfe conditions as described in [38]. The strong Wolfe conditions are used for the improved algorithm and are stated in the appendix for completeness.
Both the standard and the improved optimization algorithms described in the following paragraphs are quasi-Newton algorithms. They differ in the way they approximate H^{-1} and the way they handle the positivity constraint of the beamlet weights. All algorithms described terminate if the change of the objective function is less than 0.1% for two successive iterations.

3.1.1 Standard algorithm

The standard algorithm in KonRad is the algorithm published by Bortfeld et al. in 1990 [34]. It takes only diagonal matrix elements of H into account. The inversion of H then becomes trivial. The full Newton step length is scaled down by a fixed fraction, the damping factor α. Without this down scaling, the algorithm does not converge. Thus the step is finally given by $\vec{p_k} = -\alpha \left(\frac{\frac{\partial F}{\partial w_1}}{\frac{\partial^2 F}{\partial w_1^2}}, \ldots, \frac{\frac{\partial F}{\partial w_n}}{\frac{\partial^2 F}{\partial w_n^2}} \right)$. The damping factor α is chosen such that when starting from zero the first step results in a mean dose to the planning target volume (PTV) equal to the prescribed dose. This α is used throughout the optimization. As a useful by-product the weights obtained during the determination of α are used as the initial solution for the standard as well as for the improved optimization algorithm.
To fulfill the constraint of non-negative weights, all negative w_j (if any) are set to zero after each step. This projection results in a valid solution. However it can slow down the algorithm since it can go repeatedly in the same direction. The projection can even result in solutions which are worse than the previous iteration (see Fig. 3.2).

3.1.2 Improved algorithm (L-BFGS)

The improved algorithm is similar to the optimization algorithm described in [35]. It solves the non-negative constraint by a variable transformation: $w_j \rightarrow x_j^2$. Thus $F(\vec{w})$ changes to $\tilde{F}(\vec{x})$ with equation 2.2 changing to:

$$D_i(\vec{w}) \rightarrow \tilde{D}_i(\vec{x}) = \sum_{j}^{beamlets} x_j^2 D_{ij}. \tag{3.1}$$

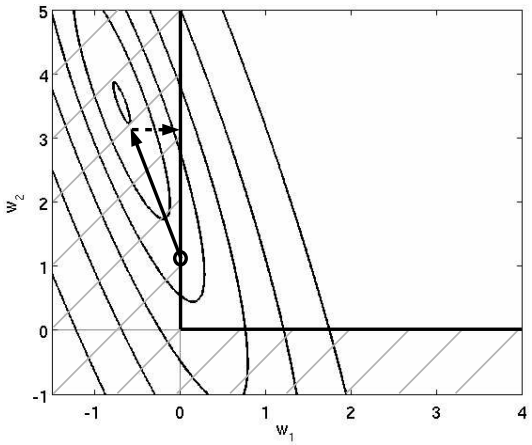

Figure 3.2: *An example objective function for two beamlets is plotted. Starting from the optimal value (circle), an iteration step will point towards the global minimum, leading to a negative weight w_1. In this case by setting w_1 to zero the resulting solution is worse than the previous iteration.*

Using this transformation each actual beamlet weight w_j is the square of an optimization variable x_j and therefore it is positive for any choice of x_j. The constrained optimization problem has thus been transformed into an unconstrained one. This is beneficial since the theory of unconstrained optimization is simpler and usually faster. However, this transformation also turns the convex objective function into a non-convex one. This issue is addressed in section 3.3.1.1.

The improved algorithm generates the iteration steps using the L-BFGS algorithm as described in [38]. This algorithm directly calculates $H^{-1}(\vec{x})\vec{\nabla}\tilde{F}(\vec{x})$ using only previously calculated positions and gradients. No second derivatives need to be calculated, nor does the algorithm need the memory to store the $n \times n$ Hessian matrix (3 GB using single precision for $n = 40000$ beamlets, a number easily reached in hadron therapy). The L-BFGS routine needs one previous step to calculate the search direction. The first iteration is thus done using the standard algorithm.

To ensure convergence, an inexact line search along the direction $\vec{p}_k$ is invoked after each iteration. Inexact means that it is not locating the minimum along the line, but only identifies a point that fulfills a convergence criteria. The commonly used strong Wolfe conditions are employed as this convergence criteria. They are stated in the appendix. By choosing the parameters as recommended by [38] for the strong Wolfe conditions, the length of the vector $\vec{p}_k$ has to be adjusted in only a few percent of all iterations. Only then an additional time-consuming function evaluation is necessary. Thus there is not much time lost due to this additional line search. The L-BFGS algorithm needs in each iteration an initial solution

for the inverse of the Hessian matrix. This initial solution can be used as a scaling factor as described in [38]. This leads to well-scaled search directions which also contributes to the fact that the line search does usually not need extra function evaluations.
The constants used in the algorithm follow the recommendation of [38], except for the number m of stored positions and gradients. There I used a value of $m = 100$ instead of the recommended $m = 15$. This number is a tradeoff between the accuracy for the estimation of $H(\vec{x})^{-1}\vec{\nabla}F(\vec{x})$ and the time to calculate this product. However, in this case the time to calculate the product is negligible compared to the time of the function evaluation. Thus going to a higher m showed better results since the improved accuracy saved extra function evaluations.

3.1.3 Conjugate gradient algorithm

The intuitive way to choose a new search direction is the direction of the steepest descend, thus along the gradient. This method known as steepest descend or simple gradient method can be "... excruciatingly slow on difficult problems." [38]. This is due to the fact that using this approach the search direction is selected independently for each iteration. This can lead to very similar search directions and thus to little progress in the optimization. The method of conjugate gradients (CG) avoids this problem by taking into account the previous search directions, resulting in a set of search directions which are said to be conjugate to each other (for a detailed description of the conjugate gradient method see e.g. [38, 40]). To generate the next search direction out of this set only the gradient at the previous and the actual position is necessary. Thus the memory required is only $2 \cdot n$, which makes this method attractive for problems with many degrees of freedom. The set of search directions are only conjugate to each other if the exact minimum along the search direction is located, thus an inexact line search as for the improved algorithm is not feasible here. In contrast to the L-BFGS method the CG method does only give a search direction. No estimation of a suitable steplength along this direction is given.
The implementation of the conjugate gradient algorithm in this thesis uses the well established routines from the "Numerical Recipes in C"[40]. I used the recommended Polak-Ribière formula to update the search direction. The exact line minimization is done using the "Numerical Recipes" routines "mnbrak" to bracket the minimum and "brent" to locate it. A tolerance of 1% was used for the line search. Since a quadratic objective function is used the routine "brent" is very suitable for the line minimization as it employs parabolic interpolation to locate the minimum. Both methods to assure positive beamlet weight were investigated in combination with the conjugate gradient method. In the following section,

negative beamlet weights are set to zero after each iteration as it is done in the standard algorithm. This approach is compared to the variable transformation employed in the improved algorithm in section 3.3.1.
Although the line minimization makes use of the quadratic form of the objective function it does need several function evaluations, ranging from five to over twenty in the examples shown in this chapter. This is the main drawback compared to the other two optimization algorithms.

3.2 Comparison of the three optimization algorithms

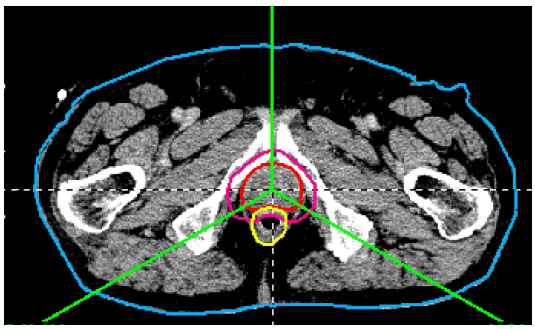

Figure 3.3: *A CT slice of the prostate patient data used for the example treatment plan. The plan consists of three coplanar, equidistant proton fields. Spacing of the Bragg peaks is 3 mm in the lateral directions as well as in depth. This results in 28586 individual beamlets.*

Several treatment plans have been optimized using all three optimization algorithms. This includes photon as well as proton treatment plans. The influence matrix has been pre-calculated before the optimization, using a pencil beam algorithm for the photon and the finite size pencil beam algorithm for the proton beamlets. Typical values of the grid sizes for the dose calculation were between 2.6 *mm* and 3.0 *mm*. One example treatment plan is shown here in detail. It is a three field IMPT treatment plan of the prostate cancer patient shown in figure 3.3. The development of the objective function value over time is plotted in figure 3.4. The dashed lines show the objective function value at termination of the respective algorithm. For all three algorithms the optimization was repeated without the termination criteria. Switching off the termination criteria shows that the standard optimization algorithm reaches the same objective function value as the improved algorithm using the termination criteria, however it converges very slowly towards the optimal solution. Obviously, the objective function value cannot be decreased to zero. However, the minimal possible objective function value is unknown. In figure 3.4, the solid black line indicates the smallest objective function value (10678) found without a termination

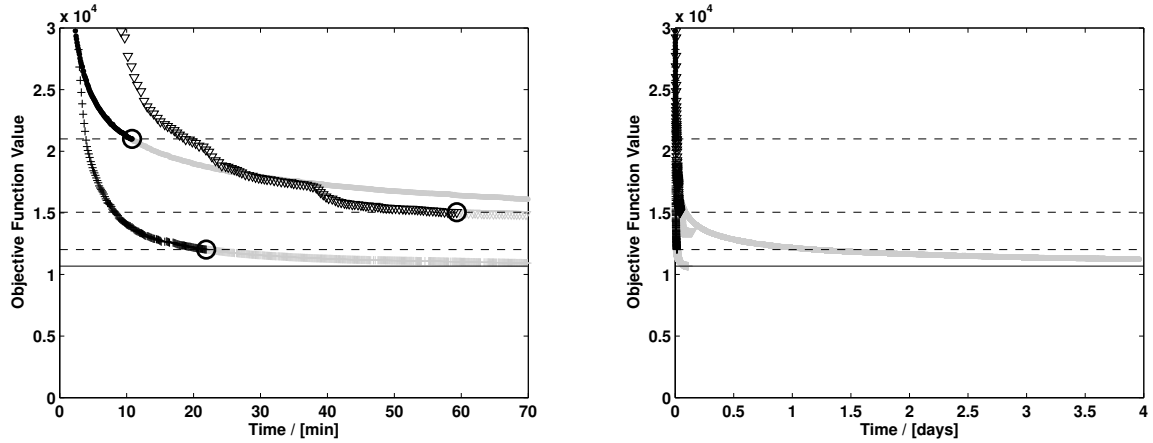

Figure 3.4: *(Left) Development of the objective function value over time for the proton treatment plan shown in figure 3.3. Dots: standard algorithm, crosses: improved algorithm, triangles: conjugate gradient. The circles indicate the points where the termination criteria was fulfilled, the dashed lines the respective values of the objective function. The grey symbols show the further behavior of the three algorithms if the termination criteria was switched off. (Right) Identical graph with an expanded time axis to show the convergence of the standard algorithm without using a termination criteria.*

criteria. At that point, numerical problems (division by zero) were observed. Such problems are expected at the true minimum of the objective function due to the limited accuracy of floating point variables, thus this value is expected to be near the lowest possible one.

In this example, the improved algorithm (*imp*) takes longer to reach the termination criteria than the standard algorithm (*std*) ($t_{imp} = 1322\ s$, $t_{std} = 652\ s$, optimization time change: $TC_{imp} = \frac{t_{std}-t_{imp}}{t_{std}} = -103\%$). However, it stops at a much lower objective function value ($F_{imp} = 12024$, $F_{std} = 21000$, objective function value improvement: $FC_{imp} = \frac{F_{std}-F_{imp}}{F_{std}} = 43\%$). To decouple these two effects I also considered the time to reach the same goal, thus the same objective function value. The objective function value at termination of the standard algorithm is taken for this purpose. While the standard optimization algorithm takes 652 s to reach this goal, the improved optimization algorithm reaches this value in 214 s. Thus the optimization can be performed faster by a factor $f_{imp} = 3.0$.

The results for the conjugate gradient algorithm (CG) are: $t_{CG} = 3128\ s$, optimization time change $TC_{CG} = -380\%$, final objective function value: $F_{CG} = 15046$, objective function value improvement: $FC_{CG} = 28\%$, $f_{CG} = 0.7$.

In addition to the decrease in optimization time, it is also worthwhile to look at the quality of the resulting treatment plans. The resulting dose volume histograms (DVH) after termination of the algorithms are plotted in figure 3.5. The improvement of the objective

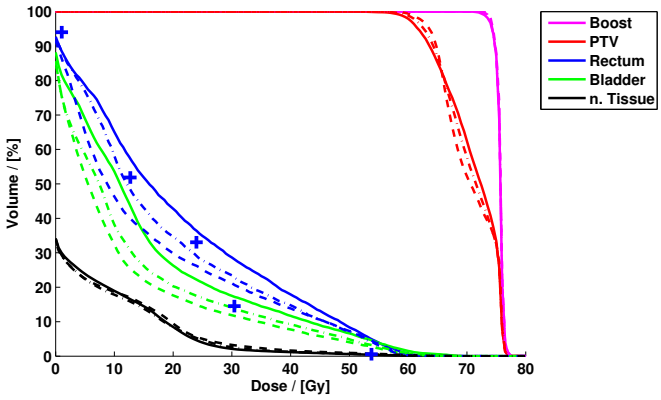

Figure 3.5: *The DVHs resulting from the optimization using three different optimization algorithms. The prescribed dose to the PTV was* 66 *Gy. It included a boost volume prescribed to* 76 *Gy. Crosses: DVH constraints for the rectum. Used dose tolerances: bladder* 30 *Gy, normal tissue* 20 *Gy. Solid line: standard algorithm; Dashed line: improved algorithm; Chain dotted line: conjugate gradient.*

function value by 43% using the improved algorithm clearly results in an improved DVH. The objective function value improvement of 28% using the conjugate gradient algorithm results in a DVH in between the DVHs from the other two optimization algorithms.
The discussed parameters (optimization time improvement, objective function value improvement and speedup factor f) were calculated for several cases. The results are summarized in table 3.1.

	Degrees of freedom	Opt. time change $TC_i = \frac{t_{std}-t_i}{t_{std}}$		Obj. function improv. $FC_i = \frac{F_{std}-F_i}{F_{std}}$		f_i	
Case		imp	CG	imp	CG	imp	CG
Head and neck, prot.	76775	+24%	−113%	+41%	+41%	5.1	2.4
Brain, protons	41946	+9%	−8%	+51%	+32%	7.1	2.6
Prostate, protons	28586	−103%	−380%	+43%	+28%	3.0	0.7
Lung, protons	3464	+47%	−123%	+61%	+58%	11.7	2.6
Prostate, photons	502	+57%	+4%	+19%	+16%	6.3	2.6
Prostate, photons	450	+34%	−14%	+6%	+3%	2.9	1.2
Average	n/a	+11%	−106%	+37%	+30%	6.0	2.0

Table 3.1: *Results for various treatment plans. Proton treatment plans employ IMPT using the full three-dimensional modulation technique, while photon treatment plans use IMRT. t and F refer to the optimization time and objective function value at termination of the respective algorithm. f is the speedup factor for the respective algorithm (see text).*

3.3 Constraints

3.3.1 Positivity constraints

Besides using the L-BFGS algorithm, the improved optimization algorithm also applies a variable transformation to turn the constrained optimization into an unconstrained one. To investigate the effect of this variable transformation the case shown in section 3.2 was optimized using the L-BFGS algorithm alone. As for the standard algorithm, negative beamlet weights are set to zero after each iteration. The results are compared to the improved algorithm in figure 3.6. Again the optimization was done with and without termination criteria.

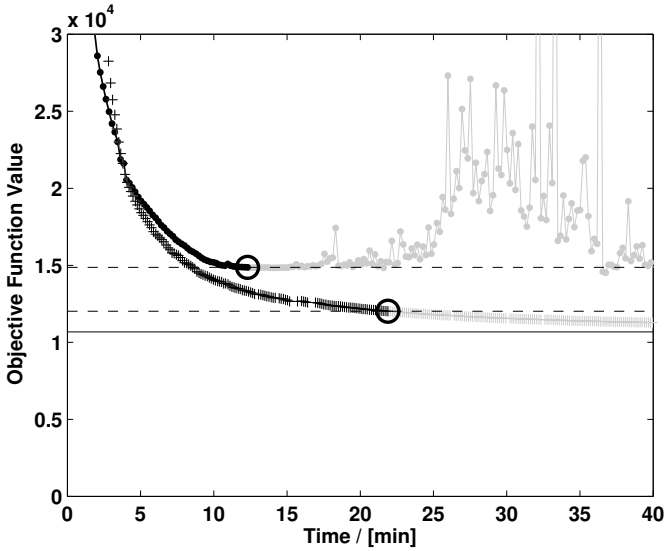

Figure 3.6: *Development of the objective function value using the improved algorithm with (crosses) and without (dots) variable transformation. After the termination criteria, the curves are shown in grey. Lines and circles as in figure 3.4. Note that the scales of the y-axis are identical in both graphs. The fluctuations raise the objective function value well above F_{std}, a value which has been shown to result in a clearly deteriorated DVH (see figure 3.5). Thus the fluctuations shown here are large.*

Without the variable transformation two effects were observed. Firstly, the optimization was terminated at a higher objective function value (14872 compared to 12024 for the improved optimization algorithm) and secondly, the objective function value increased for some iterations.

As stated before the plan consisted of 28586 individual beamlets. At the termination criteria, already more than 4000 beamlet weights had to be set to zero. After the termination criteria this number raised up to more than 10000. As sketched in figure 3.2 setting so many

beamlet weights to zero leads to the large increases of the objective function value after the termination criteria. In treatment plans for other patients increases of the objective function have even been observed before the termination criteria was reached. Despite the large fluctuations after the termination criteria, the algorithm still reached the objective function value of the improved algorithm at its termination. However, it took several days. This finding suggests that the treatment of the positivity constraint might be the reason for the slow convergence of the standard optimization algorithm towards the end of the optimization. To check this hypothesis I tried to implement the variable transformation into the standard algorithm. However it turned out that this algorithm needs a convex objective function. Using the variable transformation, the quadratic objective function is transformed into a function of fourth degree, thus it is no longer convex. $\tilde{F}(\vec{x})$ exhibits local maxima as well as inflection points where the second derivative is zero. As stated in section 3.1.1, the standard algorithm determines the value of the optimization variable x_i in step $k+1$ by: $x_i^{k+1} = x_i^k - \alpha \cdot \left(\frac{\partial \tilde{F}}{\partial x_i} / \frac{\partial^2 \tilde{F}}{\partial x_i^2} \right)$. Thus a second partial derivative close to zero can lead to arbitrary large x_i and thus to arbitrary large beamlet weights. Furthermore, the pure Newton method only locates extrema. If local maxima are present, this method will also converge towards a local maximum. Both effects have been observed when the variable transformation was used in combination with the standard optimization algorithm thus this approach was abandoned.

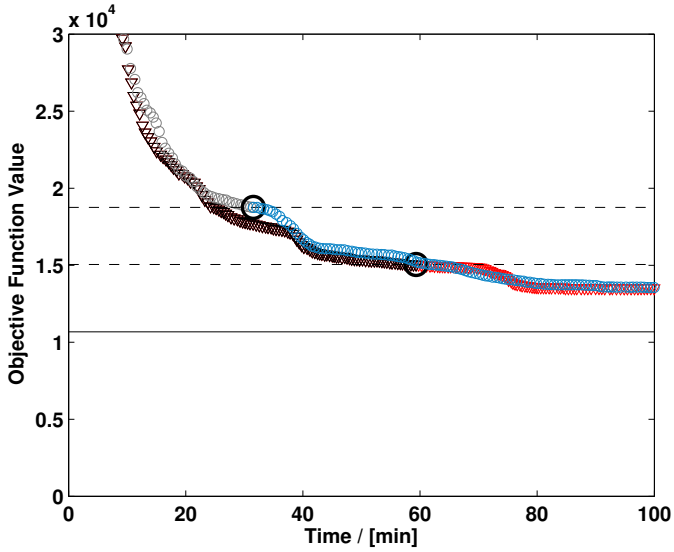

Figure 3.7: *Development of the objective function value using the conjugate gradient algorithm with (gray circles) and without (black triangles) variable transformation. After the termination criteria, the curves are shown in blue and red, respectively. Lines and circles as in figure 3.4.*

Finally, the optimization was repeated with both approaches to fulfill the positivity constraints using the conjugate gradient algorithm. The results are shown in figure 3.7. Contrary to the improved optimization algorithm, using either method resulted in a very similar development of the objective function value over time for the conjugate gradient algorithm. In figure 3.7, both variants of the CG algorithm exhibit plateaus in which the objective function value does hardly change for multiple iterations. In this case, the termination criteria was fulfilled in the first plateau using the variable transformation and in the second plateau when setting negative beamlet weights to zero. Nevertheless this does not appear to be a systematic difference between the two approaches and might be the other way around in the next optimization case. However this has not been investigated further. The conjugate gradient algorithm was not used in the subsequent chapters in this thesis since the improved optimization algorithm was found to be superior to either variant of the conjugate gradient algorithm.

Since the treatment of the positivity constraints turned out to be an important part of the optimization algorithm this issue was investigated further. For that the L-BFGS-B [41] code was implemented into KonRad as a third method to fulfill the positivity constraints. This code is an implementation of the L-BFGS algorithm which is capable to handle bound constraints on the optimization variables. Details to this algorithm are published in [42]. The L-BFGS-B code was used in three variants. As intended, the positivity bounds were explicitly specified (bounds). Obviously this algorithm can also be used without specifying bounds, treating the optimization as an unconstrained problem. This was done without the variable transformation while negative beamlet weights were set to zero after each iteration (setzero). Finally, the L-BFGS-B code was used to optimize the unconstrained objective function resulting from the variable transformation $w_j \rightarrow x_j^2$ (w2x2). The last variant is formally identical to the improved optimization algorithm. However the results vary slightly due to the different implementations. The three variants of the L-BFGS-B code were applied to a two field IMPT plan for a brain tumor patient. A CT slice of the patient including the direction of incidence of the fields is displayed in figure 3.8 (left). The development of the objective function value for all three variants is also shown this figure (right). The standard and the improved optimization algorithm are included in this graph for comparison. As expected the L-BFGS-B (w2x2) and the improved optimization algorithm terminated at a comparable objective function value, while the L-BFGS-B (w2x2) implementation was a bit faster in this case. Similar to the case shown in figure 3.6, the L-BFGS-B (setzero) terminated at a higher objective function value compared to the L-BFGS-B (w2x2) variant, and in this case it also took longer to reach the termination criteria. The L-BFGS-B (bounds) does not only take longer than the L-BFGS-B (w2x2) to reach the termination

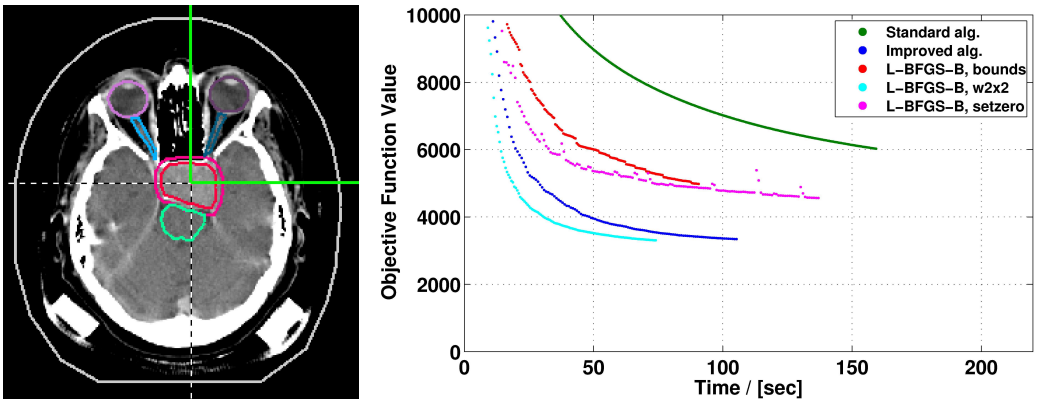

Figure 3.8: *(Right) Development of the objective function value for different optimization algorithms in a two field IMPT plan for a patient with a brain tumor (Left).*

criteria. It is remarkable that the algorithm designed to handle bound constraints also terminates at a considerably higher objective function value compared to the unconstrained algorithm using the variable transformation. Although there are some concerns against the variable transformation which will be outlined in the following section, these findings suggest the use of it to avoid the problems from the constrained optimization.
Another effect is visible for the L-BFGS-B (setzero) variant in figure 3.8. Similar to the case in figure 3.6 there are increases in the objective function value due to the projection of the negative beamlet weight to zero. However here the increases in the objective function value are present before the termination criteria is reached. Nevertheless the large fluctuation as in figure 3.6 are not present in figure 3.8. The L-BFGS-B code terminates if an increase in the objective function value is detected. The algorithm has to be restarted which erases all previously calculated positions and gradients. This restart might be the reason why the large fluctuations present in figure 3.6 are not observed using the L-BFGS-B (setzero) variant.

3.3.1.1 Local minima due to the variable transformation

As stated before, the variable transformation $w_j \rightarrow x_j^2$ turns the quadratic objective function $F(\vec{w})$ into a function of fourth degree, thus the new objective function $\tilde{F}(\vec{x})$ is no longer convex. $\tilde{F}(\vec{x})$ will have multiple minima, which is the main argument against using this variable transformation. However in this section I will show that these minima are all identical and thus are all global minima.
To show this, first of all note that due to the transformation the new objective function is symmetric, thus all orthants are mirror images of each other. (An orthant is the

	name of optimization alg.	selection search direction	selection steplength	positivity constraints
1	Standard alg.	diag. elements of H	damp. factor α	setzero
2	Standard alg.	diag. elements of H	damp. factor α	$w_i \rightarrow x_i^2$
3	CG	conjugate gradient	exact line search	setzero
4	CG	conjugate gradient	exact line search	$w_i \rightarrow x_i^2$
5	Improved alg.	L-BFGS	inexact line search	setzero
6	Improved alg.	L-BFGS	inexact line search	$w_i \rightarrow x_i^2$
7	L-BFGS-B	L-BFGS	inexact line search	setzero
8	L-BFGS-B	L-BFGS	inexact line search	$w_i \rightarrow x_i^2$
9	L-BFGS-B	L-BFGS	inexact line search	constrained opt.

Table 3.2: *The list of all optimization algorithms described in this chapter. Algorithms* $2-9$ *were implemented as part of the research for this thesis. Algorithms* 7 *and* 8 *are formally identical to* 5 *and* 6*, respectively. The L-BFGS-B package was implemented for the constrained optimization capabilities (algorithm* 9*).*

space separated by the coordinate hyperplanes.) Assume the point $P(x_1, \ldots, x_i, \ldots, x_n)$ is a minimum, then $\forall\ i \in [1..n]\ \ \tilde{P}(x_1, \ldots, -x_i, \ldots, x_n)$ is also a minimum. However, they are identical since the actual weight of beamlet i is x_i^2. Thus it is sufficient to concentrate on the positive orthant. It is left to show that due to the transformation it is not possible to generate a second minimum in the positive orthant. If this is true every minimum in the new objective function is identical and global.

An intuitive way to show this is to note that in the positive orthant the transformation $w_j \rightarrow x_j^2$ is strictly monotonically increasing. Thus in this orthant the transformation is simply a non-linear stretching. Lets assume that the graph of the old objective function in the positive orthant is given. Then the graph of the new objective function can be obtained leaving the old graph untouched simply by replacing the numbers on the coordinate axis. This cannot introduce an additional minimum.

A more formal way to show that the variable transformation cannot introduce an additional minimum in the positive orthant is given next. For simplicity first assume that $F(\vec{w})$ is strictly convex. Then this function has only one minimum and no inflection point, thus only one point with:

$$\frac{\partial F}{\partial w_i} = 0\ \forall\ i \text{ with } w_i \neq 0 \quad \text{and} \quad \frac{\partial F}{\partial w_i} \geq 0\ \forall\ i \text{ with } w_i = 0. \tag{3.2}$$

According to the chain rule the partial derivative of the new objective function is $\frac{\partial \tilde{F}}{\partial x_i} = \frac{\partial F}{\partial w_i} \cdot \frac{\partial w_i}{\partial x_i}$, thus to introduce a new stationary point due to the variable transformation $\frac{\partial w_i}{\partial x_i}$

has to be zero. However $\frac{\partial w_i}{\partial x_i} = 2 \cdot x_i$ is only zero for $x_i = 0 = w_i$. This means that new stationary points can only be introduced on the coordinate hyperplanes. But a stationary point of $\tilde{F}(\vec{x})$ on the coordinate hyperplanes can only be a minimum if the corresponding point before the transformation fulfills equation 3.2. Assume a stationary point of $\tilde{F}(\vec{x})$ with $x_i = w_i = 0$ and $\frac{\partial F}{\partial w_i} < 0$. Then $\tilde{F}(x_1, \dots, x_i = 0, \dots x_n) > \tilde{F}(x_1, \dots, x_i = \epsilon, \dots x_n)$, thus this point is either a saddle point or a local maximum.
Since there is only one point in $F(\vec{w})$ that fulfills the equation 3.2 there can also be only one minimum in the positive orthant of $\tilde{F}(\vec{x})$.
If $F(\vec{w})$ is convex but not strictly convex there can be multiple points that fulfill equation 3.2. However, these points then all belong to the same minima, thus they all have the same objective function value. □
Although the variable transformation does not introduce local minima of different objective function value it does introduce instable stationary points (saddle points, local maxima). In principle the optimization can be terminated if by chance an optimization step ends up exactly on an instable stationary point, although this was not observed. The algorithms using the variable transformation always terminated at a lower objective function value compared to the other algorithms, indicating that they did not get stuck on an instable stationary point.
Lahanas et al. [35] investigated the L-BFGS algorithm in combination with the variable transformation $w_j \rightarrow x_j^2$. They reported a case of different objective function values at the termination of the algorithm for different starting points of the optimization. They concluded this is caused by local minima introduced by the variable transformation. Since this variable transformation does not induce local minima of different objective function value these effects might have been cases where the optimization got stuck on instable stationary points. They also presented a method to eliminate this problem.

3.3.2 DVH constraints

Additionally to the quadratic objective function, DVH constraints are a valuable tool to shape the dose distribution resulting from the inverse treatment planning process. DVH constraints enforce that not more than $x\%$ volume of an organ is receiving a dose of more than y Gy. Such constraints are used e.g. in figure 3.5. A simple way to integrate the DVH constraints into the objective function was presented by Bortfeld et al. [43]. This approach is used in KonRad. While this approach is easily integrated into the objective function, it does result in discontinuities in the objective function. These discontinuities

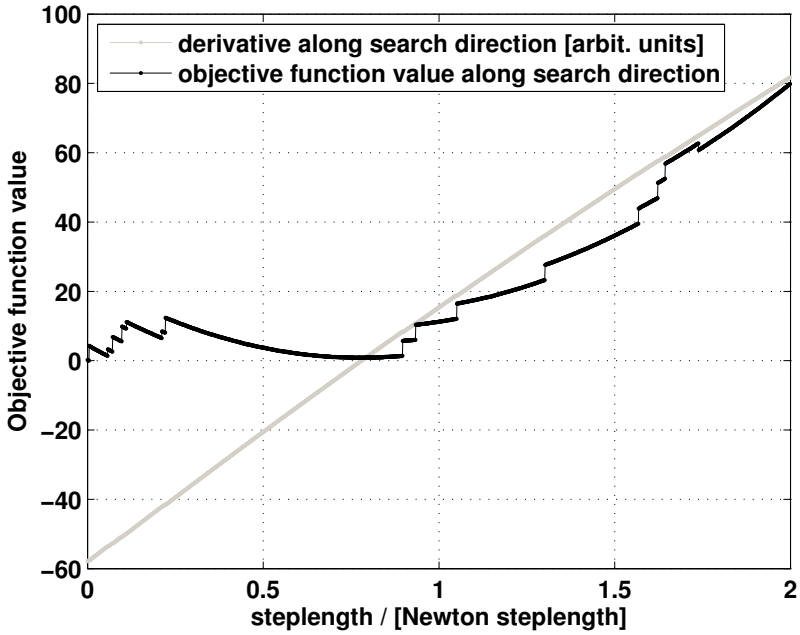

Figure 3.9: *The objective function value $\tilde{F}(\vec{x})$ and its derivative are plotted along the search direction for an iteration in the optimization of the case presented in section 3.2. The steplength is plotted in units of the newton steplength. The objective function values are shifted to zero to be able to plot the values and the derivatives into one graph. As the objective function $\tilde{F}(\vec{x})$ without the DVH constraints is a function of fourth degree, the graph of it along any direction is also a polynomial of maximal fourth degree. The DVH constraints lead to a piecewise continuous objective function where the pieces of the curve are shifted. In this example there is no steplength which fulfills the strong Wolfe conditions.*

can lead to situations where there is no point along the search direction that fulfills the strong Wolfe conditions. Such a case is shown in figure 3.9. To prevent an endless loop, the line search in the improved optimization algorithm is terminated if no suitable steplength is found within three function evaluations. Although it failed the Wolfe conditions the last trial steplength is then applied in this iteration. Up to now this approach gave good results.

3.4 Discussion

In addition to the standard optimization algorithm the conjugate gradient algorithm and the improved algorithm based on the L-BFGS method were implemented into KonRad. These algorithms were compared in section 3.2. The use of a different optimization algorithm led to a different final objective function value as well as to a different duration of

the optimization. To decouple the two effects the optimization time was compared at the point where the optimization algorithms reached the same objective function value.
The improved optimization algorithm showed the best results, leading to an average objective function improvement of 37% and an average speedup of 6.0 compared to the standard algorithm. The conjugate gradient algorithm ranked in between the other two with an average objective function improvement of 30% and an average speedup of 2.0.
The main drawback of the conjugate gradient method is the need for an exact line search. In the shown example, the conjugate gradient method needed an average of 8.9 function evaluations per iteration to perform the line search. Compared to this only 8 out of the 142 iterations needed extra function evaluations using the inexact line search of the improved optimization algorithm, leading to an average of 0.1 function evaluations per line search for this algorithm.
Beside the speedup of the optimization it is worthwhile looking at the objective function improvement. Optimization algorithms usually do not stop at the true optimum but they stop whenever they reach a termination criteria. If an algorithm is converging very slowly as it is the case for the standard algorithm this termination criteria can be fulfilled even if the resulting treatment plan is not near the optimum yet. As shown in figure 3.5, solely the change of the optimization algorithm can lead to considerably improved treatment plans. The example plans calculated for particle therapy benefited most from the improved algorithm. Intuitively, this can be understood since the standard optimization algorithm does not take into account off-diagonal elements of the Hessian matrix. Off-diagonal elements of the Hessian matrix are non-zero if two beamlets deliver dose to the same voxel. By setting these elements to zero every beamlet is treated independently during each optimization step. This is a good approximation as long as there are not too many beamlets delivering substantial dose to the same voxels. While this might be a reasonable approximation for photon treatment plans, this is not the case for particle treatment plans.
Three variants to enforce the positivity constraint on the beamlet weights have been tested for the L-BFGS algorithm in section 3.3.1: Setting negative beamlet weights to zero after each iteration, using an algorithm which can handle bound constraints (and applying the variable transformation $w_j \rightarrow x_j^2$ to turn the constrained optimization problem into an unconstrained one. Although it is common practice to set negative weights to zero it is not clear how this affects the convergence properties of the respective algorithms. While the convergence of the conjugate gradient algorithm did hardly change, the convergence of the L-BFGS algorithm was severely affected by this approach, leading to large increases of the objective function value in several iterations. In addition this variant led to a final objective function value well above the minimal value found with other methods. Surprisingly

even the algorithm designed to handle bound constraints on the optimization variables terminated at an objective function value well above the one found using the variable transformation.

In conclusion the improved optimization algorithm was found to be superior to the standard optimization algorithm. This is true for the speedup factor f of the optimization as well as for the final objective function value. The conjugate gradient algorithm ranked in between the other two algorithms. Furthermore the data presented in section 3.3.1 suggests that the variable transformation $w_j \rightarrow x_j^2$ should be used to turn the constrained optimization into an unconstrained one.

Chapter 4

Quantifying lateral tissue heterogeneities

4.1 Introduction

In proton therapy with scanned narrow beams, tissue heterogeneities lateral to the beam direction are problematic for two main reasons. One reason is the dose calculation algorithm. As described in section 2.3 the commonly used pencil beam algorithms assume a slab geometry. Thus they can only account for tissue heterogeneities in depth and not lateral to the beam axis. This can introduce substantial dose calculation errors in the presence of lateral tissue heterogeneities [45]. More accurate dose calculation algorithms like Monte Carlo algorithms could reduce this error. However these techniques usually cannot be used for routine treatment planning due to the prolonged calculation times. An other complication due to lateral tissue heterogeneities is the sensitivity of the resulting dose distribution to patient setup errors. If the central axis of a scanned narrow beam is next to e.g. a air/bone interface, a small lateral shift of the patient can cause the beam spot traversing the bone instead of the air, which can shift the Bragg peak by several centimeters.
In this chapter I derive a new quantity, the heterogeneity number H_i, as an indicator to quantify the lateral tissue heterogeneity for a single beamlet i in IMPT. The full three dimensional modulation technique is used in this chapter, thus a beamlet is defined as a single Bragg peak of a scanned narrow beam. With the help of H_i, the sensitivity to setup errors and the induced dose calculation error of the pencil beam algorithm is investigated for such single beamlets of a treatment plan.

4.2 Methods

4.2.1 The heterogeneity number H_i

4.2.1.1 Definition of H_i

The influence of tissue heterogeneities lateral to the beam direction can be investigated by assessing the water equivalent depth (WED) since the WED is an integration of the slowing down properties for all materials in the beam path encountered up to this point. The heterogeneity number H_i is introduced to quantify lateral tissue heterogeneities for a single beamlet i. The definition of H_i is sketched in figure 4.1. For the definition of H_i I used a coordinate system with its origin at the Bragg peak position of the beamlet i and the z-axis pointing towards the source.

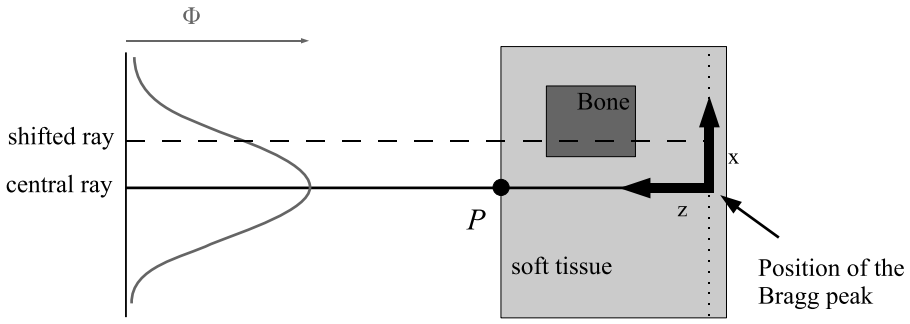

Figure 4.1: *Sketch for the definition of H_i.*

The WED is examined for all straight particle trajectories parallel to the beam direction at the anticipated depth of the Bragg peak, thus the WED on the x-y plane in figure 4.1. In a laterally homogeneous phantom, i.e. a slab geometry, the WED's of all rays should be equal throughout the x-y plane whereas in laterally inhomogeneous media variations of the WED's are expected. The square of the heterogeneity number H_i is defined as the integral over the squared differences between the WED on the x-y plane and the WED at the origin, weighted by the relative particle fluence at the entrance of the patient. Equation 4.1 gives the mathematical formula for H_i.

$$H_i = \sqrt{\frac{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \Phi_i(x,y,P_z) \cdot \left[WED_i(x,y,0) - WED_i(0,0,0)\right]^2 dxdy}{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \Phi_i(x,y,P_z) dxdy}} \tag{4.1}$$

Here, Φ is the particle fluence and P the point where the central axis of the beamlet i enters into the patient.
This definition yields H_i=0 mm for a homogeneous phantom. Heterogeneities lateral to the beamlet axis will increase the value of H_i.
The method of the heterogeneity number H_i was adopted by Soukup et al. [44] for their treatment planning system. They extended the method to quantify the total lateral heterogeneity of a treatment plan by weighting the individual H_i with the respective beamlet weights in square:

$$H_{total} = \sqrt{\frac{\sum w_i^2 \cdot H_i^2}{\sum w_i^2}} \tag{4.2}$$

This method will also be used in this chapter for the evaluation of the total lateral heterogeneity of a treatment plan.

4.2.1.2 Calculation of H_i

In practice, H_i is calculated by sampling the double integrals from equation 4.1 at discrete points, thus it transforms into equation 4.3:

$$H_i = \sqrt{\frac{\sum\limits_{j \in S_i} \Phi_i(x_j, y_j, P_z) \cdot \left[WED_i(x_j, y_j, 0) - WED_i(0,0,0)\right]^2}{\sum\limits_{j \in S_i} \Phi_i(x_j, y_j, P_z)}} \tag{4.3}$$

where S_i is an appropriate set of sampling points. Naturally, these sampling points are taken from the existing voxels of the CT grid. All voxels which are closer than $3 \cdot \sigma$ to the z-axis and closer than half the grid resolution to the x-y plane are used as sampling points, with σ being the standard deviation of the lateral Gaussian beamlet profile at P. In all cases shown in this chapter, the grid resolution for the CT data as well as for the dose calculation was $\Delta x = \Delta y = \Delta z = 1.2\ mm$. To speed up the calculation of H_i the WED of every voxel for a given beam direction is pre-calculated. The calculation of these WED-values takes about $5 - 15\ sec$. Fortunately, these values are also calculated for other reasons in the treatment planning system KonRad, so there is no additional computation time in this task. Using this method the heterogeneity numbers of roughly 500 beam spots per second can be calculated on a standard PC (3.2 GHz).

4.2.2 Beamlet dose calculation

As described in section 2.3, the finite size pencil beam algorithm used in KonRad assumes a slab geometry, thus it ignores lateral tissue heterogeneities. This will introduce dose calculation errors in the presence of such heterogeneities. To quantify these errors this pencil beam algorithm is compared to Monte Carlo calculations as a gold standard. In order to achieve a meaningful comparison of the two different dose calculation algorithms one has to make sure that both algorithms are calibrated to the same machine, thus that both algorithms use the same initial phase space. While the input data for the pencil beam algorithm can be easily measured, a major difficulty in performing Monte Carlo dose calculation is to model the unknown machine phase space. This problem was avoided by defining a virtual machine with known phase space. The parameters of this virtual machine are listed in

Depth modulation	Active Energy Selection 311 energies Range $= 4 - 35$ cm in steps of 1 mm
Machine phase space	Parallel beam Initial beam width $\sigma_x = \sigma_y = 3$ mm Gaussian energy spectrum $\sigma_E = 2$ MeV

Table 4.1: *Virtual machine parameters.*

table 4.1. Using this phase space I generated the input data needed for the pencil beam algorithm by Monte Carlo simulations. This ensured that differences between the algorithms do not result from an inconsistent modeling of the initial phase space.
As stated in section 2.3.3, the GEANT4.7.0 code was used for the Monte Carlo calculations [30]. The Monte Carlo simulations were utilized for three different applications. Firstly it was used to generate the base data for the pencil beam algorithm. Secondly it is used as the gold standard to calculate the dose deposited by single beam spots and compare it to the results obtained by the pencil beam algorithm. Finally, it is employed in section 4.3.5 to recalculate complete treatment plans in the presence of setup errors.
Commonly Monte Carlo simulations calculate the dose to the medium, whereas the pencil beam algorithm calculates dose to water. To be able to compare the two algorithms every dose deposited by the Monte Carlo algorithm was converted to dose to water by multiplying it with the mass stopping power ratio of water and the medium for the actual particle and energy.
As stated in section 2.3.3, the number of protons used in the Monte Carlo simulation de-

termines the statistical uncertainty of the calculated dose. For each beamlet I employed 300,000 primary protons. The statistical uncertainty of the calculated dose in the lateral profile for a typical beamlet is shown in figure 4.2. Two methods to generate the primary protons in the Monte Carlo simulations were tested. Usually, the particles are generated according to the generation probability defined by the initial phase space of the beamlet, e.g. with a Gaussian lateral profile and Gaussian energy spectra. The contributions of every primary particle are then summed up using an uniform weight of one for each primary particle track. However, this is equivalent to generating a uniform distribution of primary particles and weighting each particle track with its generation probability. But with the second method it is possible to calculate the dose distribution for multiple beamlets in one run. For a simulated primary particle the generation probability for every position of a beamlet can be calculated and thus each simulated primary particle can be used for several beamlets. This method can speed up the calculation if there are overlapping generation probabilities for the beamlets, which is especially the case for only slightly shifted beamlets. This second method is employed in section 4.2.4 where the impact of setup errors on the resulting beamlet dose distribution is investigated. The drawback of the second method is that it increases the statistical noise of the resulting dose distributions as shown in figure 4.2, thus more primary particles are needed to reduce the statistical uncertainty of the calculated dose to the same level.

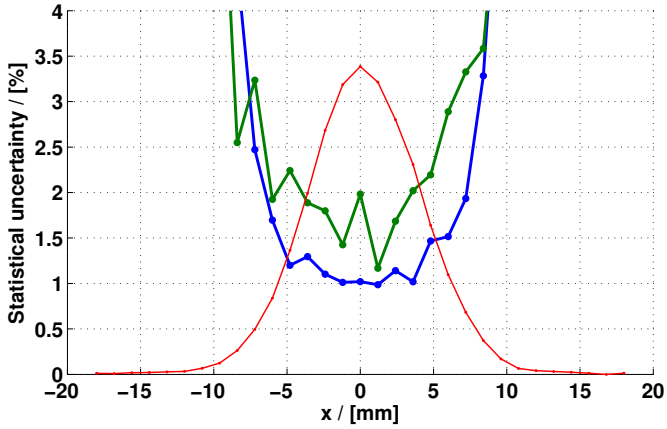

Figure 4.2: *The statistical uncertainty in the lateral profile of a typical beamlet dose distribution resulting from a Monte Carlo simulation using* 300,000 *primary protons. Generating the primary protons according to the initial phase space of the beamlet (blue) results in less statistical uncertainty compared to generating the primary protons uniformly (green). However, the second method is more flexible. The lateral dose profile (red) in arbitrary units is included in this graph for orientation.*

4.2.3 Beamlet dose comparison

In proton therapy, the protons stop completely inside the patient. Thus the total energy per proton deposited inside the patient is the initial energy E_0 of the proton beam, up to a small fraction of energy which escapes as neutrons and photons. In contrast to the integral dose the summation of the deposited energy only depends on the initial energy of the proton beam and is independent on the position of the beamlet. Thus I used the deposited energy to quantify the difference between two three-dimensional beamlet dose distributions. To convert the dose distribution into a deposited energy distribution the density of each voxel is needed. For this the same conversion from Hounsfield units to materials as already used for the Monte Carlo calculations [31] was employed.
Assume there are two dose distributions for a single beamlet, *mc* and *pb*. I defined the error between the two, ΔE, as the summation of the absolute difference in deposited energy for all voxels, divided by two:

$$\Delta E = \sum_{i}^{\text{all voxels}} \frac{|\Delta E_i|}{2} = \frac{1}{2} \sum_{i}^{\text{all voxels}} \left| E_i^{\text{mc}} - E_i^{\text{pb}} \right| \tag{4.4}$$

The factor of $\frac{1}{2}$ is introduced since the absolute value of ΔE_i is used. If for some reason the two dose distributions do not overlap at all, this definition results in an error of $\Delta E = E_0$. Thus the magnitude of the error can be readily seen, knowing that the therapeutic proton beam energy ranges from 70 MeV to 250 MeV.
Instead of using $|\Delta E_i| = \left| E_i^{\text{mc}} - E_i^{\text{pb}} \right|$, the energy difference per voxel can also be calculated using the gamma index method [46, 47]. The gamma index allows to compare the dose in two voxels, in such a way that an uncertainty in dose, ΔD, as well as a geometrical uncertainty, Δx, is tolerated. Since the Monte Carlo method has an inherent statistical uncertainty I made use of the gamma index method. $|\Delta E_i|$ was calculated the following way: if a voxel i passes the gamma test, $|\Delta E_i|$ was set to zero. If it fails, $|\Delta E_i|$ is set to the smallest energy difference found within the geometrical uncertainty: $|\Delta E_i| = \min\left(\left| E_i^{\text{mc}} - E_j^{\text{pb}} \right|\right)$ with j being all voxels closer than Δx to the voxel i. As parameters for the gamma test I took $\Delta D = 3\%$ and a geometrical uncertainty of one voxel (including those neighboring voxels that have either a face or an edge in common) thus $\Delta x = \sqrt{2} \cdot \text{grid size} = \sqrt{2} \cdot 1.2\text{ mm} = 1.7\text{ mm}$. As shown in figure 4.2, the statistical uncertainty from the Monte Carlo calculations varies for different voxels of the dose distribution. The value of $\Delta D = 3\%$ was chosen as a conservative estimate of the statistical uncertainty.

4.2.4 Beamlet dose sensitivity to setup errors

Setup errors not only lead to a shifted dose distribution, they can also change the shape of the dose distribution. While a shifted dose distribution can be handled using safety margins, margins can hardly be adopted to a changed shape of the dose distribution. In this chapter the sensitivity to setup errors refers to the change of the shape of the distribution. The sensitivity to setup errors for a single beamlet is quantified in the following way: The dose of the beamlet is calculated at the intended position, as well as in 12 laterally shifted positions as shown in figure 4.3. The resulting dose distribution of each position is shifted back to

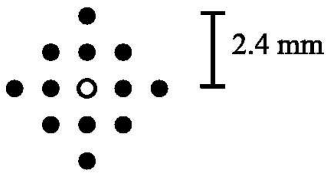

Figure 4.3: *Sampling positions in beam's eye view.*

the original position, so that the only difference to the intended dose distribution results from the different shape of the two. The back-shifted and the intended dose distribution are then compared as described in the preceding section. With this procedure an error for each sampling position in figure 4.3 is determined which only results from the changed shape of the dose distribution. These errors are then summed up, weighted with the probability for the respective setup being realized. For the setup probability I assumed a Gaussian distribution of the lateral shift with $\sigma_x = \sigma_y = 1.5$ mm.
The dose distributions were calculated using the Monte Carlo algorithm. As described in section 4.2.2, the primary protons were generated uniformly across the lateral coordinates thus each particle track calculated in the Monte Carlo simulation could be utilized for the dose calculation of each lateral sampling position sketched in figure 4.3.

4.2.5 Including H_i into the optimization

Typically, there are only a few beam spots that encounter large lateral tissue heterogeneities in a treatment plan. When using IMPT the dose contributions of these beamlets can usually be compensated for by beamlets delivered from a different beam entry angle. The easiest way to incorporate the information of H_i into the optimization is to introduce a cutoff value H_{cut} and to exclude all beamlets i with $H_i > H_{cut}$ from the optimization. Using this approach, however, all beam spots i with $H_i < H_{cut}$ are treated as equally applicable. This

is problematic especially for a large cutoff value, which is often necessary to guarantee target coverage. Thus I included the heterogeneity number into the optimization by introducing an additional term into the objective function F:

$$F_H = F + p_0 \sum_{i=1}^{N} \frac{w_i}{\sum_{j=1}^{N} w_j} \cdot p(H_i). \tag{4.5}$$

Here w_i is the weight of the beamlet i, p_0 is a constant to tune the importance of the additional term and $p(H_i)$ is the penalty for the beam spot i with heterogeneity index H_i. For $\sum_{i=1}^{N} w_i = 0$ this additional term is not defined. However, this case cannot occur during the optimization. For simplicity I chose $p(H_i) = H_i^2$. This choice guarantees a stronger suppression of beam spots with a large value of H_i. Of course, other functions of $p(H_i)$ are possible as well. To minimize this objective function, the improved optimization algorithm described in section 3.1.2 was employed.

4.3 Results

4.3.1 Examples for small and large H_i

The described methods were tested on several head and neck cases since for these cases large lateral heterogeneities due to air cavities and bony structures are expected. The patients were actually treated at the DKFZ with photon IMRT a few years ago. One of these cases is shown in figure 4.4. I created an IMPT treatment plan using three equidistant, coplanar, 3D modulated treatment fields including an anterior-posterior beam. This treatment plan had more than 7000 beamlets with energies ranging from 70 MeV to 159 MeV. In figure 4.4, two of these beamlets taken from the anterior-posterior beam are shown, one with a very low H_i-value of 2.26 mm and one with a large H_i-value of 24.07 mm. Indicated is the WED in the x-y plane minus the WED of the central ray on this plane, thus the remaining water equivalent range. For the homogeneous beamlet this is close to zero within the beamlet size, whereas there are differences up to 7 cm water equivalent range within 7.7 mm lateral distance for the beamlet with $H_i = 24.07\ mm$. Obviously, the assumption of a slab geometry completely fails in this situation. However, this assumption is crucial for the pencil beam algorithm. Consequently, doses derived with the pencil beam algorithm deviate substantially from the Monte Carlo calculations in the inhomogeneous case. However, the

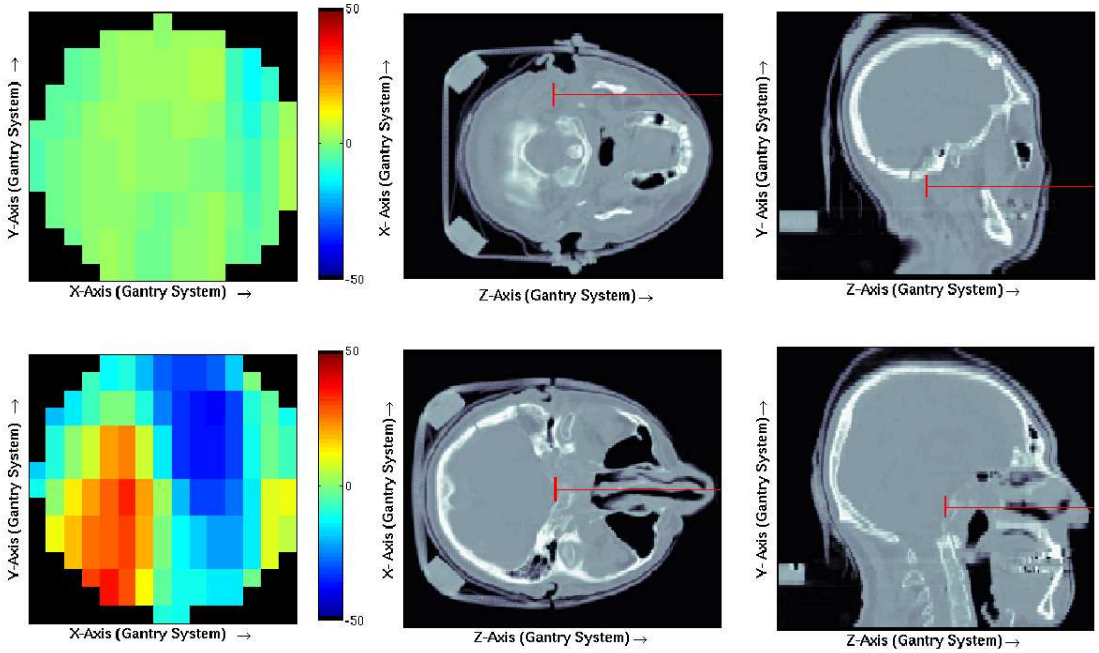

Figure 4.4: *First row: beamlet with $H_i = 2.26$ mm; Second row: beamlet with $H_i = 24.07$ mm. Plotted is the remaining water equivalent range in mm on the x-y plane and the CT slices containing the central ray of the beam spot. The radius of the disc is 3σ, σ being the lateral spread of the beam when entering the patient. The disc and the central ray of the beamlet are shown in red in the CT-slices.*

.

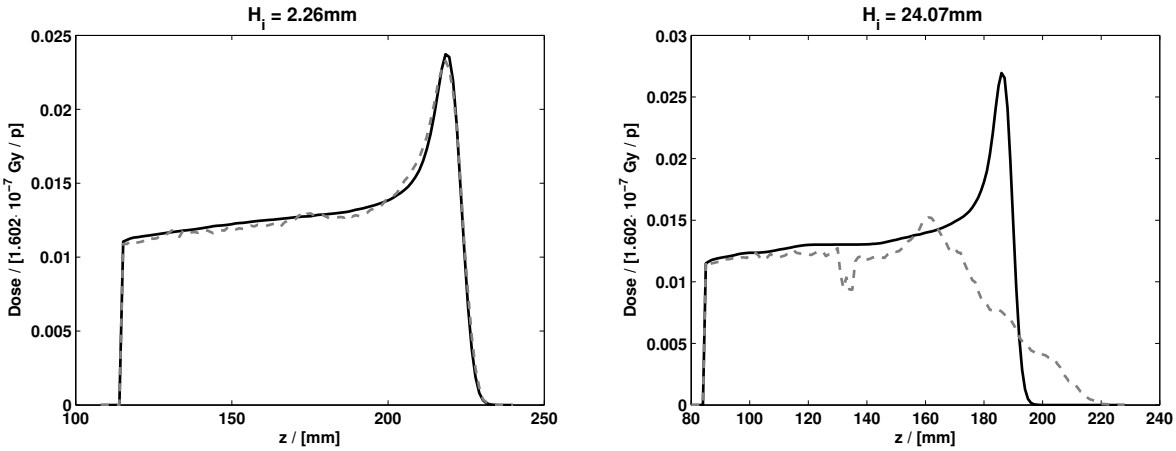

Figure 4.5: *Dose along the central beamlet axis for the two examples shown in figure 4.4. Solid line: pencil beam algorithm; dashed line: Monte Carlo.*

result of the two algorithms match nicely for the beamlet with $H_i = 2.26\ mm$ as indicated in figure 4.5.

4.3.2 Dose calculation error

For all beamlets used in the plan described in the previous paragraph I quantified the error between the pencil beam and the Monte Carlo algorithm as described in section 4.2.3. The error plotted against the heterogeneity number is shown in figure 4.6. The dependence of the dose calculation error on the heterogeneity number is clearly visible.
Since there are only a few beamlets with large lateral tissue heterogeneities there is not much data available in this region. Thus the statistics is only meaningful in the area of moderate values of H_i.

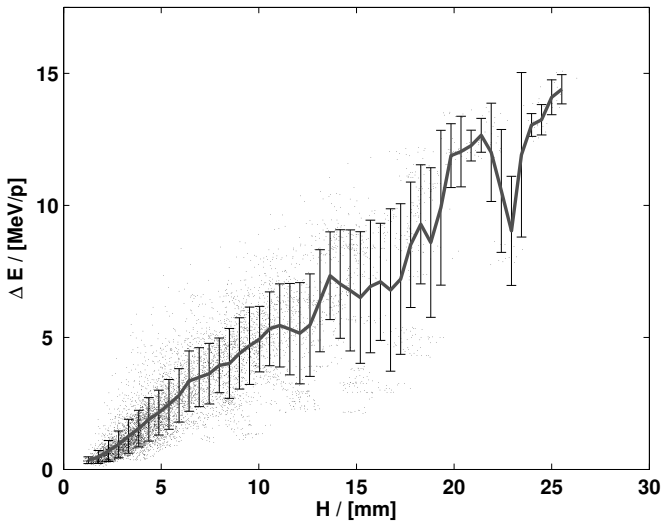

Figure 4.6: *Dose calculation error in MeV per incident proton depending on the heterogeneity number for more than 7000 beamlets. The straight line shows the mean, the error bars one standard deviation of the dose calculation error.*

4.3.3 Sensitivity to setup errors

A second patient with a brain tumor was chosen to investigate the sensitivity to setup errors using the method described in section 4.2.4. A CT slice of this patient is shown in figure 4.7. I created a treatment plan using three coplanar beams, an anterior-posterior, a left lateral and a right lateral beam. The beamlets at the distal edge of the tumor were taken for this analysis, resulting in over 600 beamlets with energies ranging from 80 MeV

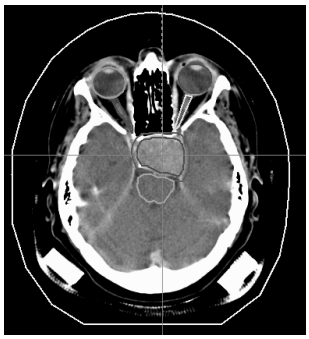

Figure 4.7: *A transversal CT slice of the patient used for the analysis in section 4.3.3 and 4.3.4.*

to 147 MeV.
Again, the calculated dose distributions are converted to energy per voxel thus the resulting error has units of MeV per incident proton. Its dependency on H_i is shown in figure 4.8. The increasing sensitivity to setup errors with increasing lateral tissue heterogeneity is evident. As stated in the last paragraph, this data is also only statistically meaningful in the range of moderate values H_i.

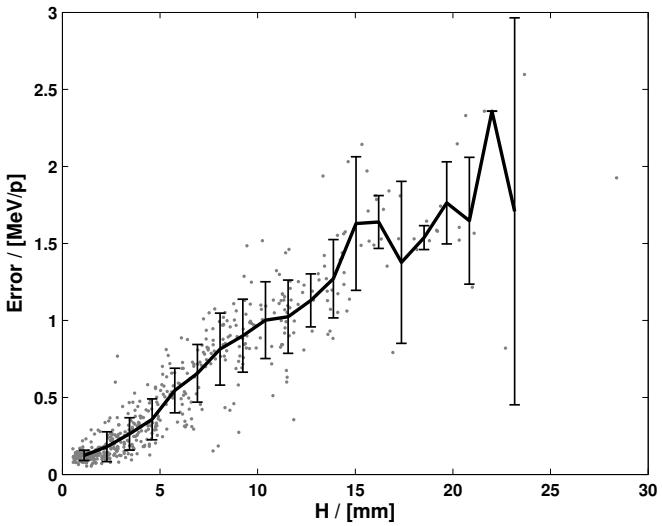

Figure 4.8: *Sensitivity to setup errors depending on H_i. Again, the straight line shows the mean, the error-bars one standard deviation of the error.*

4.3.4 Including H_i into the optimization

The new objective function (eq. 4.5) was applied to optimize a treatment plan for the brain tumor patient described in the previous paragraph. The treatment plan consists of

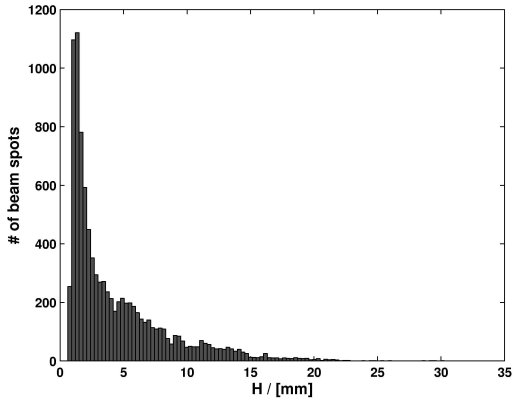

Figure 4.9: *Histogram of H_i values for the IMPT plan described in section 4.3.4.*

five coplanar, equidistant, 3D modulated fields, including an anterior-posterior beam. The energies ranged from 70 MeV to 149 MeV. The distribution of the H_i values occurring in this treatment plan is given in figure 4.9. The pencil beam algorithm was used as dose engine for the optimization. The resulting plan was then recalculated with the Monte Carlo algorithm. The results are given in figure 4.10. For the standard optimization ($p_0 = 0$), the discrepancy between the pencil beam algorithm and Monte Carlo calculation can be clearly seen in the shown example (figure 4.10a). By increasing the weight (p_0) of the additional term in the objective function and thus increasingly suppressing beam spots with a large H_i from the optimization this discrepancy can be reduced considerably (figure 4.10b). This also leads to a reduced total lateral heterogeneity (see eq. 4.2) of the treatment plan which drops from 7.5 mm ($p_0 = 0$) to 3.2 mm ($p_0 = 50$). If p_0 is selected too large, however, the quality of the resulting treatment plan will be reduced. The dosimetric data in the target for the cases shown in figure 4.10 is summarized in table 4.2. Although for the minimum dose there is no difference between the standard and the $p_0 = 50$ optimization, a big improvement can be seen for the D_{99} value. This indicates that lateral tissue heterogeneities are not the only source for dose calculation errors of the pencil beam algorithm. However, it also shows that the sensitivity to errors introduced by the use of the fast pencil beam dose calculation algorithm can be reduced by suppressing beam spots with a large H_i in the optimization.

4.3.5 Evaluation of the treatment plan in presence of setup errors

The analysis shown in the previous section only addresses the problem of the dose calculation error. This effect can be reduced if sufficient computation power is available to use an improved dose calculation algorithm during the optimization. From the data presented

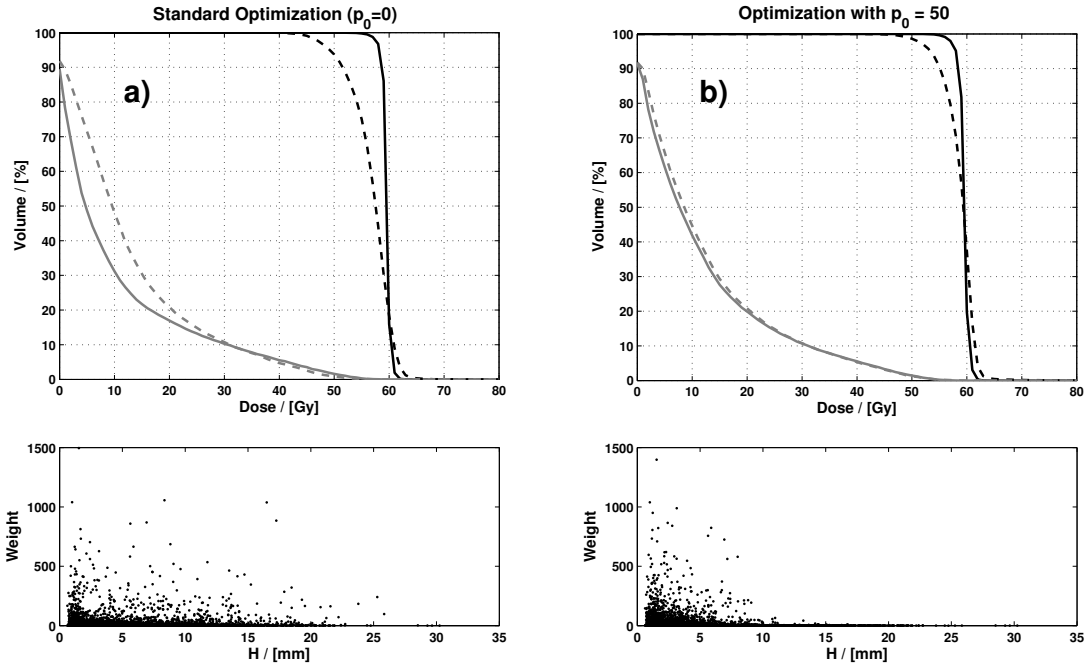

Figure 4.10: *DVH for the IMPT plan described in section 4.3.4. Black: planning target volume; gray: brainstem. Pencil beam calculations are plotted as solid, recalculations with Monte Carlo as dashed lines. In the second row the beamlet weights are plotted against the H_i value. Without the additional term in Eq. (4.5) (case a) also beam spots with a high H_i are used leading to a considerable difference between the pencil beam prediction and the recalculation. This difference is reduced by suppressing beamlets with large H_i (case b). The total heterogeneity of the treatment plan is 7.5 mm for the standard optimization and 3.2 mm for $p_0 = 50$.*

	D_{min}	D_{99}	D_{max}	D_{mean}	σ_{Dose}
$Pb_{Standard}$	50.4	57.3	63.2	60.0	0.74
$MC_{Standard}$	38.2	44.7	72.6	57.3	3.81
Pb_{p50}	50.4	56.7	63.9	60.0	0.90
MC_{p50}	38.4	49.8	79.8	59.3	2.84

Table 4.2: *Dose statistics in the planning target volume for the four examples shown in figure 4.10. All values are given in Gy. D_{99} means that 99% of the target voxels receive a dose higher than that value.*

in figure 4.8, however, an improved robustness of the treatment plan to setup errors is also expected when the beamlets with large H_i values are suppressed during the optimization. To investigate this, the treatment plan shown in the previous section was utilized. To exclude the effects described in the last section, the Monte Carlo algorithm was used here as the dose calculation engine during the optimization. The dose calculation for a single beamlet took roughly $1 - 2\ h$, while the treatment plan employed more than 9000 beamlets. Although a computer cluster was available the calculation of the Dij-matrix took several weeks. Again the optimization was done for two values of the additional penalty, $p_0 = 0$ and $p_0 = 50$. To assess the effects of setup errors, the resulting treatment plans were then recalculated with a shifted target point. The recalculation was done for 81 sampling locations of the target point, while all 81 shifted target point locations where inside a sphere with a radius of 3 mm around the original target point location. The recalculation was again done using the Monte Carlo algorithm. Similar to the method described in section 4.2.4, the primary protons were generated uniformly over the lateral coordinates of the treatment beams, thus all 81 dose distributions could be calculated in one simulation. Using this method, the dose distribution for an arbitrary number of target point locations inside the sphere can be calculated without increasing the calculation time. The restriction to 81 target point locations is not posed by the computation time but by the computer memory since all considered dose distributions have to be stored simultaneously. The DVHs of the brain stem and the CTV of all 81 dose distributions are plotted in figure 4.11 for the optimization using $p_0 = 0$ and $p_0 = 50$. The DVHs are color-coded with the distance

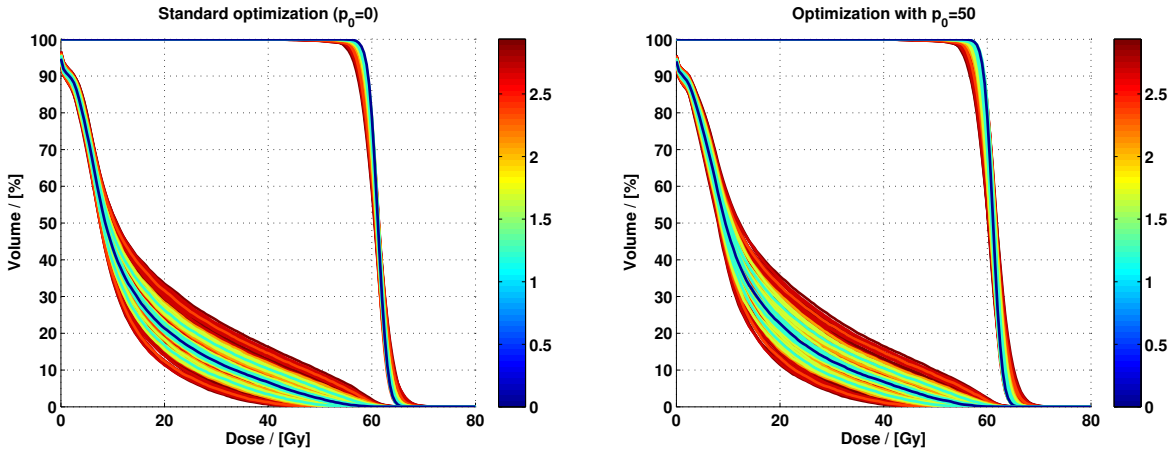

Figure 4.11: *The DVHs for the CTV and the brain stem for 81 shifted locations of the target point. The distance to the original target point location in mm is color-coded. The DVHs are plotted in order of the distance to the original target point location, from the furthest to the nearest.*

between the shifted and the original target location, while the DVHs are plotted in order from the furthest to the nearest. Thus this figure directly shows the possible variations of the DVH for a given setup error. While the optimization was done on the PTV, the DVH of the CTV is used in this figure since the safety margin between these two volumes is introduced to prevent an underdosage due to a shifted dose distribution. By respecting these margins in the evaluation of the treatment plans the remaining differences are now solely due to changes of the shape of the dose distribution. Contrary to the expectation, penalizing beamlets with a large H_i did not improve the sensitivity of the treatment plan to setup errors. As shown in figure 4.11, the dose to the CTV is hardly degraded due to the considered setup errors for $p_0 = 0$ thus large improvements are not possible. The degradation is even slightly increased if beamlets with large H_i are penalized in the optimization. The main reason for this is that reducing the weight to a single beamlet in a treatment beam will leave a "hole" in the dose delivered from this beam which has to be filled up from another direction. Obviously, such a configuration is sensitive to shifts of the target point. For the sensitivity of a treatment plan to setup errors, this effect canceled the benefit from excluding beamlets with a large lateral heterogeneity in this case.

4.4 Discussion

The heterogeneity number H_i has been introduced to quantify lateral tissue heterogeneities of a single beamlet i. From the data presented I conclude that H_i can be used as a tool to assess the risks of lateral tissue heterogeneities. H_i is a statistical quantity, in the sense that for two individual beamlets it is not guaranteed that the one with the lower H_i value will be the better one according to the error measures introduced in this chapter. Given many beamlets, however, the ensemble with a low H_i value will on average perform better than the ensemble with a high H_i value. While the calculation of the error measures for all beamlets of the shown treatment plan took several weeks on a computer cluster, the calculation of the heterogeneity numbers can be done in a few seconds on a standard PC. It is thus technically feasible to use the information of H_i during routine treatment planning. One option to use the information of H_i in the treatment planning process is to exclude beamlets with a H_i value above a cut value H_{cut}. However better results were obtained be applying a penalty to beamlets with large lateral heterogeneities in the optimization as shown in section 4.3.4. This approach considerably reduced the dose calculation error introduced by the use of the fast pencil beam algorithm. However there are several problems using these two methods. These problems and possible solutions are outlined in the following:

- The determination of the weight p_0 of the additional term in the objective function is not straightforward. The value of the additional term compared to the other terms in the objective function can be used as a hint to choose p_0. A more elegant way would be to include this parameter into the framework of a multicriterial optimization (see e.g. [48]). The prioritized prescription goals optimization recently proposed by Wilkens et al. [49] offers another interesting approach to include H_i into the optimization.

- Due to the normalization $1/\left(\sum_{i=1}^{N} w_i\right)$ the additional term in the objective function is not convex. In a simple two beamlet model, local minima have been observed. This poses a challenge to the optimization algorithm. Without this normalization, the additional term in the objective function not only penalizes the use of beamlets with a large H_i, it also aims to reduce the sum of all beamlet weights. Assuming that all beamlets result in approximately the same integral dose per unit weight this results in a penalty to the total integral dose. This effect is usually desirable, thus the normalization could also be removed, resulting again in a convex objective function with only one minimum. The normalization was included in this thesis since the goal was to investigate only the effects resulting from the use of beamlets with large H_i values.

- As described in section 4.3.5, reducing the weight of single beamlets in a treatment beam results in an inhomogeneous dose distribution from this beam. This is enhanced if single beamlets are excluded form the optimization using a cunt value H_{cut}. The "holes" generated in the dose distribution have to be filled up by a treatment beam from a different angle. However, such configurations are sensitive to setup errors thus this method can increase the sensitivity of a treatment plan to setup errors. Optimization strategies which account for setup uncertainties could be used to avoid this problem. Such a method is introduced in the following chapter.

An additional benefit of the suppression of beamlets with large H_i is a beamlet reduction. In the standard optimization, 15.5% of the beamlets had a weight below 0.001% of the maximal beamlet weight. Optimizing with the additional term F_H in the objective function, 22.3% of the beamlet weights were below this threshold. Since such small beamlet weights are usually set to zero this is expected to result in an improved delivery time for the treatment plan.
However the are also other ways to utilize the information of H_i. Obviously, the best way to avoid risks due to lateral heterogeneities is to choose a beam direction without large

lateral heterogeneities. A histogram of all H_i values for a given beam direction summarizes the heterogeneity information for this direction of incidence. This information could be used as an objective measure to decide on a beam direction in a manual or automatic beam selection tool. Selecting suitable beam directions can have a large impact on IMPT treatment planning since in IMPT only a few beam directions (typically two to maybe five) are utilized.

H_i could also be used as a criterion whether a beamlet can be calculated with the fast pencil beam algorithm or if it has to be calculated using a more sophisticated, but also more time consuming algorithm, like a subsampling of the finite pencil beam [29] or a Monte Carlo code. Using a subsampling of the finite pencil beam, the number of subsamples is a trade-off between accuracy and computation time. As H_i is correlated to the error of the fast pencil beam algorithm, H_i might be used to determine the needed accuracy and thus the number of subsamples. Since usually there are only a few beam spots with large H_i values (see figure 4.9) and since for low H_i values the fast pencil beam algorithm is a reasonable approximation (see figure 4.6), a fast and accurate dose calculation algorithm could be designed using this hybrid approach.

Chapter 5

Worst case optimization

5.1 Introduction

Proton treatment plans can be very sensitive to small uncertainties in the treatment variables such as uncertainties in the range of the individual Bragg peaks or errors in the setup of the patient. Due to the steep distal gradient of the Bragg peak a range uncertainty of a few mm can lead to an underdoseage in the radiation target by 100%, depending on the assigned safety margins. However, the concept of safety margins partly fails for setup errors since shifts of the patient not only shift the dose distribution they can also change the range of the Bragg peaks leading to a distorted dose distribution. As intensity modulated proton therapy is more complex than standard proton treatments, this sensitivity is often expected to increase for IMPT. However, as stated in section 2.2.3 there usually exist many different solutions to the inverse problem in IMPT, each represented by a set of Bragg peak weights that will deliver dosimetrically equivalent plans. This degeneracy of the solutions can be used to reduce the sensitivity of treatment plans if the uncertainties are accounted for in the optimization. In this chapter such a method to account for uncertainties in the optimization is developed. This worst case optimization method makes use of a worst case dose distribution as introduced by Lomax et al. [22]. The method has been implemented into the research version of KonRad to account for range uncertainties, setup errors and a combination of these two uncertainties. Finally an example is given how this method could be used in 4D treatment planning, that is in treatment planning which takes into account the motion of the patient. Unkelbach et al. [50] recently published two methods to account for range uncertainties in IMPT, one method using a probabilistic approach, the other applying methods from robust linear programming. Differences between the methods proposed by Unkelbach et al. and the worst case optimization will be outlined at the end

of this chapter (section 5.3.6).

5.2 Methods

Four different methods to realize intensity modulated proton therapy are described in section 2.2.3. To exploit the full degeneracy of the inverse treatment planning problem the most flexible one, the full three-dimensional modulation technique, is used in this chapter. As shown in figure 2.5 this technique places individual Bragg peaks on lines from the distal to the proximal edge of the target volume such that the whole target volume is covered with Bragg peaks. I refer to an individual Bragg peak as a beamlet and to all beamlets on such a line as a ray. In this chapter the optimization is done using the L-BFGS-B code and the variable transformation to turn the constrained optimization problem into an unconstrained one (see chapter 3).

5.2.1 Worst case dose distribution

The worst case dose distribution introduced by Lomax et al. [22] is a method to combine multiple dose distributions (e.g. dose distributions calculated for different ranges of the Bragg peaks) into a single one. For a voxel inside the target volume, the minimum dose of this voxel in all dose distributions is stored in the worst case dose distribution. For a voxel outside the target volume, the maximum is taken. The worst case dose distribution is unphysical since it treats every voxel independently. This means that while the worst case dose in a voxel i is realized under one of the considered conditions, it might not be true that the worst case dose in a voxel j can be simultaneously realized. Although being unphysical this worst case dose distribution can serve as a lower bound for the worst quality of the treatment plan.
Similar to the worst case dose distribution a best case dose distribution can be defined where inside the target volume the maximum and outside the target the minimum is taken. This best case dose distribution is not used during the worst case optimization. However, it can be used for the evaluation of the resulting treatment plan. While the worst case dose distribution is a lower bound, the best case distribution formally serves as upper bound for the achievable plan quality. When plotting the DVHs of the worst and the best case dose distribution, the DVHs of all dose distributions which were considered to calculate the worst/best case dose distribution are in between these two DVHs.

5.2.2 Uncertainties

5.2.2.1 Range uncertainties

There are multiple sources of range uncertainties in proton therapy, e.g. CT artefacts, weight-gain or -loss of a patient, conversion from Hounsfield units (HU) to stopping powers, to name a few. Schaffner et al. [33] found that solely the conversion from HU to stopping powers leads to "a range precision of about $1-3\ mm$ in typical treatment situations." However, the range uncertainties of all individual Bragg peaks are not uncorrelated. In the 3D technique, multiple beamlets are placed in rays parallel to the beam direction from the distal edge to the proximal edge of the tumor. Since all beamlets in such a ray traverse the same path in the CT these beamlets are treated as correlated, i.e. I do not consider the case where one beamlet of the ray has an increased range while an other beamlet in the ray has a decreased range. This approximation assumes that the range uncertainty is accumulated up to the target volume. Effects inside the target volume that could reduce this correlation are not considered.
To include the range uncertainties into the optimization, the dose of each beamlet has to be calculated at multiple ranges. For small range uncertainties of around $2-5\ mm$ it turned out that three range samples were sufficient, namely the nominal range, the maximal range and the minimal range. For larger range uncertainties, more samples were needed.

5.2.2.2 Setup errors

Setup errors lead to a shift of the target point relative to the patient in three dimensions. For each beam I only consider shifts perpendicular to the beam axis since these shift have the largest impact on the dose distribution. This results in a shift in only two dimensions, thus less samples are needed to include the setup errors into the optimization. For the optimization, the dose is calculated for 5 positions, namely the nominal position and a shift of $\pm$ the maximal setup error for each direction perpendicular to the beam axis.
While for the range uncertainties the rays of a beam are considered uncorrelated, the rays are certainly correlated for the setup error in a static patient geometry. However, I treat the setup error for each beam as being uncorrelated to the other beams. This is more conservative as it needs to be but this approach offers the possibility to use the projection of the setup error onto the plane perpendicular to each beam and thus allows to sample a two dimensional distribution instead of a three dimensional one. Thus the setup uncertainty can be included using only four additional sampling positions.

5.2.3 Worst case optimization

In addition to the nominal dose distribution, D_{nom}, the worst case optimization uses a second dose distribution, the worst case dose distribution D_w. D_w is included into the optimization simply by applying the used objective function F to both dose distributions:

$$\tilde{F}(\vec{w}) = F(D_{nom}(\vec{w})) + p_w \cdot F(D_w(\vec{w})) \tag{5.1}$$

Here $\vec{w}$ is the beamlet weight vector and the worst case penalty p_w is a newly introduced parameter to tune the importance of the worst case dose distribution. In particular, by setting $p_w = 0$ the conventional IMPT plan is recovered.
The worst case dose distribution D_w is calculated the following way:
For range uncertainties, the dose of each ray is calculated for every range sample, e.g. for the minimal, nominal and maximal range. From these multiple dose distributions a single worst case dose distribution is determined as described in section 5.2.1. This calculation is done ray by ray. The resulting worst case ray dose distributions are accumulated in D_w.
A similar approach is used to calculate D_w for setup errors, except that now the rays are no longer considered independent and thus the calculation described for range uncertainties is performed for each beam instead of each ray.
To calculate D_w for range uncertainties combined with setup errors I applied two simplifications. For each beam, the laterally shifted dose distributions are only calculated for the nominal beamlet range. The range uncertainties are only calculated for the unshifted beam. Furthermore I treated for simplicity the range uncertainties of all rays in a beam as correlated. With these simplifications there are seven dose distributions per treatment beam to calculated the worst case dose distribution, which are accumulated for each beam in D_w.

5.2.3.1 Worst case optimization in 4D treatment planning

Patient motion such as breathing can cause range uncertainties. If CT data for multiple phases of e.g. the breathing cycle is available such 4D-CT data can be used to estimate the range uncertainties introduced due to this patient motion. These estimates of the range uncertainties can then be accounted for in the optimization by the same methods described in section 5.2.3. A common approach in 4D treatment planning is to generate a volume which encloses the tumor in every phase of the 4D-CT, the internal target volume (ITV) (see ICRU62 [51]). Usually one phase of the 4D-CT is chosen as the reference phase. Conventional treatment planning can then be performed for the ITV on this reference

phase, including the calculation of the water equivalent range of the distal edge of the target volume for every ray of each beam. With the information of the 4D-CT, these water equivalent ranges can also be calculated for every other breathing phase available. For 4D planning different uncertainties for the maximal and the minimal range are assumed, thus instead of the uncertainty interval $[r-u, r+u]$ I consider $r \in [r-l, r+u]$. Furthermore these uncertainties are different for each ray of a treatment beam. Using the information from the 4D-CT the upper and lower range uncertainty of each ray are set to

$$\begin{aligned} u &= \max\left[\max(WER_{DE_i} - WER_{DE_{ref}}),\ \max(WER_{PE_i} - WER_{PE_{ref}})\right] \\ l &= \max\left[\max(WER_{DE_{ref}} - WER_{DE_i}),\ \max(WER_{PE_{ref}} - WER_{PE_i})\right] \end{aligned}$$

WER_{DE_i}/WER_{PE_i} is the water equivalent range of the distal/proximal edge of the ITV for breathing phase i. Using a static patient this definition yields $u = l = 0$, thus it recovers the standard IMPT.

The resulting range uncertainties are then included into the optimization by the same method described in section 5.2.2.1. If these uncertainties for each ray are accounted for during the optimization, it is possible to generate a treatment plan that ensures a target coverage independent of the actual breathing phase in which the treatment plan is delivered. However since it is assumed that the range uncertainty for all beamlets on a ray are correlated each ray would have to be applied within a single breathing phase. Furthermore interplay effects between the scanned beam and the tumor motion are not considered by this approach. This greatly restricts the possible use of this method.

5.2.4 Patient data

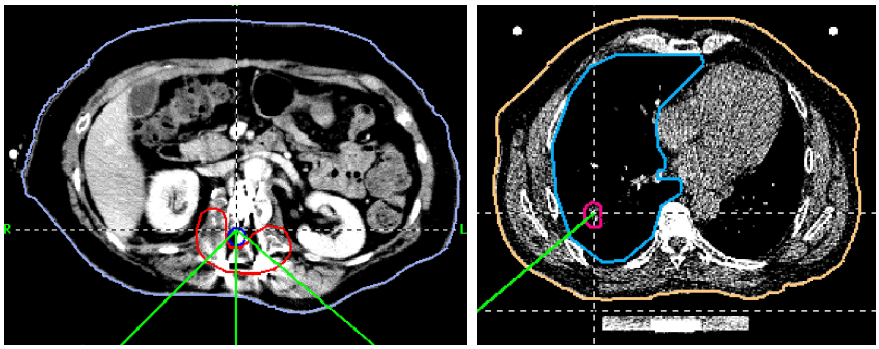

Figure 5.1: *Transversal CT slices for Patient A (left) and Patient B (right). Coplanar fields are used in this study. The direction of incidence are shown as green lines.*

Two patient data sets are used to demonstrate the potential of the worst case optimization. Patient A is a typical IMRT case for which the target volume is wrapped around the spinal cord. For simplicity the spinal cord was chosen as the only organ at risk in this study. This patient had a surgery prior to radiotherapy thus metal implants were present to stabilize the spine. It was chosen because of the similar location and shape of the target volume compared to the first clinical IMPT case reported [52] and because of the similarity to the target used by Unkelbach et al. [50]. Equivalently to these cases three coplanar fields were used as shown in figure 5.1.
As patient B a lung tumor patient was considered for whom 4D-CT data were available. CT data from six different breathing phases could be utilized for this study. To apply the method presented in section 5.2.3.1 an internal target volume was created for this patient. The ITV had a volume of 8.1 cm^3. The tumor was mainly moving in the cranio-caudal (cc) direction. The peak-to-peak difference in the cc position of the tumor was 9 mm. For this study a single field was applied as shown in figure 5.1.

5.3 Results

5.3.1 Range uncertainties

The data of patient A was used to test the worst case optimization method to include range uncertainties. Range uncertainties of ± 5 mm were considered. The range uncertainty was sampled at three positions, namely at the nominal range, the maximal range and the minimal range. The optimization was done for four values of the worst case penalty, $p_w = 0$ (conventional IMPT), $p_w = 0.05$, $p_w = 0.2$ and $p_w = 1$. The resulting dose volume histograms are shown in figure 5.2. The sensitivity to range uncertainties of the conventional IMPT treatment plan is clearly visible in figure 5.2, both as a strong underdosage in the target as well as a strong overdosage in the spinal cord for a beamlet range error of 5 mm. Another characteristic visible is that the DVH of the worst case and the best case dose distributions together serve as upper and lower bounds in which the true DVH is located for any realization of the beamlet ranges within the considered uncertainty levels. By increasing p_w these bounds are brought closer together, thus reducing the variation of DVHs from dose distributions for different realized beamlet ranges. The nominal dose distribution, where the planned beamlet range is realized for every beamlet, is shown in figure 5.3 for each beam for $p_w = 0$ and $p_w = 1$. For $p_w = 0$ (conventional IMPT) the excellent separation of the DVH of the target and the spinal cord, shown in figure 5.2, is achieved by shaping the dose gradient between the target and the organ at risk mainly with the distal

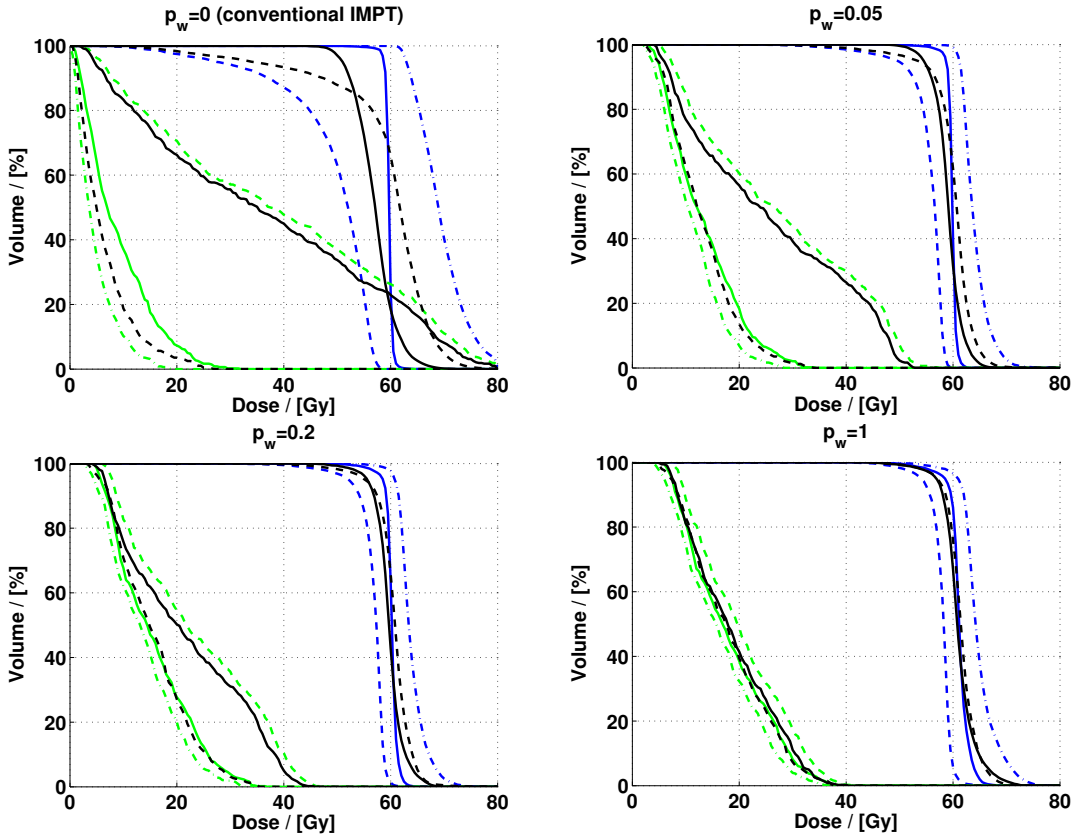

Figure 5.2: *The DVH for patient A obtained with four values of the worst case penalty p_w. Blue: CTV, green: spinal cord. The colored solid lines represent the case where the delivered range of every beamlet equals the planned range. The dashed/chain-dotted line represents the DVH of the unphysical worst/best case dose distribution. The black lines show the DVH for the two extreme cases where the delivered ranges of all beamlets are increased (solid lines) or decreased (dashed lines) by 5 mm compared to the planned range.*

gradient of the Bragg peaks. In presence of range uncertainties these promising features of the DVHs can deteriorate substantially. The worst case optimization results in a slightly deteriorated plan quality, however the obtained dose distributions are much less sensitive to range uncertainties. Three effects lead to this reduced sensitivity to range uncertainties as it is shown in figure 5.3:

(i) the dose gradients between the target and the organ at risk is shaped using the lateral instead of the distal gradient of the Bragg peaks

(ii) a "safety margin" is created automatically at the distal field edge for each treatment beam

(iii) the dose profile in depth for each treatment beam is flattened compared to the nominal plan

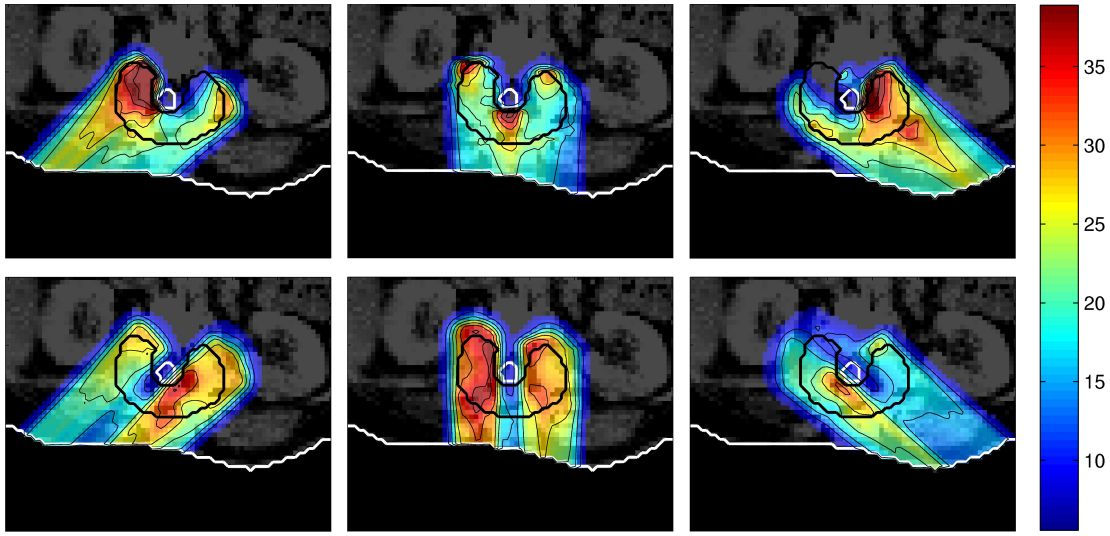

Figure 5.3: *Dose distributions in Gy for the three individual beams for $p_w = 0$ (upper row) and $p_w = 1$ (lower row). Range uncertainties of 5 mm are accounted for.*

Recalculation of the treatment plan:

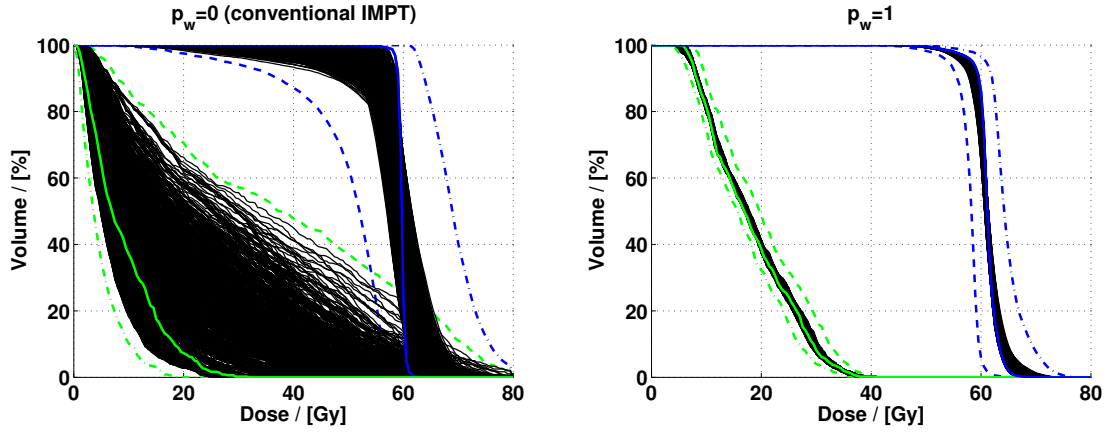

Figure 5.4: *DVH of dose distributions recalculated for 1331 different realizations of the range uncertainties (black lines). Colored solid lines: DVH for nominal dose distribution. Colored dashed/chain-dotted lines: worst/best case dose distribution.*

To show that the DVHs of the worst/best case dose distributions act as bounds for any dose distribution calculated within the considered uncertainties, the resulting treatment plan was recalculated for multiple realizations of the beamlet ranges. The dose for each treatment beam has been recalculated for 11 different realization of the ranges equally spaced between the minimal and the maximal range using the beamlet weights resulting from the optimization. Unlike in the optimization the rays in each beam are correlated in this recalculation because the use of uncorrelated rays leads to more possible dose

distributions that can be practically calculated. With three beams each having 11 different dose distributions there are $11^3 = 1331$ possibilities to combine them to the total dose. The DVHs of these 1331 dose distributions are plotted in figure 5.4. For the organ at risk they are spread almost over the entire region between the DVH of the best and the worst case dose distribution. Figure 5.4 also shows very nicely the reduced sensitivity to range uncertainties for the plan optimized with $p_w = 1$, where the DVHs for all 1331 different realizations of the beamlet ranges are very close to each other.

Comparison to non-IMPT proton therapy

IMPT treatment plans might be expected to be more sensitive to uncertainties compared to conventional proton therapy plans, where a homogeneous dose distribution in the target is applied from each direction of incidence. To investigate this, the data shown in figure 5.2 is compared to such a treatment plan. A typical conventional, non-IMPT plan for patient A would consist of a single treatment field from posterior to anterior (second beam in figure 5.3). The DVH of this conventional treatment plan is shown in figure 5.5. Again

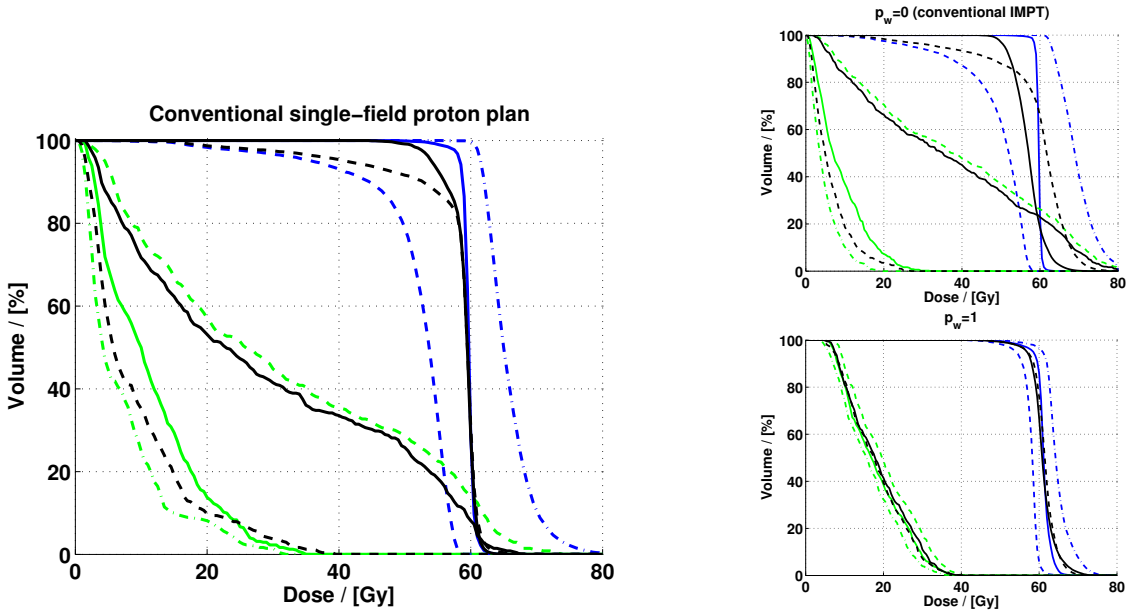

Figure 5.5: *The DVH of a conventional, non-IMPT proton plan for patient A (left). The DVH for $p_w = 0$ and $p_w = 1$ from figure 5.2 are included for comparison.*

the DVH for an increase/decrease of the beamlet ranges of 5 *mm* as well as the DVH for the worst/best case dose distributions for this uncertainty are included in the figure. In this example, the conventional non-IMPT treatment plan is indeed less sensitive to range uncertainties than the conventional IMPT plan. However, the single-field proton plan is

still very sensitive to range uncertainties. With the help of the worst case optimization, an IMPT plan can be generated which is less sensitive to range uncertainties compared to the conventional non-IMPT approach.

5.3.2 Setup errors

To account for setup errors in the optimization I applied the method described in section 5.2.2.2 to patient A. A setup error of 2 mm was assumed. Range uncertainties are not taken into account in this section. Figure 5.6 shows the resulting DVHs for two values of the worst case penalty, $p_w = 0$ (conventional IMPT) and $p_w = 1$. This figure also includes a recalculation of the optimization results on a finer sampling grid. The dose distribution of each beam was calculated for a shift of ± 1 mm and ± 2 mm in each direction perpendicular to the beam direction using the beamlet weights resulting from the optimization. Including the unshifted position this results in nine dose distributions per beam, leading to $9^3 = 729$ possible combinations to the total dose distribution. The DVHs of these 729 dose distributions are plotted as black lines in figure 5.6. Although less pronounced as

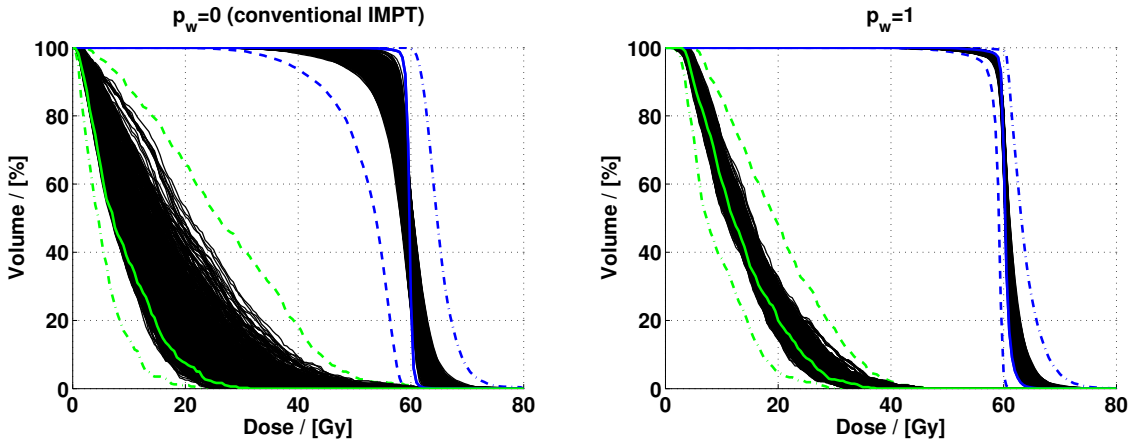

Figure 5.6: *DVH of dose distributions recalculated for 729 different shifted beam positions (black lines). Colored solid lines: DVH for nominal dose distribution. Colored dashed/chain-dotted lines: worst/best case dose distribution.*

for range uncertainties figure 5.6 shows that setup errors can still lead to a considerably deterioration of the anticipated dose distribution. The sensitivity to these errors can be decreased by increasing the worst case penalty p_w. The dose distributions of the individual beams are shown in figure 5.7. Beam doses with a reduced sensitivity to setup errors again show three characteristics:

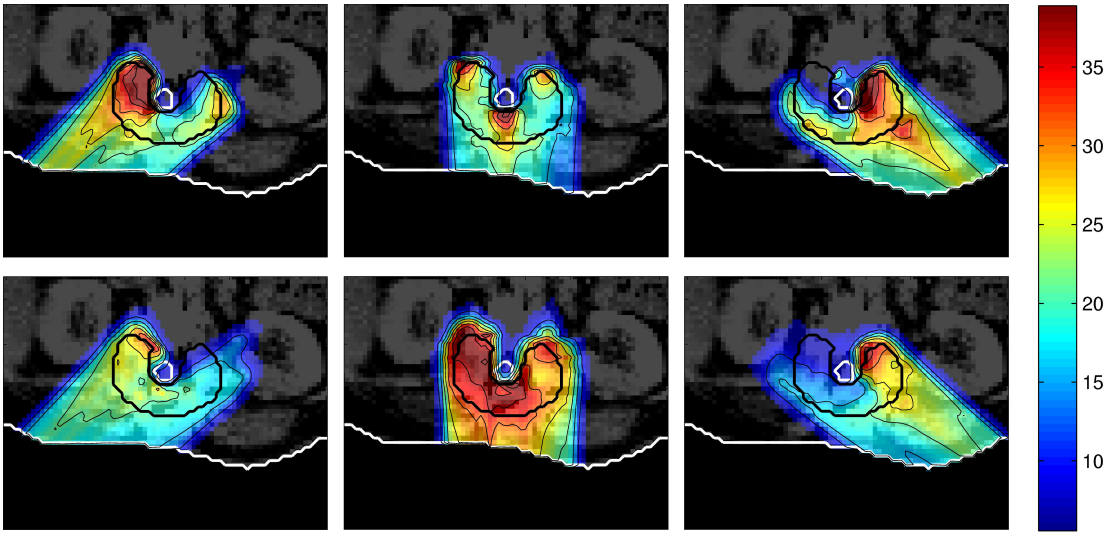

Figure 5.7: *Dose distributions in Gy for the three individual beams for $p_w = 0$ (upper row) and $p_w = 1$ (lower row). A 2 mm setup error was assumed, range uncertainties are not included.*

(i) the beam dose is reduced for parts of the beam which hit the patient at a narrow angle, as there a small shift leads to a large change in radiological depth

(ii) a "safety margin" is created automatically at the lateral field edges for each treatment beam

(iii) the lateral dose profile is flattened compared to the nominal plan

The last feature is in direct conflict with the results from the optimization including only range uncertainties, where the lateral dose gradients were explicitly used to shape the dose gradient between the target and the organ at risk. The next section will investigate the case where both uncertainties are included into the optimization. When setup errors are included into the optimization, the dose distribution of the first beam shows an extension at the distal end of the target as can be seen in figure 5.7, lower row, left picture. Switching the CT window to metal (CT window: $mean = 1500$ HU, $width = 1$ HU) as shown in figure 5.8 reveals the reason for this "dose finger". This part of the beam is passing next to a metal implant. By shifting the beam 2 *mm* to the right perpendicular to the beam direction this part transverses the metal implant which reduces the range of the dose finger towards the distal edge of the target volume. Thus range uncertainties which arise due to setup errors are automatically accounted for if setup errors are included into the optimization.

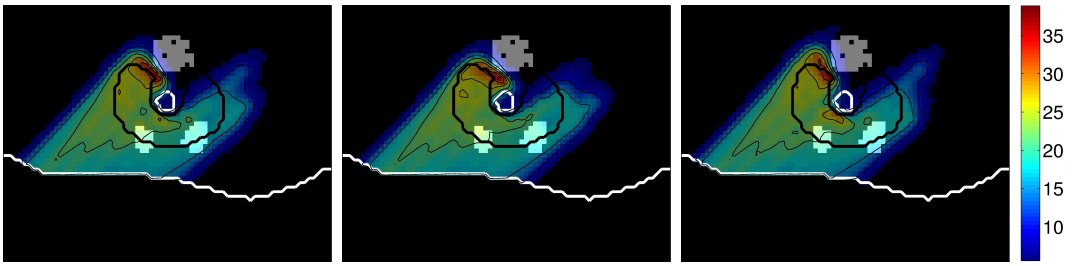

Figure 5.8: *Dose distribution of the first beam in Gy using $p_w = 1$. The applied CT window shows the metal implants only. First: nominal position of the treatment beam, second/third: beam shifted by 2 mm perpendicular to the beam direction in the image plane to the right/left.*

5.3.3 Range uncertainties and setup errors

If range uncertainties are accounted for in IMPT, the optimization shapes the dose gradient between the target and the organ at risk using preferably the lateral fall-off of the beamlets. However, the deteriorating effects of the setup errors are alleviated by softening the lateral dose gradients. If both uncertainties are accounted for simultaneously in the optimization, a compromise between these two goals has to be reached. To investigate this,

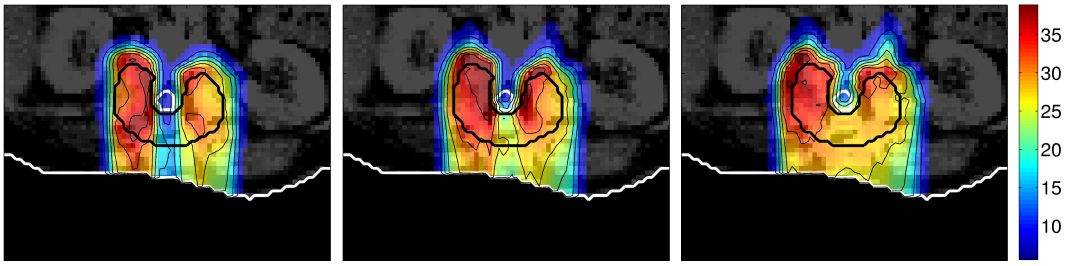

Figure 5.9: *Dose distribution of the second beam in Gy. $p_w = 1$, range uncertainty 5 mm, setup error (from left to right) 0 mm, 2 mm, 5 mm.*

both uncertainties were included into the optimization as described in section 5.2.3. As in the previous section a range uncertainty of 5 *mm* was assumed. This uncertainty was combined with a setup error of 0 *mm*, 2 *mm* and 5 *mm*. A worst case penalty of $p_w = 1$ was employed for all three cases. Exemplarily the dose distribution of the second beam for these cases is shown in figure 5.9. Using the range uncertainty only, the optimization avoids beamlets stopping in front of the spinal cord, leading to a valley in the lateral dose profile. Increasing the value for the setup error this valley is smeared out to flatten the lateral dose profile. When applying a setup error of 5 *mm*, the beamlets stopping in front of the organ at risk are again used to shape the dose gradient. However it is interesting to see that this dose gradient is not as steep as it can be using a single Bragg peak. The

beamlet weights are chosen such that the distal gradient is smeared out. The contour lines show that this gradient is comparable to the lateral dose gradient.

5.3.4 4D treatment planning with internal target volumes

To achieve a treatment plan which is applicable for every breathing phase of a 4D-CT I applied the method described in section 5.2.3.1 to patient B. Five sampling positions were used to account for the resulting range uncertainties. The CT of 0% inhale was employed as reference CT. The treatment plan was optimized for this reference CT while the other CT phases were only used to set the range uncertainties of the individual rays. After the optimization, the resulting treatment plan was then recalculated on each breathing phase of the 4D-CT. The DVHs are shown in figure 5.10. While the conventional treatment plan

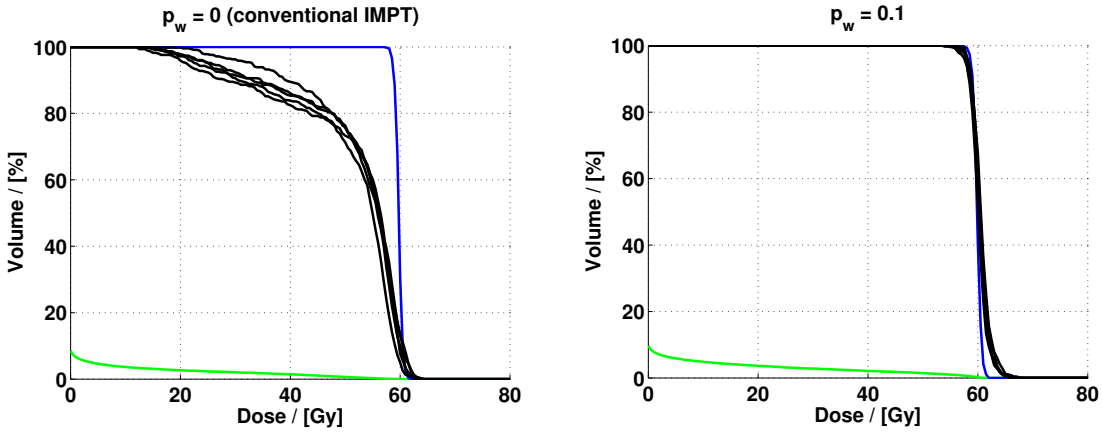

Figure 5.10: *Resulting DVHs for conventional IMPT ($p_w = 0$) and using the 4D method ($p_w = 0.1$). Blue: ITV, reference phase. Green: right lung, reference phase. Black: ITV recalculated on all other breathing phases. The DVH of the lung is only shown for the reference phase as it is nearly identical for all breathing phases.*

is only applicable for the reference CT phase, the worst case optimization resulted in a treatment plan that ensures target coverage for every CT phase used in the optimization. Since a single field was used in this study, the dose of each ray has to be a spread-out Bragg peak (SOBP). By accounting for the range uncertainties determined from the 4D-CTs in the optimization, the SOBP for each ray is enlarged such that it enclosed the ITV for all breathing phases. The enlargement of the SOBP for the individual rays ranged from 1.1 mm to 18.8 mm.

5.3.5 Benefits from precise delivery

In proton therapy, large efforts are taken to reduce uncertainties in the delivery process. With the method presented in this chapter, treatment plans can be generated which are insensitive to the remaining uncertainties. However, the quality of the resulting treatment plan will depend on the magnitude of the uncertainties considered during the optimization. By repeating the optimization for multiple values for the remaining uncertainties the potential improvement of the treatment plan quality due to an improved accuracy in the delivery process can be monitored. In figure 5.11, such a study is shown exemplarily for range uncertainties. The treatment plan for both patients described in this chapter was optimized accounting for various values of the range uncertainty. Here, patient B was treated as a static patient thus only the reference phase of the 4D CT was employed. The worst case penalty was chosen to $p_w = 1$ for all cases since this value resulted in insensitive treatment plans where the DVHs of the worst and the best case distributions are very close to each other. Since the worst case optimization resulted in treatment plans with a median dose in the target slightly above the prescribed value all treatment plans have been rescaled to a median dose of 60 Gy. The target volume for patient A is very close to

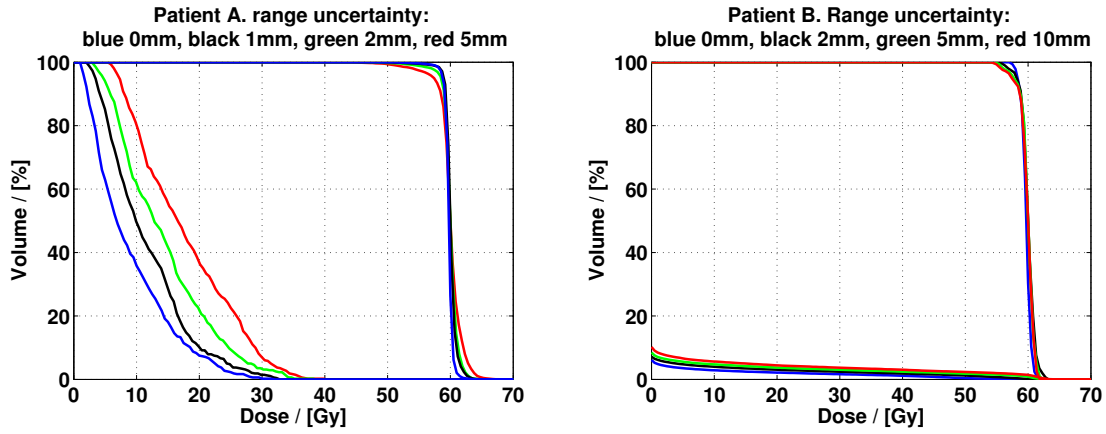

Figure 5.11: *The DVH of the nominal treatment plan for multiple values of the range uncertainty. The DVHs for the target and the brain stem are plotted for patient A (left), the ITV and the right lung were considered for patient B (right).*

the brain stem. As shown in figure 5.11, this patient can benefit from efforts to reduce the range uncertainty in the delivery process. Contrary to this, for patient B since there are no critical organs at risk in a close vicinity of the tumor thus the safety margins created by the worst case optimization can easily be expanded. Accounting for range uncertainties as large as 10 mm does not considerably reduce the treatment plan quality for this patient.

Elaborate efforts to reduce the range uncertainty which can result in additional costs and additional hazards (e.g. control CTs, general anesthesia) might not be necessary here.

5.3.6 Comparison of the three methods to account for uncertainties

Unkelbach et al. [50] recently published two methods to account for uncertainties in the optimization. Both methods as well as the worst case optimization resulted in very similar dose distributions. However they are conceptually different:
The *probabilistic method* proposed by Unkelbach considers the uncertain treatment variable as a random variable with an associated probability density function. Consequently, the dose is also an random variable which exhibits an expectation value. The objective function is then applied to this dose expectation value distribution. Doing so, the variance of the expectation value is minimized simultaneously, which means that the uncertainty of the delivered dose in minimized.
The *robust formulation* applies methods from robust linear programming. In linear programming the inverse treatment planning is treated as a feasibility problem. The linear objective function is translated into linear constraints for under- and overdosage of each voxel (see e.g. [53]). The feasible set of solutions are all beamlet weight vectors which fulfill these dose constraints. In the robust formulation of the linear programming problem, the constraints are extended for every possible value of the uncertain treatment variable such that the solution stays feasible for all cases. As stated in [50], the robust formulation can be interpreted as a worst case optimization where for each voxel the maximum absolute difference between the delivered and the prescribed dose, which can occur for all possible range combinations, is minimized.
Instead of using the complex robust formalism, the *worst case optimization* method simply calculates the worst case dose distribution. The objective function is then applied to both, the nominal and the worst case dose distribution.
Although all three methods show similar results, the different approaches have different strengths and weaknesses which will be outlined next. A summary is presented in table 5.1.

Model of the uncertainties: The robust formalism as well as the worst case optimization only require an interval of possible values for the uncertain treatment variable. Contrary to this the probabilistic method needs a detailed model of the uncertain treatment variable to determine the required probability density function. Such a model is usually not available. Unkelbach et al. employed a Gaussian distribution. This is a reasonable assump-

tions as long as the uncertainty is accumulated from multiple independent uncertainties, e.g. the total range uncertainty as a result of the accumulated range uncertainties due to the conversion from Hounsfield units to stopping powers for each voxel. Then according to the central limit theorem the resulting probability density function will be Gaussian. However this assumption might fail if the uncertainty results from a single uncertainty, e.g. the range uncertainty due to a CT artefact. However if such a detailed model for the uncertainty is available the probabilistic method can distinguish between systematic and random uncertainties which is not possible with the other two methods. This difference can be important since radiation therapy is applied in multiple fractions thus random uncertainties have less impact than systematic uncertainties.

Objective function: The robust formalism requires a linear objective function which is then translated into a constrained linear programming problem. Other objective functions like the standard quadratic objective function, objective functions to realize DVH constraints, the equivalent uniform dose (EUD) based objective function or objective functions for biological optimization cannot be used or at least have to be modified. Contrary to this the other two methods can readily be applied to any objective function.

Necessary sampling points: The uncertain treatment variable has to be sampled at discrete sampling points. The computational resources needed roughly scale with the number of sampling points, thus as few as possible should be used. To account for range uncertainties, the worst case optimization method already showed good results for three sampling points to account for a 5 mm range uncertainty. Contrary to this, Unkelbach et al. applied eleven sampling points to account for the same range uncertainty. Since the robust formulation also optimizes the worst case distributions three samples might also be sufficient for this method. The probabilistic method calculates a statistical quantity, the expectation value. For this calculation three samples can hardly suffice.

Correlation of the range uncertainties for individual rays: For both methods presented by Unkelbach et al. the range uncertainties of the individual rays in each treatment beam were assumed to be correlated. They did not observe large effects between using this assumption and employing uncorrelated range uncertainties for all rays. However the main source of range uncertainties are uncertainties of the patient model. For these uncertainties there is no physical reason to assume correlated range uncertainties. This assumption is necessary since for uncorrelated uncertainties all possible combinations have to be calculated. With k rays using N different ranges, this results in N^k possible dose distributions

which have to be calculated. One strength of the worst case optimization method is that if the worst case dose distribution can be calculated independently for parts of the treatment plan then the sum of them, D_w, is the solution for all possible combinations of the independent parts. In the shown example to account for range uncertainties I calculated the dose of 1392 independent rays at three different realizations of the beamlet range, thus the resulting D_w is the solution for 3^{1392} different total dose distributions, although only three had to be calculated.

Treatment plan evaluation: The DVHs of the worst case and the best case dose distributions can be used to visualize the sensitivity of the treatment plan to the considered uncertainties. These two dose distributions are readily available using the worst case optimization.

Complexity The worst case method is a simple method to account for uncertainties in the treatment planning process. The dose is calculated for N different realizations of the uncertain treatment variable. The resulting dose distributions are then combined to the worst case dose distribution D_w to which the objective function is applied. The computational resources needed roughly scale with N. Using three sampling points to take range uncertainties into account during the optimization resulted in an optimization time increase of only $2.3 - 3$ compared to standard IMPT. Quantitative comparisons of the optimization times are not available for the other two methods. The following informations are given by Jan Unkelbach in a personal communication. For the robust formulation, the number of constraints increases by $N \times K$, K being the number of treatment beams. The computation time is not directly proportional to this number, however it is expected to increase as the number of constraints increases. For $N = 11$ and $K = 3$, Jan Unkelbach observed an increase of the computation time of approximately 20. The computation time needed for an exact evaluation of the objective function in the probabilistic method scales with N^K. Therefore, they employed a stochastic gradient descent during the optimization. With $N = 11$ and $K = 3$, this approach yielded an increase of only approximately eleven in the computation time compared to the standard IMPT. The stochastic gradient descent might also be employed for the other two methods to reduce the computation time needed for the optimization. However, this has not been investigated.

	Probabilistic method	Robust formulation	Worst case optimization
no uncertainty model needed	-	+	+
distinction random/systematic uncer.	+	-	-
all obj. functions possible	+	-	+
needed sampling points	-	+	+
uncorrelated rays possible	-	-	+
Sensitivity analysis included	-	-	+
Complexity without/with stoch. grad.	- / +	- / ?	+ / ?

Table 5.1: *Summary of the advantages and disadvantages for the three methods to account for uncertainties.*

5.4 Discussion

5.4.1 Range uncertainties

Using the worst case optimization, the sensitivity to range uncertainties of IMPT treatment plans could be greatly reduced. For the example of range uncertainties it was also shown that the resulting IMPT treatment plans are even less sensitive to range uncertainties compared to a conventional, non-IMPT plan. Dose distributions which are insensitive to range uncertainties show three characteristics. The dose gradient between the target and the organ at risk is shaped using preferably the lateral gradient of the Bragg peaks. The dose from each treatment beam is extended at the distal edge of the target. Finally the dose profile in depth for each treatment beam is flattened compared to the standard IMPT.

5.4.2 Setup errors

Since the lateral fall-off of the beamlets is less steep than the distal fall-off setup errors are not as critical as range uncertainties. However setup errors can also lead to a change of the beamlet range. The DVHs in figure 5.6 showed that setup errors can lead to an underdosage of the target as well as to an overdosage of the organ at risk. This can be reduced if setup errors are taken into account during the optimization. Dose distributions with a reduced sensitivity to setup errors again show three characteristics. The weights of beamlets which hit the patient at a narrow angle are reduced. The dose from each treatment beam is extended at the lateral edge of the target. Finally the lateral dose profile for each treatment beam is flattened compared to the standard IMPT.
The sampling is more critical for setup errors than for range uncertainties. Next to a lateral interface an infinitesimal lateral shift can in principle lead to an arbitrarily large change

of the beamlet range. I checked the optimization results using a finer sampling of 1 mm for the setup errors. The recalculated DVHs where still inside the bounds predicted by the optimization using a sampling of 2 mm for the setup errors, thus a sampling of 2 mm seemed to be sufficient in the considered case. Since this patient had very pronounced lateral interfaces, namely the metal implants, this value might also be sufficient for other patients.

5.4.3 Range uncertainties and setup errors

Accounting only for range uncertainties in the optimization resulted in dose distributions where the lateral fall-off is used to shape the dose gradient between the target and the organ at risk, while accounting only for setup errors the lateral dose profile of each treatment beam is flattened. A compromise between these two conflicting goal has to be found if both uncertainties are accounted for in the optimization. One such compromise can be to choose the beamlet weights such that the distal dose gradient of the treatment beam is smoothed as shown in figure 5.9.

5.4.4 4D treatment planning with internal target volumes

Using the methods of range uncertainties is a very simple way to include 4D information into the proton treatment planning. There is no limit on the number of breathing phases used for this 4D treatment planning since besides the reference CT phase the other CT phases are only used to set the range uncertainties. Thus this method does not require extensive computational resources.
The method results in a treatment plan which is valid for all breathing phases. However, it is only valid if the complete dose is delivered during one breathing phase. For very small tumors this might be achievable. In 2004 the PSI was treating roughly a 300 cm^3 volume a minute [54]. The ITV of patient B has a volume of 8.1 cm^3 which results in a theoretical treatment time of about 1.6 s for this beam.
A currently pursued approach to reduce the sensitivity to motion artefacts in proton therapy is to implement a fast rescanning of the treatment beam. If the rescanning can be made fast enough to do a single scan within one breathing phase then a method which results in a treatment plan which is applicable in each breathing phase might be desirable.

5.4.5 Benefits from precise delivery

By accounting for uncertainties in the inverse treatment planning process, treatment plans which are insensitive to the considered uncertainties can be created, independently of the magnitude of the uncertainties. However the quality of the resulting treatment plan can depend on the magnitude of the uncertainties. The method shown in this section can be used to asses the potential benefit due to an improved accuracy in the delivery process. If an acceptable treatment plan can be generated considering relatively large uncertainties, costly and potentially dangerous procedures such as control CTs or general anesthesia might not be necessary. If multiple uncertainties are present, such an analysis can be used to assess the impact of the individual uncertainties. This can steer the development to reduce the remaining uncertainties.

5.5 Conclusion

The quality of proton treatment plans can be very sensitive to uncertainties in the treatment variables. Consequently all unavoidable related risks should be accounted for in the inverse treatment planning process. In this chapter I devised efficient treatment planning strategies that can account for a variety of potential treatment uncertainties and result in robust treatment plans with considerably reduced risks of respective treatment errors. The price to pay for this improved robustness of the treatment plan is a slightly deteriorated nominal dose distribution. The method presented in this chapter results in qualitatively similar treatment plans as the two methods described by Unkelbach et al.[50], suggesting that all three methods are equally applicable to reduce the sensitivity to uncertainties of IMPT plans.

Chapter 6

Summary, outlook and conclusion

In this chapter, the research done for this thesis is briefly summarized and an outlook on potential topics for further research is given.

Summary: The research towards risk-adapted optimization in IMPT presented in this thesis is divided into three parts.

In chapter 3, different optimization algorithms were investigated for the inverse treatment planning process. Additionally to the existing standard algorithm an improved optimization algorithm based on the L-BFGS algorithm and the conjugate gradient algorithm were implemented into KonRad. Furthermore the possibility to turn the constrained optimization into an unconstrained one using the a variable transformation was investigated. It turned out that the commonly employed standard optimization algorithm resulted in suboptimal treatment plans. By changing the optimization algorithm the resulting treatment plan quality could be considerably improved. The optimization algorithm based on the L-BFGS algorithm using the variable transformation to turn the constrained optimization problem into an unconstrained one was found to be superior to the other methods. Consequently, this method was used in the subsequent chapters.

Lateral tissue heterogeneities introduce uncertainties into the treatment planning process such as dose calculation errors and sensitivities to setup errors. In chapter 4, a method to quantify the lateral tissue heterogeneity of a single beamlet was developed. It was shown that this heterogeneity index H_i correlates with the dose calculation error introduced by the use of the fast pencil beam algorithm. Furthermore H_i also correlated with the sensitivity to setup errors of the dose distribution for an individual beamlet. The information of H_i can either be used prior to the optimization to avoid beam directions with large lateral heterogeneities or it can be integrated into the objective function to penalize individual beamlets. This reduced the dose calculation error introduced by the use of the fast pencil

beam dose calculation algorithm. However reducing the weight of a single beamlet can also induce a sensitivity of the resulting treatment plan to setup errors.

Finally in chapter 5 a inverse treatment planning method method was developed which accounts for uncertainties in the treatment planning and delivery process. This method is based on a worst case dose distribution D_w which is derived from all dose distributions resulting from possible realizations of an uncertain treatment variable. This worst case optimization method can considerably reduce the sensitivity of the resulting treatment plan to uncertainties. It was implemented to account for range uncertainties, setup errors and a combination of both uncertainties.

Outlook: Using the worst case optimization method, the two uncertainties with supposably the largest impact on the resulting dose distributions for a static patient have been integrated into the inverse treatment planning process. Other uncertainties in a static patient model as e.g. uncertainties in the modeling of biological input data can be integrated in a similar manner. However large uncertainties also result from the motion of the patient such as breathing, cardiac or even peristaltic motion. Current approaches to reduce these uncertainties aim to change the delivery process of particle therapy. These methods include employing a gated radiation therapy (see e.g [16]), to utilize a fast rescanning of the treatment fields (see e.g [55]) and to dynamically track the tumor volume (see e.g [56]). An interesting field of further research is to account for these uncertainties in the optimization process. An example how to use the worst case optimization for such a 4D optimization was given, however this approach can only be applied under very restricted conditions. It neglects interplay effects between the delivery and the tumor motion, furthermore it assumes that all beamlets in a ray are correlated, that is that they are applied during the same phase of the motion. These two assumptions lead to the constraint that the whole treatment plan needs to be delivered within the same phase of the motion. If this constraint is to be relaxed, other correlations have to be found in order to employ the worst case optimization method to achieve a treatment plan which is robust against patient motion. Such a correlation could be e.g. that an energy slice can be delivered within one breathing phase. If this assumption holds, D_w could be calculated as accumulation of the worst case dose distributions calculated for each energy slice determined from all breathing phases. Most likely large patient motions would decrease the nominal plan quality for such a robust treatment plan significantly. However such a method could also be used in combination with the gating technique to enlarge the gating window. If only patient motions within a gating window have to be accounted for the resulting nominal plan quality might be acceptable. But such an approach could reduce the disadvantages resulting from the gating

technique using a very narrow gating window.

Another approach to improve the worst case optimization could be to integrate the stochastic gradient optimization methods proposed by Unkelbach for the probabilistic method (see section 5.3.6). Using this approach, many additional sampling positions could be integrated into the worst case optimization.

As stated in chapter 4, the heterogeneity number H_i can be used prior to the optimization to choose beam directions which avoid large heterogeneities or during the optimization to reduce the weight of individual beamlets with large heterogeneities. A third approach to utilize the information of H_i could be a post-processing of the optimization results. The idea is outlined in the following. In the minimum, the objective function can be approximated by a quadratic function. The second order Taylor expansion $T(\vec{w})$ of the objective function $F(\vec{w})$ around the minimum $\vec{w}_m$ is given by:

$$T(\vec{w}) = F(\vec{w}_m) + \nabla F(\vec{w}_m)(\vec{w} - \vec{w}_m) + \frac{1}{2}(\vec{w} - \vec{w}_m)\nabla^2 F(\vec{w}_m)(\vec{w} - \vec{w}_m) \tag{6.1}$$

Since $F(\vec{w}_m)$ is constant and the gradient of the objective function in the minimum, $\nabla F(\vec{w}_m)$, is zero this leaves only the second order term, thus the Hessian matrix in the minimum $\nabla^2 F(\vec{w}_m)$. An eigen analysis of the Hessian matrix can now determine the degeneracy of the solution space (see e.g. [6]). Changing the solution $\vec{w}_m$ along any eigenvector of the Hessian matrix with an eigenvalue of zero will not change the objective function value, thus these vectors span the space of degenerated solutions. This space can now be searched for the solution with the minimal total heterogeneity. As stated in chapter 4, reducing the weight of individual beamlets can lead to a sensitivity to setup errors of the resulting treatment plan. Thus an objective function which accounts for setup errors as shown in chapter 5 should be used. Of course, this space could also be searched for other goals too, e.g. for a minimal delivery time. The computational effort needed to calculate the Hessian matrix is usually prohibitive due to the large number of degrees of freedom in IMPT, however if an quasi-Newton algorithm as e.g. the BFGS algorithm is used during the optimization, an approximation of the inverse Hessian matrix is available which could be employed for this task. The challenge of this approach is that the determination of the eigenvalues and -vectors are problems of third order, meaning the computational effort needed is proportional to N^3 (N being the number of beamlets).

Conclusions: In conclusion, several improvements for IMPT have been developed in this thesis. The improved optimization algorithm was shown to be superior to the standard optimization algorithm, considerably improving the resulting IMPT treatment plan quality.

This algorithm can also be employed in the optimization of IMRT treatment plans. Furthermore, the heterogeneity index H_i was developed as a valuable tool to quantify the lateral tissue heterogeneities encountered by individual beamlets. The usefulness of the heterogeneity index is stressed by the fact that it was already adopted by other researchers [44]. Finally, the importance of accounting for treatment planing and delivery uncertainties has been shown. The worst case optimization method was developed as a simple and effective method to account for such uncertainties.

Chapter 7

Appendix

Strong Wolfe conditions

The strong Wolfe conditions are used as the convergence criteria for the inexact line minimization in the improved optimization algorithm. They are: (see e.g. [38], chapter 3)

$$F(\vec{x}+\alpha\vec{p}) \leq F(\vec{x}) + c_1\alpha\vec{\nabla}F(\vec{x})\vec{p} \tag{7.1}$$

$$\left|\vec{\nabla}F(\vec{x}+\alpha\vec{p})\vec{p}\right| \leq c_2\left|\vec{\nabla}F(\vec{x})\vec{p}\right| \tag{7.2}$$

$\vec{x}$ is the position were the line search is started, $\vec{p}$ the search direction, α the step length. Equation 7.1 is also known as sufficient decrease condition, equation 7.2 as curvature condition. c_1 and c_2 are constants which can be chosen between $0 < c_1 < c_2 < 1$. The constant c_2 can be used to tune the line search. A small value of c_2 results in a step which is close to the real line minimum, a larger value of c_2 accepts steps which are further apart from it. I employed the values recommended in [38] for quasi-Newton algorithms, namely $c_1 = 0.0001$ and $c_2 = 0.9$. This large value of c_2 is recommended since for quasi-Newton methods it is in general more beneficial to continue along a new search direction instead of spending too much time to locate the exact minimum on the line.

Bibliography

[1] Krebs in Deutschland. 5. überarbeitete, aktualisierte Ausgabe. Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V. und das RKI. Saarbrücken, (2006). *see page* 1

[2] Zur Lage der deutschen Strahlentherapie. Positionspapier der Deutschen Gesellschaft für Radioonkologie e.V. (DEGRO) und des Berufsverbands Deutscher Strahlentherapeuten e.V. (BVdST). ISBN 3-934264-97-2. (2001). *see page* 1

[3] D. Pflugfelder, J. J. Wilkens, U. Oelfke: Worst case optimisation: a method to account for uncertainties in the optimisation of intensity modulated proton therapy. Phys. Med. Biol. *submitted*. *see pages* 1, 2

[4] G. T. Chen, S. Rosenthal, M. Goitein, J. C. Roeske: Quantitation of water equivalent pathlength variations from serial CT for prostate treatment with proton beams. Int. J. Radiat. Oncol. Biol. Phys. **48** (2000) 339. *see page* 1

[5] H. Szymanowski, U. Oelfke: Two-dimensional pencil beam scaling: an improved proton dose algorithm for heterogeneous media. Phys. Med. Biol. **47** (2002) 3313 – 3330. *see pages* 2, 13

[6] M. Alber, G. Meedt, F. Nüsslin: On the degeneracy of the IMRT optimization problem. Med. Phys. **29** (2002) 2584 – 2589. *see pages* 2, 75

[7] D. Pflugfelder, J. J. Wilkens, S. Nill, U. Oelfke: A comparison of three optimization algorithms for intensity modulated radiation therapy. Z. Med. Phys. *in press*. *see page* 2

[8] D. Pflugfelder, J. J. Wilkens, H. Szymanowski, U. Oelfke: Quantifying lateral tissue heterogeneities in hadron therapy. Med. Phys. **34** (2007) 1506 – 1513. *see page* 2

[9] D. Pflugfelder, J. J. Wilkens, H. Szymanowski, U. Oelfke: Towards risk adapted inverse planning for protons: Avoiding risks due to lateral tissue inhomogeneities. In: Biomedizinische Technik, **50** (Suppl.1, Part 1). Editors: U. Boenick et al. Berlin: Fachverlag Schiele & Schön GmbH, (2005) 342 – 343. *see page* 2

[10] H. Szymanowski, D. Pflugfelder, S. Nill, U. Oelfke: Practical implementation of an improved proton dose algorithm for heterogeneous media. Radiotherapy and Oncology, **76** (Suppl.2) (2005) *S*56. *see page* 2

[11] D. Pflugfelder, J. J. Wilkens, H. Szymanowski, U. Oelfke: Towards risk adapted inverse planning: Avoiding risks due to lateral tissue inhomogeneities. In: Abstracts of the 43rd PTCOG Meeting, Munich 10-14 December (2005). *see page* 2

[12] D. Pflugfelder, J. J. Wilkens, H. Szymanowski, U. Oelfke: Towards risk adapted inverse planning for protons: Avoiding risks due to lateral tissue inhomogeneities. In: E-Verhandlungen 2006. Abstracts der Frühjahrstagung in Heidelberg. Heidelberg: Deutsche Physikalische Gesellschaft (DPG) (2006) *see page* 2

[13] H. Szymanowski, T. Fuchs, S. Nill, J. J. Wilkens, D. Pflugfelder, U. Oelfke, Y. Glinec, J. Faure, V. Malka: Applications of the Monte Carlo code GEANT to particle beam therapy. In: Astroparticle, Particle and Space Physics, Detectors and Medical Physics Applications - Proceedings of the 9th Conference. Editors: M. Barone et al. World Scientific Publishing: New Jersey (2006) 758 – 767. *see page* 2

[14] D. Pflugfelder, J. Wilkens, S. Nill, U. Oelfke: Comparison of two optimization methods for inverse treatment planning. Radiotherapy and Oncology, **81** (Suppl.1) (2006) *S*29. *see page* 2

[15] R. R. Wilson: Radiological use of fast protons. Radiology **47** 487 – 491. *see page* 3

[16] W. Schlegel, A. Mahr: 3D Conformal Radiation Therapy - A Multimedia Introduction to Methods and Techniques 2nd revised and enhanced edition. Springer Verlag Berlin Heidelberg New York (2007). *see pages* 3, 5, 6, 7, 17, 74

[17] W. H. Bragg, R. Kleeman: On the Ionization Curves of Radium. Philos. Mag., **S.6**, 8 (1904) 726 – 738. *see page* 3

[18] T. Bortfeld: An analytical approximation of the Bragg curve for therapeutic proton beams. Med. Phys. **24** (1997) 2024 – 2033. *see pages* 3, 4, 12

[19] J. Sisterson (editor): Particles (Newsletter, Number 36). Particle Therapy Co-Operative Group (PTCOG), (2005). URL: `http://www.ptcog.com/particles.php`. *see page* 4

[20] R. J. Schulz, A. R. Smith, C. G. Orton: Point/CounterPoint: Proton therapy is too expensive for the minimal potential improvements in outcome claimed. Med. Phys. **34** (2007) 1135 – 1138. *see page* 4

[21] E. J. Hall: Intensity-modulated radiation therapy, protons, and the risk of second cancer. Int. J. Radiat. Oncol. Biol. Phys. **65** (2006) 1 – 7. *see page* 6

[22] A. J. Lomax, E. Pedroni, H. Rutz, G. Goitein: The Clinical Potential of Intensity Modulated Proton Therapy. Z. Med. Phys. **14** (2004) 147 – 152. *see pages* 8, 53, 54

[23] A. Lomax: Intensity modulation methods for proton radiotherapy. Phys. Med. Biol. **44** (1999) 185 – 205. *see pages* 8, 9, 17

[24] S. Nill: Development and Application of a Multi-Modality Inverse Treatment Planning System. Ph.D. thesis, University of Heidelberg. URL: `http://www.ub.uni-heidelberg.de/archiv/1802/`. *see page* 8

[25] J. O. Deasy, D. M. Shephard, T. R. Mackie: Distal edge tracking: a proposed delivery method for conformal proton therapy using intensity modulation. In: D.D. Leavitt, G. Starkschall (editors). Proceedings of the 12th International Conference on the Use of Computers in Radiation Therapy. Madison, WI, Medical Physics Publishing, (1997) 407 – 409. *see page* 8

[26] A. Brahme, P. Kallman, B. Lind: Optimization of proton and heavy ion therapy using an adaptive inversion algorithm. Radiother. Oncol. **15** (1989) 189 – 197. *see page* 8

[27] U. Oelfke: The Potential of Charged Particle Beams in Conformal Radiation Therapy. Aachen: Shaker, (2002). *see pages* 9, 10, 11, 13

[28] G. Molière: Theorie der Streuung schneller geladener Teilchen II: Mehrfach- und Vielfachstreuung. Z. Naturforschung, **2a** (1948) 133 – 145. *see page* 12

[29] M. Soukup, M. Fippel, M. Alber: A pencil beam algorithm for intensity modulated proton therapy derived from Monte Carlo simulations. Phys. Med. Biol. **50** (2005) 5089 – 5104. *see pages* 13, 51

[30] S. Agostinelli et al.: Geant4 – a simulation toolkit. Nuclear Instruments and Methods in Physics Research A 506 (2003) 250 – 303. *see pages* 13, 38

[31] W. Schneider, T. Bortfeld, W. Schlegel: Correlation between CT numbers and tissue parameters needed for Monte Carlo simulations of clinical dose distributions. Phys. Med. Biol. **45** (2000) 459 – 478. *see pages* 14, 40

[32] H. Szymanowski, U. Oelfke: CT calibration for two-dimensional scaling of proton pencil beams. Phys. Med. Biol. **48** (2003) 861 – 874. *see page* 14

[33] B. Schaffner, E. Pedroni: The precision of proton range calculation in proton radiotherapy treatment planning: experimental verification of the relation between CT-HU and proton stopping power. Phys. Med. Biol. **43** (1998) 1579 – 1592. *see pages* 14, 55

[34] T. Bortfeld, J. Bürkelbach, R. Boesecke, W. Schlegel: Methods of image reconstruction from projections applied to conformation radiotherapy. Phys. Med. Biol. **35** (1990) 1423 – 1434. *see pages* 17, 19

[35] M. Lahanas, E. Schreibmann, D. Baltas: Multiobjective inverse planning for intensity modulated radiotherapy with constraint-free gradient-based optimization algorithms. Phys. Med. Biol. **48** (2003) 2843 – 2871. *see pages* 17, 19, 30

[36] X. Zhang, H. Lui, X. Wang, L. Dong, Q. Wu, R. Mohan: Speed and convergence properties of gradient algorithms for optimization of IMRT. Med. Phys. **31** (2004) 1141 – 1152. *see pages* 17, 18

[37] J. Nocedal: Updating quasi-Newton matrices with limited storage. Mathematics of Computation **35** (1980) 773 – 782. *see page* 18

[38] J. Nocedal, S. J. Wright: Numerical Optimization. Springer Series in Operations Research. New York: Springer, (1999). *see pages* 18, 19, 20, 21, 77

[39] M. Alber, M. Birkner, W. Laub and F. Nüsslin: Hyperion: An integrated IMRT planning tool. Proceedings of the XIII International Conference on the Use of Computers in Radiation Therapy, edited by W. Schlegel and T. Bortfeld Springer, Heidelberg, (2000) 46 – 48. *see page* 18

[40] W. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling: Numerical Recipes in C. Cambridge: Cambridge University Press, (1988). *see page* 21

[41] C. Zhu, R. H. Byrd, J. Nocedal: L-BFGS-B: Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization. ACM Transactions on Mathematical Software, **23** 4 (1997) 550 – 560. *see page* 27

[42] R. H. Byrd, P. Lu and J. Nocedal. A Limited Memory Algorithm for Bound Constrained Optimization, SIAM Journal on Scientific and Statistical Computing, **16**, 5, (1995) 1190 – 1208. *see page* 27

[43] T. Bortfeld, J. Stein, K. Preiser: Clinically Relevant Intensity Modulated Optimization using Physical Criteria. In: D.D. Leavitt, G. Starkschall (editors). Proceedings of the 12th International Conference on the Use of Computers in Radiation Therapy. Madison, WI, Medical Physics Publishing, (1997) 1 – 4. *see page* 30

[44] M. Soukup, M. Alber: Influence of dose engine accuracy on the optimum dose distribution in intensity-modulated proton therapy treatment plans. Phys. Med. Biol. **52** (2007) 725 – 740. *see pages* 37, 76

[45] M. Urie, M. Goitein, W. R. Holley, G. T. Y. Chen: Degradation of the Bragg peak due to inhomogeneities. Phys. Med. Biol. **31** (1986) 1 – 15. *see page* 35

[46] D. A. Low, W. B. Harms, S. Mutic, J. A. Purdy: A technique for the quantitative evaluation of dose distributions. Med. Phys. **25**, 5, (1998) 656 – 661. *see page* 40

[47] T. Depuydt, A. Van Esch, D. P. Huyskens: A quantitative evaluation of IMRT dose distributions: refinement and clinical assessment of the gamma evaluation. Radiotherapy and Oncology **62** (2002) 309 – 319. *see page* 40

[48] K. H. Kuefer, H. W. Hamacher, T. Bortfeld: A multicriteria optimization approach for inverse radiotherapy planning, in Proceedings of the 13th International Conference on the Use of Computers in Radiation Therapy, edited by W. Schlegel and T. Bortfeld (Springer, Heidelberg, Germany) 2000, 26 – 28. *see page* 50

[49] J. J. Wilkens, J. R. Alaly, K. Zakarian, W. L. Thorstad, J. O. Deasy: IMRT treatment planning based on prioritizing prescription goals. Phys. Med. Biol. **52** (2007) 1675 – 1692. *see page* 50

[50] J. Unkelbach, T. C. Y. Chan, T. Bortfeld: Accounting for range uncertainties in the optimization of intensity modulated proton therapy. Phys. Med. Biol. **52** (2007) 2755 – 2773. *see pages* 53, 58, 67, 72

[51] International Commission on Radiation Units and Measurements, Inc. (ICRU) Report 62: Prescribing, recording, and reporting photon beam therapy (supplement to ICRU Report 50), ICRU (1999) Bethesda, MD. *see page* 56

[52] A. J. Lomax et al.: Intensity modulated proton therapy: A clinical example. Med. Phys. **28** (2001) 317 – 324. *see page* 58

[53] T. C. Y. Chan, T. Bortfeld, J. N. Tsitsiklis: A robust approach to IMRT optimization. Phys. Med. Biol. **51** (2006) 2567 – 2583. *see page* 67

[54] A. Lomax, T. Böhringer, A. Bolsi, D. Coray, F. Emert, G. Goitein, M. Jermann, S. Lin, E. Pedroni, H. Rutz, O. Stadelmann, B. Timmermann, J. Verwey, D. C. Weber: Treatment planning and verification of proton therapy using spot scanning: Initial experiences. Med. Phys. **31** (2004) 3150 – 3157. *see page* 71

[55] M. H. Philips, E. Pedroni, H. Blattmann, T. Boehringer, A. Coray, S. Scheib: Effects of respiratory motion on dose uniformity with a particle scanning method. Phys. Med. Biol. **37** (1992) 223 – 234. *see page* 74

[56] C. Bert, E. Rietzel: 4D treatment planning for scanned ion beams. Radiat. Oncol. **2**:24 (2007). *see page* 74

List of Figures

List of Tables

Acknowledgments

I would like to thank my supervisor Prof. Uwe Oelfke for the constant support and especially for giving me the opportunity to work here at the DKFZ. I am also obliged to Prof. Schlegel for acting as the second referee for this thesis. Furthermore I would like to express my gratitude to Prof. Schlegel, Prof. Uwe Oelfke and all other colleagues for creating such an exceptional working environment.

I would like to thank especially Dr. Jan Wilkens for countless suggestions, advises, proof-readings, ideas, help with formalities, etc. etc. etc.

Finally I am grateful to Lilli for awaiting me every evening and to Pia for being there.