IBGP: Imperfect Byzantine Generals Problem for Zero-Shot Robustness in Communicative Multi-Agent Systems

Extended Abstract

Yihuan Mao*
Institute for Interdisciplinary Information Sciences, Tsinghua University
Beijing, China
maoyh20@mails.tsinghua.edu.cn

Yipeng Kang*
State Key Laboratory of General Artificial Intelligence, BIGAI
Beijing, China
kangyipeng@bigai.ai

Peilun Li
Shanghai Tree-Graph Blockchain Research Institute
Shanghai, China
peilun.li@confluxnetwork.org

Ning Zhang
Washington University in St. Louis
St. Louis, United States
zhang.ning@wustl.edu

Wei Xu
Institute for Interdisciplinary Information Sciences, Tsinghua University
Beijing, China
weixu@tsinghua.edu.cn

Chongjie Zhang
Washington University in St. Louis
St. Louis, United States
chongjie@wustl.edu

ABSTRACT

As AI agents become integral to infrastructure, robust coordination and message synchronization are crucial. The Byzantine Generals Problem (BGP) models resilience in multi-agent systems (MAS) under adversarial conditions, handling scenarios with malicious agents—stemming from AI hallucinations or external attacks. Traditional BGP demands global consensus, which is often unnecessary and inefficient in practice. We introduce Imperfect BGP (IBGP), aligning with the local coordination patterns in MAS to address this gap, offering provable resilience against communication attacks and adaptability to changing environments, as validated by empirical results.

KEYWORDS

Multi-agent Systems; Zero-shot Robustness; Safety; Byzantine Generals Problem

ACM Reference Format:
Yihuan Mao*, Yipeng Kang*, Peilun Li, Ning Zhang, Wei Xu, and Chongjie Zhang. 2025. IBGP: Imperfect Byzantine Generals Problem for Zero-Shot Robustness in Communicative Multi-Agent Systems: 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

Advancements in AI are leading to an era where AI agents form a significant part of our infrastructure. Heterogeneous agents from different manufacturers will collaborate to solve tasks, making coordination essential. Message synchronization among agents is critical, exemplified in sensor networks and UAV control. However, these agents may lack reliable broadcast mechanisms, causing discrepancies due to hallucinations or malicious compromises, which can have disastrous consequences.

This issue parallels the consensus problem in distributed systems, particularly the Byzantine Generals Problem (BGP), which seeks global consistency despite failures. In multi-agent systems (MAS), though, global consensus isn't always necessary; partial consensus often suffices. For instance, in predator-prey environments or threshold public goods games, coordinating a subset of agents is adequate for achieving goals.

To tackle coordination challenges in MAS, we formalize the Imperfect Byzantine Generals Problem (IBGP), emphasizing partial consensus over global agreement. We introduce a protocol tailored for IBGP, requiring less redundancy and accommodating a higher proportion of malicious agents compared to traditional BGP protocols. In the context of general AI agents, our protocol can be integrated through simple prompt instructions, leveraging inherent agent capabilities. Experiments demonstrate the effectiveness of our consensus protocols in both IBGP scenarios and practical tasks.

2 METHOD

2.1 Problem Definition

2.1.1 Preliminaries: BGP. In BGP, there are n benign agents and t attacker agents communicating with each other. The attackers can send false messages to disturb coordination. Each agent begins with an initial message $M^0 \in \{0, 1\}$ as its initial proposal and finally makes a decision action $a \in \{0, 1\}$. Generally, a value of 1 indicates cooperation, while 0 indicates giving up. The formal definition of BGP is provided in Definition 2.1.

Definition 2.1. (BGP)

- *Agreement:* $a_1 = a_2 = \cdots = a_n \in \{0, 1\}$. (All n benign agents must agree on the same action, either 0 or 1.)

*Equal contribution.

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).

	IBGP Protocol	Recursive training	AME	ADMAC
Predator-prey(4, 1, 2, 1)	96.1 ± 5.4%	0%	62.4 ± 20.4%	25.6 ± 13.2%
Predator-prey(5, 2, 2, 1)	97.9 ± 2.9%	3.7 ± 3.5%	42.6 ± 13.0%	14.0 ± 7.7%
Predator-prey(20, 4, 2, 2)	100.0 ± 0%	0%	79.3 ± 12.1%	/
Predator-prey(20, 1, 4, 2)	100.0 ± 0%	16.5 ± 16.4%	100.0 ± 0%	54.1 ± 20.3%
Hallway(3, 1, 2, 1)	96.7 ± 4.7%	0%	6.3 ± 6.3%	6.4 ± 4.8%
Hallway(10, 1, 5, 2)	100.0 ± 0%	/	13.1 ± 7.2%	7.4 ± 10.4%
4bane_vs_1hM(4, 1, 3, 1)	98.4 ± 2.2%	20.4 ± 6.4%	92.9 ± 4.7%	64.5 ± 13.6%
3z_vs_1r(3, 1, 2, 1)	51.5 ± 2.6%	6.2 ± 0.8%	/	/

Table 1: The table illustrates the robustness percentages and their standard deviation of different environments and algorithms.

- *Consistency:* If $M_1^0 = M_2^0 = \cdots = M_n^0 = x$, then $a_1 = a_2 = \cdots = a_n = x$. (When all n agents have identical initial observations, their actions must also be identical.)

Mis-coordination describes the situation where *Agreement* is violated, and researchers have designed consensus protocols proven to prevent mis-coordination under any communicative attacks.

2.1.2 IBGP. BGP serves as a fundamental concept of extensive research within the area of Decentralized Systems, embodying the crucial attributes of agreement and consistency within a decentralized framework. Nevertheless, these properties may not always apply in many Multi-Agent Systems (MAS), where partial coordination is a common pattern rather than universal coordination. For example, in a predator-prey environment [3], only a subset of predators may be required to collaborate in hunting a particular prey, rather than involving all predators in the pursuit.

To capture this coordination pattern within MAS, we introduce IBGP in Definition 2.2, where successful agreement necessitates the cooperation of only k. Besides, only the agents with the initial observation $M^0 = 1$ are permitted to take the cooperative action $a = 1$. The two properties of *Agreement* and *Consistency* are redefined to align with the partial coordination prevalent in Multi-Agent Systems.

Definition 2.2. (IBGP)

- *Agreement:* $\#(M_i^0 = 1, a_i = 1) \in \{0\} \cup [k, n]$. (At least k agents that observe $M^0 = 1$ are required to cooperate; otherwise no agent should act.)
- *Consistency:* If $\#(M_i^0 = 1) = n$, then $\#(M_i^0 = 1, a_i = 1) \geq k$. (If the number of available agents is super-sufficient, cooperation must happen.)

Similarly, mis-coordination is defined as the situation $0 < \#(M_i^0 = 1, a_i = 1) < k$, where *Agreement* is violated. It means that some agents try to coordinate but fail. Just like BGP, the goal of IBGP is to avoid mis-coordination altogether, and although we use Reinforcement Learning to formulate BGP and IBGP in the following section, our focus is on robustness rather than expected return.

2.2 Consensus Protocol for IBGP

To solve IBGP, we propose a consensus protocol. This protocol employs a multi-round broadcast pattern and incorporates the concept of an independent global randomizer (implemented similarly with [1]). In each round, a randomized bit variable determines whether it is the last round (with a value of 1 indicating the final round and 0 indicating that the process should continue).The framework amounts to the (k, λ)-protocol listed below:

(1) The global randomizer initializes the number of rounds from a distribution $r_{tot} \sim \mathcal{R}$. ($\mathcal{R}$ is the sample distribution of the total number of rounds r_{tot} in the IBGP Protocol. r_{tot} isn't revealed until the last round arrives.)
(2) Initial round: Each agent i broadcasts its initial proposal M_i^0 to each agents j.
(3) Round $r \in \{1 \cdots r_{tot}\}$: Each agent $i \in \{i | M_i^{r-1} = 1\}$ broadcasts $M_i^r = \mathbb{1}(\#_{j \in [N]}(M_{j \to i}^{r-1} = 1) \geq k + \lambda)$.
(4) Decision making round: Each agent $i \in \{i | M_i^{r_{tot}} = 1\}$ select action $a_i = \mathbb{1}(\#_{j \in [N]}(M_{j \to i}^{r_{tot}} = 1) \geq k)$.

When we set $\lambda = t$, the following theorem illustrates the robustness of the protocol:

Theorem 2.3. *(k, t)-protocol is robust with a high level of confidence $1 - \max_r\{p(r_{tot} = r)\}$ under any attack on IBGP(t, k).*

3 EXPERIMENTS

The experiment includes four kinds of environments, denoted as Env(n, m, k, t). Here, n represents the number of benign agents, m is the number of targets, k is the coordination threshold, and t is the number of attackers. Predator-prey is modified from the well-known predator-prey environment [3], requiring several predators to hunt the prey together. Hallway is introduced in [5], requiring several agents to reach the destination simultaneously. *4bane_vs_1hM* and *3z_vs_1r* are built on the SMAC benchmark (StarCraft Multi-agent Challenge) [2].

Table 1 indicates the robustness percentage of the IBGP protocol in the first column, while the second column displays the ratio of recursive training, which means that the training process of agents and attackers is repeated. Recursive training is one of the contributions of the algorithm in [6]. In the last two columns, AME [4] proposes a defense algorithm by taking the majority of multiple randomly ablated message sets, and ADMAC [7] automatically reduces the impact of potentially harmful messages on the final decision. Table 1 reveals that the IBGP Protocol maintains its performance from training to testing phases.

REFERENCES

[1] Michael O. Rabin. 1983. Randomized byzantine generals. In *24th Annual Symposium on Foundations of Computer Science (sfcs 1983)*. 403–409. https://doi.org/10.1109/

SFCS.1983.48

[2] Mikayel Samvelyan, Tabish Rashid, Christian Schröder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob N. Foerster, and Shimon Whiteson. 2019. The StarCraft Multi-Agent Challenge. *CoRR* abs/1902.04043 (2019). arXiv:1902.04043 http://arxiv.org/abs/1902.04043

[3] Peter Stone and Manuela Veloso. 2000. Multiagent systems: A survey from a machine learning perspective. *Autonomous Robots* 8 (2000), 345–383.

[4] Yanchao Sun, Ruijie Zheng, Parisa Hassanzadeh, Yongyuan Liang, Soheil Feizi, Sumitra Ganesh, and Furong Huang. 2023. Certifiably Robust Policy Learning against Adversarial Multi-Agent Communication. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=dCOL0inGl3e

[5] Tonghan Wang*, Jianhao Wang*, Chongyi Zheng, and Chongjie Zhang. 2020. Learning Nearly Decomposable Value Functions Via Communication Minimization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HJx-3grYDB

[6] Wanqi Xue, Wei Qiu, Bo An, Zinovi Rabinovich, Svetlana Obraztsova, and Chai Kiat Yeo. 2022. Mis-Spoke or Mis-Lead: Achieving Robustness in Multi-Agent Communicative Reinforcement Learning *(AAMAS '22)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC.

[7] Lebin Yu, Yunbo Qiu, Quanming Yao, Yuan Shen, Xudong Zhang, and Jian Wang. 2024. Robust Communicative Multi-Agent Reinforcement Learning with Active Defense. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 16 (Mar. 2024), 17575–17582. https://doi.org/10.1609/aaai.v38i16.29708