Radiotherapy Dose Optimization via Clinical Knowledge Based Reinforcement Learning

Paul Dubois1,2[0009−0003−3856−8048], Paul-Henry Cournède1[0000−0001−7679−6197], Nikos Paragios2[0000−0002−9668−4763], and Pascal Fenoglietto3

1 Biomathematics, MICS, CentraleSupélec, Université Paris-Saclay, Paris, France

{p.dubois,paul-henry.cournede}@centralesupelec.fr

2 TheraPanacea, Paris, France

{p.dubois,n.paragios}@therapanacea.eu

3 Institut du Cancer de Montpellier (Val d’Aurelle), Montpellier, France

{paul.dubois,pascal.fenoglietto}@icm.unicancer.fr

**Abstract.** A radiation therapy plan finds an equilibrium between goals with no universal prioritization. The delicate balance between multiple objectives is typically 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 the preferences of radiation oncologists and medical physicists shape treatment planning. To surmount these challenges, we propose a first step towards a fully automated approach, using an innovative deep-learning framework. Using a clinically meaningful distance between doses, we trained a reinforcement learning agent to mimic a set of plans. This method allows automatic navigation toward acceptable solutions via the exploitation of optimal dose distributions found by human planners on previously treated patients. As this is ongoing research, we generated synthetic phantom patients and associated realistic clinical doses. Our deep learning agent successfully learned correct actions leading to treatment plans similar to past cases ones. The incapacity to reproduce human-like dose plans hinders adopting a fully automated treatment planning system; this method could start paving the way towards human-less treatment planning system technologies. In future work, we hope to be able to apply this technique to real cases.

**Keywords:** Radiotherapy, Dose Optimization, Reinforcement Learning, Deep Learning.

1. Introduction

In contemporary radiation therapy, photon intensity modulated radiation therapy (IMRT) is a pivotal technique to attain precise and conformal dose distributions within target volumes [18]. This achievement owes its realization to the advent of the multileaf collimator (MLC) [5]. Radiation therapy is now a reliable treatment for oncology [14]. Despite this consensus, the way to deliver radiotherapy for its best result remains very dependent upon doctors. Moreover, there appears to be a large variability across physicians and centers, but in terms of 3D structures contouring and irradiation, constrains priorities [3].

To achieve the best treatment, doctors must solve a complex inverse mathematical optimization problem with multiple trade-offs [10] [15]. However, a lack of standardized prioritization of constraints makes the optimization a real challenge. The standard procedure nowadays is to guide computer optimization manually: dosimetrists manually update the settings of an optimizing software so-called Treatment Planning System (TPS) [1].

There have been many tries to create a metric that quantifies the quality of a treatment plan, such as Normal Tissue Complication Probabilities (NTCP), target coverage, conformity index, and heterogeneity index, among others/to name a few [8] [7]. However, they have yet to satisfy all radio-oncologists, and the only reliable way to assess a doctor’s plan is to evaluate the dose-volume histograms (DVHs) themselves.

As a result, Pareto surface exploration is unsuitable due to the lack of impartial quantitative measurement for a particular plan [6]. Other meta-optimization techniques are similarly bounded for the same reason [16] [17]. 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. The context-aware acceptability criteria make the acceptability of a plan hard to define in general.

Reinforcement learning (RL) is a machine learning paradigm that trains agents to make sequential decisions in dynamic environments [2]. Agents learn to optimize their actions to achieve long-term objectives through trial and error guided by rewards or penalties. The decisions taken by dosimetrists when optimizing treatment can be formalized as an RL problem. Moreover, dosimetrists can guide the TPS towards an acceptable plan but usually struggle to explain their decision while interacting with the TPS. The difficulty in explaining why certain decisions are taken suggests using deep RL over expert-based methods. This setup is similar to image recognition, where one can say a picture represents a car or a boat but struggles to explain why.

The study’s primary hypothesis is that all the information needed to decide what weights should be changed in the objective function used by the optimizer relies on the Dose Volume Histograms (DVHs). Our hypothesis is supported by the fact that dosimetrists almost solely use DVHs plots. In order to learn the actions of dosimetrists who use a TPS to optimize doses, we leverage deep learning. This is done by training an agent that takes the DVHs as the input of the current optimized dose, and predicts the evaluation of possible weights changes.

Typically, access to the exact actions taken by human dosimetrists on the TPS is unavailable (as clinics do not usually store this data; only the final plan is held). Therefore, we only use the dose distributions of previously treated patients to train our model. This partial availability of data suggests the use of RL.

1. Materials and Methods

We introduce a new paradigm for reward-based dosimetrist RL agents. This new reward system aims to improve how human-optimized doses are mimicked.

* 1. Reinforcement Learning Reward

In classical RL, we want (so the update is ). In the context of dose optimization, the reward is defined as , where is a function that evaluates the quality of a state (such that higher is better; if lower is better, then swap and ).

The evaluation can be one or a mixture of the metrics mentioned in the introduction (Section 1) [12] [13] [9]. This setup may leverage knowledge about which actions to perform instead of guessing randomly, as a meta-optimizer would do. This could potentially gain some computation time.

However, this technique does not use past plans; it only needs the optimizer inputs (CT, structures contours). We propose using the availability of past treatment plans to more accurately reflect the complexity of decisions made by dosimetrists and better match their expectations of a fully automatic treatment planning system.

As developed in previous work, we can derive a distance between dose plans [11]. If we consider the clinical dose of past cases (used for training) as the best achievable one, we can evaluate a dose plan by computing its distance from the clinical dose plan.

Let be the dose associated with , and the clinical dose. We then define . Since, in that case, should be minimized, we will define the reward as

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

* 1. Architecture

We generated a training set of over 125k actions (this took five days on an NVIDIA GeForce GTX 1080). Despite this relatively large dataset, we have not explored exhaustively the state-actions space, and the network still lands off distribution. This can easily be spotted when the predicted value is greater than the current distance to the clinical dose; we choose to ignore those predictions, and in fact all outlier predictions. The justification is that our set of actions is limited, no action will suddenly drastically improve the plan. It is the combination of several sequential actions that allows good plan optimization. Therefore, while testing, we choose the action with the best prediction, while passing the outlier test just mentioned.

* 1. Avoiding Off-Distribution

Please note that the first paragraph of a section or subsection is not indented. The first paragraphs that follows a table, figure, equation etc. does not have an indent, either.

Subsequent paragraphs, however, are indented.

### Sample Heading (Third Level). Only two levels of headings should be numbered. Lower level headings remain unnumbered; they are formatted as run-in headings.

#### Sample Heading (Forth Level). The contribution should contain no more than four levels of headings. The following Table 1 gives a summary of all heading levels.

1. Results

Figure 3 shows how the distance between our RL agents performs over five steps on 30 test patients (unseen during the training). A lower distance is interpreted as an improved dose, since it is closer to the best dose, which is the clinical one.

* 1. Quantitative Results

The network converged on the training data, and validation showed minor overfitting. For testing, we generated 30 brand new cases that we again manually optimized. We then used the RL model to perform the optimization of these 30 unseen cases. On average, our model was able to reduce the dose distance with manually optimized dose by a factor of 3 (from 1.8 at iteration 0 to 0.6 at iteration 4), as shown in Table 3.1. We remark from the Table 3.1 that the homogeneity score and conformity score give similar results. Classical meta-optimization performs well, but needs a metric to elect the best dose (during the test, the clinical dose is unknown, so the DVHs distance metric is not available). We also observe that clinical doses are not always scoring high (in this test set, a high conformity, but low homogeneity compared to automatic techniques). This show the difficulty to create a metric that capture all the complexity of a clinically acceptable dose.

* 1. Qualitative Results

Figure 4 shows the DVHs at each of the first four optimization steps on one of the test patients, unseen by the agent during the training. Our model drastically reduced the dose distance with manually optimized doses. Visual inspection of the DVHs plot shows that the dose optimized by the RL agent is very close to the clinical (manually fine-tuned) one.

1. Discussion
2. Conclusion

Appendix

### Synthetic phantom patients

### Clinical dose

### Optimization

References

**Table 1.** Table captions should be placed above the tables.

|  |  |  |
| --- | --- | --- |
| Heading level | Example | Font size and style |
| Title (centered) | **Lecture Notes** | 14 point, bold |
| 1st-level heading | **1 Introduction** | 12 point, bold |
| 2nd-level heading | **2.1 Printing Area** | 10 point, bold |
| 3rd-level heading | **Run-in Heading in Bold.** Text follows | 10 point, bold |
| 4th-level heading | *Lowest Level Heading.* Text follows | 10 point, italic |

Displayed equations are centered and set on a separate line.

*x* + *y* = *z* ()

Please try to avoid rasterized images for line-art diagrams and schemas. Whenever possible, use vector graphics instead (see Fig. 1).

**Fig. 1.** A figure caption is always placed below the illustration. Short captions are centered, while long ones are justified. The macro button chooses the correct format automatically.

For citations of references, we prefer the use of square brackets and consecutive numbers. Citations using labels or the author/year convention are also acceptable. The following bibliography provides a sample reference list with entries for journal articles [1], an LNCS chapter [2], a book [3], proceedings without editors [4], as well as a URL [5].

References

1. Author, F.: Article title. Journal 2(5), 99–110 (2016).
2. Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016).
3. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999).
4. Author, F.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010).
5. LNCS Homepage, <http://www.springer.com/lncs>, last accessed 2016/11/21.