Multiple Plans are Better than One: Diverse Stochastic Planning

Mahsa Ghasemi, [1] Evan Scope Crafts, [2] Bo Zhao, [2, 3] and Ufuk Topcu [2, 4]

[1]Department of Electrical and Computer Engineering, [2]Oden Institute for Computational Engineering and Sciences
[3]Department of Biomedical Engineering, [4]Department of Aerospace Engineering and Engineering Mechanics
The University of Texas at Austin

MOTIVATION: THE VALUE OF DIVERSITY

- In group settings, behavioral diversity promotes complimentary skills that improve team performance
- Diversity encourages environment exploration
- **Diversity can be used to find solutions that satisfy unknown preferences (this work)**

DIVERSITY IN MDPS

- Setting: infinite horizon Markov decision process (MDP)
- Known objective:

$$\mathbb{E}_{\tau\sim\pi}\left[\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} r(s_t,a_t)\right]$$

- Want to find a set of policies that are both:
  - Representative: **diverse** and small
  - **Near-optimal** with respective known objective
- Need diversity measure to capture important properties of solutions *(behavior characterization)*
- Characterize policies by their state-action occupancy measures:

$$\rho^{\pi}(s,a) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\Pr(s_t = s, a_t = a|\pi)$$

- This characterization captures both the action choices and their resulting trajectory distribution
- We then use the **Jensen-Shannon divergence** to measure distance between the occupancy measures:

$$\mathrm{JSD}(p\|q) = \frac{1}{2}\mathrm{KL}(p\|m) + \frac{1}{2}\mathrm{KL}(q\|m),$$

$$\mathrm{KL}(p\|m) = -\sum_{x\in X} p(x)\log\left(\frac{m(x)}{p(x)}\right).$$

ACKNOWLEDGEMENTS

Grant support

- NIH: R00-EB027181
- DARPA: D19AP00004
- AFRL: FA9550-19-1-0169

PROBLEM FORMULATION

Find a set of k policies Π_k that have high reward and diversity:

$$R(\Pi_k) = \sum_{\pi\in\Pi_k}\mathbb{E}_{\tau\sim\pi}\left[\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} r(s_t,a_t)\right]$$

$$D(\Pi_k) = \sum_{\substack{\pi_i,\pi_j\in\Pi_k\\ i<j}}\mathrm{JSD}(\rho^{\pi_i}\|\rho^{\pi_j}).$$

SOLUTION APPROACH

- Use **tradeoff parameter** to address multi-objective problem:

$$\Pi_k^* = \underset{\Pi_k\in\Pi_{ss}^k}{\arg\max}\ \frac{1}{k}R(\Pi_k) + \frac{2\lambda}{k(k-1)}D(\Pi_k)$$

- Then reformulate using the **dual of the MDP linear program**:

$$\max_{\rho_{1:k}}\ \frac{1}{k}\sum_{i\in[k]}\langle\rho_i, r\rangle + \frac{2\lambda}{k(k-1)}\sum_{\substack{i,j\in[k]\\ i<j}}\mathrm{JSD}(\rho_i\|\rho_j)$$

subject to

$$\sum_{a\in A}\rho_i(s,a) = \sum_{s'\in S}\sum_{a'\in A}P(s|s',a')\rho_i(s',a') \quad \textit{for all}\quad i\in[k],\ s\in S,$$

$$\sum_{s\in S}\sum_{a\in A}\rho_i(s,a) = 1 \qquad \textit{for all}\quad i\in[k],$$

$$\rho_i(s,a)\geq 0 \qquad \textit{for all}\quad i\in[k],\ s\in S,\ a\in A$$

- Problem has linear constraints, non-linear and non-concave objective
- Solve using the **Frank-Wolfe algorithm**

CONVERGENCE GUARANTEE

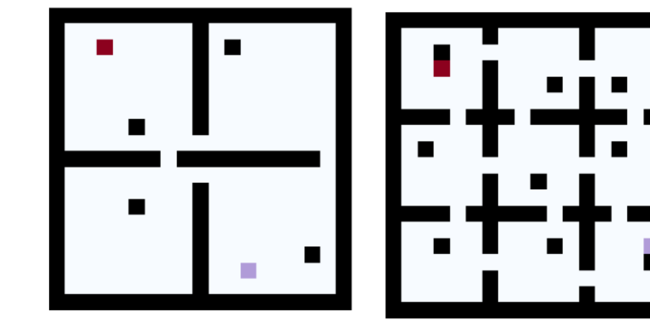
Non-asymptotic convergence guarantee:

$$g_{opt} \leq \frac{\max\{2(f_\delta^* - f(\rho_{1:k}^0)),\ \mathrm{diam}(\Delta_{M,\delta})^2 L\}}{\sqrt{T+1}}$$

Bounds first order optimality of solution after T iterates by term dependent on initial optimality gap, Lipschitz constant of gradient L, and the diameter of the feasible set

EXPERIMENTAL RESULTS

Navigation Problems



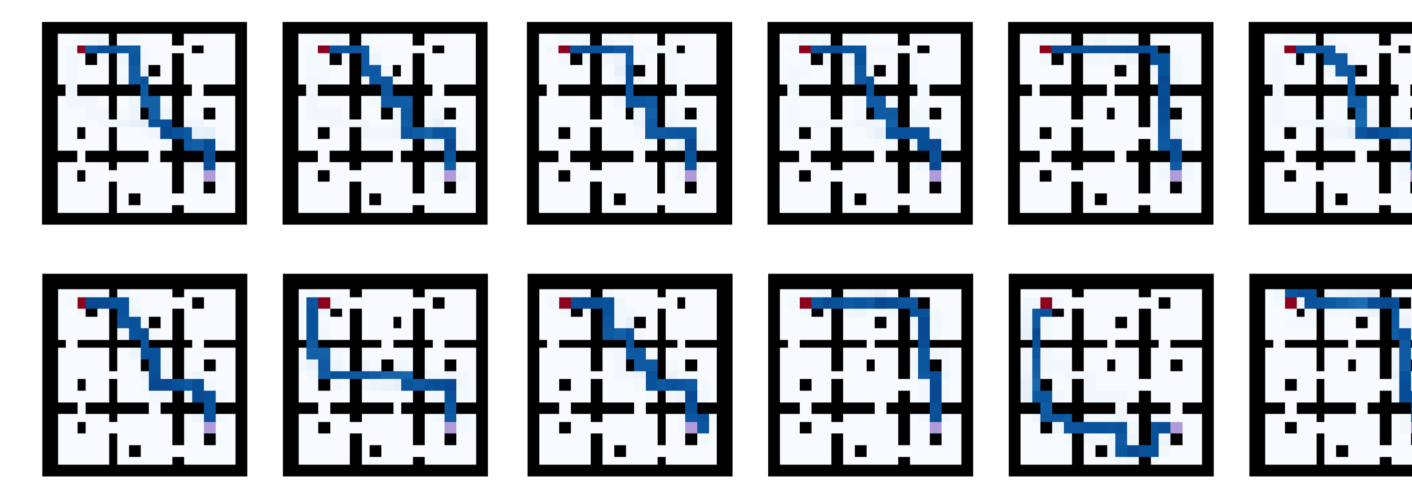
- Agent starts at red square
- Can transition to desired neighboring state with probability α
- Receives large reward for reaching goal (purple square) and loops back to start state
- Penalties for hitting walls and obstacles

Role of the Tradeoff Parameter

Results ($k = 6, \alpha = .95$):

$\lambda = 2$:

$\lambda = 8$:



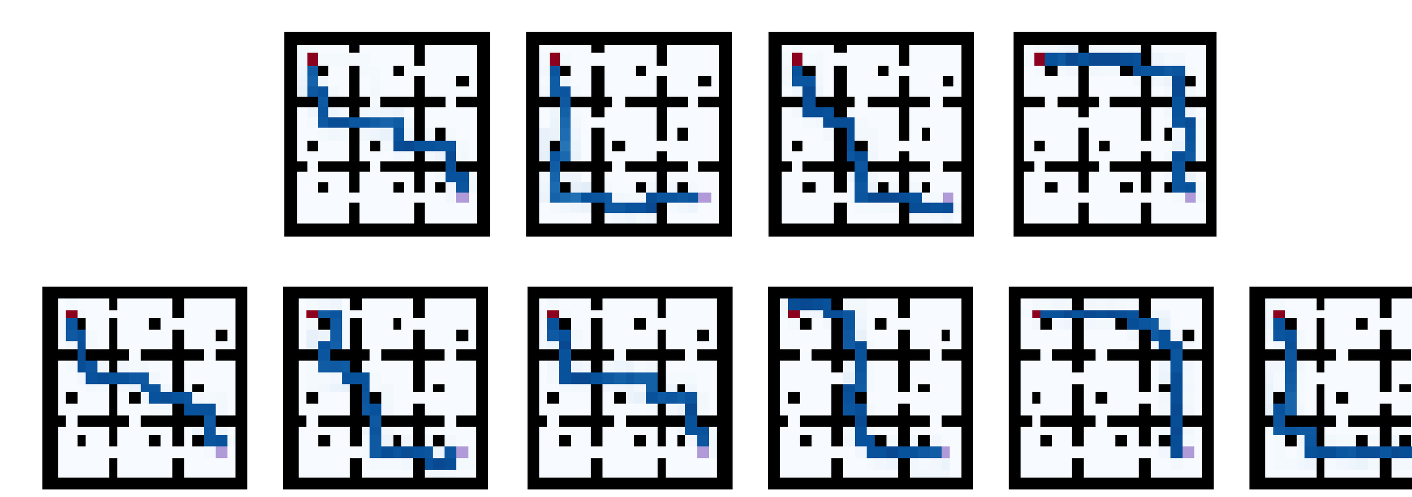
Tradeoff parameter has major effect on solution diversity

Finding more Policies

Results ($\lambda = 8, \alpha = .95$):

k = 4:

k = 6:



Diversity increases up to environment determined limit

CONCLUSIONS AND FUTURE WORK

- Considered stochastic planning problems with partially specified objectives
- Formulated a nonlinear optimization problem for finding a set of near-optimal and diverse policies and provided a solution algorithm with non-asymptotic convergence guarantees
- Demonstrated efficacy of approach using navigation problems
- **Future work:** use diversity to obtain effective collaboration among autonomous agents