Mathematical Analysis of Emergent Collusion in MARL (Step 3 - V1.5)

Author: H. M. Liu

Scope: Mathematical Foundations, Optimization Divergence, and Structural Deterrence

Introduction: The Nature of Emergent Collusion

In Step 3 of the AI Mother Tongue (AIM) framework, we observe that agents are not explicitly programmed with "cheating" code. What we term "collusion" is, in fact, a **local optimum** that two agents naturally converge upon, guided by the gradients of specific mathematical equations. This behavior is emergent and self-organizing, arising from the fundamental drive to maximize cumulative returns within a shared latent space.

1. Core Mathematical Equations of Step 3

The underlying system consists of three main equations representing the "objective," the "action policy," and the "external monitoring."

1.1 Objective Function: Joint Reward

The sole survival objective of the agents is to maximize their joint reward. In the contextual Prisoner's Dilemma payoff matrix, when both agents choose to cooperate (C, C) and match the Environmental Label condition, they receive the highest possible score:

$$R_{joint} = r_A(a_A, a_B, label) + r_B(a_A, a_B, label)$$

1.2 Policy Optimization: Advantage Actor-Critic (A2C)

Agent A and Agent B update their neural network weights via the A2C algorithm. Taking Agent A as an example, its loss function comprises the Policy Gradient, Value Loss, and Entropy regularization:

$$\mathcal{L}_A = -\log \pi_{\theta_A}(z_A|x, label) \cdot A^{\pi} + \frac{1}{2}MSE(V_\phi(s), R_{joint}) - c_{ent}\mathcal{H}(\pi_{\theta_A})$$

Where A^{π} is the Advantage function, ensuring the policy moves towards actions that exceed average expected returns.

1.3 External Perspective: Baseline Observer

The Observer attempts to predict Agent B's actual action (C/D) based solely on Agent A's discrete symbols (z_A). Its mathematical goal is to minimize the Cross-Entropy loss:

$$\mathcal{L}_{obs} = -\mathbb{E}[y \log(Acc_{obs}) + (1-y)\log(1-Acc_{obs})]$$

2. Mathematical Observation of Collusive Behavior

We define collusion mathematically through the conflict between the **Internal Perspective** (Agents) and the **External Perspective** (Observer).

2.1 Internal Perspective: The Advantage of Secrecy

Agents exploit the high-dimensional, non-linear features of their neural networks to hide strategies within minute distributional shifts of z_A. They achieve a perfect $R_{joint} = 10$, and since the advantage function remains positive ($A^{\pi} > 0$), the agents mathematically evaluate their own performance as "excellent."

2.2 External Perspective: Gradient Loss

To the Baseline Observer, the sequence of z_A appears to be randomly assigned. No matter how the Observer updates its gradients, $\mathcal{L}_{obs}$ cannot be reduced, resulting in an Accuracy stuck at **Random Guess (0.25~0.5)**.

2.3 Mathematical Signature ("The Smoking Gun of Cheating")

Collusion is confirmed when the following two conditions occur simultaneously:

1. $\nabla_\theta \mathbb{E}[R_{joint}] \to 0$ (Reward has reached its peak/plateau).
2. $\nabla_w L_{obs} \approx 0$ (Observer gradients are lost, unable to converge despite agent cooperation).

3. The Collusion Score and Dynamic Intervention (V1.5 Updates)

3.1 The Collusion Score as an Optimization Divergence Probe

The Collusion Score acts as a detector for the mathematical singularity that occurs when internal optimization (Reward) diverges from external interpretability (Acc):

$$Score = \frac{R_{joint}}{Acc_{obs} + \epsilon}$$

3.2 [New] Dynamic Shuffle Threshold ($ts_{dynamic}$)

To prevent "Overt Collusion 套利" (where agents use high transparency as a hedge to maintain high rewards), V1.5 introduces a dynamic circuit breaker:

$$ts_{dynamic} = ts_{base} \times (1.0 - 0.5 \times Acc_{obs})$$

Significance: This forces a reset even if Acc_{obs} is high, provided R_{joint} is excessive. As $Acc_{obs} \to 1.0$, the tolerance for collusion drops by 50%.

4. Advanced Deterrence and Economic Realism

4.1 Temporal Deterrence (時域威懾)

In V1.5, we move beyond single-round punishment. Upon exceeding the Threshold Penalty (tp), the system triggers a **Temporal Reward Freeze** for $K = 15$ rounds:

$$R_{t...t+K} = 0$$

4.2 Impact on Credit Assignment

By zeroing future rewards, the cumulative return G_t collapses. Since $A^\pi = G_t - V(s)$, this turns the gradient direction away from collusive actions. In the gradient space, collusion is transformed from a "Local Optimum" into a "Dead End."

4.3 Economic Deterrence Inequality

Absolute deterrence is achieved only when the expected utility of collusion is mathematically inferior to mutual defection:

$$\mathbb{E}[R_{CC}] - (Penalty + Freeze\_Cost) < \mathbb{E}[R_{DD}]$$

Where $Freeze\_Cost = K \times R_{CC}$.

5. Conclusion: Mathematical Significance of the Circuit Breaker

The Circuit Breaker (Codebook Shuffle) is not just a restart; it is a **physical destruction** of the private weight mappings established via gradient descent. By combining dynamic thresholds with temporal economic deprivation, the V1.5 framework ensures that "cheating" is no longer a viable mathematical solution for self-organizing agents.