Status-quo policy gradient in Multi-Agent Reinforcement Learning

Extended Abstract

Pinkesh Badjatiya*
Microsoft, India
pbadjatiya@microsoft.com

Mausoom Sarkar, Nikaash Puri, Jayakumar Subramanian
MDSR, Adobe, India
{msarkar,nikpuri,jasubram}@adobe.com

Abhishek Sinha*
Waymo, USA
a7b23@stanford.edu

Siddharth Singh†
University Of Maryland, USA
siddharth9820@gmail.com

Balaji Krishnamurthy
MDSR, Adobe, India
kbalaji@adobe.com

ABSTRACT

Individual rationality, which involves maximizing expected individual returns, does not always lead to high-utility individual or group outcomes in multi-agent problems. For instance, in multi-agent social dilemmas, Reinforcement Learning (RL) agents trained to maximize individual rewards converge to a low-utility mutually harmful equilibrium. In contrast, humans evolve useful strategies in such social dilemmas. Inspired by ideas from human psychology that attribute this behavior to the status-quo bias, we present a status-quo loss (*SQLoss*) and the corresponding policy gradient algorithm that incorporates this bias in an RL agent. We demonstrate that agents trained with *SQLoss* learn high-utility policies in several social dilemma matrix games (Prisoner's Dilemma, Matching Pennies, Chicken Game). To apply SQLoss to visual input games where cooperation and defection are determined by a sequence of lower-level actions, we present GameDistill, an algorithm that reduces a visual input game to a matrix game. We empirically show how agents trained with SQLoss on GameDistill reduced versions of Coin Game and Stag Hunt learn high-utility policies. Finally, we show that *SQLoss* extends to a 4-agent setting by demonstrating the emergence of cooperative behavior in the popular Braess' paradox.

KEYWORDS

Multi-Agent Learning, Reinforcement Learning, Social Dilemma, Policy gradient, Deep Learning

ACM Reference Format:
Pinkesh Badjatiya[1], Mausoom Sarkar, Nikaash Puri, Jayakumar Subramanian, Abhishek Sinha[1], Siddharth Singh[2], and Balaji Krishnamurthy. 2022. Status-quo policy gradient in Multi-Agent Reinforcement Learning: Extended Abstract. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

* Work done while at Media and Data Science Research Labs, Adobe.
† Work done during the internship at Adobe.

1 INTRODUCTION

In sequential social dilemmas, individually rational behavior can lead to low-utility outcomes for each individual in the group [3, 7, 11, 12]. Current state-of-the-art Multi-Agent Deep RL (MARL) methods train agents who play individualistically and receive lower rewards, even in simple social dilemmas [4, 10] such as the Prisoner's Dilemma and Coin Game [4].

Interestingly, humans learn cooperative strategies (without sharing rewards or having access to private information) that benefit both the individual and the group in such dilemmas. Several ideas in human psychology [8, 9, 13, 14] have attributed this behavior to the status-quo bias [6]. Inspired by this idea, we present the status-quo loss (*SQLoss*) and the corresponding status-quo policy gradient formulation for RL.

Agents with *SQLoss* achieve high rewards in multi-agent social dilemmas without sharing rewards, gradients, modeling the other agent(s), or using a communication channel. Intuitively, *SQLoss* encourages an agent to stick to past actions provided the actions did not cause significant harm. *SQLoss* requires a binary action space. To apply *SQLoss* to visual games where cooperation and defection policies are defined using a sequence of lower-level actions, we present *GameDistill*, an algorithm that reduces a visual input game to a matrix game. *GameDistill* uses self-supervision and clustering to extract distinct policies from a sequential social dilemma.

Our key contributions can be summarised as:

(1) We introduce a **Status-Quo** loss and an associated policy gradient-based algorithm to learn policies in a decentralized manner for agents playing iterated games where agents can choose between two distinct policies in each iteration. We demonstrate that agents trained with *SQLoss* achieve high rewards in several social dilemma matrix games.
(2) We propose *GameDistill*, an algorithm that reduces a visual input game to a matrix game by automatically extracting distinct policies. We show how agents trained with *SQLoss* on these *GameDistill* extracted policies obtain high rewards in the Coin Game and Stag Hunt.
(3) We demonstrate that *SQLoss* extends to games with more than two agents. We show agents trained with *SQLoss* cooperate in the 4-agent setting of the popular Braess' paradox.

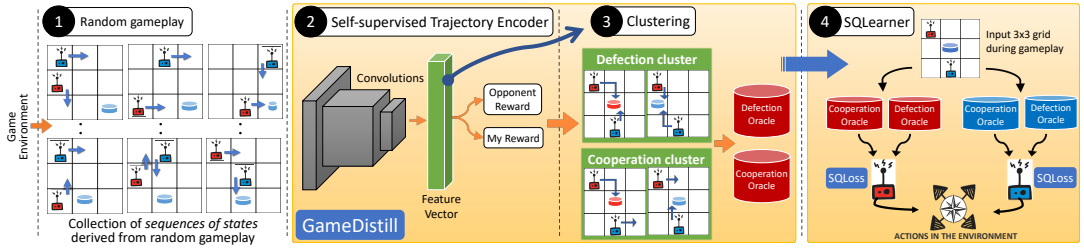

Figure 1: Extending *SQLoss* to visual input games using GameDistill: High-level architecture illustrated using coin game. Each agent runs *GameDistill* by performing steps (1), (2), (3) individually to obtain two oracles per agent. During game-play(4), each agent (with *SQLoss*) takes either the action suggested by the cooperation or the defection oracle

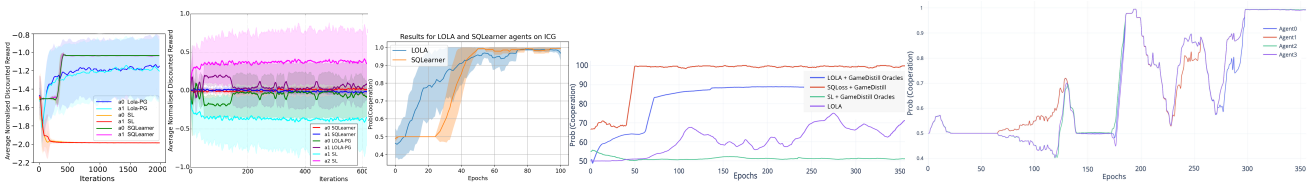

Figure 2: Results of SQLoss on (1) IPD (2) IMP (3) ICG games and (4) *GameDistill* with SQLoss on visual Coin Game.

2 APPROACH

In Infinitely Iterated Matrix Games, agents repeatedly play a matrix game against each other. In each game iteration, each agent has access to actions played by both agents in the previous iteration. We refer to infinitely iterated matrix games as iterated matrix games. In Iterated Prisoner's Dilemma (IPD), RL agents trained with the policy gradient update method converge to a sub-optimal mutual defection (*DD*) equilibrium (Figure 2, Lerer and Peysakhovich [10]) instead of cooperation (*CC*). This sub-optimal equilibrium attained by Selfish Learners motivates us to explore alternatives.

2.1 SQLoss: Motivation

The status-quo bias instills in humans a preference for the current state provided the state is not harmful to them. Inspired by this idea, we introduce a status-quo loss (*SQLoss*). The *SQLoss* encourages an agent to imagine a future episode where the status-quo (current situation) is repeated for several steps. If an agent has been exploited in the previous iteration of the game (state *DC*), then *SQLoss* will encourage the agent to imagine a continued risk of exploitation and subsequently switch to defection and move to state *DD*. Conversely, if both agents cooperated in the previous iteration of the game (state *CC*), then *SQLoss* will encourage the agent to imagine a continued gain from mutual cooperation and subsequently stick to state *CC*. Please see our full-version of the paper [1] for details.

2.2 Learning policies using *SQLoss* and *GameDistill*

Applying *SQLoss* to visual input games is not straightforward since *SQLoss* requires a binary action space that typically contains a cooperation and a defection action. To apply the Status-Quo policy gradient to such games, we propose *GameDistill*, a self-supervised algorithm that reduces a visual input game to a matrix game. *GameDistill* works as follows.

(1) We initialize agents with random weights and play them against each other in the game. During **random game-play**, whenever an agent receives a reward, we store the sequence of states and the rewards for both agents.
(2) This collection of state sequences is used to train the *GameDistill* network, which is a **self-supervised trajectory encoder**. It takes as input a sequence of states and predicts the rewards of both agents during training.
(3) We now **cluster the embeddings** extracted from the penultimate layer of the trained *GameDistill* network using Agglomerative Clustering [5]. Each embedding is a finite dimensional representation of the corresponding state sequence. For the Coin Game, we set the number of clusters to two (since *SQLoss* requires a binary action space).
(4) We **train an oracle** to predict the next action given the current state using the state sequences in each cluster. For the Coin Game, we get two oracles, one for each cluster. Each agent uses *GameDistill* independently to extract two oracles that represent a high-level behavior in the game

3 EXPERIMENTS AND RESULTS

We compare our approach Status-Quo Aware Learner or *SQLearner* to Lola-PG [4] and the Selfish Learner (SL) agents. For all experiments, we perform 20 runs and report average *NDR*, along with variance across runs. *SQLearner* agents achieve close to optimal score in the IPD, IMP and ICG (Figure 2). SL agents converge to mutually harmful selfish behavior. *SQLearner* agents move towards equilibrium faster than Lola-PG agents and also has low variance. For Visual games, Lola-PG agents trained with *GameDistill* oracles achieve higher rates of cooperation than vanilla Lola-PG agents but lower rates of cooperation than *SQLearner* agents trained with *GameDistill* oracles. Finally, using a 4-agent Braess' paradox[2] game, we extend *SQLoss* to games beyond two agents.

REFERENCES

[1] Pinkesh Badjatiya, Mausoom Sarkar, Abhishek Sinha, Siddharth Singh, Nikaash Puri, and Balaji Krishnamurthy. 2021. Status-quo policy gradient in Multi-Agent Reinforcement Learning. https://arxiv.org/abs/2111.11692. (2021).

[2] D. Braess. 1968. Über ein Paradoxon aus der Verkehrsplanung. *Unternehmensforschung* 12, 1 (01 Dec 1968), 258–268.

[3] Thomas Dietz, Elinor Ostrom, and Paul C. Stern. 2003. The Struggle to Govern the Commons. *Science* 302, 5652 (2003), 1907–1912. https://doi.org/10.1126/science.1091015

[4] Jakob Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. 2018. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 122–130.

[5] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning*. Vol. 1. Springer series in statistics New York.

[6] Begum Guney and Michael Richter. 2018. Costly switching from a status quo. *Journal of Economic Behavior & Organization* 156 (2018), 55–70.

[7] Garrett Hardin. 1968. The Tragedy of the Commons. *Science* 162, 3859 (1968), 1243–1248. https://doi.org/10.1126/science.162.3859.1243

[8] Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.

[9] Daniel Kahneman, Jack L Knetsch, and Richard H Thaler. 1991. Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic perspectives* 5, 1 (1991), 193–206.

[10] Adam Lerer and Alexander Peysakhovich. 2017. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. (2017). arXiv:arXiv:1707.01068

[11] E. Ostrom. 1990. *Governing the commons-The evolution of institutions for collective actions*. Political economy of institutions and decisions.

[12] Elinor Ostrom, Joanna Burger, Christopher B. Field, Richard B. Norgaard, and David Policansky. 1999. Revisiting the Commons: Local Lessons, Global Challenges. *Science* 284, 5412 (1999), 278–282. https://doi.org/10.1126/science.284.5412.278

[13] William Samuelson and Richard Zeckhauser. 1988. Status quo bias in decision making. *Journal of risk and uncertainty* 1, 1 (1988), 7–59.

[14] Richard H Thaler and Cass R Sunstein. 2009. *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.