COMP2550 Assignment 5: Research Project

Xin Lu (u7165475)

1 Research Project (Q1)

1.1 Research project title:

Study of Blind Value and Continuous RAVE exploration parameters' impacts on continuous MCTS performance.

1.2 Research area:

Continuous MCTS, exploration of MCTS in continuous action and state spaces.

1.3 Proposed ANU supervisor:

- Associate Professor Hanna Kurniawati
- Dr Johannes Hendriks
- Associate Professor Kee Siong Ng

2 What are you trying to do? (Q2)

I will study and analyse the performance differences of Blind Value (BV) (Couetoux et al., 2012), and Continuous Rapid Action Value Estimation (cRAVE) (Gelly et al., 2007), with respect to their usage of explored graph information, algorithm complexity, and impacts of their exploration parameters on continuous MCTS performance, through mathematical reasoning and quantitatively analysing algorithm performance differences.

Next I will propose a combined exploration method for continuous MCTS, based on the previous analysis, aimed to blend the advantages of two so as to achieve a better balance between computational cost and agent performance.

This project is motivated by the fact that there is still non-sufficient understanding of MCTS search dynamics, especially regarding parameters' impacts on MCTS's performance (Browne et al., 2012). Also, continuous MCTS is of current research interest, whose action and state space are infinite in order to align with the complexity of more realistic problems. Hence a fair comparison

between existing exploration methods will offer insights into MCTS search dynamics, and reveal directions for possible improvements in continuous MCTS designs.

3 Current state of art and brief summaries of 3 publications. (Q3)

For exploring in the infinite action and state spaces of continuous MCTS problems, there are two existing methods designed to make a more informed selection of action to explore: **(1) Blind Value (BV)** (Couetoux et al., 2012), and **(2) Continuous Rapid Action Value Estimation (cRAVE)**, extended to infinite action space by (Couetoux et al., 2012), whose original discrete version was proposed in (Gelly et al., 2007).

Couetoux, A., Doghmen, H. and Teytaud, O., 2012, January. Improving the exploration in upper confidence trees. In *International Conference on Learning and Intelligent Optimization* (pp. 366-371). Springer, Berlin, Heidelberg.

BV uses only information in child nodes of the current node, obtained from previous simulations, to select an unbiased action that balances the exploitation and exploration. It chooses an action to explore by considering both its Upper Confidence Bound (UCB) value and its euclidean distance from the explored graph. So unexplored areas will be prioritised at the early stage since there is not enough information of its UCBs, and promising areas will be focused in the later stage of MCTS when enough explorations have been done. BV's formula is of simple form and it is easy to implement. Also, its horizontal propagation structure makes continuous MCTS converges significantly faster than randomly sampling.

Gelly, S. and Silver, D., 2007, June. Combining online and offline knowledge in UCT. In *Proceedings of the 24th international conference on Machine learning* (pp. 273-280).

On the other hand, RAVE is used in a reinforcement learning MCTS setting to select a biassed action to explore. It uses action-state value pairs learned onlinely, and averages returns from all explored tree walks to output a biassed action to explore. Since the reward value of an action is closely related to its previous state in an action-state value pair, the final output action will very likely be biassed by some specific rollouts.

Both methods achieved positive results in their proposed tasks, in which RAVE was implemented as part of MoGo for playing Go. However, estimated action values of RAVE may be heavily biassed, and BV is less sufficiently studied, whose parameters leave spaces to incorporate more information from the explored tree, possibly cRAVE.

Yee, T., Lisý, V. and Bowling, M.H., 2016, January. Monte Carlo Tree Search in Continuous Action Spaces with Execution Uncertainty. In *IJCAI* (pp. 690-697).

(Yee et al., 2016) proposed a non-parametric method Kernel Regression (KR) for estimating a node's average rewards and number of visits, and then replacing the corresponding terms in the classic UCB formula to output an estimated action value for selecting the action to explore, based on information from the entire continuous action and state space. Its approach allows information sharing between considered actions with the same possible outcomes, since those actions will share the same kernel data points.

To summarise, since the above three methods all use information from the explored tree to guide the exploration stage for continuous MCTS problems, it is of great interest to analyse and compare performance differences of all to see how exploration affects continuous MCTS search dynamics.

4 What is new in your approach, how does it extend the state of the art? Why do you think it will work? (Q4)

4.1 What is new in your approach, and how does it extend the state of the art:

I will use mathematical reasoning and quantitative analysis of performance differences to show impacts of BV's and cRAVE's exploration parameters on continuous MCTS performance. Although reasoning of RAVE was mentioned in (Gelly et al., 2007), there is no such analysis for performance of Continuous RAVE (cRAVE) and BV, or for agent performance differences between two methods. The resulting analysis and conclusions are expected to offer insights of exploration parameters' impacts on MCTS search dynamics, which is still a non-sufficiently studied problem in the research field of MCTS.

For the second part of my project proposal, an exploration method combining advantages of two methods will be proposed, and it is expected to improve continuous MCTS performance given a limited computational budget, since BV is fast to converge, and cRAVE will use more information from the explored graph.

4.2 Why do you think it will work?

MCTS is well defined in its mathematical notation, and reasonings about UCB and RAVE have been established in previous works of MCTS. Thus analysing the exploration parameter behaviour of cRAVE is pushing one step forward from the previous research, and is feasible within one semester scope. Also, since BV formula uses a simple argmin expression to balance exploitation and exploration in a horizontal propagation manner, and both cRAVE and BV do not necessarily involve using domain knowledge to prune actions or states, the action and state space of combined method is expected to be convex, where nice mathematical properties and approximations can be established.

Additionally, since BV and cRAVE are both applied in the exploration stage of continuous MCTS, and they use different subsets of information from the explored graph to estimate the best action to explore, the combination of two methods are reasonable, and also interesting, bringing the expectation of combined method to be able to control over exploration biases, usage coverage of explored graph data, and algorithm converging time.

5 Why is this project interesting and important? Who can benefit if you are successful with it? (Q5)

Impact of parameters on MCTS performance is still a non-sufficiently studied question in the MCTS research community (Browne et al., 2012). Examining the mathematical behaviour of BV and cRAVE with respect to usage of graph information and algorithm complexity thus will give insights into MCTS search dynamics, especially how changes of exploration parameters will influence overall MCTS computational complexity and accuracy.

Moreover, currently there are no strong results of MCTS performance on generalised continuous problems. However, MCTS's sampling and simulation essence implies that it does not necessarily require domain-knowledge inputs for MCTS to achieve good performance, and the sequential nature of MCTS default policy makes it compatible with reinforcement learning methods, as we have seen in (Gelly et al., 2007).

These make MCTS a promising candidate for good generalizability, for which many current deep learning methods still struggle to achieve. The major difficulties facing in the construction of continuous MCTS are the infinite size of action and state spaces, and the infinite many possible transitions within state space, which makes it impossible to simulate all possible options. This is also true for large action and state space discrete MCTS, since computational budget is always limited. There are many previous works to tackle the above issues.

For example, Progressive Widening (PW), proposed in (Wang et al., 2008), addresses the issue of infinite action and state spaces by progressively increasing the size of explored states in the selection stage, and theoretically ensures PW-MCTS will converge to optimal solutions. Double Progress Widening, proposed in (Couetoux, 2013), addresses the issue of infinite many possible transitions within state space by applying PW twice in the selection stage. Also, as discussed above in Q3, there are many previous works for improving the exploration stage of continuous MCTS.

Hence, developing a method which combines strengths of the two existing exploration methods to achieve faster convergence and better performance will help to develop generalised continuous problems applicable-MCTS in the future.

6 What should you do to properly prepare for the project? (Q6)

6.1 Literature readings for mathematical proof

I will read papers that established mathematical reasonings for important MCTS formulas, which includes Upper Confidence Bound (Browne et al., 2012), Rapid Action Value Estimation (Gelly et al., 2007), and Progressive Widening (Wang et al., 2008). I will study the general proof structure of MCTS theorems, and learn the mathematical construction of VARE and cVARE. One should also prepare a proof sketch to reason about the construction of BV method, using the formula provided in (Couetoux et al., 2012).

6.2 Find benchmarks for evaluating continuous MCTS performance

Although testing tasks proposed by (Gelly et al., 2007) and (Couetoux et al., 2012) can be used as test beds to evaluate performance of different exploration methods, we cannot confidently assume that those tasks can fairly examine different algorithm's ability in solving general continuous problems. Thus a MCTS benchmark that features continuous action and state spaces is desired to be used for the evaluation stage of this project. (Yee et al., 2016) used the curling problem as its test beds, which feature a continuous physical environment. This inspires that one can consider using reinforcement learning or intuitive physics tasks to evaluare agent performances if no such continuous MCTS benchmarks are available.

6.3 Start implementing continuous MCTS using BV and cVARE

Since DPW, BV, and cVARE are established methods from previous papers, one can start to implement continuous MCTS agents using BV and cVARE with pseudo codes provided by those papers. This will ensure one can focus on the important new ideas of this project from the very beginning of next semester.

7 What are the fortnightly milestones you set yourself over the next semester in order to successfully complete the project? How will you know you succeeded? (Q7)

7.1 Fortnightly milestones

1. Construct a mathematical analysis sketch for how BV and cVARE may result in performance differences on continuous MCTS problems.
2. Finish implementing continuous MCTS agents using BV and cVARE as exploration methods separately.
3. Conduct experiments of BV and cVARE in the selected benchmark tests, and collect quantitative data.
4. Quantitatively analyse performance results to find main aspects of differences, and possible causes of such performance differences. Construct a mathematical analysis to explain the cause of such performance differences.
5. Propose an exploration method for continuous MCTS to combine the advantages of BV and cVARE, in order to converge faster and achieve better performance on the testing problems.
6. Implement the combined exploration method, and conduct experiments to collect data.
7. Quantitatively and qualitatively analyse the final performance results, formulate a research report, and propose directions for future works.

7.2 Indications of success

- Successfully observing performance differences of BV and cVARE indicates that the chosen continuous MCTS problems are suitable for serving as testing beds to gather quantitative data for evaluating performance differences.
- Quantitative results of BV and cVARE performance differences aligning with the mathematical analysis indicates that the theoretical analysis of BV and cVARE performance difference is plausible.
- Quantitative results showing combined exploration method outperforms BV and cVARE in the selected continuous MCTS problems indicates that the construction of a better exploration algorithm for continuous MCTS is successful.

8 Reference

Browne, C.B., Powley, E., Whitehouse, D., Lucas, S.M., Cowling, P.I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S. and Colton, S., 2012. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, *4*(1), pp.1-43.

Couetoux, A., 2013. *Monte Carlo tree search for continuous and stochastic sequential decision making problems* (Doctoral dissertation, Université Paris Sud-Paris XI).

Couetoux, A., Doghmen, H. and Teytaud, O., 2012, January. Improving the exploration in upper confidence trees. In *International Conference on Learning and Intelligent Optimization* (pp. 366-371). Springer, Berlin, Heidelberg.

Gelly, S. and Silver, D., 2007, June. Combining online and offline knowledge in UCT. In *Proceedings of the 24th international conference on Machine learning* (pp. 273-280).

Wang, Y., Audibert, J.Y. and Munos, R., 2008. Algorithms for infinitely many-armed bandits. *Advances in Neural Information Processing Systems*, *21*.

Yee, T., Lisý, V. and Bowling, M.H., 2016, January. Monte Carlo Tree Search in Continuous Action Spaces with Execution Uncertainty. In *IJCAI* (pp. 690-697).