

# 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

<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>xi</b>
<b>List of Notation</b>	<b>xiii</b>
<b>List of Abbreviations</b>	<b>xv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Overview . . . . .	1
1.2 Contributions . . . . .	2
1.3 Structure of Thesis . . . . .	3
1.4 Publications . . . . .	4
<b>2 Background</b>	<b>5</b>
2.1 Multi-Armed Bandits . . . . .	5
2.2 Reinforcement Learning . . . . .	5
2.3 Trial-Based Heuristic Tree Search and Monte-Carlo Tree Search . .	6
2.3.1 Trial Based Heuristic Tree Search . . . . .	7
2.3.2 Monte-Carlo Tree Search . . . . .	8
2.3.3 Maximum Entropy Tree Search . . . . .	9
2.4 Multi-Objective Reinforcement Learning . . . . .	9
2.5 Multi-Objective Monte Carlo Tree Search . . . . .	9
2.6 Sampling Random Variables . . . . .	10
<b>3 Literature Review</b>	<b>11</b>
3.1 Multi-Armed Bandits . . . . .	11
3.2 Reinforcement Learning . . . . .	12
3.3 Trial-Based Heuristic Tree Search and Monte-Carlo Tree Search . .	12
3.3.1 Trial Based Heuristic Tree Search . . . . .	12
3.3.2 Monte-Carlo Tree Search . . . . .	12
3.3.3 Maximum Entropy Tree Search . . . . .	12
3.4 Multi-Objective Reinforcement Learning . . . . .	12
3.5 Multi-Objective Monte Carlo Tree Search . . . . .	13
3.6 Sampling Random Variables . . . . .	13

<b>4</b>	<b>Monte Carlo Tree Search With Boltzmann Exploration</b>	<b>15</b>
4.1	Introduction . . . . .	15
4.2	Boltzmann Search . . . . .	16
4.3	Toy Environments . . . . .	16
4.4	Theoretical Results . . . . .	16
4.5	Empirical Results . . . . .	16
4.6	Full Results . . . . .	17
<b>5</b>	<b>Convex Hull Monte Carlo Tree Search</b>	<b>19</b>
5.1	Introduction . . . . .	19
5.2	Contextual Tree Search . . . . .	19
5.3	Contextual Zooming for Trees . . . . .	20
5.4	Convex Hull Monte Carlo Tree Search . . . . .	20
5.5	Results . . . . .	20
<b>6</b>	<b>Simplex Maps for Multi-Objective Monte Carlo Tree Search</b>	<b>21</b>
6.1	Introduction . . . . .	21
6.2	Simplex Maps . . . . .	22
6.3	Simplex Maps in Tree Search . . . . .	22
6.4	Theoretical Results . . . . .	22
6.5	Empirical Results . . . . .	22
<b>7</b>	<b>Conclusion</b>	<b>23</b>
7.1	Summary of Contributions . . . . .	23
7.2	Future Work . . . . .	23
<b>Appendices</b>		
<b>A</b>	<b>List Of Appendices To Consider</b>	<b>27</b>
<b>B</b>	<b>Convex Hull Monte Carlo Tree Search - ICAPS2020</b>	<b>29</b>
<b>C</b>	<b>Monte Carlo Tree Search with Boltzmann Exploration</b>	<b>31</b>
<b>D</b>	<b>Simplex Maps for Multi-Objecitve Monte Carlo Tree Search</b>	<b>33</b>
	<b>Bibliography</b>	<b>35</b>



## List of Figures



## List of Tables



# List of Notation

## Markov Decision Processes

$a_t$  . . . . . TODO

$r_t$  . . . . . TODO

$s_t$  . . . . . TODO

$t$  . . . . . TODO: free variable for current timestep?

## Trial Based Heuristic Tree Search

$n$  . . . . . Number of trials run

$T$  . . . . . Computation time limit



# List of Abbreviations

## **Trial Based Heuristic Tree Search**

<b>CNODE</b>	. . . . .	The chance node class.
<b>cnode</b>	. . . . .	An instance of <b>CNODE</b> , i.e. a chance node instance.
<b>DNODE</b>	. . . . .	The decision node class.
<b>dnode</b>	. . . . .	An intance of <b>DNODE</b> , i.e. a decision node instance.
<b>MCTS</b>	. . . . .	Monte Carlo Tree Search
<b>mcts_mode</b>	. . .	TODO: definition of <b>mcts_mode</b>
<b>MENTS</b>	. . . .	Maximum ENtropy Tree Search
<b>THTS++</b>	. . . . .	TODO: thts++
<b>UCT</b>	. . . . .	Upper Confidence Bound applied to Trees





# 1

## Introduction

### Contents

---

<b>1.1</b>	<b>Overview . . . . .</b>	<b>1</b>
<b>1.2</b>	<b>Contributions . . . . .</b>	<b>2</b>
<b>1.3</b>	<b>Structure of Thesis . . . . .</b>	<b>3</b>
<b>1.4</b>	<b>Publications . . . . .</b>	<b>4</b>

---

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: I'm not too happy with this section, I think most of the questions could be phrased better, but nothing I'm writing sounds quite right to me at the moment. I also just tried to come up with as many questions as I could with the idea of I don't need to keep them all

TODO: Also inline acronyms used,

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, how can an algorithm best exploit the time it is given to produce the best recommendation action?
- Q2 - Entropy:** Entropy is often used as an exploration objective in RL, but can it be used soundly in MCTS?
- Q3 - Complexity:** MCTS algorithms typically run in  $O(nAH)$ , but are there algorithms that can improve upon this?
- Q4 - Scalability:** How can the scalability (with respect to size of environments) of multi-objective MCTS methods be improved?
- Q5 - Curse of Dimensionality:** To what extent do Multi-Objective MCTS methods suffer from the curse of dimensionality? (basically Multi-Objective complexity)

**Q6 - Tree Evaluation:** How can we best evaluate a search tree produced by a Monte Carlo Tree Search algorithm? TODO: add a matching contrib

**Q7 - 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 (**Q2 - Entropy**)
- Boltzmann Search Policies - BTS and DENTS (**Q1 - Exploration, Q2 - Entropy**, and eval **Q6 - Tree Evaluation**)
- Use the alias method to make faster algorithms (**Q3 - Complexity**)
- Simple regret (**Q1 - Exploration**)
- Contextual regret introduced in CHMCTS (**Q4 - Scalability, Q7 - Multi-Objective Evaluation**)
- Contextual Zooming and CHMCTS (runtimes cover **Q5 - Dimensionality**, results **Q6 - Tree Evaluation**)
- Simplex maps (**Q4 - Scalability, Q5 - Dimensionality**)
- Contextual Simple Regret (**Q6 - Tree Evaluation, Q7 - 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 B).
- 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 C).
- 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 D).

# 2

## Background

### Contents

---

<b>2.1</b>	<b>Multi-Armed Bandits . . . . .</b>	<b>5</b>
<b>2.2</b>	<b>Reinforcement Learning . . . . .</b>	<b>5</b>
<b>2.3</b>	<b>Trial-Based Heuristic Tree Search and Monte-Carlo Tree Search . . . . .</b>	<b>6</b>
2.3.1	Trial Based Heuristic Tree Search . . . . .	7
2.3.2	Monte-Carlo Tree Search . . . . .	8
2.3.3	Maximum Entropy Tree Search . . . . .	9
<b>2.4</b>	<b>Multi-Objective Reinforcement Learning . . . . .</b>	<b>9</b>
<b>2.5</b>	<b>Multi-Objective Monte Carlo Tree Search . . . . .</b>	<b>9</b>
<b>2.6</b>	<b>Sampling Random Variables . . . . .</b>	<b>10</b>

---

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

### 2.1 Multi-Armed Bandits

TODO: Introduce tree search using multi-armed bandits?

### 2.2 Reinforcement Learning

TODO: list

- MDPs definition

- Basic results and definitions we use
- 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?)

### 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

---

<b>3.1</b>	<b>Multi-Armed Bandits . . . . .</b>	<b>11</b>
<b>3.2</b>	<b>Reinforcement Learning . . . . .</b>	<b>12</b>
<b>3.3</b>	<b>Trial-Based Heuristic Tree Search and Monte-Carlo Tree Search . . . . .</b>	<b>12</b>
3.3.1	Trial Based Heuristic Tree Search . . . . .	12
3.3.2	Monte-Carlo Tree Search . . . . .	12
3.3.3	Maximum Entropy Tree Search . . . . .	12
<b>3.4</b>	<b>Multi-Objective Reinforcement Learning . . . . .</b>	<b>12</b>
<b>3.5</b>	<b>Multi-Objective Monte Carlo Tree Search . . . . .</b>	<b>13</b>
<b>3.6</b>	<b>Sampling Random Variables . . . . .</b>	<b>13</b>

---

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: Just UCT/dont lit review?

## 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 deterministic
- 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

### 3.6 Sampling Random Variables

TODO: Just citing the alias method papers?



# 4

## Monte Carlo Tree Search With Boltzmann Exploration

### Contents

---

<b>4.1</b>	<b>Introduction</b>	<b>15</b>
<b>4.2</b>	<b>Boltzmann Search</b>	<b>16</b>
<b>4.3</b>	<b>Toy Environments</b>	<b>16</b>
<b>4.4</b>	<b>Theoretical Results</b>	<b>16</b>
<b>4.5</b>	<b>Empirical Results</b>	<b>16</b>
<b>4.6</b>	<b>Full Results</b>	<b>17</b>

---

### 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

---

<b>5.1</b>	<b>Introduction . . . . .</b>	<b>19</b>
<b>5.2</b>	<b>Contextual Tree Search . . . . .</b>	<b>19</b>
<b>5.3</b>	<b>Contextual Zooming for Trees . . . . .</b>	<b>20</b>
<b>5.4</b>	<b>Convex Hull Monte Carlo Tree Search . . . . .</b>	<b>20</b>
<b>5.5</b>	<b>Results . . . . .</b>	<b>20</b>

---

### 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

---

<b>6.1</b>	<b>Introduction</b>	<b>21</b>
<b>6.2</b>	<b>Simplex Maps</b>	<b>22</b>
<b>6.3</b>	<b>Simplex Maps in Tree Search</b>	<b>22</b>
<b>6.4</b>	<b>Theoretical Results</b>	<b>22</b>
<b>6.5</b>	<b>Empirical Results</b>	<b>22</b>

---

### 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  $\delta$  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  $\delta$  (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

---

<b>7.1</b>	<b>Summary of Contributions . . . . .</b>	<b>23</b>
<b>7.2</b>	<b>Future Work . . . . .</b>	<b>23</b>

---

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







## List Of Appendices To Consider

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



# B

## Convex Hull Monte Carlo Tree Search - ICAPS2020

TODO: something about paper appearing in ICAPS 2020 & reference to the chapter that covers this content.

TODO: c&p paper



# Monte Carlo Tree Search with Boltzmann Exploration

TODO: something about paper appearing in NeurIPS 2023 & reference to the chapter that covers this content.

TODO: c&p paper

D

# Simplex Maps for Multi-Objective Monte Carlo Tree Search

TODO: something about paper submitted to XXX & reference to the chapter that covers this content.

TODO: c&p paper



# 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>.