A Unified Diversity Measure for Multiagent Reinforcement Learning

Zongkai Liu[1], Chao Yu[1*], Yaodong Yang[2],
Peng Sun[3], Zifan Wu[1]

1 School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
2 Institute for AI Peking University, Beijing, China
3 ByteDance, Shenzhen, China

Introduction

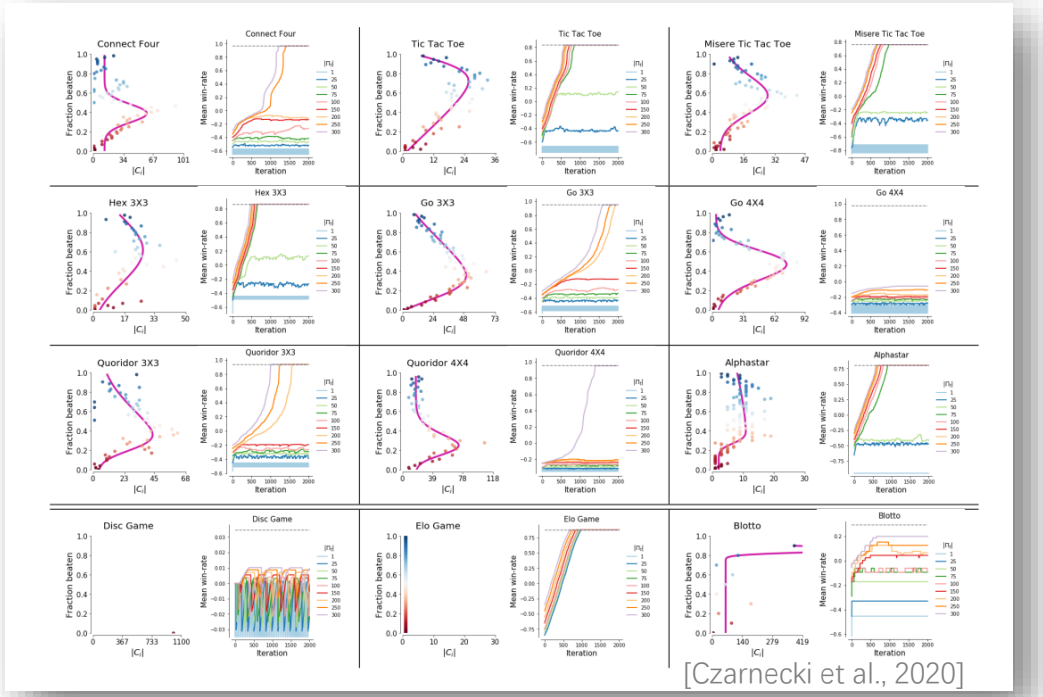
- Evaluation is of critical importance in Multiagent Reinforcement Learning.
  - Elo rating system [Elo, 1978]
  - Exploitability [Davis et al., 2014]
  - …

Arpad Elo, the inventor of the Elo rating system

A.E. Elo, The Rating of Chess players, Past and Present. Ishi Press International, 1978.
Davis et al., Using response functions to measure strategy strength. AAAI, 2014.

Introduction

- There is not a dominant strategy in non-transitive games, where the set of strategies follows a cyclic rule.
  - e.g., the strategic cycle among Rock, Paper and Scissors
- Many real-world games demonstrate strong non-transitivity [Czarnecki et al., 2020].

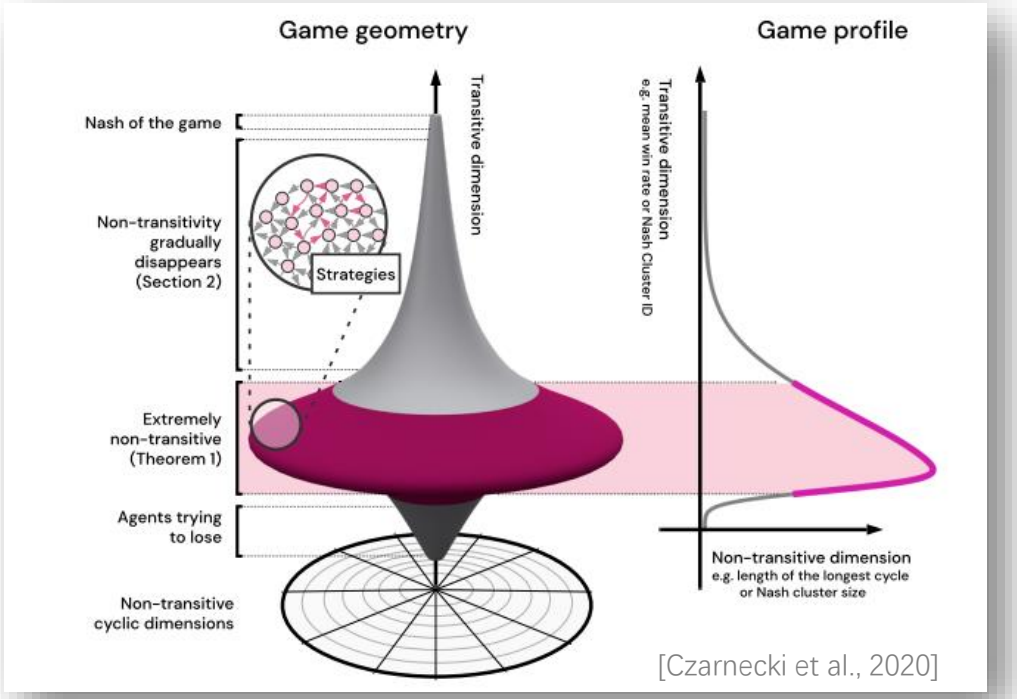


[Czarnecki et al., 2020]

Czarnecki et al., Real world games look like spinning tops. NeurIPS 2020.

Introduction

- Since the strategic cycle exists, diversity matters at each level of strategies.
- Diversity also helps agents find strategies at the higher level of the spinning top.
- How to evaluate the diversity of a set of strategies is critical for solving games with non-transitive dynamics.



Czarnecki et al., Real world games look like spinning tops. NeurIPS 2020.

Unified Diversity Measure

- Many diversity metrics have been proposed, such as *Effective Diversity* [Balduzzi et al., 2019], *Population Diversity* [Parker-Holder et al., 2020], *Expected Cardinality* [Nieves et al., 2021].
- But there are still no consistent formal definitions for diversity, making it difficult to evaluate the diverse strategies in MARL.
- We work towards offering a consistent definition for diversity, and propose a novel population-wide diversity measure called the Unified Diversity Measure (UDM) .

Balduzzi et al., Open-ended learning in symmetric zero-sum games. ICML, 2019.
Parker-Holder et al., Effective diversity in population-based reinforcement learning. NeurIPS, 2020.
Nieves et al., Modelling Behavioural Diversity for Learning in Open-Ended Games. ICML, 2021.

Unified Diversity Measure

Strategy Feature

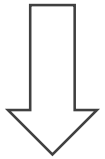
$$\boldsymbol{\phi}_i^n \in \mathbb{R}^{1\times p}, p \leq M =: |\mathbb{S}^n|$$

Diversity Kernel

$$\mathcal{L}_K^n := [K(\boldsymbol{\phi}_i^n, \boldsymbol{\phi}_j^n)]_{M\times M}$$

Function

$$f \in \boldsymbol{F} := \left\{ f : f(x) = \sum_{k=0}^{\infty} c_k x^k, f'(x) > 0, x \in R \right\}$$



Unified Diversity Measure (UDM)

$$\mathrm{UDM}(\mathbb{S}^n) := \sum_{i=1}^{M} f(\lambda_i)$$

Geometric Meaning of UDM

Equivalent Representation

$$\mathrm{UDM}(\mathbb{S}^n) := \sum_{i=1}^{M} f(\lambda_i) = \mathrm{Tr}(f(\mathcal{L}_K^n))$$

(proposition 1)

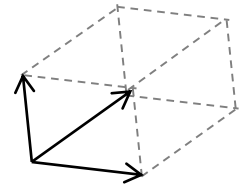
Jacobi Formula

$$\det\left(\exp\left(\mathbf{A}\right)\right) = \exp \mathbf{Tr}(\mathbf{A}), \mathbf{A} \in \mathbb{R}^{M \times M}$$

[Magnus et al., 2019]

Geometric Meaning of UDM

the volume of the exponential of the diversity kernel



Magnus et al., Matrix differential calculus with applications in statistics and econometrics. John Wiley & Sons, 2019.

Unify Existing Metrics into UDM

Effective Diversity (ED)

$$\mathrm{ED}(\mathbb{S}^n) := \boldsymbol{\pi}^{*\mathsf{T}} \lfloor \mathcal{M} \rfloor_+ \boldsymbol{\pi}^* \,,\ \lfloor x \rfloor_+ = x \text{ if } x > 0 \text{ and } \lfloor x \rfloor_+ = 0 \text{ if } x \leq 0$$

[Balduzzi et al., 2019]

Population Diversity (PD)

$$\mathrm{PD}(\mathbb{S}^n) := \det(\mathbf{K}) = \det(K(\boldsymbol{\phi}_i^n, \boldsymbol{\phi}_j^n)) \,,\ \boldsymbol{\phi}_i^n = \{\boldsymbol{\pi}_i^n(\cdot|s)\}_s$$

[Parker-Holder et al., 2020]

Expected Cardinality (EC)

$$\mathrm{EC}(\mathbb{S}^n) := \mathbb{E}_{\mathbf{Y} \sim \mathbb{P}_{\mathcal{L}^n}}[|\mathbf{Y}|] = \mathrm{Tr}(\mathbf{I} - (\mathcal{L}^n + \mathbf{I})^{-1})$$

[Nieves et al., 2021]

Unify Existing Metrics into UDM

Methods	Kernel Function $K(\cdot,\cdot)$	Function f	Strategy Feature $\boldsymbol{\phi}_i$
ED [1]	Linear Kernel	$f(x) = x$	$\boldsymbol{m}_i^*$
PD [38]	self-selected	$f(x) = \ln x$	$\{\boldsymbol{\pi}(\cdot\|s)\}_s$
RPD	self-selected	$f(x) = \ln x$	$\boldsymbol{m}_i$
EC [36]	Linear Kernel	$f(x) = \frac{x}{1+x}$	$\boldsymbol{m}_i$

(Table 2)

Remark:
Using UDM, we can also analyze the advantages and shortcomings of the existing metrics, and study why ED and PD cannot measure the diversity properly in certain cases. Please see our paper for details.

Balduzzi et al., Open-ended learning in symmetric zero-sum games. ICML, 2019.
Parker-Holder et al., Effective diversity in population-based reinforcement learning. NeurIPS, 2020.
Nieves et al., Modelling Behavioural Diversity for Learning in Open-Ended Games. ICML, 2021.

UDM-Based Algorithms

UDM Fictitious Play

$$\mathrm{BR}_{\tau_t}^n(\boldsymbol{\pi}_t^{-n}) = \arg\max_{\tilde{\boldsymbol{\pi}} \in \Delta_{\mathbb{S}_t^n}} [\boldsymbol{G}^n(\tilde{\boldsymbol{\pi}}, \boldsymbol{\pi}_t^{-n}) + \tau_t \cdot \mathrm{UDM}(\mathbb{S}_t^n \cup \{\tilde{\boldsymbol{\pi}}\})]$$

Proposition 3 (Convergence of UDM-FP). *If UDM is concave, and UDM-FP uses the update rule:*

$$\boldsymbol{\pi}_{t+1}^n \in (1-\alpha_{t+1})\boldsymbol{\pi}_t^n + \alpha_t(BR_{\tau_t}^n(\boldsymbol{\pi}_t^{-n}) + \boldsymbol{U}_{t+1}^n),$$

where $\alpha_t = o(1/\log t)$ is deterministic and perturbations $\boldsymbol{U}_{t+1}^n$ are the differences between the actual and expected changes in strategies. Then UDM-FP shares the same convergence property as GWFP: the policy sequence $\boldsymbol{\pi}_t^n$ converges to the NE on two-player zero-sum games or potential games.

UDM-PSRO

$$\mathcal{O}^n(\boldsymbol{\pi}^{-n}) = \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^d} [\sum_{S^{-n} \in \mathbb{S}^{-n}} \boldsymbol{\pi}^{-n}(S^{-n}) \cdot g(S_{\boldsymbol{\theta}}, S^{-n}) + \tau \cdot \mathrm{UDM}(\mathbb{S}^n \cup \{S_{\boldsymbol{\theta}}\})]$$

Proposition 4 (EGS Enlargement). *Adding a new (meta-)strategy $S_{\boldsymbol{\theta}}$ via Eq. (8) enlarges EGS. Formally, we have* $\mathrm{EGS}(\mathbb{S}^n) \subseteq \mathrm{EGS}(\mathbb{S} \cup \{S_{\boldsymbol{\theta}}\})$. $\mathrm{EGS}(\mathbb{S}^n) := \{\sum_i \alpha_i \cdot \boldsymbol{m}_i : \boldsymbol{\alpha} \geq 0, \boldsymbol{\alpha}^\top \cdot \mathbf{1} = 1, \boldsymbol{m}_i = \mathcal{M}_{[i,:]}\}$

Experiments

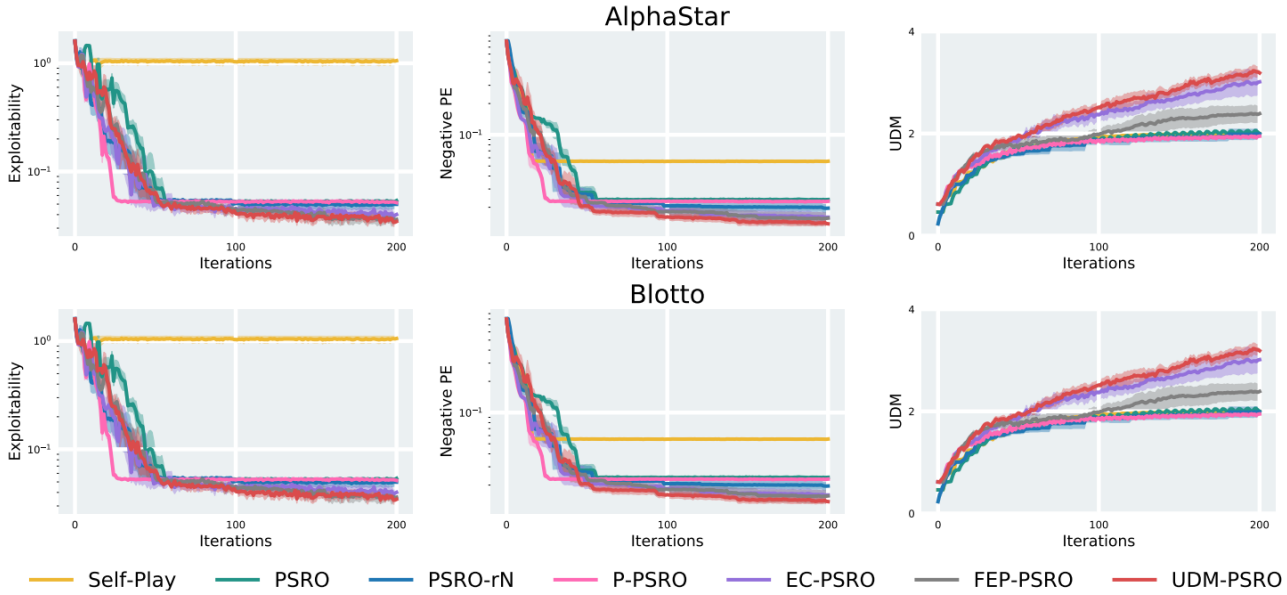


Figure 1: AlphaStar and Blotto: Exploitability & Negative PE & UDM vs. Iterations.

Table 3: The OS (Opponent Strength) associated with the PE$\times 10^2$ represents the strength of the opponent during the process of using PSRO to solve it. The last row is Exploit.$\times 10^2$.

PE(OS)	PSRO	PSRO-rN	P-PSRO	EC-PSRO	RPD-PSRO	UDM-PSRO
PE(10)	-18.18 ± 0.32	-18.18 ± 0.32	9.62 ± 0.16	9.34 ± 0.13	9.68 ± 0.19	$\mathbf{9.73 \pm 0.24}$
PE(15)	-27.28 ± 0.04	-27.28 ± 0.04	0.43 ± 0.04	0.01 ± 0.04	0.39 ± 0.05	$\mathbf{0.44 \pm 0.06}$
PE(20)	-26.73 ± 0.04	-26.73 ± 0.04	0.10 ± 0.12	0.18 ± 0.06	0.34 ± 0.16	$\mathbf{0.69 \pm 0.09}$
PE(25)	-25.47 ± 0.07	-25.47 ± 0.07	1.12 ± 0.10	1.25 ± 0.17	1.39 ± 0.14	$\mathbf{1.81 \pm 0.18}$
Exploit.	33.71 ± 0.37	35.11 ± 0.23	2.34 ± 0.43	2.05 ± 0.38	$\mathbf{1.95 \pm 0.54}$	2.07 ± 0.37

Baselines:

Self-Play [Fudenberg et al., 1998] PSRO [Lanctot et al., 2017]

PSRO-rN [Balduzzi et al., 2019] P-PSRO [McAleer et al., 2020]

EC-PSRO [Nieves et al., 2021] FEP-PSRO [Liu et al., 2021]

Fudenberg et al., The theory of learning in games. MIT press, 1998.
Lanctot et al., A unified game-theoretic approach to multiagent reinforcement learning. NeurIPS, 2017.
Balduzzi et al., Open-ended learning in symmetric zero-sum games. ICML, 2019.
McAleer et al., Pipeline PSRO: A scalable approach for finding approximate nash equilibria in large games.NeurIPS, 2020.
Nieves et al., Modelling Behavioural Diversity for Learning in Open-Ended Games. ICML, 2021.
Liu et al., Towards unifying behavioral and response diversity for open-ended learning in zero-sum games. NeurIPS, 2021.

Thank you!