# Cheap Talking Algorithms

Daniele Condorelli and Massimiliano Furlan

University of Warwick

June 12, 2024

# Introduction

We let two independent reinforcement learning agents play repeatedly a discretized version of the Crawford and Sobel (1982) (CS) game

We show agents converge to behaviour close to the ex-ante optimal or second best equilibrium of the game

Results are robust to changes in the reinforcement learning hyperparameters and to different specifications of the game

Motivation: (computational) learning-approach to equilibrium selection

# Relevant Literature

**Other computational work:**

- Evolutionary perspective on language (Skyrms, 2010);
- Communication games with aligned AI agents (Foerster et al., 2016; Lazaridou et al., 2016; Havrylov and Titov, 2017);
- Communication with partially aligned AI agents (Noukhovitch et al., 2021)

**Equilibrium Selection in Cheap Talk games:**

- Reinforcement learning to model bounded rationality (Erev and Roth, 1998);
- Equilibrium selection in games of information transmission (Chen et al. (2008), Blume et al. (1993), Gordon et al. (2022))

# Discretized Cheap Talk Game

Two agents, a sender ($S$) and a receiver ($R$)

Set of states, $\Theta$, is formed by $n$ linearly spaced points in $[0, 1]$

Set of messages, $M$, has $n$ elements

Set of actions, $A$, is formed by $2n - 1$ linearly spaced points in $[0, 1]$

Distribution of states, $p$, is known and has full support over $\Theta$

Utilities are $u_S(\theta, a) = -(a - \theta - b)^2$ and $u_R(\theta, a) = -(a - \theta)^2$; $b \geq 0$

# Discretized Cheap Talk Game (contd)

**Timing:**

A state $\theta \in \Theta$ is drawn accordig to $p$

The sender observes $\theta$ and sends a message $m \in M$ to the receiver

The receiver observes message $m$ and takes an action $a \in A$

Agents get their utilities $u_S(\theta, a)$, $u_R(\theta, a)$

**Equilibria:**

Frug (2016): If utilities are concave and the sender is upwardly biased the ex-ante receiver-optimal equilibrium is monotone partitional

In uniform-quadratic case, there is a single Pareto optimal equilibrium

# Simulations: Q-Learning

In each period $t = 1, ..., T$:

1) a state for $S$ is independently drawn from $\Theta$ according to $p$

2) $S$ chooses a message in $M$ which represents the state for $R$

3) $R$ takes an action from $A$

The choice $\pi_t(\cdot \mid s)$ of an agent at period $t$ in state $s$ is determined by

$$\pi_t(a \mid s) = \frac{e^{Q_t(s,a)/\tau_t}}{\sum_{a' \in A} e^{Q_t(s,a')/\tau_t}}$$

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha \left[ r_t(s,a) - Q_t(s,a) \right]$$

$$\tau_t = e^{-\lambda(t-1)}$$

where: $r_t(s,a)$ is the payoff in period $t$, $\alpha$ is the learning rate, $\lambda$ is the temperature decay rate and $Q_0(s,a)$ arbitrarily initialized.

# Illustration

$Q(s,a)$

| | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ |
|---|---|---|---|---|---|
| $s_1$ | 0.7 | 0.2 | 0.4 | 0.9 | 0.1 |
| $s_2$ | 0.2 | 0.4 | 0.1 | 0.7 | 0.6 |
| $s_3$ | 0.3 | 0.6 | 0.8 | 0.5 | 0.7 |
| $s_4$ | 0.9 | 0.3 | 0.6 | 0.2 | 0.8 |
| $s_5$ | 0.5 | 0.1 | 0.7 | 0.4 | 0.3 |
| $s_6$ | 0.8 | 0.5 | 0.2 | 0.4 | 0.9 |

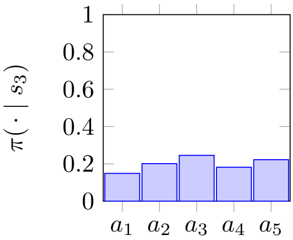



Figure: Softmax on $Q(s_3, \cdot)/\tau$ with $\tau = 1$. The probability mass is almost uniform over A.

# Illustration (contd)

$Q(s,a)$

| | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ |
|---|---|---|---|---|---|
| $s_1$ | 0.7 | 0.2 | 0.4 | 0.9 | 0.1 |
| $s_2$ | 0.2 | 0.4 | 0.1 | 0.7 | 0.6 |
| $s_3$ | 0.3 | 0.6 | 0.8 | 0.5 | 0.7 |
| $s_4$ | 0.9 | 0.3 | 0.6 | 0.2 | 0.8 |
| $s_5$ | 0.5 | 0.1 | 0.7 | 0.4 | 0.3 |
| $s_6$ | 0.8 | 0.5 | 0.2 | 0.4 | 0.9 |

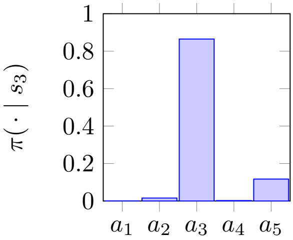



Figure: Softmax on $Q(s_3, \cdot)/\tau$ with $\tau = 0.05$. The probability mass is very concentrated on the most rewarding action.

# Analysis

We analyze behavior at convergence: $\pi^S_\infty$ and $\pi^R_\infty$

- a simulation converges if policies exhibit relative deviations in $L_{2,2}$ norm smaller than 0.1% for $10^4$ consecutive periods;

- all simulations converged

We run 1000 simulations for each bias level $b \in \{0, 0.005, \ldots, 0.495, 0.5\}$

We compare average outcomes against the equilibria for:

- ex-ante expected utilities;

- informativeness of the sender's strategy.

We also look how close to equilibrium the agents play in strategy space

# Metrics

**Ex-ante expected utility**

Ex-ante expected utility of the agents at convergence is

$$U_S = -\sum_{\theta} p(\theta) \sum_{m} \pi_{\infty}^{S}(m \mid \theta) \sum_{a} \pi_{\infty}^{R}(a \mid m)(a - \theta - b)^2$$

$$U_R = -\sum_{\theta} p(\theta) \sum_{m} \pi_{\infty}^{S}(m \mid \theta) \sum_{a} \pi_{\infty}^{R}(a \mid m)(a - \theta)^2$$

# Metrics (contd)

**Informativness of the sender's policy**

Normalized mutual information between the distribution of messages, $\sum_\theta \pi^S_\infty(m \mid \theta)p(\theta)$, and the distribution of the states, $p(\theta)$

$$I(\pi^S) = \frac{\displaystyle\sum_\theta \sum_m \pi^S_\infty(m \mid \theta)p(\theta) \log\left(\frac{\pi^S_\infty(m \mid \theta)}{\sum_\theta \pi^S_\infty(m \mid \theta)p(\theta)}\right)}{\displaystyle\sum_\theta p(\theta)\log\left(\frac{1}{p(\theta)}\right)}.$$

When $\pi^S_\infty$ is fully informative $I(\pi^S) = 1$.

When $\pi^S_\infty$ is completely uninformative $I(\pi^S) = 0$.

# Baseline Setting

**Game:**

$\Theta = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ and $A = \{0, 0.1, 0.2, \ldots, 0.8, 0.9, 1\}$

$u_S(\theta, a) = -(a - \theta - b)^2, \quad u_R(\theta, a) = -(a - \theta)^2$

$p(\theta) = 1/6$ for all $\theta \in \Theta$

**Reinforcement learning:**

$\alpha = 0.1$ and $\lambda = 5 \times 10^{-5}$

$Q_0^S(\theta, m) \sim \text{Uniform}\left(-\frac{7}{60}, 0\right)$ and $Q_0^R(m, a) \sim \text{Uniform}\left(-\frac{7}{60} - b^2, 0\right)$

# Simulation outcomes

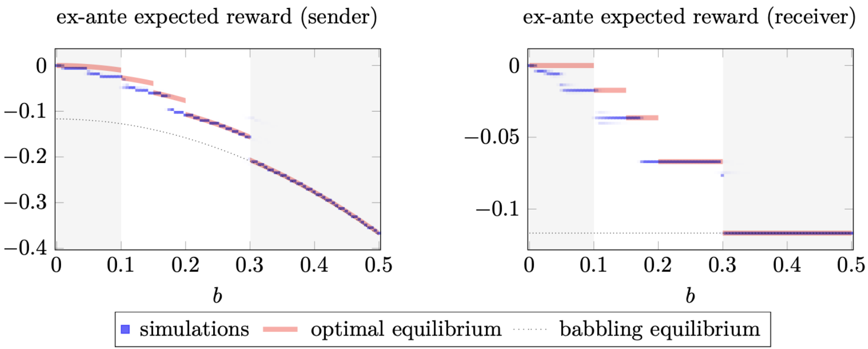



Figure: The distribution of values of 1000 simulations is shown in shades of blue. Grey shaded areas indicate where full information is optimal and when babbling is the unique equilibrium.

# Simulation outcomes (contd)

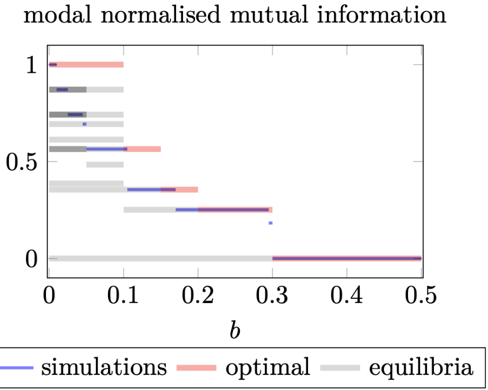



Figure: Normalised mutual information of the sender's modal policy across simulations converged to an equilibrium (maximum mass on suboptimal actions across states $< 0.01$ for both agents). The normalised mutual information of monotone partitional equilibria that exist for a given bias is shown in grey.

# Partitional equilibria

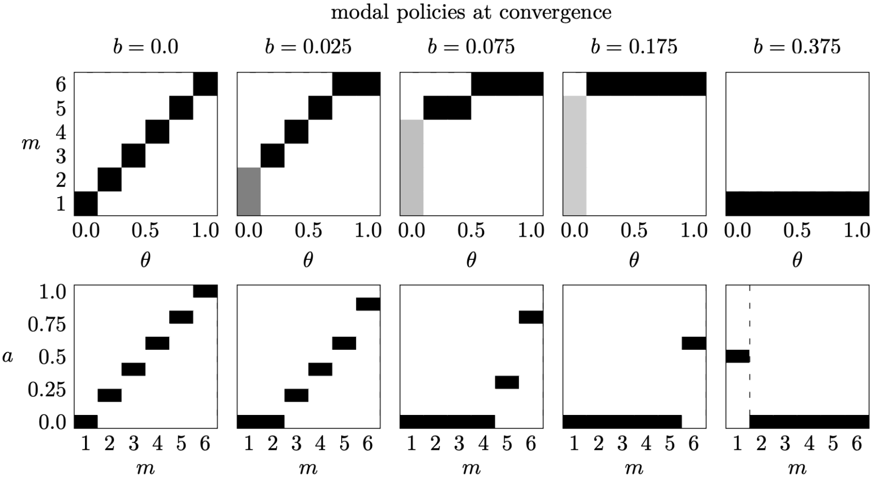


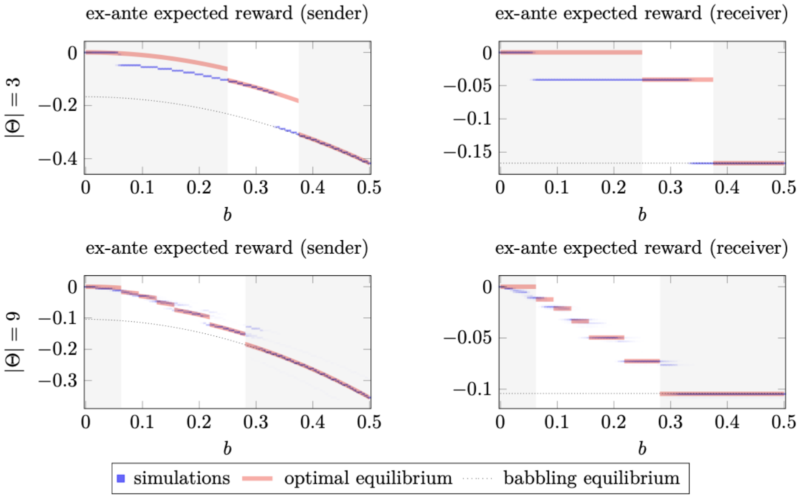

Figure: Heathmap of the modal policies of sender (top) and receiver (top) for different levels of bias over 1000 independent simulations.

# Robustness

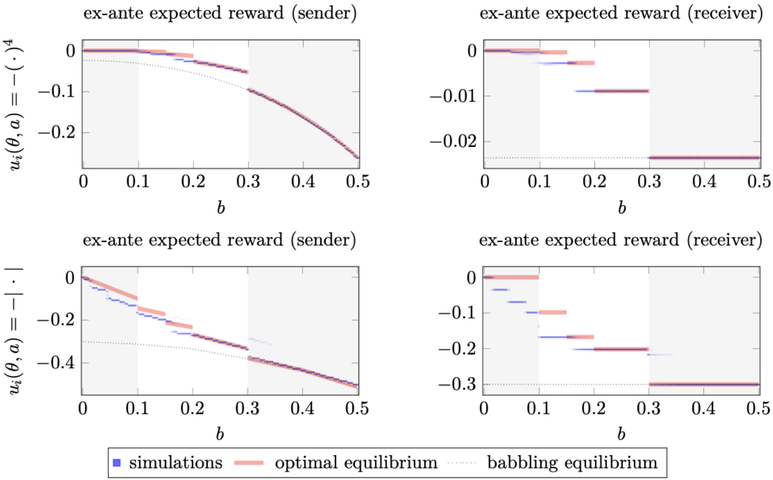
We replicate the analysis with different

- number of states of the world: $n \in \{3, 9\}$
- utilities: absolute loss, fourth power loss
- distribution of states: (linearly) increasing, (linearly) decreasing
- learning hyperparameters: $\alpha \in \{0.025, 0.050.1, 0.20.4\}$,
  $\lambda \in \{2, 1, 0.5, 0.25, 0.125\} \cdot 10^{-5}$

# Robustness: number of states



# Robustness: utilities



# Robustness: distribution of states