Valuing Sports Actions and Players with Inverse Reinforcement Learning

Yudong Luo

Oliver Schulte

SFU SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

SPORT LOGiQ

The Score Sparsity Problem

- A fundamental problem in sports analytics is <u>**valuing actions.**</u>
- In low-scoring sports (hockey, soccer), explicit values are attached only to rare goal events.

➢ Emphasis on goals and related actions (shots, assists)

➢ Bias towards offensive players

- Top-50 players for NHL 2018-19 season
  - **Scoring Impact** (SI)[Routley and Schulte, 2015] : All offensive players
  - **Goal Impact Metric** (GIM)[Liu and Schulte, 2018] : Only one Defenceman

<u>Our approach:</u> Learn a latent reward function that values a match situation

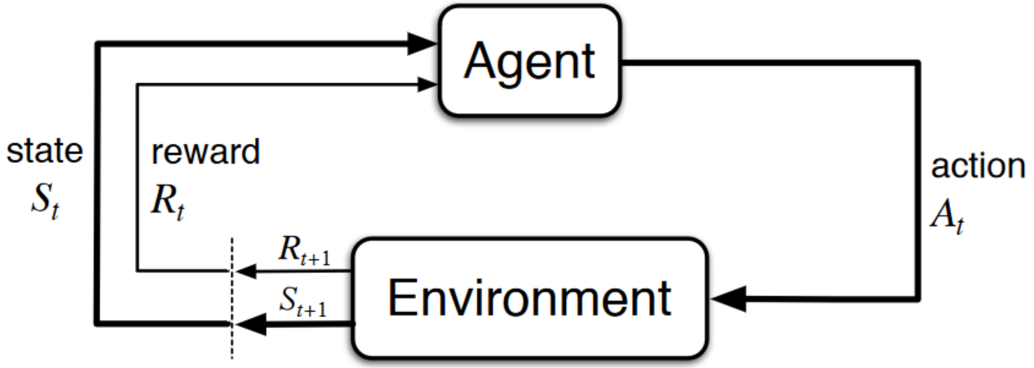
$f($ Match State $)$ ⟶ Reward Value

Overview

- Brief Intro to Inverse Reinforcement Learning (IRL)
- Our method:
  1. Alternating approach: leverage single-agent IRL by learning rewards for team A given observations of team B, then vice versa
  2. Combine learned latent rewards with observed goals by regularization
- Evaluation on ice hockey:
  - Dense reward signal
  - No bias between offensive and defensive players
  - Useful player ranking

Markov Model Setup

- Markov Decision Process



	Agent	State				Action	Observed Reward
gameId	teamId	Period	xCoord	yCoord	Manpower	Event	Score
849	15	1	-24.5	-17	Even	Carry	0
849	16	1	-75.5	-21.5	Even	Check	0
849	15	1	-79	-19.5	Even	pass	0
849	16	1	-92	-32.5	Even	Lpr	0
849	16	1	-92	-32.5	Even	Goal	1
849	15	1	-70	42	Even	Face-off	0

Inverse Reinforcement Learning

- In IRL [Ng et al., 2000], the reward function is unknown and should be inferred from demonstrations (data)
- Given $\mathrm{MDP}\backslash r$ and data, recover reward r

	Single Agent	State				Action	Reward
gameId	teamId	Period	xCoord	yCoord	Manpower	Event	
849	15	1	-9.5	1.5	Even	Lpr	?
849	15	1	-24.5	-17	Even	Carry	?

Inverse Reinforcement Learning

Maximum Entropy IRL [Ziebart et al, 2008]

- Reward is a linear function of state features, with weights $\boldsymbol{\theta} \in \mathbb{R}^k$
- The reward for a trajectory is the sum reward of visited states
- MaxEnt: the likelihood of trajectory is proportion to exponential reward

$$P(\zeta_i) \propto e^{r_{\zeta_i}}$$

- Calculate gradient of likelihood for $\boldsymbol{\theta}$, and update

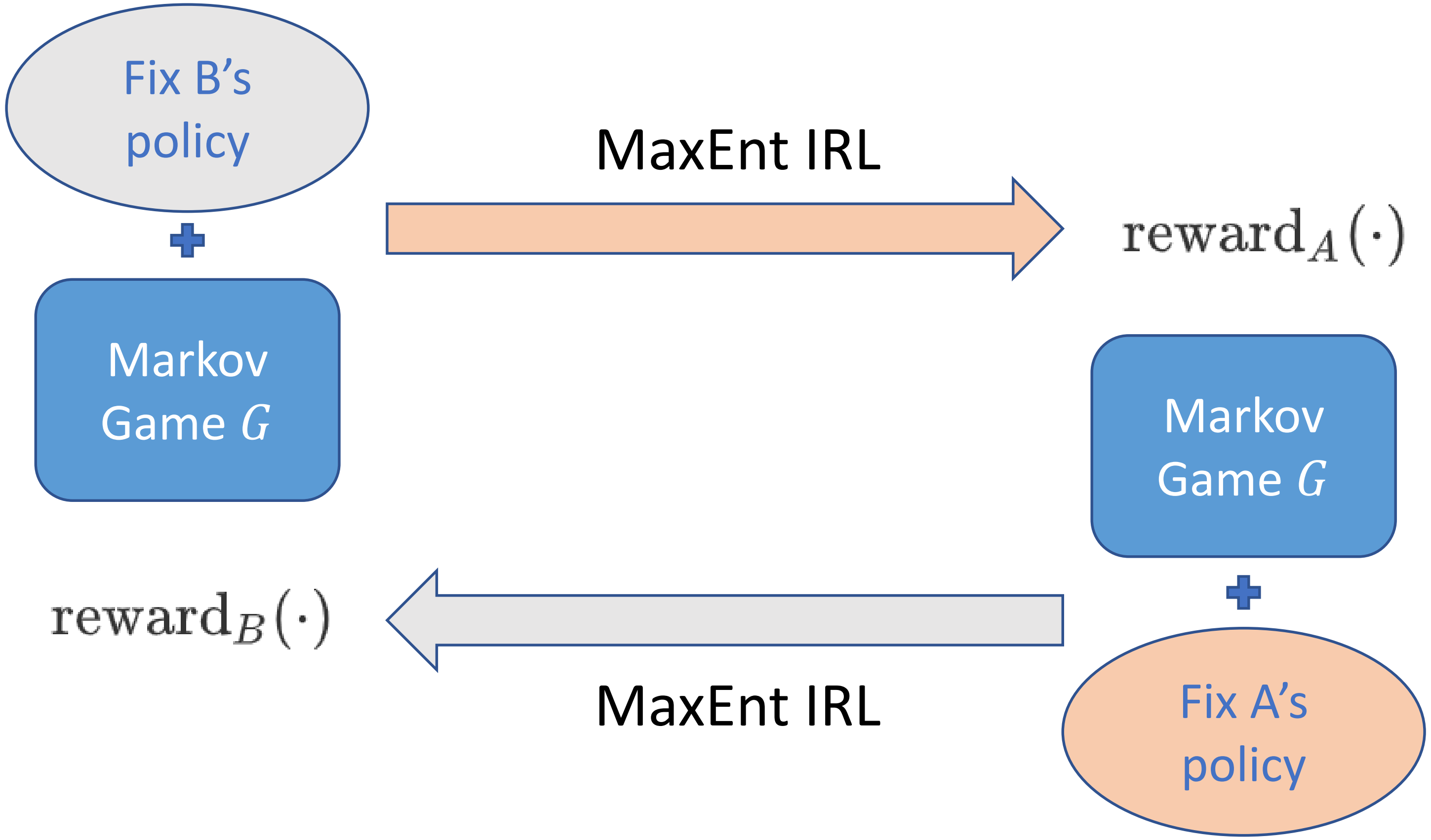
Our scenarios

- Leverage single-agent IRL for Multi-agent Markov Game (Home/Away)
- Combine knowledge between observed and unobserved reward

Multiple Agents	State				Action	Observed Reward	Unobserved Reward
teamId	Period	xCoord	yCoord	Manpower	Event	Score	
16	1	-75.5	-21.5	Even	Check	0	?
15	1	-79	-19.5	Even	pass	0	?
16	1	-92	-32.5	Even	Lpr	0	?
16	1	-92	-32.5	Even	Goal	1	?
15	1	-70	42	Even	Face-off	0	?

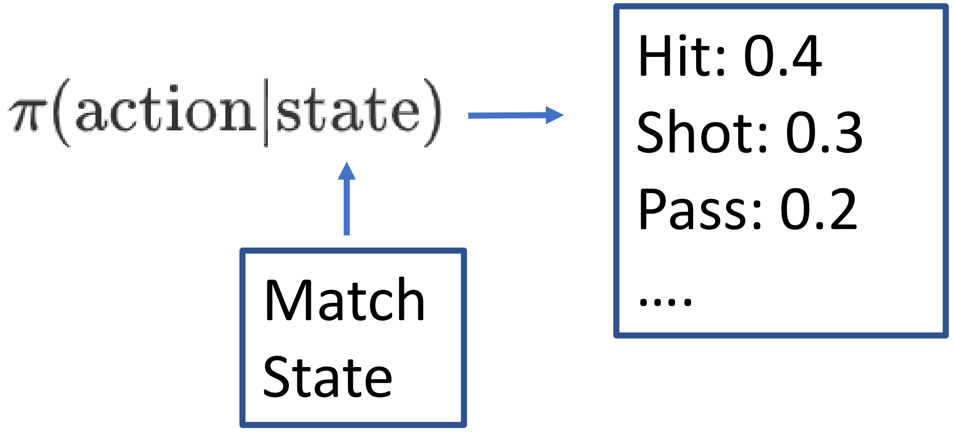
Alternating IRL

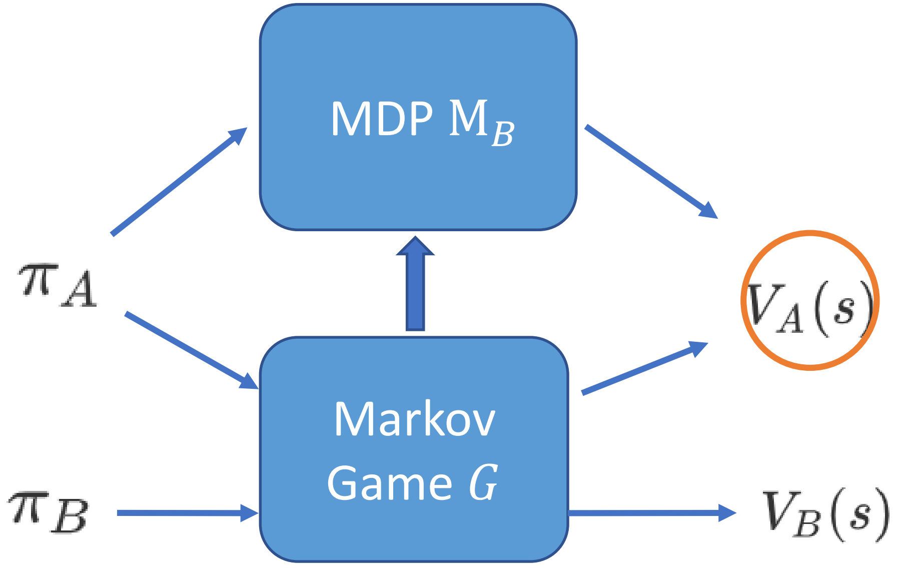
- Treat B as A's environment, learn reward for A using single-agent IRL
- Repeat the procedure with the role of teams A and B reversed



Transform Multi-agent Model to Single-agent Model

- **Proposition** Consider a two-agent Markov Game model G with two agent A, B, and a policy π_B for agent B. There is a single-agent MDP M_B such that for every policy π_A of agent A, the state value in Markov Game to A equals the state value in MDP
- Inituition: Single-agent MDP M_B treats B as part of A's environment





Combining Observed Goals and Learned Rewards

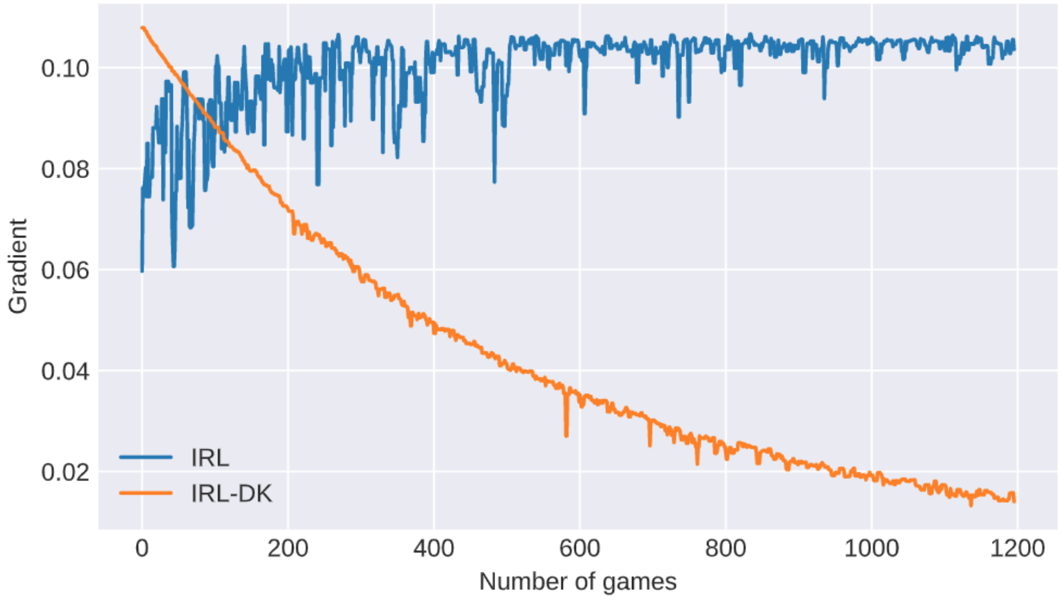
- Choose a kernel function k to measure similarity between observed scores and learned rewards
- Learning procedure is maximize regularized likelihood function

$$\arg\max L(\{\zeta\}|\text{rewards}) + \lambda k(\text{rewards}, \text{goals})$$

- Can be derived from maximum mean discrepancy [Gretton et al., 2012] framework for transfer learning

Learning Details and Performance

- MaxEnt IRL define a linear reward function with weight θ
- Pretrain a θ_0 to match goals reward, and initialize θ with θ_0
- Domain knowledge leads to much more stable and faster convergence



Learned Rewards Solve Sparsity

- Dataset
  - NHL play-by-play dataset from SPORTLOGiQ
  - Game from October 2018 to April 2019

Number of teams	31
Number of players	979
Number of games	1,202
Number of events	4,534,017

- Learned Rewards

Items	STD
Rule reward function (goals)	0.0383
IRL-DK learned reward function	0.1281
Q-values from goals (GIM)	0.0963
Q-values from IRL-DK	1.2207

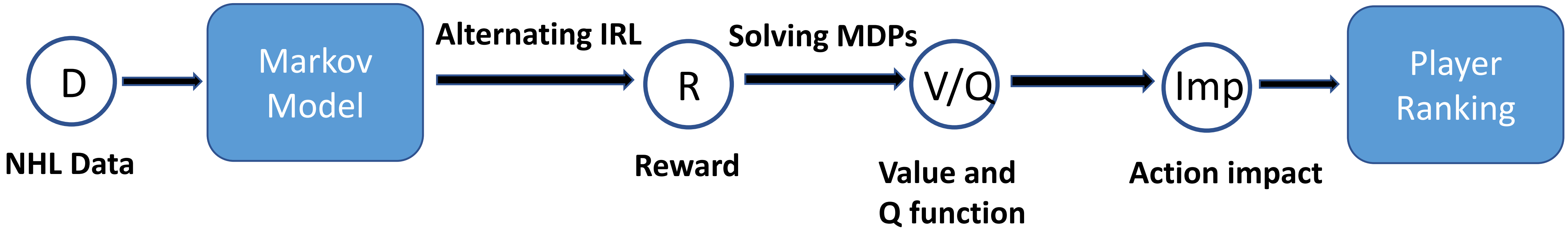
Rationalizing Player Behavior

- Given learned reward functions, solve MDPs to find optimal policy for Home/Away team.
- Negative log-likelihood (NLL) of observed trajectories under optimal policy
- Modified Hausdorff Distance (HMD) between observed trajectories and trajectoies generated by optimal policy [Kitani et al, 2012]

Methods	NLL	HMD
Rule reward function (goals)	185.0	13.37
IRL learned reward function	53.9	9.71
IRL-DK learned reward function	**49.5**	**7.77**

Player Ranking

- Value/Q function: estimates expected total future reward given current match state
- Use learned reward to calculate value function and Q function for each team (Routley and Schulte, 2015)
- Use value and Q function to assess action impact (Routley and Schulte, 2015; Liu and Schulte, 2018)



Player Ranking

- Top-10 offensive and defensive players

Name	Assists	Goals	Points	Team	Salary
Anze Kopitar	38	22	60	LA	11,000,000
Aleksander Barkov	61	35	96	FLA	6,900,000
Dylan Larkin	41	32	73	DET	7,000,000
Nathan Mackinnon	58	41	99	COL	6,750,000
Leon Draisaitl	55	50	105	EDM	9,000,000
Mark Scheifele	46	38	84	WPG	6,750,000
Jonthan Toews	46	35	81	CHI	9,800,000
Connor McDavid	75	41	116	EDM	14,000,000
Jack Eichel	54	28	82	BUF	10,000,000
Ryan O'Reilly	53	30	83	CAR	6,000,000

Table 3: 2018-19 Top-10 offensive players

Name	Assists	Goals	Points	Team	Salary
Drew Doughty	37	8	45	LA	12,000,000
Brent Burns	67	16	83	SJ	10,000,000
Roman Josi	41	15	56	NSH	4,000,000
John Carlson	57	13	70	WSH	12,000,000
Morgan Rielly	52	20	72	TOR	5,000,000
Ryan Suter	40	7	47	MIN	9,000,000
Mark Giordano	57	17	74	CGY	6,750,000
Duncan Keith	34	6	40	CHI	3,500,000
Erik Gustafsson	43	17	60	CHI	1,800,000
Miro Heiskane	21	12	33	DAL	925,000

Table 4: 2018-19 Top-10 defensive players

- No obvious bias to player position (top-50)
  - **SI** : 0 / 50 defensive players
  - **GIM** : 1 / 50 defensive players
  - **Ours** : 32 / 50 defensive players

Correlation with Success Measures

Methods	Assists	GP	Goals	GWG	SHG	PPG	S
+/-	0.269	0.086	0.282	0.278	0.118	0.124	0.156
VAEP	0.215	0.185	0.215	0.089	-0.074	0.160	0.239
WAR	0.591	0.322	0.742	0.571	0.179	0.610	0.576
EG	0.656	0.629	0.633	0.489	0.099	0.391	0.737
SI	0.717	0.633	**0.975**	**0.665**	**0.249**	**0.770**	0.860
GIM	0.757	0.772	0.781	0.518	0.147	0.477	0.795
IRL	0.855	0.872	0.812	0.587	0.123	0.513	0.901
IRL-DK	**0.882**	**0.887**	0.824	0.607	0.125	0.537	**0.907**

Methods	Points	SHP	PPP	FOW	P/GP	SFT/GP	PIM
+/-	0.285	0.179	0.157	0.012	0.306	0.109	0.100
VAEP	0.235	-0.076	0.185	0.021	0.204	0.129	0.172
WAR	0.692	0.147	0.605	0.040	0.699	0.396	0.145
EG	0.694	0.183	0.508	0.254	0.644	0.713	0.355
SI	0.869	0.204	0.708	0.135	0.728	0.639	0.361
GIM	0.818	0.151	0.561	0.289	0.705	0.751	0.372
IRL	0.891	0.207	0.696	0.294	0.741	0.818	0.437
IRL-DK	**0.908**	**0.213**	**0.734**	**0.298**	**0.769**	**0.820**	**0.446**

Table 5: Correlation with success measures (offensive)

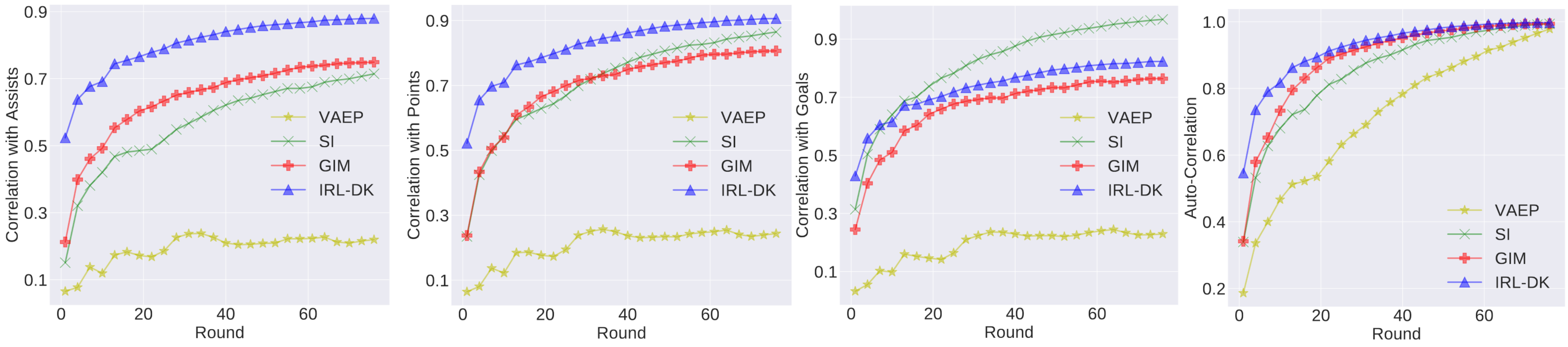
Methods	Assists	GP	Goals	GWG	SHG	PPG	S
+/-	0.173	0.132	0.144	0.177	0.235	-0.116	0.113
VAEP	0.054	-0.045	0.005	0.010	0.384	0.071	-0.016
WAR	0.204	0.028	0.365	0.275	0.097	0.246	0.186
EG	0.589	0.688	0.507	0.321	0.327	0.306	0.679
SI	0.607	0.488	**0.934**	**0.449**	**0.491**	**0.457**	0.709
GIM	0.702	0.862	0.596	0.263	0.130	0.170	0.764
IRL	0.809	0.941	0.686	0.415	0.268	0.347	0.908
IRL-DK	**0.852**	**0.959**	0.701	0.439	0.289	0.360	**0.920**

Methods	Points	SHP	PPP	FOW	P/GP	SFT/GP	PIM
+/-	0.175	0.107	-0.05	0.095	0.169	0.067	0.072
VAEP	0.042	0.065	-0.003	0.101	0.064	-0.036	-0.031
WAR	0.252	0.128	0.266	0.174	0.279	0.006	-0.089
EG	0.611	0.278	0.399	0.118	0.503	0.694	0.360
SI	0.720	0.174	0.488	0.103	0.521	0.499	0.272
GIM	0.730	0.085	0.358	0.140	0.471	0.706	0.438
IRL	0.841	0.281	0.549	0.182	0.557	0.776	0.549
IRL-DK	**0.865**	**0.307**	**0.571**	**0.185**	**0.574**	**0.778**	**0.570**

Table 6: Correlation with success measures (defensive)

Temporal Consistency

- Correlation between **first n round** value and Assists, Points, Goals
- Auto-correlation: **first n round** with **entire season** value



Conclusions

- Inverse reinforcement learning is a technique to infer reward for agent that explain its behavior
- Two innovations for our multi-agent IRL
  - Alternating learning reduces multi-agent to single-agent IRL
  - Transfer knowledge between observed goals and uobsorved rewards
- Learn dense rewards and match observed behavior
- Can be used to value actions and players, with a promising player ranking

Thank you!