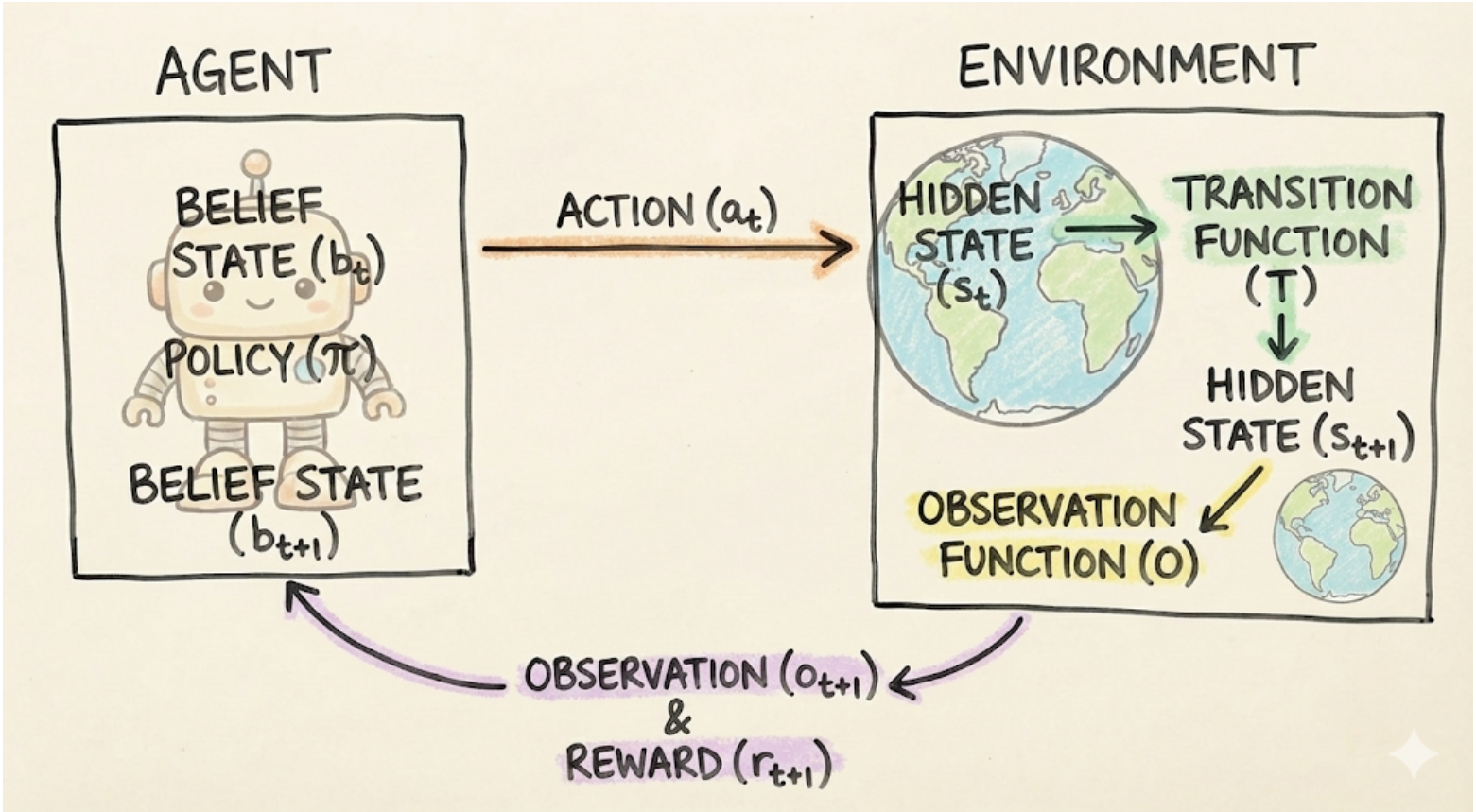
# Action-Gradient Monte Carlo Tree Search for Non-Parametric Continuous (PO)MDPs

Idan Lev-Yehudi Michael Novitsky Moran Barenboim Ron Benchetrit Vadim Indelman

ANPL
Autonomous Navigation and Perception Lab

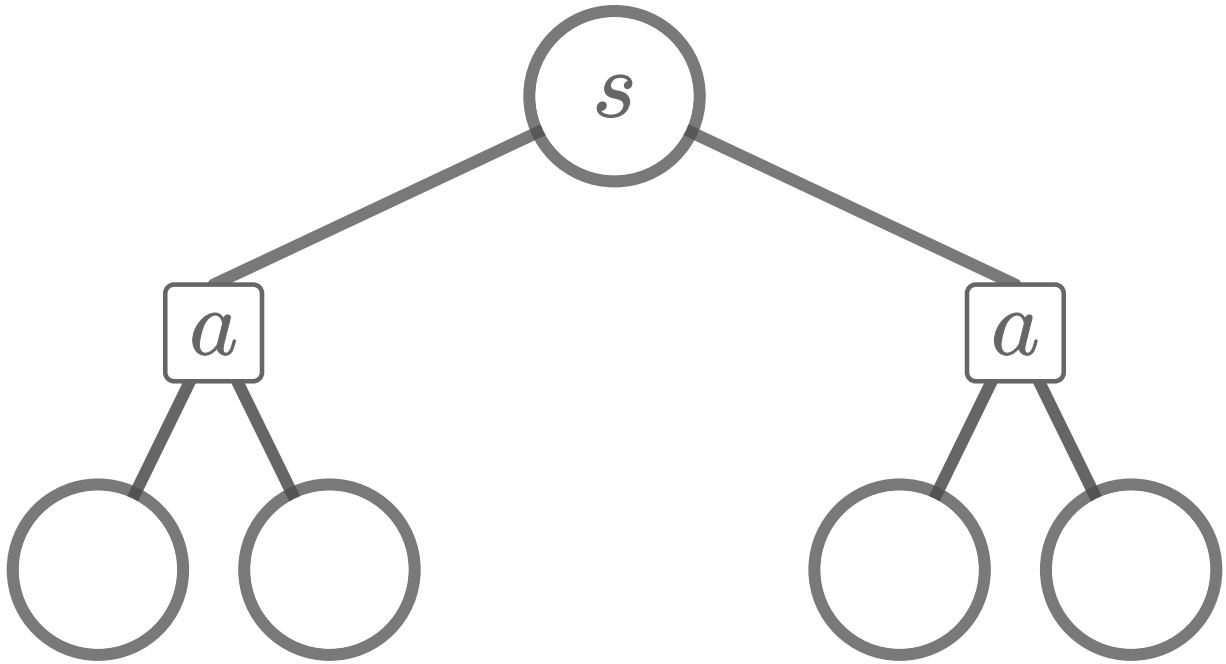
TECHNION
Israel Institute of Technology

# POMDP Introduction



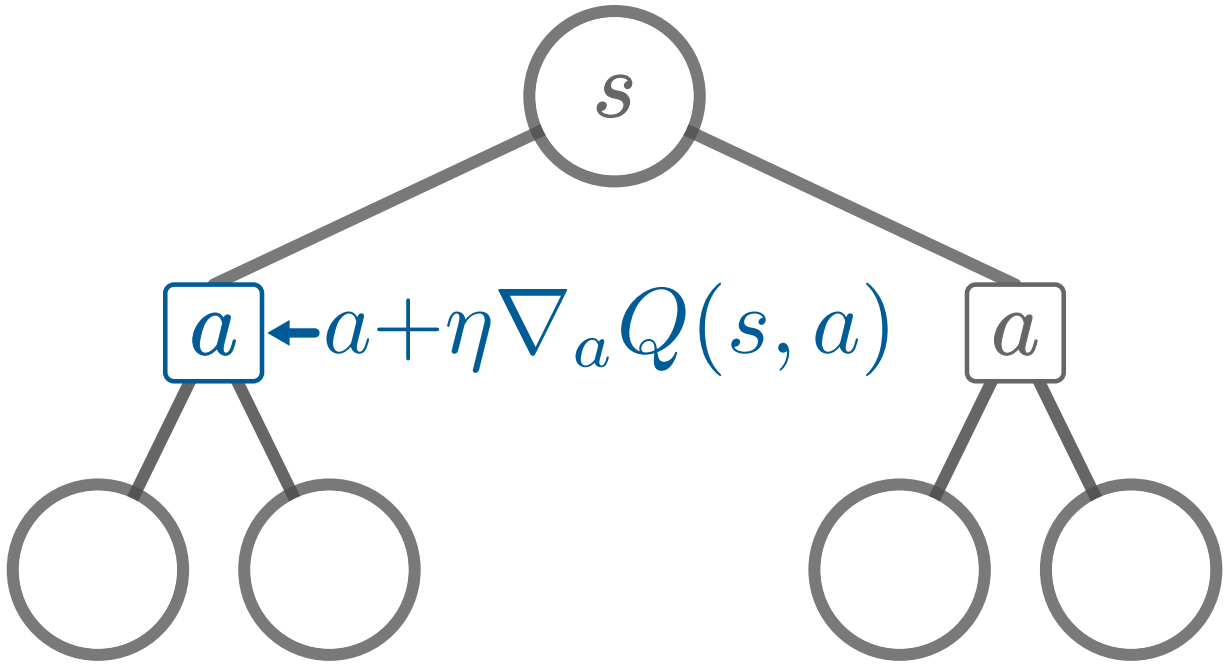

Robots reason over continuous states, actions and observations

# MCTS in Continuous Domains




- Continuous actions $\Rightarrow$ progressive widening samples *finitely many* branches
- Each branch carries its own $Q(s, a)$ estimate
- Once sampled, an action never moves

# MCTS in Continuous Domains




- Continuous actions $\Rightarrow$ progressive widening samples *finitely many* branches
- Each branch carries its own $Q(s,a)$ estimate
- Once sampled, an action never moves
- **AGMCTS:** refine the action itself, $a \leftarrow a + \eta \nabla_a Q(s,a)$

# Action-Gradient MCTS: Global Search, Local Refinement

## MDP Action Score Gradient

$$\nabla_a Q_t^\pi(s,a) = \mathbb{E}_{s'|s,a}\Big[\nabla_a \log p_T(s' \mid s,a) \cdot \big(r(s,a,s') + \gamma V_{t+1}^\pi(s') - V(s)\big) + \nabla_a r(s,a,s')\Big]$$

**Transition score**
Favor high-return successors

**Direct reward**
Immediate reward gradient

# Action-Gradient MCTS: Global Search, Local Refinement

## MDP Action Score Gradient

$$\nabla_a Q_t^\pi(s,a) = \mathbb{E}_{s'|s,a}\Big[\nabla_a \log p_T(s' \mid s,a) \cdot \big(r(s,a,s') + \gamma V_{t+1}^\pi(s') - V(s)\big) + \nabla_a r(s,a,s')\Big]$$

**Transition score**
Favor high-return successors

**Direct reward**
Immediate reward gradient

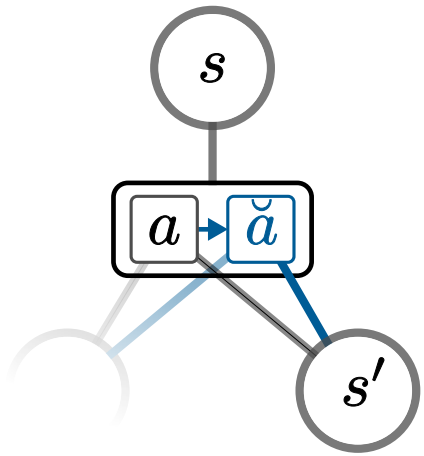
## POMDPs via Particle-Belief MDPs

MDP: $p_T(s' \mid s,a) \implies$ POMDP: $\prod_{j=1}^{J} p_T(s'^{-,j} \mid s^j, a)$

Same formula, adjusted likelihood

# Multiple Importance Sampling Trees

- Actions shift $\Rightarrow$ children were sampled under *old* actions
- Stale $Q(s,a) \Rightarrow$ reweight by Multiple Importance Sampling



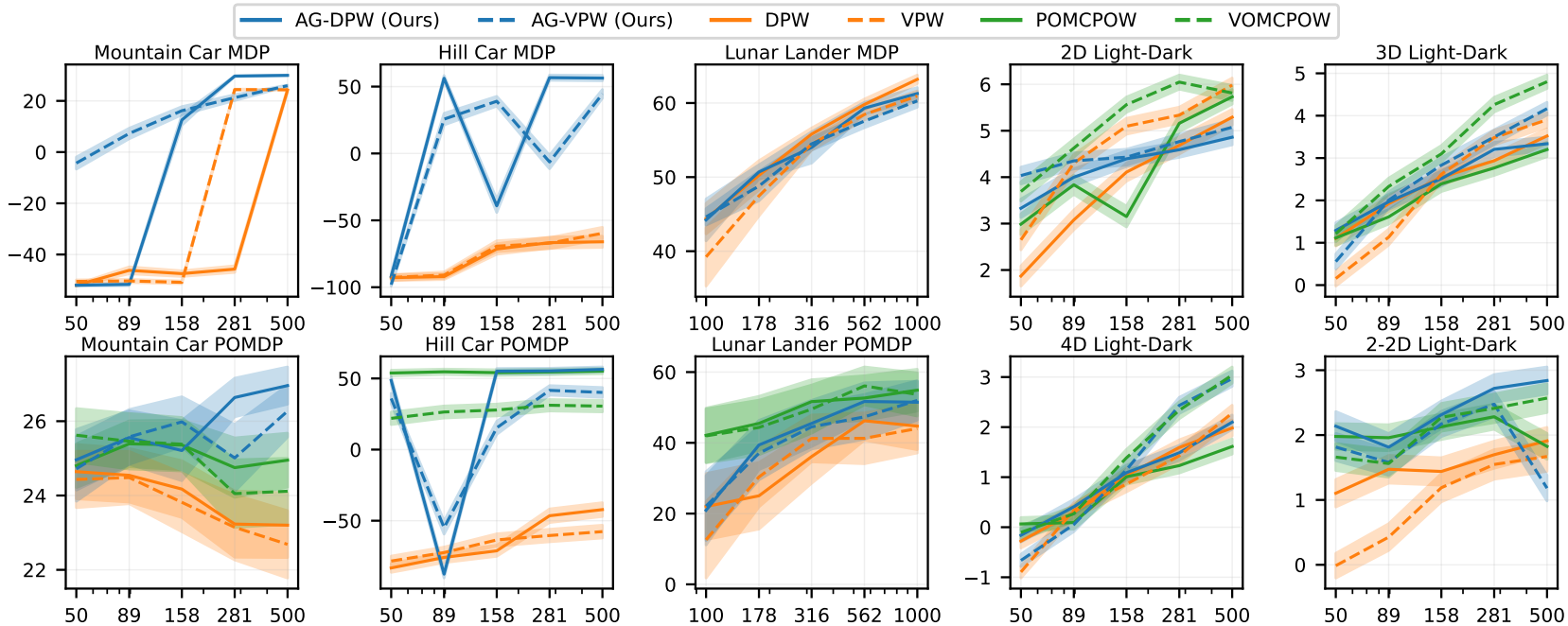

## MIS Tree estimator

$$V_f(s,a) = \frac{\sum_i N_i \, \rho_a^i(s'^i) \, V(s'^i)}{\sum_i N_i \, \rho_a^i(s'^i)}, \qquad \rho_a^i(s') = \frac{p_T(s' \mid s,a)}{p_T(s' \mid s, a^i_{\text{prop}})}$$

**Reuse correction** every successor remembers the action that proposed it

**Action update** $a \to \breve{a}$ in $O(|\mathcal{S}_{sa}|)$ • **Backprop** of $Q, V$ in $O(1)$

# Simulation Results




Average return, 1000 shared seeds, $\pm 2$ standard errors; x-axis: simulation budget

- **6/10** scenarios: an AG variant beats its matched baseline
- **5/7** POMDPs: best or tied; largest gains on Mountain/Hill Car
- **Cost:** $10$–$35\times$ (MDP) / $2$–$20\times$ (POMDP) overhead per plan

# Takeaway

- **Global search + local refinement** inside one tree
- MIS Tree keeps value estimates **consistent** after actions move
- **Best fit:** small action changes strongly affect states or rewards
- **Requires:** transition densities, tractable for smooth models

# Thanks for listening!

**Paper Link**