Uncertainty-Aware Reinforcement Learning for Risk-Sensitive Player Evaluation in Sports Game

Guiliang Liu · Yudong Luo ·
Oliver Schulte · Pascal Poupart

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

- We design an **uncertainty-aware Reinforcement Learning (RL)** framework to learn a **risk-sensitive player evaluation** metric from stochastic game dynamics.
- To capture the risk of a player's movements into the distribution of action-values, we model
  1. **aleatoric uncertainty**, which represents the intrinsic stochasticity in a sports game
  2. **epistemic uncertainty**, which is due to a model's insufficient knowledge for Out-of-Distribution (OoD) samples.
- We introduce a **Risk-sensitive Game Impact Metric (RiGIM)** that measures players' performance over a season by conditioning on a specific confidence level derived from the uncertainty estimation.

Example

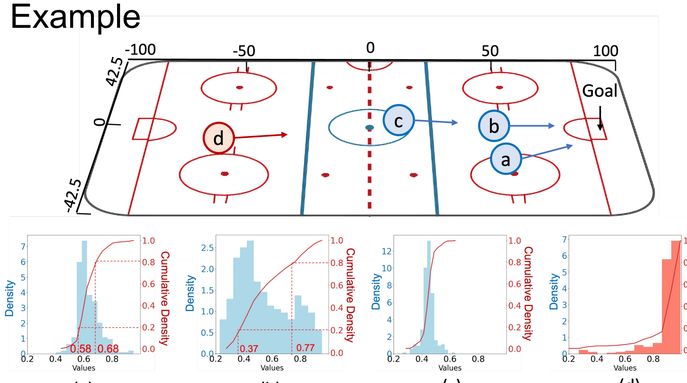

Figure 1. The predicted distribution of future goals in an ice hockey game between Blues and Coyotes, 2018-19 NHL season.

The shots are made in the positions (a) - (d).

1) **Risk-sensitive Evaluation**: Distributions (a) and (b) have the same expectation (around 0.6), but **different** impact on risk-sensitive evaluation:
   1) the first shot has a larger risk-averse estimate (at the confidence 0.8, we find 0.58 > 0.37) and
   2) a smaller risk-seeking estimate (at the confidence 0.2, we find 0.68 < 0.77)
2) **Post-hoc calibration**: the event of shooting from the position (d) is rare in an ice hockey game, and thus this event is likely to be OoD, leading to a biased prediction at (d) (the predicted scoring chances are too large).

Approach to Uncertainty-Aware RL

Epistemic Uncertainty. We perform a post-hoc calibration of the predicted action values by modeling their epistemic uncertainty (due to OoD samples).
Aleatoric Uncertainty. We estimate distributions of action values to model their aleatoric uncertainty (due to the intrinsic stochasticity in the game dynamics).

[1] Borislav Mavrin, Hengshuai Yao, Linglong Kong, Kaiwen Wu, and Yaoliang Yu. Distributional reinforcement learning for efficient exploration. In International Conference on Machine Learning (ICML), volume 97, pages 4424–4434, 2019.

Modelling the Uncertainty of Action Values

We develop a distributional-RL approach for aleatoric uncertainty and a feature-space density estimator for measuring epistemic uncertainty.

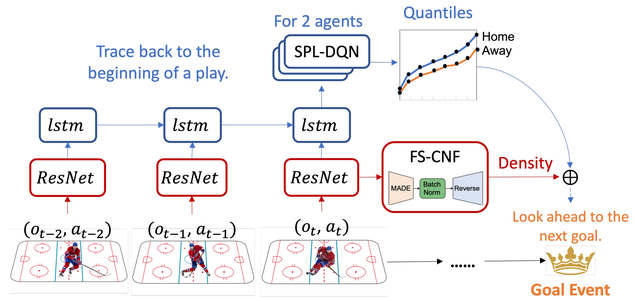

Figure 2. Model architecture. A play is a turn where one team attacks and the other defends. We add Spectral Normalization to ResNet outputs.

Distributional RL for Capturing Aleatoric Uncertainty

- Distributional RL learns the **distribution of the random variable** $Z_k(s_t, a_t)$ that corresponds to the number of future goals when a player of team k performs action a_t in state s_t.
- Following the Quantile-Regression (QR)-DQN method, we represent the distribution of Z by **a uniform mixture of N supporting quantiles**.
- When the player of a team k performs an action a_t at a state s_t, the agent receives a reward $R_k(s_t, a_t)$, moves to a future state $s_{t+1} \sim P_{\mathcal{T}}(s_{t+1}|s_t, a_t)$, where the agent's next action $a_{t+1} \sim \pi(A_{t+1}|S_{t+1})$. **This stochastic process can be captured by a distributional Bellman operator [2] $\mathcal{T}_\pi$**:

$$\mathcal{T}^\pi Z_k(s_t, a_t) \triangleq R_k(s_t, a_t) + \gamma Z_k(S_{t+1}, A_{t+1})$$

- **Temporal projection** of model outputs:

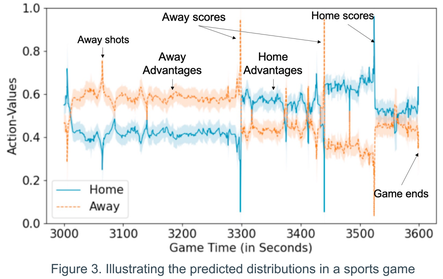

Figure 3. Illustrating the predicted distributions in a sports game

Density Estimator for Capturing Epistemic Uncertainty

We design a **Feature Space Conditional Normalizing Flow (FS-CNF)** to estimate sample density in the training distribution.

- To prevent feature collapse, the feature extractor is subjected to a bi-Lipschitz constraint (see Formula (4) in our paper).
- FS-CNF utilizes the Masked Auto-regressive Flow (MAF) design that estimates the density of input variables in the training data distribution.

[2] Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforce- ment learning with quantile regression. In AAAI, pages 2892–2901, 2018.

Risk-Sensitive Action Impact

To understand how players respond to risk, we propose a **Risk-sensitive Game Impact Metric (RiGIM)** that assigns lower weights to low-density match states:

$$RiGIM_l(c) = \sum_{(s,a)\in\mathcal{D}'} n(s,a,l) \times \phi_k(s,a,c)$$

$$\phi_k(s_{t+1}, a_{t+1}, c) = \left[\hat{Z}_k^c(s_{t+1}, a_{t+1}) - \hat{Z}_k^c(s_t, a_t)\right]\mathbb{I}_{p(\cdot|\boldsymbol{z}_E)\geq\epsilon}$$

- $\phi(s,a)$ measures how much an action a changes the return of a player's team.
- $c \in [0,1]$ is the confidence level, Z^c denotes the $(1-c)th$ quantile in $Z(\cdot)$
- $n(s,a,l)$ denotes the number of times that a player l performs action a at a state s in the testing dataset.
- $p(s,a|Z_E)$ defines the density estimator, with which we filter the OoD samples.

Empirical Evaluation

- **Dataset.** We utilize both an ice-hockey and a soccer dataset from the National Hockey League (NHL) and major European soccer leagues, which contains 9,213,371 events, covering 195 teams, 4,172 games, and 6,513 players.
- **Experiment Settings**. We divide the dataset into a training set (80%), a validation set (10%), and a testing set (10%) according to game dates.
- **Comparison Methods.** We employ the baselines[3] under an ablation design:

Method	Risk-Aware	History-Aware	RL-Based	Continuous Feature	Impact-Based	Context-Aware
+/−	✗	✗	✗	✗	✗	✗
EG	✗	✗	✗	✗	✗	✓
SI	✗	✗	✗	✗	✓	✓
VAEP	✗	✓	✗	✓	✓	✓
T0-GIM	✗	✗	✓	✓	✓	✓
GIM	✗	✓	✓	✓	✓	✓

RiGIM is our model

Player Evaluation Performance: Correlations with Standard Measures

Table 4: Correlations with standard measures in the **ice hockey** games. The *success* measures are assist, goal, Game Winning Goal (GWG), Overtime Goal (OTG), Short-handed Goal (SHG), Power-play Goal (PPG), Point (P), Short-handed Point (SHP), Power-play Point (PPP), Time On Ice (TOI), and Shots (S). The *penalty* measure is Penalty Minute (PIM).

Methods	Assist	Goal	GWG	OTG	SHG	PPG	Point	SHP	PPP	TOI	S	PIM
+/−	0.181	0.189	0.187	0.028	0.071	0.063	0.206	0.119	-0.071	0.021	0.038	-0.014
EG	0.239	0.303	0.264	0.130	-0.053	0.163	0.322	0.023	0.226	0.153	0.534	-0.112
SI	0.237	**0.596**	**0.409**	0.123	0.095	0.351	0.452	0.066	0.274	0.224	0.405	0.138
VAEP	0.238	0.454	0.225	0.06	0.053	0.326	0.382	-0.0	0.321	0.086	0.362	0.027
T0-GIM	0.397	0.394	0.139	0.16	0.151	0.216	0.455	0.153	0.295	0.356	0.387	0.058
GIM	0.456	0.408	0.167	0.158	0.134	0.246	0.501	0.137	0.345	0.395	0.431	0.061
Na-RiGIM(0.5)	0.593	0.476	0.223	0.173	**0.152**	0.313	0.625	**0.175**	0.453	0.597	0.611	0.115
GDA-RiGIM(0.5)	0.591	0.475	0.221	0.174	**0.152**	0.315	0.623	0.174	0.452	0.593	0.609	0.113
RiGIM(0.5)	0.675	0.477	0.266	0.184	0.11	0.355	0.678	0.141	0.529	0.68	0.7	0.146
RiGIM(c^*)	**0.68**	0.477	0.269	**0.187**	0.107	**0.357**	**0.681**	0.141	**0.531**	**0.685**	**0.707**	0.147

Sensitivity to Risk: Correlations Conditioning on Different Confidence Levels

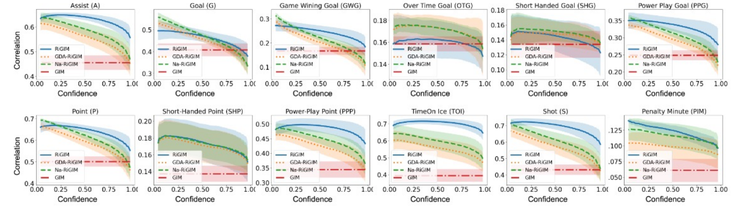

Figure 5: Correlations (Mean ± standard deviation) with success measures (the first 11 plots) and penalty measures (the last plot) at different confidence levels in **ice-hockey** games.

[3] Guiliang Liu, Oliver Schulte. Deep Reinforcement Learning in Ice Hockey for Context-Aware Player Evaluation. International Joint Conference on Artificial Intelligence (IJCAI) 2018.