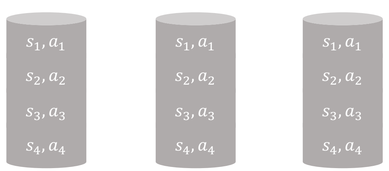
# Imitation Learning by Estimating Expertise of Demonstrators

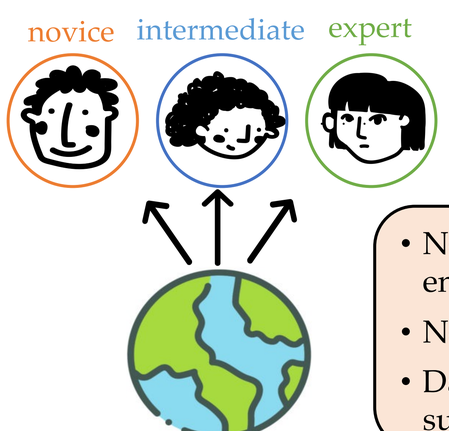
Mark Beliaev*[1], Andy Shih*[2], Stefano Ermon[2], Dorsa Sadigh[2], Ramtin Pedarsani[1]

[1]University of California, Santa Barbara, [2]Stanford University

**Problem:** Imitation learning on datasets with varying levels of suboptimality

**ILEED:** Leverage demonstrator identities to recover expertise levels and learn better policy!




- No access to environment
- No reward
- Data is suboptimal

## ILEED learns both policy and demonstrator expertise

**We have**

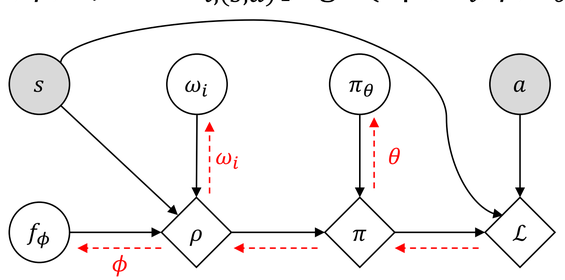
- states $s$
- actions $a$
- demonstrator identities $i$

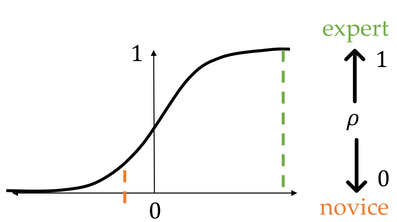
**We estimate**

- policy $\pi_\theta$
- state embedding $f_\phi$
- demonstrator embedding $\omega$

$$\mathcal{L}(\theta, \phi, \omega) = -\mathbb{E}_{i,(s,a)}[\log \pi(a|s, \omega_i, \phi, \pi_\theta)]$$

backpropagate through $\mathcal{L}$ to learn $(\theta, \phi, \omega)$




$$\rho_\phi(s, \omega_i) = \sigma(\langle f_\phi(s), \omega_i \rangle)$$

$$\pi(a|s, \omega_i, \phi, \pi_\theta) = \begin{cases} \text{discrete} & \rho_\phi(s, \omega_i)\pi_\theta(a|s) + \frac{1 - \rho_\phi(s, \omega_i)}{|A|} \\ \text{continuous} & \sum_{j=1}^{k} \alpha_j \mathcal{N}\left(a; \mu_j(s; \theta), \frac{\sigma_j(s; \theta)}{\rho_\phi(s, \omega_i)}\right) \end{cases}$$

## Minigrid

| | | Empty | | | Obstacles | |
|---|---|---|---|---|---|---|
| | $\rho$ | BC | GAIL | ILEED | BC | GAIL | ILEED |
| expert | | 0.81 | 0.96 | **0.97** | 0.18 | -0.82 | **0.91** |
| noisy | | 0.97 | 0.96 | 0.97 | 0.66 | -0.77 | **0.94** |
| | | 0.97 | 0.96 | 0.97 | 0.63 | -0.01 | **0.94** |
| | | 0.97 | 0.96 | 0.97 | 0.80 | -0.84 | **0.90** |

## Robomimic

| Dataset | BC-RNN | IRIS | ILEED (ours) |
|---|---|---|---|
| All | $\mathbf{78.0 \pm 4.3}$ | $52.7 \pm 5.0$ | $\mathbf{78.0 \pm 1.6}$ |
| Worse | $39.3 \pm 3.8$ | $38.7 \pm 0.9$ | $\mathbf{46.7 \pm 4.7}$ |
| Okay | $45.3 \pm 2.5$ | $42.0 \pm 3.3$ | $\mathbf{53.3 \pm 2.5}$ |
| Better | $66.0 \pm 2.8$ | $60.0 \pm 1.6$ | $\mathbf{72.7 \pm 3.8}$ |
| Worse-Okay | $55.3 \pm 0.9$ | $43.3 \pm 2.5$ | $\mathbf{59.3 \pm 3.8}$ |
| Worse-Better | $73.3 \pm 6.2$ | $56.7 \pm 3.4$ | $\mathbf{77.3 \pm 6.8}$ |
| Okay-Better | $74.0 \pm 2.8$ | $56.7 \pm 3.8$ | $\mathbf{77.3 \pm 0.9}$ |