f-Policy Gradients: A General Framework for Goal Conditioned RL using f-Divergences

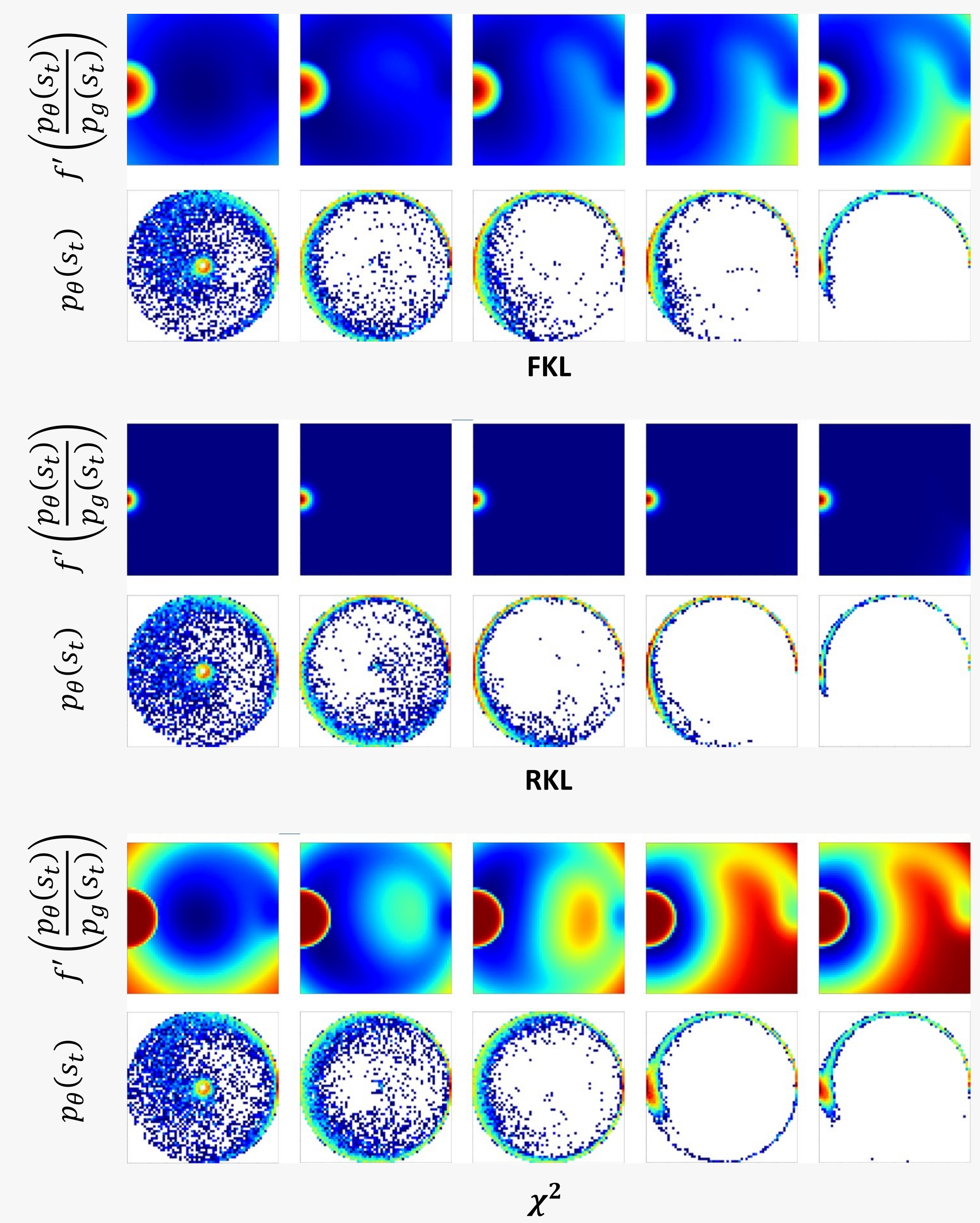
Siddhant Agarwal, Ishan Durugkar, Peter Stone, Amy Zhang

TEXAS The University of Texas at Austin
LARG Learning Agents Research Group The University of Texas at Austin
MIDI lab Machine Intelligence & Decision Making Through Interaction
Sony AI
NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

- Goal Conditioned Reinforcement Learning suffers from **sparse rewards**.
- One way to accelerate learning in sparse reward settings is using some form of reward shaping or augmenting sparse rewards with dense signals.
- Reward shaping requires domain information which is either provided by a human or learnt using expert trajectories and interactions with the environment, making it difficult to transfer to unknown environments.
- Reward shaping can be suboptimal and can lead to misalignment.
- We define a GCRL as a **distribution matching** problem as an alternate framework to the conventional reward maximization.
- Distribution matching techniques have been used in several imitation learning but they require a discriminator which are unstable and require coverage assumptions.
- **We present a general framework for GCRL that (a) produces optimal policies, (b) does not use a discriminator - stable and relaxes coverage assumptions, (c) works for any goal distribution, including Dirac, (and corresponding metric based shaping rewards) (d) provides dense signals for policy optimization even when the goal is not seen.**

Learning Signals



state-MaxEnt RL - A special case

The commonly used MaxEnt RL (π-MaxEnt RL) maximizes entropy of policy:

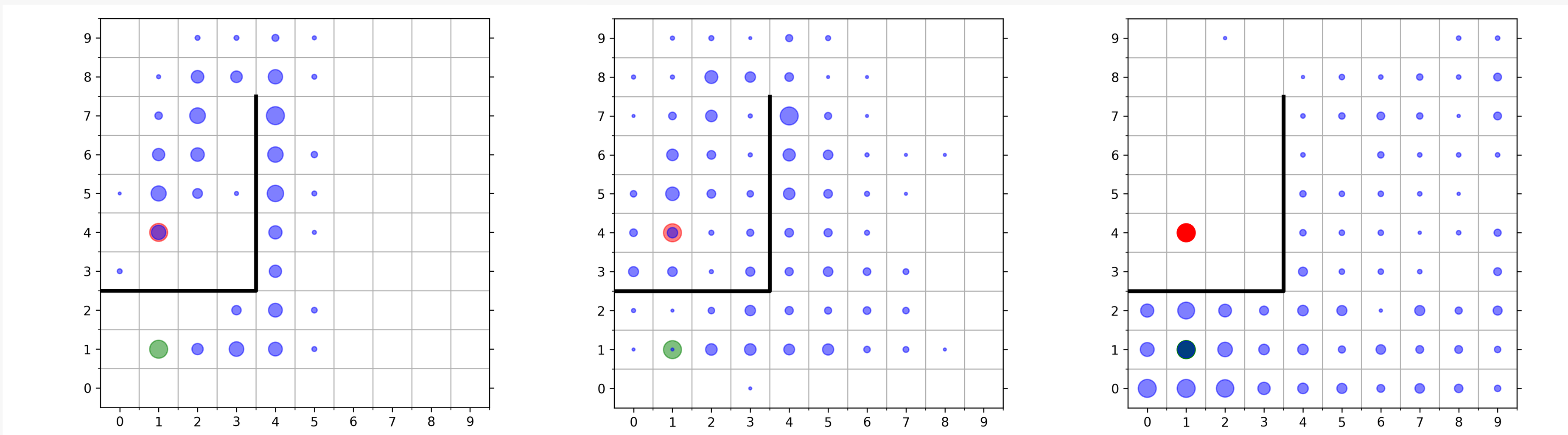
$$\max \mathbb{E}_{p_\theta(s)}[r(s)] + \mathcal{H}(\pi)$$

A special case of f-PG (using Forward KL): *state*-MaxEnt RL

$$\max \mathbb{E}_{p_\theta(s)}[\log p_g(s)] + \mathcal{H}(p_\theta(s))$$

Can be a Metric based shaping reward for some $p_g(s)$

Entropy of state visitation distribution



Reverse KL | *state*-MaxEnt RL | π-MaxEnt RL

Method

Use f-Divergence to characterize "distance" between distributions

Let $p_\theta(s)$ be the agent's state visitation distribution for a policy π_θ and $p_g(s)$ is the goal distribution

Minimize the following divergence:

$$J(\theta) = D_f(p_\theta(s)||p_g(s))$$

This means the visitation should be high at the goal states and as low as possible at all other states.

The objective produces optimal policies for some f-divergences (f-divergences with bounded $f'(\infty)$).

Optimize using gradients:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\left[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\right]\left[\sum_{t=1}^{T} f'\left(\frac{p_\theta(s_t)}{p_g(s_t)}\right)\right]\right]$$

Gradients of log probabilities

Dense learning signal

These gradients look like policy gradients but the term $f'\left(\frac{p_\theta(s_t)}{p_g(s_t)}\right)$ is not reward but simply a weight.

The value of $f'\left(\frac{p_\theta(s_t)}{p_g(s_t)}\right)$ will be low at (a) the goal and (b) states with low visitation probability.

Results