

Accelerated Multi-Criterial Optimization in Radiation Therapy
using Voxel-Wise Dose Prediction

by

Patrick James Jensen Jr.

Medical Physics Graduate Program
Duke University

Date: _____

Approved:

Q. Jackie Wu, Advisor

Qiuwen Wu

Fang-Fang Yin

Yaorong Ge

John Kirkpatrick

Dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Medical Physics Graduate Program
in the Graduate School of Duke University

2020

ABSTRACT

Accelerated Multi-Criterial Optimization in Radiation Therapy
using Voxel-Wise Dose Prediction
by

Patrick James Jensen Jr.

Medical Physics Graduate Program
Duke University

Date: _____

Approved:

Q. Jackie Wu, Advisor

Qiuwen Wu

Fang-Fang Yin

Yaorong Ge

John Kirkpatrick

An abstract of a dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Medical Physics Graduate Program
in the Graduate School of Duke University

2020

Copyright by
Patrick James Jensen Jr.
2020

Abstract

In external beam radiation therapy (EBRT) for cancer patients, it is highly desirable to completely eradicate the cancerous cells for the purpose of improving the patient's quality of life and increasing the patient's likelihood of survival. However, there can be significant side effects when large regions of healthy cells are irradiated during EBRT, particularly for organs-at-risk (OARs). Due to the juxtaposition of the cancerous and non-cancerous tissue, trade-offs need to be made between target coverage and OAR sparing during treatment planning. For this reason, the treatment planning process can be posed as a multi-criterial optimization (MCO) problem, which has previously been studied extensively with several exact solutions existing specifically for radiation therapy. Typical MCO implementations for EBRT involve creating, optimizing, and calculating many treatment plans to infer the set of feasible best radiation doses, or the Pareto surface. However, each optimization and calculation can take 10-30 minutes per plan. As a result, generating enough plans to attain an accurate representation of the Pareto surface can be very time-consuming, particularly in higher-dimensions with many possible trade-offs.

The purpose of this study is to streamline the MCO workflow by using a machine-learning model to quickly predict the Pareto surface plan doses, rather than exactly computing them. The primary focus of this study focuses on the development

and analysis of the dose prediction model. The secondary focus of this study is to develop new metrics for analyzing the similarity between different Pareto surface interpolations. The tertiary focus of this study is to investigate the feasibility of deliberately irradiating the epidural space in spine stereotactic radiosurgery (SRS), as well as estimate its potential effect on preventing tumor recurrence.

For the primary focus of this study, the model's architecture proceeds as follows. The model begins by creating an initial dose distribution via an inverse fit of inter-slice and intra-slice PTV distance maps on a voxel-wise basis. The model proceeds by extracting three sets of transverse patches from all structure maps and the initialized dose map at each voxel. The model then uses the patch vectors as inputs for a neural network which updates and refines the dose initialization to achieve a final dose prediction. The primary motivation behind our model is to use our understanding of the general shape of dose distributions to remove much of the nonlinearity of the dose prediction problem, decreasing the difficulty of subsequent network predictions. Our model is able to take the optimization priorities into account during dose prediction and infer feasible dose distributions across a range of optimization priority combinations, allowing for indirect Pareto surface inference.

The model's performance was analyzed on conventional prostate volumetric modulated arc therapy (VMAT), pancreas stereotactic body radiation therapy (SBRT), and spine stereotactic radiosurgery (SRS) with epidural space irradiation. For each of

these treatment paradigms, the Pareto surfaces of many patients were thoroughly sampled to train and test the model. On all of these cases, our model achieved good performance in terms of speed and accuracy. Overfitting was shown to be minimal in all cases, and dose distribution slices and dose-volume histograms (DVHs) were shown for comparison, confirming the proficiency of our model. This model is relatively fast (0.05-0.20 seconds per plan), and it is capable of sampling the entire Pareto surface much faster than commercial dose optimization and calculation engines.

While these results were generally promising, the model achieved lower error on the prostate VMAT treatment plans compared to the pancreas SBRT and spine SRS treatment plans. This is likely due to the existence of heavier beam streaks in the stereotactic treatment plans which are generated by a sharper control of the delivered dose distribution. However, the Pareto surface errors were similar across all three cases, so these dose distribution errors did not propagate to the Pareto objective space.

The secondary focus of this study is the development and analysis of Pareto surface similarity metrics. The dose prediction model can be used to rapidly estimate many Pareto-optimal plans for quick Pareto surface inference. This could allow for a potentially significant increase in the speed at which Pareto surfaces are inferred to provide treatment planning assistance and acceleration. However, previous investigations into Pareto surface analysis typically do not compare a ground truth Pareto surface with a Pareto surface prediction. Therefore, there is a need to develop a

Pareto surface metric in order to evaluate the ability of the model to generate accurate Pareto surfaces in addition to accurate dose distributions.

To address these needs, we developed four Pareto surface similarity metrics, emphasizing the ability to represent distances between the interpolations rather than the sampled points. The most straightforward metric is the root-mean-square error (RMSE) evaluated between matched, sampled points on the Pareto surfaces, augmented by intra-simplex upsampling of the barycentric dimensions of each simplex. The second metric is the Hausdorff distance, which evaluates the maximum closest distance between the sets of sampled points. The third metric is the average projected distance (APD), which evaluates the displacements between the sampled points and evaluates their projections along the mean displacement. The fourth metric is the average nearest-point distance (ANPD), which numerically integrates point-to-simplex distances over the upsampled simplices of the Pareto surfaces. These metrics are compared by their convergence rates as a function of intra-simplex upsampling, the calculation times required to achieve convergence, and their qualitative meaningfulness in representing the underlying interpolated surfaces. For testing, several simplex pairs were constructed abstractly, and Pareto surfaces were constructed using inverse optimization and our dose prediction model applied to conventional prostate VMAT, pancreas SBRT, and spine SRS with epidural irradiation.

For the abstract simplex pairs, convergence within 1% was typically achieved at approximately 50 and 100 samples per barycentric dimension for the ANPD and the RMSE, respectively. The RMSE and the ANPD required approximately 50 milliseconds and 3 seconds to calculate to these sampling rates, respectively, while the APD and HD required much less than 1 millisecond. Additionally, the APD values closely resembled the ANPD limits, while the RMSE limits and HD tended to be more different. The ANPD is likely more meaningful than the RMSE and APD, as the ANPD's point-to-simplex distance functions more closely represent the dissimilarity between the underlying interpolated surfaces rather than the sampling points on the surfaces. However, in situations requiring high-speed evaluations, the APD may be more desirable due to its speed, lack of subjective specification of intra-simplex upsampling rates, and similarity to the ANPD limits.

The tertiary focus of this study is the analysis of the feasibility of epidural space irradiation in spine SRS. The epidural space is a frequent site of cancer recurrence after spine SRS. This may be due to microscopic disease in the epidural space which is underdosed to obey strict spinal cord dose constraints. We hypothesized that the epidural space could be purposefully irradiated to prescription dose levels, potentially reducing the risk of recurrence in the epidural space without increasing toxicity. To address this, we sought to analyze the feasibility of irradiating the epidural space in spine SRS. Analyzing the data associated with this study is synergistic to our MCO acceleration

study, since the range of trade-offs between epidural space irradiation and spinal cord sparing represents an MCO problem which our dose prediction model may quickly solve.

Spine SRS clinical treatment plans with associated spinal PTV (PTV_{spine}) and spinal cord contours, and prior delivered dose distributions were identified retrospectively. An epidural space PTV ($PTV_{epidural}$) was contoured to avoid the spinal cord and focus on regions near the PTV_{spine} . Clinical plan constraints included PTV_{spine} constraints ($D_{95\%} = 1800$ cGy, $D_{5\%} < 1950$ cGy) and spinal cord constraints ($D_{max} < 1300$ cGy, $D_{10\%} < 1000$ cGy). Prior clinical plan doses were mapped onto the new $PTV_{epidural}$ contour for analysis. Plans were copied and revised to additionally target the $PTV_{epidural}$, optimizing $PTV_{epidural} D_{95\%}$ after meeting clinical plan constraints. Tumor control probabilities (TCPs) were estimated for the $PTV_{epidural}$ using a radiobiological linear-quadratic model of cell survival for both clinical and revised plans. Clinical and revised plans were compared according to their $PTV_{epidural}$ DVH distributions, $D_{95\%}$ distributions, and TCPs.

Seventeen SSRS plans were identified and included in this study. Revised plan DVHs demonstrated higher doses to the epidural low-dose regions, with $D_{95\%}$ improving from $10.96 \text{ Gy} \pm 1.76 \text{ Gy}$ to $16.84 \text{ Gy} \pm 0.87 \text{ Gy}$ ($p < 10^{-5}$). Our TCP modeling set the clinical plan TCP average to 85%, while revised plan TCPs were all greater than 99.99%. Therefore, irradiating the epidural space in spine SRS is likely feasible, and

purposefully targeting the epidural space in SSRS should increase control in the epidural space without significantly increasing the risk of spinal cord toxicity.

List of Contents

Abstract	iv
List of Contents	xi
List of Tables	xv
List of Figures	xvi
Acknowledgements	xix
1. Introduction	1
1.1 Radiation therapy	1
1.1.1 Stereotactic techniques.....	3
1.2 Treatment planning optimization	4
1.3 Machine learning	7
1.3.1 Traditional machine learning	7
1.3.2 Deep learning.....	8
1.4 Previous work on treatment planning acceleration and optimization	9
1.4.1 Multi-criterial optimization	9
1.4.2 Knowledge-based planning.....	14
1.4.3 Voxel-wise dose prediction.....	15
1.5 Spine SRS and its failure patterns	16
2. Study Objectives and Structure.....	19
3. Model design	22
3.1 Final iteration	22

3.1.1 Patch extraction	23
3.1.2 Dose initialization.....	24
3.1.3 Residual network.....	26
3.1.4 Model training	27
3.2 Earlier iterations.....	29
3.2.1 Neural network design.....	29
3.2.2 Regularization.....	31
3.2.3 Dose initialization fits	32
4. Pareto surface metrics	34
4.1 Introduction.....	34
4.2 Materials and methods	36
4.2.1 Similarity metrics between Pareto surfaces.....	36
4.2.1.1 Root-mean-square error	36
4.2.1.2 Hausdorff distance.....	44
4.2.1.3 Average projected distance	45
4.2.1.4 Average nearest-point distance	49
4.2.2 Theoretical Pareto surfaces for comparison.....	52
4.3 Results	66
4.3.1 Metric comparison and convergence analysis	66
4.3.2 Time analysis.....	74
4.4 Discussion.....	82
4.5 Conclusion.....	85

5. Evaluation of epidural space irradiation in spinal SBRT	86
5.1 Introduction.....	86
5.2 Materials and methods	87
5.3 Results	93
5.4 Discussion.....	98
5.5 Conclusion.....	100
6. Model application to prostate VMAT	101
6.1 Introduction.....	101
6.2 Materials and methods	102
6.3 Results	104
6.4 Discussion.....	108
6.5 Conclusion.....	112
7. Model application to pancreas SBRT	113
7.1 Introduction.....	113
7.2 Materials and methods	114
7.3 Results	116
7.4 Discussion.....	120
7.5 Conclusion.....	123
8. Model application to spine SRS with epidural space irradiation.....	124
8.1 Introduction.....	124
8.2 Materials and methods	125
8.3 Results	128

8.4 Discussion.....	132
8.5 Conclusion.....	135
9. Conclusions.....	136
References	140
Biography.....	145

List of Tables

Table 1: The number of simplex samples per barycentric dimension required to yield an ANPD within specific convergence thresholds of its limit.	73
Table 2: The number of simplex samples per barycentric dimension required to yield an RMSE within specific convergence thresholds of its limit.	73
Table 3: Time required to yield an ANPD within specific convergence thresholds of its limit.	81
Table 4: Time required to yield an RMSE within specific convergence thresholds of its limit.	81
Table 5: Summary of previous estimations of epidural space recurrence rates in spine SRS.....	93
Table 6: Pareto surface similarity metrics evaluated on the Pareto surfaces indirectly generated by our dose prediction model for prostate VMAT.	108
Table 7: Pareto surface similarity metrics evaluated on the Pareto surfaces indirectly generated by our dose prediction model for spine SRS.	120
Table 8: Pareto surface similarity metrics evaluated on the Pareto surfaces indirectly generated by our dose prediction model for spine SRS.	132

List of Figures

Figure 1: Graphical depiction of a Pareto surface in radiation therapy with two dose-volume objectives.....	10
Figure 2: Example of Pareto surface sampling in clinical implementations of MCO.	12
Figure 3: Pareto surface interpolation uncertainty generated by insufficient sampling. .	13
Figure 4: Overview of the dose prediction model architecture.....	22
Figure 5: A sample transverse prostate PTV contour (left) compared to its initialized dose (right).....	25
Figure 6: Graphical depiction of a residual block within the neural network.	27
Figure 7: Graphical depiction of the diagonal batch aggregation scheme used during model training.	28
Figure 8: Example dose distributions.....	35
Figure 9: Example visualization of distance differences between sampled points and the underlying surfaces.	38
Figure 10: Visualization of intra-simplex upsampling according to the simplex's barycentric coordinates.	40
Figure 11: Example visualization of the effect of intra-simplex upsampling on the distances.	41
Figure 12: Example of the importance of accurate matching between sets in RMSE calculations.....	43
Figure 13: Example results of Pareto surface comparison using the RMSE and APD metrics.	46
Figure 14: Case 1 for simplex testing: Distance from $\{(0, 1); (1, 0)\}$ to $\{(0.15, 1); (1, 0.3)\}$...	60
Figure 15: Case 2 for simplex testing: Distance from $\{(0, 1); (1, 0)\}$ to $\{(0.455, 0.655); (0.655, 0.455)\}$	61
Figure 16: Case 3 for simplex testing: Distance from $\{(0, 1); (1, 0)\}$ to $\{(0.6, 0.6); (1, 0.9)\}$..	62

Figure 17: Case 4 for simplex testing: Distance from $\{(1, 0, 0); (0, 1, 0); (0, 0, 1)\}$ to $\{(0, 1.1, -0.1); (1.05, 0.05, 0); (-0.05, 0, 0.85)\}$.	63
Figure 18: Case 5: Distance from $\{(1, 0, 0); (0, 1, 0); (0, 0, 1)\}$ to $\{(0.55, 0.55, 0.55); (1.55, 0.8, 0.8); (1.55, 0.8, 1.55)\}$.	64
Figure 19: Case 6 for simplex testing: Distance from $\{(1, 0, 0); (0, 1, 0); (0, 0, 1)\}$ to $\{(0.6, 0.6, 0); (0.6, 0, 0.6); (1, 0.3, 0.3)\}$.	65
Figure 20: Distance metrics for Case 1 as a function of number of samples per barycentric dimension.	67
Figure 21: Distance metrics for Case 2 as a function of number of samples per barycentric dimension.	68
Figure 22: Distance metrics for Case 3 as a function of number of samples per barycentric dimension.	69
Figure 23: Distance metrics for Case 4 as a function of number of samples per barycentric dimension.	70
Figure 24: Distance metrics for Case 5 as a function of number of samples per barycentric dimension.	71
Figure 25: Distance metrics for Case 6 as a function of number of samples per barycentric dimension.	72
Figure 26: Metric computation times for Case 1.	75
Figure 27: Metric computation times for Case 2.	76
Figure 28: Metric computation times for Case 3.	77
Figure 29: Metric computation times for Case 4.	78
Figure 30: Metric computation times for Case 5.	79
Figure 31: Metric computation times for Case 6.	80
Figure 32: An example transverse planar contour of the PTV_{spine} (red), PTV_{epidural} (blue), and spinal cord (green) contours.	89

Figure 33: Example dose distributions comparing a clinical plan (lower-left) to its corresponding revised plan (upper-right) and its anatomical contours (upper-left).....	94
Figure 34: PTV _{epidural} DVH distributions.	95
Figure 35: PTV _{epidural} D _{95%} distributions.....	96
Figure 36: Fitted values of α for each hypothetical value of clonogen density.....	97
Figure 37: Graph of dose map root-mean-square error for the training set (blue) and testing set (orange) as a function of number of iterations during model training.....	105
Figure 38: Comparison between dose distribution prediction and TPS calculation in plans prioritizing PTV HI or prioritizing rectum D _{25%}	106
Figure 39: Comparison between indirect DVH prediction and TPS calculation in plans prioritizing PTV HI or prioritizing rectum D _{25%}	107
Figure 40: Comparison between dose distribution prediction and TPS calculation in plans prioritizing PTV HI or prioritizing bowel D _{0.1cc}	118
Figure 41: Comparison between dose distribution prediction and TPS calculation in plans prioritizing PTV HI or prioritizing small bowel D _{0.1cc}	119
Figure 42: An example transverse planar contour of the PTV _{spine} (red), PTV _{epidural} (blue), and spinal cord (green) contours.	126
Figure 43: Comparison between dose distribution prediction and TPS calculation in plans prioritizing epidural space coverage or prioritizing spinal cord sparing.	130
Figure 44: Comparison between indirect DVH prediction and TPS calculation in plans prioritizing epidural space coverage or prioritizing spinal cord sparing.....	131

Acknowledgements

I would like to acknowledge and thank my advisor, Dr. Jackie Wu, for her invaluable guidance that accelerated my research and augmented my understanding of the clinical aspects of research in medical physics, for her suggestions for contacts outside of the Medical Physics Graduate Program which opened new possibilities for collaboration in research, and for her support and motivation throughout my time at Duke University. I sincerely appreciate her for being as excellent a mentor as any student could hope to have.

I would also like to thank Dr. John Kirkpatrick for assisting with project design and for lending me his insight and expertise in spinal stereotactic body radiation therapy. I would like to thank Dr. Qiuwen Wu for serving as my committee chair and for his constructive criticism regarding my research. I would like to thank Dr. Fang-Fang Yin and Dr. Yaorong Ge for their rigorous analysis of my work. I also thank Dr. Jiahua Zhang, Dr. Yang Sheng, Dr. Chunhao Wang, Dr. Kyle Lafata, Tianyi Xie, Yushi Chang, and Hunter Stephens for our informal academic discussions which have improved and fine-tuned my research. Finally, I would like to thank my professors and classmates in the Medical Physics Graduate Program for developing my technical and analytical expertise in medical physics, my professionalism, and my research capabilities.

1. Introduction

1.1 *Radiation therapy*

Cancer is a class of diseases characterized by aggressive cellular proliferation caused by malignant genetic mutations. Radiation therapy is a commonly used treatment option for managing many types of cancers. In external beam radiation therapy (EBRT), a beam of ionizing radiation strikes and penetrates through human tissue, dealing damage to both cancerous and non-cancerous cells at an atomic level. In EBRT, it is highly desirable to completely eradicate the cancerous cells for the purpose of improving the patient's quality of life and increasing the patient's likelihood of survival. However, there can be significant side effects when large regions of healthy cells are irradiated during EBRT. This is especially true for cells inside of an organ or structure which is particularly sensitive to radiation damage, such as the heart, lungs, and bowels (typically called "organs-at-risk", or "OARs"). The side effects of OAR irradiation can include heart disease, gastrointestinal dysfunction, chronic pain, fertility issues, and carcinogenesis of new cancerous cells. To mitigate these side effects, it is important to minimize the irradiation of healthy tissues and OARs while still maximizing the irradiation of the cancerous cells.

Following the advent of radiation therapy, many technologies have been developed which increase the precision with which radiation is delivered in order to more intensely irradiate the cancerous cells while avoiding the OARs and healthy

tissues. The extent of cancer can be localized by several imaging modalities, including computed tomography (CT), positron emission tomography (PET), and magnetic resonance imaging (MRI). On these images, digital contours are drawn which indicate either the various treatment target volumes or the organs-at-risk (OAR). These contours can be used in conjunction with images acquired during treatment planning to more precisely aim the radiation beam at the target volumes while aiming the beam away from the OARs. Target volume contours are differentiated by their margin composition: gross tumor volumes (GTVs) constitute the visible cancer tumor to be targeted, clinical target volumes (CTVs) extend from GTVs to include regions of likely subclinical disease, internal target volumes (ITVs) expand around CTVs to account for patient motion during treatment, and planning target volumes (PTVs) expand around ITVs to account for positioning and setup uncertainties. Several immobilization devices have been developed which increase the precision of image registration from treatment planning to delivery, decreasing the permissible margin which the PTV generates around the ITV and increasing the accuracy of the delivered radiation dose. The quality of the patient's immobilization can be monitored by external tracking devices such as infrared cameras, which further improves the overall accuracy of the treatment delivery.

Additionally, many technologies directly improve the degree of control which can be imposed on the radiation dose distributions delivered. Intensity-modulated radiation therapy (IMRT) involves passing the radiation beam through a multi-leaf

collimator (MLC), which is a set of mobile metallic leaves which allow for direct sculpting of the beam fluence. Sliding window IMRT utilizes MLCs which move during irradiation, further increasing the degree of control exerted on the radiation beam.

Volumetric modulated arc therapy (VMAT) is similar to sliding window IMRT, except the treatment gantry revolves in an arc around the patient during irradiation, creating even more possible deliverable dose distributions. All of these techniques combine to give the treatment planner immense control over the radiation doses delivered to the patient, which improves the ability to both irradiate the PTV while shielding the OARs and healthy tissue.

1.1.1 Stereotactic techniques

Stereotactic radiosurgery (SRS) and stereotactic body radiation therapy (SBRT) are methodologies of radiation therapy which deliver radiation with higher precision than conventional external beam radiation therapy. In SRS and SBRT, the treatment planner is able to impose sharper gradients on the dose distribution, which can further widen the potential therapeutic edge by simultaneously improving target volume coverage and OAR sparing. Typically, SRS and SBRT are achieved by combining tight patient immobilization techniques, high-definition MLCs with a smaller leaf width, and precise on-board image guidance to more precisely register the images used during treatment planning and treatment delivery. The improved dose distribution conformity has previously been shown to significantly improve patient outcomes for many types of

cancers and anatomical treatment sites. The primary distinction between the two techniques is that SRS is typically employed for brain and central nervous system lesions, such as spinal vertebral bodies, while SBRT is typically employed for treating lesions in other parts of the body, such as the pancreas. As a result, any potential registration or delivery errors can produce higher changes in the delivered dose distribution than they would in conventional EBRT. To account for this, SRS and SBRT typically require additional quality assurance before and after treatment delivery.

1.2 Treatment planning optimization

Despite all of the control that can be exerted on a dose distribution, there is no general guarantee that all target volumes can be uniformly irradiated to the target's prescription dose while completely avoiding the OARs. To address this, physicians will typically prescribe one or more dose-volume objectives (DVOs) to each contour. For example, a physician may prescribe to deliver at most 25% of the target's prescription dose to at most 35% of a given OAR's contour, represented as $D_{25\%}^{OAR} < 35\%$. Since these DVOs may conflict with each other, radiation therapy treatment planning is inherently a multi-criterial optimization problem, which has been abstractly studied in depth (Hwang and Masud, 1979, Miettinen, 1999).

IMRT and VMAT treatment plans usually involve high amounts of complexity with many degrees of freedom. This complexity is due primarily to the MLCs used in these treatment techniques which modulate the radiation beam fluences. Each plan

requires specific MLC positions to yield the desired dose distribution, and the number of possible MLC leaf positions is quite large. As a result, specifying these MLC positions by hand is not practical.

To overcome this immense computational complexity, computer techniques have been developed which automatically search the possible fluence combinations, determine the optimal fluence maps, calculate the MLC leaf motions which produce those fluence maps, and calculate the resulting dose distribution which will be delivered. These fluence optimization techniques rely on scalarized optimization functions which take a given dose distribution, compute the prescribed DVOs for each contour, and evaluate a weighted average of these DVOs, where the weights are determined by user-specified optimization objective priorities. These scalarized functions represent a single numerical value which can be minimized using modern univariate optimization techniques such as gradient descent, thereby pushing the dose distribution towards achieving all prescribed dosimetric goals. By specifying objective priorities, the physician is numerically weighting the trade-offs between the DVOs, i.e. the physician is specifying how much priority goes towards target coverage and how much goes towards dose sparing for each of the OARs. After specification, the doses are optimized and calculated in the treatment planning system (TPS), resulting in a deliverable plan with the required MLC movements calculated.

One downside of the DVO scalarization process is that the resulting DVOs cannot be deduced from the optimization priorities alone. As a result, it is not clear which optimization priorities should be selected for a given patient. Instead, treatment planners are required to empirically determine appropriate optimization priorities by creating several treatment plans and using their intuition and experience to guess the next appropriate set of optimization objective priorities to employ. This process can require the generation of many plans in order to create a final, clinically sufficient plan.

However, even with these powerful optimization and calculation techniques available, the treatment planning process can still be time-consuming for IMRT and VMAT treatment planning. This is primarily due to the high-resolution imaging which indirectly creates many voxels of interest within the body, the large number of feasible fluence maps which can be generated by the MLCs, and the algorithmic complexity involved in calculating the dose distribution from a given fluence map. As a result of the time requirements of dose optimization and calculation, each additional set of tested optimization priorities requires additional time. This time cost prevents the treatment planner from thoroughly testing the possible optimization priorities and reduces the amount of time which the treatment planner can spend on fine-tuning, which indirectly reduces the final plan quality and optimality.

In order to accurately calculate the dose distributions delivered by SRS and SBRT, the dose optimization and calculation engines need to operate at a higher

resolution. Typically, the optimization and calculation resolutions are raised from 2.5 mm to 1.25 mm or smaller in the X- and Y- directions, approximately quadrupling the number of voxels involved in optimization and calculation. This causes stereotactic treatment plans to require much more time to create, so treatment planners are limited to creating even fewer plans during treatment planning. Therefore, SRS and SBRT would benefit from treatment planning acceleration techniques more so than EBRT.

1.3 Machine learning

1.3.1 Traditional machine learning

Traditional machine learning techniques attempt to use human understanding and simple statistical regression to capture enough information to represent the underlying problem. In traditional machine learning, the model's architect needs to provide the model with a set of statistical features summarizing the input data which are believed to represent the underlying problem in order to handle the potentially-high dimensionality of the process to be modeled. These techniques then apply a relatively simple statistical regression, such as multi-linear regression, on the input features to produce an output prediction. The regression itself has several variables that need to be trained using input data. While traditional machine learning techniques tend to be rather simple and fast, their performance is limited by how much information is represented and the relevance of the information in the task at hand.

1.3.2 Deep learning

In contrast to traditional machine learning techniques, deep learning techniques attempt to use a more complicated regression to supplant the need for features provided by the model architect. For example, convolutional neural networks automatically generate relevant features from images by processing the images through several convolutions and activation functions (Krizhevsky et al., 2012). Many different deep learning techniques have been proposed, and there is significant research on developing models which operate under entirely different paradigms, such as inception modules (Szegedy et al., 2015), residual networks (He et al., 2016), generative adversarial networks (Goodfellow et al., 2014), and spatial transformer networks (Jaderberg et al., 2015). These techniques have proven to be very effective for image classification and computer vision (Goodfellow et al., 2014, He et al., 2016, Szegedy et al., 2015) as well as playing games such as Go with superhuman proficiency (Silver et al., 2017).

While there is a great potential for deep learning techniques to outperform traditional machine learning methods, their complexity and vast numbers of trainable parameters impose significant requirements during training. Generally, deep learning models need very large training datasets to generalize well outside of the training dataset. For example, effective image classification techniques typically use hundreds of millions of images in their training datasets. This quantity of information simply does not exist for radiation therapy, so it is challenging to create models which generalize

well. Additionally, the complexity of deep learning models requires large amounts of computer memory, high computer speed, and graphical processing unit (GPU) parallelization to train in a feasible time period.

1.4 Previous work on treatment planning acceleration and optimization

To address the computational requirements of treatment planning, many techniques have been proposed to accelerate and optimize the treatment planning process. The ultimate goal of all of these techniques is to circumvent the shortcomings that arise from the lack of knowledge about a patient's feasible dose-volume objectives and optimization priorities.

1.4.1 Multi-criterial optimization

The first of these techniques is multi-criterial optimization (MCO), which is a mathematical framework for handling situations in which it is desirable to optimize multiple objectives which potentially conflict with each other. MCO has a natural translation to treatment planning because target irradiation typically conflicts with normal tissue sparing. In MCO, the objective is to infer the set of points which cannot be strictly improved in every objective, historically called the Pareto surface (Hwang and Masud, 1979, Miettinen, 1999). An example of a Pareto surface is shown in Figure 1.

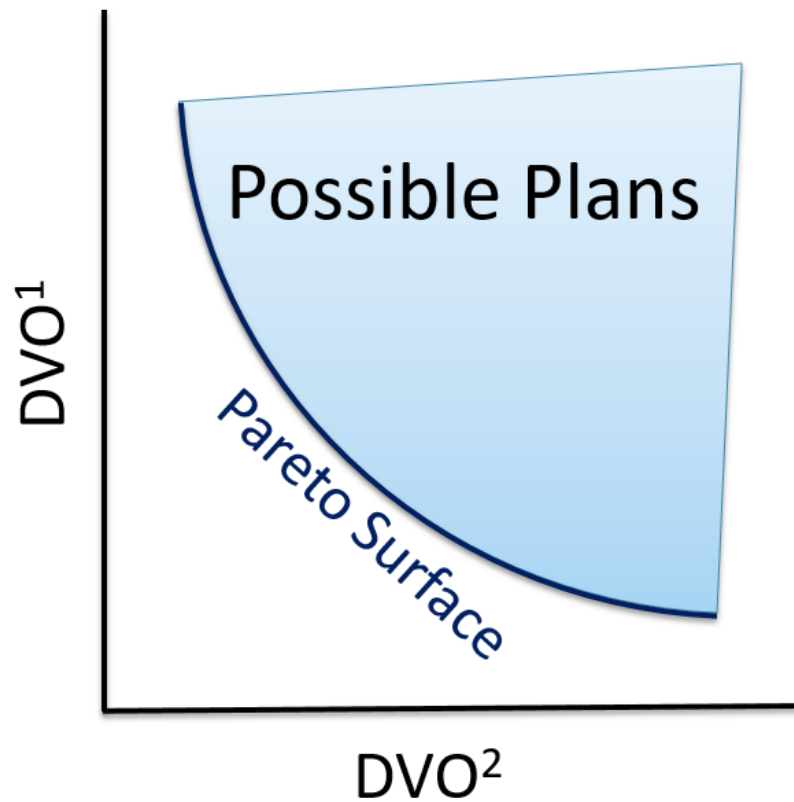


Figure 1: Graphical depiction of a Pareto surface in radiation therapy with two dose-volume objectives.

A priori knowledge of a given patient's Pareto surface can greatly accelerate the treatment planning process by providing the treatment planner an understanding of what DVOs are feasible along with which plan parameters create a dose distribution with those DVOs. Pareto surface knowledge allows the planner to immediately infer what DVOs should be pursued, which indirectly creates more time for the planner to

finely tune non-dosimetric aspects of the treatment plan, such as deliverability.

Therefore, MCO can indirectly improve the final plan quality.

Several MCO algorithms have been developed specifically for radiation therapy (Bokrantz and Forsgren, 2013, Craft et al., 2007). These algorithms typically involve the creation of an exhaustive sampling of the Pareto surface. An example of this sampling is shown in Figure 2. In this context, the samples from the Pareto surface form a simplicial complex which can be linearly interpolated to approximate the underlying Pareto surface between the sampled points. These algorithms use the convexity of DVO optimization to generate upper and lower bounds for the Pareto surface based on the current sampling and create subsequent samples in the regions of greatest distance between the upper and lower bounds. This greedily minimizes the Pareto surface uncertainty with each additional sample created, with the interpolated surfaces converging to the actual Pareto surface as more and more samples are taken.

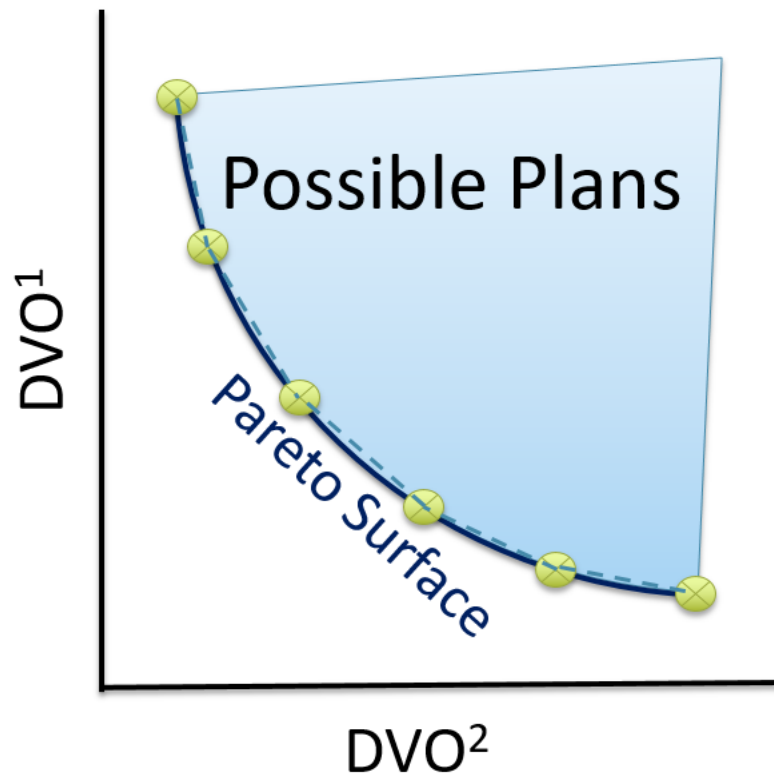


Figure 2: Example of Pareto surface sampling in clinical implementations of MCO.

To generate the sampled points on the Pareto surface, these MCO algorithms need to optimize and calculate the plans which correspond to those points in Pareto space. Since the dose optimization and calculation process can be quite time-consuming, these MCO algorithms inherently create a computational bottleneck in the treatment planning workflow. Time costs can vary significantly depending on the complexity of the treatment, the resolution at which dose optimization and calculation occur, and the computational power available to the treatment planning system, but the approximate range is 10 minutes to 30 minutes. Several commercial MCO software packages which

employ these algorithms typically mitigate the time cost of generating sampled points by reducing the number of sampled points. Although this allows the algorithms to run in a more reasonable time frame, the decreased sampling rates result in a larger error due to the linear interpolation of the Pareto surface. This effect is exemplified in Figure 3.

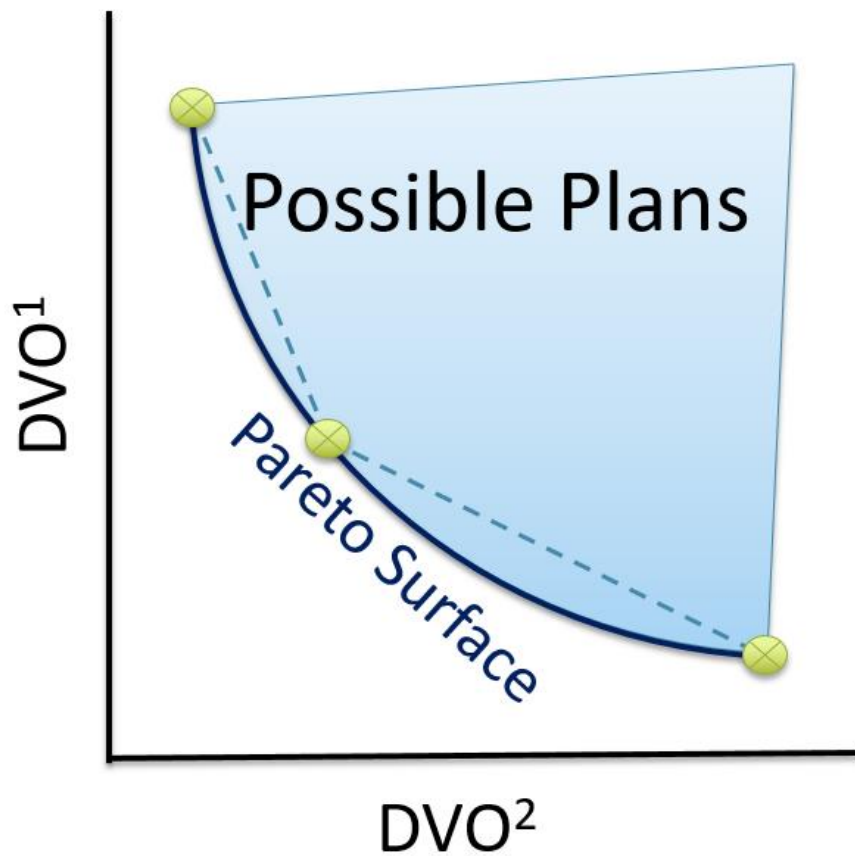


Figure 3: Pareto surface interpolation uncertainty generated by insufficient sampling. Here, we can see that the interpolations between the sampled points can be quite different from the actual Pareto surface.

Theoretically, the need to undersample the Pareto surface can be circumvented by significantly increasing the speed at which dose distributions are created. However, such speed-ups might not be attainable using contemporary dose optimization and calculation engines and algorithms. Therefore, it is desirable to develop an alternative method of dose distribution creation which operates significantly faster than current dose optimization and calculation engines.

1.4.2 Knowledge-based planning

Knowledge-based planning (KBP) is a class of treatment planning optimization methods which attempt to utilize prior knowledge to guide future treatment planning efforts. While the prior knowledge to be utilized can take many forms, most KBP methods rely on statistical modeling or machine learning to incorporate prior knowledge from previous patients and make inferences on new patients.

KBP techniques typically require a specific model architecture designed to capture the information required to make sufficiently accurate predictions. For KBP models that employ traditional machine learning, the model's architect needs to provide the model with a set of statistical features summarizing the input data which are believed to represent the underlying problem in order to handle the potentially-high dimensionality of the process to be modeled. These techniques then apply a relatively simple statistical regression, such as multi-linear regression, on the input features to produce an output prediction. The regression itself has several variables that need to be

trained using input data. Therefore, the quality of the predictions produced by traditional machine learning KBP techniques depend both on the quantity and quality of the input data to ensure generalizability and appropriateness of the predictions. In contrast, MCO techniques do not suffer from a need for quality training data because they obtain optimal solutions spread over a range of possible plan trade-offs.

One recently successful application of KBP focuses on predicting the feasible dose-volume histograms (DVHs) for a given patient (Appenzoller et al., 2012, Fogliata et al., 2015, Yuan et al., 2012). This application involves reducing the patient's anatomy to a set of distance-to-contour histograms which aggregate every voxel within the body and fitting these histograms to a skew-normal distribution. A commercial implementation of this application, RapidPlan™, has been implemented and made widely available (Varian Medical Systems, Palo Alto, CA). Although this KBP application has been successful, RapidPlan™ and similar techniques are only capable of predicting DVHs, which aggregate and summarize their underlying dose distributions. As such, these DVH-based techniques discard much of the spatial localization of the doses delivered to each structure, which could have been used to further assist in the treatment planning process.

1.4.3 Voxel-wise dose prediction

Another type of KBP application is voxel-wise dose prediction, where a machine learning model tries to predict a patient's potential dose distribution. Significant

research has been performed on the potential to predict dose distributions using deep learning (Babier et al., 2020, Kajikawa et al., 2019, Mardani et al., 2016, Nguyen et al., 2019). However, many of these techniques attempt to take models designed for image classification and apply them to the dose prediction problem. Since dose prediction is closer to a regression than a classification, the applicability of these models is limited. Additionally, almost all dose prediction models have been trained and tested on clinical, previously-delivered plans. Since physicians tend to have different preferences and priorities when it comes to treatment planning, clinical datasets of dose distributions tend to have inherent variance and irregular plan quality. This problem is similar to the plan quality requirement of RapidPlan™, which is supplanted by the uniform optimality of plans generated by MCO algorithms. Therefore, there is a need for further investigations into potential dose prediction models which overcome these hindrances.

1.5 Spine SRS and its failure patterns

Spine stereotactic radiosurgery (SRS) is a common option for treating metastases in the vertebral bodies and posterior elements of the spine. Several studies have demonstrated the ability of spine SRS to simultaneously yield higher local control rates and lower normal tissue complication rates than conventional EBRT (Kirkpatrick et al., 2014).

One of the most dreaded potential side effects of spine SRS is radiation-induced myelopathy of the spinal cord. In spine SRS, the proximity of the planning target

volume to the spinal cord typically results in some dose being incidentally delivered to the spinal cord. If the dose delivered to the spinal cord is sufficiently high, it may deal enough damage to impair the neurological function of the spinal cord, potentially resulting in partial or complete paralysis.

To minimize the possibility of radiation-induced myelopathy, physicians have historically been extremely conservative regarding spinal cord dose tolerances in spine SRS. Commonly-used conservative spinal cord dose tolerances for single-fraction spine SRS are at $D_{\max} < 13$ Gy and $D_{10\%} < 10$ Gy; as a result, the rate of radiation-induced myelopathy are extremely low (Grimm et al., 2016, Kirkpatrick et al., 2010, Sahgal et al., 2010).

The strictness of the dosimetric constraints imposed on the spinal cord require treatment planners to take strong precautions to minimize the probability of unintentional spinal cord overdosing. Treatment planners typically achieve this by not placing dosimetric constraints on the epidural space between the spinal cord and the vertebral metastasis. As a result, the doses delivered to the spinal cord can vary widely from patient to patient between the target prescription dose and the spinal cord dose constraints. Incidentally, approximately 50% of local control failures occur in the epidural space in spine SBRT, while approximately 5%-20% of all cases result in epidural failure (Chang et al., 2007, Garg et al., 2011, Nelson et al., 2009, Oinam et al., 2011, Thibault et al., 2015). We suspect that these high epidural failure rates are caused by the

lack of uniform epidural space coverage combined with the imperceptible subclinical disease in the epidural space. Importantly, failures in the epidural space can cause metastatic epidural spinal cord compression (MESCC), where the recurring tumor applies pressure to the spinal cord, causing loss of neurologic function and potential paralysis. To address the disparity between radiation-induced myelopathy and MESCC, researchers will likely conduct clinical trials on epidural space irradiation in spine SBRT in the future. These trials would be further justified by dosimetric analyses investigating the feasibility and range of feasible dose targets for epidural space irradiation in spine SBRT.

2. Study Objectives and Structure

As discussed in Chapter 1.4.1, Pareto surface inference can be beneficial for the treatment planning process by reducing the amount of trial-and-error required to determine the range of feasible doses for a given patient. However, there is a significant time cost associated with optimizing and calculating a dose distribution, slowing down the exact Pareto surface generation methods in contemporary multi-criterial optimization (MCO) algorithms in radiation therapy. Therefore, there is a strong clinical need to accelerate the speed at which Pareto surfaces can be inferred.

The purpose of this study is to streamline the MCO workflow by using a machine-learning model to quickly predict the Pareto surface plan doses, rather than exactly computing them. The model will then be used to rapidly estimate many Pareto optimal plans for quick Pareto surface inference. This could allow for a potentially significant increase in the speed at which Pareto surfaces are inferred to provide treatment planning assistance and acceleration. However, previous investigations into Pareto surface analysis typically do not compare a ground truth Pareto surface with a Pareto surface prediction. Therefore, there is a need to develop a Pareto surface metric in order to evaluate the ability of the model to generate accurate Pareto surfaces in addition to accurate dose distributions.

As discussed in Chapter 1.5, Spine SRS has historically suffered from epidural space local failures occurring in approximately 5%-20% of patients, which has been

caused by restrictive, conservative spinal cord constraints. In order to decrease the epidural local recurrence rate, there is a clinical need to quantitatively estimate the level to which the epidural space can be irradiated while still observing PTV and spinal cord constraints. A clinical trial investigating the potential for epidural space irradiation would be assisted with justification by a treatment planning dosimetric study on epidural space irradiation. This would essentially be a multi-criterial optimization (MCO) exploration of the range of possible dosimetric trade-offs between epidural space irradiation, PTV coverage, and spinal cord sparing. Since there is a natural synergy between epidural space irradiation studies and MCO studies, a dosimetric investigation on epidural space irradiation will be included in this work. Additionally, our model will be applied to this specific case to infer the ability of the model to maintain its accuracy and generalizability on a more diverse patient cohort.

The specific objectives of this study can be summarized as follows:

Objective 1: Develop a fast, accurate voxel-wise dose prediction machine learning model. This objective will be met in Chapter 3, where the model's final design is presented. Additionally, earlier iterations of the model will be discussed, including why specific elements of the model were replaced over the course of development.

Objective 2: Develop a stable, meaningful similarity metric to be used in Pareto surface similarity analyses. This objective will be met in Chapter 4, where several Pareto surface similarity metrics will be analyzed and compared, including an analysis of

metric convergence and computational time requirements. Theoretical, abstract simplex pairs will be used for testing.

Objective 3: Analyze the range of possible doses for epidural space irradiation in spine SRS and conclude whether epidural space irradiation is feasible in spine SRS. This objective will be met in Chapter 5, which presents the results of repeated treatment planning with epidural space targeting on retrospective spine SRS patients. The effectiveness of these levels of epidural space irradiation will be estimated using a simplistic radiobiological model.

Objective 4: Implement this dose prediction model in Pareto surface estimation for several treatment paradigms. This objective will be met in Chapters 6, 7, and 8, which present the results of applying our dose prediction model to conventional prostate VMAT, pancreas SBRT, and spine SRS, respectively. Conventional prostate VMAT will assess our model on relatively lower-resolution dose distributions with patient cohorts of low variance in order to assist a treatment planner with a common, standardized treatment technique. Pancreas SBRT will assess our model on high-resolution dose distributions with patient cohorts of moderate variance in order to assist a treatment planner with a more specialized, more difficult treatment technique. Spine SRS will assess our model on high-resolution dose distributions with patient cohorts of high variance in order to accelerate theoretical dosimetric studies for new treatment paradigms, such as epidural space irradiation.

3. Model design

3.1 Final iteration

An overview of the model's architecture is shown in Figure 4. The model's inputs are the objective priorities and the structure maps of the PTV and all OARs which are relevant to optimization. These structure maps are binary image-domain representations of the corresponding structures, indicating for each voxel whether that voxel is inside the structure's contour. These structure maps have been scaled by their corresponding objective priorities for each plan. This is a straightforward way to incorporate objective trade-off priorities without complicating the model's architecture.

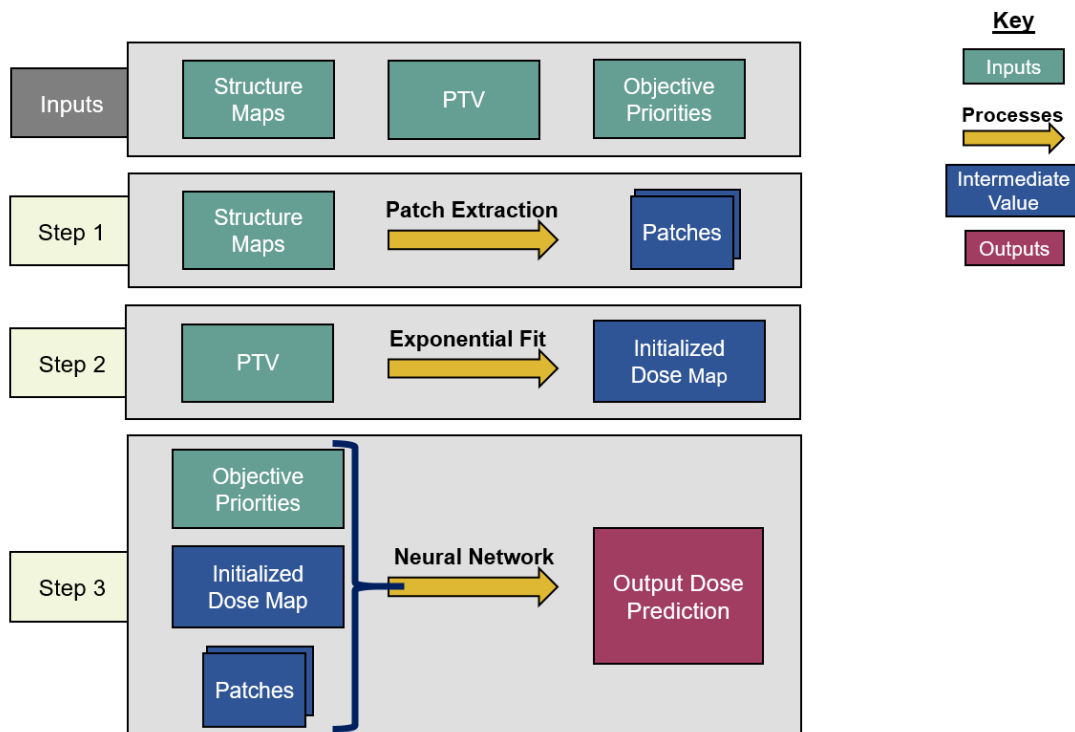


Figure 4: Overview of the dose prediction model architecture.

3.1.1 Patch extraction

The model begins by extracting three sets of 9×9 transverse patches from all structure maps at each voxel. Each set of patches has a different atrous rate, which is the number of skipped voxels between the sampled voxels. The first patches have an atrous rate of 1, i.e. they do not skip any voxels and are contiguous. These patches allow the model to infer local structure information near the pixels on which they are centered at the same resolution as the underlying image. For the second patches, the structure maps are smoothed by convolution with a uniform 3×3 kernel, and the patches are extracted with an atrous rate of 3. Similarly, the third patches are extracted with an atrous rate of 10 from the structure maps after smoothing by a 10×10 kernel. The purpose of the second and third patches is to coarsely infer wider-range information. The smoothing convolutions are performed to make each voxel sampled by the atrous patches contain structure information from the nearby voxels that the atrous sampling skips. The combination of average smoothing and atrous sampling essentially increase the model's receptive field size per layer, so that the model can infer the presence of critical structures at both large and short distances without significantly increasing the amount of memory or model parameters required. The patches are cast into 81-element vectors per voxel, and the vectors and optimization priorities are all concatenated voxel-wise to serve as input for the residual network.

3.1.2 Dose initialization

The model proceeds by creating an initial dose distribution via an inverse fit of inter-slice and intra-slice PTV distance maps on a voxel-wise basis. The functional form of the initialized dose fit is

$$D_i = [1 + a_1 * ISD_1^{a_2} + a_3 * ISD_2^{a_4}]^{-1}$$

where ISD_1 refers to the intra-slice distance from the voxel to the nearest PTV location within the voxel's slice, ISD_2 refers to the inter-slice distance from the voxel to the nearest PTV location at the voxel's row and column, and a_1, a_2, a_3 , and a_4 are variables that need to be fitted. Figure 5 depicts the effect of initializing a dose distribution using this fit. The purpose of this initialization is to allow the subsequent neural network to predict the shift between the initialization and the TPS-calculated dose distribution rather than the dose distribution itself. We hypothesize that these shifts are likely more linear than the dose distribution itself and therefore more easily learned.

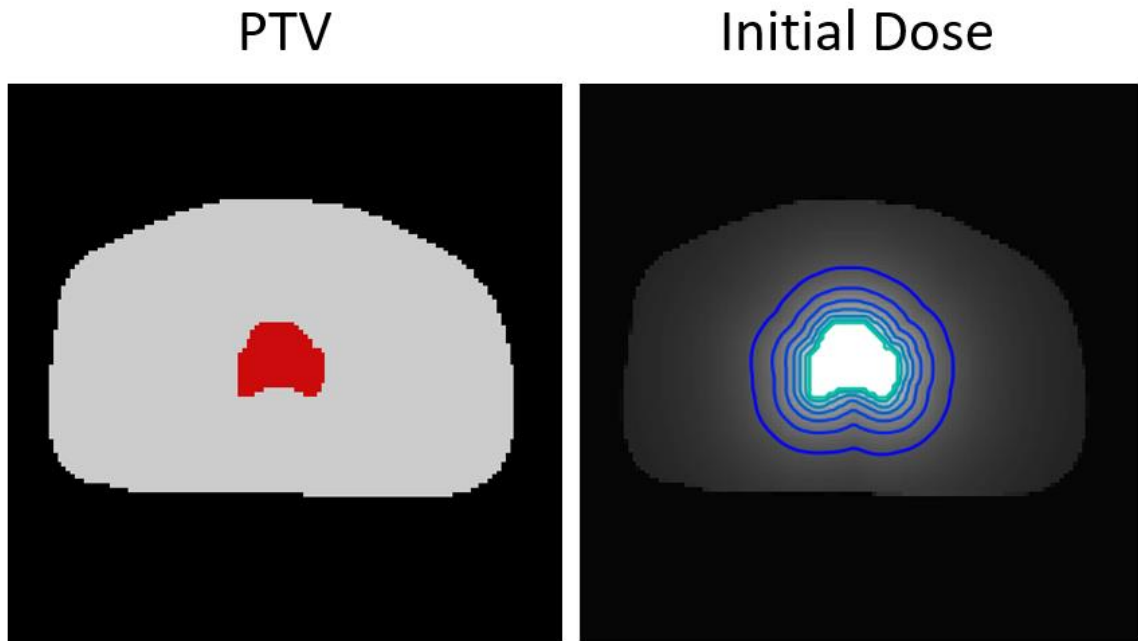


Figure 5: A sample transverse prostate PTV contour (left) compared to its initialized dose (right).

This dose initialization process makes several assumptions. Most importantly, the initialized dose is isotropic as a function of distance from the PTV. This isotropic assumption is most applicable for coplanar, concentric VMAT arcs which sweep across nearly all 360 degrees of rotation. Although this assumption narrows the applicability of our model, we anticipate that similar dose initialization processes can extend the model to partial VMAT arcs or IMRT beams. Additionally, this initialization process completely disregards all OARS, making the initialized dose very different from an actual plan. However, the distribution does resemble the general shape of a real dose distribution, which our model uses to kick-start the residual network.

3.1.3 Residual network

The model then uses the extracted patches and the initialized dose as inputs for a neural network, which is inspired by the recently developed ResNet (He et al., 2016). Our network consists of a series of 6 residual blocks that sequentially update the initialized dose map. Each residual block, depicted in Figure 6, consists of three fully connected layers. Similarly to the anatomical contours, patches are extracted from this initial dose distribution for use in the residual network. All anatomical and dose patches are then concatenated and processed through the first two neural network layers, which have 100 output units and leaky rectified linear unit (L-ReLU) activations, defined as $L-ReLU(x) = x$ when $x > 0$ and $L-ReLU(x) = 0.2x$ when $x \leq 0$. These first two layers extract quasi-linear features from the patch vectors which promote well-behaved gradients for training. The last layer has a single output and scaled softsign (SS) activation, defined as $SS(x) = 0.5x / (1 + |x|)$. The purpose of this last activation function is to take the quasi-linear combinations from the previous layer and map them to a suitable dose shift with a limited range. Since each residual block changes the initialized dose map, the dose map patches need to be re-extracted after each block. The number of residual blocks, layers per block, and output units per layer were chosen subjectively through trial-and-error, and we suspect that the accuracy achieved by this neural network can be achieved through similar network designs and hyperparameter tunings.

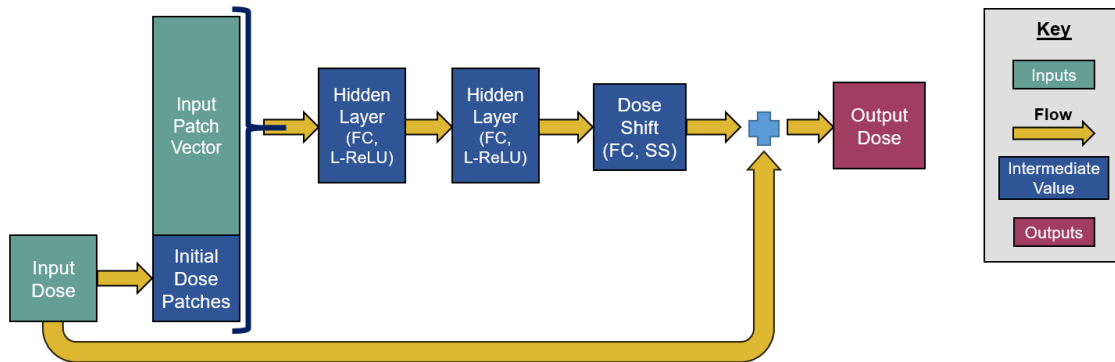


Figure 6: Graphical depiction of a residual block within the neural network. (FC = fully connected, L-ReLU = leaky rectified linear unit activation, SS = scaled softsign activation).

3.1.4 Model training

The training loss function was the root-mean-square error (RMSE) between the predicted dose map and TPS- calculated dose map, restricted to voxels within the body contour and restricted to slices containing at least one critical structure. Dose initialization variables were fit according to the RMSE between the initialized dose map and the TPS- calculated dose, and these variables were trained before the residual network variables. Gradients for the loss function were estimated using batches of training data, with each batch containing a number of slices approximately equal to the typical number of slices that a patient would have. Slices for the batches were sampled diagonally, such that the batch slices were located at different levels within different patients. This sampling means that each batch contains slices from most patients at most

slice positions, such that each batch is a good representation of the entire cohort. Therefore, the gradients computed from the batches were close approximations to the gradients of the loss function applied to the entire cohort, improving optimization convergence and stability. A graphical representation of this diagonal batch aggregation scheme is shown in Figure 7.

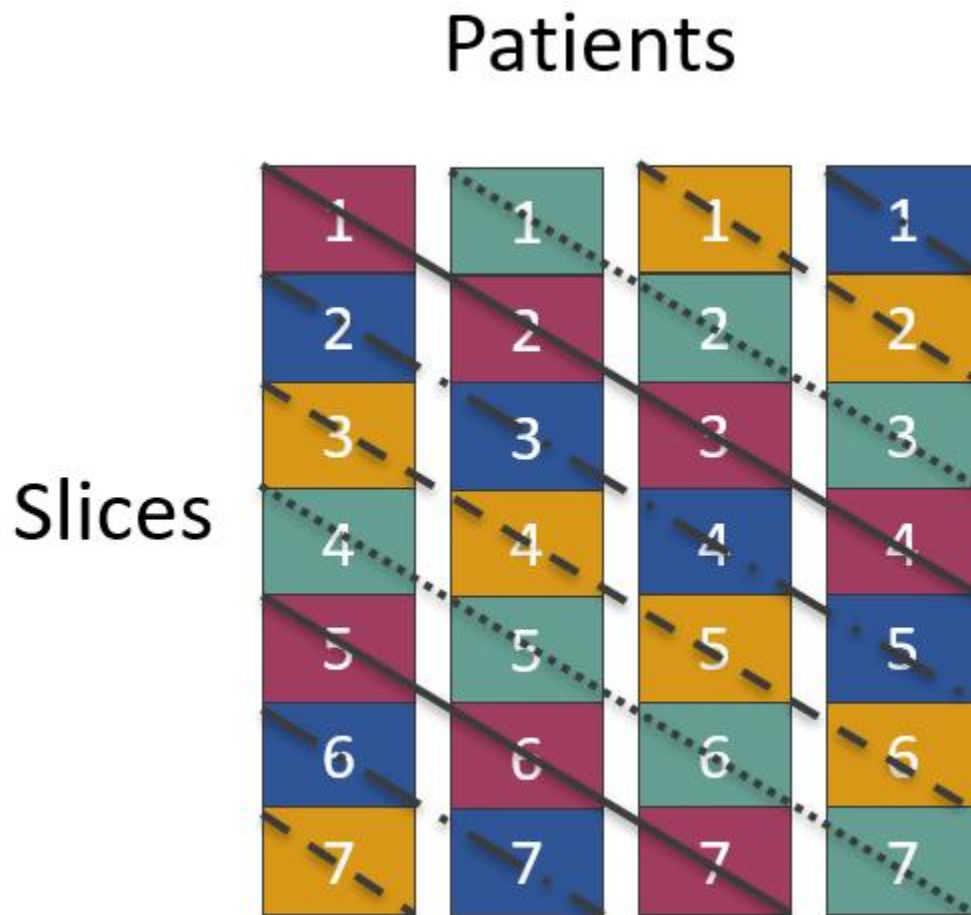


Figure 7: Graphical depiction of the diagonal batch aggregation scheme used during model training. Here, different colors refer to different diagonal batches.

The model was trained using the Adam optimization algorithm, which was designed for stochastic gradient-based optimization (Kingma and Ba, 2014). Kingma and Ba recommend specific hyperparameters, including step size $\alpha = 0.001$, decay hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and error epsilon $\epsilon = 10^{-8}$, all of which are used in the training of our model. The Adam optimizer is particularly appropriate here, since the batch gradient computations are stochastic. The trainable parameters in each layer were initialized using the Glorot uniform initializer, which initializes variables by sampling randomly from a uniform distribution bounded by

$\pm \sqrt{6 \div ((number\ of\ inputs + number\ of\ outputs))}$ (Glorot and Bengio, 2010). The

Glorot uniform initializer was designed to model the inherent variance of rectified linear unit activation functions, similar to the activation functions used in our residual network. All aspects of the model, including optimization and evaluation, were implemented using the Tensorflow machine learning platform (Abadi et al., 2015) with an NVIDIA Quadro M4000 graphics processing unit. Optimization proceeded for 2000 iterations before termination.

3.2 Earlier iterations

3.2.1 Neural network design

During the model design process, it was not evident which model design would produce the best dose predictions. To address this, several network architectures were

implemented to determine the optimal network architecture. The first network architecture was a classical convolutional neural network (CNN) implementation. This model operated directly on the input data, recast as a multi-channel data frame with each channel referring to either the PTV, an OAR, or the body. This network did not require or accept a dose initialization. Next, we attempted to use the CNN as an update to a dose initialization, which will be discussed in Chapter 3.2.3. However, in both cases, the resulting dose predictions resembled blurs around the PTV, rather than isotropic fall-off, which was likely due to gradient vanishing. We finally chose to use a residual network for producing dose shifts in order to prevent those gradient vanishing problems within the PTV, and we achieved a noticeable improvement to the model.

Additionally, there was some subjectivity when deciding on the optimal activation functions. Before settling on the combination of L-ReLU (leaky rectified linear unit) and scaled softsign activation functions, our model architecture first contained just normal ReLU activation. We additionally tried using just L-ReLU activation functions for comparison. However, both cases resulted in dose predictions that were rather linear. To address this, we implemented the scaled softsign activation at the end of each residual block to cast the model's processed information from a quasi-linear combination into an actual dose shift. This produced a measurable improvement on our model, hence our final model architecture.

3.2.2 Regularization

Regularization is a method of reducing the potentially-saturating degrees of freedom contained by an ill-defined problem. In optimization, regularization is applied to a problem by adding a regularization term to the loss function being optimized: $L'(x) = L(x) + \lambda R(x)$, where λ is a parameter which controls the magnitude of the effect of regularization. Typically for neural networks, functional forms of the regularization term $R(x)$ include linear combinations of the L^n -norms of the weights defining each of the convolution layers, most commonly lasso (L^1) and ridge (L^2) regularization.

Although lasso and ridge regularization have previously been shown to be effective for linear regression models, they have had mixed results in their application to non-linear deep learning models. When applied specifically to our model, both lasso and ridge regularization decreased the final training errors and final testing errors, regardless of the free parameter λ . During this analysis, λ was sequentially divided by 10, exponentially approaching zero to floating-point precision, and the training errors and testing errors for all values of λ were higher than the errors achieved without regularization, with the difference approaching zero as λ became small.

Additionally, regularization can commonly be applied to a problem by inducing randomness in the problem's information. For example, the diagonal batch aggregation scheme presented in Chapter 3.1.4 causes the loss function gradients to be stochastic

from batch to batch, so it is difficult to overfit to any single batch. The results of Chapters 6, 7, and 8 will confirm that the regularization inherent in batch aggregation is enough to prevent the model from overfitting to the training dataset too heavily. Therefore, functional regularization techniques were discarded from the final model iteration.

3.2.3 Dose initialization fits

One of the more unique elements of this machine learning model is the inclusion of the dose initialization structure. Most model frameworks attempt to generate the output directly from the input. Our model instead attempts to use the input information to morph an initialized guess of the dose distribution, which is a rudimentary fit of distance from the PTV. However, it was not immediately clear from the problem as to which initialization best serves to streamline the prediction process.

Originally, the dose initialization was set to be the binary mask of the PTV, with zero dose outside of the PTV and prescription dose uniformly delivered to the PTV. Although this does not resemble a feasible dose distribution at all, the principle behind this selection was to use a set of residual convolutions to spread the dose away from the PTV. However, this process was hindered by the functional performance on the PTV itself, since the blurring process tended to reduce the PTV dose below prescription. Although the dose distribution root-mean-square error was within a few percentage points of the dose prescription, the resulting DVHs differed significantly from the actual DVHs, particularly in the PTV.

The second initialization attempt was a mask consisting of just ones at every location in the body. The principle behind this initialization is to add more mass to the dose distribution, so that the doses in the PTV do not decay as the information exchanges from layer to layer. However, the dose predictions were prone to errors at a greater distance to the PTV, suggesting that the information of PTV location was not able to properly propagate through the network. This problem persisted when adding another image as input, where the values in the image corresponded to the voxel-wise distance to the nearest voxel within the PTV.

In order to induce PTV distance more directly, the model also tried to initialize a distribution according to an exponentially-decaying fit as a function of distance to the PTV: $D_i = \exp(a_1 * ISD_1^{a_2} + a_3 * ISD_2^{a_4})$, where ISD_1 refers to the intra-slice distance from the voxel to the nearest PTV location within the voxel's slice, ISD_2 refers to the inter-slice distance from the voxel to the nearest PTV location at the voxel's row and column, and a_1, a_2, a_3 , and a_4 are variables that need to be fitted. While these results were reasonable, there was a slight improvement when testing the inverse-decaying fit described in Chapter 3.1.2, so the initialization was finally determined to be an inverse-decaying fit.

4. Pareto surface metrics

4.1 Introduction

Chapter 3 focused on the creation and presentation of a dose prediction model which can take optimization objective priorities into account. This model is indirectly capable of predicting Pareto surfaces in radiation therapy treatment planning by predicting many dose distributions for many objective priority combinations and computing the relevant dosimetric quantities. These dosimetric quantities are then grouped according to their plan to form sampled points along the Pareto surface.

Theoretically, a dose prediction model which exactly replicates the clinical dose distributions would also exactly replicate the clinically created Pareto surface. However, when the dose distribution predictions contain errors, the predicted Pareto surfaces also contain errors. Importantly, the magnitude of errors in the dose distribution predictions cannot be used to infer or estimate the magnitude of errors in the Pareto surface predictions. This is exemplified in Figure 8, which demonstrates different dose distributions with identical dose-volume metrics. Therefore, it is important to be able to quantify uncertainty in Pareto surface predictions apart from analyzing the underlying dose distributions.

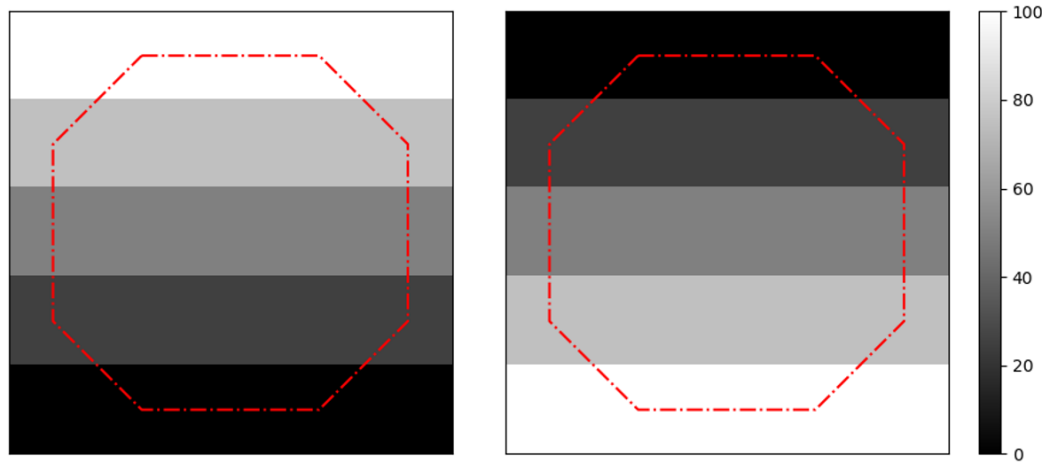


Figure 8: Example dose distributions. Here, the root-mean-square difference between these dose distributions is 70.7%, although the distributions have identical DVHs within the contour.

The focus of this Chapter is to present and analyze several metrics for comparing predicted Pareto surfaces with real Pareto surfaces generated from simulation in a treatment planning system (TPS). There has been essentially zero research into this topic, since most MCO algorithms operate by iteratively solving the underlying multi-criterial optimization problem to exactly sample the Pareto surface with real plans. Instead, Pareto surface comparison research has focused on evaluating which points were Pareto-optimal compared to both surfaces (Berezkin and Lotov, 2014, Teichert et al., 2011). To address the lack of research on Pareto surface similarity metrics, this Chapter presents four different Pareto surface metrics: the root-mean-square error (RMSE), the Hausdorff distance (HD), the average projected distance (APD), and the average nearest-point distance (ANPD). This Chapter examines the benefits and drawbacks of

each metric, tests these metrics on theoretical Pareto surface examples, analyzes the convergence of these metrics with respect to upsampling parameters, and concludes with the range of appropriate scenarios of each metric.

4.2 Materials and methods

4.2.1 Similarity metrics between Pareto surfaces

4.2.1.1 Root-mean-square error

The first metric which we will consider is the root-mean-square error (RMSE) between the sampled points of the Pareto surfaces. This metric evaluates the distance between matched points on two Pareto surfaces (labeled A and B), and computes the root-mean-square of these distances:

$$RMSE(X, Y) = \sqrt{\frac{\sum_{i=1}^s \|\bar{x}_i - \bar{y}_i\|_2^2}{s}}, \quad \bar{x}_i \in X, \quad \bar{y}_i \in Y$$

Here, X is a set of sampled points from Pareto surface A , Y is a set of sampled points from Pareto surface B , " s " is the number of samples on each surface, and the index " i " refers to a joint enumeration between the sets X and Y which matches points between the sets. Essentially, this metric examines pairs of points between the two surfaces which are believed to be similar and aggregates their Euclidean distances.

The primary benefit of this metric is that it is easy to interpret and implement. Euclidean distances between points are straightforward to compute, and human interpreters of this metric can easily apply their intuition about distance to understand

this metric. This method is also the easiest from a computational standpoint, with typical computation times much less than one millisecond.

The primary downside of this metric is that it only calculates errors on the sampled points and not on the underlying interpolations of the Pareto surfaces. In general, these can be different, as is exemplified in Figure 9. This is particularly important for Pareto surface prediction because the final plan resulting from an MCO algorithm in radiation therapy is usually an interpolation between previously sampled plans. Therefore, the RMSE between sampled points does not reflect the uncertainty of the final dose predictions in Pareto space.

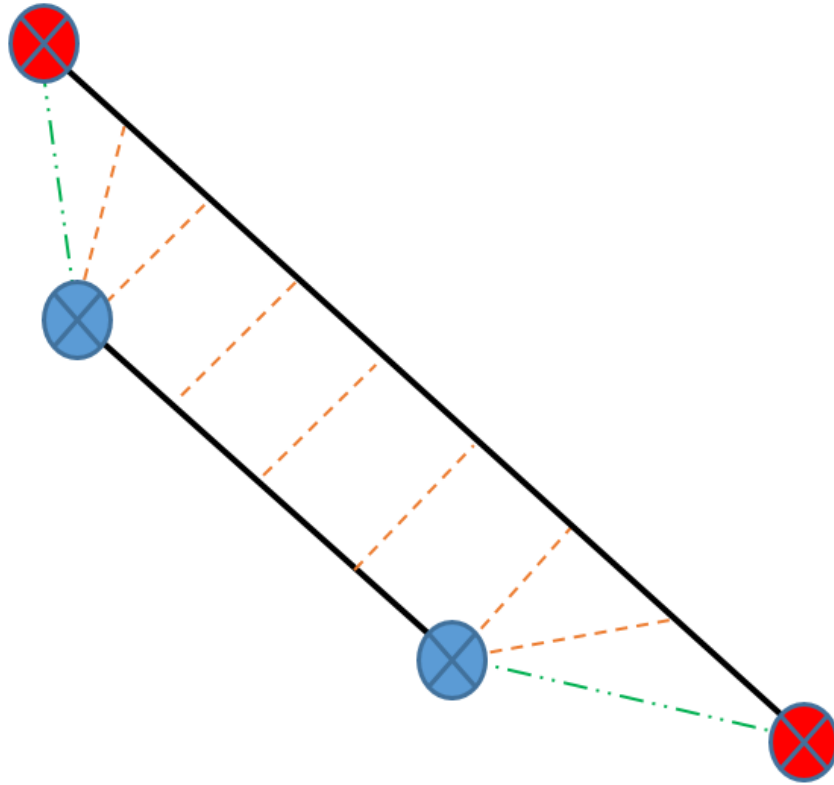


Figure 9: Example visualization of distance differences between sampled points and the underlying surfaces. Here, the distances between the sampled points (green) are larger than the distances between all of the interpolations (orange).

This metric can be made to more closely represent the distances between the interpolations by having the sampled sets X and Y include interpolations from the samples on the Pareto surfaces. These interpolations can be generated by considering the Pareto surfaces as simplicial complexes and interpolating within a given simplex by uniformly sampling from its barycentric coordinates, i.e. generating uniformly

distributed sets of parameters $\{\lambda_{ij}\}_{i=1}^n$, where "n" is the number of objectives, which meet the following constraints:

$$\lambda_{ij} \in [0,1]$$

$$\sum_{i=1}^n \lambda_{ij} = 1$$

These coordinates can then be used to generate the relevant points:

$$\vec{x}_j = X_M \vec{\lambda}_j$$

$$\vec{y}_j = Y_M \vec{\lambda}_j$$

where X_M and Y_M are matrix representations of the sets X and Y , and $\vec{\lambda}_j$ is a vector composed of all of the λ_{ij} elements. For example, in a two-objective MCO problem ($n=2$), three valid sets of parameters are (1, 0), (0, 1), and (0.5, 0.5), which refer to the minimization of objective 1, the minimization of objective 2, and the minimization of the average of objective 1 and 2, respectively. Figure 10 provides a visualization of this intra-simplex interpolation. For an n -dimensional MCO problem, sets of barycentric coordinates can be generated procedurally by specifying a number of samples to obtain per barycentric dimension, m , generating all possible tuples of the form $\{i_1, \dots, i_{n-1} \mid i_k \in \{0, \dots, m-1\}\}$, dividing the values of those tuples by $m-1$, and then

excluding all tuples which do not add up to 1. Finally, when aggregating the distances from these points over multiple simplices, each simplex is weighted by its n -volume to ensure that densely-sampled regions are not more heavily represented than sparsely-sampled regions.

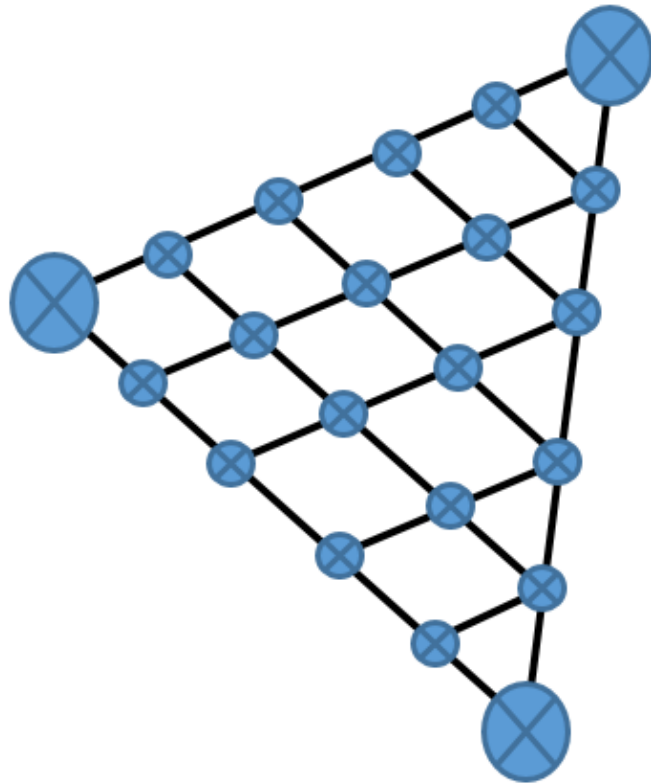


Figure 10: Visualization of intra-simplex upsampling according to the simplex's barycentric coordinates. Here, the number of samples obtained per barycentric dimension is $m = 6$, while the number of samples total is $s = 21$.

The optimization priorities which underlie these interpolations can be approximated as linear interpolations between the priorities which underlie the original sampled points, extending the original joint enumeration to include more matched pairs of sampled points. This approximation of priorities is not necessarily accurate, but it is likely to yield higher accuracy compared to the error evaluations including just the original samples. An example of the improvement is shown in Figure 11. These intra-simplex interpolations also create some subjectivity in the metric, as it is not immediately clear from the metric's definition exactly how finely these interpolations should be sampled. To address this subjectivity, an analysis of metric convergence will be included later in this Chapter.

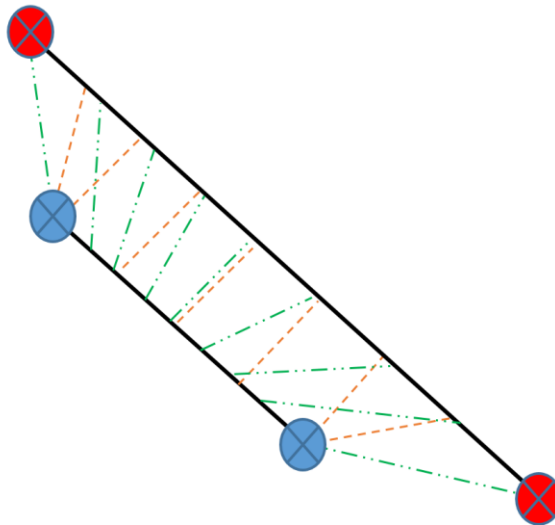


Figure 11: Example visualization of the effect of intra-simplex upsampling on the distances. Here, the distances between the interpolations between the sampled points (green) are still larger than the distances between the underlying surfaces (orange), but the differences are less drastic on average.

The secondary downside of this method is that it assumes the existence of a joint enumeration between the sampled sets X and Y , i.e. that it is possible to match points between the Pareto surfaces. The importance of the appropriateness of this enumeration is demonstrated in Figure 12. Due to the mismatching of the central sampled points of the Pareto surfaces, the RMSE is artificially high, despite the Pareto surfaces being quite similar. In the context of evaluating our dose prediction model, there is a natural joint enumeration, since the model requires an input set of priorities to perform a dose prediction. When comparing the Pareto surface generated by our model to the Pareto surface formed by a set of clinical plans, it suffices to take the optimization priorities which correspond to the clinical plans and use them to make equally many dose predictions. An enumeration can then be constructed by matching clinical distributions with predicted dose distributions that share optimization priorities. Since plans created using the same priorities would ideally be identical, this enumeration should yield an exact match between the sets of sampled points.

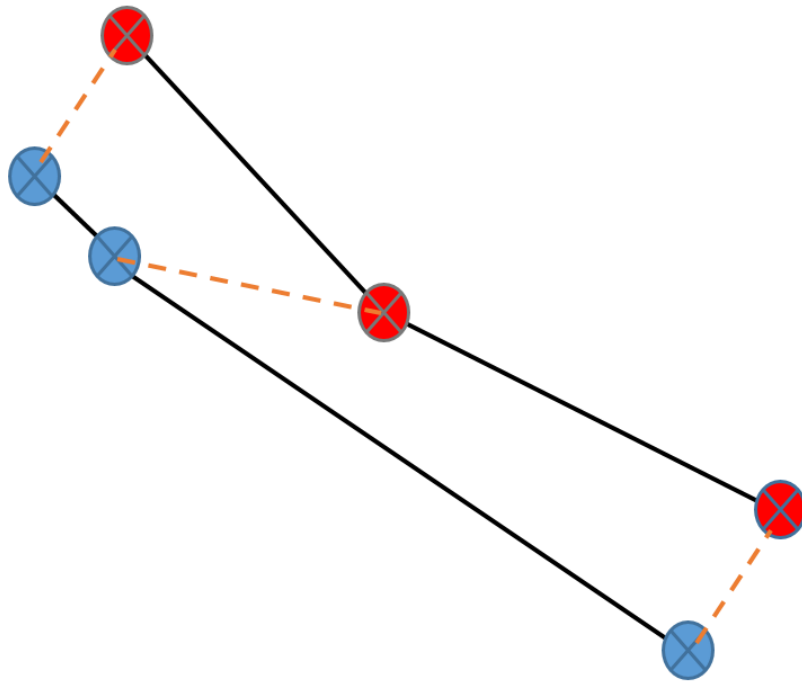


Figure 12: Example of the importance of accurate matching between sets in RMSE calculations. Here, the central pair of points is mismatched, resulting in artificially high distance evaluations.

However, in general, there does not always exist a joint enumeration between sampled points on Pareto surfaces. For example, we cannot match points between Pareto surfaces which are generated using clinical plans created by different algorithms, since the clinical plans very likely use different optimization priorities. As such, this metric cannot be used to evaluate the similarity or difference between Pareto surfaces originating from different software. Therefore, future studies which seek to compare

radiation therapy MCO software i.e. that of Varian Medical Systems and Raystation Laboratories, must use alternative Pareto surface similarity metrics.

4.2.1.2 Hausdorff distance

The second metric which we will consider is the Hausdorff distance between sampled points on the Pareto surfaces. The Hausdorff distance was originally created as a similarity metric between arbitrary sets. This metric evaluates the distances between sampled points of the Pareto surfaces and computes the greatest minimum distance:

$$HD(X, Y) = \max \{ \sup_{y \in Y} \inf_{x \in X} \|\vec{x} - \vec{y}\|_2, \sup_{x \in X} \inf_{y \in Y} \|\vec{x} - \vec{y}\|_2 \}.$$

Essentially, the metric first takes each sampled point from X , finds the distance to the closest sampled point on Y , finds the largest such distance among all points in X . The metric then finds the largest such distance among all points in Y and returns the larger value. Bokrantz and Forsgren use the Hausdorff distance as their loss function which is greedily minimized by sampling more and more points on the Pareto surfaces (Bokrantz and Forsgren, 2013). The most notable feature of this metric is its large dependence on outliers in the Pareto surfaces. This sensitivity to outliers causes a lack of dependence on the inclusion of interpolations. Specifically, since the sampled points are inherently more extreme than their interpolations, their distances are significantly more likely to dominate than the distances of the interpolations. This effect will be demonstrated during the error convergence analysis later in this Chapter.

One benefit of this metric is that it does not require an explicit joint enumeration between the sampled sets X and Y . Since every pair of points between the surfaces is utilized, the metric automatically finds the most appropriate pairs of points for evaluation. However, the lack of matched points may still result in an artificially high Hausdorff distance between sets. This is visualized in Figure 12 (Chapter 4.2.1.1), where the central pair of points dominates the Hausdorff distance despite the centers of the Pareto surfaces having distances similar to the edges of the surfaces. Additionally, since the Hausdorff distance heavily emphasizes outliers, any internal simplicial upsampling tends to have little effect on the Hausdorff distance in the absence of mismatching.

4.2.1.3 Average projected distance

The third metric which we will consider is the average projected distance (APD). This metric is similar to the RMSE described in Chapter 4.2.1.1, except the displacements are first projected to the mean displacement before aggregation:

$$\overline{\mu_{XY}} = \frac{\sum_{j=1}^s \overline{x_j} - \overline{y_j}}{s}$$

$$APD(X, Y) = \begin{cases} \frac{\sum_{j=1}^s ((\overline{x_j} - \overline{y_j}) \cdot \overline{\mu_{XY}})}{s \|\overline{\mu_{XY}}\|_2}, & \text{if } \overline{\mu_{XY}} \neq 0 \\ 0, & \text{otherwise} \end{cases}$$

Here, X is a set of s sampled points from Pareto surface A , Y is a set of s sampled points from Pareto surface B , the index i refers to a joint enumeration between the sets X

and Y which matches points between the sets, and $\overline{\mu_{XY}}$ is the mean displacement between the Pareto surfaces samples.

The purpose of this metric is to allow the joint enumeration to contain mismatching points by removing the component of the displacement along the average displacement before evaluating the distance. The primary motivation behind applying this metric to Pareto surface comparisons is to reduce the effect of using only sets of sampled points to infer inter-surface distance, rather than directly using the interpolated surfaces. This is shown in Figure 13, where point shifts along the surfaces more heavily affect the RMSE/Hausdorff contributions than the APD contributions.

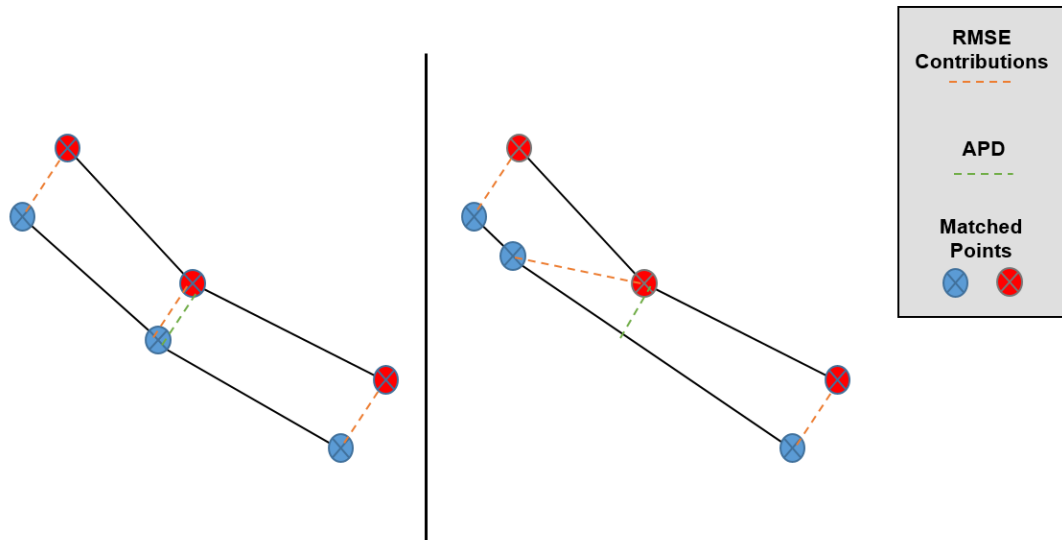


Figure 13: Example results of Pareto surface comparison using the RMSE and APD metrics. Here, we can see that the Pareto surfaces are relatively similar, but their central pair of sampled points is heavily mismatched. In this case, the RMSE contribution from the central pair is abnormally high. However, this component is first removed when computing the APD.

One seemingly viable method to improve the APD's accuracy would be to up-sample the sampled sets of points X and Y , similar to the upsampling technique used for the RMSE. This upsampling process is described in Chapter 4.2.1.1, and its mathematical implementation for the APD would be identical to its implementation for the RMSE, with the symbols applying exactly. However, the APD is mathematically invariant to upsampling, so the same value would be achieved for every number of samples taken. A proof can be construed as follows: since the tuple coordinates are generated to uniformly sample the barycentric dimensions, we know that $\sum_{j=1}^s \vec{\lambda}_j = Q\vec{1}$ and $\sum_{i=1}^n \sum_{j=1}^s \lambda_{ij} = \sum_{i=1}^n Q = nQ$, where Q is a constant and s is the number of coordinates generated by the intra-simplex upsampling process. Since $\sum_{i=1}^n \lambda_{ij} = 1$, we know that

$$\begin{aligned}
Q &= \frac{nQ}{n} \\
&= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^s \lambda_{ij} \\
&= \frac{1}{n} \sum_{j=1}^s \sum_{i=1}^n \lambda_{ij} \\
&= \frac{1}{n} \sum_{j=1}^s 1 \\
&= \frac{s}{n}.
\end{aligned}$$

We can therefore rewrite the APD as:

$$\begin{aligned}
APD(X, Y) &= \frac{\sum_{j=1}^s \left((X_M \bar{\lambda}_j - Y_M \bar{\lambda}_j) \cdot \overrightarrow{\mu_{XY}} \right)}{s \|\mu_{XY}\|_2} \\
&= \frac{(X_M - Y_M) \sum_{j=1}^s (\bar{\lambda}_j \cdot \overrightarrow{\mu_{XY}})}{s \|\mu_{XY}\|_2} \\
&= \frac{(X_M - Y_M) (\sum_{j=1}^s \bar{\lambda}_j) \cdot \overrightarrow{\mu_{XY}}}{s \|\mu_{XY}\|_2} \\
&= \frac{\frac{s}{n} (X_M - Y_M) \bar{\mathbf{1}} \cdot \overrightarrow{\mu_{XY}}}{s \|\mu_{XY}\|_2} \\
&= \frac{\frac{1}{n} (X_M - Y_M) \bar{\mathbf{1}} \cdot \overrightarrow{\mu_{XY}}}{\|\mu_{XY}\|_2}
\end{aligned}$$

where

$$\begin{aligned}
\overrightarrow{\mu_{XY}} &= \frac{\sum_{j=1}^s \bar{x}_j - \bar{y}_j}{s} \\
&= \frac{\sum_{j=1}^s (X_M \bar{\lambda}_j - Y_M \bar{\lambda}_j)}{s} \\
&= (X_M - Y_M) \frac{\sum_{j=1}^s \bar{\lambda}_j}{s} \\
&= \frac{1}{n} (X_M - Y_M) \bar{\mathbf{1}}.
\end{aligned}$$

Here, both μ_{XY} and $APD(X, Y)$ are independent of s and of all λ_j parameters, so the APD does not depend on the level of intra-simplex upsampling. Therefore, the APD is invariant to any intra-simplex upsampling performed on the sampled sets X and Y . This fact will be confirmed experimentally during our metric convergence analysis later in this Chapter.

One benefit of this metric is that it is easy to implement. Euclidean distances between points are straightforward to compute. This method is also very easy from a computational standpoint, with typical computation times much less than one millisecond. However, it is not guaranteed from its formulation that the APD computes appropriate distances for comparing Pareto surfaces. To address this, we will compare the APD with the metric discussed in Chapter 4.2.1.4, which does have a theoretical foundation supporting its appropriateness for Pareto surface comparison.

4.2.1.4 Average nearest-point distance

The fourth metric which we will consider is the average nearest-point distance (ANPD). Like the average projected distance discussed in Chapter 4.2.1.3, the ANPD has been created specifically to address the lack of useful Pareto surface metrics for machine learning model evaluation. This metric somewhat resembles the Hausdorff distance in that it takes individual points on one Pareto surface and finds the nearest point on the other surface as an internal step during calculation. However, rather than computing the supremum over the set of sampled points, the ANPD computes the average over the sampled points. Additionally, the ANPD considers all possible interpolations of the sampled points when finding the nearest points:

$$ANPD(X, Y) = avg \left\{ \text{avg}_{y \in S(Y)} \inf_{x \in S(X)} \|x - y\|_2, \text{avg}_{x \in S(X)} \inf_{y \in S(Y)} \|x - y\|_2 \right\}$$

Here, X is a set of sampled points from Pareto surface A , Y is a set of sampled points from Pareto surface B , and $S(X)$ is the simplicial complex formed by the sampled points in X , i.e. $S(X) = \{\sum_{i=1}^n \lambda_i x_i \mid \lambda_i \in [0, 1] \mid \sum_{i=1}^n \lambda_i = 1 \mid \{x_i\} \text{ all belong to one simplex}\}$. In the context of Pareto surface evaluation, $S(X)$ refers to the portion of A which is bounded by the sampled points of X .

The main hurdle of this metric is the requirement of computing the internal point-to-simplex distances of the form $\inf_{x \in S(X)} \|x - y\|_2$. Originally, Johnson proposed an exhaustive search algorithm for computing point-to-simplex distances, which proceeds by checking every facet of the simplex and checking whether its closest point satisfies a set of conditions (Gilbert et al., 1988). This algorithm was developed as an intermediate step in the Gilbert-Johnson-Keerthi algorithm, which computes the distance between arbitrary convex objects. Although Johnson's algorithm is accurate in arbitrary dimensions, it operates on brute force and can be computationally expensive. However, most of the more recent research on point-to-simplex distance algorithms has focused on restrictions to at most three dimensions after Johnson's algorithm was proposed. This is because most of the interest in point-to-simplex distance algorithms is in the application to collision detection in computer simulations (Ericson, 2004). For this reason, we will be employing Johnson's distance algorithm as an intermediary step in the ANPD calculation.

A second hurdle is created when calculating point-to-simplex distances using an iterative algorithm such as Johnson’s distance algorithm, which is that an exact integration of these distances is difficult to perform over simplices, i.e. computing $\text{avg}_{y \in S(Y)} \inf_{x \in S(X)} \|x - y\|_2$. However, distance functions tend to be well-behaved and smooth, leading to fast convergence in numerical integration. Therefore, we propose to numerically integrate these averages by upsampling the simplicial complexes according to their barycentric coordinates, similar to Chapter 4.2.1.1. An analysis of numerical integration convergence will be provided later in this Chapter.

The primary benefit of this metric is that it represents exactly what we desire in a Pareto surface similarity metric. This metric does not explicitly require a joint enumeration between the sampled sets X and Y . Since every pair of points between the surfaces is utilized and all possible interpolations are considered, the metric automatically finds the most appropriate pairs of points for evaluating the distances. As such, this metric is ideal for comparing Pareto surfaces without a joint enumeration.

The primary cost of this metric is the computational complexity of the interior point-to-distance algorithm proposed by Johnson. Each internal iteration of this algorithm relies on solving a set of equations which depend on the input point. When running this algorithm with large numbers of input points, each different point requires a different solution set, which can cause the algorithm to impose a significant

computational cost. An analysis of the time requirements of the ANPD using this algorithm is included later in this Chapter.

4.2.2 Metric classification proofs

Based on their definitions from Chapter 4.2.1 alone, it is not immediately clear that the proposed Pareto surface similarity metrics are in fact metrics in the formal, mathematical sense. This section contains mathematical proof that the metrics presented in Chapter 4.2.1 are distance metrics in the conventional sense, i.e. they have the properties of non-negativity, the identity of indiscernibles (the metrics equal zero only when comparing identical Pareto surfaces), symmetry (the order of the Pareto surfaces A and B does not matter), and sub-additivity ($d(A, B) \leq d(A, C) + d(C, B)$ for every group of Pareto surfaces A, B , and C).

4.2.2.1 Root-mean-square error

Proof of non-negativity: By definition, the l^2 -norm $\|\vec{x}_i - \vec{y}_i\|_2$ is non-negative, so the summation $\sum_{i=1}^s \|\vec{x}_i - \vec{y}_i\|_2^2$ is also non-negative. Since square roots on non-negative numbers are also non-negative, the RMSE is also non-negative.

Proof of identity of indiscernibles: By definition,

$$RMSE(X, X) = \sqrt{\frac{\sum_{i=1}^s \|\vec{x}_i - \vec{x}_i\|_2^2}{s}} = \sqrt{\frac{\sum_{i=1}^s 0^2}{s}} = \sqrt{\frac{0}{s}} = 0$$

Now, assume that X and Y are Pareto surface samplings satisfying $RMSE(X, Y) = 0$. Then,

$$\sqrt{\frac{\sum_{i=1}^s \|\vec{x}_i - \vec{y}_i\|_2^2}{s}} = 0 \Rightarrow \sum_{i=1}^s \|\vec{x}_i - \vec{y}_i\|_2^2 = 0 \Rightarrow \|\vec{x}_i - \vec{y}_i\|_2 = 0 \forall i \in \{1, \dots, s\}$$

Since the l^2 -norm has the identity of indiscernibles, this implies that $\vec{x}_i = \vec{y}_i \forall i \in \{1, \dots, s\}$, so $X = Y$. Therefore, the RMSE has the identity of indiscernibles.

Proof of symmetry: By definition,

$$\begin{aligned} RMSE(X, Y) &= \sqrt{\frac{\sum_{i=1}^s \|\vec{x}_i - \vec{y}_i\|_2^2}{s}} \\ &= \sqrt{\frac{\sum_{i=1}^s \|-(\vec{y}_i - \vec{x}_i)\|_2^2}{s}} \\ &= \sqrt{\frac{\sum_{i=1}^s \|\vec{y}_i - \vec{x}_i\|_2^2}{s}} \\ &= RMSE(Y, X) \end{aligned}$$

Therefore, the RMSE is symmetric.

Proof of sub-additivity: By definition,

$$\begin{aligned} (RMSE(X, Z) + RMSE(Z, Y))^2 &= \left(\sqrt{\frac{\sum_{i=1}^s \|\vec{x}_i - \vec{z}_i\|_2^2}{s}} + \sqrt{\frac{\sum_{i=1}^s \|\vec{z}_i - \vec{y}_i\|_2^2}{s}} \right)^2 \\ &= \frac{\sum_{i=1}^s \|\vec{x}_i - \vec{z}_i\|_2^2}{s} + \frac{\sum_{i=1}^s \|\vec{z}_i - \vec{y}_i\|_2^2}{s} + 2 \sqrt{\frac{\sum_{i=1}^s \|\vec{x}_i - \vec{z}_i\|_2^2}{s}} \sqrt{\frac{\sum_{i=1}^s \|\vec{z}_i - \vec{y}_i\|_2^2}{s}} \\ &\geq \frac{\sum_{i=1}^s \|\vec{x}_i - \vec{z}_i\|_2^2}{s} + \frac{\sum_{i=1}^s \|\vec{z}_i - \vec{y}_i\|_2^2}{s} \\ &= \frac{\sum_{i=1}^s \left(\|\vec{x}_i - \vec{z}_i\|_2^2 + \|\vec{z}_i - \vec{y}_i\|_2^2 \right)}{s} \end{aligned}$$

$$\geq \frac{\sum_{i=1}^s \|\bar{x}_i - \bar{y}_i\|_2^2}{s} = RMSE(X, Y)^2$$

Therefore, $(RMSE(X, Z) + RMSE(Z, Y))^2 \geq RMSE(X, Y)^2$, so $RMSE(X, Y) \leq MSE(X, Z) + RMSE(Z, Y)$, proving the sub-additivity of the root-mean-square error.

4.2.2.2 Hausdorff distance

Proof of non-negativity: By definition, the l^2 -norm $\|\bar{x} - \bar{y}\|_2$ is non-negative, so the infimum $\inf_{x \in X} \|\bar{x} - \bar{y}\|_2$ is also non-negative and the supremum $\sup_{y \in Y} \inf_{x \in X} \|\bar{x} - \bar{y}\|_2$ is also non-negative. Therefore, the Hausdorff distance is non-negative.

Proof of identity of indiscernibles: Since X is a finite set, infima on X are attained, so $\inf_{x' \in X} \|\bar{x} - \bar{x}'\|_2 = 0$. So, by definition,

$$\begin{aligned} HD(X, X) &= \max \left\{ \sup_{x \in X} \inf_{x' \in X} \|\bar{x} - \bar{x}'\|_2, \sup_{x \in X} \inf_{x' \in X} \|\bar{x} - \bar{x}'\|_2 \right\} \\ &= \max \left\{ \sup_{x' \in X} 0, \sup_{x \in X} 0 \right\} = 0 \end{aligned}$$

Now, suppose that X and Y are Pareto surface samplings such that $HD(X, Y) = 0$.

Then, $\sup_{y \in Y} \inf_{x \in X} \|\bar{x} - \bar{y}\|_2 = 0$ and $\sup_{x \in X} \inf_{y \in Y} \|\bar{x} - \bar{y}\|_2 = 0$. Since X and Y are finite sets,

suprema and infima of continuous functions are attained, so $\forall y \in Y, \exists x \in X$ such that

$\|\bar{x} - \bar{y}\|_2 = 0$, i.e. $y = x$, so $Y \subset X$. Similarly, $X \subset Y$, so $X = Y$. Therefore, the Hausdorff

distance has the identity of indiscernibles.

Proof of symmetry: By definition,

$$HD(X, Y) = \max \left\{ \sup_{y \in Y} \inf_{x \in X} \|\bar{x} - \bar{y}\|_2, \sup_{x \in X} \inf_{y \in Y} \|\bar{x} - \bar{y}\|_2 \right\}$$

$$\begin{aligned}
&= \max \left\{ \sup_{x \in X} \inf_{y \in Y} \|\vec{x} - \vec{y}\|_2, \sup_{y \in Y} \inf_{x \in X} \|\vec{x} - \vec{y}\|_2 \right\} \\
&= HD(Y, X)
\end{aligned}$$

Therefore, the Hausdorff distance is symmetric.

Proof of sub-additivity: Since X, Y , and Z are finite sets, suprema and infima of continuous functions are attained on these sets. So, consider a point $x_0 \in S(X)$, and let $z_0 \in S(Z)$ such that $\|x_0 - z_0\|_2 = \inf_{z \in S(Z)} \|x_0 - z\|_2$. Then let $y_0 \in Y$ such that $\inf_{y \in S(Y)} \|y - z_0\|_2 = \|y_0 - z_0\|_2$. Then,

$$\begin{aligned}
\inf_{y \in S(Y)} \|x_0 - y\|_2 &\leq \|x_0 - y_0\|_2 \\
&\leq \|x_0 - z_0\|_2 + \|y_0 - z_0\|_2 \\
&= \inf_{z \in S(Z)} \|x_0 - z\|_2 + \inf_{y \in S(Y)} \|y - z_0\|_2 \\
&\leq \inf_{z \in S(Z)} \|x_0 - z\|_2 + \sup_{z \in S(Z)} \inf_{y \in S(Y)} \|y - z\|_2 \\
&\leq \sup_{x \in S(X)} \inf_{z \in S(Z)} \|x - z\|_2 + \sup_{z \in S(Z)} \inf_{y \in S(Y)} \|y - z\|_2 \\
&\leq HD(S(X), S(Z)) + HD(S(Y), S(Z))
\end{aligned}$$

Since this is true for every point within $S(X)$, the inequality extends over the set:

$$\sup_{x \in S(X)} \inf_{y \in S(Y)} \|x - y\|_2 \leq HD(S(X), S(Z)) + HD(S(Y), S(Z))$$

Similar logic applies to the other half of the maximum function in the Hausdorff distance:

$$\sup_{y \in S(Y)} \inf_{x \in S(X)} \|x - y\|_2 \leq HD(S(X), S(Z)) + HD(S(Y), S(Z))$$

Combining these two results applies the inequality to the Hausdorff distance

itself:

$$\begin{aligned} HD(S(X), S(Y)) &= \max \left\{ \sup_{x \in S(X)} \inf_{y \in S(Y)} \|x - y\|_2, \sup_{y \in S(Y)} \inf_{x \in S(X)} \|x - y\|_2 \right\} \\ &\leq HD(S(X), S(Z)) + HD(S(Y), S(Z)) \end{aligned}$$

Therefore, the Hausdorff distance has the sub-additivity property.

4.2.2.3 Average projected distance

Proof through re-definition: As shown in Chapter 4.2.1.3, the APD can be written

as

$$APD(X, Y) = \frac{\frac{1}{n}(X_M - Y_M)\vec{1} \cdot \overline{\mu_{XY}}}{\|\mu_{XY}\|_2}$$

where

$$\overline{\mu_{XY}} = \frac{1}{n}(X_M - Y_M)\vec{1}$$

Therefore, we can write

$$APD(X, Y) = \frac{\overline{\mu_{XY}} \cdot \overline{\mu_{XY}}}{\|\mu_{XY}\|_2} = \|\mu_{XY}\|_2$$

Since the average projected distance is an l^2 -norm, it is a distance metric in the formal, mathematical sense.

4.2.2.4 Average nearest-point distance

Proof of non-negativity: By definition, the l^2 -norm $\|\vec{x} - \vec{y}\|_2$ is non-negative, so the infimum $\inf_{x \in S(X)} \|\vec{x} - \vec{y}\|_2$ is also non-negative and the average $\text{avg}_{y \in S(Y)} \inf_{x \in S(X)} \|\vec{x} - \vec{y}\|_2$ is also non-negative. Therefore, the average nearest-point distance is non-negative.

Proof of identity of indiscernibles: By definition,

$$\begin{aligned} ANPD(S(X), S(X)) &= \text{avg} \left\{ \text{avg}_{x \in S(X)} \inf_{x' \in S(X)} \|x - x'\|_2, \text{avg}_{x \in S(X)} \inf_{x' \in S(X)} \|x - x'\|_2 \right\} \\ &= \text{avg} \left\{ \text{avg}_{x \in S(X)} 0, \text{avg}_{x' \in S(X)} 0 \right\} = 0 \end{aligned}$$

Now, suppose that X and Y are Pareto surface samplings such that

$$ANPD(S(X), S(Y)) = 0. \text{ Then, } \text{avg}_{y \in S(Y)} \inf_{x \in S(X)} \|\vec{x} - \vec{y}\|_2 = 0 \text{ and } \text{avg}_{x \in S(X)} \inf_{y \in S(Y)} \|\vec{x} - \vec{y}\|_2 = 0.$$

Since the l^2 -norm is non-negative, this implies that $\inf_{x \in S(X)} \|\vec{x} - \vec{y}\|_2 = 0 \forall y \in S(Y)$. Since $S(X)$ and $S(Y)$ are closed sets, suprema and infima of continuous functions are attained, so $\forall y \in S(Y), \exists x \in S(X)$ such that $y = x$, so $S(Y) \subset S(X)$. Similarly, $S(X) \subset Y$, so $S(X) = S(Y)$. Therefore, the average nearest-point distance has the identity of indiscernibles.

Proof of symmetry: By definition,

$$\begin{aligned} ANPD(S(X), S(Y)) &= \text{avg} \left\{ \text{avg}_{x \in S(X)} \inf_{y \in S(Y)} \|x - y\|_2, \text{avg}_{y \in S(Y)} \inf_{x \in S(X)} \|x - y\|_2 \right\} \\ &= \text{avg} \left\{ \text{avg}_{y \in S(Y)} \inf_{x \in S(X)} \|x - y\|_2, \text{avg}_{x \in S(X)} \inf_{y \in S(Y)} \|x - y\|_2 \right\} \\ &= ANPD(S(Y), S(X)). \end{aligned}$$

Therefore, the average nearest-point distance is symmetric.

Proof of sub-additivity: Since the simplicial complexes formed by the samples X, Y , and Z are closed sets, infima of continuous functions are attained on them. So, consider a point $x \in S(X)$, and let $z_0 \in Z$ such that $\|x - z_0\|_2 = \inf_{z \in Z} \|x - z\|_2$. Then let $y_0 \in Y$ such that $\inf_{y \in S(Y)} \|y - z_0\|_2 = \|y_0 - z_0\|_2$. Then,

$$\begin{aligned} \inf_{y \in S(Y)} \|x - y\|_2 &\leq \|x - y_0\|_2 \\ &\leq \|x - z_0\|_2 + \|y_0 - z_0\|_2 \\ &= \inf_{z \in S(Z)} \|x - z\|_2 + \inf_{y \in S(Y)} \|y - z_0\|_2 \end{aligned}$$

Since x was chosen arbitrarily, this extends across $S(X)$:

$$\begin{aligned} \text{avg}_{x \in S(X)} \inf_{y \in S(Y)} \|x - y\|_2 &\leq \text{avg}_{x \in S(X)} \inf_{z \in S(Z)} \|x - z\|_2 + \text{avg}_{x \in S(X)} \inf_{y \in S(Y)} \|y - z_0(x)\|_2 \\ &\leq \text{avg}_{x \in S(X)} \inf_{z \in S(Z)} \|x - z\|_2 + \text{avg}_{z \in S(Z)} \inf_{y \in S(Y)} \|y - z\|_2 \end{aligned}$$

By applying similar logic to both sides of the average in the definition of the average nearest-point distance, we have that $ANPD(X, Y) \leq ANPD(X, Z) + ANPD(Y, Z)$, proving that the average nearest-point distance has the property of sub-additivity.

4.2.3 Theoretical Pareto surfaces for comparison

All of the anatomical treatment sites included in this study have either two or three trade-off objectives. Specifically, prostate VMAT includes PTV homogeneity index, bladder $D_{25\%}$, and rectum $D_{25\%}$ (three objectives); spine SRS includes PTV homogeneity index, epidural space $D_{95\%}$, and spinal cord $D_{10\%}/D_{\max}$ (three objectives, with the spinal cord objectives merged to share the same objective priorities); and pancreas SBRT includes PTV homogeneity and small bowel/stomach $D_{0.5cc}$ (two objectives, with the

small bowel and stomach objectives merged to share the same objective priorities).

Therefore, it is particularly important for this study to analyze and compare our Pareto surface metrics in the two-dimensional and three-dimensional cases.

To address these needs, three pairs of simplices were generated in two dimensions and three dimensions, which will be used for metric comparison and analysis later in this Chapter. Since simplices are the constituents of Pareto surfaces, the errors between simplices should extend to Pareto surfaces in general. The time analysis between simplices also extends to Pareto surfaces linearly, i.e. the time requirement of computing the distance between two Pareto surfaces, each with M and N sampled points, is MNT , where T is the time requirement of computing that distance for a pair of simplices. Moreover, the range of possible Pareto surfaces is very large, so it is not feasible to create enough cases to test the possible scenarios. Metrics were computed using Python and an Intel Xeon CPU.

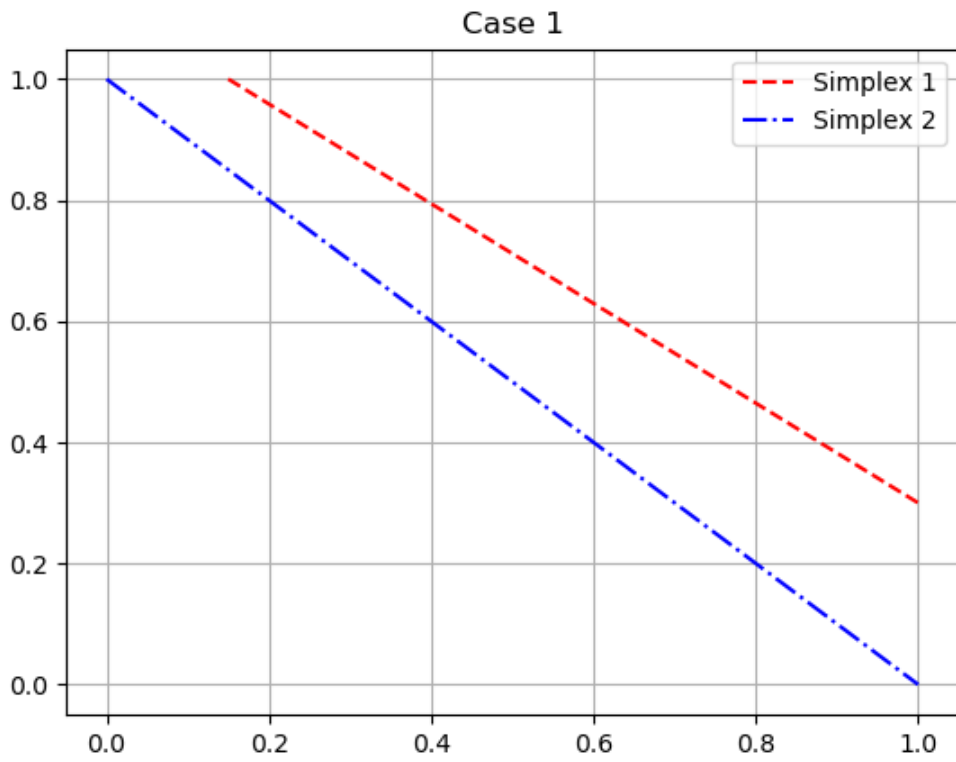


Figure 14: Case 1 for simplex testing: Distance from $\{(0, 1); (1, 0)\}$ to $\{(0.15, 1); (1, 0.3)\}$. The purpose of including this case is to test our metrics on Pareto surfaces of similar sizes, shapes, and orientations in two dimensions.

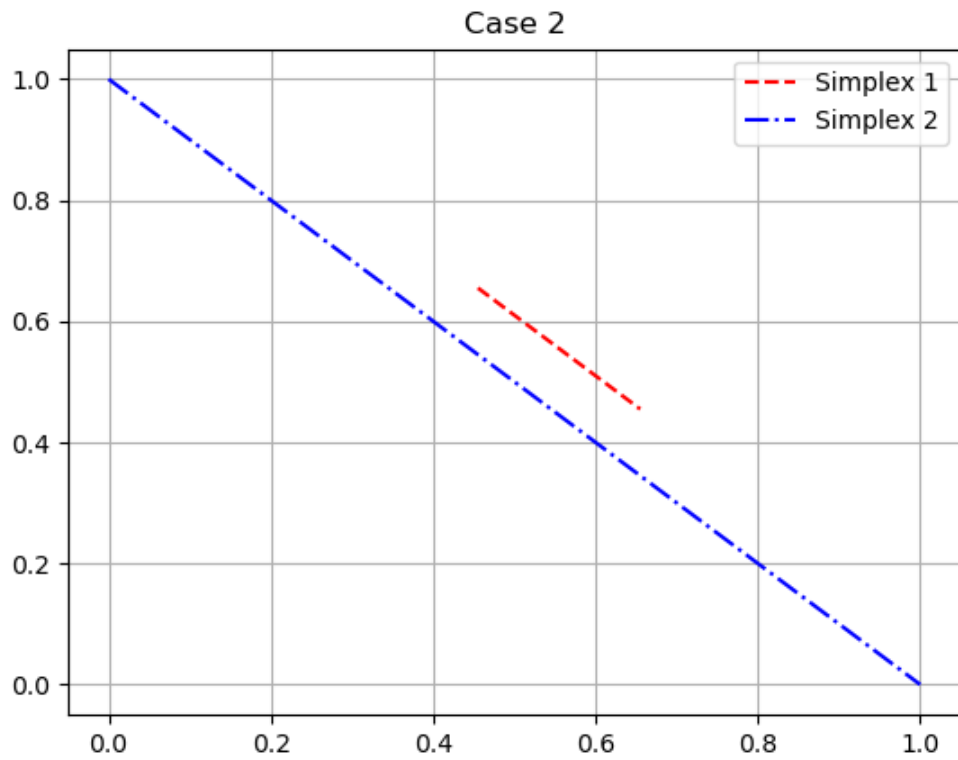


Figure 15: Case 2 for simplex testing: Distance from $\{(0, 1); (1, 0)\}$ to $\{(0.455, 0.655); (0.655, 0.455)\}$. The purpose of including this case is to determine the effect of significantly different surface size on ANPD convergence.

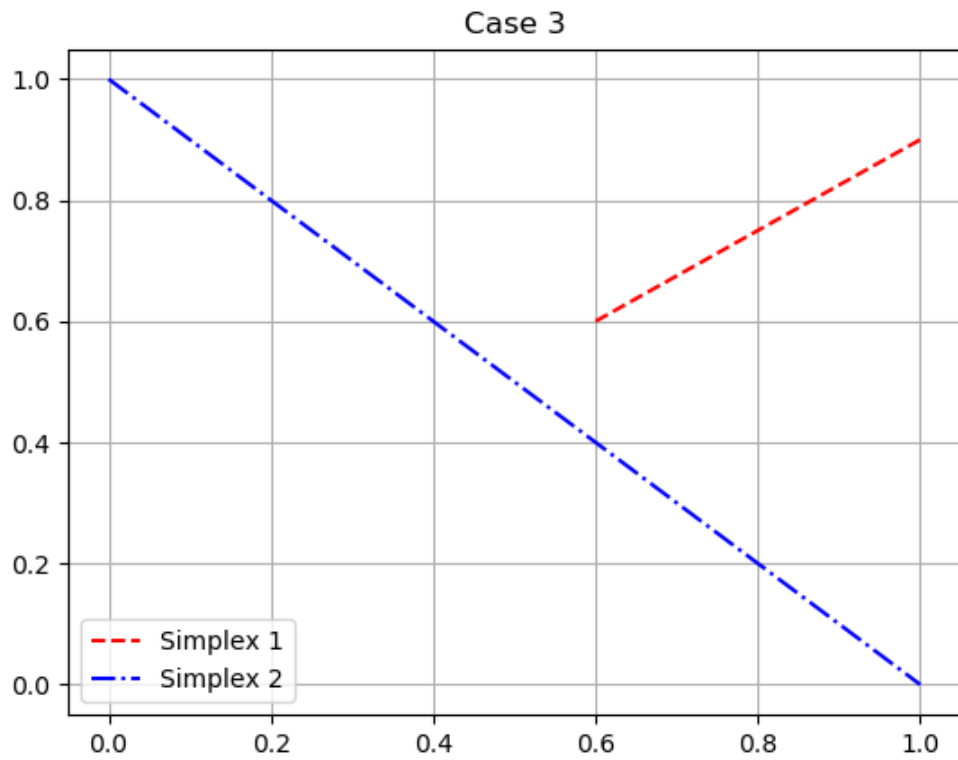


Figure 16: Case 3 for simplex testing: Distance from $\{(0, 1); (1, 0)\}$ to $\{(0.6, 0.6); (1, 0.9)\}$. The purpose of including this case is to determine the effect of significantly different orientations on ANPD convergence.

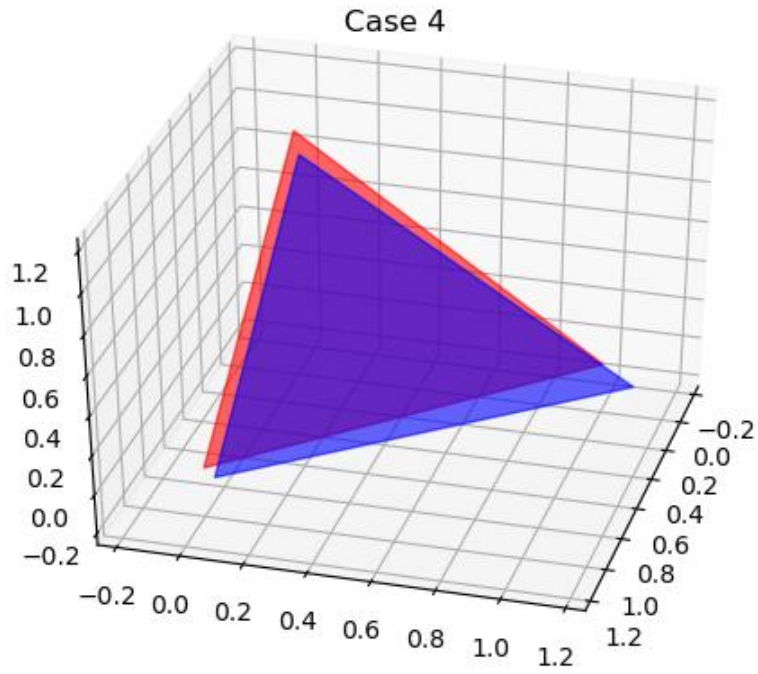


Figure 17: Case 4 for simplex testing: Distance from $\{(1, 0, 0); (0, 1, 0); (0, 0, 1)\}$ to $\{(0, 1.1, -0.1); (1.05, 0.05, 0); (-0.05, 0, 0.85)\}$. The purpose of including this case is to test our metrics on Pareto surfaces of similar sizes, shapes, and orientations in three dimensions.

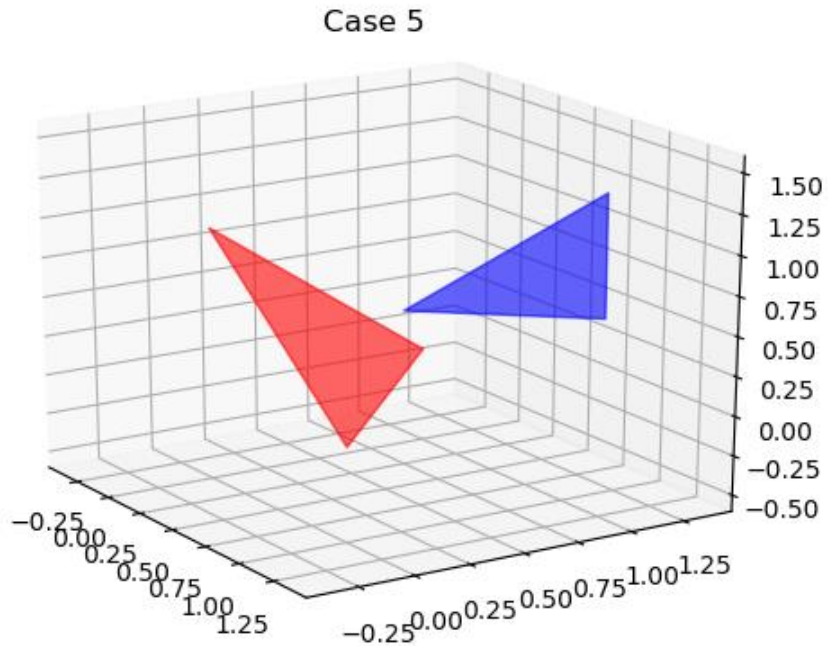


Figure 18: Case 5: Distance from $\{(1, 0, 0); (0, 1, 0); (0, 0, 1)\}$ to $\{(0.55, 0.55, 0.55); (1.55, 0.8, 0.8); (1.55, 0.8, 1.55)\}$. The purpose of including this case is to determine the effect of variable surface distances and significantly different surface orientations on ANPD convergence.

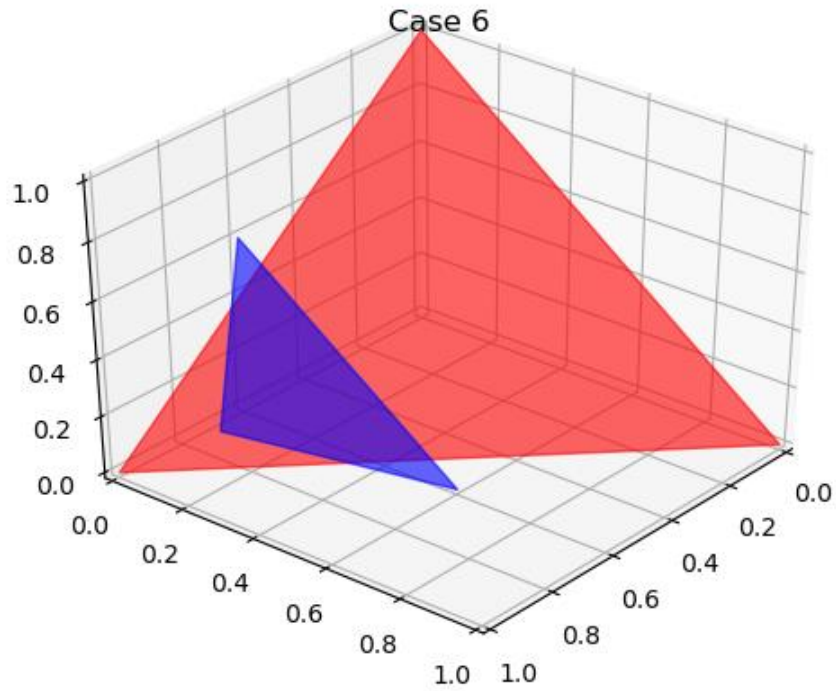


Figure 19: Case 6 for simplex testing: Distance from $\{(1, 0, 0); (0, 1, 0); (0, 0, 1)\}$ to $\{(0.6, 0.6, 0); (0.6, 0, 0.6); (1, 0.3, 0.3)\}$. The purpose of including this case is to determine the effect of significantly different surface orientations and sizes on ANPD convergence.

4.3 Results

4.3.1 Metric comparison and convergence analysis

Figures 20 - 25 depict the Pareto surface distances for the four metrics presented in Chapter 4.2.1 on the six cases presented in Chapter 4.2.3. Since the average nearest point distance (ANPD) depends on the number of intra-simplex samples taken during computation, the ANPD is displayed as a graphical function of the number of samples per barycentric dimension. Since the root-mean-square error (RMSE), Hausdorff distance (HD), and average projected distance (APD) do not require an internal sampling, these metrics are represented by dashed horizontal lines for ease of visual comparison. However, these metrics can only take a single possible value and are not actually functions of the number of samples. Tables 1 and 2 contain the number of samples required for the ANPD and RMSE to be within a range of tolerances of their limits. Here, we can see that convergence within 1% is typically achieved at approximately 100 samples per barycentric dimension for the ANPD and at approximately 50 samples per barycentric dimension for the RMSE.

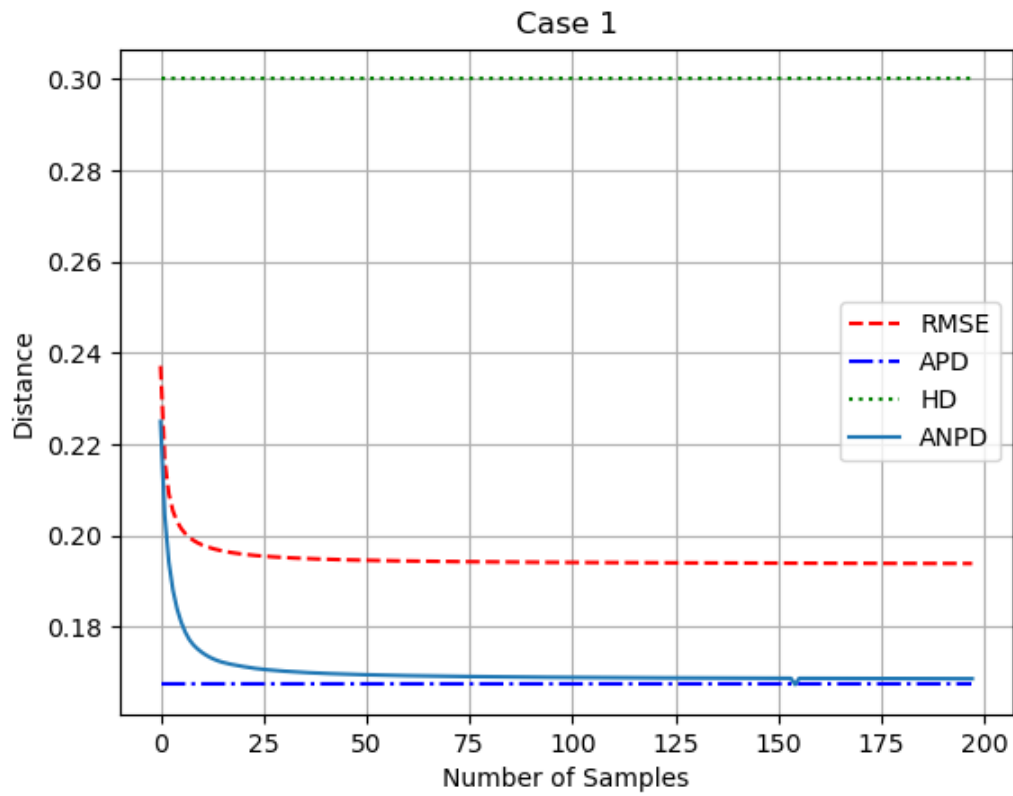


Figure 20: Distance metrics for Case 1 as a function of number of samples per barycentric dimension. Here, the ANPD appears to converge quickly, and the APD appears to be similar to the final value of the ANPD.

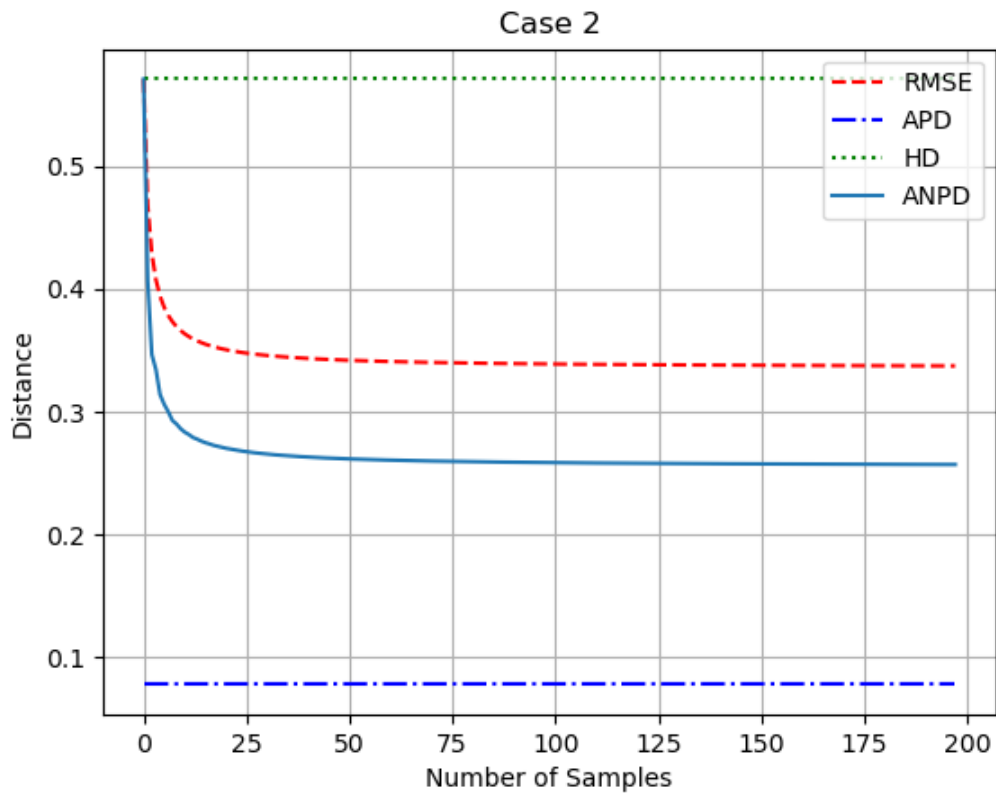


Figure 21: Distance metrics for Case 2 as a function of number of samples per barycentric dimension. Here, the ANPD appears to converge quickly. However, the APD appears to be different than the limit of the ANPD, and the RMSE converges closer to the limit of the ANPD.

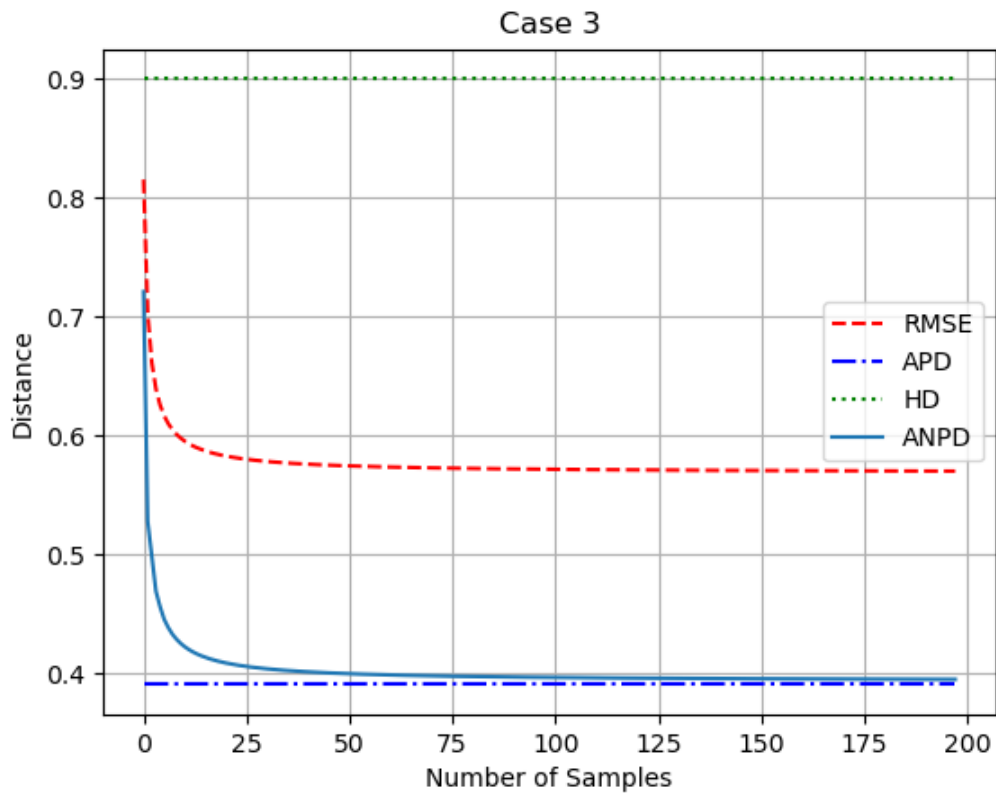


Figure 22: Distance metrics for Case 3 as a function of number of samples per barycentric dimension. Here, the ANPD appears to converge quickly, and the APD appears to be similar to the limit of the ANPD.

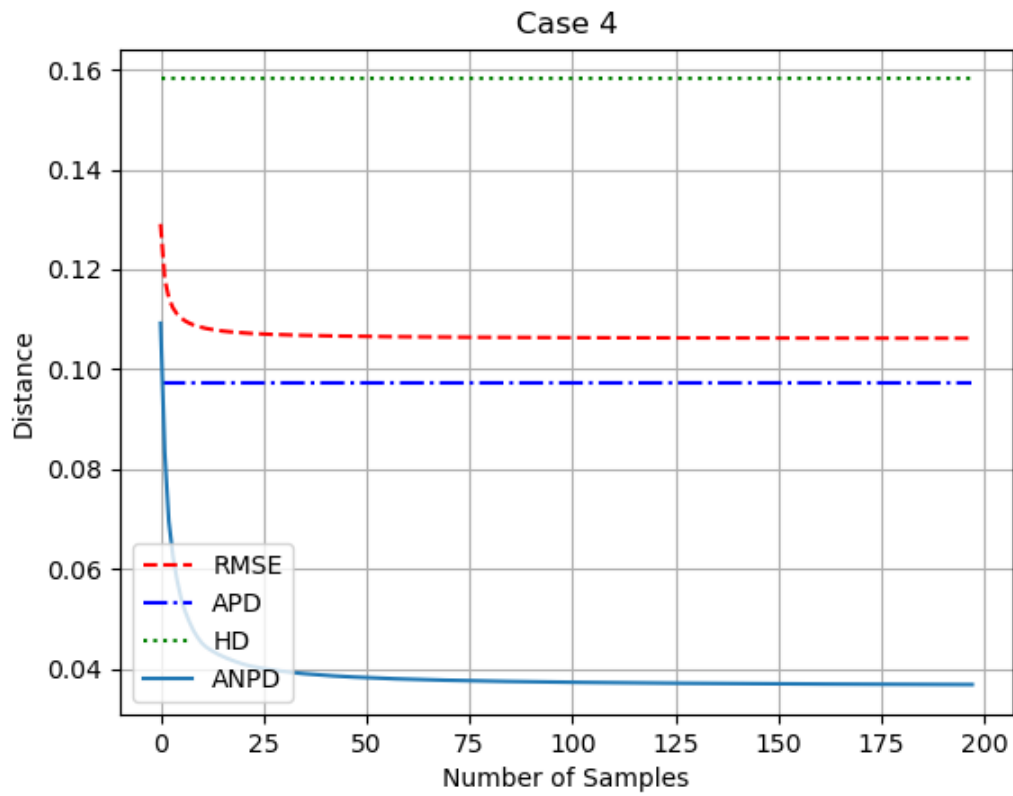


Figure 23: Distance metrics for Case 4 as a function of number of samples per barycentric dimension. Here, the ANPD appears to converge quickly, and the APD appears to be more similar to the limit of the RMSE than the limit of the ANPD.

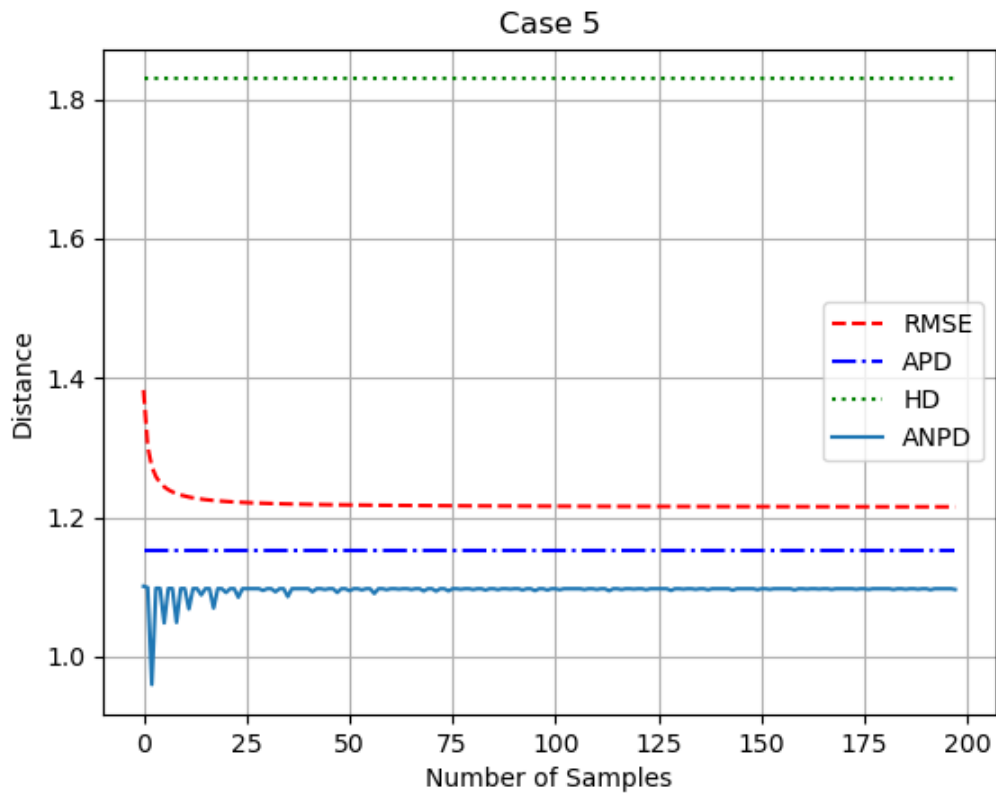


Figure 24: Distance metrics for Case 5 as a function of number of samples per barycentric dimension. Here, the ANPD appears to converge quickly, and the APD appears to be similar to the limit of the ANPD.

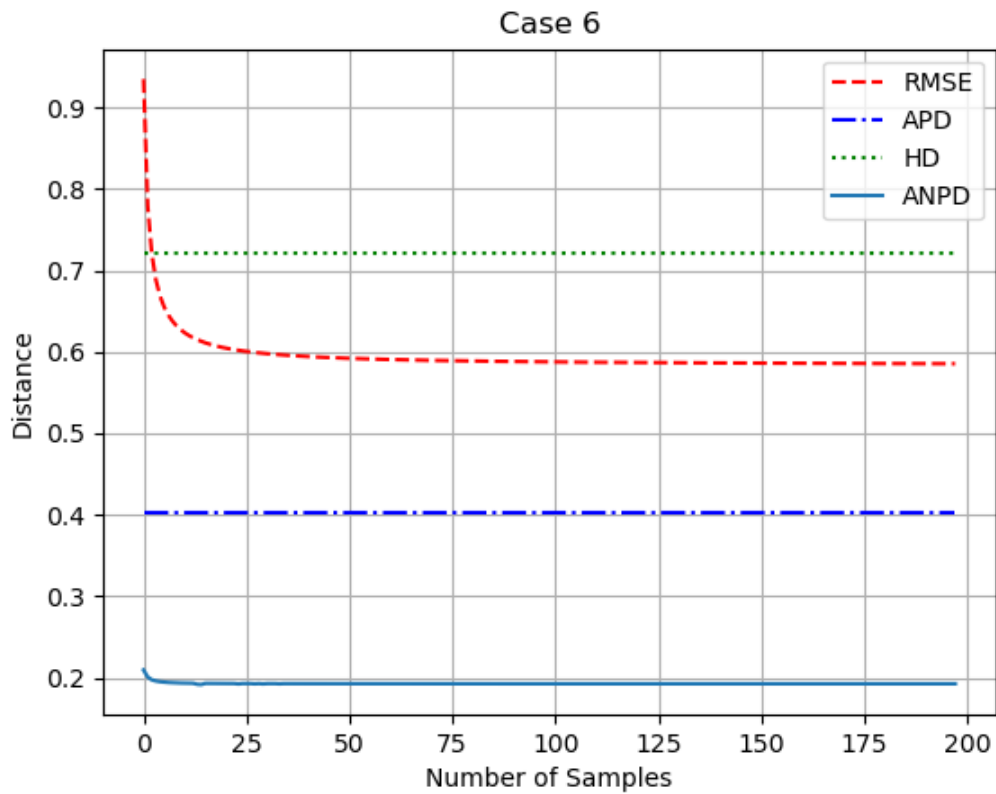


Figure 25: Distance metrics for Case 6 as a function of number of samples per barycentric dimension. Here, the ANPD appears to converge quickly. However, the APD appears to be different than the final value of the ANPD. Additionally, ANPD convergence is particularly fast for this case.

Table 1: The number of simplex samples per barycentric dimension required to yield an ANPD within specific convergence thresholds of its limit.

Convergence	10%	3%	1%	0.3%
Threshold				
Case 1	5	11	29	154
Case 2	12	34	76	134
Case 3	8	24	59	116
Case 4	23	60	112	161
Case 5	4	10	25	58
Case 6	2	3	7	16

Table 2: The number of simplex samples per barycentric dimension required to yield an RMSE within specific convergence thresholds of its limit.

Convergence	10%	3%	1%	0.3%
Threshold				
Case 1	2	8	23	59
Case 2	9	27	64	122
Case 3	5	16	42	93
Case 4	3	8	22	57
Case 5	2	5	14	40
Case 6	8	21	56	113

4.3.2 Time analysis

Figures 26 - 31 depict the time required to compute the Pareto surface distances for the four metrics presented in Chapter 4.2.1 on the six cases presented in Chapter 4.3. Since the average nearest point distance (ANPD) and root-mean-square error (RMSE) depend on the number of intra-simplex samples taken during computation, they are displayed as a graphical function of the number of samples per barycentric dimension. Since Hausdorff distance (HD) and average projected distance (APD) do not require an internal sampling, these metrics are represented by dashed horizontal lines for ease of visual comparison. However, these metrics can only take a single possible value and are not actually functions of the number of samples. Tables 3 and 4 contain the computation time required for the ANPD and RMSE to be within a range of tolerances of their limits, coinciding with Tables 1 and 2 in Chapter 4.3.1. Here, we can see that convergence within 1% is typically achieved in less than 3 seconds for the ANPD and less than 50 milliseconds for the RMSE. Therefore, HD and APD calculations tend to be much faster than RMSE calculations, which in turn tend to be much faster than ANPD calculations.

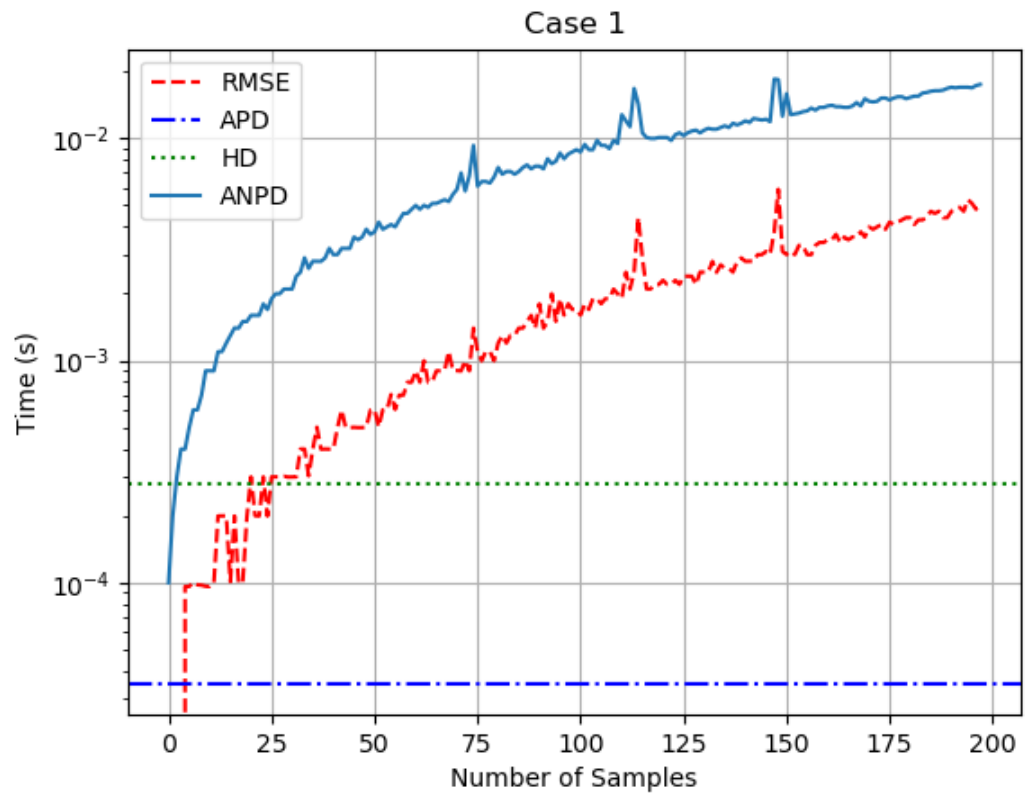


Figure 26: Metric computation times for Case 1.

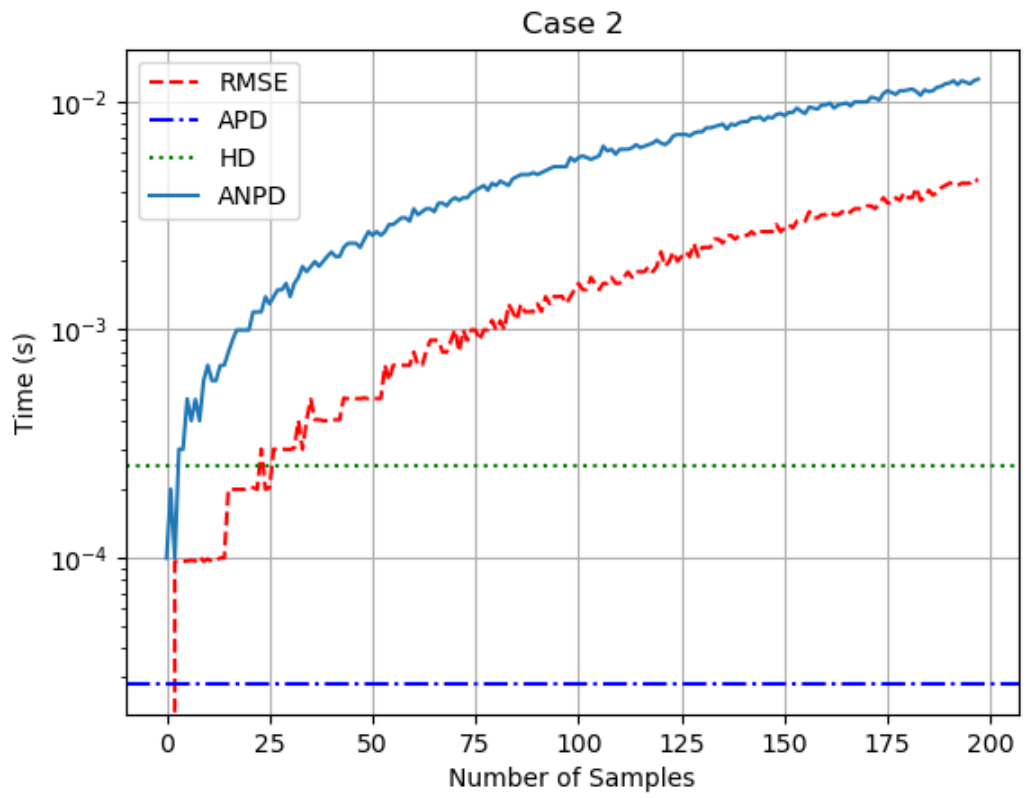


Figure 27: Metric computation times for Case 2.

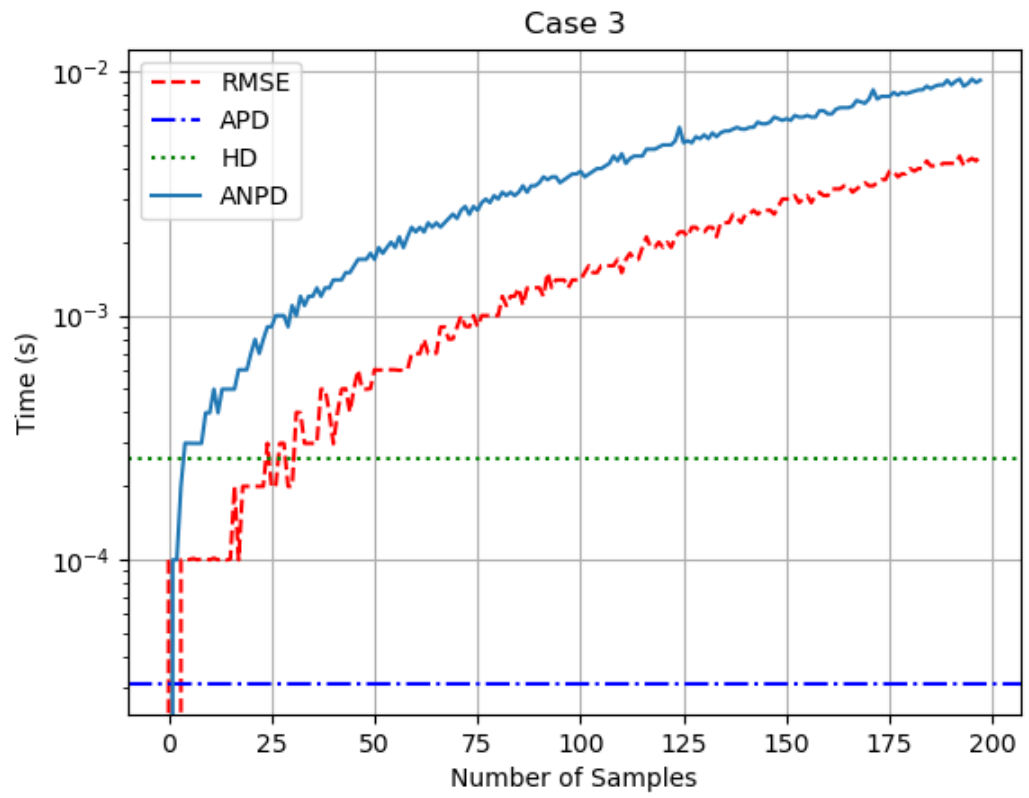


Figure 28: Metric computation times for Case 3.

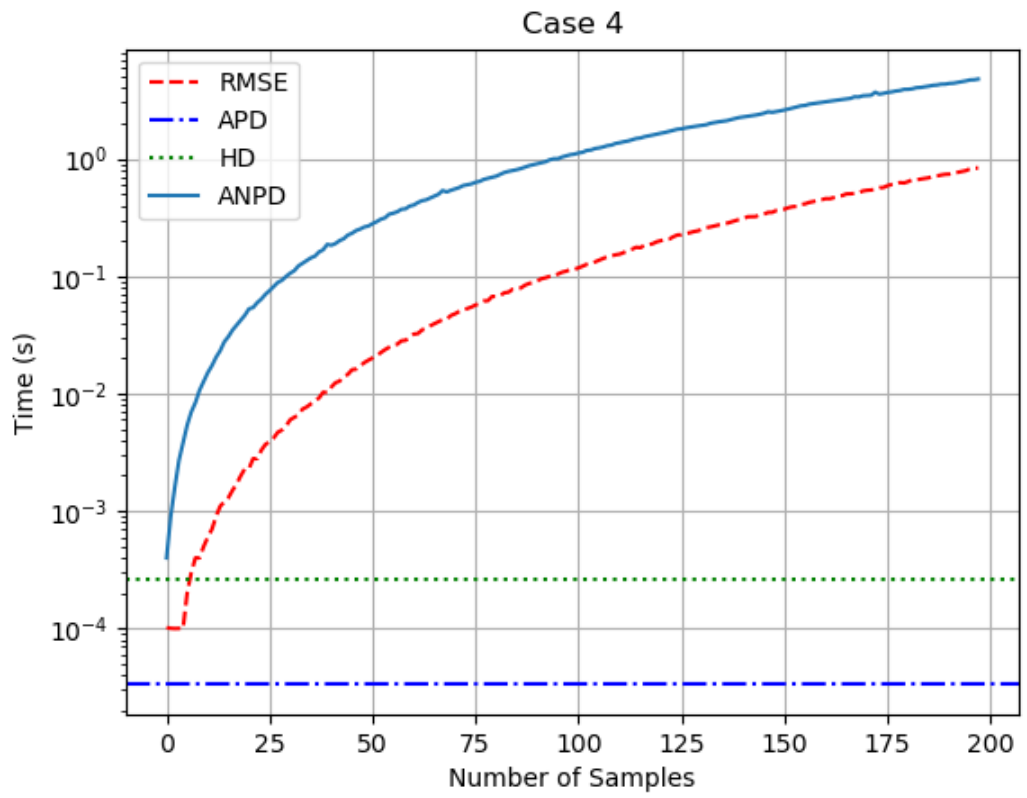


Figure 29: Metric computation times for Case 4.

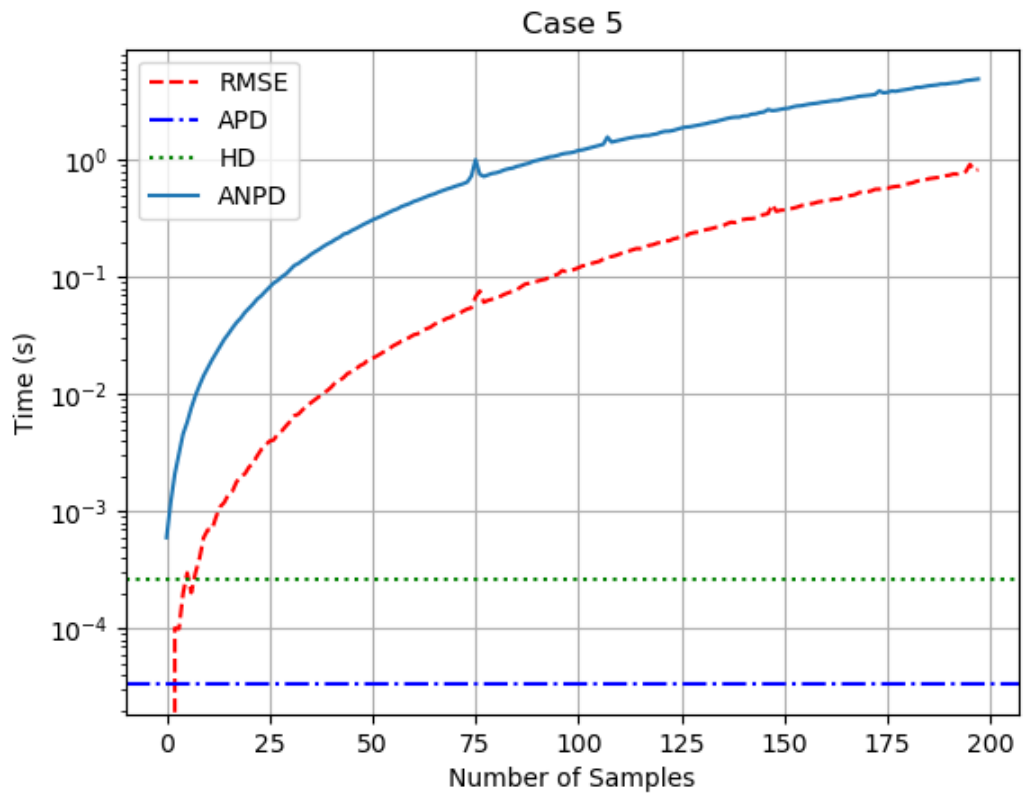


Figure 30: Metric computation times for Case 5.

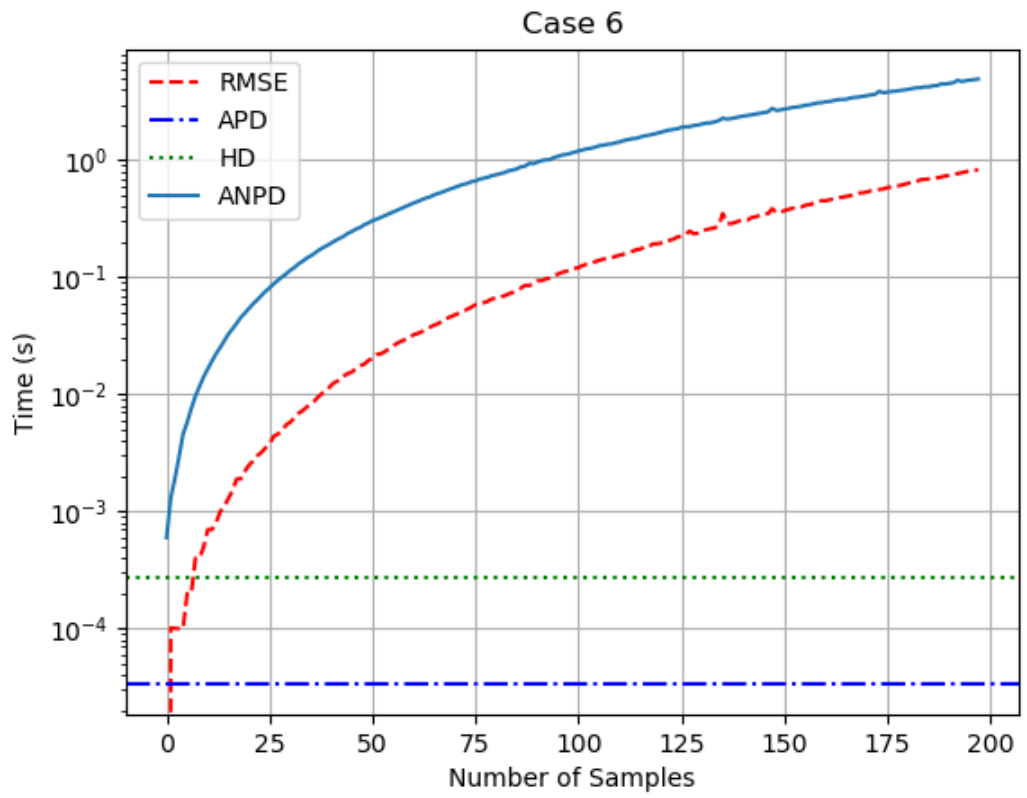


Figure 31: Metric computation times for Case 6.

Table 3: Time required to yield an ANPD within specific convergence thresholds of its limit.

Convergence	10%	3%	1%	0.3%
Threshold				
Case 1	1.02 ms	2.03 ms	4.97 ms	32.9 ms
Case 2	0.998 ms	3.04 ms	7.99 ms	19.9 ms
Case 3	0.986 ms	1.03 ms	4.02 ms	9.97 ms
Case 4	0.103 s	0.742 s	2.63 s	5.69 s
Case 5	3.99 ms	23.0 ms	136 ms	707 ms
Case 6	1.03 ms	1.99 ms	11.0 ms	54.9 ms

Table 4: Time required to yield an RMSE within specific convergence thresholds of its limit.

Convergence	10%	3%	1%	0.3%
Threshold				
Case 1	0.942 ms	0.967 ms	0.997 ms	1.00 ms
Case 2	0.996 ms	9.98 ms	20.0 ms	39.9 ms
Case 3	0.997 ms	1.03 ms	0.998 ms	4.99 ms
Case 4	0.995 ms	1.00 ms	3.99 ms	52.9 ms
Case 5	0.998 ms	0.997 ms	20.0 ms	21.9 ms
Case 6	0.997 ms	5.98 ms	46.9 ms	316 ms

4.4 Discussion

Our results show that convergence to within 1% is typically achieved in less than 3 seconds for the ANPD and less than 50 milliseconds for the RMSE. Based on this analysis, the root-mean-square error (RMSE) calculations tend to be much faster than the average nearest-point distance (ANPD) calculations. However, in general, the RMSE did not converge to the limit of the ANPD, indicating that the RMSE may not be applicable in general. However, the convergence for the RMSE is much faster than the convergence for the ANPD, so the RMSE may be an applicable choice when the computation times for the ANPD impractical. For the purpose of this work, there were no significant time constraints for evaluating the ANPD, so the ANPD will be used to analyze the quality of predicted Pareto surfaces in future Chapters.

The average projected distance (APD) had average computation times less than 1 millisecond and it did not require any simplicial upsampling during calculation. Compared to the metrics presented in this Chapter, the APD seems to be significantly faster than the RMSE and ANPD when upsampled to create 99% convergence of their limits. The APD also appears to be approximately ten times faster to compute than the Hausdorff distance. While this is desirable, the APD may be more or less appropriate for evaluation Pareto surface similarities. Particularly for case 4, the APDs were much smaller than both the ANPDs and the RMSEs. Since the APD is not guaranteed to

produce an appropriate distance, it should be used primarily in cases where the computation times for the ANPD and RMSE are too large for practical application.

In this study, we proposed to calculate the ANPD by upsampling one Pareto surface and calculating the distance from each of the upsampled points to the simplices on the other Pareto surface. Fortunately, distance functions tend to be smooth and well-behaved, so they tend to converge quickly when computed via numerical integration. We were able to confirm this for both the RMSE, which tended to converge within 1% of its limiting value after taking approximately 50 samples per barycentric dimension, as well as the ANPD, which tended to converge within 1% of its limiting value after taking approximately 100 samples per barycentric dimension. However, performing an actual analytic integration of these point-to-simplex functions would be beneficial to implement because an analytic integration would be significantly faster, and it would remove the subjectivity involved in specifying the number of samples to acquire.

It is unfortunate that Johnson's distance algorithm was the only point-to-simplex distance calculation method available to us. Although the time cost of the algorithm was feasible in the two- and three-dimensional cases studied in this Chapter, the time cost very quickly becomes not feasible in higher-dimensional cases. Therefore, while Johnson's distance algorithm was usable in this study, it is desirable for a faster point-to-simplex distance algorithm to be developed. However, Johnson's algorithm is numerically accurate, so the convergence analysis presented in this Chapter will likely

apply to any other point-to-simplex distance algorithm that may be implemented in the future.

Our results generally indicate that the ANPD is stable with respect to intra-simplex upsampling, and the ANPD convergence rates are rather fast. Therefore, the ANPD can be used as an appropriate metric for evaluating Pareto surface interpolation similarities. Additionally, the ANPD would likely be useful for powering future studies which seek to compare different MCO algorithms, such as those provided by Varian Medical Systems and Raysearch Laboratories.

Bokrantz and Forsgren originally used the Hausdorff distance as their metric of choice to compare upper and lower bounds of the Pareto surface as an intermediate step in their MCO software (Bokrantz and Forsgren, 2013). Their Pareto surface generation algorithm greedily minimizes the Hausdorff distance between these upper and lower bounds during Pareto surface construction in MCO. Their framework assumes that the dose optimization is convex, allowing the upper and lower bounds of the Pareto surface to be inferred exactly. As such, the Hausdorff distance is a particularly appropriate metric for efficient minimization of Pareto surface uncertainty.

The Hausdorff distance is also particularly appropriate for use in quality assurance in the clinical setting while evaluating the maximum error that a machine learning product may generate. Clinically, much of the focus in quality assurance lies on controlling the maximum error of a given process rather than the average error.

However, due to the stochastic nature of outliers in machine learning models and the sensitivity of the Hausdorff distance to outliers, the randomness in the Hausdorff distances between Pareto surfaces may outweigh the differences in quality between the models. Therefore, the Hausdorff distance is a poor metric for comparing and evaluating models during the early stages of model development.

4.5 Conclusion

In this Chapter, we have presented, compared, and analyzed several Pareto surface metrics for use in MCO prediction assessment. Based on our analysis, we believe that the average nearest-point distance is a suitable metric for Pareto surface comparison, with metric convergence being reached at approximately 100 interior samples per barycentric dimension per simplex. This metric evaluation costs approximately 10 milliseconds per pair of simplices in 2D and 5 seconds per pair of simplices in 3D. When possible, it is recommended to use the ANPD to achieve the most accurate and appropriate distance metric. However, significant accelerations may be achieved by using either the RMSE or APD metrics for more time-sensitive applications.

5. Evaluation of epidural space irradiation in spinal SBRT

5.1 Introduction

Metastatic disease to the spine is common, presenting in up to 40% of all cancer patients with about half of these patients ultimately becoming symptomatic (Klimo and Schmidt, 2004). Metastatic epidural spinal cord compression (MESCC) can be particularly devastating, producing severe pain and disability, impairing quality of life and decreasing probability of survival (Barzilai et al., 2019, Loblaw et al., 2005). While timely intervention with surgery, radiation therapy or a combination of the two can effectively and safely treat spinal metastases, local recurrence of disease following these treatments is common and challenging to address.

Spinal stereotactic radiosurgery (SSRS) to osseous metastases offers potential advantages over conventionally fractionated external beam radiotherapy (EBRT) in that a high dose of radiation is delivered to the vertebral body at the involved level while minimizing dose to the cord (Kirkpatrick et al., 2014). In contrast, conventional EBRT delivers the full dose of radiation to the osseous spine and its contents (the epidural space and spinal cord or the spinal nerve roots, depending on the level). In our experience, most practitioners of SSRS are extraordinarily conscientious about respecting the recommended dose limitations to the spinal cord due to the severity of spinal cord complications such as radiation-induced myelopathy. Consequently, the

reported rates of spinal cord toxicity with SSRS are extremely low (Grimm et al., 2016, Kirkpatrick et al., 2010, Sahgal et al., 2010).

The desire to minimize the dose to the cord potentially results in the minimization of the dose to the epidural space, which may extend 3-4 mm from the surface of the cord to the interior surface of the vertebral body. Thus, we speculate that microscopic disease in the epidural space may be inadvertently underdosed, even when all dose constraints are met in an SSRS treatment plan. In turn, this may contribute to the moderately high rate of epidural space “failures” observed in SSRS, with reported crude recurrence rates in this region ranging from 6 to 20% (Chang et al., 2007, Garg et al., 2011, Nelson et al., 2009, Oinam et al., 2011, Thibault et al., 2015). We hypothesized that the epidural space could be purposefully irradiated to dose levels near the planning target volume (PTV) prescription, reducing rates of failure in the epidural space without increasing the risk of spinal cord toxicity. To address this question, we conducted the treatment planning study described below, in which we modified SSRS treatment plans to enhance epidural coverage without exceeding spinal cord dose constraints. We then assessed the potential impact on local recurrence using a simplistic radiobiological model.

5.2 Materials and methods

Spinal stereotactic radiosurgery (SSRS) clinical plans from our institution were identified retrospectively for this analysis. All patients had undergone computed

tomography (CT) and magnetic resonance (MR) imaging, as well as clinical target volume (CTV) and spinal cord contouring on the registered images. The CTV encompassed the entire anterior vertebral body, the posterior elements or the whole vertebral body, according to consensus contouring guidelines (Cox et al., 2012). The spine planning target volume (PTV_{spine}) was generated from the CTV, avoiding the spinal cord expanded by 2mm to account for our institution's previously reported on-board imaging accuracy and associated immobilization accuracy for SSRS (Nelson et al., 2009). All spinal cord contours ended 6mm above/below the PTV_{spine} in the craniocaudal direction to ensure that all irradiated regions in the spinal cord were taken into account without excessively skewing relative volume-dose metrics. Patients were excluded if their prior clinical treatment did not target a cervical, thoracic or lumbar vertebral body. For eligible patients, the epidural space PTV ($PTV_{epidural}$) was contoured to include the epidural space adjacent to the PTV_{spine} , while excluding the PTV_{spine} and the 2mm spinal cord expansion. Patients were excluded if the $PTV_{epidural}$ was completely contained within the PTV_{spine} . Additionally, the $PTV_{epidural}$ contours were only drawn for subsets of the epidural space that share a border with the PTV_{spine} and appear sufficiently close to the PTV_{spine} . An example of a $PTV_{epidural}$ contour is shown in Figure 32. The average $PTV_{epidural}$ contour volume was 2.85 ± 1.85 cc.

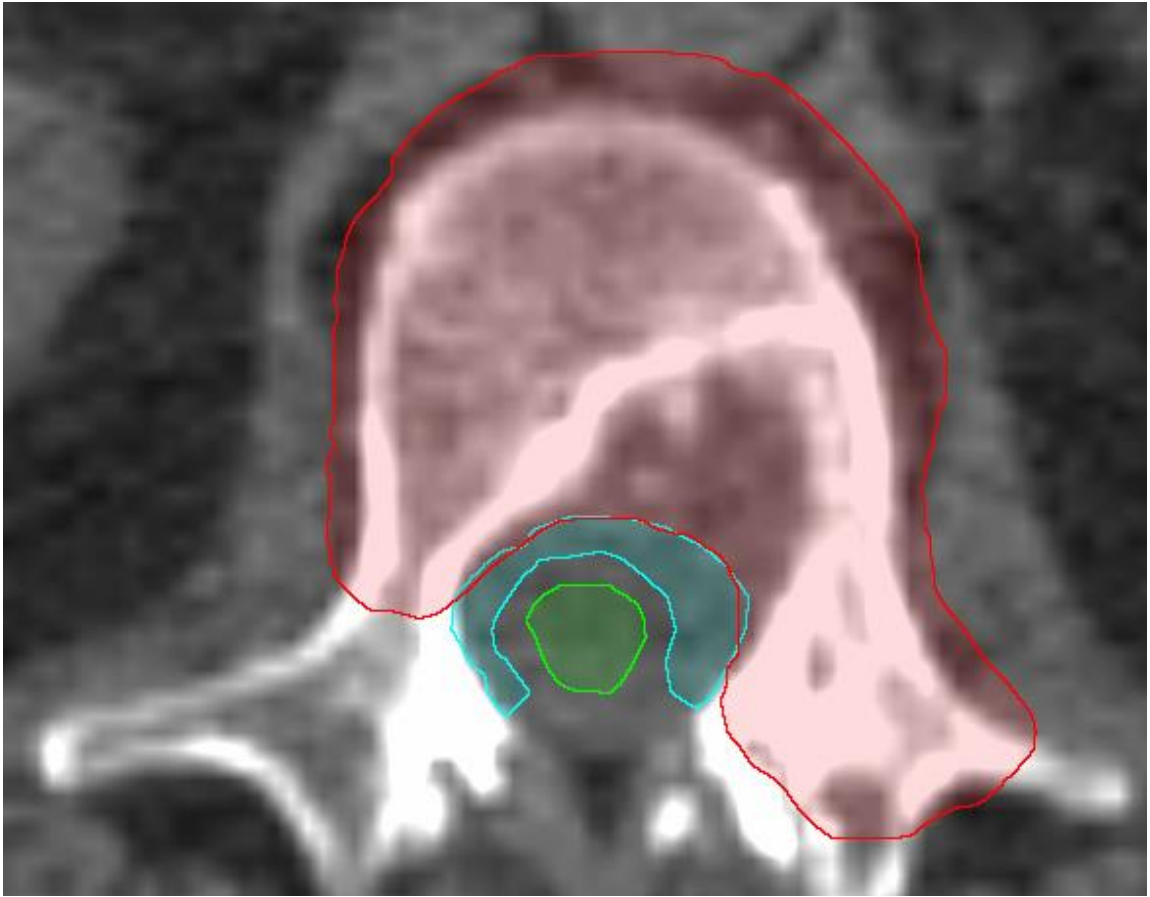


Figure 32: An example transverse planar contour of the PTV_{spine} (red), PTV_{epidural} (blue), and spinal cord (green) contours.

Prior clinical plans were renormalized for standardization such that their PTV_{spine} prescription was 1800 cGy in one fraction. Clinical plan constraints included PTV_{spine} constraints ($D_{95\%} = 1800$ cGy, $D_{5\%} < 1950$ cGy) and spinal cord constraints ($D_{\text{max}} < 1300$ cGy, $D_{10\%} < 1000$ cGy). These spinal cord constraints are commonly used in SSRS planning and have previously been categorized as low-risk limits, with an estimated risk of treatment-related spinal cord myelopathy of less than 1% (Grimm et al., 2016,

Kirkpatrick et al., 2010, Sahgal et al., 2010). Prior clinical plan doses were mapped onto the new epidural space contour, and dose-volume histograms (DVHs) were exported for analysis. For comparison, revised plans were created using three coplanar volumetric-modulated arc therapy (VMAT) beams. Beam energy was set to 10 MV, and collimator angles were set to 10°, 80°, and 100°; these configurations are typical beam settings for our clinical practice. Each revised plan always met the clinical plan spinal cord constraints ($D_{\max} < 1300$ cGy, $D_{10\%} < 1000$ cGy) and PTV_{spine} constraints ($D_{95\%} = 1800$ cGy, $D_{5\%} < 1950$ cGy). Once these constraints were met, the PTV_{epidural} $D_{95\%}$ was maximized. DVHs for the revised plans were exported for analysis.

Analysis of both the prior clinical plans and revised plans allowed for estimation of the distribution of feasible doses to the PTV_{epidural}, as well as the distribution of doses received incidentally when the PTV_{epidural} was not targeted. For each patient, the prior clinical and new plans were compared by their DVHs and estimated control probabilities for the PTV_{epidural}. $D_{95\%}$ values were compared individually, and all DVHs were accumulated in an average DVH for visual comparison. Control probabilities were estimated using the linear-quadratic (LQ) dose response model of cell survival and the Poisson-based model of tumor control probability (TCP) applied to the DVH of the PTV_{epidural} (Brenner, 2008, Hall and Giaccia, 2019, Oinam et al., 2011). The LQ model assumes that cell survival fractions are given as

$$f(d) = \exp(-\alpha d - \beta d^2)$$

where f is the surviving fraction of cells irradiated to a uniform dose d , with α and β being fitted parameters. Assuming that many voxels of volumes v_i with uniform clonogen density ρ are irradiated to doses d_i , the expected total number of cells surviving is

$$N_f = \sum_i \rho v_i f(d_i)$$

where i iterates over the voxels. Finally, the model assumes that the number of cells surviving is Poisson-distributed with a mean of N_f , so the probability of zero cells surviving is

$$TCP = \exp(-N_f).$$

No data regarding LQ model parameters exists pertaining specifically to the epidural space, so parameters were estimated as follows. First, the α/β ratio was taken to be 3 Gy to represent a range of published α/β ratio estimates for SSRS targets, similarly to Nelson et al. (Nelson et al., 2009). Next, ρ was assumed to be within a wide range, with the upper bound and lower bound set such that the range likely included the

clonogen densities typically occurring in the epidural space. Since the epidural space was not previously targeted in the clinical plans, any disease in this region is subclinical, with a tumor clonogen density significantly lower than the typical cell density in gross tumors visible on MR images. Therefore, our choice for the upper bound for this range was the typical density of cells in tumors, which is approximately equal to the density of cells in human tissue. Sender et al. estimated the number of human cells in the body as 3.0×10^{13} and the average human volume as approximately 70 L, so our upper bound was chosen to be 10^9 cells/cc (Sender et al., 2016). The lower bound for the range was set such that each voxel contained at least one cell. With an approximate voxel size of 0.001 cc, the lower bound was 10^3 cells/cc. 50 values of ρ were logarithmically sampled from our range of possible ρ values, with each value analyzed separately. Finally, α was extrapolated for each ρ such that the cohort's clinical PTV_{epidural} TCP would equal 100% minus published rates of recurrence in the epidural space. A collection of published epidural recurrence rates is shown in Table 5. For analysis, a representative epidural recurrence rate of 15% was chosen to represent the range of rates shown in Table 5. Fitting our TCP model to our clinical data yields α for each ρ , which was then used for estimating the TCPs for the revised plans.

Table 5: Summary of previous estimations of epidural space recurrence rates in spine SRS.

Reference	All Patients	All Failures	Epidural Failures	Percentage Epidural Failures
(Nelson et al., 2009)	33	4	2	6.06%
(Garg et al., 2011)	63	15	6	9.52%
(Chang et al., 2007)	74	17	8	10.81%
(Al-Omair et al., 2013)	80	21	15	18.75%
(Thibault et al., 2015)	56	13	11	19.64%

5.3 Results

A total of seventeen treatment plans satisfying the above criteria were identified. Sample dose distributions for a random patient are shown in Figure 33, including both the clinical and revised plan doses. The 1800 cGy isodose line stops exactly at the PTV_{spine} border in the clinical dose distribution, while it extends over most of the $PTV_{epidural}$ in the revised dose distribution. Additionally, the dose distribution for the epidural space is much more heterogeneous in the clinical dose distribution, with isodose lines passing through the epidural space ranging from 1000 to 1800 cGy. This visually demonstrates the significance of targeting the epidural space in SSRS. When not

explicitly targeted, the epidural space receives doses which are coincidental, spatially inconsistent, and potentially much lower than the prescription dose. However, targeting the epidural space yields doses which are consistently closer to the prescription dose.

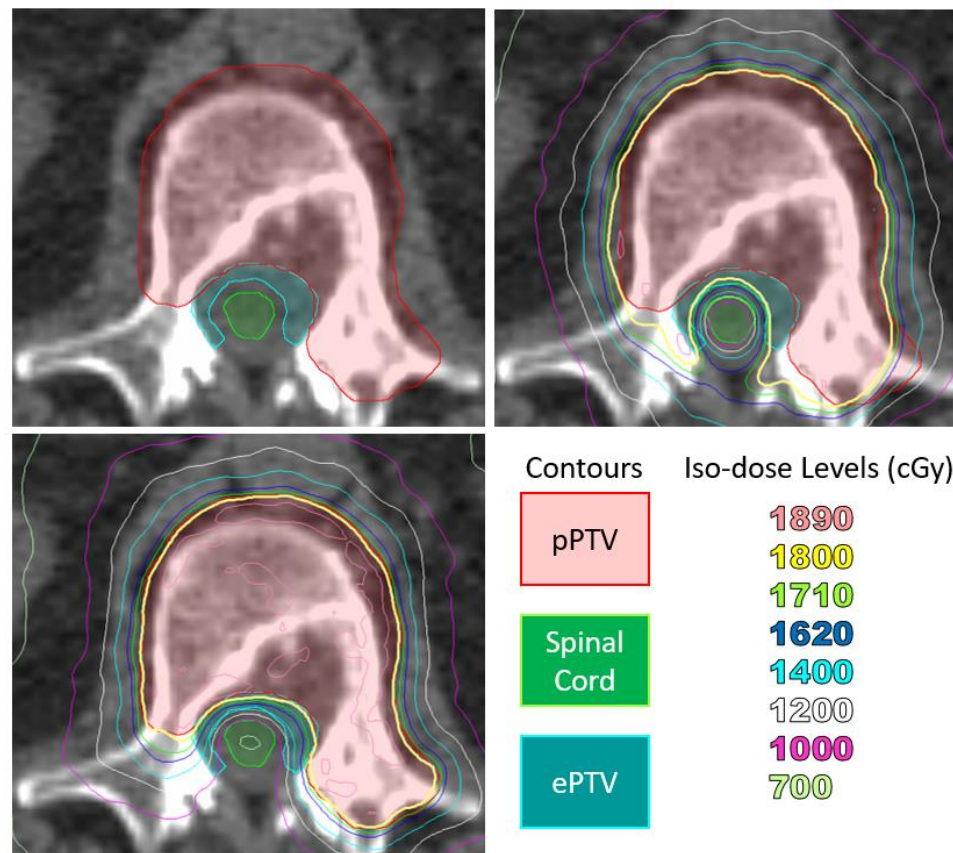


Figure 33: Example dose distributions comparing a clinical plan (lower-left) to its corresponding revised plan (upper-right) and its anatomical contours (upper-left).

The distributions of PTV_{epidural} DVHs are shown in Figure 34 for prior clinical and revised plans. The distribution of revised plan DVHs shows increased doses at all

volume fractions compared to the clinical DVHs, and this increase is greater at larger volume fractions. This demonstrates that the strongest improvement occurs at the low-dose regions of the epidural space, a result which generalizes our observations from Figure 33 to the entire cohort.

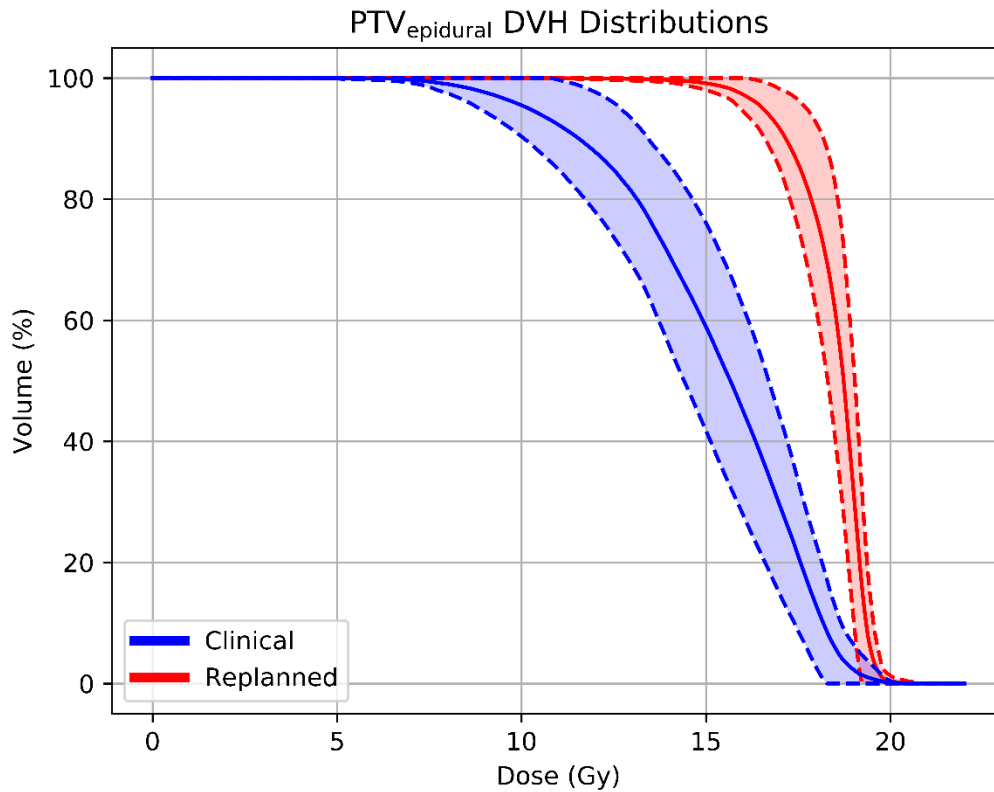


Figure 34: PTV_{epidural} DVH distributions. The blue and red regions indicate the prior clinical and revised distributions, respectively. The solid lines in the centers of the regions mark the average volume at a given dose value, while the dashed lines mark the volume average \pm standard deviation for a given dose value.

The PTV_{epidural} D_{95%} distributions are shown in Figure 35. The average D_{95%} for the prior clinical plans and revised plans were 10.96 Gy ± 1.76 Gy and 16.84 Gy ± 0.87 Gy (p < 10⁻⁵) respectively. In addition to achieving higher D_{95%} doses than the prior clinical plans, the revised plans achieved a tighter D_{95%} distribution – approximately half as wide. The doses delivered by the revised plans are higher and more consistent than the doses delivered by the clinical plans.

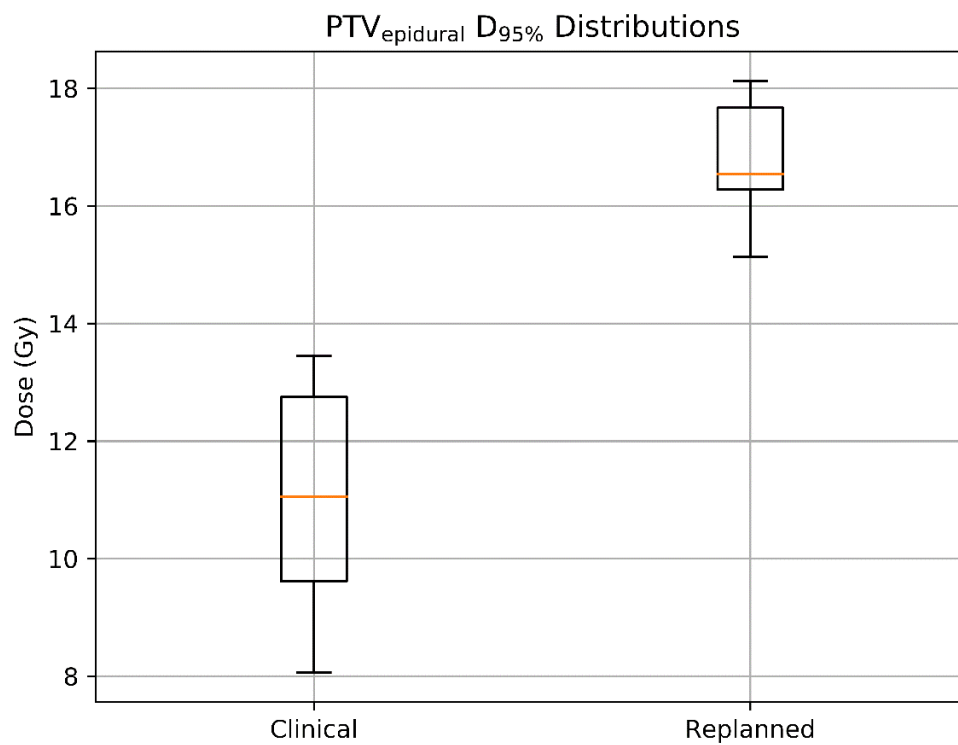


Figure 35: PTV_{epidural} D_{95%} distributions. The boxplots indicate 0, 25, 50, 75, and 100 percentile within the distribution.

Fitted values of α for each hypothetical value of ρ are shown in Figure 36.

Aggregating results over all hypothetical values of ρ , the average modeled PTV_{epidural}

TCP for the revised plans was $99.99985\% \pm 0.00035\%$, while the minimum was

99.99276%. This is a significant improvement over the average modeled epidural space

TCP for the clinical plans, which was set at 85%.

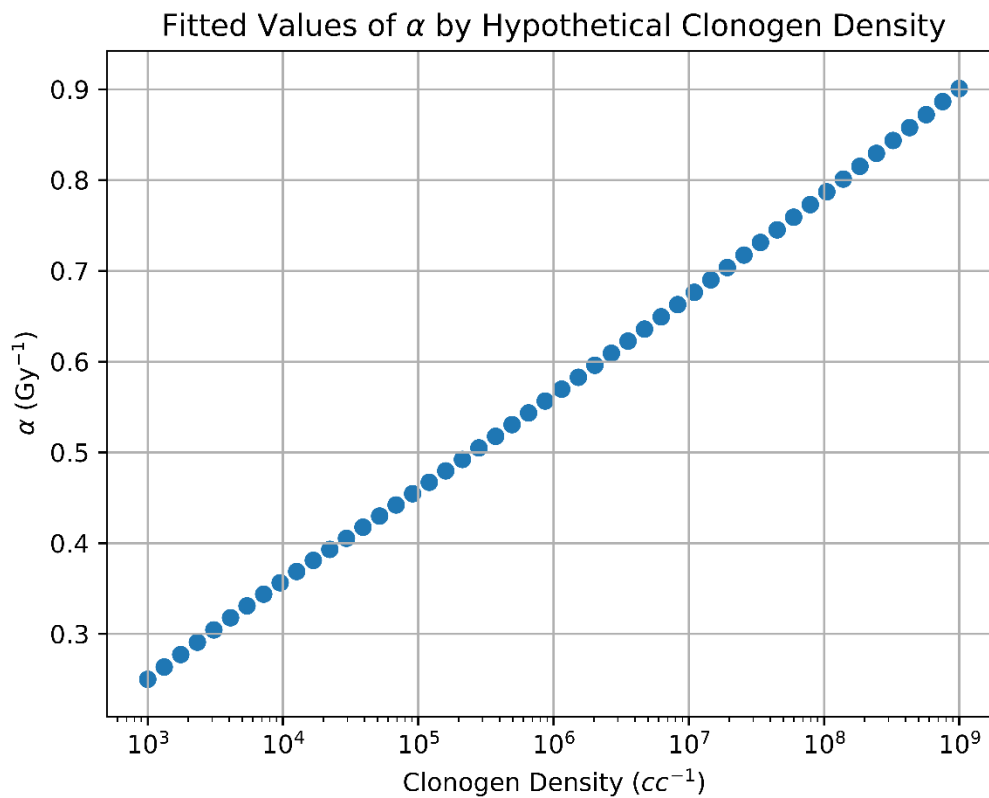


Figure 36: Fitted values of α for each hypothetical value of clonogen density.

5.4 Discussion

All SSRS plans in this analysis exhibited the potential to significantly increase the dose to the epidural space while obeying spinal cord constraints. Using our simplistic radiobiological model, our results suggest that the increased doses may be clinically significant, yielding a substantially increased probability of epidural space tumor control. Specifically targeting the epidural space involved modest modifications of the original clinical plans. In all cases, established low-risk spinal cord dose limits were respected while taking our spatial accuracy capabilities into account. Therefore, we expect that the additional effort required to purposefully irradiate the epidural space in SSRS is low in terms of planning and delivery time and complexity, while the clinical benefits could be significant. Moreover, spinal cord dose constraints are not violated by this technique, so targeting the epidural space should not significantly increase the risk of radiation induced spinal cord toxicity. However, this approach results in a more intimate association of the higher isodose lines with the spinal cord throughout the entire involved level, increasing the sensitivity of spinal cord dose to spatial deviations in targeting and positioning. To account for this, it is essential to maintain a high degree of vigilance regarding consistent and accurate patient positioning, immobilization and imaging, in keeping with standard practice in SSRS.

Thoughtful and cautious relaxation of spinal cord constraints will also increase the dose to the epidural space, improving tumor control in the spine and at least

partially accomplishing one of the goals of epidural space targeting. However, relaxing the dose constraint alone does not necessarily result in the maximum safe irradiation of the entire epidural space and would likely increase the risk of spinal cord toxicity. The technique for spinal metastasis planning and treatment described in this study presents an opportunity in selected patients to improve treatment efficacy without compromising patient safety.

Our simplistic radiobiological modeling involved many assumptions which qualify the accuracy of our extrapolated PTV_{epidural} TCPs. Our TCP calculations assumed a uniform clonogen density within the epidural space, which is likely not true. However, we believe that the effect of this assumption is reduced by the uniformity of doses delivered to the epidural space on the revised plans. We also used a representative literature value of 85% to model clinical epidural TCP, but this might not be applicable to our cohort. Most significantly, our PTV_{epidural} contours did not reflect the entire epidural space. We excluded a 2mm ring around the spinal cord from our contours to account for positioning variation, and we excluded regions of the epidural space that subjectively seemed sufficiently far from the PTV_{spine} . Our assumptions ignore the presence of clonogens within these regions. Despite this, our TCP calculations assumed that local epidural failure is entirely due to clonogens within the PTV_{epidural} contours at the time of treatment. Due to the potential presence of clonogens outside the PTV_{epidural} contours and heterogeneity of clonogen sensitivity to radiation, we expect actual

epidural TCPs to be less than our modeled average result of 99.99985% when the epidural space is targeted in SSRS. Instead, our results should be interpreted as suggesting that epidural TCPs can be improved through deliberate targeting of the epidural space, rather than precisely predicting the achievable epidural TCPs with epidural space targeting. A clinical trial on epidural space targeting in SSRS would more precisely determine the actual average epidural TCP. From a clinical perspective, epidural space targeting could result in a decreased risk of epidural failure while retaining the same risk of radiation-induced spinal cord myelopathy, leading to improved patient outcomes.

5.5 Conclusion

Targeting the epidural space in SSRS to account for subclinical epidural disease is feasible without significantly increasing the risk of radiation-induced myelopathy. In turn, this approach could reduce the risk of local recurrence in the epidural space, potentially enhancing survival and quality of life. It would be worthwhile to test this approach in a clinical trial of epidural targeting in SSRS, utilizing the precise attention to treatment planning, patient positioning, target visualization and sophisticated dose delivery established over 20 years of image-guided stereotactic body radiotherapy.

6. Model application to prostate VMAT

6.1 Introduction

The dose prediction model developed in Chapter 3 will first be analyzed on a patient cohort consisting of prostate cancer patients. Volumetric modulated arc therapy (VMAT) is a common treatment option for prostate cancer, with a well-established workflow for contour delineation, dosimetric guidelines, and treatment protocol. VMAT deliveries involve gantry rotation during irradiation, increasing the possible ranges of deliverable doses while also decreasing the delivery time (Otto, 2008, Teoh et al., 2011). Dual-arc VMAT has been shown to be an effective treatment technique specifically in prostate cancer, so we will be analyzing patients with dual-arc prostate VMAT (Guckenberger et al., 2009, Zhang et al., 2010).

Using current commercial treatment planning systems, prostate VMAT treatment planning can take 10-30 minutes per plan to optimize and calculate. Due to this time cost, commercial MCO techniques can be require large amounts of time to create enough plans to achieve sufficient Pareto surface interpolation accuracy. Therefore, prostate VMAT could particularly benefit from a fast, accurate dose prediction model. Additionally, this dose prediction application be an example of the model's ability to assist a treatment planner for typical treatments in simple treatment paradigms in the clinical workflow.

6.2 Materials and methods

90 prostate cancer patients were retrospectively included in this study. Each patient's dataset consisted of an abdominal computed tomography (CT) scan and contours of their planning target volume (PTV), bladder, rectum, left femoral head and right femoral head. After anonymization, patient datasets were imported to a commercial treatment planning system for fluence optimization and dose calculation. The PTV dose prescription was set to 70 Gy in 29 fractions, as is the current standard for clinical practice at our institution. During treatment planning, each plan included two concentric, coplanar volumetric modulated arc therapy (VMAT) beams centered on the PTV, with field sizes set to encompass the PTV during a 358-degree beam rotation. Beam collimators were set at 15 and 345 degrees to reduce the effect of collimator leaf gap overlap. During optimization, priorities were placed on the PTV homogeneity index ($HI = D_{2\%} - D_{98\%}$), bladder $D_{25\%}$, and rectum $D_{25\%}$. These objectives were chosen to represent the dimensions of trade-off during treatment planning, since priority in prostate VMAT is primarily divided between uniform PTV coverage and bladder and rectum sparing. Therefore, these objectives had different optimization priority combinations for each plan to sample the Pareto surface of dose trade-offs. After optimization, plans were normalized such that PTV $D_{95\%}$ equaled 100% of the target's dose prescription. Fixed constraints for each plan optimization included PTV $D_{93\%} < 101\%$ to reduce the dose-shifting effect of plan normalization, as well as $D_{0.01cc} < 65\%$ for both femoral heads in

accordance of our institution's standard practice for normal critical structure constraints. Variable constraints included PTV HI < 10%, bladder $D_{25\%}$ < 30% of prescription, and rectum $D_{25\%}$.

For each patient, the Pareto surface was sampled by optimizing and calculating 25 plans as follows. The Pareto space dimensions were PTV HI, bladder $D_{25\%}$, and rectum $D_{25\%}$. Each plan had a different optimization priority combination and therefore sampled a different location on the Pareto surface. Bounding points on the surface were chosen through manual plan optimization such that the bounding points represented clinically feasible plans. Subsequent points on the surface were created using linear combinations of the objective priorities of the bounding points; this ensured that all interior points also represented clinically feasible plans on the Pareto surface. Beamlet fluence optimization and dose calculation was performed with a commercial treatment planning system. After each plan was calculated, the corresponding dose map, critical structure maps, and optimization priority combination were exported for use during model training and evaluation.

The model's performance was assessed by its dose map root-mean-square error, visual comparison of the dose map prediction and TPS- calculated dose map for a random patient, visual comparison of the corresponding DVH prediction and TPS- calculated DVH, and model evaluation speed. Additionally, the Pareto surfaces inferred from the dose predictions was compared using the four Pareto surface

similarity metrics developed in Chapter 4.2.1, namely the Pareto space RMSE with 50 samples per barycentric dimension, the Hausdorff distance (HD), the average projected distance (APD), and the average nearest-point distance (ANPD) with 100 samples per barycentric dimension. The number of samples per barycentric dimension were chosen to coincide with the number of samples shown to achieve convergence in Chapter 4.3.1.

6.3 Results

Figure 37 shows the dose map RMSE for the training and testing sets during model training. After training, the mean dose map RMSE was $2.31\% \pm 0.41\%$ and $2.60\% \pm 1.23\%$ for the training and testing set dose predictions, respectively. These errors demonstrate that the model is capable of achieving good prediction accuracy on a voxel-by-voxel basis. The training set error was somewhat lower than the testing set error, indicating that the model overfit to the training data by about 12%. However, 12% overfitting to training data is reasonable, since the testing set predictions were still only a few percentages of dose prescription away from the TPS-calculated dose maps. The dose map RMSE due to the initialization fit alone is $4.98\% \pm 0.68\%$ and $5.47\% \pm 1.55\%$ for the training and testing sets, respectively. This indicates that the residual network makes a measurable improvement to the dose initialization. Note that these values differ from the model's general dose map RMSE at 0 iterations into training because the residual network's parameters are nonzero and randomly generated, decreasing accuracy since the subsequent updates are random. For comparison, the International Commission on

Radiation Units and Measurements (ICRU) and Task Group 142 of the American Association of Physicists in Medicine (AAPM) have stated that a 5% maximum dosimetric uncertainty is appropriate for standard IMRT treatments such as ours (ICRU, 1976, Klein et al., 2009). Therefore, these dose map RMSEs are comparable to the maximum error permitted in treatment delivery.

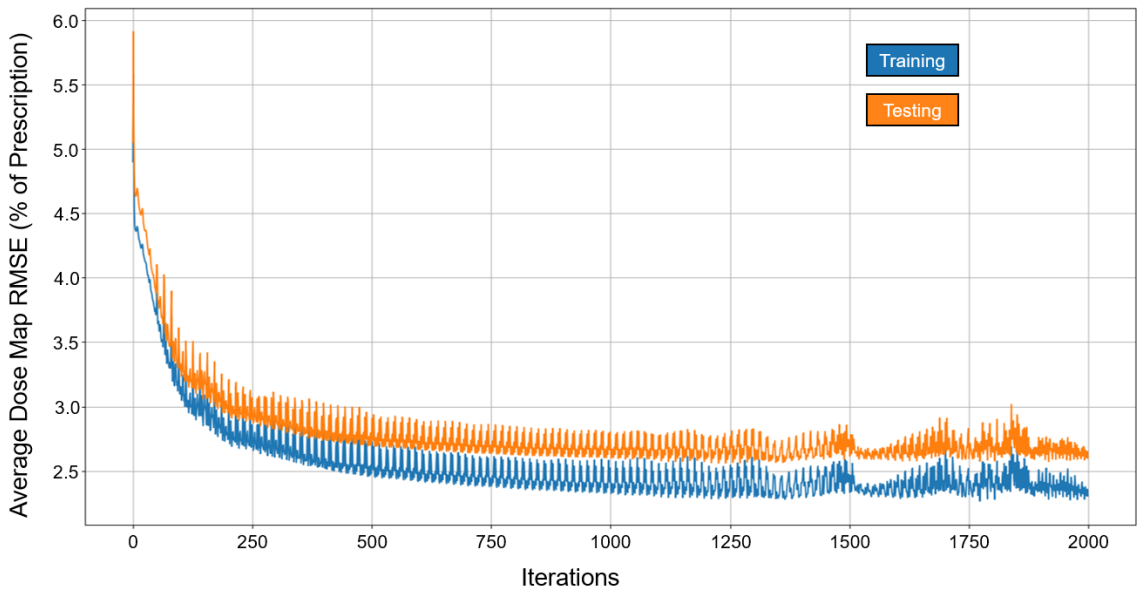


Figure 37: Graph of dose map root-mean-square error for the training set (blue) and testing set (orange) as a function of number of iterations during model training.

Figures 38 and 39 show side-by-side comparisons between the effect of prioritizing PTV HI or prioritizing rectum D25% in a dose map prediction and its corresponding TPS calculation. Visually, we can see that the dose map predictions are jagged compared to their respective TPS-calculated dose maps, specifically around the

30% isodose line. We expect this to be the case, since the neural network's architecture does not explicitly promote local smoothness in the dose distribution predictions. Additionally, we see that the region of largest isodose displacement is the low dose region superior to the PTV. Note that this region is not near the PTV or surrounding critical structures.

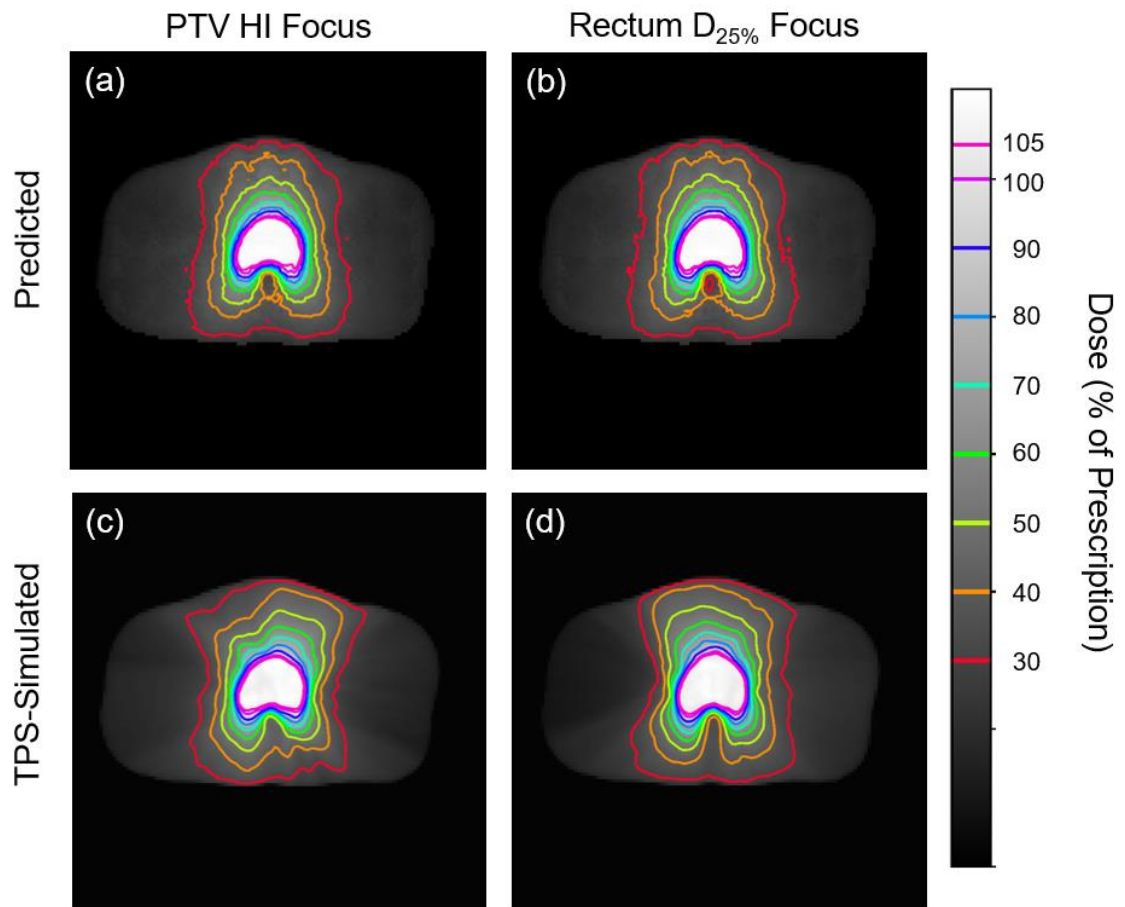


Figure 38: Comparison between dose distribution prediction and TPS calculation in plans prioritizing PTV HI or prioritizing rectum $D_{25\%}$. Transverse slices are taken from the center of the PTV, and the patient was randomly sampled from the testing dataset.

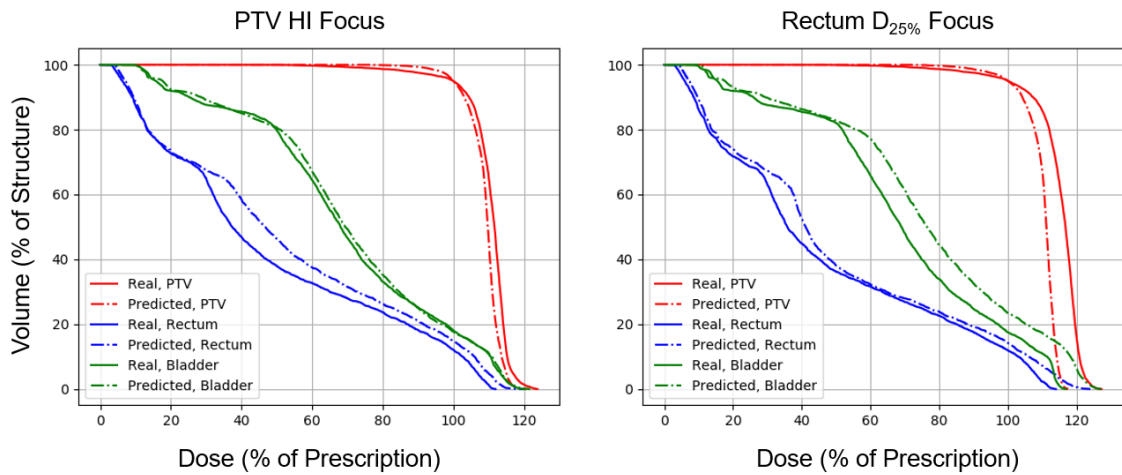


Figure 39: Comparison between indirect DVH prediction and TPS calculation in plans prioritizing PTV HI or prioritizing rectum D_{25%}.

Table 6 compares the four Pareto surface similarity metrics described in Chapter 4.2.1, namely the Pareto space RMSE with 50 samples per barycentric dimension, the Hausdorff distance (HD), the average projected distance (APD), and the average nearest-point distance (ANPD) with 100 samples per barycentric dimension. The number of samples per barycentric dimension were chosen to coincide with the number of samples shown to achieve convergence in Chapter 4.3.1. Table 6 indicates that there is a modest difference between all four metrics, suggesting that the choice of metric is a meaningful decision when comparing Pareto surfaces.

Table 6: Pareto surface similarity metrics evaluated on the Pareto surfaces indirectly generated by our dose prediction model for prostate VMAT.

Pareto surface metric	Training dataset	Testing dataset
RMSE (50 samples)	10.98% \pm 3.56%	10.84% \pm 2.65%
HD	17.98% \pm 5.99%	18.31% \pm 7.07%
APD	9.93% \pm 3.80%	10.76% \pm 4.18%
ANPD (100 samples)	8.80% \pm 3.37%	9.72% \pm 3.85%

Additionally, the amount of time required to evaluate our model on all Pareto surface points for all patients in the training set was about 57 seconds. Therefore, predicting a single dose distribution required about 0.05 seconds per plan. This is much faster than current optimization and calculation techniques, which take approximately 10-30 minutes per plan.

6.4 Discussion

Chapter 3 presented a novel machine learning dose prediction model which takes optimization objective priorities into account, allowing for indirect Pareto surface estimation. Our results indicate that the model is able to predict doses with good accuracy, as the root-mean-square predicted dose map errors are a few percentages of their corresponding TPS- calculated doses. These dose map RMSEs are less than the maximum error tolerance proposed by the ICRU and AAPM TG 142, suggesting that our predictions may be appropriate for clinical dose distribution estimation. Moreover, the

model produces just a dose distribution without actually creating a plan, so the model requires a final real plan optimization and dose calculation which will correct these dose map prediction errors prior to treatment delivery. Pareto surface metrics indicate that these dose map predictions make reasonable translations in Pareto space. Our results also indicate that the model's overfitting to training data dose map RMSE is modest, and that overfitting to dose map RMSE does not appear to result in overfitting in any of the Pareto space metrics.

The prediction speed of our model is particularly encouraging. By predicting each plan in approximately 0.05 seconds, our model may be used for real-time treatment planning without needing to interpolate between previously sampled points, allowing the treatment planner to very quickly estimate the doses produced by a given optimization priority combination. This indirectly gives the planner more time to plan per patient, which may improve final plan quality. Moreover, our model only requires patient anatomy and optimization priorities, so it is capable of generating many samples from the Pareto surface automatically. This is potentially useful for large-scale automatic theoretical dosimetric investigations of new treatment planning paradigms, such as testing the effects of pushing a dose limit past its historical value or determining the feasibility of treating new structures. More research is needed to investigate these possibilities.

We believe that our model's speed, accuracy, and mitigated overfitting are due to the model's design. The combination of contiguous and atrous patches during contour processing increases the effective receptive field size of each layer in the residual network. Achieving a similar effective field-of-view in a more traditional convolutional neural network would involve either increasing size of each convolution kernel or adding many more layers to the network. However, both of these options involve more model parameters, have increased computational requirements, and are more prone to overfitting. Our model's patch extraction process innovates by incorporating local and global information within each layer without increasing computational requirements or promoting overfitting.

Despite its potential advantages, our model has some limitations which hinder its accuracy and utility. The model's dose initialization assumes an isotropic inverse exponential decay of dose as a function of inter-slice and intra-slice distances from the PTV. Although this assumption is only appropriate for VMAT plans which involve beam arcs wrapping nearly 360 degrees around the patient, it is likely that other dose initializations exist which are appropriate for IMRT or VMAT with significantly fewer than 360 degrees per arc. Additionally, the model required several hyperparameters (i.e. 6 residual blocks in the neural network, 100 output units for the first two layers in each block, atrous rates of 3 and 10 in patch sampling, etc.), and it is not immediately clear how to determine the optimal values for these hyperparameters aside from trial and

error. However, we expect that slight adjustments from our chosen values for the hyperparameters should not significantly change model performance. Finally, since the model's output is a dose distribution without an actual plan optimization or dose calculation, the model can only be used to determine the subjectively optimal optimization priorities, which then need to be used in a real plan optimization and dose calculation to actually create a deliverable plan.

We have implemented several metrics for evaluating the error between a predicted Pareto surface and its corresponding TPS- calculated Pareto surface. Our metrics reported similar values around 10-15% of dose prescription for both training and testing sets. Again, it is worth noting that these metrics accumulate the errors from each dimension rather than averaging them, which is why these surface metrics are significantly larger than the dose map RMSE of 2-3%. Of these metrics, we hypothesize that the average nearest-point distance is the most appropriate of these metric due to its removal of error contributions orthogonal to the direction between the Pareto surfaces. The distinction appears to be meaningful as well, with the metrics all attaining somewhat different values on our training and testing datasets. Also, to our knowledge, no other body of research has applied Pareto space metrics to evaluate the Pareto surfaces of radiation therapy dose predictions. This prevents us from comparing our Pareto space results with previous dose prediction research. To account for this, we have included all of these metrics for ease of comparison with future research.

6.5 Conclusion

We have applied the dose prediction model developed in Chapter 3 to prostate VMAT dose prediction. The model's error is modest when applied to our prostate VMAT cases, with average dose map root-mean-square errors of $2.60\% \pm 1.23\%$ over all patients and all optimization priority combinations in the patient testing dataset. Therefore, our model may be used to accelerate the prostate VMAT treatment planning process and prostate VMAT MCO.

7. Model application to pancreas SBRT

7.1 Introduction

The second patient cohort used to test our dose prediction model consists of pancreas SBRT patients. One of the primary challenges of radiation therapy for pancreas cancers specifically is the very close proximity of the pancreas to the duodenum and small bowel generally (Trakul et al., 2014). Due to this proximity, it is difficult to deliver radiation to the pancreas with uniform coverage while also minimizing the radiation delivered to the duodenum. For this reason, pancreas radiation therapy is particularly dependent on the size of the treatment margins, so SBRT is appropriate for this case. SBRT has also been experimentally shown to produce higher local control rates and lower normal tissue side effects (Trakul et al., 2014).

To optimize and calculate SBRT dose distributions with enough precision to accurately compute the doses which these tight margins receive, it is important to calculate the dose distribution at an increased resolution. Typically, the dose distribution resolution changes from 2.5 mm to 1.25 mm in the X- and Y- directions to yield sufficient accuracy. This quadruples the number of voxels to be optimized and calculated per plan during treatment planning, which results in plan creation time requirements approximately four times longer than that of conventional EBRT.

As a result, the number of plans that the treatment planner can create imposes an upper limit on the precision with which such an analysis can estimate feasible dose

targets and predict local control rates. To bypass this limitation, a dose prediction model may be used to significantly increase the rate at which plans can be produced, increasing the number of plans available for analysis and improving the precision of subsequent feasible dose targets and local control predictions. It is therefore of particular interest to be able to predict SBRT dose distributions.

In this Chapter, we will analyze the feasibility of applying our dose prediction model to the prediction of dose distributions of pancreas SBRT treatment plans. This analysis will determine the ability of our model to generalize its performance to increased dose distribution resolutions, different anatomical treatment sites, and increased inter-patient heterogeneity compared to prostate VMAT.

7.2 Materials and methods

20 pancreas SBRT patients from our institution were identified retrospectively for this analysis. During all of these cases, the critical structures relevant to treatment planning were the PTV, small bowel, stomach, kidneys, liver, and spinal cord. However, for multi-criterial optimization, the Pareto space dimensions were the PTV homogeneity index ($HI = D_{2\%} - D_{98\%}$) and the small bowel/stomach $D_{0.1cc}$. These dimensions were chosen to reflect the primary trade-off made between PTV coverage and small bowel sparing during pancreas SBRT. During plan generation, fixed constraints were placed on the kidneys, liver, and spinal cord to ensure doses to these OARs were below a set of dose constraints typically used at our institution. Additionally, variable constraints were

placed on the PTV HI and the small bowel/stomach $D_{0.1cc}$ to sample the possible trade-offs and populate the Pareto surface.

The model's application to pancreas SBRT uses the same architecture as the architecture described in Chapter 3.1 and used in Chapter 6. However, compared to conventional prostate VMAT, pancreas SBRT operates at a higher resolution, quadrupling the number of voxels per slice. As a consequence, performing dose predictions on pancreas SBRT require four times as much computer memory. To compensate for this requirement while preserving the model's parameters, each batch contains one-quarter as many slices as would be contained in prostate VMAT. Additionally, at each optimization iteration, the loss function gradients are calculated for four batches and aggregated before updating the model's parameters. Although this quadruples the amount of time required for model training and evaluation, the actual performance of the model is identical to that for prostate VMAT.

For each patient, the Pareto surface was sampled by optimizing and calculating 10 plans as follows. The Pareto space dimensions were PTV HI and the small bowel/stomach $D_{0.1cc}$. Each plan had a different optimization priority combination and therefore sampled a different location on the Pareto surface. Bounding points on the surface were chosen through manual plan optimization such that the bounding points represented clinically feasible plans. Subsequent points on the surface were created using linear combinations of the objective priorities of the bounding points; this ensured

that all interior points also represented clinically feasible plans on the Pareto surface. Beamlet fluence optimization and dose calculation was performed with a commercial treatment planning system. After each plan was calculated, the corresponding dose map, critical structure maps, and optimization priority combination were exported for use during model training and evaluation. For analysis, 10 patients were included in the training dataset and 10 were included in the testing dataset.

The model's performance was evaluated by dose map root-mean-square error, visual comparison of the dose map prediction and TPS- calculated dose map for a random patient, visual comparison of the corresponding DVH prediction and TPS- calculated DVH, and model evaluation speed. Additionally, the Pareto surfaces inferred from the dose predictions will be compared using the four Pareto surface similarity metrics developed in Chapter 4.2.1, namely the Pareto space RMSE with 50 samples per barycentric dimension, the Hausdorff distance (HD), the average projected distance (APD), and the average nearest-point distance (ANPD) with 100 samples per barycentric dimension. The number of samples per barycentric dimension were chosen to coincide with the number of samples shown to achieve convergence in Chapter 4.3.1.

7.3 Results

After training, the mean dose map RMSE was $3.77\% \pm 0.79\%$ and $5.34\% \pm 1.57\%$ for the training dataset and testing dataset, respectively. Although these errors are larger the dose map RMSE during prostate VMAT prediction, they are still relatively modest.

For comparison, the International Commission on Radiation Units and Measurements (ICRU) and Task Group 142 of the American Association of Physicists in Medicine (AAPM) have stated that a 5% maximum dosimetric uncertainty is appropriate for standard IMRT treatments such as ours (ICRU, 1976, Klein et al., 2009). However, overfitting was more dominant, with the testing errors being 40% larger than the training set errors. Despite this increase in overfitting, the errors are still comparable to the typical dosimetric uncertainty limit of 5%.

Figure 40 shows side-by-side comparisons between the effect of prioritizing PTV coverage or prioritizing small bowel sparing in a dose map prediction and its corresponding TPS calculation. Importantly, the lower isodose lines in the predictions appear to be significantly more smooth and circular than the TPS calculations, which have beam streaks distorting their shapes.

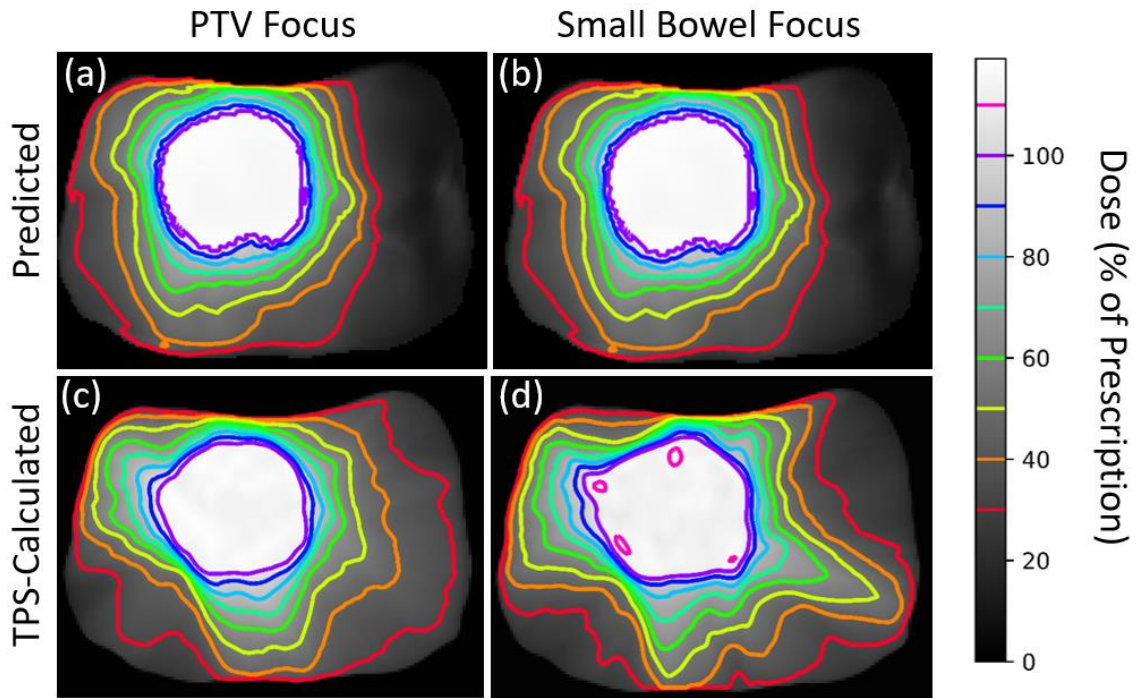


Figure 40: Comparison between dose distribution prediction and TPS calculation in plans prioritizing PTV HI or prioritizing bowel $D_{0.1cc}$. Transverse slices are taken from the center of the PTV, and the patient was randomly sampled from the testing dataset.

Figure 41 shows side-by-side comparisons between the effect of prioritizing PTV coverage or prioritizing small bowel sparing in a DVH prediction and its corresponding TPS calculation. Despite the differences between the dose map predictions and dose map calculations in Figure 40, their corresponding DVHs are much more similar. This is likely because the exterior distorted isodose shape in the TPS calculations do not directly affect the dose distributions within the critical structures. Additionally, Figure 41 indicates that the prediction model is able to predict DVHs with fundamentally different

shapes, as the epidural space DVH changes noticeably between the epidural space prioritization and spinal cord sparing prioritization.

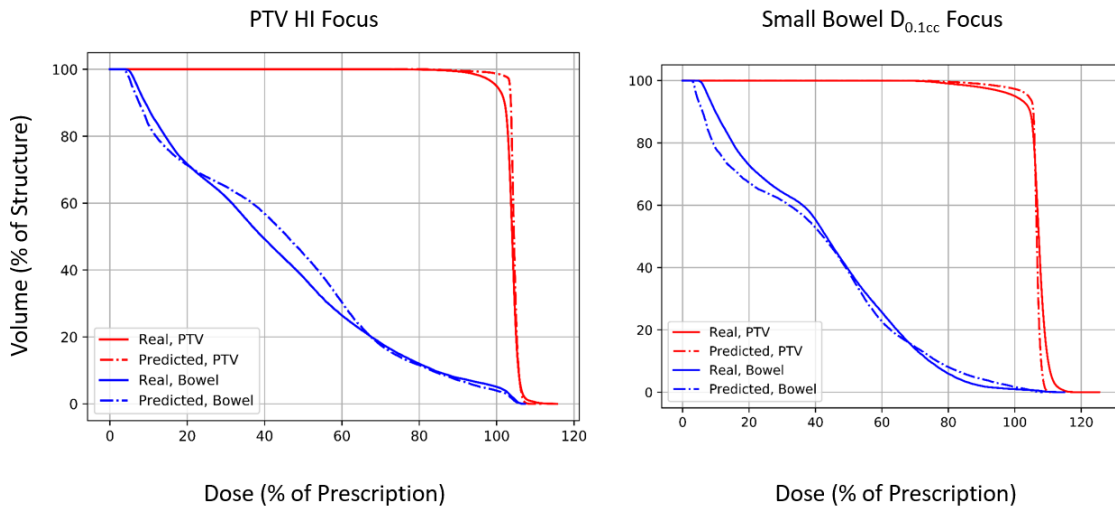


Figure 41: Comparison between dose distribution prediction and TPS calculation in plans prioritizing PTV HI or prioritizing small bowel $D_{0.1cc}$.

Table 8 compares the four Pareto surface similarity metrics described in Chapter 4.2.1, namely the Pareto space RMSE with 50 samples per barycentric dimension, the Hausdorff distance (HD), the average projected distance (APD), and the average nearest-point distance (ANPD) with 100 samples per barycentric dimension. The number of samples per barycentric dimension were chosen to coincide with the number of samples shown to achieve convergence in Chapter 4.3.1. Table 7 indicates that there is a modest difference between all four metrics.

Table 7: Pareto surface similarity metrics evaluated on the Pareto surfaces indirectly generated by our dose prediction model for spine SRS.

Pareto surface metric	Training dataset	Testing dataset
RMSE (50 samples)	4.25 % \pm 0.84%	6.39% \pm 2.37%
HD	5.88% \pm 1.32%	7.74% \pm 4.04%
APD	3.10% \pm 1.13%	5.59% \pm 2.74%
ANPD (100 samples)	2.54% \pm 0.76%	4.65% \pm 2.23%

Additionally, the amount of time required to evaluate our model on all Pareto surface points for all patients in the training set was about 23 seconds. Therefore, predicting a single dose distribution required about 0.21 seconds per plan. This is about 4 times slower than our prostate VMAT dose prediction speed, which agrees with the doubled underlying dose distribution resolution. However, this is much faster than current optimization and calculation techniques, which take approximately 10-30 minutes per plan.

7.4 Discussion

Chapter 3 presented a novel machine learning dose prediction model which takes optimization objective priorities into account, allowing for indirect Pareto surface estimation. Our results indicate that the model is able to predict doses with good accuracy, as the root-mean-square predicted dose map errors are a few percentages of their corresponding TPS- calculated doses. However, these errors are notably higher

than the errors observed in Chapter 6's analysis of prostate VMAT dose prediction. Despite this, the dose map RMSEs are still comparable the maximum error tolerance proposed by the ICRU and AAPM TG 142, indicating that our model may be useful for clinical application in pancreas SBRT. However, the model produces just a dose distribution without actually creating a plan, so the model requires a final real plan optimization and dose calculation which will correct these dose map prediction errors prior to treatment delivery.

We have implemented several metrics for evaluating the error between a predicted Pareto surface and its corresponding TPS- calculated Pareto surface. Our metrics reported similar values around 4-7% of dose prescription for both training and testing sets. Again, it is worth noting that these metrics accumulate the errors from each dimension rather than averaging them, which is why these surface metrics are slightly larger than the dose map RMSE of 4-5%. Of these metrics, we hypothesize that the average nearest-point distance is the most appropriate of these metric due to its removal of error contributions orthogonal to the direction between the Pareto surfaces. The distinction appears to be meaningful as well, with the metrics all attaining somewhat different values on our training and testing datasets. Also, to our knowledge, no other body of research has applied Pareto space metrics to evaluate the Pareto surfaces of radiation therapy dose predictions. This prevents us from comparing our Pareto space

results with previous dose prediction research. To account for this, we have included all of these metrics for ease of comparison with future research.

Interestingly, the Pareto surface similarity metrics evaluated on our pancreas SBRT predictions are actually lower than the metrics evaluated on prostate VMAT predictions, despite the prostate dose map RMSEs being noticeably lower. As seen in Figure 40, it is likely that the pancreas SBRT dose map RMSEs are more heavily influenced by the beam streaks which occur. The problem of predicting these beam streaks is somewhat similar to the problem of predicting optimal beam angles in IMRT, which our model's architecture is not well-designed to handle. However, this problem is less impactful on the critical structures than the low-dose regions far from the critical structures, so it makes sense that the DVHs are not heavily impacted by these streaks. Although the dose distribution predictions are not as good, our model may be useful for inferring Pareto surfaces. Additionally, our results indicate that there is a modest difference between the four Pareto surface similarity metrics, suggesting that the choice of metric is significant when evaluating Pareto surface similarity. Our results also indicate that the model's overfitting to training data dose map RMSE is modest, and that overfitting to dose map RMSE does not appear to result in overfitting in any of the Pareto space metrics.

The prediction speed of our model is particularly encouraging. Although the dose prediction times for pancreas SBRT are four times higher than conventional

prostate VMAT dose prediction times, the model still predicts each plan in approximately 0.21 seconds. With this speed, our model may be used for real-time treatment planning without needing to interpolate between previously sampled points, allowing the treatment planner to very quickly estimate the doses produced by a given optimization priority combination. This indirectly gives the planner more time to plan per patient, which may improve final plan quality. Moreover, our model only requires patient anatomy and optimization priorities, so it is capable of generating many samples from the Pareto surface automatically.

7.5 Conclusion

We have applied the dose prediction model developed in Chapter 3 to predicting the dose distributions of pancreas SBRT treatment plans. The model's error is modest, with average dose map root-mean-square errors of $5.34\% \pm 1.57\%$ over all patients and all optimization priority combinations in the patient testing dataset. Although these errors are higher than the errors observed in prostate VMAT dose prediction, the Pareto surface similarity metrics indicate that the model generalizes well to Pareto surface prediction. Therefore, our model may be used to accelerate MCO techniques for pancreas SBRT treatment planning assistance.

8. Model application to spine SRS with epidural space irradiation

8.1 Introduction

Chapter 5 presented an analysis of the potential to irradiate the epidural space in spine stereotactic radiosurgery (SRS), concluding that such irradiation is feasible in typical spine metastasis cancer patients and has the potential to improve local control rates while preserving the low probability of radiation-induced spinal cord myelopathy. However, like SBRT, SRS is a special radiation therapy procedure which involves relatively small margins for positioning. To optimize and calculate SRS dose distributions with enough precision to accurately compute the doses which these tight margins receive, it is important to calculate the dose distribution at an increased resolution. Typically, the dose distribution resolution changes from 2.5 mm to 1.25 mm in the X- and Y- directions to yield sufficient accuracy. This quadruples the number of voxels to be optimized and calculated per plan during treatment planning, which results in plan creation time requirements approximately four times longer than that of conventional EBRT.

As a result, the number of plans that the treatment planner can create imposes an upper limit on the precision with which such an analysis can estimate feasible dose targets and predict local control rates. To bypass this limitation, a dose prediction model may be used to significantly increase the rate at which plans can be produced, increasing the number of plans available for analysis and improving the precision of subsequent

feasible dose targets and local control predictions. It is therefore of particular interest to be able to predict SRS dose distributions.

In this Chapter, we will analyze the feasibility of applying our dose prediction model to the prediction of dose distributions of spine SRS with epidural space irradiation. This analysis will determine the ability of our model to generalize its performance to increased dose distribution resolutions, different anatomical treatment sites, and increased inter-patient heterogeneity compared to prostate VMAT.

8.2 Materials and methods

17 spinal stereotactic radiosurgery (SSRS) clinical plans from our institution were identified retrospectively for this analysis. All patients had undergone computed tomography (CT) and magnetic resonance (MR) imaging, as well as clinical target volume (CTV) and spinal cord contouring on the registered images. The CTV encompassed the entire anterior vertebral body, the posterior elements or the whole vertebral body, according to consensus contouring guidelines (Cox et al., 2012). The spine planning target volume (PTV_{spine}) was generated from the CTV, avoiding the spinal cord expanded by 2mm to account for our institution's previously reported on-board imaging accuracy and associated immobilization accuracy for SSRS (Nelson et al., 2009). All spinal cord contours ended 6mm above/below the PTV_{spine} in the craniocaudal direction to ensure that all irradiated regions in the spinal cord were taken into account without excessively skewing relative volume-dose metrics. Patients were excluded if

their prior clinical treatment did not target a cervical, thoracic or lumbar vertebral body. For eligible patients, the epidural space PTV (PTV_{epidural}) was contoured to include the epidural space adjacent to the PTV_{spine} , while excluding the PTV_{spine} and the 2mm spinal cord expansion. Patients were excluded if the PTV_{epidural} was completely contained within the PTV_{spine} . Additionally, the PTV_{epidural} contours were only drawn for subsets of the epidural space that share a border with the PTV_{spine} and appear sufficiently close to the PTV_{spine} . An example of a PTV_{epidural} contour is shown in Figure 42. The average PTV_{epidural} contour volume was 2.85 ± 1.85 cc.



Figure 42: An example transverse planar contour of the PTV_{spine} (red), PTV_{epidural} (blue), and spinal cord (green) contours.

The model's application to spine SRS uses the same architecture as the architecture described in Chapter 3.1 and used in Chapters 6 and 7. However, compared to conventional prostate VMAT, spine SRS operates at a higher resolution, quadrupling the number of voxels per slice. As a consequence, performing dose predictions on spine SRS require four times as much computer memory. To compensate for this requirement while preserving the model's parameters, each batch contains one-quarter as many slices as would be contained in prostate VMAT. Additionally, at each optimization iteration, the loss function gradients are calculated for four batches and aggregated before updating the model's parameters. Although this quadruples the amount of time required for model training and evaluation, the actual performance of the model is identical to that for prostate VMAT.

For each patient, the Pareto surface was sampled by optimizing and calculating 25 plans as follows. The Pareto space dimensions were $PTV_{\text{spine HI}}$, $PTV_{\text{epidural D}_{95\%}}$, and spinal cord D_{max} . Each plan had a different optimization priority combination and therefore sampled a different location on the Pareto surface. Bounding points on the surface were chosen through manual plan optimization such that the bounding points represented clinically feasible plans. Subsequent points on the surface were created using linear combinations of the objective priorities of the bounding points; this ensured that all interior points also represented clinically feasible plans on the Pareto surface. Beamlet fluence optimization and dose calculation was performed with a commercial

treatment planning system. After each plan was calculated, the corresponding dose map, critical structure maps, and optimization priority combination were exported for use during model training and evaluation. For analysis, 8 patients were included in the training dataset and 9 were included in the testing dataset.

The model's performance was evaluated by dose map root-mean-square error, visual comparison of the dose map prediction and TPS- calculated dose map for a random patient, visual comparison of the corresponding DVH prediction and TPS- calculated DVH, and model evaluation speed. Additionally, the Pareto surfaces inferred from the dose predictions will be compared using the four Pareto surface similarity metrics developed in Chapter 4.2.1, namely the Pareto space RMSE with 50 samples per barycentric dimension, the Hausdorff distance (HD), the average projected distance (APD), and the average nearest-point distance (ANPD) with 100 samples per barycentric dimension. The number of samples per barycentric dimension were chosen to coincide with the number of samples shown to achieve convergence in Chapter 4.3.1.

8.3 Results

After training, the mean dose map RMSE was $5.10\% \pm 0.84\%$ and $6.63\% \pm 1.65\%$ for the training dataset and testing dataset, respectively. While these errors are approximately twice as large as the dose map RMSE during prostate VMAT prediction, they are still relatively modest. For comparison, the International Commission on Radiation Units and Measurements (ICRU) and Task Group 142 of the American

Association of Physicists in Medicine (AAPM) have stated that a 5% maximum dosimetric uncertainty is appropriate for standard IMRT treatments such as ours (ICRU, 1976, Klein et al., 2009). Similarly, overfitting was approximately twice as dominant, with the testing errors being 30% larger than the training set errors.

Figure 43 shows side-by-side comparisons between the effect of prioritizing epidural space coverage or prioritizing spinal cord sparing in a dose map prediction and its corresponding TPS calculation. Importantly, the lower isodose lines in the predictions appear to be significantly more smooth and circular than the TPS calculations, which have beam streaks distorting their shapes. Additionally, the TPS-calculated dose maps appear to be different, with the spinal-cord-focused plan having a noticeably higher dose in a beam anterior to the PTV.

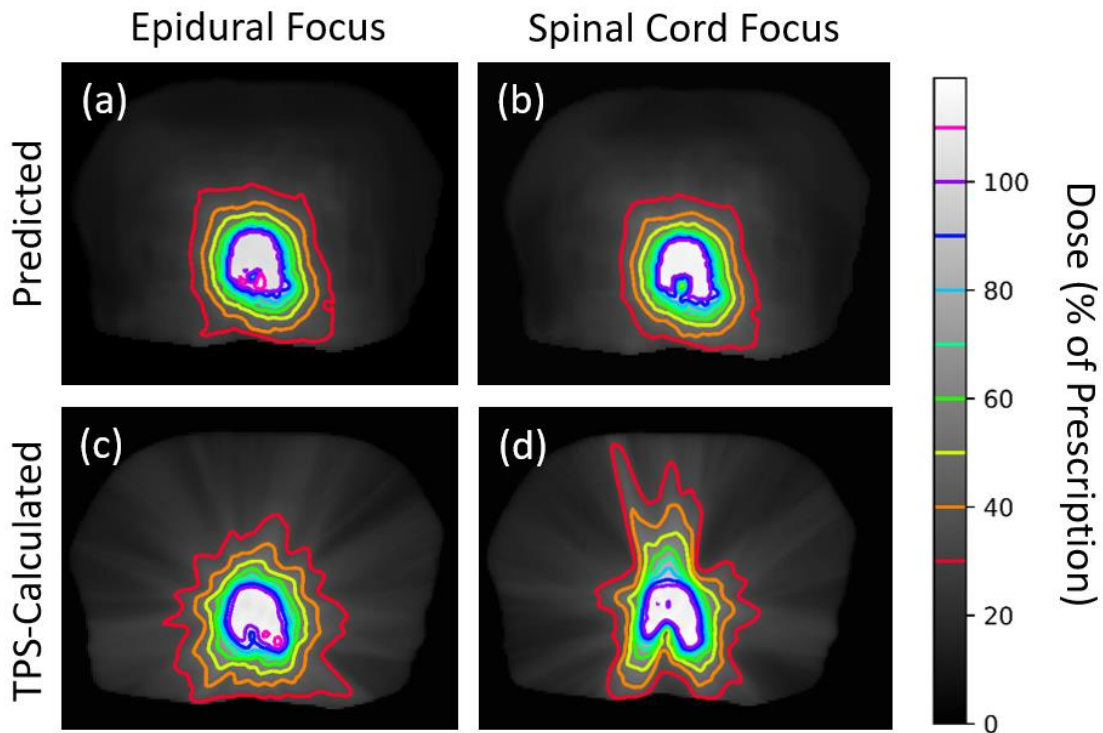


Figure 43: Comparison between dose distribution prediction and TPS calculation in plans prioritizing epidural space coverage or prioritizing spinal cord sparing. Transverse slices are taken from the center of the PTV, and the patient was randomly sampled from the testing dataset.

Figure 44 shows side-by-side comparisons between the effect of prioritizing epidural space coverage or prioritizing spinal cord sparing in a DVH prediction and its corresponding TPS calculation. Despite the differences between the dose map predictions and dose map simulations in Figure 43, their corresponding DVHs are much more similar. This is likely because the exterior distorted isodose shape in the TPS calculations do not directly affect the dose distributions within the PTV, epidural space, or spinal cord. Additionally, Figure 44 indicates that the prediction model is able to

predict DVHs with fundamentally different shapes, as the epidural space DVH changes noticeably between the epidural space prioritization and spinal cord sparing prioritization.

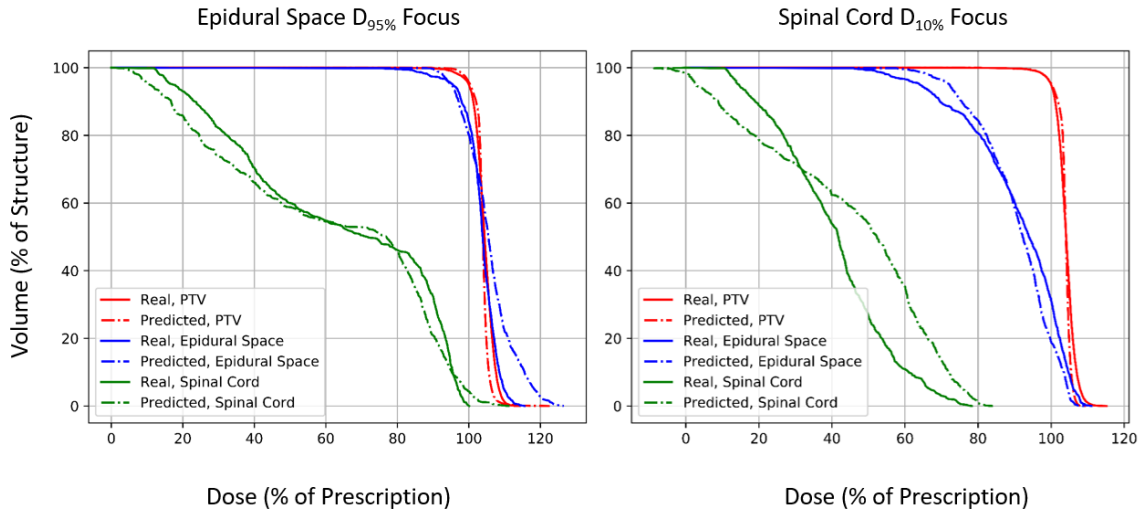


Figure 44: Comparison between indirect DVH prediction and TPS calculation in plans prioritizing epidural space coverage or prioritizing spinal cord sparing.

Table 8 compares the four Pareto surface similarity metrics described in Chapter 4.2.1, namely the Pareto space RMSE with 50 samples per barycentric dimension, the Hausdorff distance (HD), the average projected distance (APD), and the average nearest-point distance (ANPD) with 100 samples per barycentric dimension. The number of samples per barycentric dimension were chosen to coincide with the number of samples shown to achieve convergence in Chapter 4.3.1. Table 8 indicates that there is a modest difference between all four metrics.

Table 8: Pareto surface similarity metrics evaluated on the Pareto surfaces indirectly generated by our dose prediction model for spine SRS.

Pareto surface metric	Training dataset	Testing dataset
RMSE (50 samples)	10.13% \pm 2.28%	10.26% \pm 1.44%
HD	13.01% \pm 2.70%	13.00% \pm 2.08%
APD	9.04% \pm 2.28%	8.70% \pm 1.99%
ANPD (100 samples)	8.22% \pm 2.61%	7.59% \pm 1.92%

Additionally, the amount of time required to evaluate our model on all Pareto surface points for all patients in the training set was about 32 seconds. Therefore, predicting a single dose distribution required about 0.22 seconds per plan. This is about 4 times slower than our prostate VMAT dose prediction speed, which agrees with the doubled underlying dose distribution resolution. However, this is much faster than current optimization and calculation techniques, which take approximately 10-30 minutes per plan.

8.4 Discussion

Chapter 3 presented a novel machine learning dose prediction model which takes optimization objective priorities into account, allowing for indirect Pareto surface estimation. Our results indicate that the model is able to predict doses with good accuracy, as the root-mean-square predicted dose map errors are a few percentages of their corresponding TPS-calculated doses. However, these errors are approximately

twice as high as the errors observed in Chapter 6's analysis of prostate VMAT dose prediction. The dose map RMSEs are comparable the maximum error tolerance proposed by the ICRU and AAPM TG 142, indicating that our model is not quite sufficient for clinical application in spine SRS. However, the model produces just a dose distribution without actually creating a plan, so the model requires a final real plan optimization and dose calculation which will correct these dose map prediction errors prior to treatment delivery.

We have implemented several metrics for evaluating the error between a predicted Pareto surface and its corresponding TPS-calculated Pareto surface. Our metrics reported similar values around 10-15% of dose prescription for both training and testing sets. Again, it is worth noting that these metrics accumulate the errors from each dimension rather than averaging them, which is why these surface metrics are significantly larger than the dose map RMSE of 5-7%. Of these metrics, we hypothesize that the average nearest-point distance is the most appropriate of these metric due to its removal of error contributions orthogonal to the direction between the Pareto surfaces. The distinction appears to be meaningful as well, with the metrics all attaining somewhat different values on our training and testing datasets. Also, to our knowledge, no other body of research has applied Pareto space metrics to evaluate the Pareto surfaces of radiation therapy dose predictions. This prevents us from comparing our

Pareto space results with previous dose prediction research. To account for this, we have included all of these metrics for ease of comparison with future research.

Interestingly, the Pareto surface similarity metrics evaluated on our spine SRS predictions are actually comparable to the metrics evaluated on prostate VMAT predictions, despite the prostate dose map RMSEs being half as high. As seen in Figure 43, it is likely that the spine SRS dose map RMSEs are more heavily influenced by the beam streaks which occur. The problem of predicting these beam streaks is somewhat similar to the problem of predicting optimal beam angles in IMRT, which our model's architecture is not well-designed to handle. However, this problem is less impactful on the critical structures than the low-dose regions far from the critical structures, so it makes sense that the DVHs are not heavily impacted by these streaks. Although the dose distribution predictions are not as good, our model may be useful for inferring Pareto surfaces. Additionally, our results indicate that there is a modest difference between the four Pareto surface similarity metrics, suggesting that the choice of metric is significant when evaluating Pareto surface similarity. Our results also indicate that the model's overfitting to training data dose map RMSE is modest, and that overfitting to dose map RMSE does not appear to result in overfitting in any of the Pareto space metrics.

The prediction speed of our model is particularly encouraging. Although the dose prediction times for spine SRS are four times higher than conventional prostate

VMAT dose prediction times, the model still predicts each plan in approximately 0.22 seconds. With this speed, our model may be used for real-time treatment planning without needing to interpolate between previously sampled points, allowing the treatment planner to very quickly estimate the doses produced by a given optimization priority combination. This indirectly gives the planner more time to plan per patient, which may improve final plan quality. Moreover, our model only requires patient anatomy and optimization priorities, so it is capable of generating many samples from the Pareto surface automatically. In the context of spine SRS, our model may be used for very quickly and very finely estimating the range of possible dose distributions and DVHs for very large patient datasets.

8.5 Conclusion

We have applied the dose prediction model developed in Chapter 3 to predicting the dose distributions of spine SRS treatment plans with epidural irradiation. The model's error is modest, with average dose map root-mean-square errors of $6.63\% \pm 1.65\%$ over all patients and all optimization priority combinations in the patient testing dataset. Although these errors are higher than the errors observed in prostate VMAT dose prediction, the Pareto surface similarity metrics indicate that the model generalizes well to Pareto surface prediction. Therefore, our model may be used to accelerate dosimetric studies regarding the coverage of the epidural space in spine SRS.

9. Conclusions

The primary focus of this study was to develop, present, and analyze a machine-learning, voxel-wise dose prediction model. Such a dose prediction model would be clinically significant for accelerating the treatment planning process, both in the context of assisting treatment planners with clinical cases and advancing large-scale dosimetric studies into new treatment paradigms. In both contexts, the dose prediction model would greatly speed up the rate at which multi-criterial optimization (MCO) information can be inferred by allowing for rapid, fine sampling of the Pareto surface.

In this study, our dose prediction model was presented and developed in Chapter 3; subsequently, the model was analyzed in Chapters 6, 7, and 8 on conventional prostate VMAT treatment planning, pancreas SBRT treatment planning, and spine SRS treatment planning with epidural irradiation, respectively. For these cases, our model had modest errors, ranging between 2% and 7% of dose prescription. Additionally, our model appeared to be robust against overfitting, with testing dataset errors within 15% and 30% relative to training dataset errors. These results indicate that our model is proficient at predicting doses for a variety of treatment planning paradigms. However, the errors were worse for the pancreas SBRT and spine SRS cases than for the prostate VMAT cases. We believe that this is due to the presence of beam streaks in the stereotactic treatment plans which influence the dose distribution errors more heavily than the Pareto surface errors.

The secondary focus of this study was to develop, present, and analyze Pareto surface similarity metrics. This focus was essential to the proper evaluation of our dose prediction model because the Pareto surface errors do not perfectly correlate with dose distribution errors. Additionally, this focus has significant clinical significance beyond machine learning evaluation, since these metrics would likely be useful for powering future studies which seek to compare different MCO algorithms.

During development, four metrics were proposed to measure Pareto surface similarity and compared using abstract, theoretical simplex pairs in Chapter 4, namely the root-mean-square error (RMSE), the Hausdorff distance (HD), the average projected distance (APD), and the average nearest-point distance (ANPD). Subsequently, all four metrics were used to compare the Pareto surfaces which were indirectly generated by the dose prediction applications in Chapters 6, 7, and 8 for conventional prostate VMAT treatment planning, pancreas SBRT treatment planning, and spine SRS treatment planning with epidural irradiation, respectively. In these cases, we confirmed that Pareto surface similarity was not necessarily correlated with dose distribution RMSE; for example, prostate VMAT treatment planning had lower dose distribution RMSE, while pancreas SBRT and spine SRS had lower Pareto surface errors. Our analysis confirmed that these metrics converge rather quickly, and times for evaluating these metrics to convergence ranged from milliseconds to seconds. Ultimately, we concluded that the ANPD was the most appropriate for evaluating Pareto surface similarity. Additionally,

the numerical difference between the ANPD and the other metrics indicated that the choice of metric had a measurable impact on the surface similarity inferred.

The tertiary focus of this study was to determine the feasibility of deliberately irradiating the epidural space during spine SRS treatment planning. This focus was clinically significant because it presented an alternative treatment planning contouring paradigm which may raise epidural space local control rates while preserving current spinal cord toxicity rates. A dosimetric investigation of the spinal cord was included in Chapter 5, in which many treatment plans were created to evaluate the potential to increase epidural space coverage while preserving PTV coverage constraints and spinal cord sparing constraints. The investigation concluded that the epidural space $D_{95\%}$ could be raised to $16.84 \text{ Gy} \pm 0.87 \text{ Gy}$, which is significantly higher than the average incidental epidural space $D_{05\%}$ of $10.96 \text{ Gy} \pm 1.76 \text{ Gy}$. A simplistic biological model was implemented to estimate the significance of this dose escalation, concluding that the deliberate doses near 17 Gy would very likely increase local control rates. This study supports the justification of potential future clinical trials on epidural space irradiation in spine SRS which seek to preserve constant dosimetric spinal cord constraints.

The overall impact of this study is to streamline the multi-criterial optimization workflow. All three focuses of this study were successfully met, with thorough investigation and analysis into the significance of their results. By using our dose prediction model, it is possible to greatly accelerate the treatment planning process,

indirectly improving the quality of the patient care provided by plans which use our model. Given the lowered survival rates and worsened qualities of life which cancer patients must endure, this body of research can have a significant, meaningful impact for cancer patients by improving their probabilities of survival as well as improving their qualities of life.

References

- Al-Omair, A., Masucci, L., Masson-Cote, L., Campbell, M., Atenafu, E. G., Parent, A., Letourneau, D., Yu, E., Rampersaud, R., Massicotte, E., Lewis, S., Yee, A., Thibault, I., Fehlings, M. G. & Sahgal, A. 2013. Surgical resection of epidural disease improves local control following postoperative spine stereotactic body radiotherapy. *Neuro Oncol*, 15, 1413-9.
- Appenzoller, L. M., Michalski, J. M., Thorstad, W. L., Mutic, S. & Moore, K. L. 2012. Predicting dose-volume histograms for organs-at-risk in IMRT planning. *Med Phys*, 39, 7446-61.
- Babier, A., Mahmood, R., McNiven, A. L., Diamant, A. & Chan, T. C. Y. 2020. Knowledge-based automated planning with three-dimensional generative adversarial networks. *Medical Physics*, 47, 297-306.
- Barzilai, O., Boriani, S., Fisher, C. G., Sahgal, A., Verlaan, J. J., Gokaslan, Z. L., Lazary, A., Bettogowda, C., Rhines, L. D. & Laufer, I. 2019. Essential Concepts for the Management of Metastatic Spine Disease: What the Surgeon Should Know and Practice. *Global Spine J*, 9, 98S-107S.
- Berezkin, V. & Lotov, A. 2014. Comparison of two Pareto frontier approximations. *Computational Mathematics and Mathematical Physics*, 54, 1402-1410.
- Bokrantz, R. & Forsgren, A. 2013. An Algorithm for Approximating Convex Pareto Surfaces Based on Dual Techniques. *INFORMS Journal on Computing*, 25, 377-393.
- Brenner, D. J. 2008. The linear-quadratic model is an appropriate methodology for determining isoeffective doses at large doses per fraction. *Semin Radiat Oncol*, 18, 234-9.
- Chang, E. L., Shiu, A. S., Mendel, E., Mathews, L. A., Mahajan, A., Allen, P. K., Weinberg, J. S., Brown, B. W., Wang, X. S., Woo, S. Y., Cleeland, C., Maor, M. H. & Rhines, L. D. 2007. Phase I/II study of stereotactic body radiotherapy for spinal metastasis and its pattern of failure. *J Neurosurg Spine*, 7, 151-60.
- Cox, B. W., Spratt, D. E., Lovelock, M., Bilsky, M. H., Lis, E., Ryu, S., Sheehan, J., Gerszten, P. C., Chang, E., Gibbs, I., Soltys, S., Sahgal, A., Deasy, J., Flickinger, J., Quader, M., Mindea, S. & Yamada, Y. 2012. International Spine Radiosurgery Consortium Consensus Guidelines for Target Volume Definition in Spinal

Stereotactic Radiosurgery. *International Journal of Radiation Oncology*Biological*Physics*, 83, e597-e605.

Craft, D., Halabi, T., Shih, H. A. & Bortfeld, T. 2007. An approach for practical multiobjective IMRT treatment planning. *Int J Radiat Oncol Biol Phys*, 69, 1600-7.

Ericson, C. 2004. *Real-Time Collision Detection*, CRC Press, Inc.

Fogliata, A., Nicolini, G., Bourgier, C., Clivio, A., De Rose, F., Fenoglietto, P., Lobefalo, F., Mancosu, P., Tomatis, S., Vanetti, E., Scorsetti, M. & Cozzi, L. 2015. Performance of a Knowledge-Based Model for Optimization of Volumetric Modulated Arc Therapy Plans for Single and Bilateral Breast Irradiation. *PLoS One*, 10, e0145137.

Garg, A. K., Wang, X. S., Shiu, A. S., Allen, P., Yang, J., McAleer, M. F., Azeem, S., Rhines, L. D. & Chang, E. L. 2011. Prospective evaluation of spinal reirradiation by using stereotactic body radiation therapy: The University of Texas MD Anderson Cancer Center experience. *Cancer*, 117, 3509-16.

Gilbert, E. G., Johnson, D. W. & Keerthi, S. S. 1988. A fast procedure for computing the distance between complex objects in three-dimensional space. *IEEE Journal on Robotics and Automation*, 4, 193-203.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. 2014. Generative adversarial nets. *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. Montreal, Canada: MIT Press.

Grimm, J., Sahgal, A., Soltys, S. G., Luxton, G., Patel, A., Herbert, S., Xue, J., Ma, L., Yorke, E., Adler, J. R. & Gibbs, I. C. 2016. Estimated Risk Level of Unified Stereotactic Body Radiation Therapy Dose Tolerance Limits for Spinal Cord. *Semin Radiat Oncol*, 26, 165-71.

Guckenberger, M., Richter, A., Krieger, T., Wilbert, J., Baier, K. & Flentje, M. 2009. Is a single arc sufficient in volumetric-modulated arc therapy (VMAT) for complex-shaped target volumes? *Radiother Oncol*, 93, 259-65.

Hall, E. J. & Giaccia, A. J. 2019. *Radiobiology for the radiologist*, Philadelphia, Wolters Kluwer.

- He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 27-30 June 2016 2016. 770-778.
- Hwang, C. L. & Masud, A. S. M. 1979. *Multiple objective decision making, methods and applications: a state-of-the-art survey*, Springer-Verlag.
- Jaderberg, M., Simonyan, K., Zisserman, A. & Kavukcuoglu, K. 2015. Spatial transformer networks. *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. Montreal, Canada: MIT Press.
- Kajikawa, T., Kadoya, N., Ito, K., Takayama, Y., Chiba, T., Tomori, S., Nemoto, H., Dobashi, S., Takeda, K. & Jingu, K. 2019. A convolutional neural network approach for IMRT dose distribution prediction in prostate cancer patients. *Journal of Radiation Research*, 60, 685-693.
- Kirkpatrick, J. P., Kelsey, C. R., Palta, M., Cabrera, A. R., Salama, J. K., Patel, P., Perez, B. A., Lee, J. & Yin, F. F. 2014. Stereotactic body radiotherapy: a critical review for nonradiation oncologists. *Cancer*, 120, 942-54.
- Kirkpatrick, J. P., van der Kogel, A. J. & Schultheiss, T. E. 2010. Radiation dose-volume effects in the spinal cord. *Int J Radiat Oncol Biol Phys*, 76, S42-9.
- Klimo, P., Jr. & Schmidt, M. H. 2004. Surgical management of spinal metastases. *Oncologist*, 9, 188-96.
- Krizhevsky, A., Sutskever, I. & Hinton, G. 2012. ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems*, 25.
- Loblaw, D. A., Perry, J., Chambers, A. & Laperriere, N. J. 2005. Systematic review of the diagnosis and management of malignant extradural spinal cord compression: the Cancer Care Ontario Practice Guidelines Initiative's Neuro-Oncology Disease Site Group. *J Clin Oncol*, 23, 2028-37.
- Mardani, M., Dong, P. & Xing, L. 2016. Deep-Learning Based Prediction of Achievable Dose for Personalizing Inverse Treatment Planning. *International Journal of Radiation Oncology • Biology • Physics*, 96, E419-E420.
- Miettinen, K. 1999. *Nonlinear Multiobjective Optimization*, Springer US.

- Nelson, J. W., Yoo, D. S., Sampson, J. H., Isaacs, R. E., Larrier, N. A., Marks, L. B., Yin, F. F., Wu, Q. J., Wang, Z. & Kirkpatrick, J. P. 2009. Stereotactic body radiotherapy for lesions of the spine and paraspinal regions. *Int J Radiat Oncol Biol Phys*, 73, 1369-75.
- Nguyen, D., Long, T., Jia, X., Lu, W., Gu, X., Iqbal, Z. & Jiang, S. 2019. A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning. *Scientific reports*, 9, 1076-1076.
- Oinam, A. S., Singh, L., Shukla, A., Ghoshal, S., Kapoor, R. & Sharma, S. C. 2011. Dose volume histogram analysis and comparison of different radiobiological models using in-house developed software. *J Med Phys*, 36, 220-9.
- Otto, K. 2008. Volumetric modulated arc therapy: IMRT in a single gantry arc. *Medical Physics*, 35, 310-317.
- Sahgal, A., Ma, L., Gibbs, I., Gerszten, P. C., Ryu, S., Soltys, S., Weinberg, V., Wong, S., Chang, E., Fowler, J. & Larson, D. A. 2010. Spinal cord tolerance for stereotactic body radiotherapy. *Int J Radiat Oncol Biol Phys*, 77, 548-53.
- Sender, R., Fuchs, S. & Milo, R. 2016. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol*, 14, e1002533.
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T. & Hassabis, D. 2017. Mastering the game of Go without human knowledge. *Nature*, 550, 354-359.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. 2015. *Going deeper with convolutions*.
- Teichert, K., Süß, P., Serna, J. I., Monz, M., Küfer, K. H. & Thieke, C. 2011. Comparative analysis of Pareto surfaces in multi-criteria IMRT planning. *Physics in medicine and biology*, 56, 3669-3684.
- Teoh, M., Clark, C. H., Wood, K., Whitaker, S. & Nisbet, A. 2011. Volumetric modulated arc therapy: a review of current literature and clinical use in practice. *Br J Radiol*, 84, 967-96.

- Thibault, I., Campbell, M., Tseng, C. L., Atenafu, E. G., Letourneau, D., Yu, E., Cho, B. C., Lee, Y. K., Fehlings, M. G. & Sahgal, A. 2015. Salvage Stereotactic Body Radiotherapy (SBRT) Following In-Field Failure of Initial SBRT for Spinal Metastases. *Int J Radiat Oncol Biol Phys*, 93, 353-60.
- Trakul, N., Koong, A. C. & Chang, D. T. Stereotactic body radiotherapy in the treatment of pancreatic cancer. *Seminars in radiation oncology*, 2014. Elsevier, 140-147.
- Yuan, L., Ge, Y., Lee, W. R., Yin, F. F., Kirkpatrick, J. P. & Wu, Q. J. 2012. Quantitative analysis of the factors which affect the interpatient organ-at-risk dose sparing variation in IMRT plans. *Med Phys*, 39, 6868-78.
- Zhang, P., Happersett, L., Hunt, M., Jackson, A., Zelefsky, M. & Mageras, G. 2010. Volumetric modulated arc therapy: planning and evaluation for prostate cancer cases. *Int J Radiat Oncol Biol Phys*, 76, 1456-62.

Biography

Patrick James Jensen Jr. joined the University of Chicago in 2012 to pursue his Bachelor's degree. He graduated in the June of 2016 and earned his Bachelor of Arts degree in physics with honors and his Bachelor of Science degree in mathematics. Soon after graduation, he came to Duke University and joined the Medical Physics Graduate Program to pursue his Doctoral degree in medical physics. During his four years of study at Duke, James published two first-author, peer-reviewed articles: "A fast, novel machine learning model for dose prediction in prostate volumetric modulated arc therapy using output initialization and optimization priorities" in *Physics and Medicine and Biology (PMB)*, and "Purposeful Irradiation of the Epidural Space to Enhance Local Control without Compromising Cord Sparing in Spine Radiosurgery" in the *International Journal of Radiation Oncology • Biology • Physics (IJROBP)*. He has given four presentations, including three oral presentations, at the annual meetings of the American Association of Physicists in Medicine (AAPM), the American Society of Radiation Oncology (ASTRO), and the International Stereotactic Radiosurgery Society (ISRS). He has also been awarded the James B. Duke Fellowship from the Graduate School of Duke University.