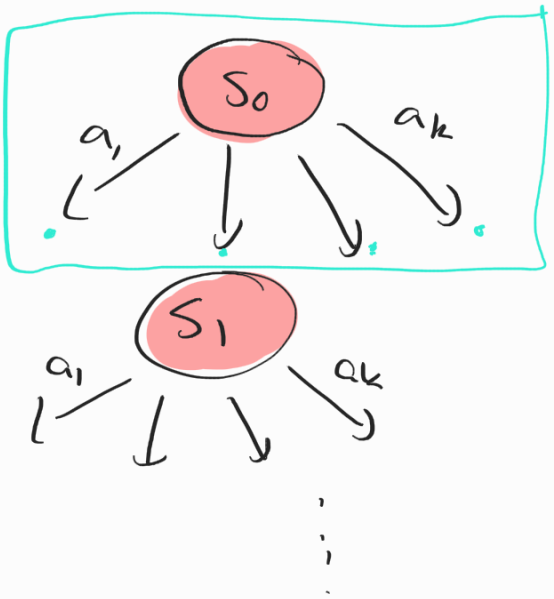
② Monte Carlo Tree Search (MCTS)

UCT → extension of UCB

↳ Upper Confidence (Bounds) for Trees



← a multi-armed bandit question

We are essentially solving nested multi armed bandit problems !

What do we learn ?

- v · values ≡ state value function
- q · values ≡ action state value function

Data structure :

Plays : $S \rightarrow \mathbb{N}$

$s \in S$ Plays(s) how many plays I have simulated that contained s (?)

Values: $S \to \mathbb{R}$

Values(s) our current estimates for $val_*(s)$

Play(s): return an action to play from s.

(1) **Naive version**:

We could simply "trust" our value function:

$$\text{play} \quad \underset{a \in A}{\operatorname{argmax}} \underbrace{\sum_{s', r} \Delta(s,a)(s',r)\,(r + \gamma \text{val}(s'))}_{q(s,a)}$$

(2) Actual algorithm:

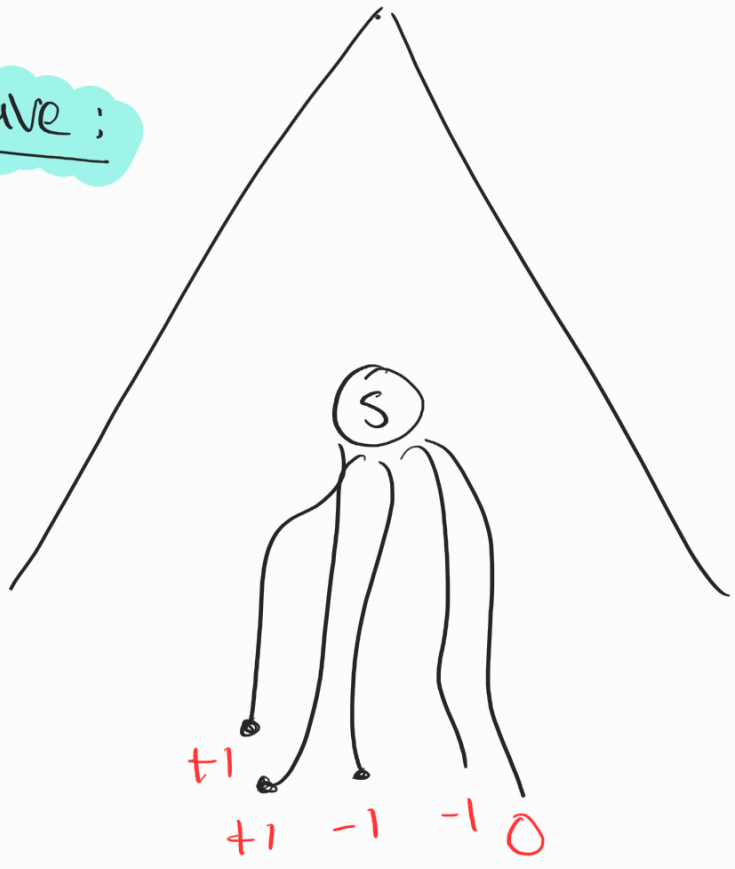
(1) **Run a number of simulations from s** (set a timer)

(2) Now I can trust my value

function, and do as in the naive version above for choosing the action.

Simulation(s) : Repeat { · construct a trajectory from s · update our statistics using the trajectory }

Naive :



val(s) updated using (+1, +1, -1, -1, 0)

Actual MCTS

4 phases :

as long as we have value estimates for all successors

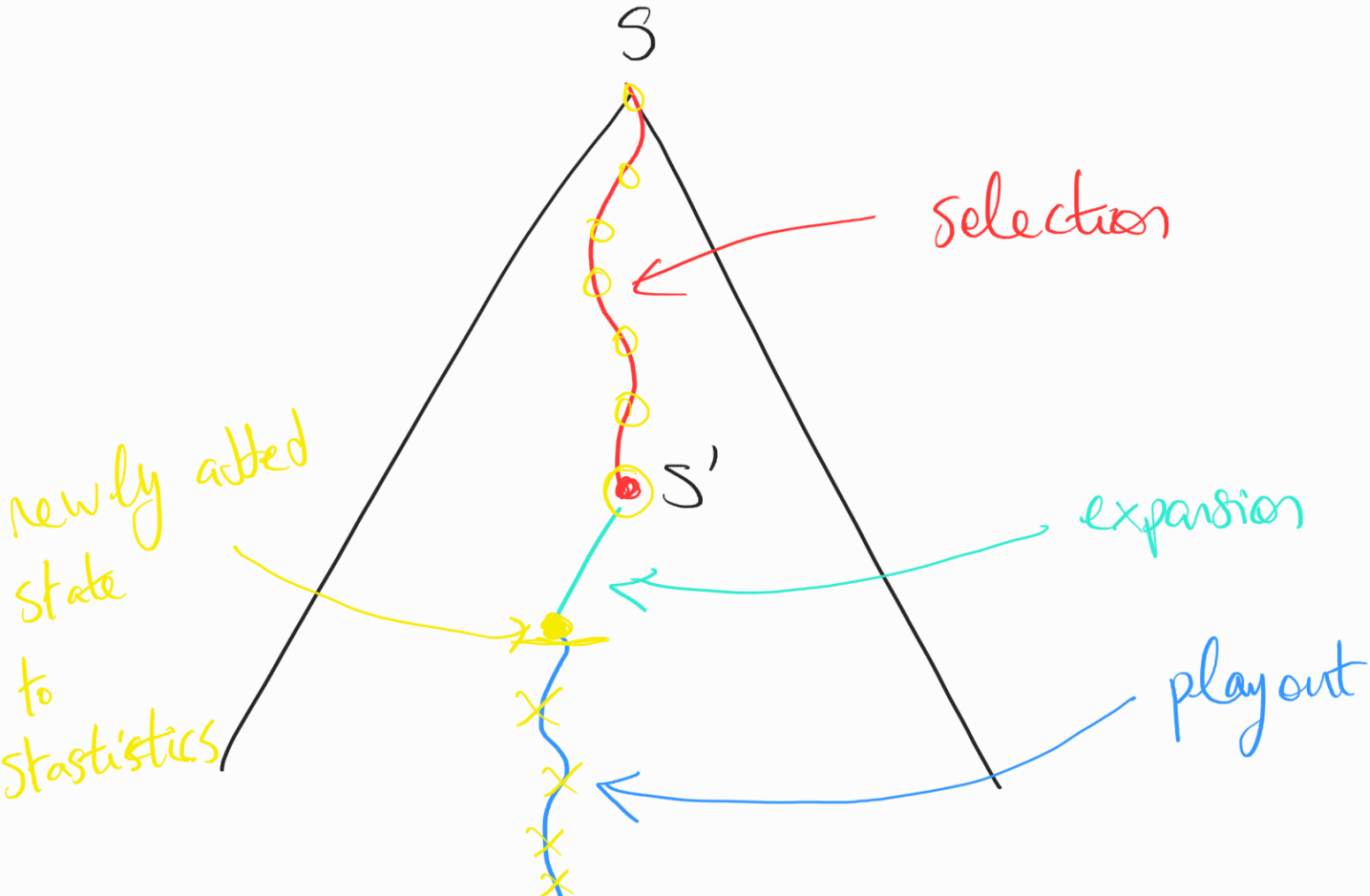
[SELECTION : as long as possible, use the statistics (values)

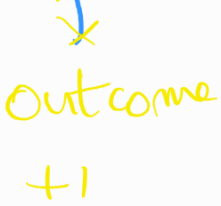
using ε-greedy or UCB

EXPANSION:
pick a random successor for which we don't have value estimate

PLAYOUT:
play randomly until trajectory is complete

BACKPROPAGATION
update values of all states visited DURING SELECTION AND EXPANSION





$S \in$

$Plays(s) \quad += 1$

$Values(s)$
- average
- step size
- temporal difference