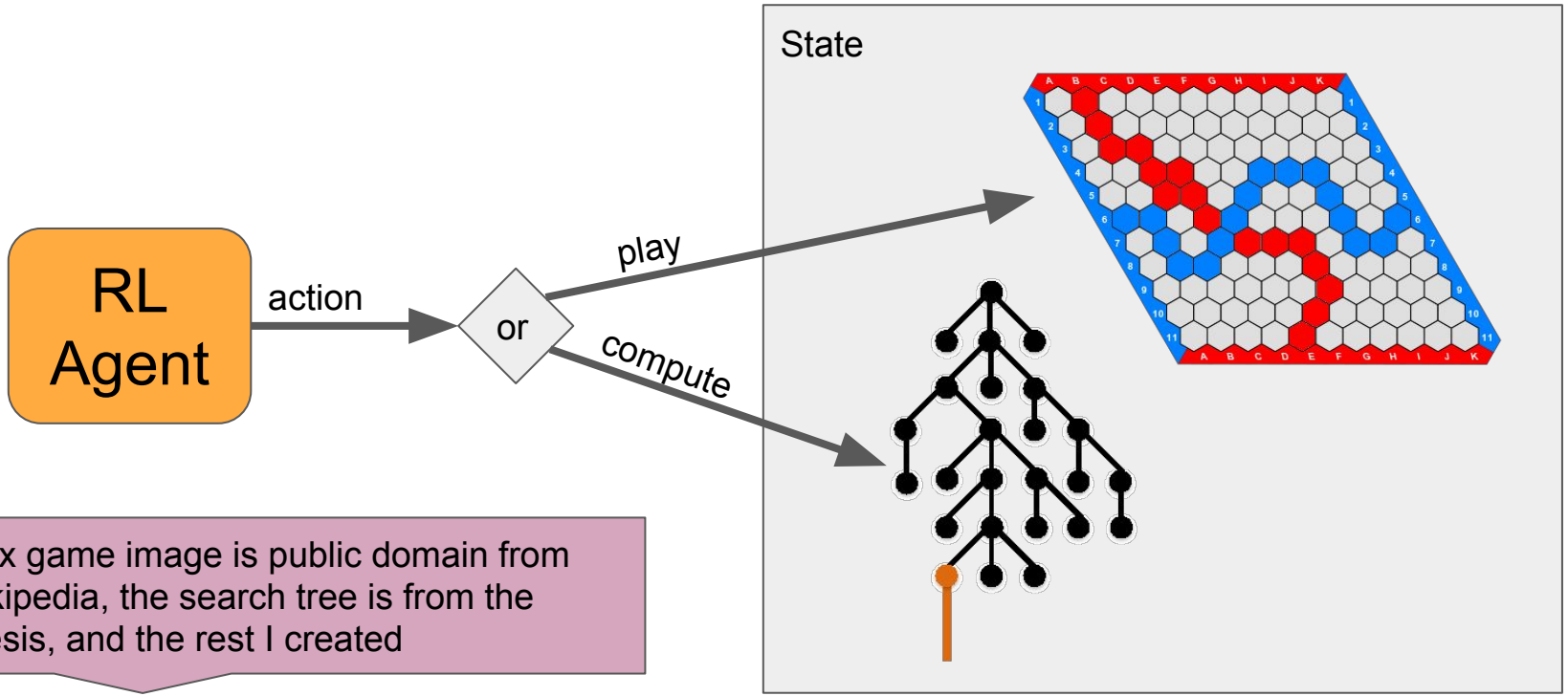
Metalevel RL for MCTS

Principles of Metalevel Control
by Nicholas Hay

Presented by Eric Langlois

Metalevel Monte Carlo Tree Search



Hex game image is public domain from wikipedia, the search tree is from the thesis, and the rest I created

Monte Carlo Tree Search

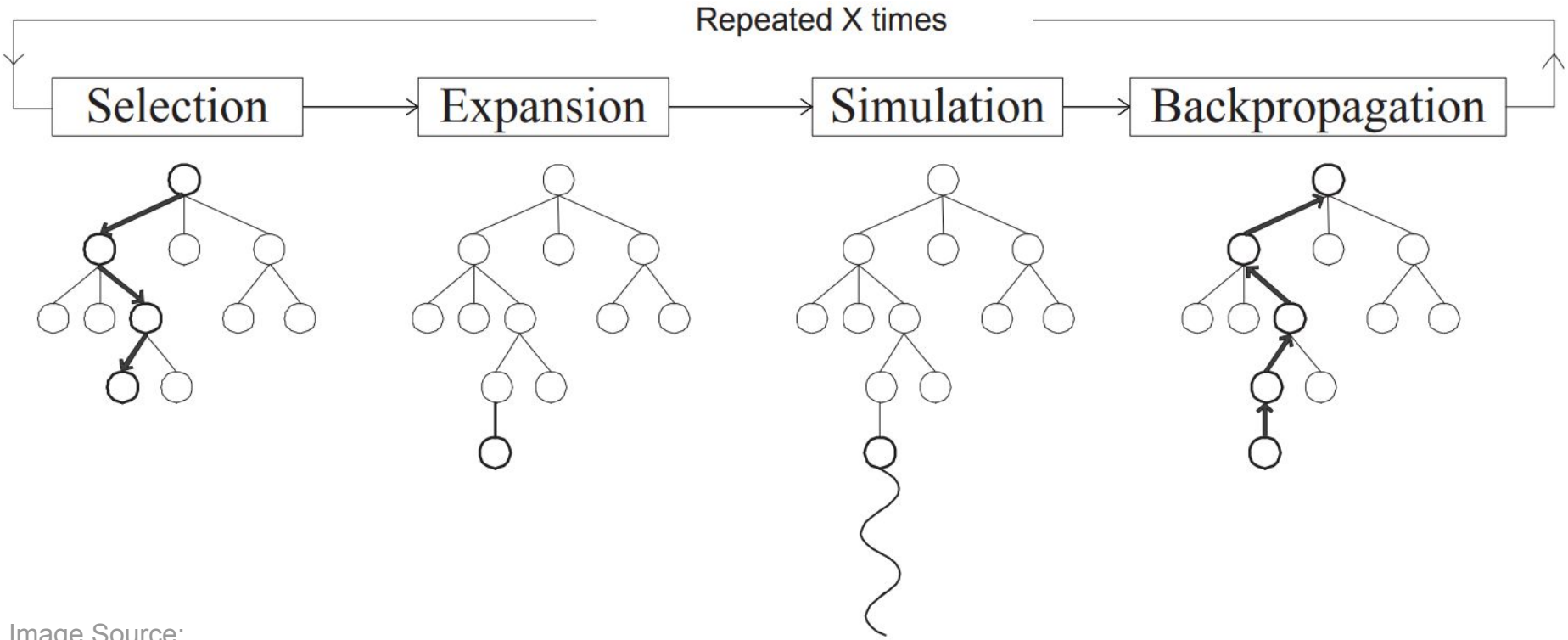
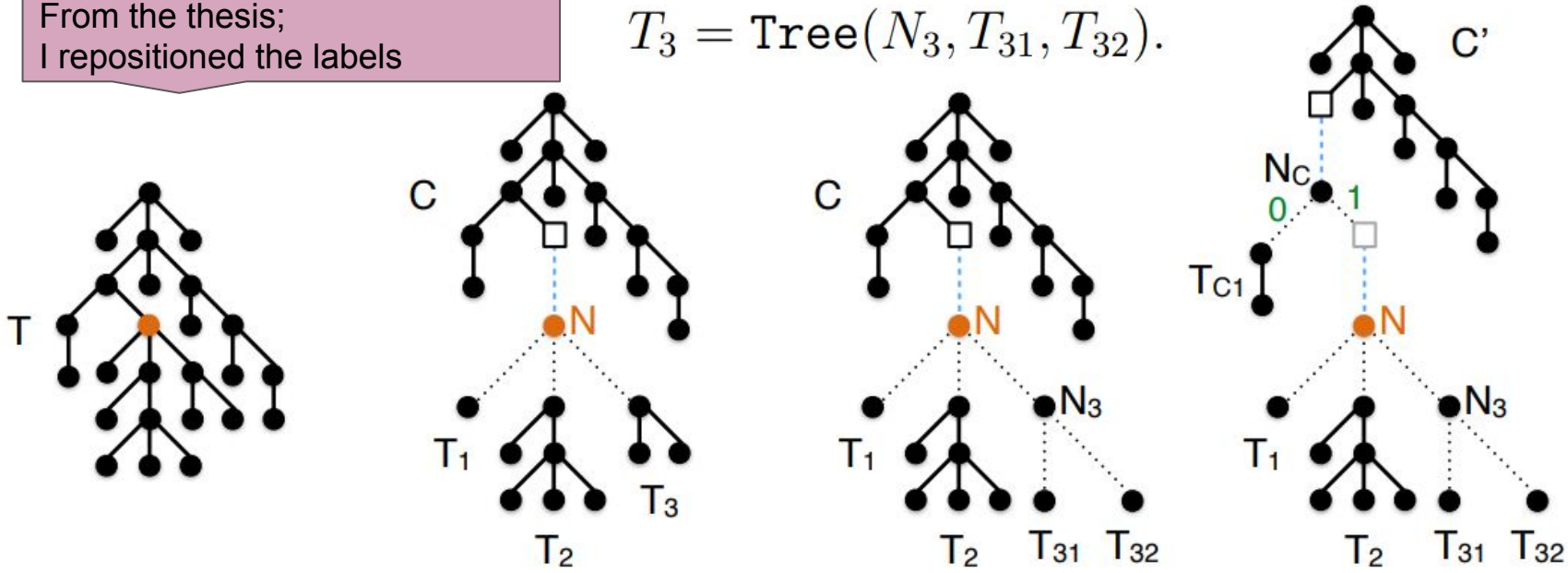


Image Source:
Chaslot et al (2008). Progressive Strategies for Monte-Carlo Tree Search. New Mathematics and Natural Computation.

Pointed Trees

From the thesis;
I repositioned the labels

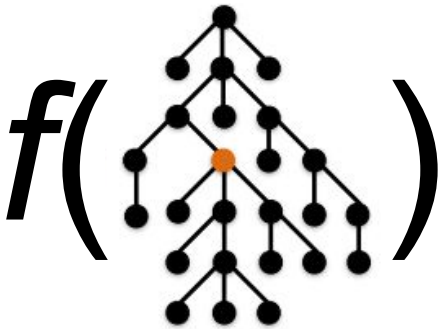
$$T_3 = \texttt{Tree}(N_3, T_{31}, T_{32}).$$



$$T = \texttt{PointedTree}(C, N, T_1, T_2, T_3).$$

$$C = \texttt{Context}(C', N_C, 1, T_{C1})$$

Recursive Functions on Pointed Trees



=

I created the formula display using copied bits of the tree image from the thesis

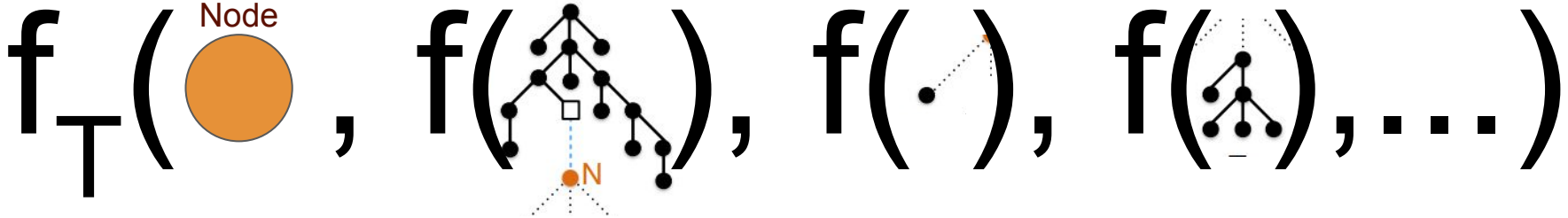
$(f_\theta, \frac{\partial f_\theta}{\partial \theta})$ is also recursive!

Pointer Node

Context

1st Subtree

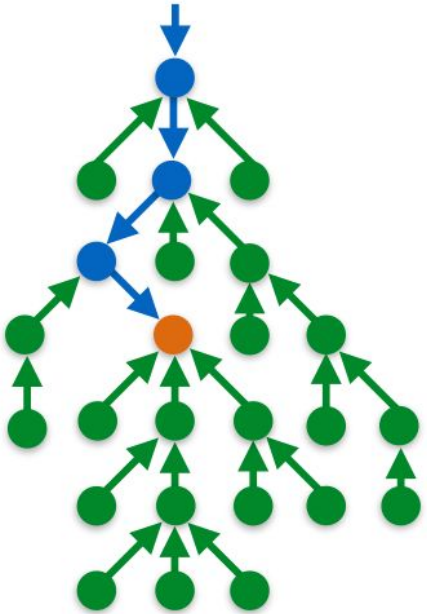
2nd Subtree



Local Computations on Pointed Trees

Operations

- `up(T)`
- `down(T, i)`
- `modify(T, N')`
- `insert(T, i, T')`

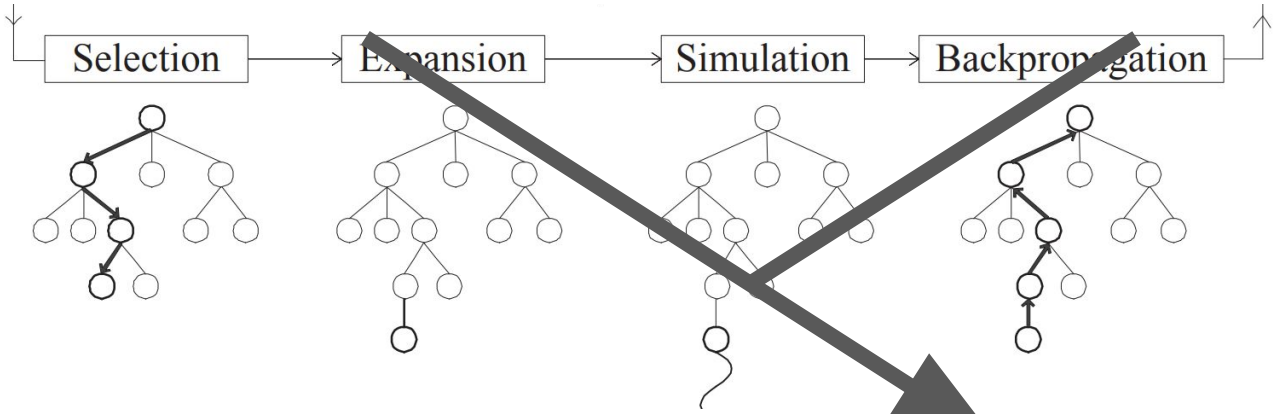


f(Context)

f(Tree)

From the thesis

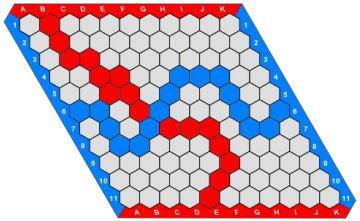
MCTS as a Metalevel MDP



Simulated Actions

Metalevel MDP Action Space

Real Actions



Fixed Implementation with Pointed Tree Operations

MCTS figure is from another paper and Hex board is public domain image. I added the rest

Learning a Metalevel Agent

$$\pi_\theta(T) = g_\theta(f_{\text{fixed}}(T))$$

Tree Functions

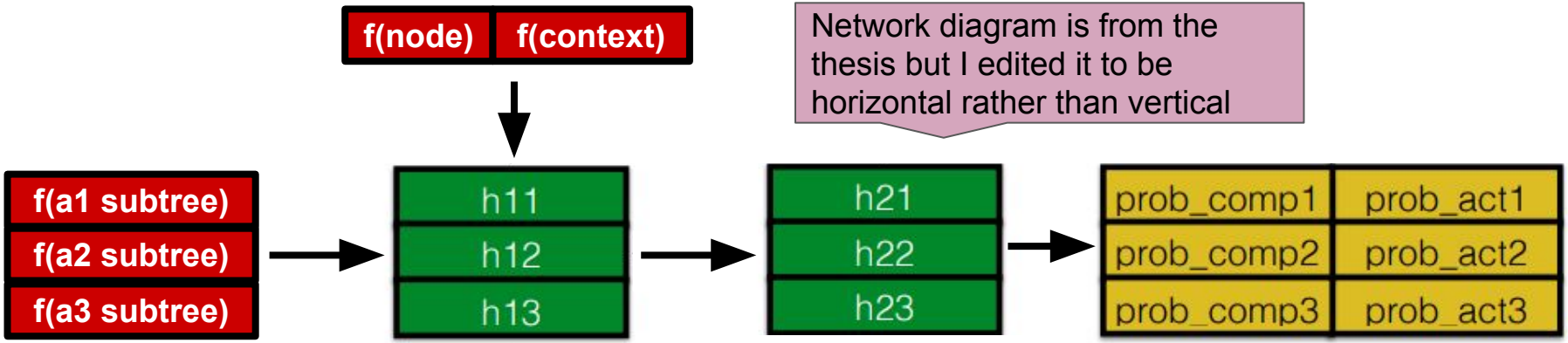
- `num_visits(T)`
- `all_done(T)`
- `average_rollout_value(T)`
- `average_estimate_value(T)`
- `p_over_n+1(T)`
- `minimax(T)`

Context Functions

- `depth(C)`
- `alpha(C)` *utility lower bound*
- `beta(C)` *utility upper bound*

Policy Network g_θ

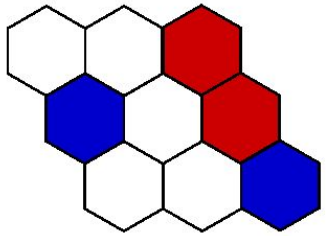
$$\pi_\theta(T) = g_\theta(f_{\text{fixed}}(T))$$



Training Algorithm: TRPO with Generalized Advantage Estimation

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel.
Highdimensional continuous control using generalized advantage estimation. ICLR, 2016

Experiments



Hex with a L x L board, maximum of n simulated actions per move

Average win rate against a UCT baseline

Random Initialization

	n=10	n=20	n=50	n=100
L=3	**0.60**	**0.44**	-0.59	-0.60
L=5	**0.50**	**0.48**	-0.42	-0.91
L=7	**0.35**	**0.50**	-0.63	-0.54

Initialized to UCT

	n=10	n=20	n=50	n=100
L=3	**0.58**	**0.47**	**0.47**	**0.33**
L=5	**0.70**	**0.60**	**0.51**	-0.89
L=7	**0.37**	**0.47**	**0.36**	-0.45

Related Work

Metareasoning and Bounded Optimality

- Stuart J. Russell and Eric H.Wefald. Decision-theoretic control of search: General theory and an application to game-playing.
- Stuart Russell. *Rationality and intelligence: A brief update.* In Fundamental Issues of Artificial Intelligence, pages 7–28. Springer, 2014
- Eric Horvitz. *Models of continual computation.* In AAAI/IAAI, pages 286–293, 1997.

Monte-Carlo Tree Search

- Levente Kocsis and Csaba Szepesvari. *Bandit Based Monte-Carlo Planning.* ECML, 2006.