Safe Multi-Agent Navigation guided by Goal-Conditioned Safe Reinforcement Learning

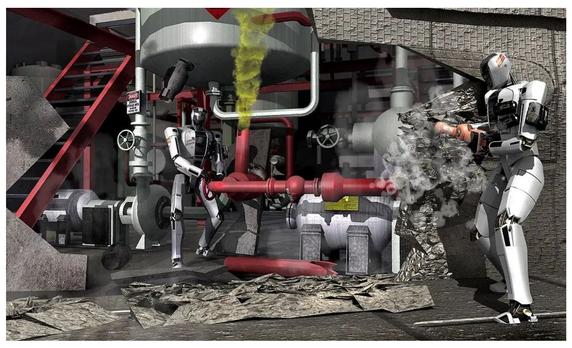
Meng Feng*, Viraj Parimi*, Brian Williams
Massachusetts Institute of Technology

SCAN ME

Enabling scalable coordination of multi-agent systems relying on image-based observations to solve long-horizon safe navigation tasks, while balancing goal achievement with risk mitigation.

Motivation

- Autonomous Disaster Response
  - Time-critical missions
  - Safety-critical missions
- Time-critical missions demand multi-agent coordination for rapid task completion
- Safety-critical missions require navigating risks while meeting safety thresholds



Goal: Coordinate multiple agents to reach their goals in a safe manner

Problem

How can we advance goal-conditioned RL to enable safer deployment in multi-agent, sparse-reward, long-horizon navigation problems?

Problem Features

- Multiple agents
- Long-horizon visual navigation
- Avoid risky behaviors like getting too close to obstacles
- Sparse-reward settings

Key Insights

- Goal-conditioned RL handles local goal achievement with safe policy
- Planning handles multi-agent coordination to ensure global goal achievement

Approach

Goal-Conditioned Risk-Aware Policy

$$\pi^* = \arg\max_{\pi \in \Pi} J_r(\pi_\theta) \quad \text{s.t } J_c(\pi_\theta) \leq \Delta$$

$$\pi^*, \lambda^* = \arg\min_{\lambda \geq 0} \arg\max_{\pi \in \Pi} J_r(\pi_\theta) - \lambda(J_c(\pi_\theta) - \Delta)$$

$$J_r(\pi_\theta) = \mathbb{E}_{\substack{s_1 \sim \rho(s),\ a_t \sim \pi(a_t|s_t,s_g),\\ s_{t+1} \sim p(s_{t+1}|s_t,a_t)}} \left[\sum_{t=1}^{T} r(s_t, a_t, s_g) \right] \quad J_c(\pi_\theta) = \mathbb{E}_{\substack{s_1 \sim \rho(s),\ a_t \sim \pi(a_t|s_t,s_g),\\ s_{t+1} \sim p(s_{t+1}|s_t,a_t)}} \left[\sum_{t=1}^{T} c(s_t, a_t, s_g) \right]$$

Graph Generation

$$d_\pi(s_i, s_j) = Q_d(s_i, \pi(s_i, s_j), s_j) \qquad D \sim \mathrm{Cat}(N_D, p_\pi(d \mid s, a, s_g))$$

$$r_\pi(s_i, s_j) = Q_r(s_i, \pi(s_i, s_j), s_j) \qquad R \sim \mathrm{Cat}(N_R, p_\pi(r \mid s, a, s_g))$$

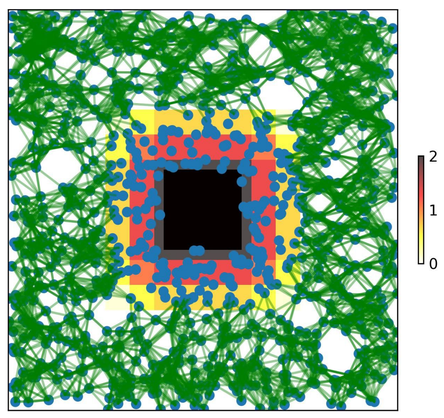
$$\mathcal{G} \triangleq (\mathcal{V}, \mathcal{E}, \mathcal{W}_d, \mathcal{W}_r)$$

where $\mathcal{V} = \mathcal{B}$

$$\mathcal{E} = \mathcal{B} \times \mathcal{B} = \left\{ e_{s_i \to s_j} | s_i, s_j \in \mathcal{B} \right\}$$

$$\mathcal{W}_d\left(e_{s_i \to s_j}\right) = \begin{cases} d_\pi(s_i, s_j) & \text{if } d_\pi(s_i, s_j) < d_{\max} \\ \infty & \text{otherwise} \end{cases}$$

$$\mathcal{W}_r\left(e_{s_i \to s_j}\right) = \begin{cases} r_\pi(s_i, s_j) & \text{if } r_\pi(s_i, s_j) < r_{\max} \\ \infty & \text{otherwise} \end{cases}$$



Goal Sampling

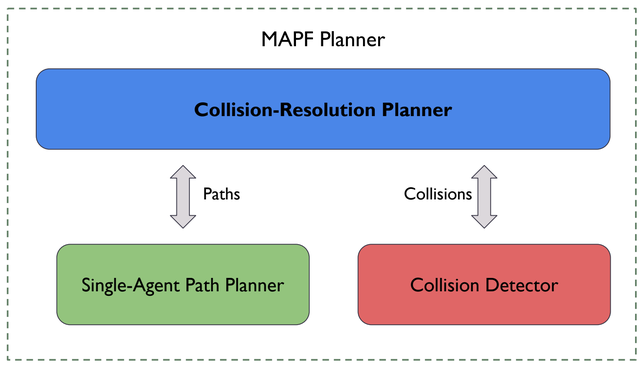
Algorithm 1 Self-Sampling and Training of Goal-Conditioned Actor Critic

Inputs: Environment E, Goal-Conditioned Policy π, Associated Q-functions for rewards or auxiliary costs $Q(s, a, s_g)$, Desirable sample target for rewards or costs τ, Population size N, Number of training problems per batch K

Outputs: Updated $\pi(s, a, s_g)$, $Q(s, a, s_g)$

$\mathcal{P} \leftarrow$ Initialize an empty training set
for each batch **do**
  $\{(s_i, s_j)\}_N \leftarrow$ Randomly sample state pairs from E
  $\{v_{ij}\}_N \leftarrow Q(s_i, \pi(s_i, a, s_j), s_j)$
  $\{l_{ij}\}_N \leftarrow L^2$ distance to τ
  Find $\{(s_i, s_j)\}$ corresponding to lowest K $\{l_i j\}_N$
  Add the selected $\{(s_i, s_j)\}$ to the training set
  Train the agent with $\mathcal{P}$ until $\mathcal{P}$ is depleted

Multi-Agent Path Finding



Plan and Act

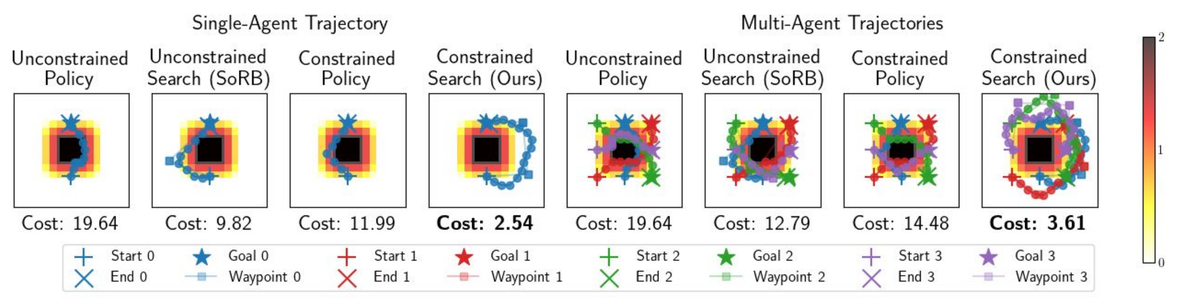
Inputs: Agents N, Current States $\mathbf{s}$, Goal States $\mathbf{s_g}$, Buffer of observations $\mathcal{B}$, Learned constrained policy π_c, Q-functions Q_d^π, Q_r^π of the unconstrained policy π

Output: Action a

if $N > 1$ **then**
  $s_{w_1}, \ldots \leftarrow \mathrm{CBS}(s, s_g, \mathcal{B}, Q_d^\pi, Q_r^\pi)$
else
  $s_{w_1}, \ldots \leftarrow \mathrm{SHORTEST\_PATH}(s, s_g, \mathcal{B}, Q_d^\pi, Q_r^\pi)$
if $d_\pi(s \to s_{w_1}) < d_\pi(s \to s_g)$ or $d_\pi(s \to s_g) > \mathrm{d_{max}}$ **then**
  $a \leftarrow \pi_c(a \mid s, s_{w_1})$
else
  $a \leftarrow \pi_c(a \mid s, s_g)$

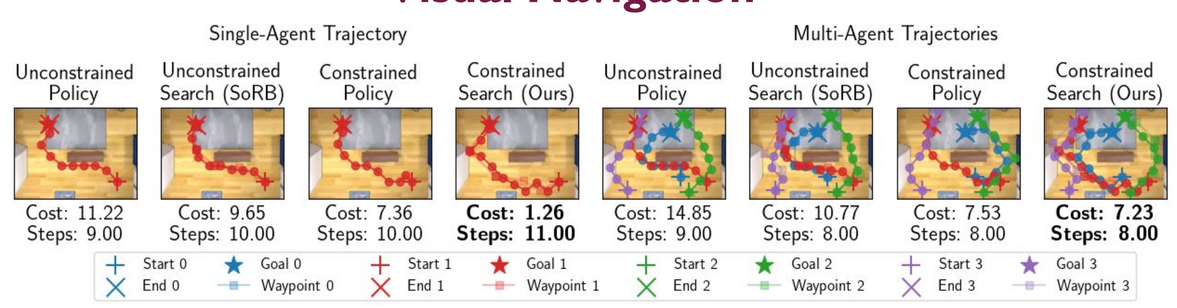
Experiments

2D Navigation



Problem Configurations		Methods (Cumulative Cost (Success Rate))			
Problem Type	Agents	Unconstrained Policy	Unconstrained Search (SoRB)	Constrained Policy	Constrained Search (Ours)
Easy	1	**0.46 ± 1.05 (100%)**	1.38 ± 2.62 (100%)	0.47 ± 1.07 (100%)	0.49 ± 1.06 (100%)
	5	N/A	6.39 ± 4.66 (100%)	N/A	**1.68 ± 1.21 (100%)**
	10	N/A	8.69 ± 4.05 (100%)	N/A	**2.24 ± 1.03 (100%)**
	20	N/A	10.51 ± 3.57 (100%)	N/A	**2.72 ± 0.80 (100%)**
Medium	1	1.58 ± 2.03 (100%)	2.30 ± 3.02 (100%)	**1.58 ± 2.01 (100%)**	1.59 ± 1.99 (100%)
	5	N/A	5.12 ± 3.65 (100%)	N/A	**3.09 ± 1.50 (100%)**
	10	N/A	7.27 ± 3.66 (98%)	N/A	**4.11 ± 1.48 (98%)**
	20	N/A	8.73 ± 3.73 (98%)	N/A	**4.78 ± 1.33 (98%)**
Hard	1	3.98 ± 4.40 (100%)	4.05 ± 4.19 (100%)	4.19 ± 4.23 (100%)	**3.09 ± 3.87 (100%)**
	5	N/A	8.58 ± 3.79 (100%)	N/A	**6.01 ± 1.96 (100%)**
	10	N/A	10.77 ± 3.51 (100%)	N/A	**7.24 ± 1.52 (100%)**
	20	N/A	11.96 ± 3.31 (100%)	N/A	**8.36 ± 1.58 (100%)**

Visual Navigation



Problem Configurations			Methods (Cumulative Cost (Success Rate))			
Map	Problem Type	Agents	Unconstrained Policy	Unconstrained Search (SoRB)	Constrained Policy	Constrained Search (Ours)
SC2 Staging 08	Easy	1	3.16 ± 9.40 (100%)	2.14 ± 2.22 (98%)	1.85 ± 1.80 (100%)	**1.28 ± 1.69 (100%)**
		5	N/A	4.32 ± 2.29 (98%)	N/A	**3.35 ± 1.48 (100%)**
		10	N/A	5.36 ± 2.60 (96%)	N/A	**4.02 ± 1.15 (100%)**
	Medium	1	4.98 ± 3.46 (98%)	5.15 ± 3.32 (98%)	4.77 ± 3.25 (100%)	**3.35 ± 3.23 (100%)**
		5	N/A	9.74 ± 11.13 (90%)	N/A	**6.18 ± 1.64 (98%)**
		10	N/A	14.23 ± 23.53 (84%)	N/A	**7.30 ± 1.69 (98%)**
	Hard	1	13.41 ± 4.00 (96%)	12.35 ± 5.10 (90%)	10.43 ± 3.07 (100%)	**9.62 ± 3.66 (100%)**
		5	N/A	21.23 ± 14.64 (84%)	N/A	**13.02 ± 2.38 (98%)**
		10	N/A	27.37 ± 19.05 (62%)	N/A	**13.87 ± 2.17 (92%)**
SC3 Staging 05	Easy	1	1.87 ± 1.65 (98%)	2.23 ± 1.85 (98%)	1.80 ± 1.58 (100%)	**1.25 ± 1.60 (100%)**
		5	N/A	3.71 ± 1.96 (98%)	N/A	**2.76 ± 1.36 (100%)**
		10	N/A	4.20 ± 1.73 (98%)	N/A	**3.32 ± 1.11 (100%)**
	Medium	1	5.06 ± 3.27 (100%)	4.87 ± 3.15 (100%)	4.38 ± 2.91 (100%)	**2.66 ± 2.32 (100%)**
		5	N/A	7.58 ± 2.16 (98%)	N/A	**5.88 ± 1.38 (100%)**
		10	N/A	8.83 ± 1.75 (98%)	N/A	**7.28 ± 1.53 (96%)**
	Hard	1	14.86 ± 4.9 (98%)	13.16 ± 4.24 (100%)	8.59 ± 2.21 (98%)	**6.47 ± 3.30 (98%)**
		5	N/A	17.79 ± 2.68 (96%)	N/A	**15.10 ± 2.19 (96%)**
		10	N/A	19.56 ± 2.66 (94%)	N/A	**16.71 ± 2.28 (94%)**
SC3 Staging 11	Easy	1	1.20 ± 1.56 (100%)	1.27 ± 1.35 (100%)	1.20 ± 1.55 (100%)	**0.78 ± 1.12 (100%)**
		5	N/A	2.32 ± 1.11 (100%)	N/A	**1.88 ± 0.90 (100%)**
		10	N/A	3.00 ± 1.07 (100%)	N/A	**2.54 ± 1.15 (100%)**
	Medium	1	3.66 ± 2.95 (98%)	4.19 ± 3.23 (98%)	3.40 ± 2.83 (98%)	**1.95 ± 2.24 (100%)**
		5	N/A	8.33 ± 2.37 (98%)	N/A	**5.93 ± 2.40 (96%)**
		10	N/A	10.23 ± 2.27 (96%)	N/A	**8.13 ± 2.18 (94%)**
	Hard	1	16.41 ± 6.50 (100%)	14.75 ± 5.43 (100%)	11.59 ± 3.60 (100%)	**10.93 ± 3.94 (100%)**
		5	N/A	18.67 ± 2.51 (96%)	N/A	**15.78 ± 2.16 (98%)**
		10	N/A	20.37 ± 2.41 (92%)	N/A	**16.94 ± 1.66 (92%)**

References

[1] B. Eysenbach, R. R. Salakhutdinov, and S. Levine, "Search on the replay buffer: Bridging planning and reinforcement learning," Advances in Neural Information Processing Systems, vol. 32, 2019.
[2] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in International conference on machine learning. PMLR, 2017, pp. 449–458.
[3] R. Stern, N. Sturtevant, A. Felner, S. Koenig, H. Ma, T. Walker, J. Li, D. Atzmon, L. Cohen, T. Kumar, et al., "Multi-agent pathfinding: Definitions, variants, and benchmarks," in Proceedings of the International Symposium on Combinatorial Search, vol. 10, 2019, pp. 151–158
[4] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," 2019.

Acknowledgements

This work was supported by the Defence Science and Technology Agency (DSTA). Any opinions, findings and conclusions or recommendations in this material are those of the author(s) and do not necessarily reflect the views of DSTA.