Optimistic Linear Support and Successor Features as a Basis for Optimal Policy Transfer

Lucas N. Alegre[1] Ana L. C. Bazzan[1] Bruno C. da Silva[2]

[1]Institute of Informatics, UFRGS, Brazil [2]CICS, University of Massachusetts, MA

Contributions

We solve an open problem in transfer learning:

Optimal Policy Transfer

How to construct a set of policies, such that combining them directly leads to the optimal policy for *any* novel tasks?

Successor Features Optimistic Linear Support (SFOLS)

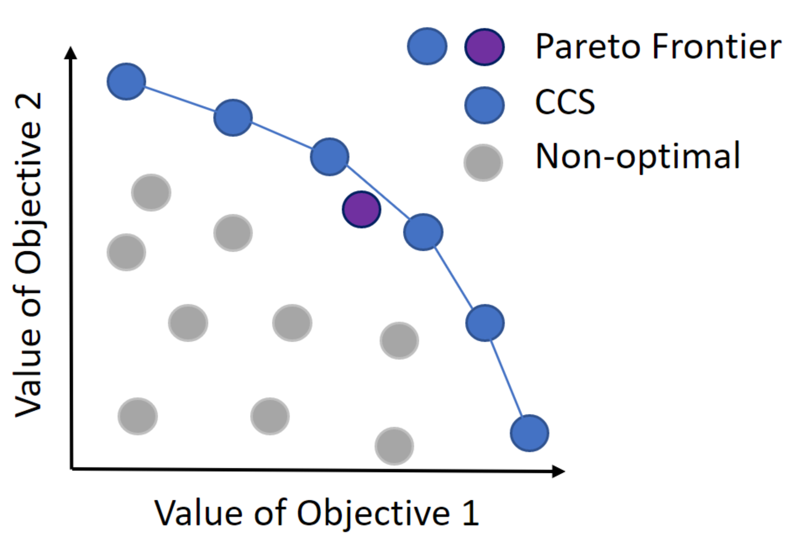
- extends theoretical guarantees from Transfer Learning and Multi-Objective RL (MORL)
- solves the Optimal Policy Transfer problem
- identifies optimal policies for *any* new tasks

MORL

Multi-objective reward:

$$\mathbf{r}: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto \mathbb{R}^m$$

Goal: find optimal policies for all convex combinations of the rewards/objectives



The optimal policy w.r.t. *any* convex combination of the rewards is in the **CCS**!

Policy Transfer

Reward linear in $\phi \in \mathbb{R}^d$:

$$r_{\mathbf{w}}(s,a,s') = \phi(s,a,s') \cdot \mathbf{w}$$

Successor Features (SFs)

$$\psi^\pi(s,a) \equiv \mathbb{E}_\pi\left[\sum_{i=0}^{\infty} \gamma^i \phi_{t+i} | S_t = s, A_t = a\right]$$

$$q^{\pi_i}_{\mathbf{w}}(s,a) = \psi^{\pi_i}(s,a) \cdot \mathbf{w}$$

Generalized Policy Improvement (GPI)

$$\pi^{\mathrm{GPI}}(s;\mathbf{w}) \in \arg\max_{a \in \mathcal{A}} \max_{\pi \in \Pi} q^\pi_{\mathbf{w}}(s,a)$$

No guarantees that π^{GPI} will be optimal for the new task!

Theoretical Results

Given an MDP, we define a MOMDP where each i-th objective/reward function is equal to the i-th reward feature:

$$R_i(s,a,s') \equiv \phi_i(s,a,s') \Rightarrow \mathbf{q}^\pi(s,a) \equiv \psi^\pi(s,a)$$

We can then define a CCS over SFs!

$$\mathrm{CCS} = \{\psi^\pi \mid \exists \mathbf{w} \text{ s.t. } \forall \psi^{\pi'}, \psi^\pi \cdot \mathbf{w} \geq \psi^{\pi'} \cdot \mathbf{w}\}$$

Intuitively: If we learn a set of policies whose SFs form a CCS, we can directly identify optimal policies for any new tasks!

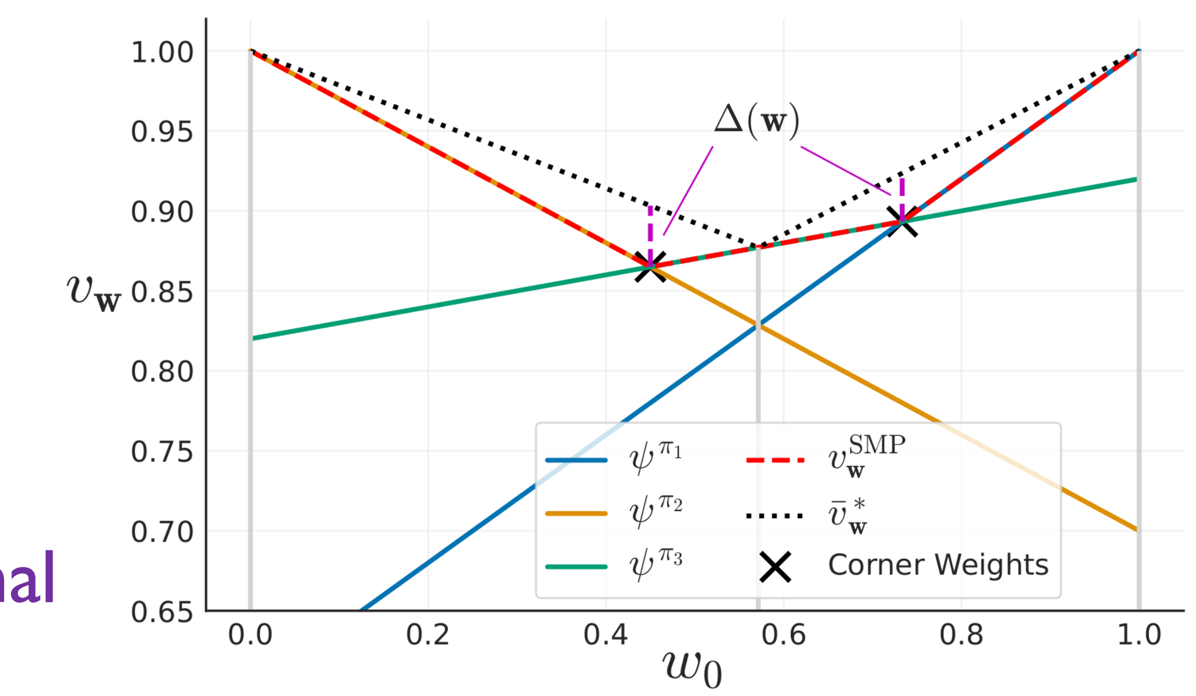
Theorem 3.2

If the SF set $\Psi = \{\psi^{\pi_i}\}_{i=1}^n$ of a policy set $\Pi \equiv \{\pi_i\}_{i=1}^n$ is a CCS, then, given any task $\mathbf{w} \in \mathcal{W}$, the GPI policy $\pi^{GPI}(s;\mathbf{w})$ is optimal with respect to $r_{\mathbf{w}}$.

SFOLS: SFs Optimistic Linear Support

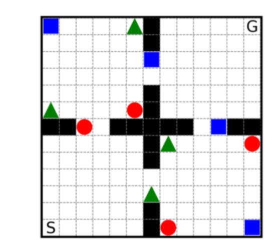
- Extension of the OLS algorithm (Roijers, 2016)
- Guaranteed to identify a CCS over SFs
  (key for solving the Optimal Policy Transfer problem)

- Iteratively learns policies for tasks defined by corner weights
- Corner weights: tasks with optimistic maximal improvement

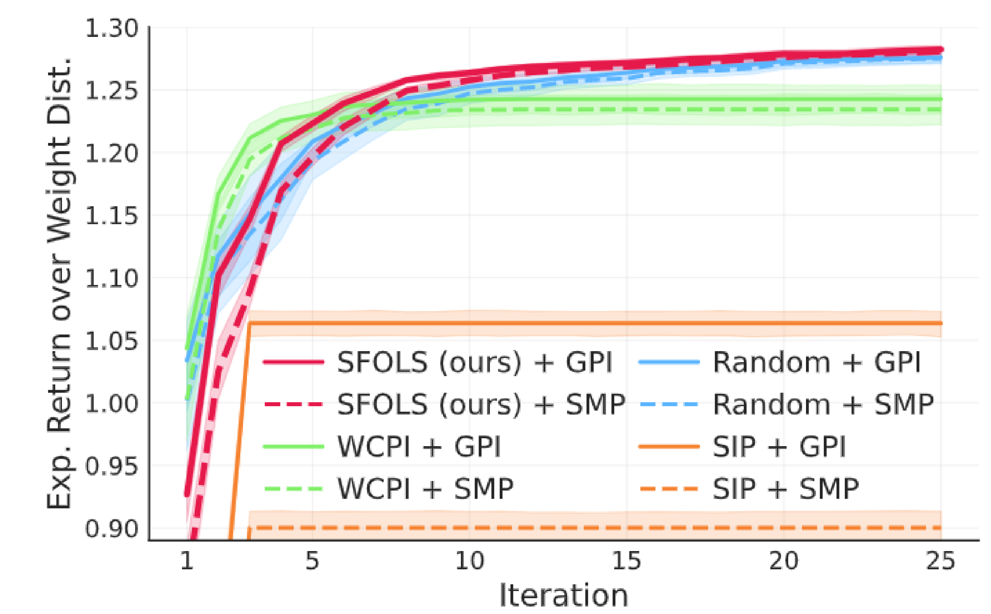


Experiments & Results

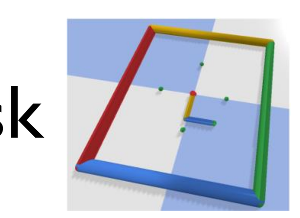
Four Room



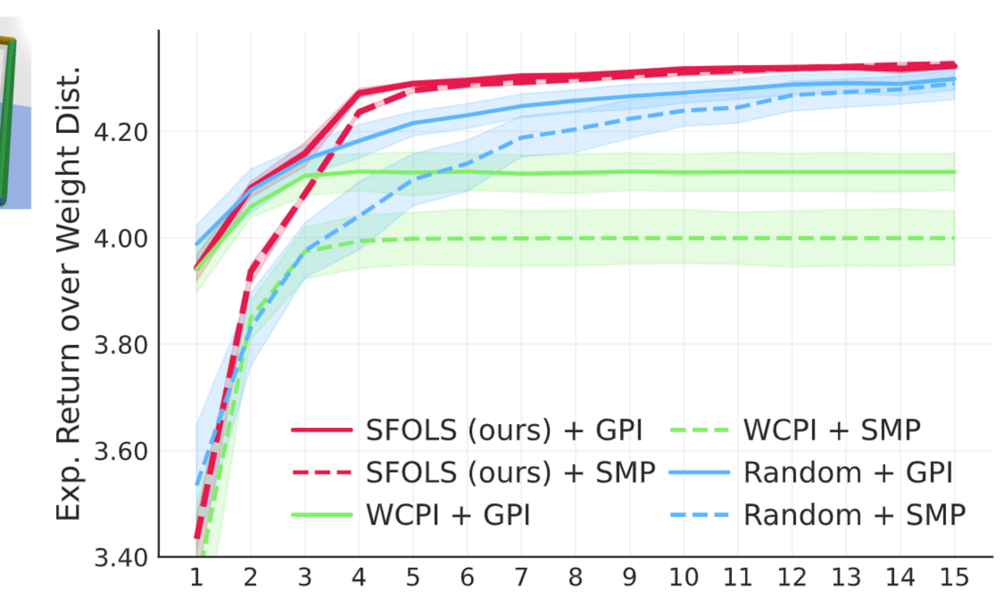
- SFOLS rapidly learns base policies that perform well over all tasks



Reacher Task

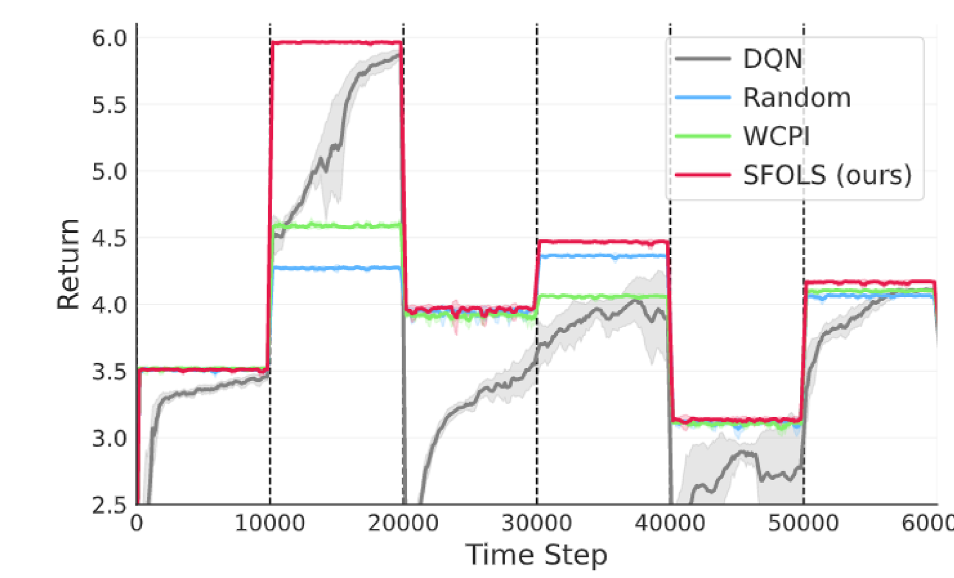


- Outperforms state-of-the-art competing algorithms and baselines



Lifelong RL

- Zero-shot policy transfer setting
- SFOLS *immediately* adapts to novel tasks



Discussion & Conclusions

- We formally characterize the connection between transfer learning and MORL
- SFOLS solves the Optimal Policy Transfer problem
  - identifies optimal policies for *any* new tasks
- Theoretical/empirical findings relevant to the MORL and transfer learning communities