ENDOGENOUS COMMUNICATION IN REPEATED GAMES WITH LEARNING AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Communication among learning agents often emerges without explicit supervision. We study endogenous protocol formation in infinitely repeated stage games with a costless pre play channel. Each agent has a representation map that compresses private signals into messages subject to an information budget. Agents update strategies by no regret learning with stochastic approximation and choose representation maps by a myopic objective that trades off predictive value and encoding cost. We provide three main results. First, if the stage game admits a folk theorem set and the information budget exceeds a task specific threshold, there exists a stable communication equilibrium in which messages are sufficient statistics for continuation payoffs. Second, when the budget is below the threshold, any stable equilibrium must be pooling on a finite partition that we characterize with a minimax information bound. Third, we give polynomial sample complexity guarantees for convergence to an approximately efficient communicating equilibrium under mild regularity. The framework connects cheap talk, representation learning with information constraints, and multi agent no regret dynamics, and yields predictions for when emergent messages are interpretable, when they collapse, and how much data is needed for stable coordination.

1 INTRODUCTION

Agents trained in multi agent environments often invent discrete protocols that carry actionable information useful for coordination. When do such protocols become informative and stable. When do they collapse into uninformative chatter. We provide a simple theory that integrates incentives from repeated games with compression constraints from representation learning.

We model a costless pre play channel with agent specific encoders that map private signals to messages under an information budget. Strategic behavior unfolds in the continuation game. Learning plays two roles. Policies adapt through no regret updates. Encoders adapt to improve value prediction under a penalty for information usage. This produces a coupled dynamical system whose stable points we analyze.

Contributions

- A formal model of endogenous protocol formation in repeated games with information constrained representation maps.
- Existence of stable communication equilibria with value sufficient messages once budgets exceed a task specific threshold.
- A lower budget regime where any stable equilibrium must pool states; we bound the inevitable welfare gap with a minimax information argument.
- Finite sample convergence guarantees to near efficient communicating equilibria under standard assumptions for no regret dynamics.

2 RELATED WORK

Our work connects cheap talk in economics, learning in repeated games, information bottleneck style representation learning, and emergent communication in multi agent learning. We use classical

repeated game incentives and rate distortion style bounds to obtain sharp information requirements for informative communication and combine them with regret guarantees for convergence of adaptive play. Prior work on emergent communication focuses on empirical protocols and differentiable channels. We complement this with theory that explains when protocols carry value sufficient information and when they must pool.

3 MODEL

There are $N \geq 2$ agents indexed by i. Time is discrete $t = 1, 2, \ldots$. In each period a state $\theta_t \in \Theta$ is drawn i.i.d. from P. Agent i observes a private signal $s_{i,t} \sim P(\cdot \mid \theta_t)$. Before choosing actions, agents simultaneously send costless messages $m_{i,t} = \phi_i(s_{i,t})$ through a public channel with alphabet $\mathcal{M}_i$. After observing $m_t = (m_{1,t}, \ldots, m_{N,t})$ agents play a stage game $G(\theta_t)$ with action sets A_i and payoffs $u_i(a_t, \theta_t) \in [0, 1]$. The discounted value uses $\delta \in (0, 1)$.

Representation maps and budgets Each encoder ϕ_i belongs to a class Φ_i of measurable maps. Communication is limited by an information budget

$$I\left(S_i; M_i\right) \leq \kappa_i, \tag{1}$$

where the mutual information is computed under the stationary distribution induced by (P, ϕ) and the policy profile. We also use an equivalent cardinality constraint $\log |\mathcal{M}_i| \leq B_i$ when helpful. Budgets capture attention, bandwidth, or architectural limits.

Policies and learning A stationary policy for agent i is a map $\sigma_i : \mathcal{M} \to \Delta(A_i)$. Agents update σ_i by a no regret algorithm such as mirror descent on bandit feedback. Encoders update to optimize a myopic objective

$$J_i(\phi_i; \sigma) = \mathbb{E}\left\{V_i^{\sigma,\phi}(M_t)\right\} - \lambda_i I\left(S_i; M_i\right), \tag{2}$$

where $V_i^{\sigma,\phi}$ approximates the continuation value used by the learner and λ_i enforces the budget κ_i.

Stability

Definition 1 (Stable communicating equilibrium)**.** *A profile (ϕ, σ) is a stable communicating equilibrium if: (i) given (ϕ_{-i}, σ_{-i}) each σ_i is a best response in the repeated game supported by continuation strategies; (ii) each ϕ_i is a maximizer of* (2) *subject to* (1)*; (iii) the coupled learning dynamics converge in probability to a neighborhood of (ϕ, σ) from a set of initial conditions with positive measure.*

Value sufficient statistics

Definition 2 (Value sufficiency and threshold)**.** *A statistic $T_i(s_i)$ is value sufficient if $V_i^{\sigma,\phi}$ is conditionally independent of s_i given $T_i(s_i)$. The information threshold $\kappa_i^\star$ is the infimum of $I\left(S_i; T_i(S_i)\right)$ over value sufficient T_i.*

4 MAIN RESULTS

We state our main theorems. Proofs appear in Section 7 and the appendix.

Assumption 1 (Regularity)**.** *The stage game $G(\theta)$ satisfies standard folk theorem conditions. Signals are conditionally independent across agents given θ. Payoffs are Lipschitz in mixed actions. The learning rate sequence for policy updates satisfies $\sum_t \eta_t = \infty$ and $\sum_t \eta_t^2 < \infty$. Encoders are selected from classes with finite metric entropy under total variation.*

Theorem 1 (Sufficient statistic communication above threshold)**.** *If Assumption 1 holds and $\kappa_i \geq \kappa_i^\star$ for all i, then there exists a stable communicating equilibrium in which each encoder implements a value sufficient statistic and the joint message is fully revealing of the value relevant state. The equilibrium achieves the efficient payoff vector in the feasible folk theorem set.*

Theorem 2 (Mandatory pooling below threshold)**.** *If Assumption 1 holds and $\kappa_j < \kappa_j^\star$ for some agent j, then any stable communicating equilibrium must induce a finite partition of the private signal*

space of agent j with at most $\exp(\kappa_j)$ cells. The welfare loss relative to the efficient communicating benchmark is lower bounded by

$$\Delta \geq c \cdot \inf_{\Pi} \sup_{\theta,\theta'} \|v(\theta) - v(\theta')\|_1 \quad \textit{subject to } \Pi \textit{ respecting the information budget}$$

for a problem dependent curvature constant $c > 0$, where $v(\theta)$ denotes the continuation value vector.

Theorem 3 (Sample complexity of convergence)**.** *Under Assumption 1 with bounded stochastic gradients for value learners and uniform ergodicity of the message augmented Markov chain, mirror descent style no regret updates with step size $\eta_t = \Theta(t^{-1/2})$ reach an ε approximate stable communicating equilibrium in $\tilde{\mathcal{O}}(\varepsilon^{-2})$ samples with probability at least $1 - \delta$.*

Theorems 1 and 2 give a sharp qualitative picture. Above the threshold, messages can be interpreted as value sufficient statistics and informative communication is stable. Below the threshold, messages must pool states on a coarse partition and a welfare gap is unavoidable. Theorem 3 shows that standard no regret dynamics are sufficient to reach a near stable point with polynomial data.

5 Algorithms and diagnostics

We use an alternating scheme that interleaves policy and encoder updates. The encoder step treats the continuation value as a prediction target with an information penalty. The policy step uses no regret updates on the stage game conditional on messages.

Algorithm 1 Alternating Learning with Information Constrained Encoders

1: Initialize policies $\sigma_i^{(0)}$ and encoders $\phi_i^{(0)}$
2: **for** $t = 1, 2, \ldots$ **do**
3: Observe $m_t = \phi^{(t-1)}(s_t)$ and play $a_t \sim \sigma^{(t-1)}(\cdot \mid m_t)$
4: Receive payoff u_t
5: Policy update: for each i, perform a no regret step $\sigma_i^{(t)} \leftarrow \mathrm{MirrorDescent}\left(\sigma_i^{(t-1)}, g_{i,t}\right)$
6: Encoder update: for each i, update $\phi_i^{(t)} \in \arg\max_{\phi_i \in \Phi_i} \mathbb{E}\left\{\hat{V}_i^{(t)}(M)\right\} - \lambda_i I\left(S_i; M_i\right)$
7: **end for**

Diagnostics for practice follow directly. If performance jumps once the empirical estimate of $I\left(S_i; M_i\right)$ exceeds a threshold, the learned mapping is likely value sufficient. If increasing the alphabet does not improve value, the regime is pooling constrained.

6 Toy example

Two agents face a coordination game with private binary state $\theta \in \{H, L\}$. Payoffs are 1 for coordinated actions that match the state and 0 otherwise. Signals satisfy $\Pr[s_i = \theta \mid \theta] = 1 - \epsilon$ with $\epsilon \in (0, 1/2)$. Each agent has an encoder with budget κ.

Threshold The value sufficient statistic needs to disambiguate the state at the level required to achieve the efficient continuation payoff with grim trigger support. The mutual information needed is at least the Bayes risk reduction between the prior and posterior over θ. This yields a threshold $\kappa^\star = H(\Theta) - H(\Theta \mid T)$ where T is sufficient for the continuation value. For the binary case this equals the reduction from the posterior error probability under the optimal partition. When $\kappa \geq \kappa^\star$ the joint message makes the efficient outcome sustainable. When $\kappa < \kappa^\star$ any stable equilibrium pools some posteriors which creates a positive coordination error with a bound that matches Theorem 2.

7 Proofs

We summarize the main ideas. Full derivations are self contained.

7.1 Preliminaries

We use standard tools from repeated games, information theory, and online convex optimization. The folk theorem provides continuation payoffs that deter deviations with discount factor δ close to one. The data processing inequality and rate distortion bounds translate encoder budgets into partition limits. Regret bounds for mirror descent give convergence of adaptive play to coarse correlated equilibria which coincide with best responses in the continuation game when punishments are available.

7.2 Proof of Theorem 1

Let T_i be a value sufficient statistic with $I(S_i; T_i(S_i)) \leq \kappa_i$. Set $\phi_i = T_i$. The joint message m is then sufficient for the continuation values. Construct strategies that achieve the efficient payoff vector within the feasible folk theorem set using standard trigger strategies. Under value sufficiency, deviations that misreport messages do not improve continuation value because beliefs and punishments are computed from m. The encoder objective in (2) is maximized at value sufficient maps since any further compression that loses value would reduce $\mathbb{E}[V_i]$ more than it saves in the penalty when λ_i is set to enforce the budget. The coupled dynamics converge to a neighborhood due to standard stochastic approximation results since the best response correspondence is upper hemicontinuous and the encoder step is a contraction near the maximizer.

7.3 Proof of Theorem 2

Suppose $\kappa_j < \kappa_j^\star$. For any encoder ϕ_j the image of S_j has effective cardinality at most $\exp(\kappa_j)$ by standard rate distortion arguments. Hence ϕ_j induces a partition Π with at most $\exp(\kappa_j)$ cells. Two states that fall into the same cell are value indistinguishable under the induced beliefs. The continuation value cannot condition on distinctions that the message does not carry. A minimax argument yields a lower bound on the welfare loss that depends on the maximum value difference between states that are pooled by any partition respecting the budget. This proves the bound with constant c that depends on payoff curvature through a Lipschitz to value gap conversion.

7.4 Proof of Theorem 3

Mirror descent with unbiased gradient estimates achieves regret $R_T = \mathcal{O}(\sqrt{T})$. Under uniform ergodicity of the message augmented process, averaging arguments deliver concentration around the stable set defined by Definition 1. Standard reductions transform regret bounds into convergence rates for approachability of the equilibrium set. Choosing $\eta_t = \Theta(t^{-1/2})$ yields the claimed $\tilde{\mathcal{O}}(\varepsilon^{-2})$ sample complexity.

8 Limitations and scope

The analysis assumes i.i.d. states, synchronous and costless channels, and stationary limits. Extending to adversarial or Markovian states, delays, and private communication costs is an interesting direction. Our learning results use convexity and Lipschitz assumptions that hold for common no regret updates but may not hold for all deep learning implementations.

Ethics statement

This paper develops a theory of communication among learning agents and does not use human subjects or sensitive data. We discuss potential misuse in safety critical settings and recommend clear reporting of limitations. No new datasets are introduced.

Reproducibility statement

We provide complete proof sketches in the main text and full proofs are included in this self contained document. The toy example has all quantities defined in closed form. No external code is required.

REFERENCES

[1] V. P. Crawford and J. Sobel. Strategic information transmission. *Econometrica*, 50(6):1431–1451, 1982.

[2] D. Fudenberg and E. Maskin. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3):533–554, 1986.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

[4] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. *Allerton Conference*, 1999.

[5] F. Orabona. *A Modern Introduction to Online Learning*. Now Publishers, 2019.

[6] N. Cesa Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[7] J. R. Marden and J. S. Shamma. Game theory and distributed control. *Handbook of Game Theory*, 2012.

[8] I. Mordatch and P. Abbeel. Emergence of grounded compositional language in multi agent populations. *AAAI*, 2018.

[9] J. N. Foerster et al. Learning to communicate with deep multi agent reinforcement learning. *NeurIPS*, 2016.

[10] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

A ADDITIONAL PROOFS

A.1 DETAILS FOR THEOREM 1

We formalize value sufficiency by requiring that for almost all messages m the posterior over states induced by m is a sufficient statistic for continuation values. Let $\mathcal{V}$ denote the set of continuation value functions induced by feasible strategies. Existence of a measurable selector for the encoder that attains the infimum in the threshold follows from standard compactness under finite entropy classes. The rest follows by construction of grim trigger paths that achieve the efficient value and by verifying one step deviation conditions.

A.2 DETAILS FOR THEOREM 2

Let Π be any partition of the signal space that obeys the budget. The induced posterior set is a finite subset of the probability simplex whose size is bounded by $\exp(\kappa_j)$. The best possible mapping under the budget selects the partition that minimizes the worst case value distortion. A data processing argument shows that any policy profile that conditions only on Π suffers at least the claimed gap when payoffs are Lipschitz in beliefs. The constant c captures the modulus of continuity between belief differences and value differences.

A.3 DETAILS FOR THEOREM 3

We view the coupled learning process as a two time scale stochastic approximation. The policy step runs on the fast time scale with step size η_t. The encoder step runs on a slower schedule so that the policy process tracks the best response dynamics induced by a near stationary encoder. Standard results yield convergence to an invariant set whose distance to the stable communicating equilibrium is $\mathcal{O}(\varepsilon)$ when the regrets scale as $\mathcal{O}(\sqrt{T})$.

B LLM USAGE DISCLOSURE

A general purpose language model assisted with editing and organization. All problem statements, assumptions, and proofs were authored and verified by the authors. The model did not contribute original scientific ideas and is not an author.