Iowa Initiative for Artificial Intelligence
Final Report

Project title:	Reinforcement Learning Based Adaptive Treatment Planning for Prostate Cancer
Principal Investigator:	St-Aubin Joel, PhD, FCCPM
Prepared by (IIAI):	Avinash Mudireddy
Other investigators:	Shaffer Nathan
Date:	06/22/2024

Were specific aims fulfilled:	Ongoing
Readiness for extramural proposal?	No
If yes … Planned submission date	
Funding agency	
Grant mechanism	
If no … Why not? What went wrong?	The project aims that would have provided the preliminary data for an external grant were not achieved at this point. The main aim of generating a RL-network to solve the given problem proved to be much more challenging than originally anticipated. Significant time was spent learning the subtleties of designing the architecture for the RL network, iterating the network structure based on gathered results, and shaping the rewards. However, significant progress has been made and the expectation is that this preliminary data will be generated within the upcoming year.

Brief summary of accomplished results:

This ongoing, multi-year project has explored various approaches in the realm of reinforcement learning (RL). Initially, we experimented with Multi-Agent Deep Deterministic Policy Gradient (MADDPG) Actor-Critic networks and tested several custom loss functions to enhance performance. Recently, we have transitioned to a Policy Optimization framework. This new direction utilizes a neural network initialized through supervised learning, which is showing hopeful results.

Research report:

Aims (provided by PI):

Background

Figure 1: Schematic cutaway of the Elekta Unity MR-Linac used for adaptive radiotherapy (Elekta.com).

The purpose of radiation therapy treatments is to kill tumor cells. High energy x-ray radiation (photons), interact with patient tissues and sets in motion high energy electrons which deposit energy in the tissue causing DNA damage. When tumor cell DNA is damaged, they generally cannot repair due to their mutations. The amount of tumor cell DNA damage is proportional to the energy deposited by the electrons. Healthy tissue also undergoes DNA damage by radiation, so the primary goal of radiation therapy is to maximize the radiation dose to the tumor while minimizing dose to the healthy tissues. The process of calculating radiation dose deposition in human tissue requires solving the linear Boltzmann transport equation (LBTE) which is a complex integro-partial differential equation. Furthermore, maximizing radiation dose to the tumor while minimizing dose to healthy tissue is complex and requires sophisticated optimization algorithms. The process of dose calculation and optimization is known as treatment planning. The standard treatment planning process currently requires significant user interaction, is time consuming (many hours to complete), and the results of the treatment process is highly dependent on the skill of the user. The purpose of this research study is to develop a method of treatment planning that is fast and consistent without user interaction. The ability to rapidly generate a high quality radiation treatment plan is especially important for adaptive radiotherapy where the plan is created based on the patient anatomy visualized while the patient lies on the treatment table waiting for treatment. An example technology that enables adaptive radiotherapy is the Elekta Unity MR-Linac (Fig. 1).

Radiation Technology

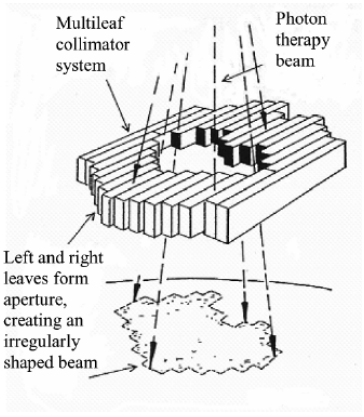

Figure 2: Schematic diagram of multileaf collimators (Romeijn et al.,SIAM J. Optim, (2005), 838-862).

Radiation therapy is delivered using medical linear accelerator (linac) technology. Because tumors can be all shapes and sizes, the linac is capable of shaping the radiation beam using a collimation system known as multileaf collimators (MLCs) (Fig. 2). The Elekta Unity system, which is the system focus of this work, uses an MLC system is comprised of two parallel banks of 80 MLCs for a total of 160 independent leaves. In order to maximize tumor dose and minimize normal tissue dose, the radiation beam is delivered to the patient from many different angles and the MLCs shape the radiation beam at each angle. To add additional conformality to the treatments, the MLCs can also modulate the radiation intensity at each beam angle. This modulation of radiation intensity for a treatment is known as intensity modulated radiation therapy (IMRT) and is the standard treatment practice. Thus, from each beam angle the MLCs can move to create multiple independent shapes with each shape having a set amount of radiation intensity. Each independent shape and associated radiation intensity combination is known as a *segment*. The machine parameters to deliver an IMRT plan are the MLC positions for each segment, the amount of radiation intensity delivered in each segment, and the number of segments per beam angle. In order to ensure that the linac can accurately deliver the radiation to the patient, there are two constraints applied to each segment: (1) a minimum segment open area (calculated from the MLC leaf positions), and (2) an minimum radiation intensity. A pictorial example of the segment open area is given by the orange

shaded regions in Fig. 3. Quantitatively it would be calculated by adding the difference between opposing leaf positions (MLC bank A and MLC bank B) and multiplying by the known MLC leaf width (Eq. 1).

$$open\ area = \sum_i width_{MLC} \times (MLC_{A,i} - MLC_{B,i})$$

Treatment Planning

The purpose of treatment planning is to leverage the ability of linacs to produce complex radiation patterns and develop a radiation plan that maximize tumor dose while minimizing healthy tissue dose. As part of the treatment planning process, the physician will specify the amount of radiation dose required to treat a tumor and the maximum dose each healthy organ can receive without permanent damage. These specifications form the *constraints* of the optimization problem. A treatment plan is considered ideal if the tumor dose is achieved while the dose to healthy organs is minimized. An ideal treatment plan would have full dose to the tumor and zero dose to healthy tissue but based on the physics of radiation transport (governed by the LBTE) this is impossible. Thus, in practice there are always compromises in achieving the maximum tumor dose and minimizing healthy tissue dose. This is where experience and skill of the user plays a large role in the treatment planning process.

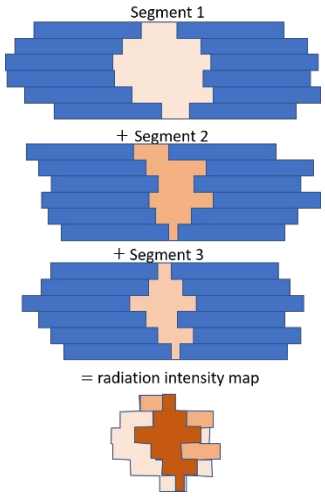

Figure 3: Schematic example of different segments that add together to make a radiation fluence map.

During the optimization process, the number of segments per beam and the individual segment shapes with their radiation intensities are determined. When all segments from a particular beam angle are added together, a radiation fluence map is generated (Fig. 3). For the prostate treatment plans that this project will be focusing on, there are 9 radiation beam angles with each beam angle comprised of a varying number of segments. Although a maximum of 90 segments is used for each radiation plan, the number of segments per beam can vary from 5 – 15 depending on the plan. Once the radiation fluence maps are generated within the treatment planning system, a final dose calculation is performed using those radiation fluence maps as the radiation source. Ultimately, the optimization process begins with a calculation of dose in the patient based on an initial radiation fluence map guess. The results of the dose calculation are compared to the constraints provided and the segments are individually adjusted in a manner to more fully meet the constraints. Once the segments are adjusted, a new radiation fluence map is created, and the dose is calculated in the patient again. This process continues until the constraints are met, or the maximum number of iterations is met.

Research Goals

The overarching goal of this research is to develop an artificial intelligence (AI) network that rapidly predicts the machine parameters to deliver a clinically acceptable prostate cancer treatment. As this can require compromises (e.g. allowing lower target dose to spare healthy tissues), a method that allows the clinician to review multiple plans with varying objectives would allow the clinician to select the plan that suits clinical judgement in the absence of a 'perfect plan'. The AI-based approach to develop multiple plans with varying objectives will be termed *AI-based multicriterial optimization (AI-MCO)*. Ultimately, a reinforcement learning (RL) network that uses synthetic data (i.e. different optimization constraints) seems to be the most appropriate option for AI-MCO, but such a network has yet to be developed. Thus, in order to achieve the goal of an AI-MCO method using RL, the research is broken down into phases, with the first aim the subject of this grant.

Aim 1: Development of a RL network to predict machine parameters for prostate cancer treatments The number of radiation segments per beam angle, individual position of the multileaf collimator (MLC) leaves used to shape the radiation beam, and dose per radiation segment are to be

predicted based on anatomical and tumor contours and tissue density information provided by CT or density assigned MRI.

Phase I – IIAI pilot grant

The aim of this phase is to develop the foundation of a RL-based network that ultimately could be used for AI-MCO. The ability to generate different plans based on different optimization constraints for the RL network has not been built yet and will be the subject of a future R01 application. However, the premise for this phase is that a RL-based network can be built to predict a single clinically acceptable prostate cancer treatment plan. For this work, we have access to clinically approved treatment plans that treat prostate cancer for 100 patients. Each treatment plan is composed of 9 radiation beam angles, and roughly 90 segments for each plan.

The input to the RL-based network would be anatomical contours and CT data projected to the beams-eye-view of each of the 9 radiation beam angles for a given patient. The output of the network would be the machine parameters (number of segments, radiation intensity per segment, and MLC positions per segment) (Fig. 4).

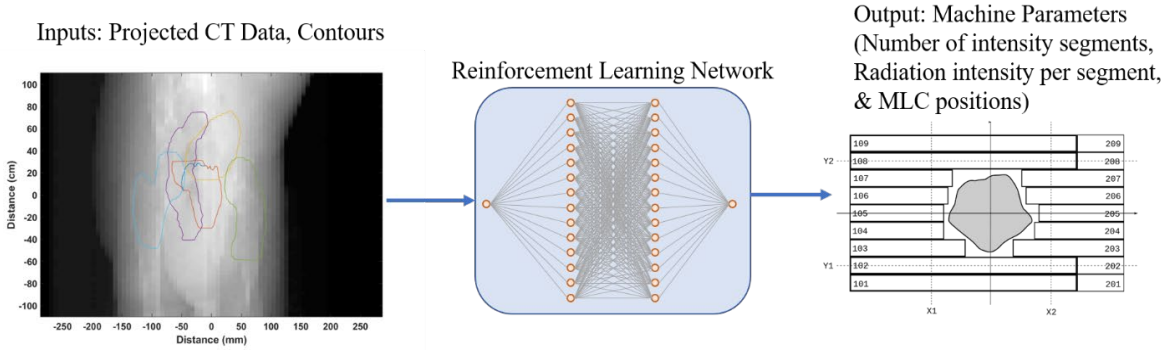

Figure 4: Schematic diagram of the inputs and outputs of the proposed RL network

The two physical constraints that would be added to the network would be the minimum open area size, and the minimum radiation intensity per segment.

Calculation of the reward for the RL-network will be performed by comparing the dosimetric results of the prediction to the physician requirement. In order to do this, a rapid dose calculation has been developed to calculate the radiation dose for each predicted machine parameter set from the RL-network. It is recognized that it is possible that many different combinations of machine parameters (i.e., MLC positions and segment intensity) can lead to a very similar dosimetric results. However, as long as each segment meets the constraints of minimum open area and minimum radiation intensity, only the final radiation dose matters. In order to reduce the number of variables the network has to predict, the following is proposed:

1. Each beam angle for any patient is assumed to have a maximum of 15 segments.
2. Only the central 26 MLC leaf pairs need to be used for an optimal radiation plan (20 cm opening). This is a reasonable assumption for prostate treatments.

This reduces the number of variables to 7,155 (instead of 21,600 for 160 MLC leaves).

The goal of this phase, for which the pilot grant focuses, is to generate a RL-network that uses the patient tumor and organ contours with CT data to predict the machine parameters that yields a clinically acceptable prostate cancer treatment plan.

AI/ML Approach:

Phase 1: MADDPG Actor-Critic Networks

Objective: To implement and evaluate the effectiveness of Multi-Agent Deep Deterministic Policy Gradient (MADDPG) in a multi-agent environment.

Approach: The MADDPG algorithm was chosen for its capability to handle environments with multiple agents. This algorithm uses separate actor-critic networks for each agent, which allows for learning in a coordinated yet decentralized manner.

Environment Setup:

- We used a custom multi-agent environment designed to simulate complex interactions between agents.
- The environment's state space was designed to include a state which represents initial openings of the leaves along with the contours of the target and surrounding organs near the region of interest.
- The action space for each agent was defined to predict the change of leaf openings for each of 15 segments per agent, as well as the dosage. This complex action space required precise adjustments to optimize the collective performance of all agents.

MADDPG Implementation:

- **Actor Networks**: The actor network for each agent was implemented using Convolutional Neural Networks (CNNs) and Dense layers. Each actor network was designed to predict the best action given the current state.
- **Critic Networks**: The critic network was designed to evaluate the action taken by the actor. It was implemented using Dense layers to predict a Q-value.

Training Process:

1. **Experience Replay**: A replay buffer was used to store experiences (state, action, reward, next state) for training. This helps in breaking the correlation between consecutive samples, leading to more stable learning.
2. **Actor-Critic Update**:
    - **Critic Network Update**: The critic network update is based on Temporal Difference (TD) error. The target Q-value is computed as the sum of the immediate reward and the discounted Q-value of the next state (estimated by the target critic network). The critic network minimizes the mean squared error between its Q-value prediction and the target Q-value.
    - **Actor Network Update**: Update using policy gradient method to maximize expected return by maximizing the Q-value
3. **Target Networks**: Both actor and critic target networks were periodically updated using a soft update mechanism to improve training stability. This involves slowly updating the target network parameters to track the learned networks.

Challenges:

- **Exploration vs. Exploitation**: Balancing the trade-off between exploring new actions and exploiting known rewarding actions was critical. Initially, a high exploration rate was used, which decayed over time.
- **Training Stability**: Given the complexity of multi-agent interactions, maintaining stability in training required careful tuning of hyperparameters and loss functions.

Results:

- The initial experiments demonstrated the ability of MADDPG to learn cooperative behaviors among agents.
- The convergence was slow and non-optimal, with occasional instability due to the high dimensionality of the action space and complex agent interactions.
- Custom loss functions were tested to improve training stability, which showed some improvements in specific scenarios.

Visualization and Metrics:

- Training progress was monitored using episodic reward plots.
- The performance of each agent was evaluated based on the average reward over time.
- Convergence was assessed by observing the reduction in loss values and the stabilization of policy performance.

Conclusion: Phase 1 of the project demonstrated the potential of MADDPG in multi-agent environments. Despite the challenges in training stability and convergence speed, the algorithm showed promise in learning complex interactions among agents. This phase provided valuable insights and set the foundation for exploring more advanced techniques in subsequent phases.

Phase 2: Custom Loss Functions

Objective: To improve our reinforcement learning models' training stability and convergence rate by designing and experimenting with custom loss functions tailored to our specific problem domain.

Approach: During this phase, we explored various loss functions to enhance the performance and stability of the MADDPG Actor-Critic networks.

Initial Loss Function:

- **Comparison of Predicted and Actual Fluence Maps**: Our initial approach involved designing a loss function that compared the predicted fluence maps generated by the model with the actual fluence maps. The error was calculated as the mean squared error (MSE) between the predicted and actual values. This approach aimed to directly minimize the discrepancy between the model's predictions and the desired outcomes.

Exploration of Combined Loss Functions:

- **Linear Weighted Combination**: To capture more aspects of the problem, we experimented with a linear weighted combination of various loss components:

    - **Expected Reward**: This component aimed to maximize the expected return, encouraging the agent to perform actions that would yield higher rewards.
    - **Boundary Loss**: This loss component was designed to penalize actions that resulted in fluence map values exceeding predefined boundaries, ensuring the generated maps stayed within acceptable limits.
    - **Step Loss**: This component focused on minimizing the number of steps taken to achieve the desired outcome, promoting efficiency in reaching the goal state.

Challenges and Insights:

- **Complexity in Multi-Agent Settings**: Designing effective loss functions for multi-agent environments proved to be complex. Each agent's actions could influence the overall outcome, requiring a delicate balance between individual and collective performance.
- **Trade-Offs in Loss Components**: Balancing the different components of the loss function was challenging. Over-emphasizing one component could lead to suboptimal performance in other areas, highlighting the need for a holistic approach to loss design.
- **Iterative Process**: The process of refining the loss functions was iterative, involving continuous experimentation and evaluation. While some combinations showed improvements in specific scenarios, a generalized optimal loss design remained elusive.

Current Experiments: We are continuing to experiment with different loss functions, adjusting weights, and exploring new formulations to find an optimal design. The quest for an effective and generalized loss function remains a work in progress.

Conclusion: Phase 2 highlighted the importance of carefully designed loss functions in reinforcement learning. While significant progress was made in improving training stability and convergence, the quest for an optimal loss design continues. The insights gained from this phase provide a valuable foundation for further refinement and experimentation in future phases.

Phase 3: Policy Optimization Framework

Objective: To enhance the overall performance and reliability of our reinforcement learning models by transitioning to a Policy Optimization framework. This phase focuses on implementing a neural network with supervised learning initialization and plans to incorporate Proximal Policy Optimization (PPO) in future iterations.

Approach: In this phase, we shifted to a Policy Optimization framework with the following key elements:

Current Neural Network Design:

- **Initialization Using Supervised Learning**: We are in the process of initializing the neural network using supervised learning. This aims to provide a strong starting policy, reducing the need for extensive exploration and helping to stabilize the initial training phase.
- **Neural Network Architecture**:
    - **Input Layers**: The model accepts three primary inputs:
        1. **State Inputs**: Time series data of previous and current states, which include relevant parameters for each agent.
        2. **Reward Inputs**: Time series data of previous rewards.
        3. **Action Inputs**: Time series data of previous actions taken by the agents.

- **Convolutional and GRU Layers**: The network features Conv3D layers for processing spatial data from the states and GRU layers for handling sequential data from previous action and reward arrays.
- **Dynamic Masked Dense Layers**: These custom layers allow the network to dynamically adjust based on the input data, providing more flexibility in handling various scenarios.
- **ResNet Blocks**: Utilized for enhancing feature extraction capabilities, ResNet blocks help in learning deeper representations from the input data.
- **Output Layers**: The network outputs predictions for changes in leaf openings and dosages, designed to be applied to the agents' actions.

Training Process:

1. **Data Preparation**: Initial states, actions, and RP arrays are loaded to set up the environment and agents.
2. **Experience Collection**: Agents interact with the environment to collect experiences, including states, actions, rewards, and next states.
3. **Policy Update**: Currently, the focus is on stabilizing the training process using supervised initialization. Future plans include incorporating PPO for more robust policy updates.
4. **Training Stability**: The supervised initialization is expected to enhance stability during the initial training phases.

Challenges and Insights:

- **Balancing Exploration and Exploitation**: While supervised initialization reduces the need for extensive exploration, ensuring that agents continue to explore sufficiently remains a challenge.
- **Training Stability**: Achieving stable training through supervised initialization and preparing for future PPO implementation is critical.

Conclusion: Phase 3 represents an important step in transitioning to a Policy Optimization framework. The ongoing implementation of a neural network with supervised learning initialization is expected to provide a strong foundation for more stable and efficient training. Future incorporation of PPO will further enhance the robustness and effectiveness of the models in achieving desired outcomes. Continuous experimentation and refinement are key to driving improvements in model performance in this complex multi-agent environment.

DISCUSION

Although the main aim of this project has not yet been achieved, significant progress has been made in the development of a RL-network that can rapidly predict the machine parameters for a clinically acceptable prostate cancer treatment. Specific improvements that have been achieved are,

1. Understanding of network architecture: The move towards Policy Optimization utilizing temporal information shows significant promise.
2. Reward shaping: It remains an art in many RL projects to define the proper reward function that avoids unexpected and unwanted behavior. Significant progress has been made in the shaping of a dosimetric reward for this project.
3. Supervised learning initialization: Experience gained early in this work showed that significant exploration led to training convergence issues. Thus, utilizing a supervised learning initialization approach is expected to improve convergence substantially.

The development of a RL-network for the creation of clinically acceptable treatment plans will provide two major improvements over the current research in this area. (1) Directly predicting the machine

parameters for a clinical prostate cancer treatment will guarantee that the results are deliverable by the current linac technology. This is not true of other approaches where only the treatment plan is produced without the machine parameters. (2) RL learns the policy to optimally move through the environment. By tying this policy to real dosimetric results (though the rewards) that clinicians understand, generalization to other objectives and treatment sites is possible.

Ideas/aims for future extramural project:

With the development of a RL-network that predicts the machine parameters for a clinical prostate cancer treatment plan, further development allowing for the AI-MCO will be possible. By allowing the RL-network to explore additional states that encompass different optimization criteria, the clinician will be able to adjust the AI-based results in real-time – something that has never been achieved previously but is critical to clinical deployment. This work would be the subject of a future NIH R01 grant.

Additionally, with the RL-network trained to generate prostate cancer treatment plans, transfer learning will be utilized to broaden the scope of the network to predict treatment plans for other sites (e.g., head and neck, lung, liver, etc.).