Provably Efficient Long-Horizon Exploration in Monte Carlo Tree Search through State Occupancy Regularization

Liam Schramm, Abdeslam Boularias

RUTGERS THE STATE UNIVERSITY OF NEW JERSEY

Definitions

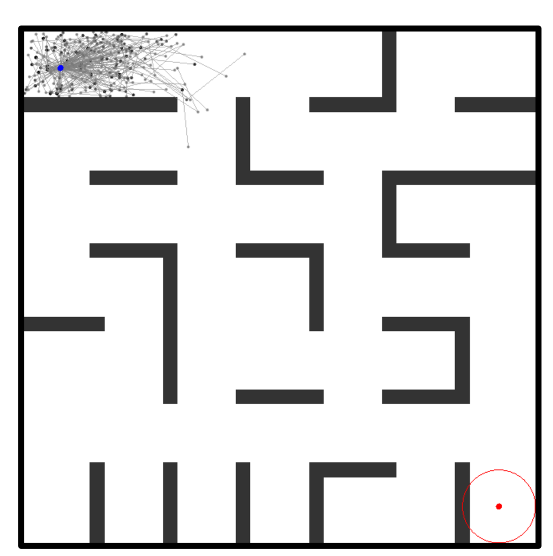
- S – State space
- $U(S)$ – Uniform distribution on state space
- π – Policy
- $\hat{\pi}$ – Empirical policy
- $d^{\pi}(n)$ – State occupancy measure/Probability of expanding node
- $\rho(d^{\pi})$ – Density estimate of tree in space
- $E[R]$ – Expected Reward
- $D_f(\pi || \pi_\theta)$ – f-divergence between π and neural net policy π_θ
- $\mathcal{V}(n)$ – Expected reward of path through n

Connection to RRT and Count-based exploration

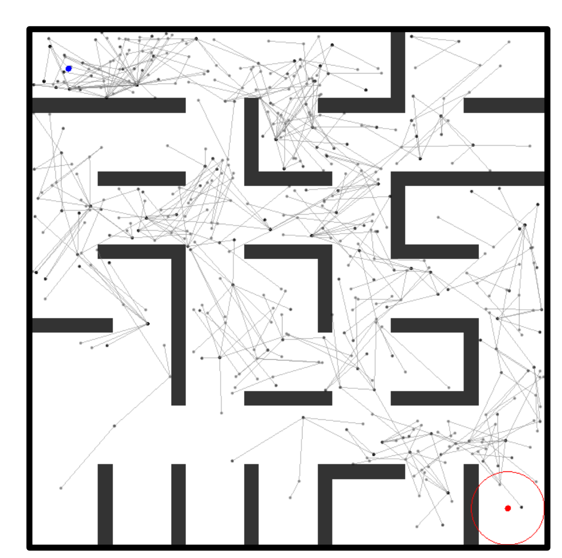
- AlphaZero equivalent to optimizing $E_{\hat{\pi}}[R] - \lambda\, KL(\hat{\pi}\, || \pi_\theta)$
- We regularize $\rho(d^{\pi})$, density of tree in state space
- RRT = $argmin_{d^{\pi}}\, KL\,(\rho(d^{\pi}) || U(S))$
- Count-based exploration ≈ $argmax_{\widehat{d^{\pi}}}\, E_{\widehat{d^{\pi}}}[R] - \lambda\, D_f(\rho(\widehat{d^{\pi}}) || U(S))$
- Propose Volume-MCTS, optimizing $E_{d^{\pi}}[R] - \, KL\,(\rho(d^{\pi}) || Uniform(S))$
- Solve analytically,
- $d^{\pi^*}(n) = \frac{\lambda Vol(n)}{\alpha - \mathcal{V}(n)}$
  - $Vol(n)$ – Volume of n's Voronoi region

- Regularizing state-occupancy measure instead of policy makes MCTS exponentially faster at exploration
  - $O(\exp(N)) \rightarrow O(N^2)$
- Beat MCTS on range of hard exploration tasks
- RRT and count-based exploration equivalent to MCTS with state-occupancy regularization

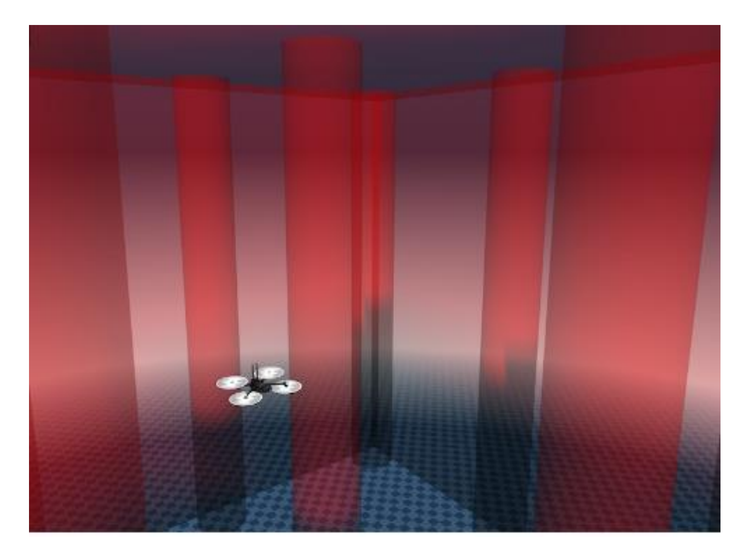
Experiments



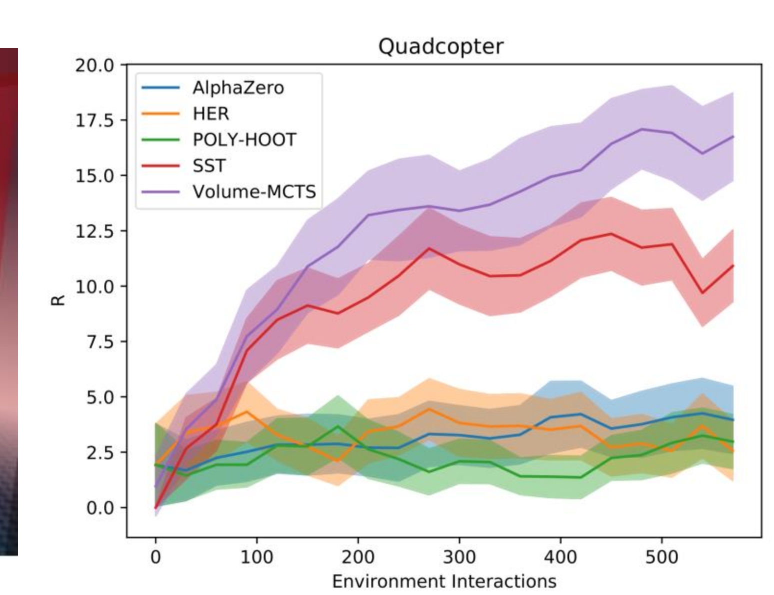
AlphaZero



Volume-MCTS



Quadcopter environment



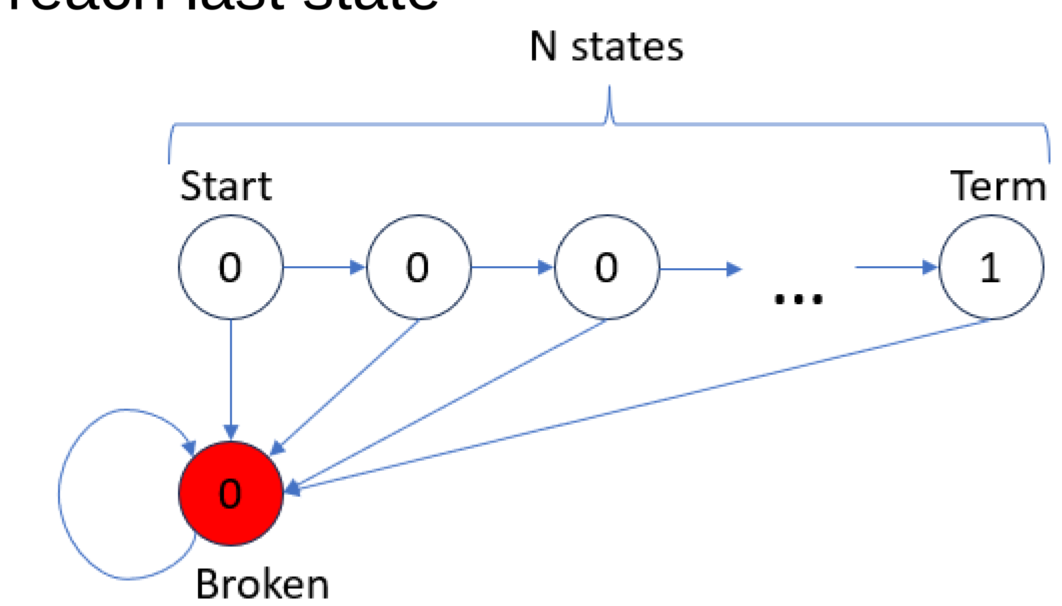
Quadcopter Results

- Volume-MCTS covers entire space, while AlphaZero stays close to starting location
- Volume-MCTS (purple) outperforms both AlphaZero and sampling-based motion planners on quadcopter environment

Theoretical guarantees

Thm 1: If a region with radius δ is reachable in N steps by a path with tolerance of σ, then with probability > 0.5 it will be reached in less than $c^2(1-\gamma)^2\left(\frac{1}{2}\, N\, \left|\mathcal{B}_{\frac{\delta}{5}}\right| \sigma d^A + 1\right)^2$ steps

"Tightrope" problem. MCTS takes $O(2^N)$ to reach last state



MCTS explores both actions equally.
Each state visited ½ as often as previous
$O(2^N)$ time needed to reach last state.

Additional details

Open to work