REINFORCEMENT LEARNING: BEYOND OPTIMALITY

Contents

1. Research Statement

Reinforcement learning (RL) is the standard model for learning some desired target behavior in an unknown environment. The learning agent can observe her current state and can choose among different actions that give some reward and lead to new states. In standard RL rewards are quantifyable and the learner's goal is to maximize her total reward over time.

In recent years, applied RL can refer to several success stories like Google's Alpha-Go [10], various applications in robotics [4], or autonomous driving. However, in spite of this success the respective RL theory seems to be still in its infancy. That is, RL applications typically rely on tailor-made algorithms and use problem specific customized representations, while there are no guarantees with respect to the reliability or the safety of the employed methods. We claim that one of the reasons for this large gap between RL practice and RL theory is the *optimization paradigm* that is at the core of current RL theory:

The model usually considered in RL theory is that of *optimization* in the framework of Markov decision processes (MDPs). The latter have been developed in the 1950s and examined in the field of operations research. Optimization in MDPs has yielded a rich theory that in turn has been adapted in RL theory. While it is obviously convenient to have access to that theory for the purpose of application in RL, we think that the paradigm of optimization is one of the reasons why current RL theory is incapacle of keeping pace with RL applications.

That is, most RL applications deal with successfully learning to perform a certain task. While there are obviously better and worse ways to perform a task (like picking up an object, or driving from A to B) there is usually not an *optimal* way. In further consequence the focus on optimization does not only lead to models that are inadequate for the respective applications (and hence have to be adpated by experts by hand). Moreover, learning to do things in an optimal way is also much more costly. Thus requiring to drive to work *as fast as possible* or picking up an object using the *least* possible amount of energy is obviously much more demanding than the same tasks without the optimization component. Accordingly, the computation of such optimal behavior is not only disproportionate with respect to a given task, it is usually also computationally intractable for most RL applications.

The goal of this project is to develop RL theory that replaces the optimality paradigm with an alternative approach based on *satisficing*. That is, the goal is to find a *sufficient* strategy instead of an *optimal* one. This alternative approach not only makes learning obviously easier, it also better fits most RL applications. Thus, if successful our research project will finally develop RL theory that is more useful and applicable to practical RL problems.

2. Research Approach

The paradigm of optimization and its downsides

The standard model for formalizing reinforcement learning (RL) problems are *Markov decision processes* (MDPs) [8]. These have originally been developed by Richard Bellman in the 1950s and have been used since the 1980s in *artificial intelligence* (AI) research to model environments that are unknown to the learning agent. As MDPs had been introduced to describe optimization problems, the focus of MDP theory (a branch in the field of *operations research*) has always been on the respective questions of conditions for the existence of optimal policies and algorithms that are able to compute them. Obviously, it has been quite natural for research in AI (and in consequence in particular for RL as well) to adopt this focus and concentrate on learning *optimal* policies in an unknown environment.

In spite of the fact that most RL applications (like in robotics or autonomous driving) are task-oriented, the emphasis of RL theory has remained on the the question how to learn an optimal policy. While there are usually some constraints on the way the task shall be performed (e.g. with respect to safety, reliability, or efficiency) there is usually not an *optimal* way, at least not in a way that one would consider relevant for fulfilment of a task. Accordingly, application of RL theory to respective practical problems is not straightforward, leading to branches like *inverse RL* [7] where the aim is (put in a very simple way) to find a representation of the problem so that the optimal solution corresponds to the behavior one aims to simulate.

Another obvious and crucial downside of the focus on optimization is that in theory problems become much harder than in practice. The additional optimization criterion thus not only leads to models that are inadequate for many applications, it is obviously also much more involved and costly to determine an *optimal* solution than a *sufficient* one. Accordingly, the gap between RL theory and RL practice has been growing in recent years. On the one hand, RL applicationers manage to engineer algorithms capable of beating world class players in difficult strategy games (like Go and chess). However, these algorithms are tailor-made solutions for each problem at hand, have no theoretical guarantees with respect to reliability or robustness, and do not generalize to other problem domains. On the other hand, theory provides algorithms that provably learn optimal behavior but are not easily applicable to real-world scenarios, also because they are usually computationally much too costly to implement for an actual problem.

Therefore, we suggest to drop the paradigm of optimization and develop theory for *satisficing* in MDPs. Based on this, new RL theory shall be developed that is on the one hand more directly

applicable to actual RL applications and moreover offers solutions that are efficiently computable and hence ready to use. To illustrate the principled problem of the optimization paradigm and the potential of the satisficing approach note that when aiming at *optimal* policies the whole state space has to be explored. Consequently, the necessary computation time as well as time till convergence to an optimal policy must depend on the size of the state space (assuming the tabular case with finite state space for the moment). However, the state space of typical RL applications is huge so that it is hopeless to directly employ methods analysed by RL theory. On the other hand, under a satisficing criterion it may be sufficient to explore only a small part of the state space until determining a suitable policy that successfully solves the set task.

Satisficing in MDPs and RL

While we think that optimization is a disputable paradigm for applications in RL, we are convinced that the MDP framework is still a suitable model for representing RL problems. Thus, in the following we will consider the standard MDP framework. That is, an MDP $M = (S, A, p, r)$ is defined by a given set of states S, a set of actions A, transition probabilities $p(s'|s, a)$ that define the probability of a transition from some state s into some other state s' when selecting a particular action a, and a reward function $r(s, a)$ that defines the mean reward when choosing action a in state s.

Formally, the most natural way to represent a satisficing framework for MDPs is to consider a standard MDP and instead of aiming to find a *policy* selecting actions that maximizes the rewards[1] the goal is to find a policy satisfying some more modest criterion with respect to reward. The most straightforward choice is obviously the satisfaction of some minimum reward we want to achieve. This is a setting commonly found in problems of operations research and game theory, however we are not aware of respective theory developed for the MDP setting. Obvious questions to be addressed are how for a given and known MDP one can efficiently determine whether a respective policy exists and how to compute such a policy. We note that in a next step one could think of more complex criteria like policies obtaining a minimum reward in a certain amount of time, or adding further constraints, something that is obviously important in applications. In any case, an important aspect of this part of the project will be to determine which parameters of an MDP are relevant for the computational complexity of the developed algorithms. We expect that some new structural parameters of the MDP may turn out to be crucial with respect to the satisficing criterion, which will also be an interesting addition to the operations research literature.

[1]For the specification of the precise reward criterion see the next section.

After development of some basic theory for satisficing in MDPs, the second step will be application of the findings to RL. This part can adapt already known approaches, but we expect that much simpler methods can be successfully applied in the satisficing setting (cf. next section). We note that for RL in the much simpler multi-armed bandit setting satisficing is not without predecessor [9], however not with a focus on development of theory for applications. (Unlike for general RL, computation is not an issue for bandit problems.) Another related direction is that of multi-criterion RL for which however there isn't any fully developed theory available except again for the simple bandit setting [2].

Odds and ends

Reinforcement learning is a complex field. While we expect that our project will contribute to a "*satisficing turn*" in the discipline, it would of course be naive to believe that this in itself will be sufficient to solve all problems in RL at one go. Rather, we think that this would be a first step to develop an alternative RL theory that has the potential to close the gap to RL applications. For this, research in other branches of RL has to be continued and adapted to the setting of satisficing. This includes topics like the choice and learning of suitable representations [5] and the implementation of previous knowledge into RL algorithms to mention just two subjects that we consider to be of particular importance.

On the other hand, RL approaches that have not been able to prove themselves to be essential so far could get boosted with the new underlying paradigm. One example that comes to mind is that of *relational* RL (research on which recently also has been continued and generalized under the term of *object-oriented* RL). While this approach is indeed very natural it has turned out that —intuitively speaking— it does not fit well with the optimization paradigm [11].

While the goal is to finally develop some RL theory that can be applied to practical problems the project does not aim at actually implementing the developed theory to solve some actual real-world problems. That would be the task of a successor project. However, it would be suitable to start with implementations of developed algorithms for smaller toy examples as they are regularly used in the RL literature. This could be done e.g. within a Master thesis.

Ethical and Gender-Related Aspects

This is a theoretical computer science project, so there are neither any ethical, safety-related, nor regulatory issues to be considered. In case some algorithms are tested, this will be done on artificial sample domains that do not involve any data privacy issues. For the same reason, the project and its implementation do not involve any sex-specific or gender-related issues.

3. Project Implementation

Phase I: Theory for Satisficing in MDPs

The implementation of the project will be done along the outline given in the previous section. That is, in the first part of the project, we will concentrate on the development of theory for satisficing in MDPs. More precisely, we will start with the already mentioned questions of (i) how in a given and known MDP one can determine whether a respective policy exists that guarantees some minimum reward and (ii) how to efficiently compute such a policy. As reward criterion, we will consider *average* as well as *total T-step* reward, as we think that these are more relevant for RL applications than *discounted* and *episodic* settings as currently often considered in RL theory.

Depending on findings for this base case, we consider to extend our research to alternative settings. This may either be necessary if the base case cannot be solved as efficiently as necessary for successful implementation into the RL setting. However, even if the base case turns out to be simple, it will pay off to consider alternative or more general scenarios that are relevant for RL applications. For example, adding further constraints on the target policy is a natural extension that we will have a look at. We note that in case these constraints can be expressed via state-visits there is already some theory developed under the optimization criterion [1], that could be the base for an adaptation to the satisficing setting. However, different constraints (e.g. with respect to reliability) will be considered as well.

Phase II: Development of RL Theory

In the second part of the project, we shall implement the results and algorithms developed for the case when a known MDP is given to the case where the MDP has to be explored by a learning agent whose aim is to learn successful behavior. The latter is now not determined by some optimization but by a satisfaction criterion.

The application of theory developed in Phase I to find respective RL algorithms and theory can in principle be done along the guidelines of current RL theory. That is, the various principled approaches to estimate (or approximate) MDP parameters (e.g., either directly rewards and transition probabilities, or value functions of the respective policy) can be simply extended to our approach of satisficing. However, while this is a natural first-step, there is no need to stick to traditional RL approaches, as we expect that our findings will allow the development of new methods that are more efficient and easier applicable than traditional approaches in RL. Note in particular that the heavily cited *exploration-exploitation dilemma* every optimizing RL algorithm

has to deal with is by no means as significant in a satisficing setting. Once a satisficing policy has been found (handling the respective estimation error e.g. by the application of confidence intervals) there is no need of further expensive exploration and exploitation will be good enough. Accordingly, we expect to be able to show that much simpler algorithms will be able to succeed in a satisficing RL setting. Still, even if it is easier to separate exploration from exploitation, exploration remains an important topic [6].

In any case, it will be not only necessary to think about algorithms and analysis of their properties, the project may also have to consider new measures for the performance of an algorithm in the satisficing RL setting. For example, it is apriori not clear whether the popular *regret* criterion [3] that considers the loss with respect to an optimal policy in the worst case can be sensibly adapted to the setting of satisficing RL. Other measures like sample complexity bounds that bound the number of time steps till a respective policy is found seem to be easier to apply, but in any case it will be necessary to consider suitable criteria. Of course, efficient computability will be an essential property of all developed algorithms.

Time schedule

Obviously, we will begin with research in Phase I, while the start of Phase II will depend on the progress in Phase I. In an optimistic scenario we should be able to come up with a first basic theory for satisficing in known MDPs within a few months when we could already start with work on Phase II. In general we expect that there will be an ongoing interplay between Phases I and II. For each piece of theory developed for the MDP satisficing setting in Phase I, we can immediately start checking how it can be used for RL in Phase II. Theory that can be successfully implemented in Phase II will consequently be extended and generalized in promising directions back in Phase I.

In the worst case scenario (cf. also next section on risk assessment) it may turn out to be difficult to develop a useful theory on satisficing in MDPs in Phase I. In that case we will try to at least develop some theory in a constrained setting (like certain classes of MDPs with target policies having certain properties making things simpler) within the first ten months, so that we can start with Phase II at least in this constrained setting. As mentioned before, results in Phase II may then help with identifying directions to be followed in further research in Phase I.

A set of small experiments in toy domains can be implemented as soon as we have developed suitable algorithms in Phase II. As experiments are only of complementary character, there is no fixed schedule for this optional part of the project.

4. Risk Assessment and Learning Potential

Risks in Phase I

The biggest overall risk we see for our project is that satisficing for MDPs turns out to be a problem that has basically the same complexity as finding an optimal policy. We think that this risk is definitely not negligible, as it is obvious that e.g. just the existence of a suitable policy that satisfies some minimum reward criterion will depend on the particular structure of the MDP. Thus the most difficult scenario is when the single policy fulfilling the criterion will be identical to the optimal policy. Therefore, it will be crucial to identify the respective relevant parameters of the MDP that hopefully allow a quick assessment of the hardness of the respective MDP.

In the worst case it may turn out that the complexity of satisficing entails that there are no computational savings to be earned with respect to the case of optimality. While this is definitely an interesting result in itself, we consider two resorts to be able to still make further progress beyond Phase I. The first one is to consider alternative criteria that may be useful in the RL context. A weaker criterion can be e.g. obtained if one does not fix a minimum reward one wants to achieve but instead formulates a policy-relative goal of finding a policy that is among the top $x\%$ of all policies. This approach may be able to avoid that there are too many relevant parameters of the MDP to be determined and would allow a (more modest) continuation of the project. A second approach to make the development of theory easier would be to consider special cases, that is, MDPs and/or target policies with certain properties. This will reduce the number of relevant parameters and may be also a good starting point for an extension to more general cases.

Risks in Phase II

Once the problems of Phase I are solved, we think that development of respective RL theory will involve less risk. At least adapting known approaches to the satisficing criterion in principle should be straightforward. More critical is the question about genuinely new approaches that will result from the findings in Phase I, as these will depend on the concrete results that we are able to obtain in Phase I. It may well be that the latter do not allow for the general development of transformationally new methods for RL that we aim at. In that case we would try to use additional natural assumptions on the learning setting (i.e., the underlying MDP and the target policy) that allow to come up with methods that are simple and successful. We think that also constrained methods will be a valuable addition to the field of RL that has some potential in making RL theory more accessible for RL applications.

References

[1] Eitan Altman. *Constrained Markov Decision Processes*. Chapman & Hall, 1999.

[2] Peter Auer, Chao-Kai Chiang, Ronald Ortner, and Madalina M. Drugan. Pareto front identification from stochastic bandit feedback. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016*, volume 51 of *JMLR Workshop and Conference Proceedings*, pages 939–947. JMLR.org, 2016. URL http://proceedings.mlr.press/v51/auer16.html.

[3] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, 2010. URL https://www.jmlr.org/papers/volume11/jaksch10a/jaksch10a.pdf.

[4] Jens Kober and Jan Peters. Reinforcement learning in robotics: A survey. In Marco A. Wiering and Martijn van Otterlo, editors, *Reinforcement Learning*, volume 12 of *Adaptation, Learning, and Optimization*, pages 579–610. Springer, 2012. URL https://doi.org/10.1007/978-3-642-27645-3_18.

[5] Timothée Lesort, Natalia Díaz Rodríguez, Jean-François Goudou, and David Filliat. State representation learning for control: An overview. *Neural Networks*, 108:379–392, 2018. URL https://doi.org/10.1016/j.neunet.2018.07.006.

[6] Shiau Hong Lim and Peter Auer. Autonomous exploration for navigating in MDPs. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *COLT 2012 - The 25th Annual Conference on Learning Theory*, volume 23 of *JMLR Proceedings*, pages 40.1–40.24. JMLR.org, 2012. URL http://proceedings.mlr.press/v23/lim12/lim12.pdf.

[7] Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In Pat Langley, editor, *Proceedings of the Seventeenth International Conference on Machine Learning, ICML 2000*, pages 663–670. Morgan Kaufmann, 2000.

[8] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1994.

[9] Paul Reverdy, Vaibhav Srivastava, and Naomi Ehrich Leonard. Satisficing in multi-armed bandit problems. *IEEE Trans. Autom. Control.*, 62(8):3788–3803, 2017. URL https://doi.org/10.1109/TAC.2016.2644380.

[10] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):

484–489, 2016. URL https://doi.org/10.1038/nature16961.

[11] Martijn van Otterlo. Solving relational and first-order logical Markov decision processes: A survey. In Marco A. Wiering and Martijn van Otterlo, editors, *Reinforcement Learning*, volume 12 of *Adaptation, Learning, and Optimization*, pages 253–292. Springer, 2012. URL https://doi.org/10.1007/978-3-642-27645-3_8.