Aux-AIRL: End-to-End Self-Supervised Reward Learning for Extrapolating beyond Suboptimal Demonstrations

Yuchen Cui*, Bo Liu*, Akanksha Saran, Stephen Giguere, Peter Stone, Scott Niekum

The University of Texas at Austin
Department of Computer Science
College of Natural Sciences

Abstract

Real-world human demonstrations are often *suboptimal*. How to extrapolate beyond suboptimal demonstration is an important open research question.

In this ongoing work, we analyze the success of a previous state-of-the-art self-supervised reward learning method that requires four sequential optimization steps, and propose a simple end-to-end imitation learning method Aux-ARIL that extrapolates from suboptimal demonstrations without requiring multiple optimization steps.

Aux-AIRL

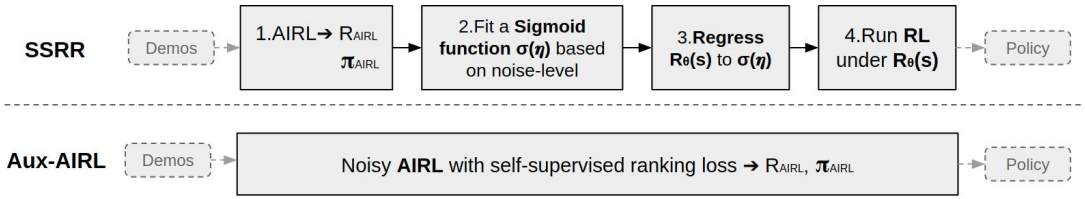
$$L_{\text{aux}}(\theta) = \mathbb{E}_{\eta \geq \eta_0}\left[V_\theta^{\pi^\eta} - V_\theta^{\pi}\right] - \mathbb{E}_{\eta < \eta_0}\left[V_\theta^{\pi^\eta} - V_\theta^{\pi}\right]$$
$$= \mathbb{E}_{\eta \geq \eta_0, \tau \sim \pi^\eta}\left[\sum_{t=0}^{\infty} \gamma^t A_\theta^\pi(s,a)\right]$$
$$- \mathbb{E}_{\eta < \eta_0, \tau \sim \pi^\eta}\left[\sum_{t=0}^{\infty} \gamma^t A_\theta^\pi(s,a)\right].$$

$$A_\theta^\pi(s,a) = \mathbb{E}_{s' \sim T(\cdot|s,a)}[f_{\xi,\phi}(s,a,s')]$$
$$= \mathbb{E}_{s' \sim T(\cdot|s,a)}[g_\xi(s,a) + \gamma h_\phi(s') - h_\phi(s)].$$

Method	HalfCheetah-v3	Hopper-v3
Demonstration	1085	1130
AIRL (Fu et al., 2017)	1872.81±87.13	1188.93±31.00
Aux-AIRL	**2191.64**±103.34	**1453.61**±15.09

Table 3. Imitation learning performance of Aux-AIRL and AIRL evaluated on the ground-truth reward throughout the training.

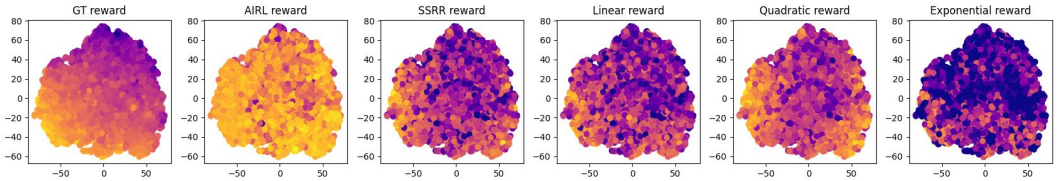
Aux-AIRL vs SSRR



Analysis on SSRR

Domain	Metric	Target Regression Function Form (fitted)			
		Sigmoid (SSRR)	Linear	Quadratic	Exponential
HalfCheetah-V3	Avg. Return	**1148.98 ±945.36**	821.34 ±208.60	774.66 ±418.97	476.42 ±881.49
	GT Corr.	**0.965**	0.934	0.956	0.952
Hopper-V3	Avg. Ret.	1916.72 ±102.36	2447.04 ±199.35	1630.26 ±339.61	**2529.09 ±315.39**
	GT Corr.	0.948	**0.966**	0.949	**0.966**

Table 1. Ground truth reward (with standard error) and correlation coefficients of different target regression functions.



Target Regression Function Form (hand-picked)			
Sigmoid	Linear	Quadratic	Exponential
3375.67	**4946.85**	2520.87	2045.53
±638.00	±514.95	±798.11	±1228.59
0.977	0.978	0.958	**0.983**

Table 2. Ground truth reward (with standard error) and correlation coefficients of different hand-picked regression functions (no fitting) in HalfCheetah-V3.