Negotiated Reasoning: On Provably Addressing Relative Over-Generalization

Extended Abstract

Junjie Sheng
East China Normal University
Shanghai, China
jarvis@stu.ecnu.edu.cn

Wenhao Li*
Tongji University
Shanghai, China
whli@tongji.edu.cn

Bo Jin
Tongji University
Shanghai, China
bjin@tongji.edu.cn

Hongyuan Zha
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
zhahy@cuhk.edu.cn

Jun Wang
East China Normal University
Shanghai, China
jwang@cs.ecnu.edu.cn

Xiangfeng Wang*
East China Normal University,
Shanghai Formal-Tech Information Technology Co., Lt
Shanghai, China
xfwang@cs.ecnu.edu.cn

ABSTRACT

We focus on the *relative over-generalization* (RO) issue in fully cooperative multi-agent reinforcement learning (MARL). Existing methods show that endowing agents with *reasoning* can help mitigate RO empirically, but there is little theoretical insight. We first prove that RO is avoided when agents satisfy a *consistent reasoning* requirement. We then propose a new *negotiated reasoning* framework connecting reasoning and RO with theoretical guarantees. Based on it, we develop an algorithm called *Stein variational negotiated reasoning* (SVNR), which uses Stein variational gradient descent to form a negotiation policy that provably bypasses RO under maximum-entropy policy iteration. SVNR is further parameterized with neural networks for computational efficiency. Experiments demonstrate that SVNR significantly outperforms baselines on RO-challenged tasks, confirming its advantage in achieving better cooperation.

KEYWORDS

Multi-Agent Reinforcement Learning; Relative Overgeneralization

ACM Reference Format:
Junjie Sheng, Wenhao Li*, Bo Jin, Hongyuan Zha, Jun Wang, and Xiangfeng Wang*. 2025. Negotiated Reasoning: On Provably Addressing Relative Over-Generalization: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

We focus on the fully cooperative setting in Multi-agent reinforcement learning (MARL) [2, 4, 6, 12], where agents optimize team performance. A crucial challenge here is *relative over-generalization* (RO). People and animals can suffer from over-generalization in various tasks [1, 7, 11], e.g., a "once bitten, twice shy" effect. In MARL, RO similarly causes sub-optimal behaviors by over-fitting to partial exploration from others [8]. For instance, in a Particle Gather environment, pairs of agents controlling each dimension of movement often avoid success due to negative experiences where only one arrived at the target. Researchers address RO through credit assignment (e.g., lenient learning [9, 10] or shaped values [13, 15, 18]) and *reasoning*-based methods [14, 16, 17]. Despite empirical advances, few works offer solid theoretical guarantees or a formal definition of RO. We ask: (1) *Can methods provably avoid RO?* and (2) *How to design a method that meets such guarantees?*

We first formalize *perceived RO* (PRO) and *executed RO* (ERO), showing that provably avoiding RO requires *consistent reasoning*: agents must model other agents' behaviors consistently with optimal policies, both in training and execution. We then introduce *negotiated reasoning* (NR), inspired by negotiation in human cooperation [3, 5]: agents repeatedly refine action beliefs until reaching an agreement, ensuring consistent reasoning under mild conditions.

Further, we develop a novel *Stein variational negotiated reasoning* (SVNR) that applies (MP)SVGD [19] to improve perceived policies. SVNR leverages maximum-entropy policy iteration to guarantee convergence to a global optimum in cooperative settings. By annealing the entropy regularization, SVNR yields ERO-free policies in decentralized execution. Finally, we present a practical implementation based on neural networks and *amortized* optimization to handle continuous state-action spaces. Experiments in challenging tasks (e.g., differential games and Particle Gather) confirm that SVNR consistently finds global-optimal cooperation.

2 RELATIVE OVER-GENERALIZATION

We formalize RO under the CTDE (centralized training, decentralized execution) paradigm. Standard definitions view RO as converging to a sub-optimal equilibrium because each agent's marginal policy outperforms the optimal equilibrium policy if opponents vary unpredictably [16]. We introduce two refined concepts for training vs. execution:

Definition 2.1 (Executed RO (ERO))**.** Agent i suffers *executed* RO if the executed joint policy can be improved by revealing other

*Corresponding authors.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

agents' actions. More formally, if

$$\max_{\pi_i} \left\{ U^{\pi_i(u^i|s,\boldsymbol{u}^{-i})} \right\} > U^{\prod_j \bar{\pi}_j^*(u^j|s)},$$

where $\bar{\pi}_j^*$ is the executed policy for agent j.

Definition 2.2 (Perceived RO (PRO))**.** Agents suffer *perceived* RO if there exists an agent i whose policy is regretful when not knowing the optimal opponent policy:

$$\min_{\pi_i} D_{KL}(\pi_i \rho_i \| \pi_\alpha^*) > \min_{\pi_i} D_{KL}(\pi_i \pi_\alpha^*(\boldsymbol{u}^{-i}) \| \pi_\alpha^*),$$

where π_α^* is the optimal joint policy (with entropy factor α).

We prove that avoiding both PRO at training and ERO at execution ensures no RO. In practice, we show *consistent reasoning* (agents modeling opponents accurately and making deterministic actions when $\alpha \to 0$) leads to RO-free outcomes.

3 NEGOTIATED REASONING FRAMEWORK

Our key idea is to equip each agent with *negotiation policies* that iteratively establish consistent beliefs. Each agent i maintains $f_i(u_i \mid \boldsymbol{u}_{C_i}, s)$, where C_i indicates whose actions agent i observes in negotiation. We show that if negotiation policies converge to identity maps while the perceived joint policy converges to the optimal policy, the method is PRO-free. Further, if at execution each agent's action is consistent with the negotiation outcome, ERO is avoided.

Theorem 3.1 (PRO-free Negotiated Reasoning). *If for any state s, after K negotiation steps we have*

$$\lim_{k \to K} p(\boldsymbol{u}^k \mid \boldsymbol{s}) = \pi^*(\boldsymbol{u} \mid \boldsymbol{s}),$$

then no agent suffers PRO in that state.

Theorem 3.2 (ERO-free Negotiated Reasoning). *When the above PRO-free condition holds and we anneal $\alpha \to 0$, agents are also free of ERO by executing their final negotiated actions.*

Hence, consistent reasoning emerges if each agent's negotiation policy makes them reach agreement identical to the optimal joint policy and use that in decentralized execution.

4 STEIN VARIATIONAL NEGOTIATED REASONING

We now propose *Stein Variational Negotiated Reasoning* (SVNR), the first MARL method that provably avoids RO.

Negotiation policy derivation. We minimize the KL divergence to the optimal π_α^* via a single-agent "perturbation" that fixes others. This matches the (MP)SVGD [19] approach. We combine local improvements under a strict nested negotiation structure to ensure convergence to the optimal joint policy.

SVNR policy iteration. We integrate negotiation with *maximum entropy* policy iteration. By applying repeated soft Bellman backups and nested negotiation updates, we prove that the perceived joint policy converges to the global optimum:

Lemma 4.1 (Joint Policy Evaluation)**.** For a mapping $Q^0 : \boldsymbol{\mathcal{S}} \times \boldsymbol{\mathcal{U}} \to \mathbb{R}$, the iteration $Q^{k+1} = \Gamma_{\hat{\pi}} Q^k$ converges to the joint soft Q-function of $\hat{\pi}$ for $|\boldsymbol{\mathcal{U}}| < \infty$.

Algorithm 1 SVNR: Stein Variational Negotiated Reasoning

Input: Initial policies f^{ψ_i}, critic Q^θ, nested sets $\{C_i\}$, replay buffer $\mathcal{D}$, kernel κ_i, etc.
while not converged **do**
  Collect Experience:
    Each agent samples noise ξ_i, outputs $u_i = f^{\psi_i}(\xi_i; \xi_{C_i}, s)$, executes $\boldsymbol{u}$, observes $r, \boldsymbol{s'}$, stores $(s, \boldsymbol{u}, r, s')$ in $\mathcal{D}$.
  Sample from $\mathcal{D}$:
    Update critic θ by minimizing Bellman error.
    Update policies via amortized (MP)SVGD:

$$\frac{\partial J(\psi_i)}{\partial \psi_i} \propto \mathbb{E}_\xi \left[\Delta f_i^{\boldsymbol{\psi}}(\xi) \cdot \partial f_i^{\boldsymbol{\psi}}(\xi) / \partial \psi_i \right].$$

Lemma 4.2 (Policy Improvement)**.** Under strict nesting, updating $\hat{\pi}$ with (MP)SVGD toward $\tilde{\pi} \propto \exp\left(\frac{1}{\alpha} Q - \frac{1}{\alpha} V\right)$ leads to $Q^{\hat{\pi}'} \geq Q^{\hat{\pi}}$.

Theorem 4.3 (SVNR Policy Iteration). *Repeated joint policy evaluation and improvement converges to π^* such that $Q^{\pi^*} \geq Q^{\hat{\pi}}$ for all $\hat{\pi} \in \Pi$.*

With α annealed down, the execution policy avoids ERO.

5 A PRACTICAL IMPLEMENTATION OF SVNR

To handle continuous spaces and large domains, we *amortize* the negotiation process using neural networks. Each agent i has a functional mapping $u_i = f^{\psi_i}(\xi_i; \boldsymbol{\xi}_{C_i}, s)$ producing actions from shared noise. The objective is to match the final negotiated distribution. We adopt (MP)SVGD to estimate gradients w.r.t. ψ_i:

$$\Delta f_i^{\boldsymbol{\psi}} = \mathbb{E}_{\boldsymbol{u} \sim p^{\boldsymbol{\psi}}} \left[k_i(\cdot) \nabla_{u_i'} Q^\theta(\boldsymbol{u'}) + \alpha_i \nabla_{u_i'} k_i(\cdot) \right].$$

Then we backpropagate to ψ_i. Combining this with a centralized soft critic and Bellman error minimization yields *SVNR*, summarized in Algorithm 1.

6 EXPERIMENTS

We evaluate SVNR against strong baselines (MADDPG, MASQL, PR2, ROMMEO, Lenient MADRL) in a **differential games** (Max Of Three) and a **particle gather** task. Our method consistently escapes local sub-optima and achieves **near-optimal** final returns under various difficulty factors (Table 1).

Table 1: Execution performances in testings. The proposed SVNR achieves the highest returns in all scenarios.

Methods / Scenarios	*Max Of Three* ($s_2 = 3.0$)	*Max Of Three* ($s_2 = 2.0$)	*Max Of Three* ($s_2 = 1.5$)	*Particle Gather*
SVNR (Ours)	**9.60 ± 0.30**	**9.64 ± 0.17**	**9.71 ± 0.20**	**4.76 ± 0.20**
MADDPG	2.08 ± 4.63	−0.66 ± 0.67	−0.64 ± 0.43	0.00 ± 0.00
MASQL	8.92 ± 0.37	−0.58 ± 0.24	−0.34 ± 0.12	−0.54 ± 0.20
PR2	4.76 ± 3.64	−0.64 ± 0.45	−0.29 ± 0.10	0.00 ± 0.02
ROMMEO	6.14 ± 4.82	1.59 ± 5.03	−0.59 ± 0.25	−0.87 ± 0.22
L-MADRL	9.54 ± 0.13	1.63 ± 2.51	−0.07 ± 0.04	−0.75 ± 0.00

ACKNOWLEDGMENTS

This work is supported by NSFC (62406270), STCSM (22QB1402100, 24YF2748800), and the "Sino-German Cooperation 2.0" Funding Program of Tongji University.

REFERENCES

[1] Jonathan Baron. 2000. The effects of overgeneralization on public policy. In *Proceedings of the Intervento Presentato All'Experimental Method Conference*.

[2] Jeancarlo Arguello Calvo and Ivana Dusparic. 2018. Heterogeneous Multi-Agent Deep Reinforcement Learning for Traffic Lights Control.. In *AICS*.

[3] Peter JD Carnevale and Edward J Lawler. 1986. Time pressure and the development of integrative agreements in bilateral negotiations. *Journal of Conflict Resolution* 30, 4 (1986), 636–659.

[4] Guohui Ding, Joewie J Koh, Kelly Merckaert, Bram Vanderborght, Marco M Nicotra, Christoffer Heckman, Alessandro Roncone, and Lijun Chen. 2020. Distributed reinforcement learning for cooperative multi-robot object manipulation. In *AAMAS*.

[5] Jeong-Yoo Kim. 1996. Cheap talk and reputation in repeated pretrial negotiation. *The RAND Journal of Economics* (1996), 787–802.

[6] Karol Kurach, Anton Raichuk, Piotr Stanczyk, Michal Zajkac, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, et al. 2020. Google research football: A novel reinforcement learning environment. In *AAAI*.

[7] Offir Laufer, David Israeli, and Rony Paz. 2016. Behavioral and neural mechanisms of overgeneralization in anxiety. *Current Biology* 26, 6 (2016), 713–722.

[8] Gregory Palmer. 2020. *Independent learning approaches: Overcoming multi-agent learning pathologies in team-games*. The University of Liverpool (United Kingdom).

[9] Gregory Palmer, Karl Tuyls, Daan Bloembergen, and Rahul Savani. 2017. Lenient multi-agent deep reinforcement learning. *arXiv preprint arXiv:1707.04402* (2017).

[10] Liviu Panait, Keith Sullivan, and Sean Luke. 2006. Lenient learners in cooperative multiagent systems. In *AAMAS*. 801–803.

[11] David G Rand, Alexander Peysakhovich, Gordon T Kraft-Todd, George E Newman, Owen Wurzbacher, Martin A Nowak, and Joshua D Greene. 2014. Social heuristics shape intuitive cooperation. *Nature Communications* 5, 1 (2014), 1–12.

[12] Tabish Rashid, Philip Torr, Gregory Farquhar, Chia-Man Hung, Tim Rudner, Nantas Nardelli, Shimon Whiteson, Christian Schroeder de Witt, Jakob Foerster, and Mikayel Samvelyan. 2019. The StarCraft Multi-Agent Challenge. In *AAMAS*.

[13] Lin Shi and Bei Peng. 2022. Curriculum Learning for Relative Overgeneralization. *arXiv preprint arXiv:2212.02733* (2022).

[14] Zheng Tian, Ying Wen, Zhichen Gong, Faiz Punakkath, Shihao Zou, and Jun Wang. 2019. A Regularized Opponent Model with Maximum Entropy Objective. In *IJCAI*.

[15] Lipeng Wan, Zeyang Liu, Xingyu Chen, Han Wang, and Xuguang Lan. 2022. Greedy-based value representation for optimal coordination in multi-agent reinforcement learning. In *ICML*.

[16] Ermo Wei, Drew Wicke, David Freelan, and Sean Luke. 2018. Multiagent soft Q-learning. In *AAAI Spring Symposium Series*.

[17] Ying Wen, Yaodong Yang, Rui Luo, Jun Wang, and Wei Pan. 2019. Probabilistic Recursive Reasoning for Multi-Agent Reinforcement Learning. In *ICLR*.

[18] Wenshuai Zhao, Yi Zhao, Zhiyuan Li, Juho Kannala, and Joni Pajarinen. 2023. Optimistic Multi-Agent Policy Gradient for Cooperative Tasks. *arXiv preprint arXiv:2311.01953* (2023).

[19] Jingwei Zhuo, Chang Liu, Jiaxin Shi, Jun Zhu, Ning Chen, and Bo Zhang. 2018. Message passing Stein variational gradient descent. In *ICML*.