Efficiency of analytical and sampling-based uncertainty propagation in intensity-modulated proton therapy

N Wahl[1], P Hennig[2], HP Wieser[1] and M Bangert[1]

[1] Department of Medical Physics in Radiation Oncology, German Cancer Research Center - DKFZ, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany and Heidelberg Institute for Radiation Oncology - HIRO, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany
[2] Max-Planck Institute for Intelligent Systems, Spemannstraße, 72076 Tübingen, Germany

E-mail: n.wahl@dkfz.de

January 2017

Submitted to: *Phys. Med. Biol.*

Abstract. The sensitivity of intensity-modulated proton therapy (IMPT) treatment plans to uncertainties can be quantified and mitigated with robust/min-max and stochastic/probabilistic treatment analysis and optimization techniques. Those methods usually rely on sparse random, importance, or worst-case sampling. Inevitably, this imposes a trade-off between computational speed and accuracy of the uncertainty propagation.

Here, we investigate analytical probabilistic modeling (APM) as an alternative for uncertainty propagation and minimization in IMPT that does not rely on scenario sampling. APM propagates probability distributions over range and setup uncertainties via a Gaussian pencil-beam approximation into moments of the probability distributions over the resulting dose in closed form. It supports arbitrary correlation models and allows for efficient incorporation of fractionation effects regarding random and systematic errors.

We evaluate the trade-off between run-time and accuracy of APM uncertainty computations on three patient datasets. Results are compared against reference computations facilitating importance and random sampling. Two approximation techniques to accelerate uncertainty propagation and minimization based on probabilistic treatment plan optimization are presented. Runtimes are measured on CPU and GPU platforms, dosimetric accuracy is quantified in comparison to a sampling-based benchmark (5000 random samples).

APM accurately propagates range and setup uncertainties into dose uncertainties at competitive run-times (GPU ≤ 5 min). The resulting standard deviation (expectation value) of dose show average global $\gamma_{3\,\%/3\,\mathrm{mm}}$ pass rates between 94.2 % and 99.9 % (98.4 % and 100.0 %). All investigated importance sampling strategies provided less accuracy at higher run-times considering only a single fraction. Considering fractionation, APM uncertainty propagation and treatment plan optimization was proven to be possible at constant time complexity, while run-times of sampling-based computations are linear in the number of fractions. Using sum sampling within APM, uncertainty propagation can only be accelerated at the cost of reduced accuracy in variance calculations. For probabilistic plan optimization, we were able to approximate the necessary pre-computations within seconds, yielding treatment plans of similar quality as gained from exact uncertainty propagation.

APM is suited to enhance the trade-off between speed and accuracy in uncertainty propagation and probabilistic treatment plan optimization, especially in the context of fractionation. This brings fully-fledged APM computations within reach of clinical application.

1. Introduction

The physical characteristics of dose deposition in intensity-modulated proton therapy (IMPT) allow for better dose conformity than in photon therapy, however, at the cost of an increased sensitivity to uncertainties (Lomax, 2008a,b). While photon dose distributions subject to uncertainty can be described by the static dose cloud approximation (Bortfeld et al., 2004; Herk et al., 2000), the assumption of a spatially invariant dose distribution covering a variable patient anatomy usually fails for protons. Consequently, conventional photon margin concepts alone do not suffice to achieve robustness for protons.

Besides extensions or generalizations of photon margin concepts for application with protons in combination with single-field-uniform-dose techniques (Fredriksson and Bokrantz, 2016; Knopf et al., 2013), many approaches to achieve robustness rely on the computation of a case specific uncertainty metric. This can be a worst-case dose distribution to be used for robust optimization (Chen et al., 2012; Chu et al., 2005; Fredriksson et al., 2011; Liu et al., 2012; Pflugfelder et al., 2008) or an expected dose distribution along with its standard deviation to be used for probabilistic optimization (Unkelbach et al., 2009).

While those techniques generate more robust treatment plans and can also provide robustness indicators, both rely on the repeated simulation of different treatment scenarios. Such a sampling process inevitably imposes limitations regarding the accuracy of the uncertainty metric due to computational cost. E. g., worst case dose distributions are solely calculated for fully correlated shift scenarios (more sophisticated correlation models easily result in prohibitive combinatorics), expected dose distributions and dose standard deviations are estimated on ten samples during each optimization step (Unkelbach et al., 2009) (higher sampling rates linearly increase computation time), and the interplay of systematic and random errors in the context of fractionated radiation therapy are often neglected (because they would have to be modeled separately (Lowe et al., 2016)). So far, methods addressing these computational and conceptual limitations are sparse and still rely on sampled scenarios at their basis; i. e. Sobotta et al. (2012) and Perkó et al. (2016) build surrogate models from sampled scenarios, enabling for example the incorporation of fractionation effects, while not necessarily generalizing to optimization.

Previously, Bangert et al. (2013) introduced Analytical Probabilistic Modeling (APM) for probabilistic IMPT planning, enabling closed-form uncertainty propagation for range and setup errors without scenario sampling. APM uses an analytical, Gaussian representation of a pencil beam dose calculation algorithm facilitating analytical integration against Gaussian probability densities. Thereby, APM yields an analytical mapping of pencil beam weights to the expectation value and covariance of dose in IMPT, directly generalizing to probabilistic optimization. APM implicitly supports the implementation of arbitrary correlation assumptions on the uncertainty model. Additionally it differentiates between systematic and random sources of uncertainty in the context of fractionation at constant computational complexity.

There are three main contributions in this paper. (1) We investigate the accuracy for APM computations of the expectation value and the standard deviation of an IMPT dose distribution subject to range and setup uncertainties. (2) We study the computational complexity of APM moment calculations as well as probabilistic optimization with a particular focus on fractionation. (3) We introduce new approximations that achieve a significant reduction in computational cost in APM

while incuring only minor decrease in quality. For all investigations, sampling-based uncertainty propagation techniques serve as benchmark solutions. Hence, this paper also provides a quantitative basis for a general discussion of the trade-off between speed and accuracy in uncertainty quantification.

2. Analytical probabilistic modeling

The overarching idea of APM is to describe uncertainty in treatment plan quality indicators by the moments of the underlying probability distribution (Bangert et al., 2013). E. g. uncertainty in dose is described by its expectation value and (co-)variance. The APM formalism is based on a functional approximation of the dose calculation algorithm that enables closed-form propagation of Gaussian probability distributions describing uncertain input parameters (i.e. range Δ^z and patient setup $\Delta^{x/y}$) through all numerical computations.

A detailed introduction to APM is given by Bangert et al. (2013). At this point, we only summarize the main findings which are directly relevant for present study.

2.1. Moment calculation

Conventional dose optimization usually employ a dose influence matrix D whose elements D_{ij} map the weight w_j of pencil beam j to the dose d_i in voxel i. With APM, it is possible to compute a probabilistic analogue $\mathcal{D}$ that provides a linear mapping from pencil beam weights to the expected dose $\mathrm{E}[\boldsymbol{d}]$. Given the probability densities $p(\boldsymbol{\Delta}^{x/y})$ over setup errors and $p(\boldsymbol{\Delta}^z)$ over range errors, for the first central moment, we need to solve

$$\mathcal{D}_{ij} = \iiint p(\boldsymbol{\Delta}^x)p(\boldsymbol{\Delta}^y)p(\boldsymbol{\Delta}^z)D_{ij}\mathrm{d}\boldsymbol{\Delta}^x\mathrm{d}\boldsymbol{\Delta}^y\mathrm{d}\boldsymbol{\Delta}^z \ , \tag{1}$$

to obtain an expected dose influence matrix $\mathcal{D}_{ij}$, and map the spot weights $\boldsymbol{w}$ to the expectation value of dose $\mathrm{E}[\boldsymbol{d}]$ via

$$\mathrm{E}[d_i] = \sum_j w_j \mathcal{D}_{ij} \ . \tag{2}$$

For the second central moment, we compute the raw moment contribution $\widetilde{\mathcal{V}}_{ijlm}$ to get a 4-th order (co)variance influence tensor

$$\mathcal{V}_{ijlm} = \widetilde{\mathcal{V}}_{ijlm} - \mathcal{D}_{ij}\mathcal{D}_{lm} = \iiint p(\boldsymbol{\Delta}^x)p(\boldsymbol{\Delta}^y)p(\boldsymbol{\Delta}^z)D_{ij}D_{lm}\mathrm{d}\boldsymbol{\Delta}^x\mathrm{d}\boldsymbol{\Delta}^y\mathrm{d}\boldsymbol{\Delta}^z - \mathcal{D}_{ij}\mathcal{D}_{lm} \tag{3}$$

describing the influence of a pairwise pencil beam combination jm in the (co)variance of dose on voxels i and l via

$$\mathrm{Cov}[d_i, d_l] = \sum_{jm} w_j w_m \mathcal{V}_{ijlm} \ , \tag{4}$$

$$\sigma^2[d_i] = \mathrm{Cov}[d_i, d_i] = \sum_{jm} w_j w_m \mathcal{V}_{ijim} \ . \tag{5}$$

We solve the integrals (1) and (3) by assuming Gaussian probability densities and accurately approximating Hong et al.'s (1996) pencil beam algorithm by solely relying on (superpositions of) Gaussians. This results in $\mathcal{D}_{ij}$ and $\mathcal{V}_{ijlm}$ taking the form of a sum over one and two dimensional Gaussian Kernels as detailed in Appendix A.

2.2. Probabilistic Optimization

Equations (2) and (4) represent a closed-form expression of the expectation value and covariance of the dose. Together they can be used to model uncertainty in the dose with a multivariate normal distribution. This allows for the closed form formulation (and optimization) of the expected value $\mathrm{E}[\mathcal{F}]$ of a piece-wise squared objective function $\mathcal{F} = (\boldsymbol{d} - \boldsymbol{d}^*)^T P (\boldsymbol{d} - \boldsymbol{d}^*)$ (Oelfke and Bortfeld, 2001)

$$\mathrm{E}[\mathcal{F}(\boldsymbol{w})] = \underbrace{\mathrm{tr}(P\Sigma^{\boldsymbol{d}})}_{\sum_i p_i \sum_{jm} \mathcal{V}_{ijim} w_j w_m} + \underbrace{(\mathrm{E}[\boldsymbol{d}] - \boldsymbol{d}^*)^T P (\mathrm{E}[\boldsymbol{d}] - \boldsymbol{d}^*)}_{\sum_i p_i \left(\sum_j \mathcal{D}_{ij} w_j - d_i^*\right)^2} \tag{6}$$

where $P = \mathrm{diag}\,(p_1, p_2, \cdots, p_{n_V})$ is a diagonal optimization penalty matrix and $\boldsymbol{d}^*$ is the prescribed or desired dose distribution (compare Bangert et al., 2013, eq. 36).

Storage of the full $(n_V \times n_B \times n_V \times n_B)$-sized tensor $\mathcal{V}$, where n_V and n_B are number of voxels and beamlets, respectively, is despite its sparsity and symmetry infeasible. For the optimization of $\mathrm{E}[\mathcal{F}]$, however, we do not need to compute the elements where $i \neq l$, and can omit storage of the remaining voxel dimension by performing the product with the penalty matrix P on the fly (compare Bangert et al., 2013, eq. 38).

$$\Omega_{jm} = \sum_i p_i \mathcal{V}_{ijim} \ , \tag{7}$$

The resulting matrix Ω is only of size $n_B \times n_B$ and maps pairwise spot-combinations to the penalty-weighted integral variance in dose via the quadratic form

$$\mathrm{tr}\left(P\Sigma^{\boldsymbol{d}}\right) = \sum_i p_i \sum_{jm} \mathcal{V}_{ijim} w_j w_m = \boldsymbol{w}^T \Omega \boldsymbol{w} = \sum_{\mathrm{VOI} \in \mathrm{VOIs}} p^{\mathrm{VOI}} \boldsymbol{w}^T \Omega^{\mathrm{VOI}} \boldsymbol{w} \tag{8}$$

which is used to compute the variance term in eq. (6) for optimization. By separating into VOI-based matrices Ω^{VOI} during summation, re-optimizations with changed optimization penalties p^{VOI} can be performed without recomputing Ω.

2.3. Uncertainty Model

In this paper, for all computations except for the validation (see section 3) we adopt the "ray-wise" uncertainty model described in Unkelbach et al. (2009) and Bangert et al. (2013); errors are spatially independent, all beamlets belonging to the same beam underlie perfectly correlated lateral setup errors (corresponding to full beam shifts) and all beamlets on the same central ray (thus penetrating the same tissue) underlie perfectly correlated range errors. If not stated otherwise, we assume random setup errors of 2 mm with a random range error of 1 mm and a systematic setup error of 1 mm with a relative systematic depth error of 3.5 %.

2.4. Fractionation

Within the APM formalism, fractionation effects can be incorporated with a correlation model over two different sources of uncertainty (Bangert et al., 2013, Section 3.5). Therefore, systematic preparation errors are considered as perfectly correlated through all fractions while random execution errors are considered as uncorrelated through all fractions. While the expectation value is independent from

the correlation assumptions and therefore F, the raw moment contribution $\widetilde{\mathcal{V}}_{ijlm}$ in eq. (3) is now composed out of two terms weighted by the fraction count F:

$$\mathcal{V}_{ijlm} = \frac{\widetilde{\mathcal{V}}^{\text{corr}}_{ijlm} + (F-1)\,\widetilde{\mathcal{V}}^{\text{uncorr}}_{ijlm}}{F} - \mathcal{D}_{ij}\mathcal{D}_{lm}\ . \tag{9}$$

The term $\widetilde{\mathcal{V}}^{\text{corr}}_{ijlm}$ includes the systematic and random covariance components, i. e. the full uncertainty model with potential correlation for a single fraction, while $\widetilde{\mathcal{V}}^{\text{uncorr}}_{ijlm}$ is evaluated based on the systematic correlation model but assuming always uncorrelated random components to model behavior across fractions.

For $F > 1$ the respective Gaussian kernels have to be evaluated twice and not F times. Hence, the evaluation of $\mathcal{V}$ (and therefore $\sigma^2[\boldsymbol{d}]$) is independent of the number of fractions F.

2.5. Implementation

We implemented APM within a `C++` research treatment planning plugin for MITK (Wolf et al., 2005). The radiological depth is modeled with a point-of-interest (POI) ray-casting (Schaffner et al., 1999; Siggel et al., 2012). For optimization we use IPOPT (Wächter and Biegler, 2006). Section 4 and Appendix B elaborate on the computationally challenging implementation of $\mathcal{V}$ and Ω on CPU and GPU.

2.6. APM computations in 3D

Earlier, APM computations have only been demonstrated for one- or two-dimensional problems (Bangert et al., 2013). To prove the applicability of APM to 3D patient data, we analyzed three patient datasets: an intra-cranial case, a para-spinal case and a prostate case which are explained in Appendix C.

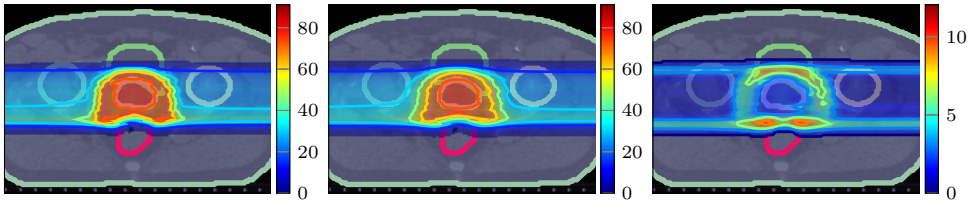
Figure 1 shows a physical dose distribution, its expectation value and standard deviation before and after probabilistic optimization for 5 fractions, demonstrating the reduction of standard deviation within the target and the OARs.

To emphasize the effect of the fraction number on $\sigma\,[\boldsymbol{d}]$ with and without probabilistic planning, we show standard deviation volume histograms (SDVH) (as used by Perkó et al., 2016; Unkelbach et al., 2009) of all three patient cases in figure 2 for two different fractionation schemes. As expected, the standard deviation in the target is significantly decreased by probabilistic optimization. Due to a consistent influence of systematic uncertainties, we also observe considerable uncertainty decrease in the delivered dose for 30 fractions.

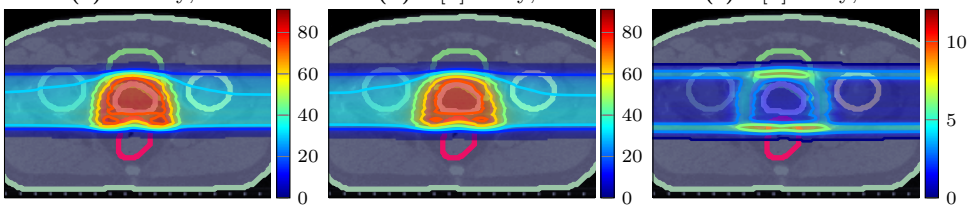
3. Validation of APM moment calculations

3.1. Methods

While Bangert et al. (2013) validated the Gaussian approximation of the proton depth dose component, the accuracy of pencil beam algorithms is generally limited in the presence of pronounced density heterogeneities in the beam path – for particles this is especially reflected around the Bragg peak region. However, our use of a POI ray-casting yields reasonable first order approximation of the Bragg peak distortions (Schaffner et al., 1999). For APM dose variance calculations there is, however, an implicit additional approximation that may compromise accuracy: As the calculation

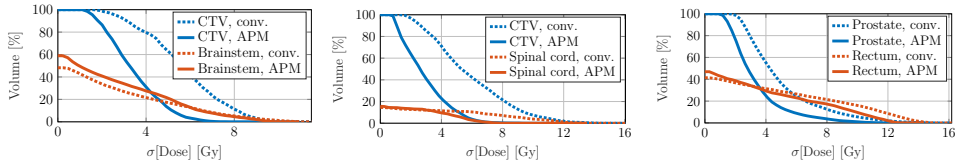


(a) $\boldsymbol{d}$ in Gy, conv. **(b)** $\mathrm{E}[\boldsymbol{d}]$ in Gy, conv. **(c)** $\sigma[\boldsymbol{d}]$ in Gy, conv.

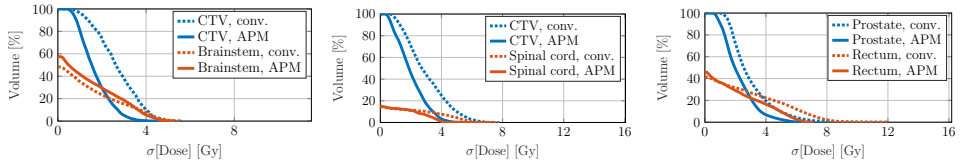


(d) $\boldsymbol{d}$ in Gy, APM opt. **(e)** $\mathrm{E}[\boldsymbol{d}]$ in Gy, APM opt. **(f)** $\sigma[\boldsymbol{d}]$ in Gy, APM opt.

Figure 1: Exemplary axial slice for a prostate plan displaying $\boldsymbol{d}$, $\mathrm{E}[\boldsymbol{d}]$ and $\sigma[\boldsymbol{d}]$ for a conventional optimization (a-c) and a probabilistic optimization (d-f) for $F = 5$ fractions.



(a) intra-cranial, 1 fraction **(b)** para-spinal, 1 fraction **(c)** prostate, 1 fraction



(d) intra-cranial, 30 fractions **(e)** para-spinal, 30 fractions **(f)** prostate, 30 fractions

Figure 2: Standard deviation volume histograms for a respective target (blue) and an OAR (orange) of all three patient cases. Curves for a conventional plan (dotted) and a probabilistic plan (solid) are compared for 1 fraction (a-c) and 30 fractions (d-f)

of the radiological depth is limited by the POI ray-cast in the nominal scenario, potential range changes induced by lateral positioning errors are neglected.

Here, we investigate the accuracy of this approximation in comparison to sampling based approaches, where radiological depth and the dose influence are explicitly recomputed for each sampled scenario. Multivariate random sampling with large sample number n_S serves as a benchmark. We use the global γ-criterion (Low et al., 1998), providing a distance-to-agreement metric between two distributions, for validation and comparison of the moment computations and measure their bias.

3.1.1. Random sampling benchmark The number n_S of random samples has to be chosen high enough to ensure a sufficiently accurate result for the sample mean $\bar{\boldsymbol{d}}$ and standard deviation $\sigma[\boldsymbol{d}]$ of dose. For the sample mean this is given by its standard

error $\sigma\left[\bar{\boldsymbol{d}}\right] = \sigma\left[\boldsymbol{d}\right] / \sqrt{n_S}$. The error $\sigma\left[\sigma\left[\boldsymbol{d}\right]\right]$ of the sample standard deviation $\sigma[\boldsymbol{d}]$ is obtained by the approximation $\sigma\left[\sigma\left[\boldsymbol{d}\right]\right] \approx \sigma\left[\boldsymbol{d}\right] / \sqrt{2\left(n_S - 1\right)}$ valid for large n_S (Squires, 2001). We use $n_S = 5000$ samples, where the expected relative error of the standard deviation drops below $\approx 1\,\%$ and the standard error of the mean drops below $\approx 1.4\,\%$ of $\sigma[\boldsymbol{d}]$.

To restrict computations for sampling, we only use non-fractionated treatment ($F = 1$) for the validation. Also, we use a "beam-wise" correlation model where all pencil beam uncertainties are perfectly correlated within the same beam, and uncorrelated otherwise, enabling use of simple spatial shifts and radiological depth scaling for the individual beams.

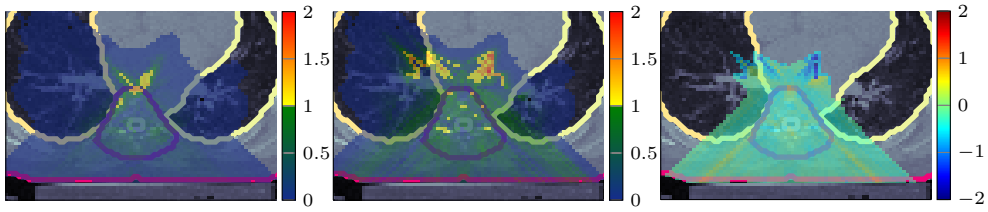
3.2. Importance sampling in the beam space

To decrease the number of samples, importance sampling is often used instead of random sampling, where the samples are chosen systematically from a grid in a multivariate uniform distributed error space and weighted with the probability density of the corresponding uncertainty model. In random sampling, the multivariate space is covered inherently, since the sample is selected from the full multivariate distribution. Multivariate importance sampling using grids, however, is not trivial; coverage of the complete multivariate space requires the evaluation of combinations of the scenarios along each independent grid dimension.

In case of independent beam shifts or, more general, uncertainty models that implement non-perfect correlation assumptions, the amount of combinations quickly becomes infeasible for sampling, and some combinations need to be neglected. In the presented case, the uncertainty space is composed of n_B independent beams from which each can be shifted or scaled along $d = 3$ independent dimensions. The number of scenarios per dimension is denoted per n_P. For $n_P = 3$, we investigate the influence of neglecting spatial and beam-wise combinations completely ($n_S = n_P \cdot d \cdot n_B - (d-1)$, labeled $\mathrm{W^I}$ from here on), using only spatial combinations ($n_S = n_P^d \cdot n_B$, $\mathrm{W^{II}}$) and sampling spatial and beamwise combinations ($n_S = n_P^{d \cdot n_B}$, $\mathrm{W^{III}}$), on accuracy of the resulting central moments and its influence on computation time. Applying $\mathrm{W^{III}}$ on the para-spinal case with $n_B = 3$, however, would already require $n_S \approx 2 \times 10^4$, and will therefore not be evaluated.

3.3. Results

Figure 3 shows $\gamma_{3\,\%/3\,\mathrm{mm}}$-index calculations for all cases of $\mathrm{E}\left[\boldsymbol{d}\right]$ and $\sigma[\boldsymbol{d}]$ of APM and the sample mean and standard deviation of the importance sampling strategies $\mathrm{W^{I\text{-}III}}$ with respect to the benchmark values gained from randomly sampling $n_S = 5000$ treatment scenarios, as well as the mean error to quantify bias. γ-distributions and an absolute difference map from the para-spinal case were selected for illustration, since this case showed the most significant difference; violations of the γ-criterion are mostly clustered in regions, where the analytical calculation of pencil-beams passing lung-tissue was cut off in depth. In most cases the mean error indicates a systematic underestimation of the standard deviation, which is especially pronounced for the more sparse importance sampling strategies $\mathrm{W^{I,II}}$. For APM, the underestimation can be mainly attributed to cut-off effects from the analytical dose engine, while for importance sampling this directly results from neglecting (parts of) the multivariate nature of the underlying uncertainty model. The expectation value/sample mean of



(a) γ-distribution for $\mathrm{E}[\boldsymbol{d}]$ **(b)** γ-distribution for $\sigma[\boldsymbol{d}]$ **(c)** difference in [Gy] for $\sigma[\boldsymbol{d}]$

patient & moment		γ pass rates [%]				mean error / bias [Gy]			
		APM	$\mathrm{W^{I}}$	$\mathrm{W^{II}}$	$\mathrm{W^{III}}$	APM	$\mathrm{W^{I}}$	$\mathrm{W^{II}}$	$\mathrm{W^{III}}$
intra-cranial	$\mathrm{E}[\boldsymbol{d}]$	100.0	99.8	99.9	100.0	−0.06	0.05	0.02	−0.03
	$\sigma[\boldsymbol{d}]$	99.9	57.5	83.4	84.7	−0.14	−1.00	−0.48	0.31
para-spinal	$\mathrm{E}[\boldsymbol{d}]$	98.4	96.3	96.0	–	−0.24	−0.05	−0.05	–
	$\sigma[\boldsymbol{d}]$	94.2	13.4	25.3	–	−0.20	−1.67	−1.17	–
prostate	$\mathrm{E}[\boldsymbol{d}]$	100.0	99.0	99.7	100.0	−0.20	0.03	0.01	0.03
	$\sigma[\boldsymbol{d}]$	99.3	59.2	72.7	89.9	−0.11	−1.44	−0.74	0.31

(d) γ pass rates and mean absolute differences for APM and importance sampling

Figure 3: γ-index distributions of $\mathrm{E}[\boldsymbol{d}]$ (a) and $\sigma[\boldsymbol{d}]$ (b) of APM compared to random sampling for the para-spinal case. For both $\mathrm{E}[\boldsymbol{d}]$ and $\sigma[\boldsymbol{d}]$ the global γ-criterion 3 % / 3 mm is used. (c) shows the absolute difference of APM's $\sigma[\boldsymbol{d}]$ to random sampling. (d) adds comparison to importance sampling ($\mathrm{W^{I\text{-}III}}$) and shows γ pass rates and the mean error / bias for all three cases.

dose is independent of the correlation model, and therefore accurately represented by importance sampling as well as by APM.

Validation of APM with a more strict γ-criterion, i. e. 2 %/2 mm, led to slightly lower γ pass rates (99.9 %, 96.9 % and 99.6 % for intra-cranial, para-spinal and prostate case) for $\mathrm{E}[\boldsymbol{d}]$, and a decrease to 98.5 %, 82.4 % and 91.3 % for $\sigma[\boldsymbol{d}]$. Those pass rates, however, need to be critically interpreted under consideration of data resolution (since no interpolation was used for γ-calculation) and, especially for $\sigma[\boldsymbol{d}]$, the standard error of the sample standard deviation ($\approx 1\,\%$) and mean ($\approx 1.4\,\%$ of $\sigma[\boldsymbol{d}]$).

4. Computational performance of APM

4.1. Methods

For calculation of the variance influence elements $\mathcal{V}_{ijim}$, we use a voxel-based approach to avoid finding the voxels belonging to a single pencil beam combination jm. To enable an efficient voxel-based variance influence calculation, we cache the respective pencil beam-specific voxel properties from the $\mathcal{D}_{ij}$-calculations. Even though the computation of $\mathcal{D}$ still underlies the same complexity ($\mathcal{O}(n_b n_v)$ with number of pencil beams n_b and number of voxels n_v), caching and uncertainty evaluation increases the run-time compared to D. The computation of the variance influence elements $\mathcal{V}_{ijim}$, however, is of complexity $\mathcal{O}(n_v n_b^2)$ and thus squared in number of pencil beams n_b. As the computation of $\mathcal{V}$ utilizes the same bivariate Gaussian kernel for each element, it is especially suitable for parallelization. We implemented a CPU and a GPU algorithm to investigate speed-up. A more detailed description of both implementations is given in Appendix B.

4.1.1. Fast approximations via sampling Both variance and the integral variance influence matrix are computed by inner products of $\mathcal{V}$ with $\boldsymbol{w}$ (eqs. (5) and (7)), resulting in a large sum along one or more tensor dimensions. These sums can be estimated by taking random samples from the respective sum. The following paragraphs detail these sampling processes for computing approximations of (1) the standard deviation distribution $\sigma[\boldsymbol{d}]$ or (2) the optimization matrix Ω.

Approximating $\sigma[\boldsymbol{d}]$ For estimating the standard deviation of dose $\sigma[d_i]$ in a voxel i, we define a set JM containing all pencil beam combinations where $j > m$ (since $\mathcal{V}_{ijim} = \mathcal{V}_{imij}$). First, eq. (5) is approximated by evaluating all summands where $j = m$ (since they are consistent along any correlation assumptions and the evaluation is only of linear complexity), and estimate the contribution from the other summands JM by randomly sampling a subset $\widehat{JM}$. The estimated variance $\hat{\sigma}^2[d_i]$ in voxel i is then given by

$$\hat{\sigma}^2[d_i] = \sum_j w_j w_j \mathcal{V}_{ijij} + 2 \cdot \frac{n_{JM}}{n_{\widehat{JM}}} \sum_{jm \in \widehat{JM}} w_j w_m \mathcal{V}_{ijim} \tag{10}$$

where n_{JM} and $n_{\widehat{JM}}$ are the number of combinations in the population and the sample, respectively. Due to sampling we might have negative results for $\hat{\sigma}^2[\boldsymbol{d}]$ which are projected to zero before computing the approximate standard deviation $\hat{\sigma}[\boldsymbol{d}] = \sqrt{\hat{\sigma}^2[\boldsymbol{d}]}$. The noise in $\hat{\sigma}[\boldsymbol{d}]$ may be reduced by application of a denoising filter. In this work, we heuristically chose the total variation denoising implemented in MITK‡ with only 2 iterations and $\lambda = 0.1$.

Theoretically, computation of an estimate of Ω is feasible via importance sampling in parallel. We choose, however, not to elaborate this issue since the random selection of index combinations jm does not guarantee the estimate of Ω to be positive semi-definite, which is required for optimization.

Approximating Ω We can estimate the Ω-matrix corresponding to a VOI, or more general a voxel population I by only evaluating a sample $\hat{I}$ from the sum over all voxels $\Omega^I = \sum_{i \in I} p_i \mathcal{V}_{ijim}$. The estimated matrix $\hat{\Omega}^I$ is then given by

$$\hat{\Omega}^I = \frac{n_I}{n_{\hat{I}}} \sum_{i \in \hat{I}} p_i \mathcal{V}_{ijim} \tag{11}$$

where n_I and $n_{\hat{I}}$ are the number of voxels in the population and the sample, respectively. By choosing adequate voxel populations, we can vary the sample size across different VOIs or VOI types (i. e. targets and OARs). We investigate the difference between the resulting plans computed with Ω and $\hat{\Omega}$ as well as the performance gain depending on sample size.

4.2. Results

4.2.1. Performance of APM We evaluate the performance of probabilistic planning workflows with APM on a desktop machine§ (PC) with 8 parallel threads and 1 GPU and a workstation‖ (WS) with 32 parallel threads and 2 GPUs.

‡ `http://docs.mitk.org/2016.03/classitk_1_1TotalVariationDenoisingImageFilter.html`
§ Windows 7, Intel® Core™ i7-2600 CPU @ 3.4 GHz, nVidia GTX 970
‖ 2 × Intel® Xeon® E5-2687W0 @ 3.10 GHz, 2 × nVidia Tesla K20

Runtimes Table 1 summarizes the run-time analysis for full APM calculations (without sampling approximations) on all three patient cases for a fractionated treatment (as presented in figure 1).

patient & machine		$\boldsymbol{d}$	$\mathrm{E}[\boldsymbol{d}]$	$\sigma[\boldsymbol{d}]$ (Ω)				$\mathcal{F}$ opt.	$\mathrm{E}[\mathcal{F}]$ opt.
				CPU		GPU			
intra-cranial	PC	10.1	32.1	490.0	(75.2)	85.0	(14.7)	17.6	9.6
	WS	8.9	32.0	117.8	(17.7)	63.6	(11.7)	15.9	9.0
para-spinal	PC	16.2	40.3	1173.2	(512.9)	223.9	(10.8)	30.1	47.1
	WS	13.8	38.7	269.0	(127.8)	176.5	(83.8)	27.5	33.2
prostate	PC	31.2	111.2	1492.1	(472.7)	202.1	(75.3)	41.8	33.7
	WS	29.6	101.7	364.2	(116.2)	155.7	(58.8)	37.9	28.9

Table 1: Runtimes (in seconds) for nominal and probabilistic dose/moment calculation and optimization for fractionated treatment plans of all three patient cases. The table differentiates between PC and Workstation (WS) run-times and between CPU and GPU implementations for the full $\sigma[\boldsymbol{d}]$-calculation and the Ω-matrix calculation without normal tissue for optimization (in parentheses).

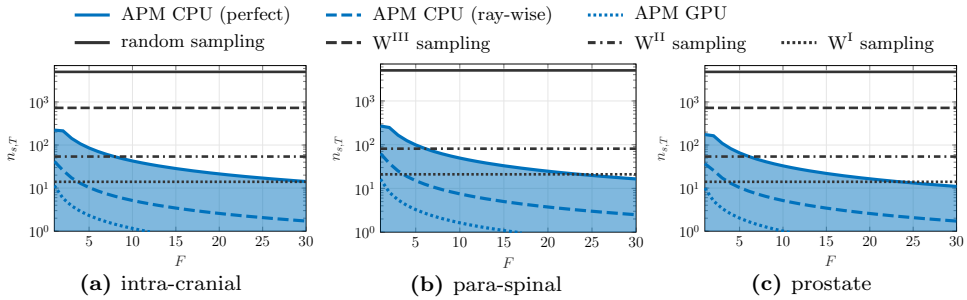
The computations underlying $\sigma[\boldsymbol{d}]$ and Ω are in principle identical. However, as we neglect unclassified normal tissue for optimization, we see a reduction in computation time by approximately a factor three to eight. This does not directly correspond to the reduction factor in number of voxels, since target voxels may exhibit many more pencil-beam contributions (and thus combinations), making those voxels more computationally demanding than others.

Ω- and $\sigma[\boldsymbol{d}]$-calculations prove to scale very well between PC and WS; their ratio corresponds roughly to the ratio of number of threads times ratio of CPU clock. The GPU code is up to a magnitude faster than the CPU code, while the speed-up is more pronounced on the PC, despite using two older high-performance GPUs on the WS. Due to the older architecture, this also indicates reasonable scaling of the GPU code.

Runtime differences between PC and WS for $\boldsymbol{d}$ and $\mathrm{E}[\boldsymbol{d}]$ and for optimizations are neglectable, since less distinguished effort was made to tune the respective algorithms. Comparing optimization of $\mathcal{F}$ and $\mathrm{E}[\mathcal{F}]$ run-times range in the same time scale, while $\mathrm{E}[\mathcal{F}]$ optimization showed, in terms of iterations, faster convergence with the same convergence criteria. Since $\mathcal{D}$ has about 25 % more non-zero elements than the nominal D matrix, time for objective function evaluations increase accordingly. A single $\boldsymbol{w}^T\Omega^{\mathrm{VOI}}\boldsymbol{w}$-matrix product adds about additional 20 % (of time needed for a $D\boldsymbol{w}$ product) per objective function and gradient evaluation, depending on the number of spots.

Performance in the context of fractionation compared to sampling In table 1 run-times are shown for a fractionated treatment ($F > 1$) with APM. On the CPU, this corresponded to an average overhead of 75 % to computations with $F = 1$, on the GPU the difference was neglectable. Figure 4 compares run-times of full APM calculations to the number of dose calculations needed for the sampling strategies depending on F. Using our ray-wise correlation model, APM run-time on the CPU equals 30 to 40 dose calculations on average. Applying the beam-wise perfect correlation model (as in the validation computations) the run-time rises approximately tenfold. In all cases, the most accurate sampling computations obtained by the random benchmark or $\mathrm{W}^{\mathrm{III}}$ are clearly outperformed. The GPU implementation was faster than any investigated

sampling set-up and is independent of the correlation model.



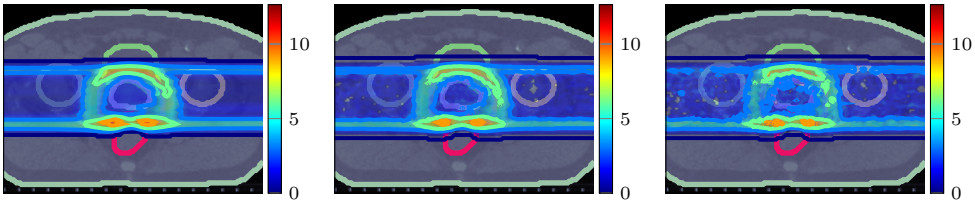
(a) intra-cranial **(b)** para-spinal **(c)** prostate

Figure 4: Run-time equivalent curves for APM and sampling strategies. A point $(F, n_{s,T})$ in the diagram corresponds to the run-time of calculating $n_{s,T} \cdot F$ dose distributions, corresponding to sampling $n_{s,T}$ treatments with F fractions. Hence the filled blue area contains all $n_{s,T}$ and F combinations that can be computed faster than an APM $\mathrm{E}[\boldsymbol{d}] + \sigma[\boldsymbol{d}]$ CPU calculation with a beam-wise perfect correlation model (solid blue line). The dashed blue line shows the run-time equivalent for the ray-wise correlation model. The GPU break-even is displayed as dotted blue line and independent of the correlation model. Since APMs moment calculations are independent of F, its run-time equivalent curves decrease with F, while the sampling strategies realize as horizontal (black) lines.

4.2.2. Quality of approximated σ-distributions We estimated standard deviation distributions $\sigma[\boldsymbol{d}]$ for different beamlet combination sample sizes with eq. (10). We present exemplary transversal slices for the prostate case from the respective $\sigma[\boldsymbol{d}]$-distributions in figures 5a to 5c to illustrate the increasing noise. Root-mean-square error (RMSE) and $\gamma_{3\,\%/3\,\mathrm{mm}}$ pass rates of the sampled distributions with respect to the exact calculations for all cases are shown in figure 5d. The noise, represented by the RMSE, increases with decreasing sample size, as expected, reaching values of $\approx 3\,\%$ of the maximum standard deviation. γ pass rates are comparable or better than the ones achieved with importance sampling approaches in section 3.3. Before application of the total variation filter, RMSE was approximately three times higher, with 3–10 % smaller γ pass rates.

4.2.3. Quality of probabilistic plans optimized with approximated Ω-matrix Figure 6 examines the differences between plans resulting from probabilistic optimization with an exactly computed Ω-matrix to the ones achieved with an estimated $\hat{\Omega}$-matrix via voxel sampling as described in eq. (11). While the sampling error in the estimated $\hat{\Omega}$ is random in nature, each realization of $\hat{\Omega}$ systematically influences subsequent probabilistic optimization. The resulting variations in $\boldsymbol{w}$ manifest as lateral difference "streaks" in figures 6a to 6c. For the prostate case, variations up to $\pm 2\,\mathrm{Gy}$ in $\boldsymbol{d}$ (less present in $\mathrm{E}[\boldsymbol{d}]$ and $\sigma[\boldsymbol{d}]$) occured across multiple sampling runs of $\hat{\Omega}$. Yet global plan variability is maintained, as seen in the SDVHs in figures 6d to 6e. Since $\sigma[\boldsymbol{d}]$ is not available from the voxel sampling, it needed to be recomputed for analysis.

4.2.4. Performance of the sampling methods We evaluated the performance of the approximation methods for the CPU implementation on the PC by linear regression, since in the optimal case relative sample size s should depend on relative run-time t via $s = mt + c$ with $m = 1$ and $c = 0\,\%$. For approximation of Ω with voxel sampling,

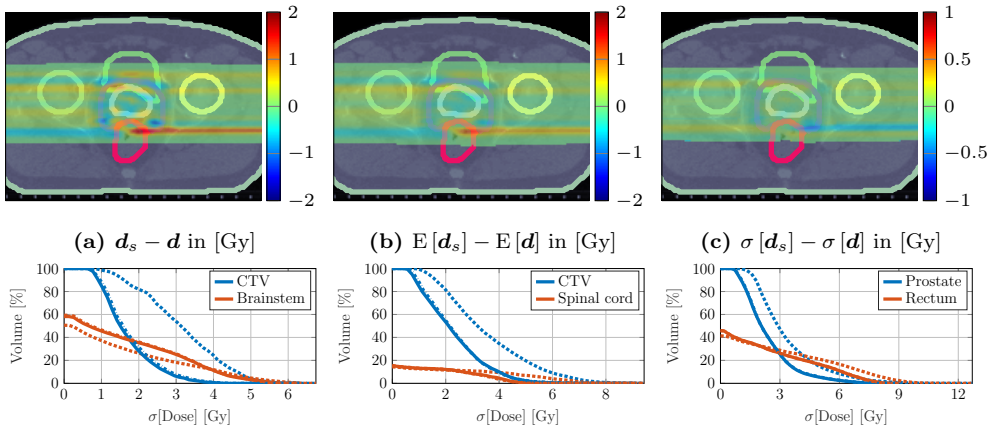


(a) exact $\sigma\,[\boldsymbol{d}]$ **(b)** $\sigma\,[\boldsymbol{d}]$, $s = 25\,\%$ **(c)** $\sigma\,[\boldsymbol{d}]$, $s = 5\,\%$

patient	$s = 50\,\%$		$s = 25\,\%$		$s = 10\,\%$		$s = 5\,\%$	
	RMSE	γ	RMSE	γ	RMSE	γ	RMSE	γ
intra-cranial	0.16	95.5	0.17	95.4	0.27	95.0	0.36	94.7
para-spinal	0.16	91.2	0.27	85.5	0.42	73.1	0.57	67.3
prostate	0.31	94.3	0.37	90.2	0.48	85.4	0.58	81.0

(d) RMSE [Gy] and γ pass rates for approximated $\sigma\,[\boldsymbol{d}]$-distributions of all cases

Figure 5: $\sigma[\boldsymbol{d}]$ distributions from exact (a) and sampled (b-c) calculations for the prostate case at relative sample sizes s. (d) compares to exact calculation for all patient cases depending on s.



(a) $\boldsymbol{d}_s - \boldsymbol{d}$ in [Gy] **(b)** $\mathrm{E}\,[\boldsymbol{d}_s] - \mathrm{E}\,[\boldsymbol{d}]$ in [Gy] **(c)** $\sigma\,[\boldsymbol{d}_s] - \sigma\,[\boldsymbol{d}]$ in [Gy]

(d) intra-cranial **(e)** para-spinal **(f)** prostate

Figure 6: Difference (in Gy) of nominal dose (a), expected dose (b) and standard deviation (c) between a probabilistic prostate plan $\boldsymbol{d}$ optimized with exactly computed Ω-matrix and a probabilistic plan $\boldsymbol{d}_s$ optimized with an estimated $\hat{\Omega}$-matrix. The relative sample size was 1.6 % and $F = 5$ fractions were used. (d-f) show corresponding SDVHs for all three patients optimized with Ω (solid) and with $\hat{\Omega}$ (dash-dotted) compared to the conventional plan (dotted).

we found nearly linear decrease ($m = 0.99$) with minimum overhead ($c = 1.02\,\%$) averaged over all three cases. The pencil-beam combination sampling shows more significant run-time overhead ($c = 7.33\,\%$) and therefore also suboptimal decrease ($m = 0.93$). The lowest tested relative sample size for the Ω-estimation was 1.6 %, leading to absolute run-times of a few seconds.

We omit detailed analysis on GPU and WS since the sampling implementation does not affect interior APM computations and should thus scale similarly on the WS and, for voxel sampling of the Ω-matrix, also on the GPUs.

5. Discussion

5.1. Accuracy

Section 3 with figure 3 in particular show that APM accurately propagates the uncertainty to expectation value and standard deviation of dose. For the head-case and the prostate case, γ pass rates for expectation value and standard deviation of dose were almost 100 %. Only for the para-spinal case, the global gamma pass-rate dropped to 94 %, with violations of the criterion mainly in low-dose regions outside the target. For the standard deviation calculation, APM outperformed any applied importance sampling approach, which was only able to reproduce accurate expectation value of dose and underestimated the standard deviation significantly when neglecting combinations of errors in the multivariate space. Due to the otherwise too large number of samples, we did only use three scenarios per dimension. Using more scenarios per dimension would improve accuracy, but not reduce the bias for $W^{\text{I-II}}$.

Arguably, the sample numbers do not escalate the same way when shifts in the patient coordinate frame are used. For example, Perkó et al. (2016) showed that only a few hundred dose calculations are sufficient to construct an accurate meta-model of dose for a fractionated treatment with polynomial chaos expansion by numerical integration. It is however debatable if a low-dimensional uncertainty model in the patient coordinate frame is representative in modern clinical systems combining gantries with robotic patient tables with many degrees of freedom. In these scenarios, APM facilitates accurate impact analysis of different uncertainty models on conventional and probabilistic planning through its ability to allow high-dimensional linearly correlated multivariate random variables as input.

Our study examined the accuracy of the uncertainty propagation with respect to the underlying pencil beam dose model, which could be refined further by using multiple lateral Gaussian components (Parodi et al., 2013) up to fine-/sub-sampling methods using a tight lateral grid of narrow Gaussian components (Soukup et al., 2005). Extension to non-Gaussian input uncertainties could also be approached by superimposing multiple Gaussians. Yet APM stays, contrary to most sampling based techniques, limited to the class of analytically approximable dose models. Thus further validation, preferably by sampling Monte-Carlo engines, has to be performed to weigh accuracy of the dose model against accuracy of uncertainty propagation, apart from run-time issues discussed in section 5.3.

5.2. Performance

While APMs computation of the expected dose influence matrix is of same complexity as a nominal dose influence matrix computation, the variance calculation of APM is of squared complexity in the number of pencil beams, making it is sensitive to the corresponding plan parameters. APM proved, however, to be well suited for parallelization, leading to run-times within a few minutes in all three presented cases with comparably cheap desktop GPUs, while also maintaining scalability to high-end hardware in workstations (table 1). Once variance influence matrix are available, they enable probabilistic optimization at minimum overhead compared to conventional optimization. Thus, all computations for a probabilistic planning workflow can be carried out in minutes (up to 30 min for the in our study most demanding prostate case on the PC, and in less than 10 min on a workstation). As a consequence, the presented run-times are not only competitive in a research environment but also

regarding clinical systems, where calculating hundreds or thousands of samples might not be applicable.

Additionally, APM efficiently incorporates fractionation effects, outperforming even the most sparse importance sampling approach considered for treatments with more than approximately 4 fractions, as shown in figure 4. APM can directly map the random and systematic errors of the input space to the expectation value and variance of dose for any given number of fractions, without the need of neglecting systematic or random components (i. e. Lowe et al., 2016) or sampling fractions (i. e. Park et al., 2013). Technically, models of the complete fractionation spectrum of $\sigma\,[\boldsymbol{d}]$ and Ω could be calculated in a single pass with minimum overhead, making APM an interesting tool to investigate the dose uncertainty depending on the fractionation schemes, or perform probabilistic optimizations for several fractionation schemes at low computational cost.

5.3. Accuracy to run-time trade-off

Uncertainty propagation and probabilistic or robust treatment plan optimization inevitably define a trade-off between speed and accuracy. In an ideal world, we would simply generate a sufficiently large number of random Monte Carlo samples within seconds (here we used 5000 to achieve a 1 % standard error of the standard deviation for a single fraction treatment). While this may be feasible in clouds or high-performance computing centers, on available desktop machines or workstations in the clinic and common research environments, other solutions at decreased accuracy must be accepted. Even for conventional treatment plan optimization, Monte Carlo computations are often considered too costly and optimization is performed with dose influence computed by pencil-beam approximations.

In contrast to sampling approaches, APM introduces a new layer of abstraction, which may complicate its application within established systems. Yet it broadens flexibility for planning and analysis within a high-dimensional multivariate space of linearly correlated input random variables. APM showed better trade-off in accuracy and speed than importance sampling approaches, often outperforming them in both quantities, especially in fractionated treatments. Further, APM computations are not subject to statistical variations compared with the results from sampling techniques.

With the introduced sub-sampling methods for standard deviation and variance influence estimation, different trade-offs can be explicitly defined depending on application and performance of the available system. Approximated $\sigma\,[\boldsymbol{d}]$-distributions were noisy at suboptimal run-time benefit and may be useful for fast qualitative analysis, but not for quantitative study. For optimization with approximated Ω-matrices, however, figure 6 and section 4.2.4 showed that even with low-accuracy approximations nearly similarly robust plans would be generated within $\approx 2\,\%$ of the run-time needed for an exact calculation. Random differences across resulting plans could be minimized or controlled by using regionally weighted sample sizes or other, e. g. grid-based, voxel selection schemes. The potential of using very sparse voxel information for estimating variance contributions also needs further investigation; it relativizes the importance of variance information with respect to the expectation value, which APM provides at comparably low computational cost, in probabilistic optimization. Therefore, sampling approximations might also prove valuable regarding the application of APM for more complex and clinically relevant objective functions, e. g. based on dose-volume points or equivalent uniform dose. These objectives,

however, may require the solution and/or approximaiton of novel Gaussian integrals that also depend on covariance elements of $\mathcal{V}$, which is part of ongoing research in our group.

6. Conclusion

We have evaluated accuracy and performance of a robustness analysis and probabilistic planning tool using analytical probabilistic modeling for three patient datasets. The expectation value $\mathrm{E}[\boldsymbol{d}]$ and standard deviation $\sigma[\boldsymbol{d}]$ calculations were benchmarked with 5000 random samples, showing γ pass-rates between 94.2 % and 100.0 %. We showed that APM is suited for fast probabilistic planning work-flows with run-times of ≈1–3.5 min by exploiting sampling approximations. During all APM computations, random and systematic errors in fractionated treatment could be efficiently incorporated at constant time complexity in the number of fractions. The results showed a competitive accuracy to run-time trade-off for research and clinical use cases.

Acknowledgements

The authors would like to thank Ralf Floca, Caspar Goch, Stefan Kislinskiy and Hendrik Heinrich for their support with the implementation of the MITK research planning plugin for APM. This work was supported by the German Research Foundation, Grant No. BA 2279/3-1.

References

Bangert, M. et al. (2013). "Analytical probabilistic modeling for radiation therapy treatment planning." In: *Phys. Med. Biol.* 58.16, pp. 5401–19.

Bortfeld, T. et al. (2004). "Effects of motion on the total dose distribution". In: *Semin. Radiat. Oncol.* 14.1, pp. 41–51.

Chen, W. et al. (2012). "Including robustness in multi-criteria optimization for intensity-modulated proton therapy". In: *Phys. Med. Biol.* 57, pp. 591–608.

Chu, M. et al. (2005). "Robust optimization for intensity modulated radiation therapy treatment planning under uncertainty". In: *Phys. Med. Biol.* 50.23, pp. 5463–77.

Dagum, L. and R. Menon (1998). "OpenMP: an industry standard API for shared-memory programming". In: *IEEE Comput. Sci. Eng.* 5.1, pp. 46–55.

Fredriksson, A. and R. Bokrantz (2016). "The scenario-based generalization of radiation therapy margins". In: *Phys. Med. Biol.* 61.5, pp. 2067–2082.

Fredriksson, A. et al. (2011). "Minimax optimization for handling range and setup uncertainties in proton therapy". In: *Med. Phys.* 38.3, p. 1672.

Herk, M. van et al. (2000). "The probability of correct target dosage: dose-population histograms for deriving treatment margins in radiotherapy". In: *Int. J. Radiat. Oncol.* 47.4, pp. 1121–1135.

Hong, L. et al. (1996). "A pencil beam algorithm for proton dose calculations". In: *Phys. Med. Biol.* 41.8, pp. 1305–1330.

Knopf, A.-C. et al. (2013). "Adequate margin definition for scanned particle therapy in the incidence of intrafractional motion". In: *Phys. Med. Biol.* 58.17, pp. 6079–6094.

Liu, W. et al. (2012). "Robust optimization of intensity modulated proton therapy". In: *Med. Phys.* 39.2, p. 1079.

Lomax, a. J. (2008a). "Intensity modulated proton therapy and its sensitivity to treatment uncertainties 1: the potential effects of calculational uncertainties." In: *Phys. Med. Biol.* 53.4, pp. 1027–1042.

— (2008b). "Intensity modulated proton therapy and its sensitivity to treatment uncertainties 2: the potential effects of inter-fraction and inter-field motions." In: *Phys. Med. Biol.* 53.4, pp. 1043–1056.

Low, D. A. et al. (1998). "A technique for the quantitative evaluation of dose distributions." In: *Med. Phys.* 25.5, pp. 656–61.

Lowe, M. et al. (2016). "Incorporating the effect of fractionation in the evaluation of proton plan robustness to setup errors". In: *Phys. Med. Biol.* 61.1, pp. 413–429.

Nickolls, J. et al. (2008). "Scalable parallel programming with CUDA". In: *Queue* 6.2, p. 40.

Oelfke, U. and T. Bortfeld (2001). "Inverse planning for photon and proton beams." In: *Med. Dosim.* 26.2, pp. 113–24.

Park, P. C. et al. (2013). "Statistical assessment of proton treatment plans under setup and range uncertainties". In: *Int. J. Radiat. Oncol. Biol. Phys.* 86.5, pp. 1007–1013.

Parodi, K. et al. (2013). "Monte Carlo-based parametrization of the lateral dose spread for clinical treatment planning of scanned proton and carbon ion beams". In: *J. Radiat. Res.* 54.Suppl 1, pp. 91–96.

Perkó, Z. et al. (2016). "Fast and accurate sensitivity analysis of IMPT treatment plans using Polynomial Chaos Expansion." In: *Phys. Med. Biol.* 61.12, pp. 4646–64.

Pflugfelder, D. et al. (2008). "Worst case optimization: a method to account for uncertainties in the optimization of intensity modulated proton therapy." In: *Phys. Med. Biol.* 53.6, pp. 1689–700.

Schaffner, B. et al. (1999). "Dose calculation models for proton treatment planning using a dynamic beam delivery system: an attempt to include density heterogeneity effects in the analytical dose calculation". In: *Phys. Med. Biol.* 44.1, pp. 27–41.

Siggel, M. et al. (2012). "Boosting runtime-performance of photon pencil beam algorithms for radiotherapy treatment planning". In: *Phys. Medica* 28.4, pp. 273–280.

Sobotta, B. et al. (2012). "Accelerated evaluation of the robustness of treatment plans against geometric uncertainties by Gaussian processes". In: *Phys. Med. Biol.* 57.23, pp. 8023–8039.

Soukup, M. et al. (2005). "A pencil beam algorithm for intensity modulated proton therapy derived from Monte Carlo simulations". In: *Phys. Med. Biol.* 50.21, pp. 5089–5104.

Squires, G. L. (2001). *Practical physics*. 4th ed. Cambridge University Press, p. 212.

Unkelbach, J. et al. (2009). "Reducing the sensitivity of IMPT treatment plans to setup errors and range uncertainties via probabilistic treatment planning." In: *Med. Phys.* 36.2009, pp. 149–163.

Wächter, A. and L. T. Biegler (2006). "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming". In: *Math. Program.* 106.1, pp. 25–57.

Wolf, I. et al. (2005). "The medical imaging interaction toolkit". In: *Med. Image Anal.* 9.6, pp. 594–604.

Appendix A. Gaussian kernels for analytical probabilistic modeling

Appendix A.1. Expectation value of dose influence $\mathcal{D}$

A matrix element $\mathcal{D}_{ij}$ describing the influence of pencil beam j on the expectation value of dose in voxel i is given by

$$\mathcal{D}_{ij} = \mathcal{L}^x_{ij}\mathcal{L}^y_{ij}\mathcal{Z}_{ij} \tag{A.1}$$

with the lateral components

$$\mathcal{L}^x_{ij} = \frac{1}{\sqrt{2\pi\left(\lambda^2_{ij} + \Sigma^x_{jj}\right)}} \cdot \mathrm{e}^{\frac{\left(x_{ij}-\mu^x_j\right)^2}{2\left(\lambda^2_{ij}+\Sigma^x_{jj}\right)}} \tag{A.2}$$

where λ^2_{ij} corresponds to the local variance of the Gaussian component describing the lateral dose fall off of pencil beam j in voxel i.

Σ^x is the covariance matrix of the pencil beam setup error in the spatial dimension x. The depth component is evaluated with the superposition of 10 Gaussian components and given by

$$\mathcal{Z}_{ij} = \sum_k \frac{\omega_{jk}}{\sqrt{2\pi\left(\delta^2_{jk} + \Sigma^z_{jj}\right)}} \cdot \mathrm{e}^{\frac{\left(z_{ij}-\mu^z_{jk}\right)^2}{2\left(\delta^2_{jk}+\Sigma^z_{jj}\right)}} \tag{A.3}$$

where now δ^2_{jk} and μ_{jk} are the variance and mean, respectively, of the k-th Gaussian depth component and for spot j.

Appendix A.2. Covariance of dose influence: $\mathcal{V}$

An element of the covariance influence tensor $\mathcal{V}_{ijlm}$ describing the influence of pencil beams j and m to the covariance in voxels i and l is given by

$$\mathcal{V}_{ijlm} = \Upsilon^x_{ijlm}\Upsilon^y_{ijlm}\Xi^x_{ijlm} - \mathcal{D}_{ij}\mathcal{D}_{lm} \tag{A.4}$$

with the lateral components

$$\Upsilon^x_{ijlm} = \frac{\mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}_{ijlm}-\boldsymbol{\mu}^x_{jm})^\top(\Lambda^{ijlm}+\Sigma^{jm})^{-1}(\boldsymbol{x}_{ijlm}-\boldsymbol{\mu}^x_{jm})}}{2\pi\sqrt{|\Lambda^{ijlm}+\Sigma^{jm}|}} \ . \tag{A.5}$$

where

$$\begin{aligned} \boldsymbol{x}_{ijlm} = \begin{pmatrix} x_{ij} \\ x_{lm} \end{pmatrix}, \ \boldsymbol{\mu}^x_{jm} = \begin{pmatrix} \mu^x_j \\ \mu^x_m \end{pmatrix}, \ \boldsymbol{\Delta}^x_{jm} = \begin{pmatrix} \Delta^x_j \\ \Delta^x_m \end{pmatrix}, \\ \Lambda^{ijlm} = \begin{pmatrix} \lambda^2_{ij} & 0 \\ 0 & \lambda^2_{lm} \end{pmatrix}, \ \Sigma^{jm} = \begin{pmatrix} \Sigma^x_{jj} & \Sigma^x_{jm} \\ \Sigma^x_{mj} & \Sigma^x_{mm} \end{pmatrix} \ . \end{aligned} \tag{A.6}$$

The depth component is given by

$$\Xi_{ijlm} = \sum_{kn} \omega_{jk}\omega_{mn} \ \frac{\mathrm{e}^{-\frac{1}{2}(\boldsymbol{z}_{ijlm}-\boldsymbol{\mu}^z_{jkmn})^\top(\Theta^{jkmn}+\Sigma^{jm})^{-1}(\boldsymbol{z}_{ijlm}-\boldsymbol{\mu}^z_{jkmn})}}{2\pi\sqrt{|\Theta^{jkmn}+\Sigma^{jm}|}} \tag{A.7}$$

where

$$\boldsymbol{z}_{ijlm} = \begin{pmatrix} z_{ij} \\ z_{lm} \end{pmatrix},\ \boldsymbol{\mu}^z_{jkmn} = \begin{pmatrix} \mu_{jk} \\ \mu_{mn} \end{pmatrix},\ \vec{\Delta}^z_{jm} = \begin{pmatrix} \Delta^z_j \\ \Delta^z_m \end{pmatrix}$$
$$\Theta^{jkmn} = \begin{pmatrix} \delta^2_{jk} & 0 \\ 0 & \delta^2_{mn} \end{pmatrix},\ \Sigma^{jm} = \begin{pmatrix} \Sigma^z_{jj} & \Sigma^z_{jm} \\ \Sigma^z_{mj} & \Sigma^z_{mm} \end{pmatrix}. \tag{A.8}$$

Appendix B. Details on the CPU and GPU implementations of APM

Appendix B.1. CPU implementation of the variance calculation

On the CPU, we exploit parallelization with OpenMP 2.0¶ (e. g. Dagum and Menon, 1998), by by distributing the individual voxels across multiple threads. The assignment of voxels to threads is handled dynamically, since the run-time per voxel differs based on the number of impinging pencil beams and their respective correlations. Additionally, we can exploit symmetry in $\mathcal{V}$ for evaluating eqs. (5) and (7) since $\mathcal{V}_{ijim} = \mathcal{V}_{imij}$.

Appendix B.2. GPU implementation of the variance calculation

For implementation on the GPU we use nVidia CUDA 7.5 (e. g. Nickolls et al., 2008). Contrary to the CPU implementation, for each voxel we sequentially start a kernel that evaluates all impinging spot combinations in parallel. While a kernel for a voxel is running, the data for the next voxel is prepared in host memory, leaving only the memory management overhead of a single memory copy operation per voxel to the device. If multiple devices are available, the voxels are in parallel dynamically distributed to the devices. The GPU implementation does not exploit correlation models but evaluates every single pair-wise pencil-beam combination influencing the variance in a voxel. Consequently, the runtime is independent from the underlying correlation model.

Appendix C. Patient data information

patient	intra-cranial	para-spinal	prostate
beam angles	60°, 120°	135°, 180°, 225°	90°, 270°
prescribed dose	60 Gy	60 Gy	70 Gy (76 Gy)
spot distance	3 mm	4 mm	5 mm
resolution	$(1.2 \times 1.2 \times 3)\,\mathrm{mm}^3$	$(3 \times 3 \times 3)\,\mathrm{mm}^3$	$(2 \times 2 \times 3)\,\mathrm{mm}^3$
#spots	1705	13274	6803
# eval. voxels σ	1.0×10^5	0.5×10^5	4.1×10^5
# eval. voxels Ω	0.7×10^4	2.0×10^4	6.9×10^4

Table C1: Information about size, complexity and planning parameters of the patient cases used. For the prostate case, the prescribed dose in parentheses applies to the target boost. The last two rows show the number of dose points used for the computation of the full σ-distributions and the Ω-matrices neglecting the unclassified normal tissue.

¶ Specifications by the OpenMP Architecture Review Board at `http://www.openmp.org/drupal/mp-documents/cspec20.pdf` (accessed 2016-22-11)