Monte-Carlo Tree Search

An introduction

Jérémie DECOCK

Inria Saclay - LRI

May 2012

Introduction

Monte-Carlo Tree Search (MCTS)

- MCTS is a recent algorithm for *sequential decision making*
- It applies to *Markov Decision Processes* (MDP)
  - discrete-time t with finite horizon T
  - state $\mathbf{s}_t \in \mathcal{S}$
  - action $\mathbf{a}_t \in \mathcal{A}$
  - transition function $\mathbf{s}_{t+1} = \mathcal{P}(\mathbf{s}_t, \mathbf{a}_t)$
  - cost function $r_t = \mathcal{R}_{\mathcal{P}}(\mathbf{s}_t)$
  - reward $R = \sum_{t=0}^{T} r_t$
  - policy function $\mathbf{a}_t = \pi_{\mathcal{P}}(\mathbf{s}_t)$
  - we look for the policy π^* that maximizes expected R

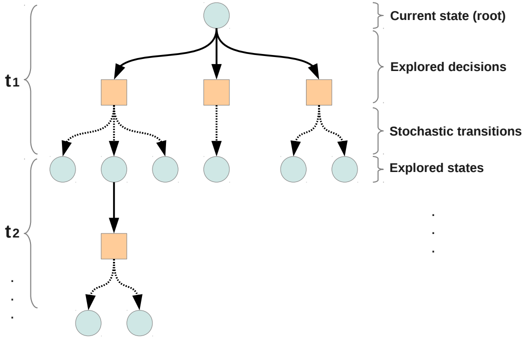
MCTS strength

- Mcts is a versatile algorithm (it does not require knowledge about the problem)
- especially, does not require any knowledge about the Bellman value function
- stable on high dimensional problems
- it outperforms all other algorithms on some problems (difficult games like Go, general game playing, ...)

MCTS

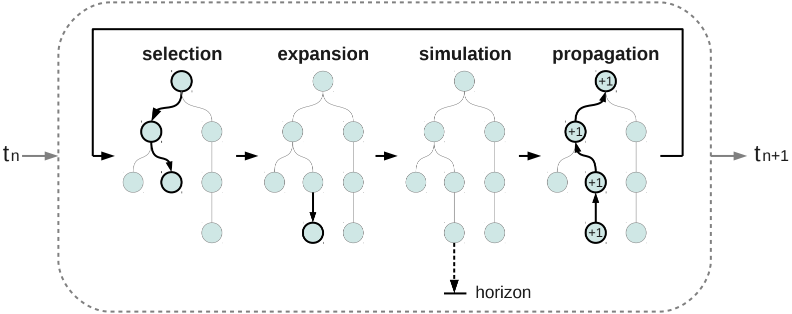
Problems are represented as a tree structure:

- blue circles = states
- plain edges + red squares = decisions
- dashed edges = stochastic transitions between two states



Monte-Carlo Tree Search

Main steps of MCTS

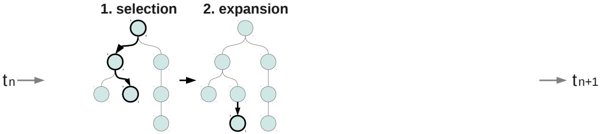


Main steps of MCTS

Starting from an initial state:

1. select the state we want to expand from

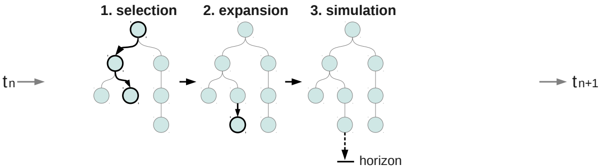
Main steps of MCTS



Starting from an initial state:

1. select the state we want to expand from
2. add the generated state in memory

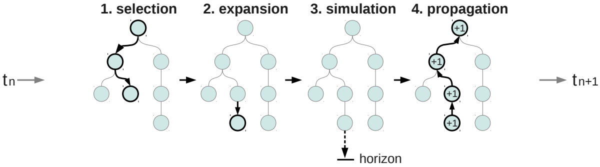
Main steps of MCTS



Starting from an initial state:

1. select the state we want to expand from
2. add the generated state in memory
3. evaluate the new state with a default policy until horizon is reached

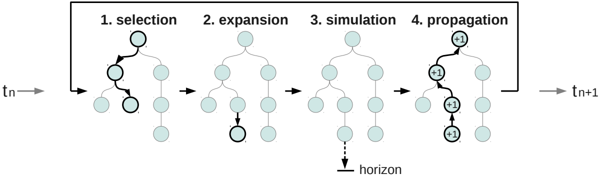
Main steps of MCTS



Starting from an initial state:

1. select the state we want to expand from
2. add the generated state in memory
3. evaluate the new state with a default policy until horizon is reached
4. back-propagation of some information:
   - $n(\mathbf{s}, \mathbf{a})$: number of times decision **a** has been simulated in **s**
   - $n(\mathbf{s})$: number of time **s** has been visited in simulations
   - $\hat{Q}(\mathbf{s}, \mathbf{a})$: mean reward of simulations where **a** was whosen in **s**

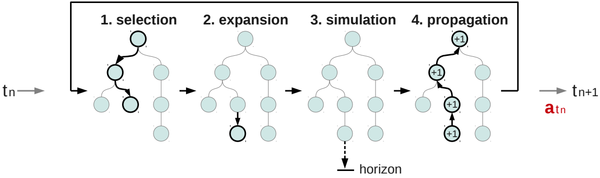
Main steps of MCTS



Starting from an initial state:

1. select the state we want to expand from
2. add the generated state in memory
3. evaluate the new state with a default policy until horizon is reached
4. back-propagation of some information:
   - $n(\mathbf{s}, \mathbf{a})$: number of times decision **a** has been simulated in **s**
   - $n(\mathbf{s})$: number of time **s** has been visited in simulations
   - $\hat{Q}(\mathbf{s}, \mathbf{a})$: mean reward of simulations where **a** was whosen in **s**

Main steps of MCTS



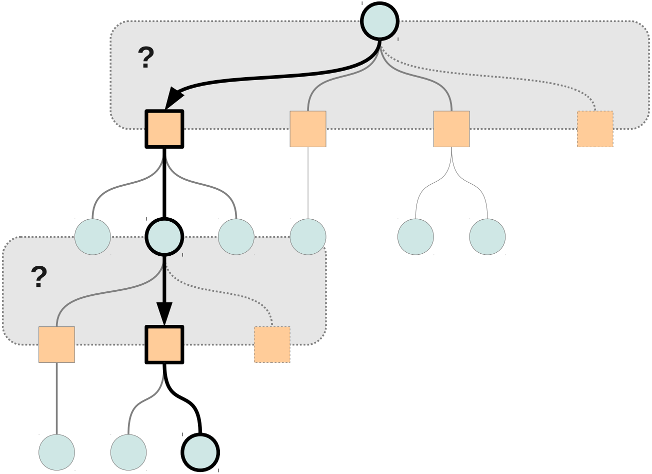
The selected decision

$\mathbf{a}_{t_n}$ = the most visited decision form the current state (root node)

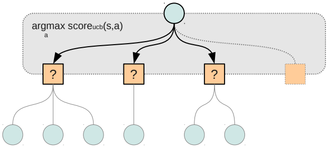
Selection and expansion

Selection step

How to select the state to expand ?



How to select the state to expand ?

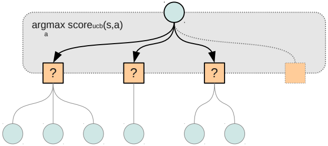


The *selection* phase is driven by *Upper Confidence Bound*

$$\text{score}_{\text{ucb}}(\mathbf{s},\mathbf{a}) = \underbrace{\hat{Q}(\mathbf{s},\mathbf{a})}_{1} + \underbrace{\sqrt{\frac{\log(2+n(\mathbf{s}))}{2+n(\mathbf{s},\mathbf{a})}}}_{2}$$

1. mean reward of simulations including action **a** in state **s**
2. the uncertainty on this estimation of the action's value

How to select the state to expand ?



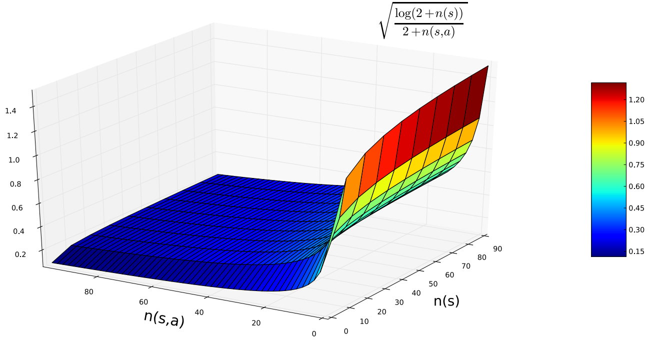
The *selection* phase is driven by *Upper Confidence Bound*

$$\text{score}_{\text{ucb}}(\mathbf{s},\mathbf{a}) = \underbrace{\hat{Q}(\mathbf{s},\mathbf{a})}_{1} + \underbrace{\sqrt{\frac{\log(2+n(\mathbf{s}))}{2+n(\mathbf{s},\mathbf{a})}}}_{2}$$

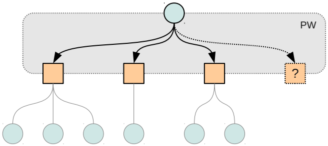
The selected action:

$$\mathbf{a}^{\star} = \arg\max_{\mathbf{a}} \; \text{score}_{\text{ucb}}(\mathbf{s},\mathbf{a})$$

How to select the state to expand ?



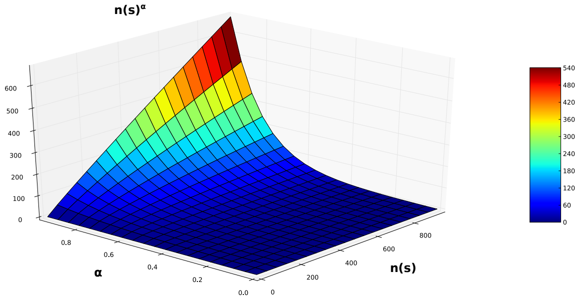
When should we expand?



One standard way of tackling the exploration/exploitation dilemma is *Progressive Widening*.

A new parameter $\alpha \in [0; 1]$ is introduced, to choose between exploration (add a decision to the tree) and exploitation (go to an existing node)

How to select the state to expand ?

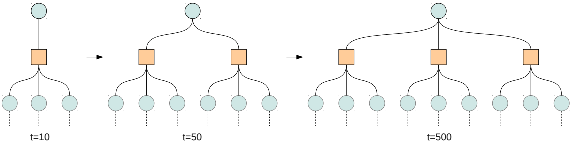


- if($|\mathcal{A}'_{\mathbf{s}}| < n(\mathbf{s})^{\alpha}$) then we explore a new decision
- else we simulate a known decision

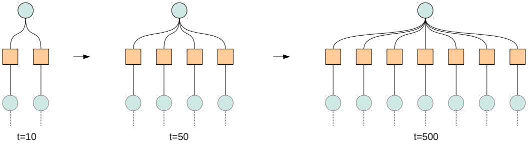
With $|\mathcal{A}'_{\mathbf{s}}|$ the number of legal actions in state $\mathbf{s}$

When should we expand?

$\alpha = 0.2$



$\alpha = 0.8$



References I