On Monte Carlo Tree Search With Multiple Objectives

Michael Painter

Pembroke College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2024

Acknowledgements

TODO: acknowledgements here

Abstract

TODO: abstract here

Contents

List of Figures

List of Tables

List of Notation

$\mathbb{1}$ . . . . . The indicator function, where $\mathbb{1}(A) = 1$ when A is true, and $\mathbb{1}(A) = 0$ when A is false.

Markov Decision Processes (Section 2.2)

$\mathcal{A}$ . . . . . Set of actions in an MDP.

$\mathcal{B}$ . . . . . TODO: thts backups.

H . . . . . The finite-horizon time bound of an MDP.

p . . . . . The next-state probability distribution of an MDP. $p(\cdot|s, a) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$.

$\mathcal{R}$ . . . . . The reward function of in an MDP: $\mathcal{R}(s, a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.

$\mathcal{S}$ . . . . . Set of states in an MDP.

t . . . . . TODO: free variable for current timestep? And generally add a list of free variables?

Trial Based Heuristic Tree Search (Section 2.3)

a_t . . . . . A random variable for the tth action sampled in a trial of `THTS++`.

n . . . . . Number of trials run.

π . . . . . A search policy, TODO: define, this is a parameter of thts++. TODO: handle π^k

r_t . . . . . A random variable for the tth reward sampled in a trial of `THTS++`.

s_t . . . . . A random variable for the tth state sampled in a trial of `THTS++`.

$\mathcal{T}$ . . . . . A THTS search tree. TODO: With: $\mathcal{T} \subseteq \mathcal{S} \cup \mathcal{S} \times \mathcal{A}$. TODO: handle $\mathcal{T}_k$

T . . . . . Computation time limit.

τ . . . . . A trajectory for a trial of `THTS++`. I.e. $\tau = (s_0, a_1, s_1, ..., s_{h-1}, a_{h-1}, s_h)$. TODO: handle τ_k

$V_{\textbf{init}}$ . . . . . The initialisation function used in `THTS++`, used to initialise the value of a new decision node.

List of Abbreviations

Markov Decision Processes (Section 2.2)

MDP . . . . . Markov Decision Process.

Trial Based Heuristic Tree Search (Section 2.3)

`CNODE` . . . . . The chance node class.

`cnode` . . . . . An instance of `CNODE`, i.e. a chance node instance.

`DNODE` . . . . . The decision node class.

`dnode` . . . . . An intance of `DNODE`, i.e. a decision node instance.

MCTS . . . . . Monte Carlo Tree Search.

`mcts_mode` . . . TODO: definition of `mcts_mode`

MENTS . . . . Maximum ENtropy Tree Search.

`node` . . . . . A mapping from states and state-action pairs to their corresponding decision and chance nodes respectively.

THTS . . . . . Trial-based Heuristic Tree Search.

`THTS++` . . . . . TODO: thts++

UCT . . . . . Upper Confidence Bound applied to Trees/

1 Introduction

Contents

TODO: chapter structure (i.e. in the introduction section I give some background in the field(s), cover the main contributions of this thesis, etc, etc).

1.1 Overview

TODO: list

- Give some context around MCTS (and talk about exploration and exploitation), and why we might use it
  - Larger scale than tabular methods
  - Can do probability and theory stuff (and some explainability, by looking at stats in the tree the agent used)
  - Can use tree search with neural networks to get some of the above (and use for neural network training as in alpha zero)

- Argument from DENTS paper for exploration > exploitation (in context of planning in a simulator)
- Give high level overview of Multi-Objective RL, and why it can be useful
- Give an idea of how my work fits into MCTS and MORL as a whole
- Discuss research questions/issues with current literature (i.e. introduce some of the ideas from contributions section below)

1.2 Contributions

TODO: Inline acronyms used, or make sure that they're defined before hand

Throughout this thesis, we will consider the following questions related to Monte Carlo Tree Search and Multi-Objective Reinforcement Learning:

Q1 - Exploration: When planning in a simulator with limited time, how can MCTS algorithms best explore to make good decisions?

Q1.1 - Entropy: Entropy is often used as an exploration objective in RL, but can it be used soundly in MCTS?

Q1.2 - Multi-Objective Exploration: How can Multi-Objective MCTS methods explore to find optimal actions for different objectives?

Q2 - Scalability: How can the scalability of (multi-objective) MCTS methods be improved?

Q2.1 - Complexity: MCTS algorithms typically run in $O(nAH)$, but are there algorithms that can improve upon this?

Q2.2 - Multi-Objective Scalability: With respect to the size of environments, how scalable are Multi-Objective MCTS methods?

Q2.3 - Curse of Dimensionality: With respect to the number of objectives, to what extent do Multi-Objective MCTS methods suffer from the curse of dimensionality?

Q3 - Evaluation: How can we best evaluate a search tree produced by a Monte Carlo Tree Search algorithm?

Q3.1 - Tree Policies: Does it suffice to extract a policy from a single search tree for evaluation? TODO: going to have to run some extra experiments for that, but I probably should do that for completeness anyway

Q3.2 - Multi-Objective Evaluation: Can we apply methods from the MORL literature to theoretically and empirically evaluate Multi-Objective MCTS?

TODO: some words about how below is the contributions we're making in this thesis and expand these bullets a bit more

- Max Entropy can be misaligned with reward maximisation (**Q1.1 - Entropy**)
- Boltzmann Search Policies - BTS and DENTS (**Q1.1 - Entropy**, and with extra results **Q3.1 - Tree Policies**)
- Use the alias method to make faster algorithms (**Q2.1 - Complexity**)
- Simple regret (**Q1 - Exploration**)
- Use of contexts in THTS to make consistent decisions in each trial (**Q1.2 - Multi-Objective Exploration**)
- Contextual regret introduced in CHMCTS (**Q2.2 - Multi-Objective Scalability**, **Q3.2 - Multi-Objective Evaluation**)
- Contextual Zooming and CHMCTS (designed for **Q1.2 - Multi-Objective Exploration**, runtimes cover **Q2.3 - Curse of Dimensionality**, results **Q3.2 - Multi-Objective Evaluation**)
- Simplex maps (**Q1.2 - Multi-Objective Exploration**, **Q2.2 - Multi-Objective Scalability**, **Q2.3 - Curse of Dimensionality**)
- Contextual Simple Regret (**Q3.2 - Multi-Objective Evaluation**)

1.3 Structure of Thesis

TODO: a paragraph with a couple lines to a paragraph about each chapter. This is the high level overview/intro to the thesis paragraph. I.e. this section is "this is the story of my thesis in a page or two"

1.4 Publications

TODO: update final publication when submit

The work covered in this thesis also appears in the following publications:

- Painter, M; Lacerda, B; and Hawes, N. "Convex Hull Monte-Carlo Tree-Search." In *Proceedings of the international conference on automated planning and scheduling. Vol. 30. 2020*, ICAPS, 2020, (see Appendix **??**).
- Painter, M; Baioumy, M; Hawes, N; and Lacerda, B. "Monte Carlo Tree Search With Boltzmann Exploration." In *Advances in Neural Information Processing Systems, 36, 2023*, NeurIPS, 2023, (see Appendix **??**).
- Painter, M; Hawes, N; and Lacerda, B. "Simplex Maps for Multi-Objective Monte Carlo Tree Search." In *TODO*, **Under Review at conf_name**, (see Appendix **??**).

2
Background

Contents

TODO: Introduce that going to introduce notation and give the building blocks this thesis builds off

TODO: R.e. presenting RL before tree search (if I forget to mention it when I message), I was hoping to present THTS and MCTS as part of reinforcement learning, and try to forgo the whole planning vs RL conversation. If it's ok, I'll write it as is, keep the below todo, and then revisit it when proof reading.

TODO: Nicks comment to revisit: Wouldn't it be better to start with single objective MDPs, then introduce methods of solving them (planning, bandits, rl) and the different assumptions they make?)

2.1 Multi-Armed Bandits

TODO: Introduce tree search using multi-armed bandits? TODO: list

- $R(s, a)$ is a random variable in MAB literature, but we're assuming it's a fixed value in RL
- Multi-Armed Bandits routines algos
- Exploring Bandits routines and algos
- Contextual Bandits routines and algos

2.2 Markov Decision Processes and Reinforcement Learning

TODO: list

- Typical agent interacting with environment diagram
- Agent planning with simulator
- MDPs definition
- Value functions (single and multi-objective)
- Basic results and definitions we use (tabular planning algorithms)
- Talk about entropy and some of that work (probably a subsection)

2.3 Trial-Based Heuristic Tree Search and Monte-Carlo Tree Search

TODO: list

- Give high level overview of MCTS (why use it etc)
- Outline that I'll present this as here is THTS, and then here's the THTS routines for MCTS

2.3.1 Trial Based Heuristic Tree Search

TODO: list

- Present thts++
- Indicate what parts are new versus the original paper (context function, optionally running `mcts_mode` and mutli-threading)
- Define terms precisely and consistently, for example `mcts_mode` (say that notation and terminology varies widely in literature, e.g. does uct run in `mcts_mode` or not?)
- Mention that V_{init} can be implemented as V_θ to be used with deep RL methods

2.3.2 Monte-Carlo Tree Search

TODO: list

- Give overview of MCTS
- Give UCT in terms of THTS schema
- Define terms precisely and consistently in terms of THTS functions, maybe `mcts_mode` should go here
- Define the value initialisation of THTS using a rollout policy for MCTS
- Talk about the things that are ambiguous from literature (e.g. people will just say UCT, which originally presented doesn't run in `mcts_mode`, but often assumed it does)
- Should talk about multi-armed bandits here?

2.3.3 Maximum Entropy Tree Search

TODO: list

- Define MENTS here

2.4 Multi-Objective Reinforcement Learning

TODO: list

- MOMDP definition
- (Expected) utility
- Define an interface for pareto front and convex hull objects
- Define CHVI
- Should talk about multi-objective and/or contextual multi-armed bandits here?
- I'm planning on aligning this section with the recent MORL survey [1]
- Mention some deep MORL stuff, say that this work (given AlphaZero) is adjacent work

2.5 Multi-Objective Monte Carlo Tree Search

TODO: I think this whole section can just go in litrev

TODO: list

- Define the old methods (using the CH object methods, so clear that not doing direct arithmetic)
- Mention that old method could be written using the arithmetic of CHMCTS (but they don't)
- Different flavours copy UCT action selection, but with different variants
- Link back to contributions and front load our results showing that all of the old methods don't explore correctly

2.6 Sampling Random Variables

TODO: list

- Talk about the alias method here
- Reference to chapter 4 section where talk about using this with THTS

3 Literature Review

Contents

TODO: currently this is a copy and paste of what I originally wrote for background chapter 2. Deleted parts which are irrelevant for litreview here (and vice versa for the background section).

TODO: I'm also going to use this as a space to paste papers I should write about as they come up while writing later chapters

3.1 Multi-Armed Bandits

TODO: Maybe dont need to cover this in litrev, but should talk about exploring bandits, UCT and contextual bandits either in background or in litrev

3.2 Reinforcement Learning

TODO: list

- Talk about entropy and some of that work (probably a subsection)

3.3 Trial-Based Heuristic Tree Search and Monte-Carlo Tree Search

3.3.1 Trial Based Heuristic Tree Search

TODO: THTS paper

3.3.2 Monte-Carlo Tree Search

TODO: list

- Talk about the things that are ambiguous from literature (e.g. people will just say UCT, which originally presented doesn't run in `mcts_mode`, but often assumed it does)
- Should talk about multi-armed bandits here?

3.3.3 Maximum Entropy Tree Search

TODO: MENTS

3.4 Multi-Objective Reinforcement Learning

TODO: list

- Should talk about multi-objective and/or contextual multi-armed bandits here?
- Bunch of the work covered in recent MORL survey [1]
- Mention some deep MORL stuff, say that this work (given AlphaZero) is adjacent work

3.5 Multi-Objective Monte Carlo Tree Search

TODO: I think this whole section can just go in litrev

TODO: list

- Define the old methods (using the CH object methods, so clear that not doing direct arithmetic)
- Mention that old method could be written using the arithmetic of CHMCTS (but they don't)
- TODO: write about & make sure its implemented - its because just updating for 1 is more efficient in deterministic, and say that the additions can be implemented as updating for 1 value when determinstic
- Different flavours copy UCT action selection, but with different variants
- Link back to contributions and front load our results showing that all of the old methods don't explore correctly

4

Monte Carlo Tree Search With Boltzmann Exploration

Contents

4.1 Introduction

TODO: list

- high level overview of DENTS work
- discuss how DENTS answers the research questions from introduction chapter
- state clearly that we're in single objective land here
- Comment about work exploring multi-armed bandits motivating this work

4.2 Boltzmann Search

TODO: list

- Recall MENTS
- Define BTS using THTS functions
- Define DENTS using THTS functions
- Discuss alias method variant (and complexity analysis) in a subsection?

4.3 Toy Environments

TODO: list

- Define D-chain stuff from the paper
- Define the D-chain with entropy trap
- Front load some results still

4.4 Theoretical Results

TODO: list

- add theoretical results

4.5 Empirical Results

TODO: list

- DChain
- GridWorlds
- Go

4.6 Full Results

TODO: there's a lot of figures for the D-chain environment, work out how to best fit them in? Or put them in this seperate section?

5 Convex Hull Monte Carlo Tree Search

Contents

5.1 Introduction

TODO: list

- high level overview of CHMCTS work
- discuss how CHMCTS answers the research questions from introduction chapter
- moving into multi-objective land now
- Comment about CHVI and prior MOMCTS work motivating this

5.2 Contextual Tree Search

TODO: list

- Discuss need for context when doing multi-objective tree Search
  - Use an example env where left gives (1,0) and right gives (0,1), optimal policy picks just left or just right, but hypervolume based methods wont
  - Use previous work on these examples and show they dont do well bad
- Discuss how UCT = running a non-stationary UCB at each node, so given above discussion, there is work in contextual MAB
- Introduce contextual regret here

5.3 Contextual Zooming for Trees

TODO: list

- Give contextual zooming for trees algorithm
- Discussion on the contextual MAB to non-stationary contextual MAB stuff (CZT is to CZ what UCT is to UCB) (and what theory carry over)

5.4 Convex Hull Monte Carlo Tree Search

TODO: list

- Give convex hull monte carlo tree search
- Contextual zooming with the convex hull backups

5.5 Results

TODO: list

- Results from CHMCTS paper
- Get same plots from C++ code, but compare expected utility, rather than the confusing hypervolume ratio stuff

6

Simplex Maps for Multi-Objective Monte Carlo Tree Search

Contents

6.1 Introduction

TODO: list

- high level overview of simplex maps work
- discuss how simplex maps answer the research questions from introduction chapter
- staying in multi-objective land now
- Motivated by CHMCTS being slow

6.2 Simplex Maps

TODO: list

- Define simplex map interface
- Give details on how to efficiently implement the interface with tree structures
- (Good diagram is everything here I think)

6.3 Simplex Maps in Tree Search

TODO: list

- Come up with better title for section
- Use simplex maps interface to create algorithms from the dents work
- Give a high level idea of what δ parameter is (used in theory section)

6.4 Theoretical Results

TODO: list

- Convergence can build ontop of DENTS results
- Runtime bounds (better than $O(2^D)$ which is what using convex hulls has)
- Simplex map has a diameter δ (i.e. the furthest away a new context could be from a point in the map)
- Bounds can then come from that diameter (which is a parameter of the simplex map/algorithm) and DENTS results

6.5 Empirical Results

TODO: list

- Results from MO-Gymnasium
- Compare algorithms using expected utility

7 Conclusion

Contents

TODO: Something about we'll conclude by looking back at contributions and possible future work.

7.1 Summary of Contributions

TODO: go through each of the research questions and contributions, and write about how the work answers the research questions

7.2 Future Work

TODO: outline some avenues of potential future work

Appendices

A

List Of Appendices To Consider

- Multi Armed Bandits, maybe
- MAybe from of the things in background are more appropriate as appendices?

Bibliography

[1] Conor F Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M Zintgraf, Richard Dazeley, Fredrik Heintz, et al. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1):26, 2022.

[2] Thomas Keller and Malte Helmert. Trial-based heuristic tree search for finite horizon mdps. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 23, pages 135–143, 2013.

[3] Michael Painter. THTS++,https://github.com/MWPainter/thts-plus-plus.