Clinical Knowledge Based Reinforcement Learning for Radiotherapy Treatment Planning System

Paul Dubois[1,2], Nikos Paragios[1], and Paul-Henry Cournède[2]

[1]TheraPanacea, Paris, France
[2]MICS, CentraleSupélec, Université Paris-Saclay, Paris, France

Abstract Achieving optimal dose distribution in radiation therapy planning is a complex task, with contradicting goals. Yet, this step is crucial with profound implications for patient treatment and toxicity management.

The absence of universally agreed-upon constraints prioritization in radiation therapy planning complicates the definition of an optimal plan, requiring a delicate balance between multiple objectives. This balanced usually ends up being done manually.

The optimization process is further hindered by complex mathematical aspects, involving non-convex multi-objective inverse problems with a vast solution space. Expert bias introduces variability in clinical practice, as treatment planning is shaped by the preferences and expertise of radiation oncologists and medical physicists.

To surmount these challenges, we propose a first step towards a fully automated approach, using an innovative deep learning method that allows automatic navigation towards acceptable solution. We successfully trained an agent evaluating actions of a human dosimetrist, in order to reach a plan similar to past history. As this is very new and ongoing research, we generated synthetic phantom patients and associated trust-able clinical dose. In future work, we hope to be able to apply this technique to real cases.

1 Introduction

In contemporary radiation therapy, photon intensity modulated radiation therapy (IMRT) stands as a pivotal technique utilized to attain precise and conformal dose distributions within target volumes [16]. This achievement owes its realization chiefly to the advent of the multileaf collimator (MLC) [4].

Radiation therapy is now a reliable treatment for oncology [12]. Despite this consensus, the way to deliver radiotherapy for its best result remain very dependent upon doctors. Moreover, it appears that there is a large variability across physicians and centers, bot in terms of 3D structures contouring and irradiation constraints priorities [3].

To achieve the best treatment, doctors need to solve a complex inverse mathematical optimization problem with multiples trade-offs [9] [13]. There is a lack of standardized prioritization of constraints make the optimization a real challenge. The standard procedure nowadays is to manually guide computer optimization: dosimetrists manually update the settings of an optimizing software (so called Treatment Planning System) [1].

There has been many tries to create a metric that quantify the quality of a treatment plan: Normal tissue complication probabilities (NTCP), Target coverage, Conformity index, Heterogeneity index (non-exhaustive list) [7] [6]. However, none of them has been able to satisfy all radio-oncologists, and the only reliable way of assessing a plan for doctors is to check out the dose-volume histograms (DVHs) them-self.

As a result, Pareto surface exploration are doomed to failure due to the lack of impartial quantitative measurement for a particular plan [5]. Other meta-optimization techniques are similarly bounded, for the same reason [14] [15]. An extra challenge to attend for those is the fact that not all cases have the same "difficulty". Hence, for an "easy" case, doctors will require an excellent dose (in terms of the metrics mentioned above), while they can be more permissive for "harder" cases. This make the acceptability of a plan hard to define in general.

Reinforcement learning is a machine learning paradigm concerned with training agents to make sequential decisions in dynamic environments [2]. Through a process of trial and error guided by rewards or penalties, agents learn to optimize their actions to achieve long-term objectives. It appears that the decisions taken by dosimetrists when performing the optimization of a treatment can be formalized as a reinforcement learning problem. Moreover, dosimetrists can guide the TPS towards an acceptable plan, but they usually struggle explaining their decision while interacting with the TPS. This suggest the use of deep reinforcement learning, over expert base methods.

2 Materials and Methods

We introduce a new paradigm in reinforcement learning (RL), based on the evaluation of states, rather than the reward.

2.1 Reinforcement Learning Reward

In classical RL, we want $V(S_t) = R_t + \gamma V(S_{t+1})$ (so the update is $V(S_t) \leftarrow (1-\alpha)V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1})]$). In the context of dose optimization, the reward R_t is defined as $R_t = \mathscr{E}(S_{t+1}) - \mathscr{E}(S_t)$. Where $\mathscr{E}$ is a function that evaluates the quality of a state (such that higher is better; if lower is better, then swap s_t and S_{t+1}).

The evaluation $\mathscr{E}$ can be one, or a mixture of the metrics mentions in introduction (Section 1) [10] [11] [8]. This setup may leverage knowledge about which actions to perform, instead of guessing randomly as a meta optimizer would do. We can hope to gain some computation time.
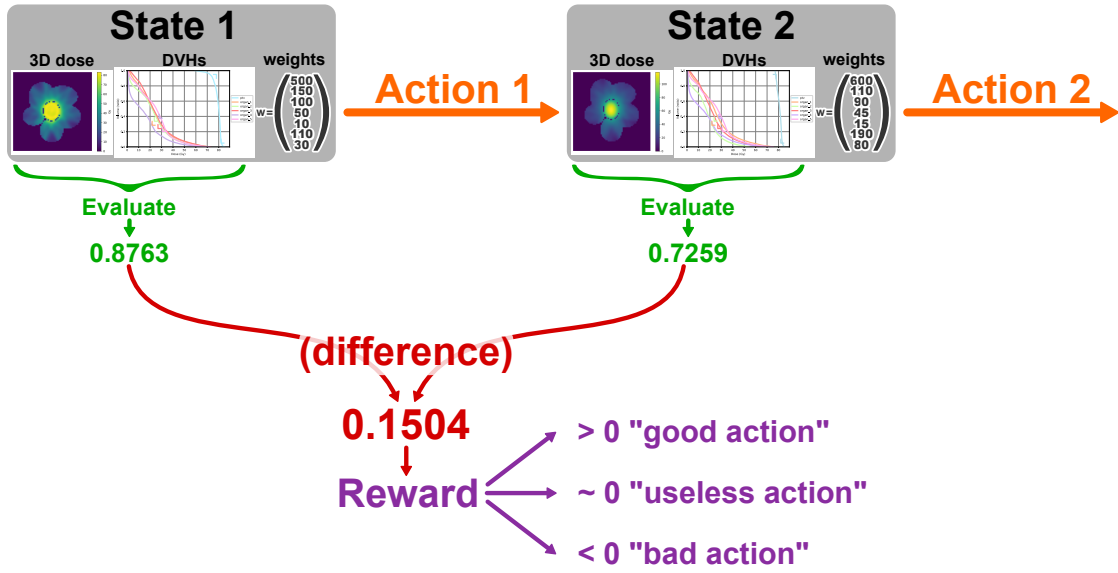
Figure 1

However, this technique is not using the plan used in past cases; it only needs the optimizer inputs (CT, structures contours). We propose to use the availability of past treatment plans, to better catch the complexity of decision taken by dosimetrists, and match better their expectations of a fully automatic treatment planning system.

As developed in previous work, we can derive a distance between doses plans [add citation]. If we consider the clinical dose of past cases (used for training) as the best achievable one, then we can evaluate a dose plan by computing its distance with the clinical dose plan.

Letting D_t be the dose associated with S_t, and D_C the clinical dose. We then define $\mathscr{E}(S_t) = \mathscr{D}(D_t, D_C)$. Since in that case, lower is better, we will define the reward as

$$R_t = \mathscr{E}(S_t) - \mathscr{E}(S_{t+1}) = \mathscr{D}(D_t, D_C) - \mathscr{D}(D_{t+1}, D_C).$$

This reward can be interpreted as the "distance gained to the clinical dose".

2.2   Reward-Free Reinforcement Learning

Since the reward is based on an evaluation of the state, one may consider dropping all the reward machinery, and directly use the state evaluation. This allows us to better capture the signal: it is not embedded in a reward system, and the network can therefore learn faster.

The reason why reward system is used in many games (such as go or chess) is that state evaluation is extremely difficult, if not impossible; there is no function that, given the board, gives the quality of white/black's position. Reward is only given at the end of a game, usually $+1$ for winning, -1 for losing. The states evaluation are deduced by the reinforcement learning agent while learning to maximize the reward. However, in our dosimetrist case, we exactly have such an evaluation function. Embedding this evaluation in a reward,

by defining the reward as the evaluation difference turns out to dilute the signal, and the agent learns slower.

By predicting the next state evaluation instead of using the Bellman optimality equation, we lose the ability to make short term concessions in order to obtain reward later on. In our case, making short term concessions would mean that we trigger the weights in a way that the optimizer performs worse, but such that triggering a little more makes the optimizer perform better.

Because optimizers have the bad habit od getting stuck in local minimum, we can not use last optimization result as a "warm start" after modifying optimization weights. Hence, restarting optimization from zero is forced after every weight modification.

Given the statement of the optimization problem, and the fact that we can not use warm start, making short term compromises does not make much sense. It is therefore safe to predict next state evaluation directly, without the use of the Bellman optimality equation.

3   Results

4   Discussion

Appendix

Synthetic phantom patients

As this is very new and ongoing research, we generated synthetic phantom patients and associated trust-able clinical dose. In future work, we hope to be able to apply this technique to real cases.

Clinical dose

Optimization

Evaluation

References

[1] Treatment Planning System Basics | Oncology Medical Physics.

[2] Ruth Brooks. What is reinforcement learning?, December 2021.

[3] Indra J Das, Julia J Compton, Amishi Bajaj, and Peter A Johnstone. Intra- and inter-physician variability in target volume delineation in radiation therapy. *Journal of Radiation Research*, page rrab080, September 2021.

[4] James M. Galvin, Alfred R. Smith, and Brian Lally. Characterization of a multileaf collimator system. *International Journal of Radiation Oncology*Biology*Physics*, 25(2):181–192, January 1993.

[5] Charles Huang, Yong Yang, Neil Panjwani, Stephen Boyd, and Lei Xing. Pareto Optimal Projection Search (POPS): Automated Radiation Therapy Treatment Planning by Direct Search of the Pareto Surface. *IEEE Transactions on Biomedical Engineering*, 68(10):2907–2917, October 2021.

[6] Xinyi Li, Yaorong Ge, Qiuwen Wu, Chunhao Wang, Yang Sheng, Wentao Wang, Hunter Stephens, Fang-Fang Yin, and Q. Jackie Wu. Input feature design and its impact on the performance of deep learning models for predicting fluence maps in intensity-modulated radiation therapy. *Physics in Medicine & Biology*, 67(21):215009, November 2022.

[7] John T. Lyman. Normal tissue complication probabilities: Variable dose per fraction. *International Journal of Radiation Oncology*Biology*Physics*, 22(2):247–250, January 1992.

[8] Grégoire Moreau, Vincent François-Lavet, Paul Desbordes, and Benoît Macq. Reinforcement Learning for Radiotherapy Dose Fractioning Automation. *Biomedicines*, 9(2):214, February 2021.

[9] U Oelfke and T Bortfeld. Inverse planning for photon and proton beams. *Medical Dosimetry*, 26(2):113–124, June 2001.

[10] Chenyang Shen, Liyuan Chen, and Xun Jia. A hierarchical deep reinforcement learning framework for intelligent automatic treatment planning of prostate cancer intensity modulated radiation therapy. *Physics in Medicine & Biology*, 66(13):134002, July 2021.

[11] Chenyang Shen, Yesenia Gonzalez, Peter Klages, Nan Qin, Hyunuk Jung, Liyuan Chen, Dan Nguyen, Steve B. Jiang, and Xun Jia. Intelligent Inverse Treatment Planning via Deep Reinforcement Learning, a Proof-of-Principle Study in High Dose-rate Brachytherapy for Cervical Cancer. *Physics in Medicine & Biology*, 64(11):115013, May 2019. arXiv:1811.10102 [physics].

[12] Vincenzo Valentini, Francesco Cellini, Bruce D. Minsky, Gian Carlo Mattiucci, Mario Balducci, Giuseppe D'Agostino, Elisa D'Angelo, Nicola Dinapoli, Nicola Nicolotti, Chiara Valentini, and Giuseppe La Torre. Survival after radiotherapy in gastric cancer: Systematic review and meta-analysis. *Radiotherapy and Oncology*, 92(2):176–183, August 2009.

[13] S Webb. The physical basis of IMRT and inverse planning. *The British Journal of Radiology*, 76(910):678–689, October 2003.

[14] Xingen Wu and Yunping Zhu. An optimization method for importance factors and beam weights based on genetic algorithms for radiotherapy treatment planning. *Physics in Medicine and Biology*, 46(4):1085–1099, April 2001.

[15] L Xing, J G Li, S Donaldson, Q T Le, and A L Boyer. Optimization of importance factors in inverse planning. *Physics in Medicine and Biology*, 44(10):2525–2536, October 1999.

[16] Dandan Xu, Guowen Li, Hongfei Li, and Fei Jia. Comparison of IMRT versus 3D-CRT in the treatment of esophagus cancer: A systematic review and meta-analysis. *Medicine*, 96(31):e7685, August 2017.