# COLA: Consistent Learning with Opponent-Learning Awareness

Timon Willi*, Alistair Letcher*, Johannes Treutlein*, Jakob Foerster

UNIVERSITY OF OXFORD

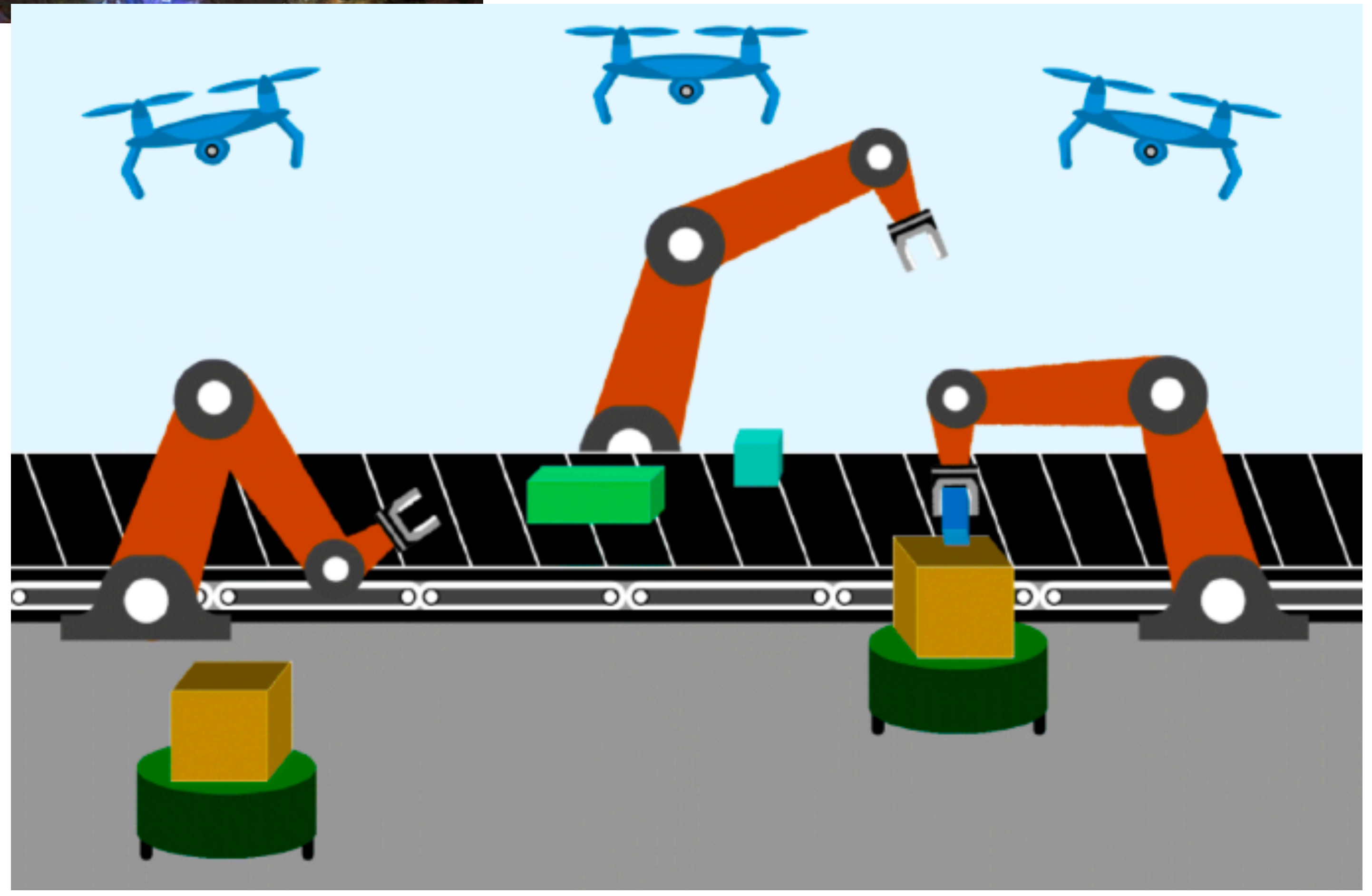
FLAIR Foerster Lab for AI Research

VECTOR INSTITUTE

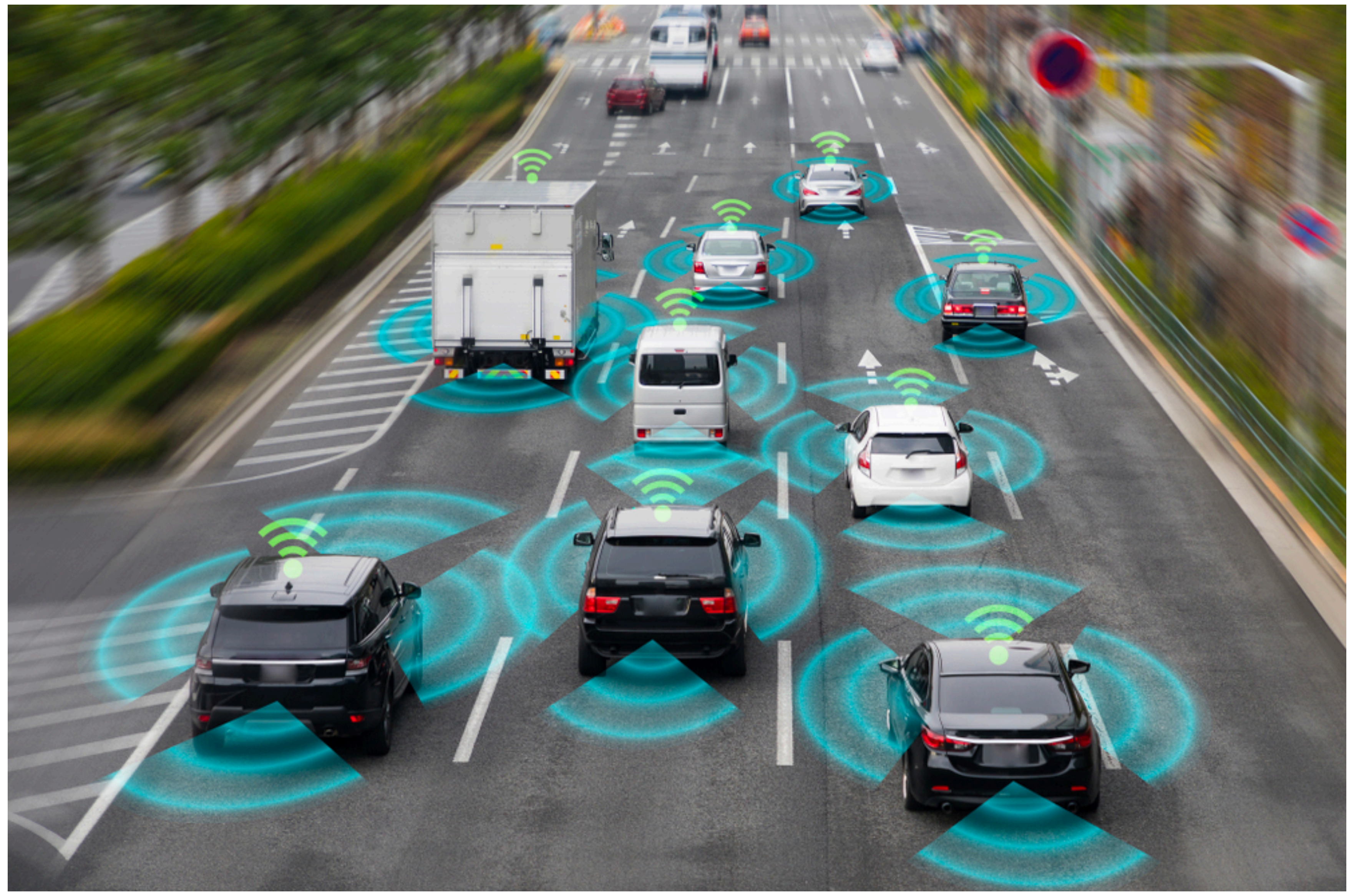
UNIVERSITY OF TORONTO

* Equal Contribution

# Multi-Agent Reinforcement Learning



# General-Sum Games



# Opponent Shaping & LOLA

- **LOLA**[1]**:** $\Delta\theta_1 = -\alpha \nabla_1 \left( L^1 \left( \theta_1, \theta_2 + \widehat{\Delta\theta_2} \right) \right)$, where $\widehat{\Delta\theta_2} = -\alpha \nabla_2 L^2$
- LOLA finds **Tit-for-Tat** in the Iterated Prisoner's Dilemma.
- LOLA is **inconsistent:** it assumes that the opponent is a **Naive Learner.**
- LOLA does not find Stable Fixed Points (SFPs).
- Stable Fixed Points are a popular solution concept.

[1] Foerster et al., 2017

# Contributions

# Higher-Order LOLA (HOLA) & iLOLA

- **HOLAn:**

$$h_1^{n+1} := -\alpha \nabla_1 \Big( L^1 \left(\theta_1, \theta_2 + h_2^n\right) \Big)$$

$$h_2^{n+1} := -\alpha \nabla_2 \Big( L^2 \left(\theta_1 + h_1^n, \theta_2\right) \Big)$$

**where** $h_1^{-1} = h_2^{-1} = 0$

- **iLOLA :=** $\lim_{n \to \infty} \begin{pmatrix} h_1^n \\ h_2^n \end{pmatrix}$ **if** $\mathrm{HOLA}n = \left(h_1^n, h_2^n\right)$ **converges pointwise as** $n \to \infty$

# Consistency

- **Consistency:** Any update functions $f_1 : \mathbb{R}^d \to \mathbb{R}^{d_1}$ and $f_2 : \mathbb{R}^d \to \mathbb{R}^{d_2}$ are consistent if they satisfy:

$$f_1\left(\theta_1, \theta_2\right) = -\alpha \nabla_1\left(L^1\left(\theta_1, \theta_2 + f_2\left(\theta_1, \theta_2\right)\right)\right)$$

$$f_2\left(\theta_1, \theta_2\right) = -\alpha \nabla_2\left(L^2\left(\theta_1 + f_1\left(\theta_1, \theta_2\right), \theta_2\right)\right)$$

- **Proposition:** *iLOLA is consistent under mutual opponent shaping*

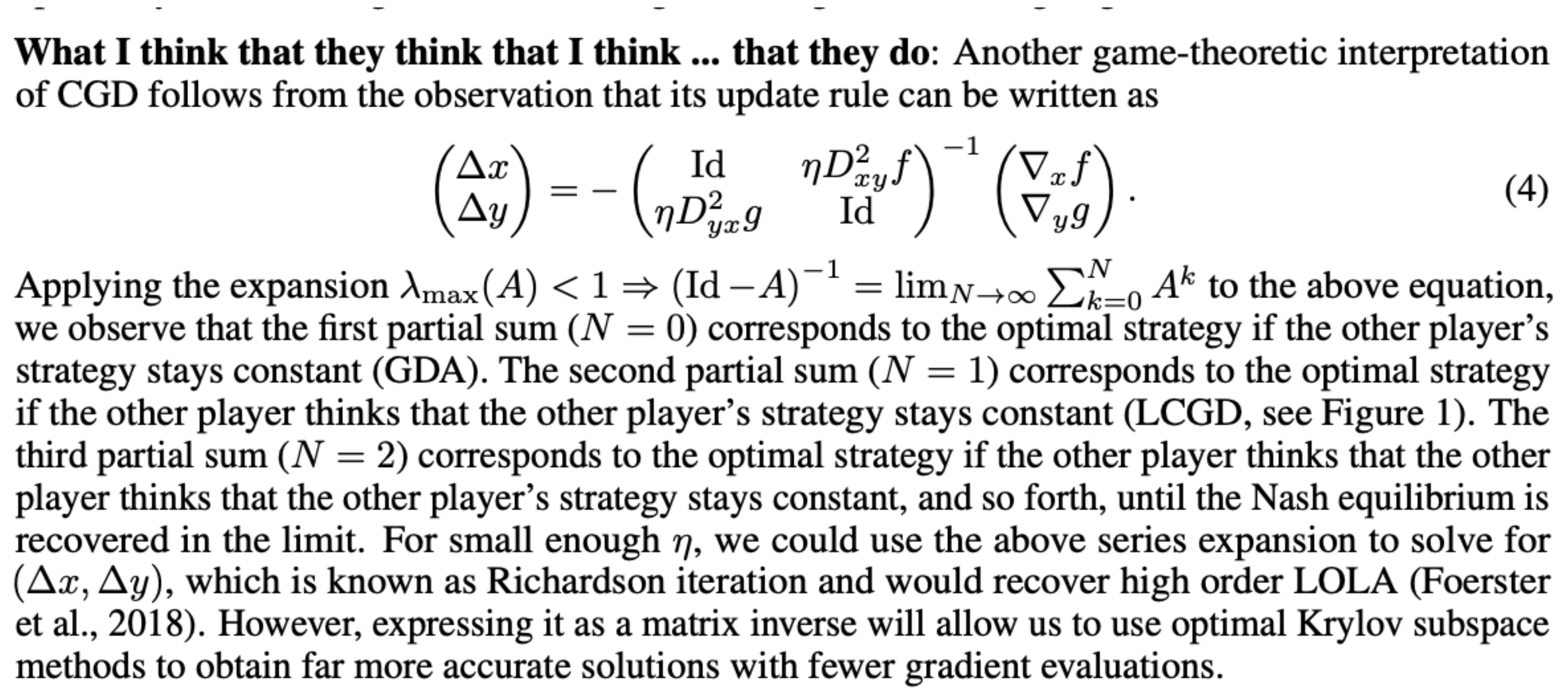
# Competitive Gradient Descent ≠ iLOLA

The authors of CGD[1] (2019) claim that:

1. the higher-order ``series-expansion [of CGD] would recover higher-order LOLA” (page 4)

2. ``LCGD [Linearized CGD] coincides with first-order LOLA” (page 6)

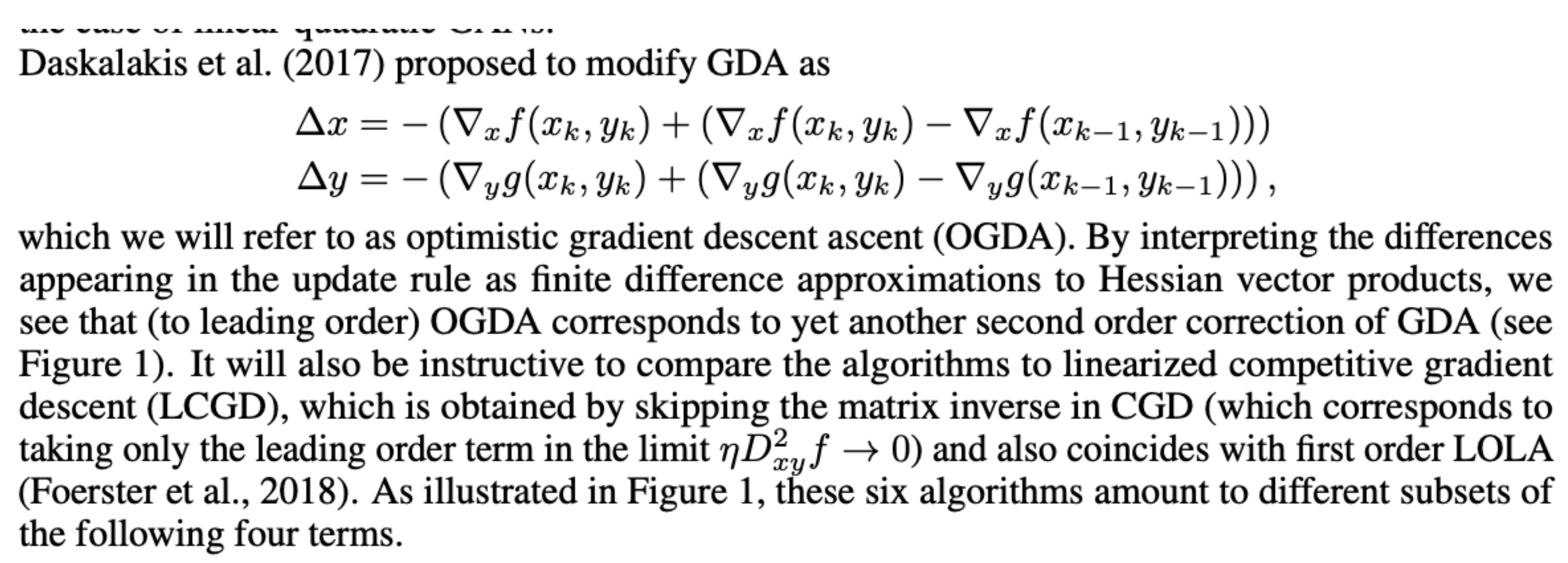
- **Proposition:** **CGD does not coincide with iLOLA and does not solve the inconsistency problem**



**What I think that they think that I think ... that they do**: Another game-theoretic interpretation of CGD follows from the observation that its update rule can be written as

$$\begin{pmatrix}\Delta x\\ \Delta y\end{pmatrix} = -\begin{pmatrix}\mathrm{Id} & \eta D_{xy}^2 f\\ \eta D_{yx}^2 g & \mathrm{Id}\end{pmatrix}^{-1}\begin{pmatrix}\nabla_x f\\ \nabla_y g\end{pmatrix}. \qquad (4)$$

Applying the expansion $\lambda_{\max}(A) < 1 \Rightarrow (\mathrm{Id} - A)^{-1} = \lim_{N\to\infty}\sum_{k=0}^{N} A^k$ to the above equation, we observe that the first partial sum ($N = 0$) corresponds to the optimal strategy if the other player's strategy stays constant (GDA). The second partial sum ($N = 1$) corresponds to the optimal strategy if the other player thinks that the other player's strategy stays constant (LCGD, see Figure 1). The third partial sum ($N = 2$) corresponds to the optimal strategy if the other player thinks that the other player thinks that the other player's strategy stays constant, and so forth, until the Nash equilibrium is recovered in the limit. For small enough $\eta$, we could use the above series expansion to solve for $(\Delta x, \Delta y)$, which is known as Richardson iteration and would recover high order LOLA (Foerster et al., 2018). However, expressing it as a matrix inverse will allow us to use optimal Krylov subspace methods to obtain far more accurate solutions with fewer gradient evaluations.



Daskalakis et al. (2017) proposed to modify GDA as

$$\Delta x = -\left(\nabla_x f(x_k, y_k) + (\nabla_x f(x_k, y_k) - \nabla_x f(x_{k-1}, y_{k-1}))\right)$$
$$\Delta y = -\left(\nabla_y g(x_k, y_k) + (\nabla_y g(x_k, y_k) - \nabla_y g(x_{k-1}, y_{k-1}))\right),$$

which we will refer to as optimistic gradient descent ascent (OGDA). By interpreting the differences appearing in the update rule as finite difference approximations to Hessian vector products, we see that (to leading order) OGDA corresponds to yet another second order correction of GDA (see Figure 1). It will also be instructive to compare the algorithms to linearized competitive gradient descent (LCGD), which is obtained by skipping the matrix inverse in CGD (which corresponds to taking only the leading order term in the limit $\eta D_{xy}^2 f \to 0$) and also coincides with first order LOLA (Foerster et al., 2018). As illustrated in Figure 1, these six algorithms amount to different subsets of the following four terms.

[1] Schäfer et al., 2019

# COLA: Consistent Learning with Opponent-Learning Awareness

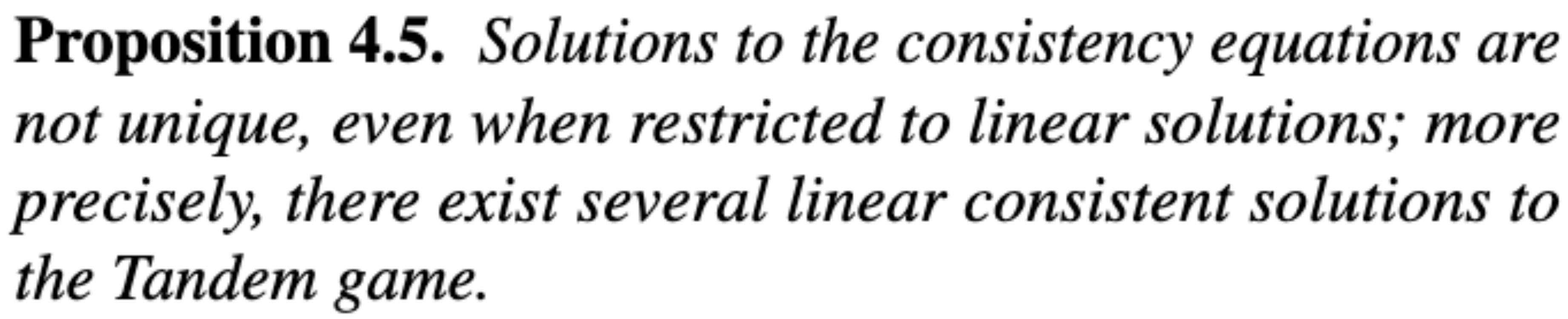
- Naive Solution to consistency: Iteratively compute higher orders of LOLA until convergence.
    - May diverge and requires arbitrarily high derivatives: expensive!
- We propose **COLA:** Learn $f_1$ and $f_2$ using neural networks

$$C_1\left(\phi_1, \phi_2, \theta_1, \theta_2\right) = \left\| f_1 + \alpha \nabla_1 \Big( L^1\left(\theta_1, \theta_2 + f_2\right) \Big) \right\|$$

$$C_2\left(\phi_1, \phi_2, \theta_1, \theta_2\right) = \left\| f_2 + \alpha \nabla_2 \Big( L^2\left(\theta_1 + f_1, \theta_2\right) \Big) \right\|$$

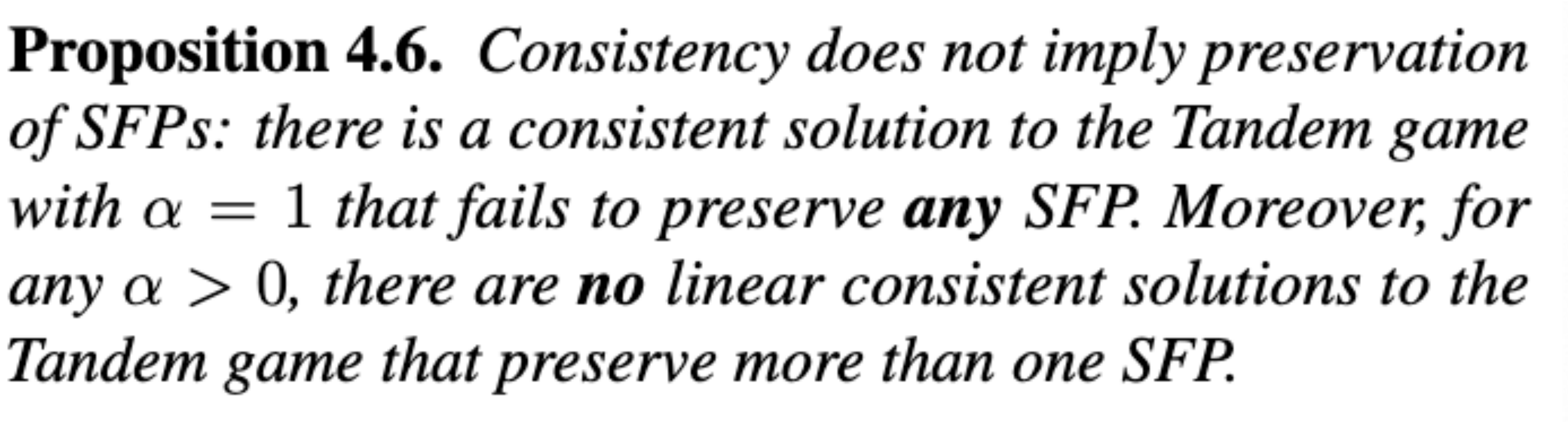
# COLA: Theoretical Results

COLA's solutions are not necessarily unique



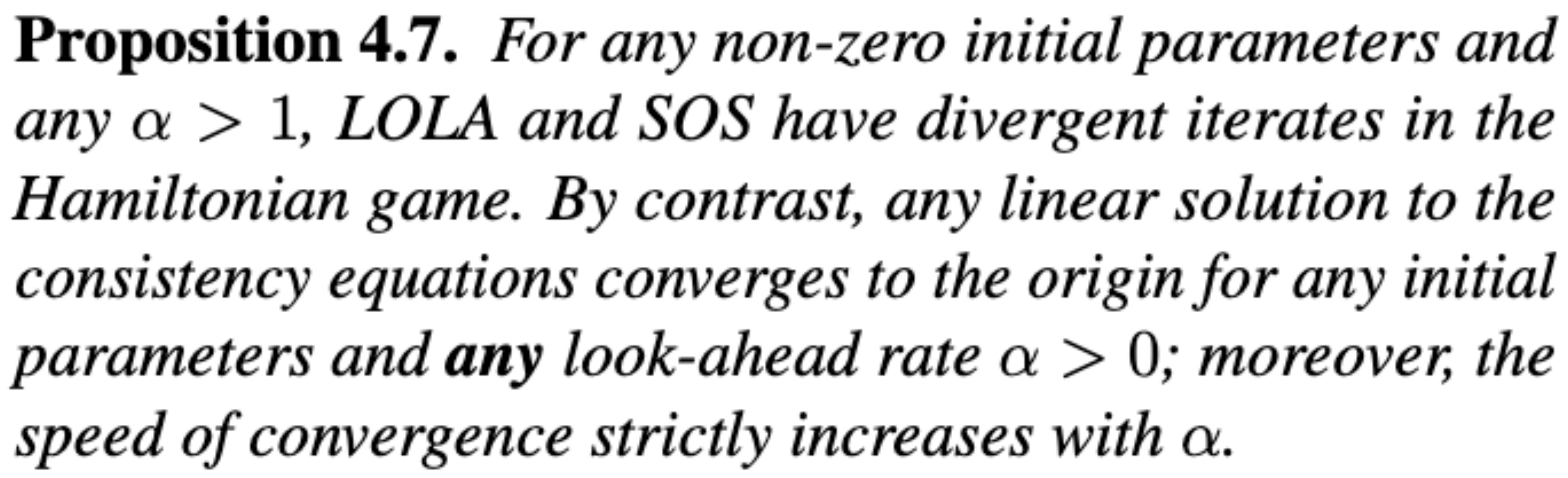
**Proposition 4.5.** *Solutions to the consistency equations are not unique, even when restricted to linear solutions; more precisely, there exist several linear consistent solutions to the Tandem game.*

Consistency does not imply preservation of SFPs



**Proposition 4.6.** *Consistency does not imply preservation of SFPs: there is a consistent solution to the Tandem game with $\alpha = 1$ that fails to preserve **any** SFP. Moreover, for any $\alpha > 0$, there are **no** linear consistent solutions to the Tandem game that preserve more than one SFP.*

COLA is more robust than LOLA



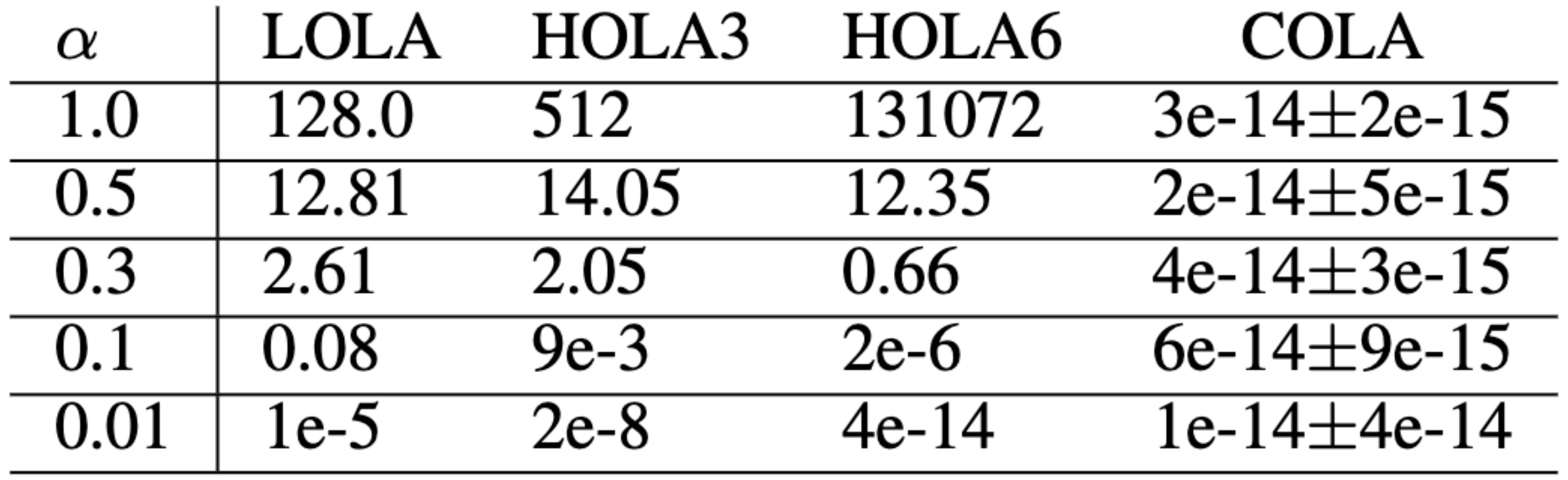
**Proposition 4.7.** *For any non-zero initial parameters and any $\alpha > 1$, LOLA and SOS have divergent iterates in the Hamiltonian game. By contrast, any linear solution to the consistency equations converges to the origin for any initial parameters and **any** look-ahead rate $\alpha > 0$; moreover, the speed of convergence strictly increases with $\alpha$.*

# Experiments

# COLA's Consistency

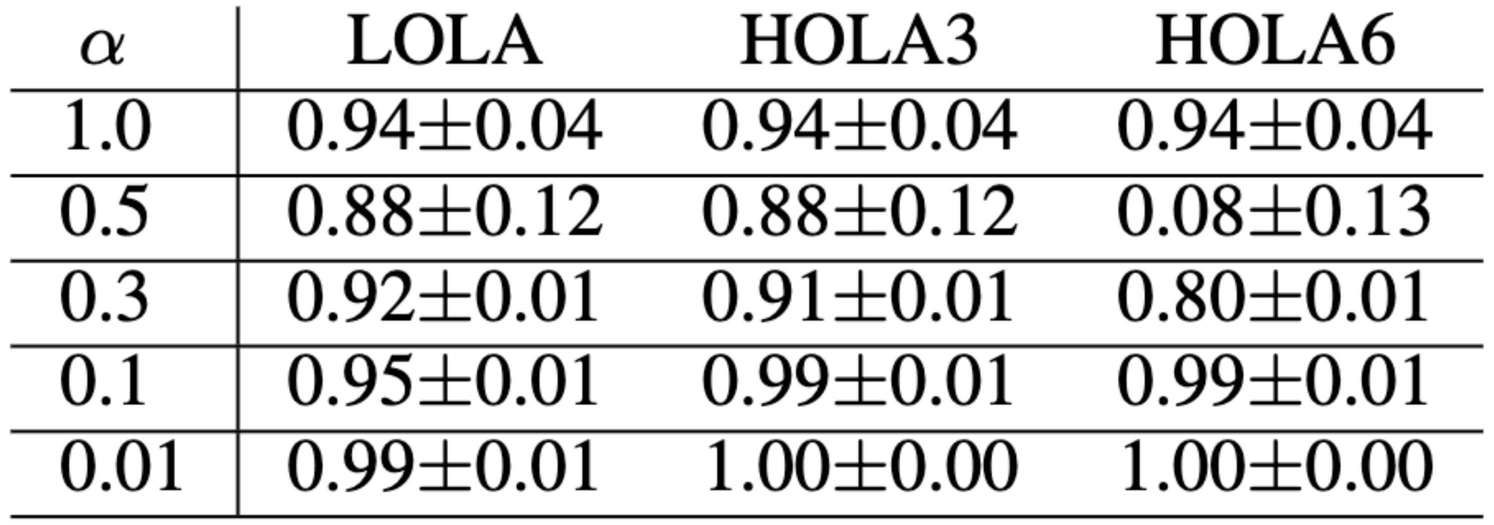
**COLA's solutions are not necessarily unique**

- COLA finds consistent update functions *even when HOLA diverges.*
- COLA's updates are *similar to iLOLA when HOLA converges*.
- COLA tends to find similar solutions over different runs (despite the theoretical result)



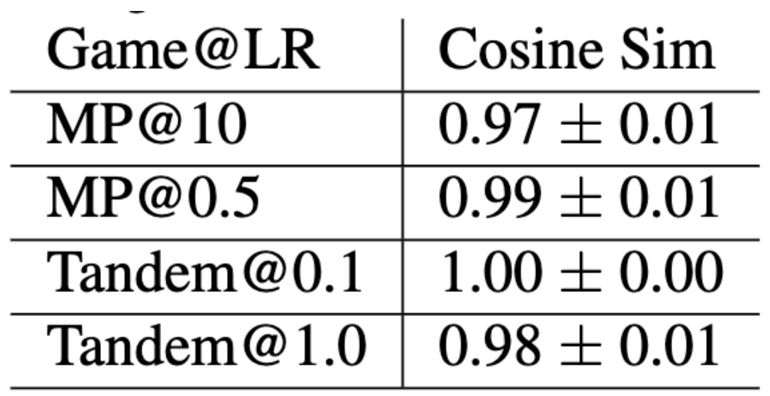
| $\alpha$ | LOLA | HOLA3 | HOLA6 | COLA |
|---|---|---|---|---|
| 1.0 | 128.0 | 512 | 131072 | 3e-14±2e-15 |
| 0.5 | 12.81 | 14.05 | 12.35 | 2e-14±5e-15 |
| 0.3 | 2.61 | 2.05 | 0.66 | 4e-14±3e-15 |
| 0.1 | 0.08 | 9e-3 | 2e-6 | 6e-14±9e-15 |
| 0.01 | 1e-5 | 2e-8 | 4e-14 | 1e-14±4e-14 |

**Table 1: Consistency Loss on Tandem**



| $\alpha$ | LOLA | HOLA3 | HOLA6 |
|---|---|---|---|
| 1.0 | 0.94±0.04 | 0.94±0.04 | 0.94±0.04 |
| 0.5 | 0.88±0.12 | 0.88±0.12 | 0.08±0.13 |
| 0.3 | 0.92±0.01 | 0.91±0.01 | 0.80±0.01 |
| 0.1 | 0.95±0.01 | 0.99±0.01 | 0.99±0.01 |
| 0.01 | 0.99±0.01 | 1.00±0.00 | 1.00±0.00 |

**Table 2: Similarity Scores on Tandem**



| Game@LR | Cosine Sim |
|---|---|
| MP@10 | 0.97 ± 0.01 |
| MP@0.5 | 0.99 ± 0.01 |
| Tandem@0.1 | 1.00 ± 0.00 |
| Tandem@1.0 | 0.98 ± 0.01 |

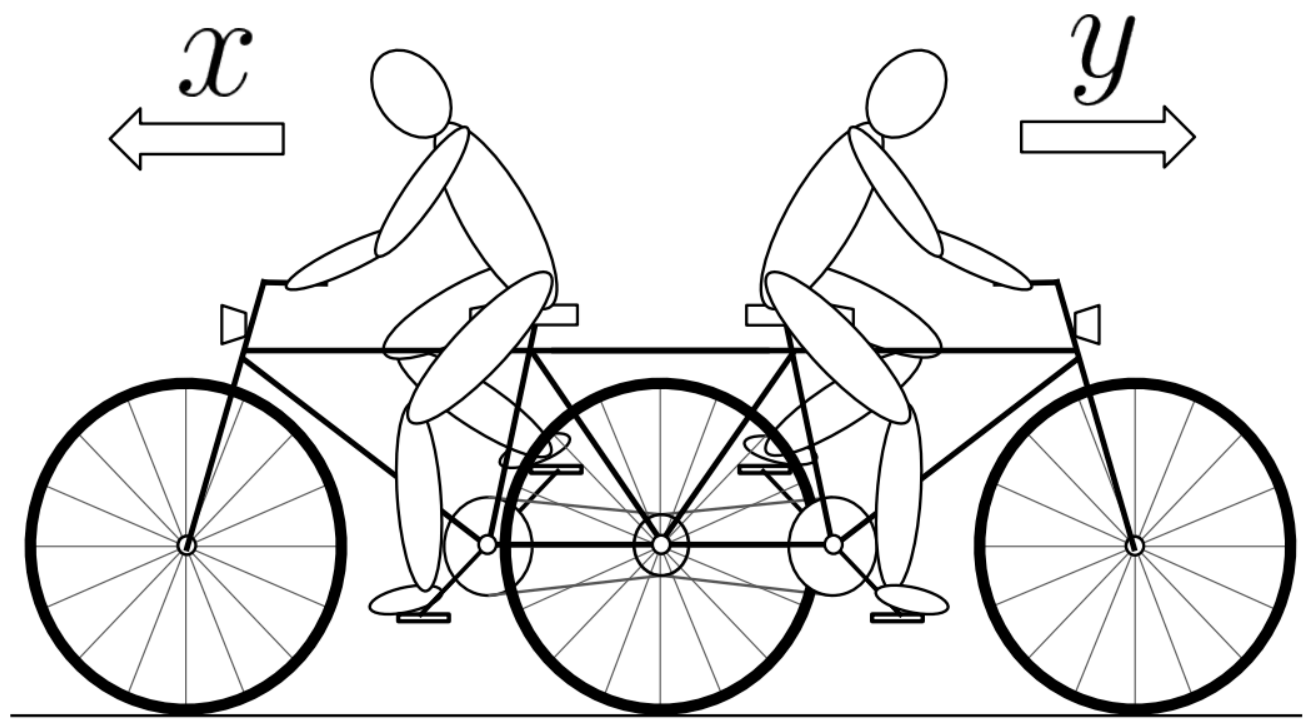
**Table 3: Self-Similarity Scores**

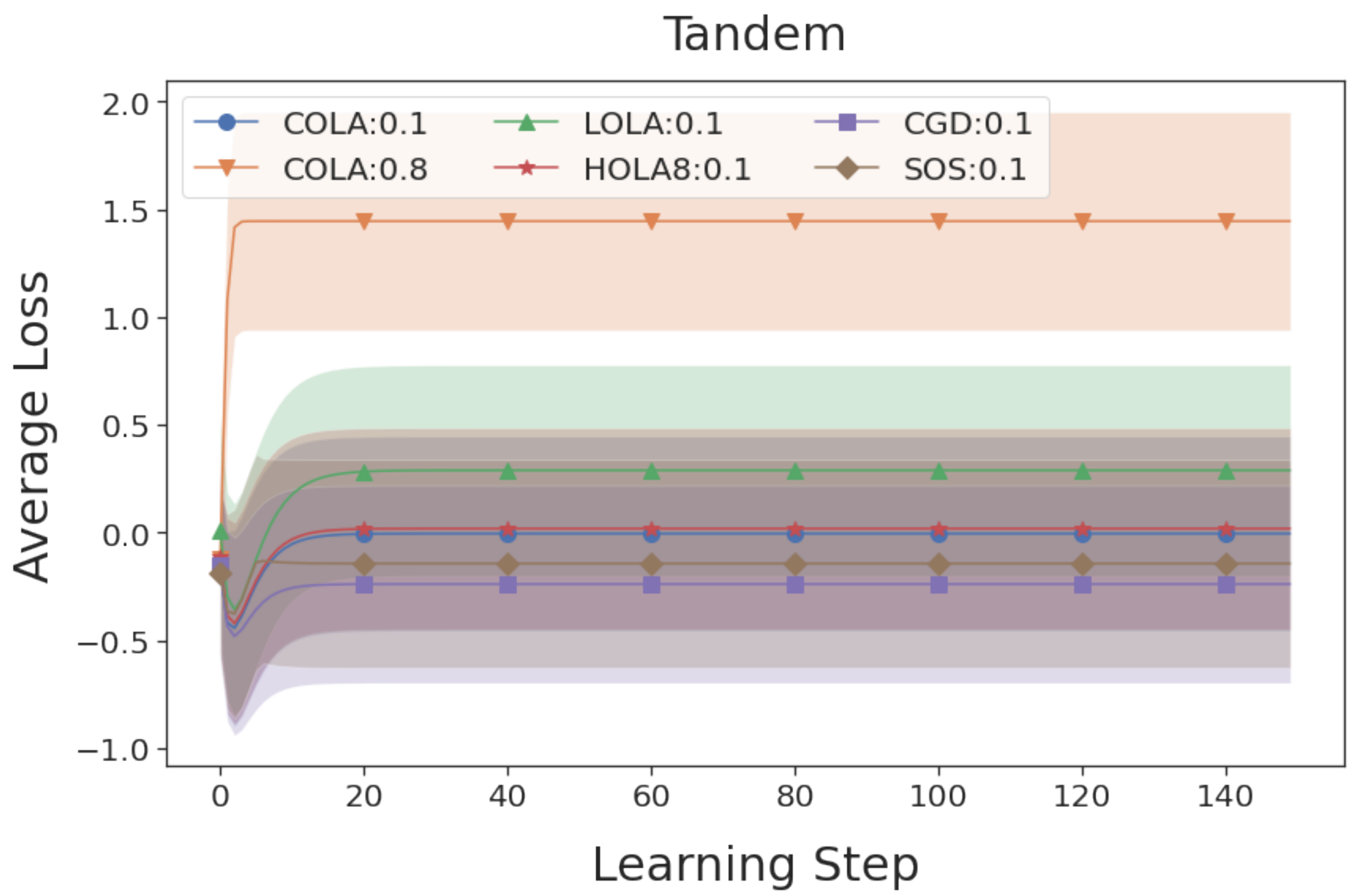
# Tandem Game

**Consistency does not imply preservation of SFPs**

$$L^1(x,y) = (x+y)^2 - 2x$$

$$L^2(x,y) = (x+y)^2 - 2y$$





- COLA and HOLA converge to similar solutions
- CGD converges to a different solution than COLA and HOLA8
- COLA does not recover SFP

# Iterated Prisoner’s Dilemma

**CGD does not coincide with iLOLA and does not solve the inconsistency problem**

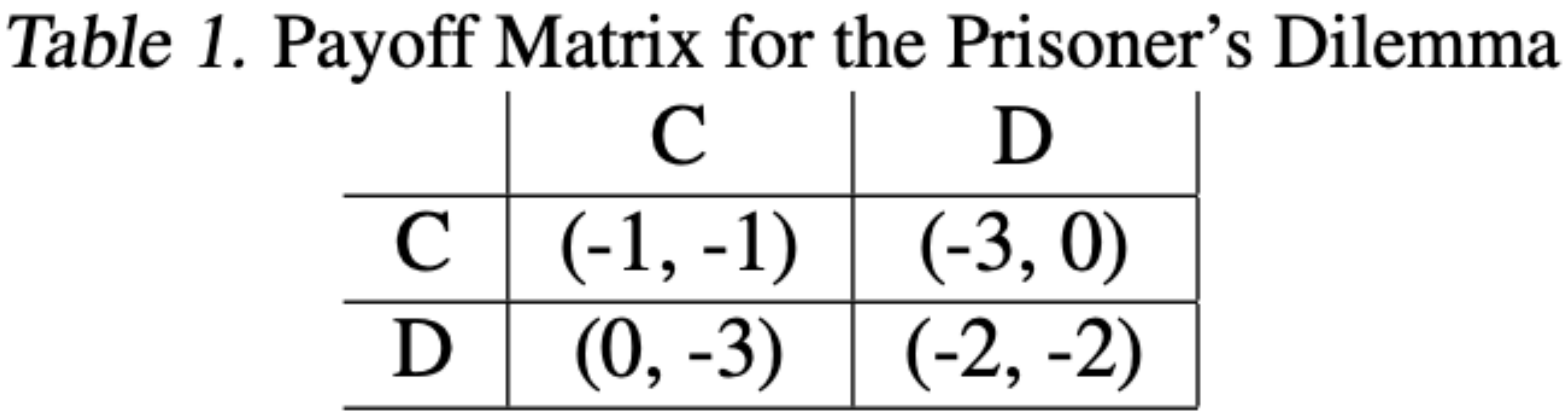
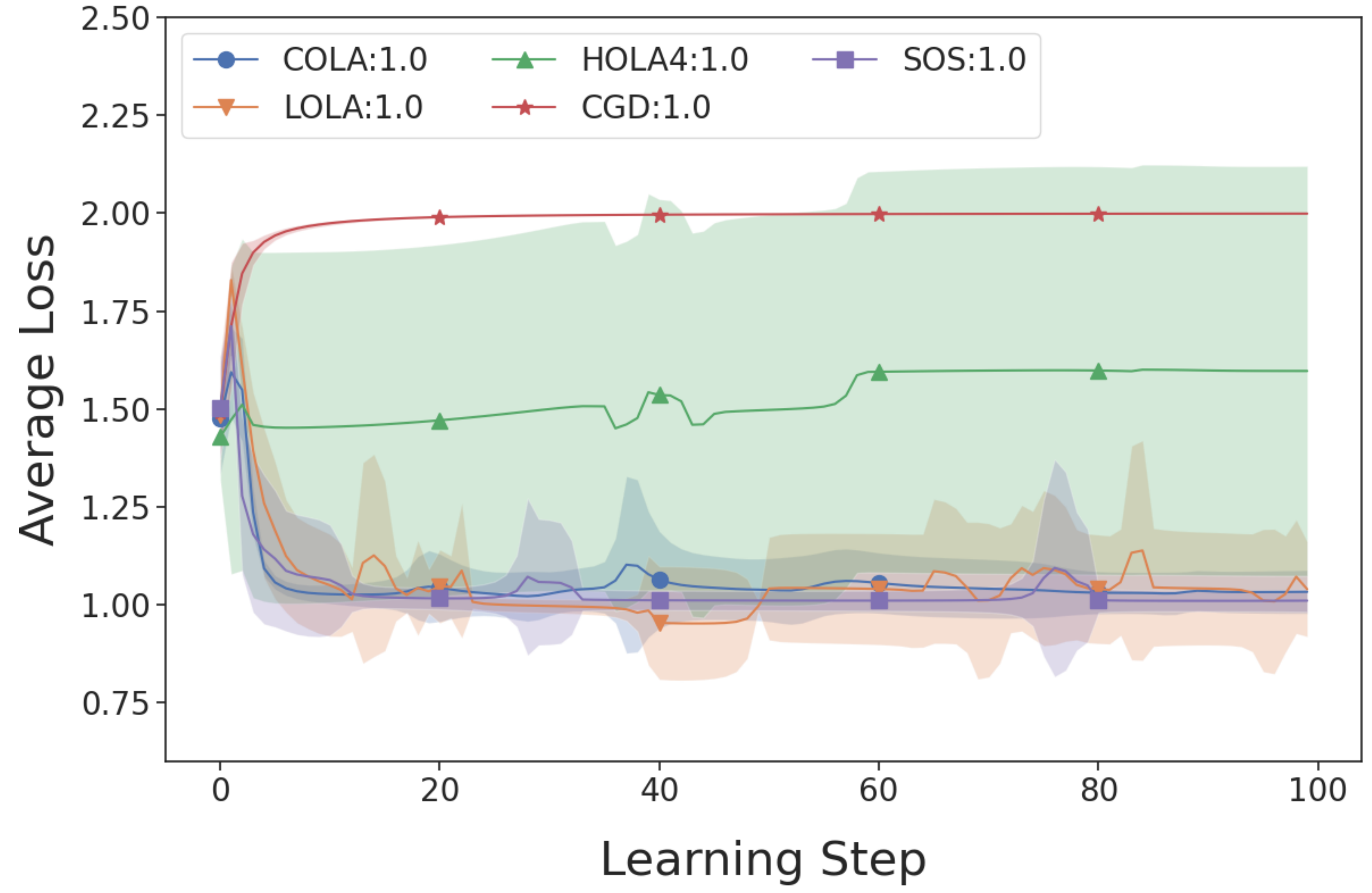
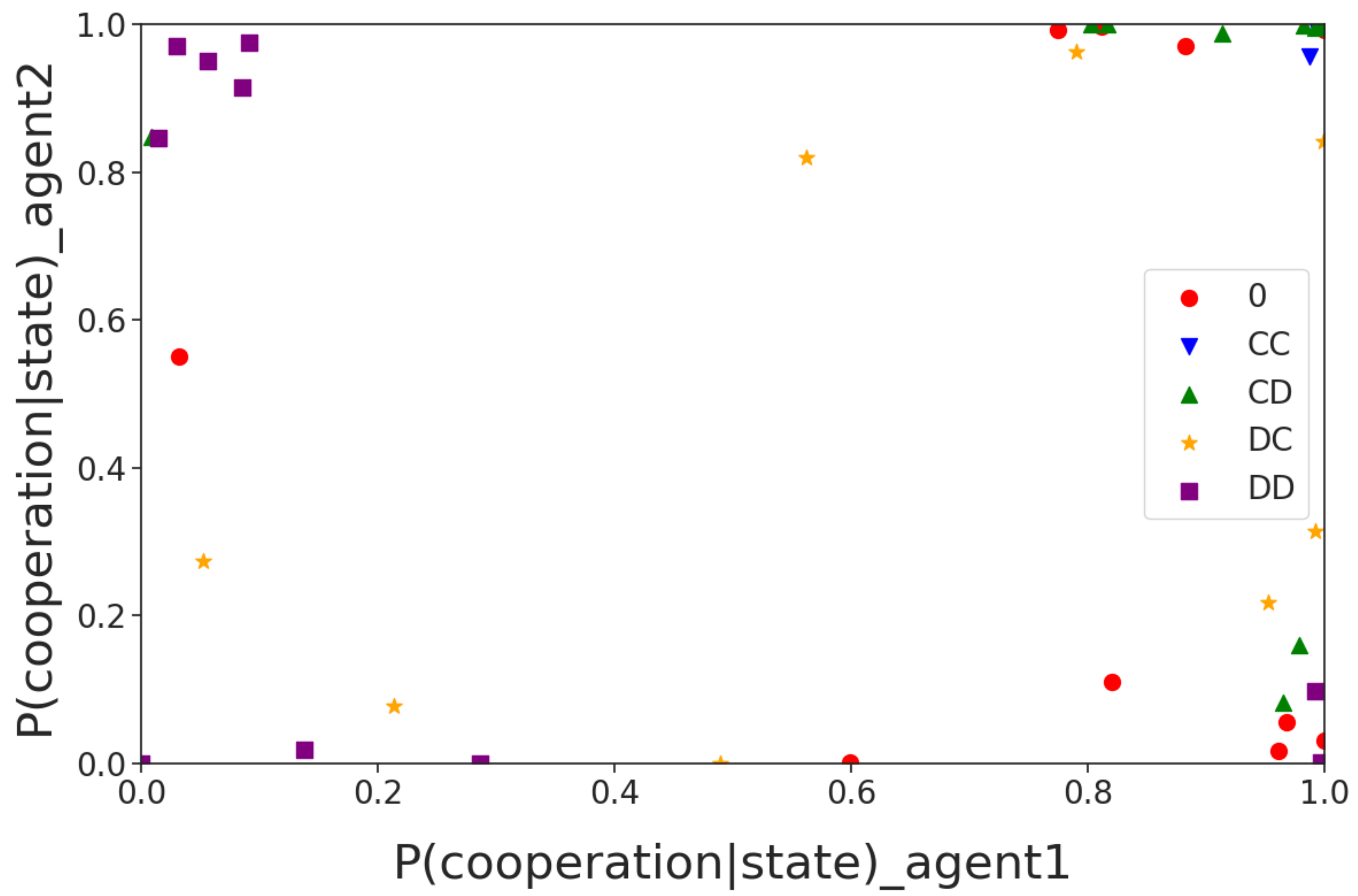


*Table 1.* Payoff Matrix for the Prisoner’s Dilemma

|   | C | D |
|---|---|---|
| C | (-1, -1) | (-3, 0) |
| D | (0, -3) | (-2, -2) |

- COLA finds prosocial solution
- COLA’s policy is similar to Tit-for-Tat
- CGD does not find prosocial solution



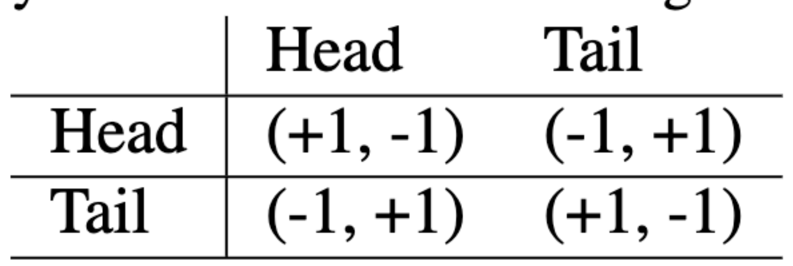


# Matching Pennies

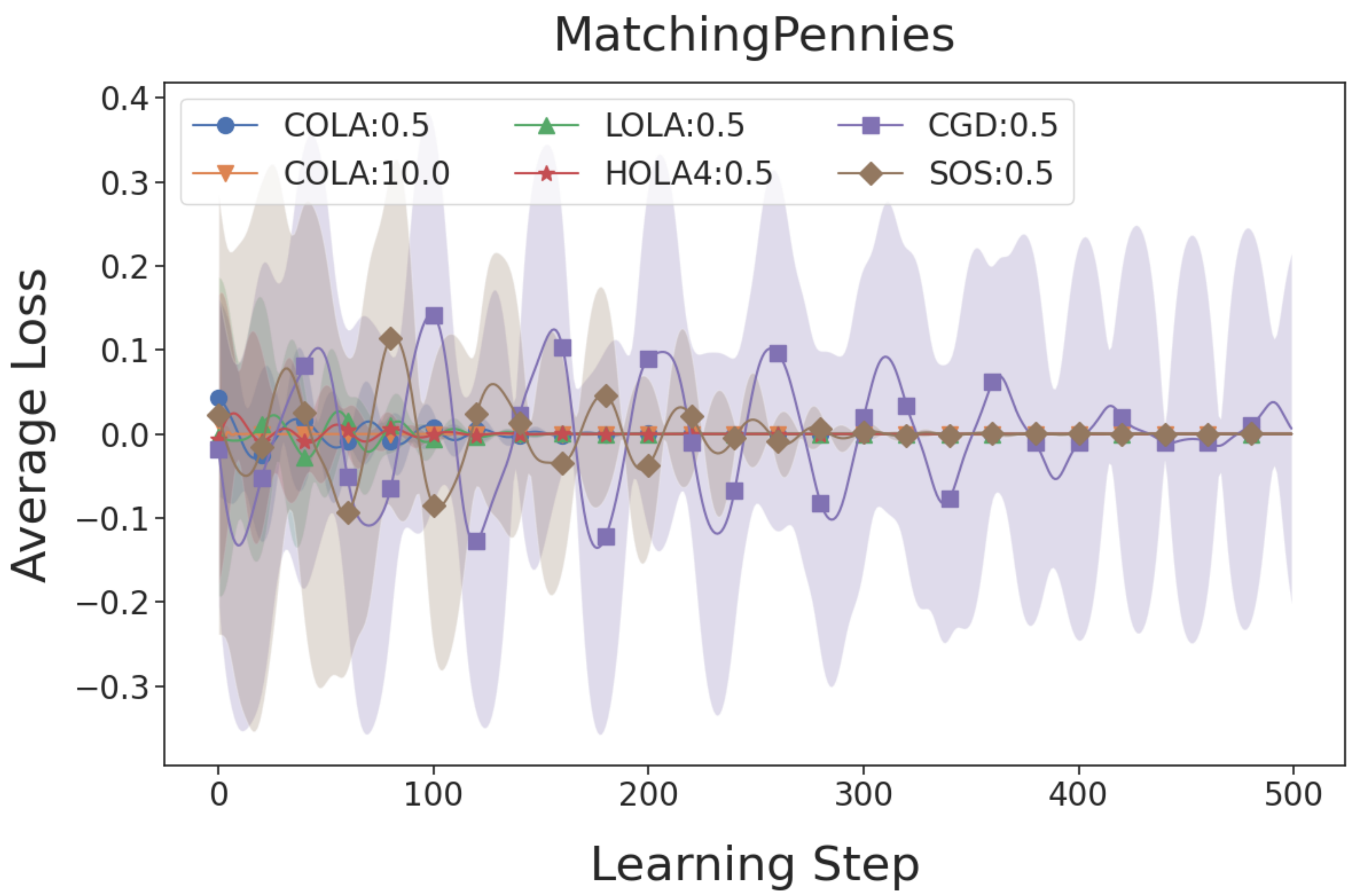
**COLA is more robust than LOLA**

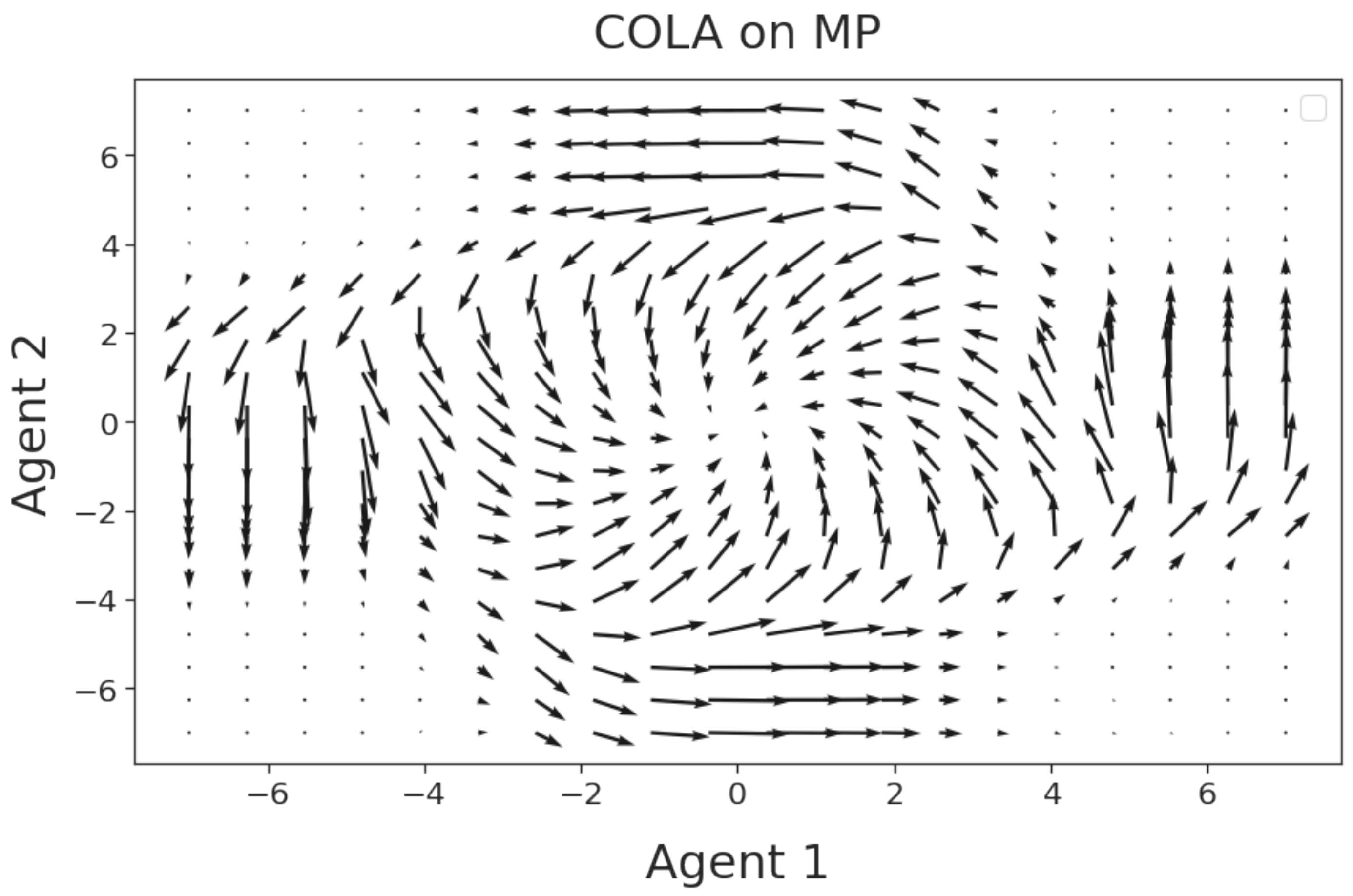
- COLA converges robustly and fast to a wide range of hyperparameter values

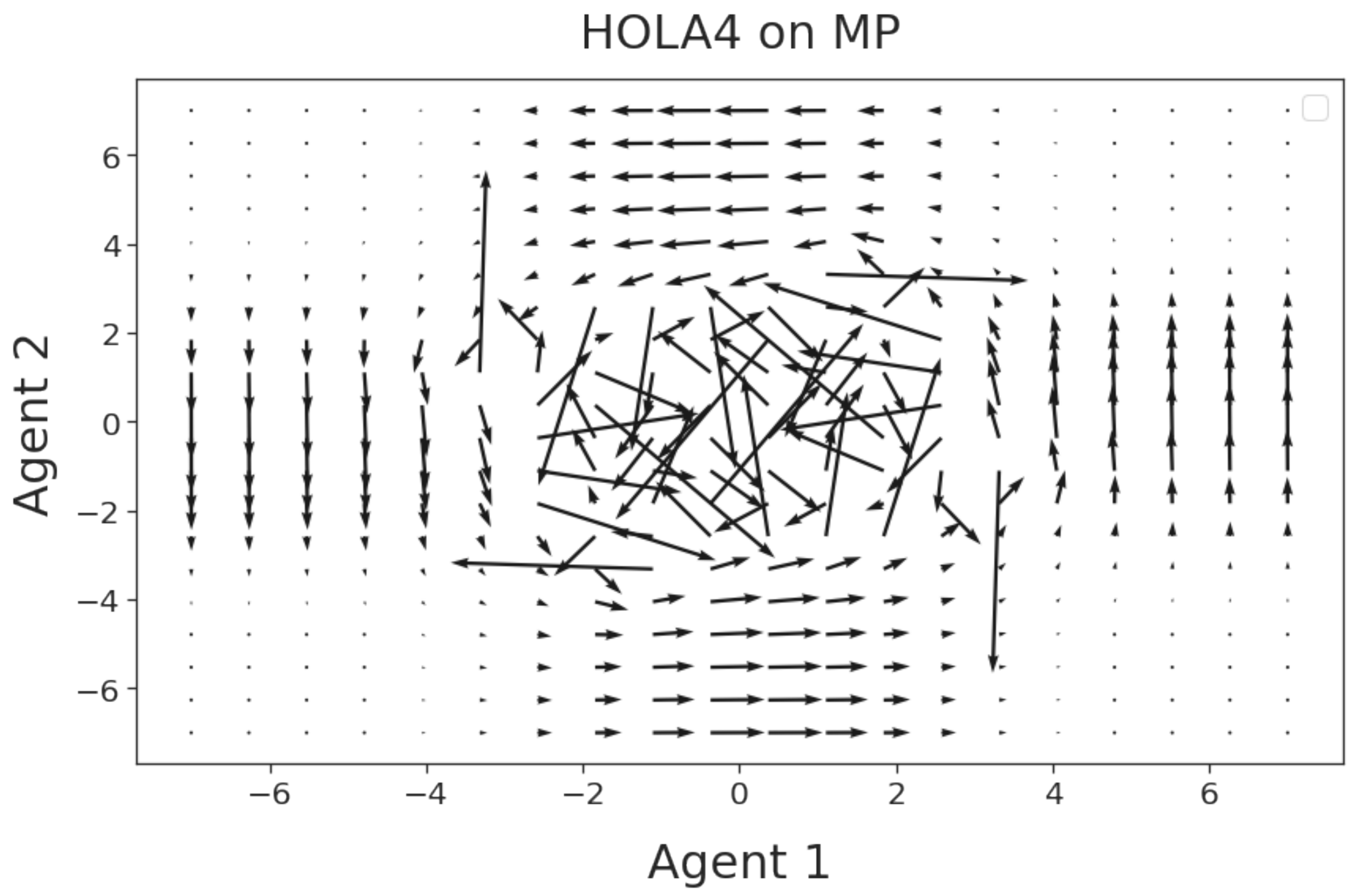
*Table 6.* Payoff Matrix for the Matching Pennies game.



| | Head | Tail |
|---|---|---|
| Head | (+1, -1) | (-1, +1) |
| Tail | (-1, +1) | (+1, -1) |





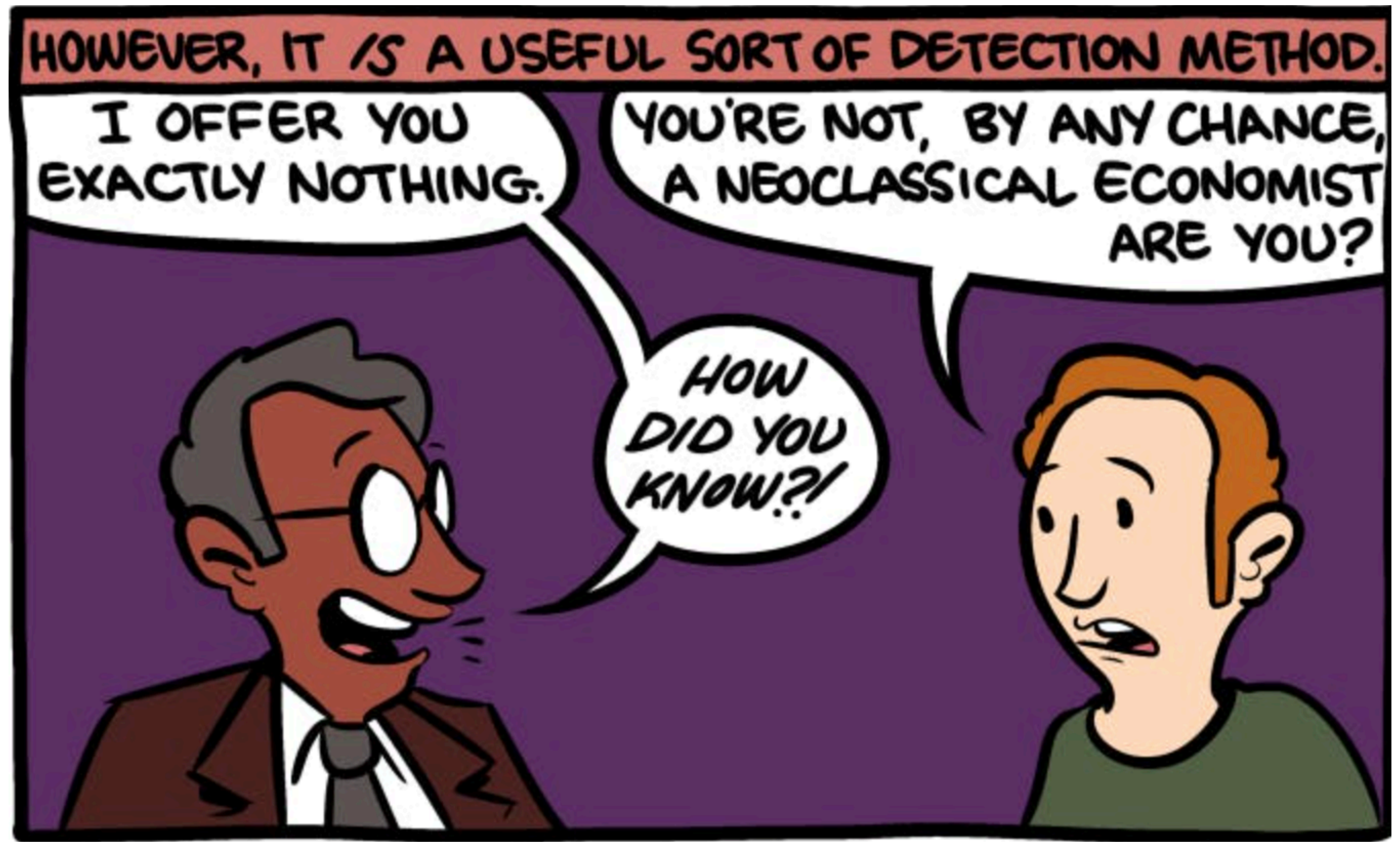


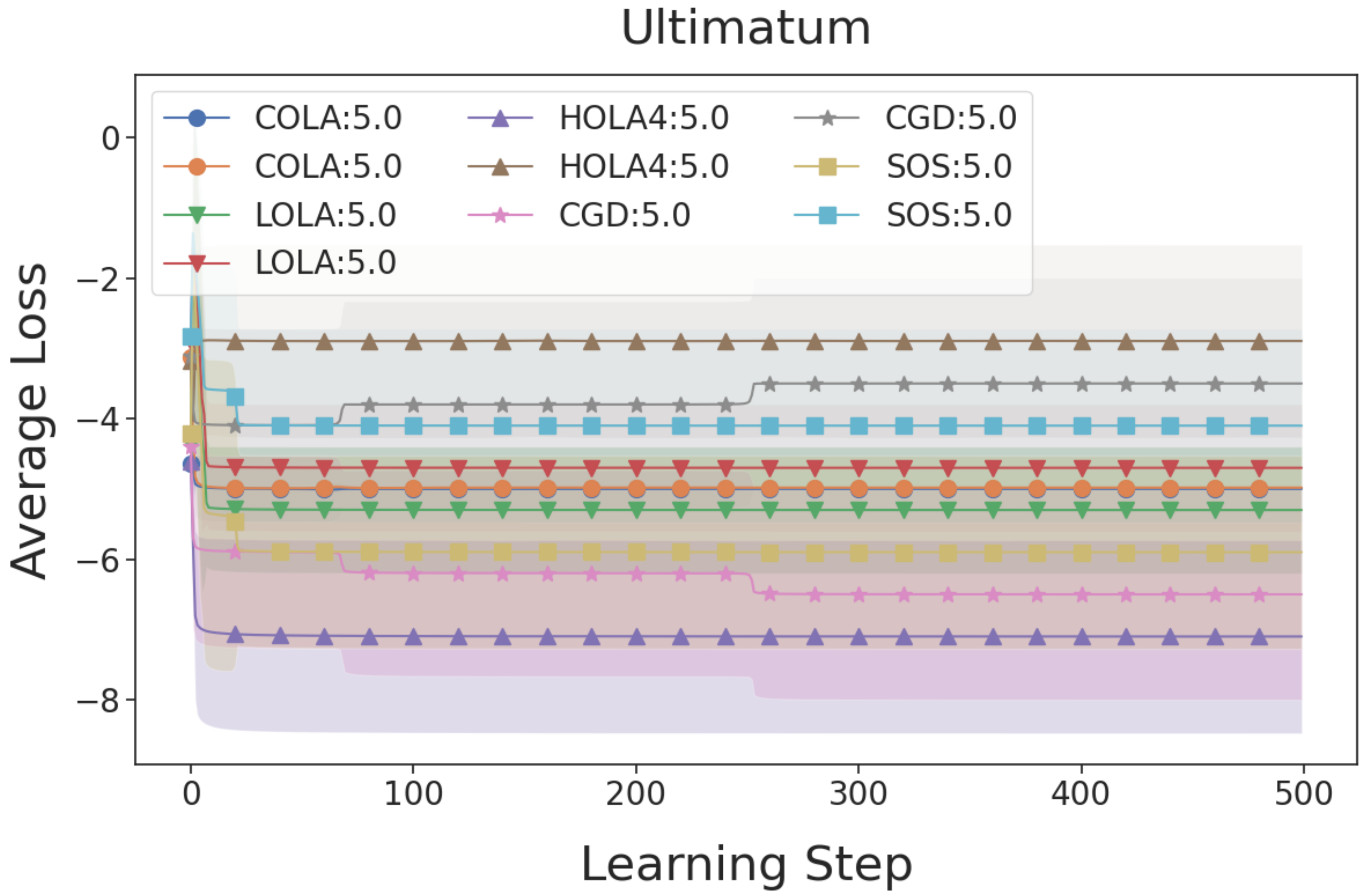
# Ultimatum Game

$$L_1 = -\left(5p_{\textbf{fair}} + 8\left(1 - p_{\textbf{fair}}\right)p_{\textbf{accept}}\right)$$

$$L_2 = -\left(5p_{\textbf{fair}} + 2\left(1 - p_{\textbf{fair}}\right)p_{\textbf{accept}}\right)$$

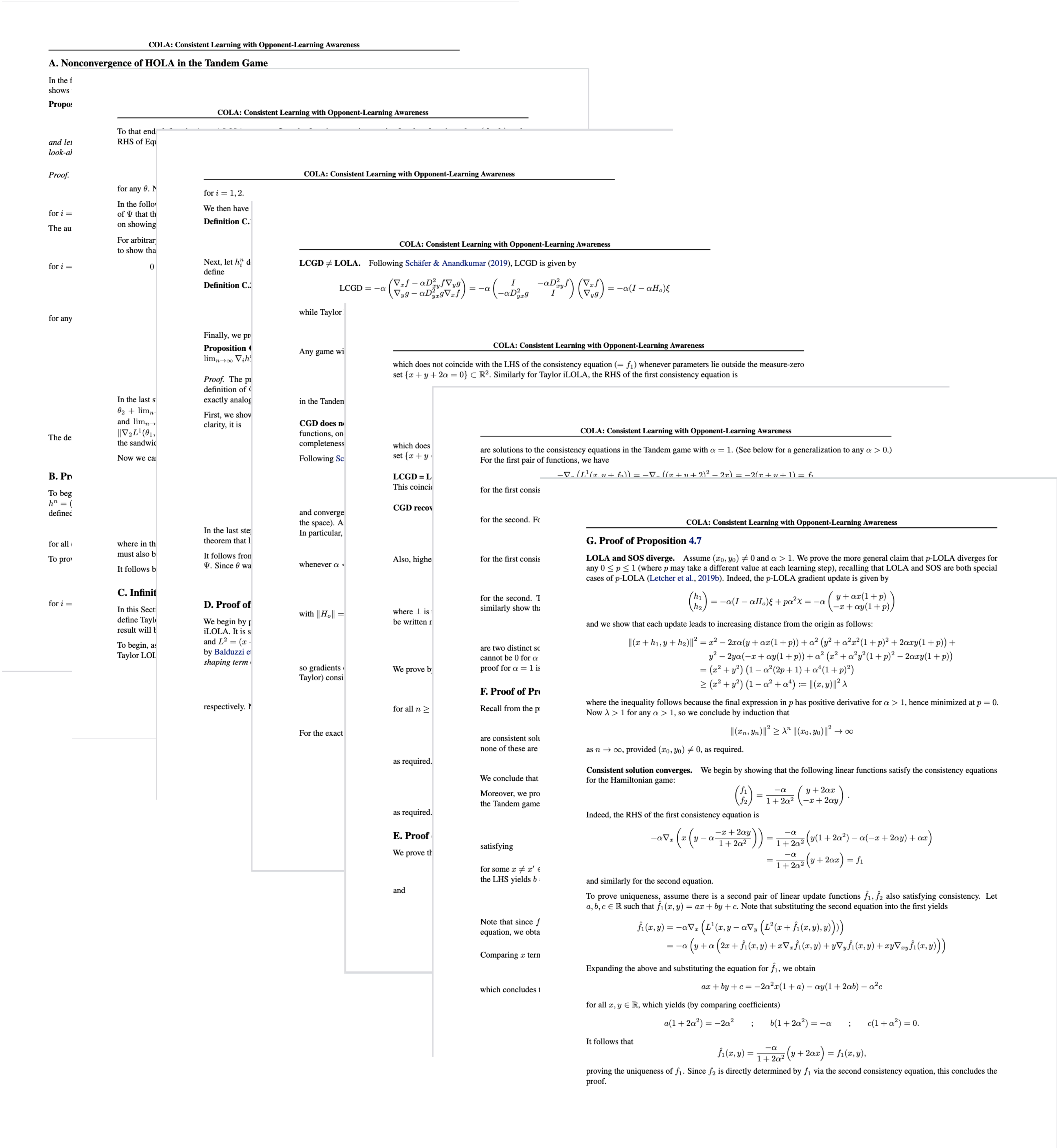
- Player 1 has access to 10$
  - Option 1: Split money 50/50
  - Option 2: Split money 80/20
- Player 2 can accept or decline
- *COLA is the only algorithm consistently converging to the fair solution*


HOWEVER, IT IS A USEFUL SORT OF DETECTION METHOD.
I OFFER YOU EXACTLY NOTHING.
YOU'RE NOT, BY ANY CHANCE, A NEOCLASSICAL ECONOMIST ARE YOU?
HOW DID YOU KNOW?!



# Conclusion

- Corrected a claim made in prior work
- iLOLA solves part of the consistency problem of LOLA
- Even with consistency, opponent shaping does not preserve SFPs
- We introduced **COLA**
- COLA tends to find prosocial solutions



COLA: Consistent Learning with Opponent-Learning Awareness

## G. Proof of Proposition 4.7

**LOLA and SOS diverge.** Assume $(x_0, y_0) \neq 0$ and $\alpha > 1$. We prove the more general claim that $p$-LOLA diverges for any $0 \leq p \leq 1$ (where $p$ may take a different value at each learning step), recalling that LOLA and SOS are both special cases of $p$-LOLA (Letcher et al., 2019b). Indeed, the $p$-LOLA gradient update is given by

$$\begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = -\alpha(I - \alpha H_o)\xi + p\alpha^2\chi = -\alpha \begin{pmatrix} y + \alpha x(1+p) \\ -x + \alpha y(1+p) \end{pmatrix}$$

and we show that each update leads to increasing distance from the origin as follows:

$$\begin{aligned} \|(x+h_1, y+h_2)\|^2 &= x^2 - 2x\alpha(y + \alpha x(1+p)) + \alpha^2\left(y^2 + \alpha^2 x^2(1+p)^2 + 2\alpha xy(1+p)\right) + \\ &\quad y^2 - 2y\alpha(-x + \alpha y(1+p)) + \alpha^2\left(x^2 + \alpha^2 y^2(1+p)^2 - 2\alpha xy(1+p)\right) \\ &= \left(x^2+y^2\right)\left(1 - \alpha^2(2p+1) + \alpha^4(1+p)^2\right) \\ &\geq \left(x^2+y^2\right)\left(1 - \alpha^2 + \alpha^4\right) := \|(x,y)\|^2 \lambda \end{aligned}$$

where the inequality follows because the final expression in $p$ has positive derivative for $\alpha > 1$, hence minimized at $p = 0$. Now $\lambda > 1$ for any $\alpha > 1$, so we conclude by induction that

$$\|(x_n, y_n)\|^2 \geq \lambda^n \|(x_0, y_0)\|^2 \to \infty$$

as $n \to \infty$, provided $(x_0, y_0) \neq 0$, as required.

**Consistent solution converges.** We begin by showing that the following linear functions satisfy the consistency equations for the Hamiltonian game:

$$\begin{pmatrix} f_1 \\ f_2 \end{pmatrix} = \frac{-\alpha}{1+2\alpha^2} \begin{pmatrix} y + 2\alpha x \\ -x + 2\alpha y \end{pmatrix}.$$

Indeed, the RHS of the first consistency equation is

$$\begin{aligned} -\alpha\nabla_x\left(x\left(y - \alpha\frac{-x+2\alpha y}{1+2\alpha^2}\right)\right) &= \frac{-\alpha}{1+2\alpha^2}\left(y(1+2\alpha^2) - \alpha(-x+2\alpha y) + \alpha x\right) \\ &= \frac{-\alpha}{1+2\alpha^2}\left(y + 2\alpha x\right) = f_1 \end{aligned}$$

and similarly for the second equation.

To prove uniqueness, assume there is a second pair of linear update functions $\hat{f}_1, \hat{f}_2$ also satisfying consistency. Let $a, b, c \in \mathbb{R}$ such that $\hat{f}_1(x,y) = ax + by + c$. Note that substituting the second equation into the first yields

$$\begin{aligned} \hat{f}_1(x,y) &= -\alpha\nabla_x\left(L^1(x, y - \alpha\nabla_y\left(L^2(x + \hat{f}_1(x,y), y)\right))\right) \\ &= -\alpha\left(y + \alpha\left(2x + \hat{f}_1(x,y) + x\nabla_x\hat{f}_1(x,y) + y\nabla_y\hat{f}_1(x,y) + xy\nabla_{xy}\hat{f}_1(x,y)\right)\right) \end{aligned}$$

Expanding the above and substituting the equation for $\hat{f}_1$, we obtain

$$ax + by + c = -2\alpha^2 x(1+a) - \alpha y(1+2\alpha b) - \alpha^2 c$$

for all $x, y \in \mathbb{R}$, which yields (by comparing coefficients)

$$a(1+2\alpha^2) = -2\alpha^2 \quad ; \quad b(1+2\alpha^2) = -\alpha \quad ; \quad c(1+\alpha^2) = 0.$$

It follows that

$$\hat{f}_1(x,y) = \frac{-\alpha}{1+2\alpha^2}\left(y + 2\alpha x\right) = f_1(x,y),$$

proving the uniqueness of $f_1$. Since $f_2$ is directly determined by $f_1$ via the second consistency equation, this concludes the proof.

# Thank you!