Effects of Supervision, Population Size, and Self-Play on Multi-Agent Reinforcement Learning to Communicate

Marina Dubova¹ and Arseny Moskvichev²

¹Cognitive Science Program, Indiana University, Bloomington, IN 47408
²Department of Cognitive Sciences, University of California, Irvine, CA 92617
mdubova@iu.edu

Abstract

Learning to communicate in adaptive multi-agent populations introduces instability challenges at the individual and population levels. To develop an effective communication system, a population must converge on a shared and sufficiently stable vocabulary. We explore the factors that affect the symmetry and effectiveness of the communication protocols developed by deep reinforcement learning agents playing a coordination game. We looked at the effects of bottom-driven supervision, agent population size, and self-play ("inner speech") on the properties of the developed communication systems. To analyse the resulting communication protocols and derive appropriate conclusions, we developed a set of information-theoretic metrics, which has been a major underdevelopment in the field. We found that all the manipulated factors greatly affect the decentralized learning outcomes of the adaptive agents. The populations with more than 2 agents or with a self-play learning mode converge on more shared and symmetric communication protocols than the 2-agent (no self-play) groups. Bottom-driven supervising feedback, in turn, augments the learning results of all groups, helping the agents learning in bigger populations or with self-play to coordinate and converge on maximally homogeneous and symmetric communication systems. We discuss the implications of our results for future work on modeling language evolution with multi-agent reinforcement learning.

Introduction

Human shared communication systems (languages) evolved through complex interaction of cultural and individual adaptation processes. There are many factors that determine the trajectory of language evolution, such as developmental, morphological, and cognitive constraints of the communicating agents; the dynamics of their learning processes; and population-level factors, such as geographical constraints, social structure, and many others (Steels, 2000; Christiansen and Kirby, 2003; Oudeyer, 2006; The Five Graces Group et al., 2009; Hashimoto, 2006).

Even though the mapping of words to their meanings is arbitrary, individuals in a society must converge on some fixed set of shared mappings to have an effective communication system. It is not surprising that even in animals, a wide array of strategies are used to ensure consistency of such mappings within agent communities. For example, finch birds change their call adaptation strategy depending on their status in the flock hierarchy (i.e. more dominant birds adapt their calls to a lesser extent, essentially anchoring the group communication system) (Mundinger, 1970). In human communities, the role of communication is, arguably, more diverse, and a broad range of tools is employed, helping individuals learn a consistent and shared mapping from signals to meaning. The existence of various cultural transmission mechanisms, asymmetric adaptation in parent-child dyads, and more generally, the education system as a whole, can all be seen as examples of tools helping us to ensure the consistency of our communication system.

In our work, we study communication learning in a simplified setting, and investigate the factors that drive symmetry, homogeneity, and consistency in the communication patterns of artificial agent populations.

Related work

Existing language evolution models include genetic algorithms (Nowak et al., 2002; Nowak and Krakauer, 1999; Nowak et al., 1999; Reali and Griffiths, 2010), iterative learning (the model of cultural evolution) (Kirby et al., 2014; Griffiths and Kalish, 2007), multi-agent models (Gong, 2009, 2011; Hashimoto and Ikegami, 1996), robotic simulations (Steels, 2015; Spranger and Steels, 2012; Oudeyer, 2005), as well as hybrid models (i.e. combining genetic algorithms and neural networks) (Batali, 2019; Cangelosi and Parisi, 1998), and others (Kirby, 2002). While language change is induced by both individual-level and population-level mechanisms, these models often focus on formalizing the adaptation and transmission processes on only one of these levels. For instance, genetic, iterative learning, and most multi-agent models formalize the transmission on the population-level, sacrificing the details of individual learning processes. Robotic simulations, in turn, focus more closely on individual and dyadic learning mechanisms and dynamics, and are often difficult to scale to bigger populations.

In this context, neural network models provide a good compromise. Their expressiveness allows us to model rich
individual learning dynamics. At the same time, with enough computational resources, such models can be scaled to model moderate-size population dynamics, as opposed to, for example, robotic simulations, which necessarily run in real time. It is not surprising, therefore, that rise of deep learning research prompted a new wave of language evolution studies. A number of works have demonstrated the promise of this approach, showing that the deep learning agents can learn communication protocols in various game scenarios (such as Lazaridou et al. (2016); Evtimova et al. (2018); Lowe et al. (2017); Foerster et al. (2016); Sukhbaatar et al. (2016); Mordatch and Abbeel (2018), and many others).

**Instability of Multi-Agent Reinforcement Learning**

The work on training deep neural network agents to communicate demonstrates the main challenge of this approach: the instability of independent multi-agent reinforcement learning. In this case, the individual agents adapt simultaneously and determine the reward and training data for the other agents, making the optimization problem for each of them non-stationary and non-Markov (Bernstein et al., 2002; Laurent et al., 2011; Matignon et al., 2012). Such a population is not guaranteed to converge on shared and stable strategies, such as a consistent homogeneous communication system.

To overcome this issue, the aforementioned works have restricted the learning situations to the fully-supervised or semi-supervised interactions of 2 agents with fixed or flexible communication roles (speaker and listener) (Lazaridou et al., 2016; Havrylov and Titov, 2017; Chaabouni et al., 2019; Lazaridou et al., 2018; Evtimova et al., 2018; Graesser et al., 2020); centralized the optimization problem by fitting a shared controller for all the agents (Sukhbaatar et al., 2016); adopted inter-agent weight sharing (Foerster et al., 2016); used centralized critic in the actor-critic algorithm (Lowe et al., 2017), and allowed shared collective memory or access of agents to each other agent’s exact weights (Pesce and Montana, 2020; Foerster et al., 2018). Some of the latter works report a comparison of the chosen centralization method with the independent learners’ scenario, demonstrating that the alternative model performs poorly and fails to converge. Moreover, the multi-agent learning instability dramatically increases with the population size, often restricting the "decentralized" (independent) learning simulations to 2-4 agents.

The challenges with decentralized optimization of multiple deep learning agents raise the question of how humans, flexible and independent in their learning, have successfully converged on shared communication systems.

**Stabilizing naive multi-agent communication learning with realistic interventions**

We hypothesize that human learners have access to various sources for supervising feedback on their communication, which amplifies and disambiguates the reinforcement signals. For example, humans can infer potentially rewarding actions or symbol meanings even after an unsuccessful interaction, by observing the action of their partner. Therefore, in our simulations, we test how the generic bottom-driven supervising feedback helps to overcome the "sparsity of rewards" and instability problems in multi-agent reinforcement learning case.

Secondly, a small number of adaptive agents playing a communication game encounter no pressure preventing them from partitioning the signaling space and developing an asymmetric communication system (Evtimova et al., 2018). It is particularly true for the 2-agent setup when the agents do not experience any randomness in who they are going to play with. To address this issue, we are looking at two potential factors enforcing the development of symmetric communication protocols: population size and self-play ("inner speech"). We hypothesize that human languages have originally emerged and developed in groups of more than 2 individuals. If such a case is applied to the modeling setup, each agent tries to communicate with different partners, and, potentially, with new ones. This makes specialization of agents’ vocabularies non-effective, especially when the partner's identity could not be reliably inferred. The self-play ("inner speech") manipulation also constrains the learning problem towards developing symmetric meanings. In this case, the agents communicate with the other agents as well as with themselves and try to coordinate their responses in these two cases. Therefore, the self-play setting forces them to coordinate their own responses to communicative signals with the other agent’s responses to them.

A couple of other works have tested some realistic factors stabilizing independent communication learning. In particular, Eccles et al. (2019) explored the effects of the biases for positive signaling and positive listening and Bogin et al. (2019) looked at the effect of the bias for speaker consistency on the overall scores of the communication game learning.

We view our work as a continuation of this exciting line of research, and extensively test whether the population size, self-play, and supervision improve the results of multi-agent communication learning. To analyze the results of our simulations, we develop a set of metrics to isolate the properties of communication that are affected by these manipulations.

**Problem Formulation**

We consider a number of multi-agent communication settings in which two or more agents have to coordinate their actions in order to receive reward. In our problem setting, we aimed to study a minimal example representing the challenges of multi-agent communication learning.

In this section, we formalize the notion of a communication game, and formally describe the settings that we considered.
One-way communication

Before we describe the multi-way multi-agent settings that constitute the main body of our work, let’s first consider the simpler situation of one-way two-agent communication.

In this simplified scenario, there are only two agents: the speaker and the listener. The speaker chooses an action and a message to communicate, the listener receives the message and chooses an action. If the actions match, both agents receive a reward.

The setting for every agent is close to that of one-step Markov Decision Process (MDP), which is a four-tuple: \((S, A, P, R)\). We will use a standard notation, denoting the set of states as \(S\), the set of actions as \(A\), the transition probabilities as \(P\), and the reward function as \(R\).

These processes are different for the speakers and the listeners. The action space is \(A_1 = \{1...K\} \subset N\) for the listeners. In other words, the listeners pick a unique action among \(K\) available options. For the speakers, \(A_2 = A_1 \times A_1\), where \(A_1 = \{1...T\} \subset N\). That is, in addition to picking an action from the same state as the listeners, speakers also select a message to communicate among \(T^2\) available options. If the chosen actions coincide, the agents receive a fixed reward for successful cooperation.

Agents also receive an additional penalty if the action they chose was overused in the past. The penalty is asymmetric: agents receive no bonus for using rare actions. The penalty is determined by \(\beta(1/K - \hat{p}_k)\), where \(K\) is the number of possible actions and \(\hat{p}_k\) is the proportion of times this action was selected in the last \(H\) episodes, where \(H\) is a hyperparameter\(^3\). This reward is utilized to make sure that agents learn to cooperate using a communicative channel, instead of relying on behavioral predictability of the other agent.

The state spaces are different for speakers and listeners as well. For the speaker, the state space includes a random input (in order to allow for non-deterministic policies), as well as the average frequencies of different actions. The listener receives only the speaker’s message and the frequency history of its actions.

Despite the interface similarity between MDPs and the setting describe above, the problems that the agents are facing do not fall under the MDP framework. The problem is that the reward depends on the policies (and, hence, internal states) of other agents, which are not observed. This makes the environment only partially observable. Moreover, the actions that the agents pick affect what the other agents are going to learn, which, together with the fact that the states of other agents are not observed, makes the environment non-Markov.

\(^3\) was set to 4 in all simulations

Two-way communication

One-way communication setting described above introduces unrealistic restrictions: one agent is always speaking, while another is always listening. This presents a sharp contrast with most naturalistic interactions.

We, therefore, use an extension of the problem posed above, with only one key difference: on each game episode, the speaker and listener roles are assigned randomly.

To account for that difference, the state space is modified accordingly. More specifically, \(S = S_r \times S_l \times S_h\), where \(S_r = \{0,1\}\) is a binary indicator of the current role. \(S_l = R^K\) is an input vector, that either contains random noise (for speakers) or the speaker utterance (for listeners). Lastly, \(S_m = [0,1]^K \subset R^K\), represents the memory of the agent, specifically, the action choice proportions over the last \(H\) episodes.

The action state is now the same for both agents and coincides with \(A_2\) in the one-way communication setting. The communication part of the action, however, is not used if the agent is assigned the listener role.

Multi-agent two-way communication

In most experiments we consider a direct generalization of the two-way communication setting. Instead of two agents, we have a pool of agents \(P\). In every episode, two agents are randomly selected, and then the episode proceeds as described in the two-way-communication section.

Model

Our choice of the model and its training procedure is also motivated by the considerations of simplicity and representativeness to the main multi-agent communication learning challenges.

By focusing on a simple problem, we can reasonably hope that any difficulties in agent’s learning are caused by the dynamics of multi-agent interaction, and not by limitations of the underlying RL-algorithm or network architecture. In general, the policies that must be learned to solve our task are very simple, the only challenge is that all agents need to learn compatible policies (e.g. Q-functions that would allow them to behave in a coordinated manner).

Therefore, we opted for using a vanilla Deep Q-learning algorithm as the main workhorse for training the agent networks. All agents are represented by feedforward neural networks with two hidden layers containing 25 and 15 neurons respectively. We used the ReLU (Rectified Linear Unit) activation function.

We used additional bottom-driven supervised updates to train the agents. Supervised feedback for a given agent corresponds to saving its partner’s action following a specific signal as a correct response for that symbol. In this way, we don’t impose any pre-defined signal-to-action mappings, since the supervision signals are generated by the agents themselves. The supervised feedback is implemented...
by changing a certain proportion of negative ("miscoordinated") experiences to the "supervised" ones (i.e. the agent’s action in a miscoordinated memory episode is changed to a corresponding action by the other agent and the reward is changed accordingly). We decided to control for supervision in our experiments, since partial supervising feedback in different forms is often available in realistic language learning situations. By amplifying the reinforcement signal with more “positive” examples, supervising feedback may help to overcome the “sparsity of rewards” and instability problems in multi-agent reinforcement learning.

**Metrics**

Following Lowe et al. (2019), we used a number of specific metrics in addition to the reward scores to uncover the properties of learned communication systems.

Most of the metrics need a number of probability distributions to be computed. We estimate all such probability distributions empirically.

1. **Speaking Consistency and Listening Consistency.**

   First of all, it is important to establish whether the actions that the agents perform are in any way related to the signals that they send or receive. Lowe et al. (2019) proposed to use the normalized mutual information between the distribution over messages and actions that an agent defines. Formally, the metric is defined as

   \[
   \text{S/LC} = \sum_{a \in A_i} \sum_{m \in A_c} p_{a,m}(a, m) \log \frac{p_{a,m}(a, m)}{p_a(a)p_m(m)} / Z
   \]

   Here, \( Z \) is the average entropy of the two marginal distributions: \( Z = \frac{H(p_a) + H(p_m)}{2} \). As in the previous sections, \( A_i \) and \( A_c \) denote the set of available actions and the set of available messages respectively.

   Note that this metric is computed twice for every agent (separately for the speaker and the listener roles). To obtain the final metric characterising the behaviour of an agent population, we average individual scores over all agents and roles.

2. **Communication Asymmetry:**

   a. **Between-agent signal-action mapping divergence (homogeneity).** Even if the messages and actions are coordinated in each agent, it may still be the case that different agents learn different “languages”.

   To quantify the extent of such communication difference, we propose to use the average Jensen-Shannon pairwise divergence between distributions of agents’ actions following a specific signal (averaged over all signals and pairs of agents).

   For a pair of agents 1 and 2, the metric is defined as follows

   \[
   \sum_{m \in A_c} JSD(p_{a_1|m}, p_{a_2|m}) / |A_c| \]  

   Where \( p_{a_i|m} \) are action distributions of agent 1 conditional on the message (received or sent) being equal to \( m \). We report this metric separately for all pairs of speakers and for all pairs of listeners.

   b. **Talking divergence.**

   To measure the variation in individual agents’ messaging preferences, we report the average pairwise Jensen-Shannon divergence of marginal signaling distributions of different agents.

3. **Behavioral Predictability.** Lastly, to assess whether agents diversify their actions and messages, we report the Jensen-Shannon divergences between marginal distributions of agent’s actions and messages and the uniform distribution.

**Experiments**

For agents’ optimization, we used the Adam method with learning rate 0.0001 and momentum (\( \beta \)) 0.3. The memory and replay sizes were set to 10; the exploration decay and the minimal exploration rate were set to 0.999 and 0.1 respectively. For each simulation, the agent population played 60000 coordination games total, and the agents updated their weights after each game they played. Each agent’s weights were updated independently from the other agents.

**Preliminary experiments: model capacity assessment and calibration**

Before evaluating the effects of the main variables we wanted to study (self-play, supervision rate and population size), we performed a number of calibration experiments, in order to make sure that the proposed model architecture and learning algorithms are adequate for the task at hand. For
these preliminary experiments, we used a one-way communication problem setting where one of the agents (either the speaker or the listener) is fixed to a hard-coded, fixed mapping (language).

**Experiment 1: self-play and supervision rate**

We simulated 2-agent coordination game learning with and without fictitious self-play for supervision rate varying from 0 to 0.9 (20 simulations per condition). In the self-play condition, the agents were randomly assigned to either play with another agent or with their own decisions in the previous game. The agents played 60000 games, which includes the self-play games.

We excluded the self-play games’ data from the subsequent analysis. We computed the average agents’ scores for the last 2000 games, and the average between- and within-agent signal-mapping divergence, “talking” divergence, speaker and listener consistencies, as well as the average divergence of agents’ signaling and acting distributions from the uniform rates for the last 4000 games that the agents played with each other (Figure 1).

**Experiment 2: population size and supervision rate**

We simulated the coordination game learning of 2 to 6 agents with supervision rate varying from 0 to 0.9 (20 samples per condition). For each condition, the agents played 60000 games total, indicating that the number of training interactions and weight updates per agent decreased with the increasing population size (\(\frac{60000}{n}\)).

We analyzed the data from this experiment in the same way as for the experiment 1. All the metric scores were averaged for each agent/pair of agents and for each signal (when necessary).

**Results**

**Preliminary experiments**

Preliminary experiments demonstrated that the agents quickly reach ceiling performance when one of them is exhibiting arbitrary deterministic performance. The type of function that the model needs to represent in order to solve the problem in multi-agent independent learning setting is the same as in our preliminary experiments. Therefore, we can be confident that any difficulties that agents may encounter in other experiments can be seen as a consequence...
of the multi-agent interactions, and not the model’s representational power.

**Supervision rate**

Supervising feedback improved the overall performance of all the groups in our experiments (Figure 1a and 2a). Supervision rate was reversely related to the asymmetry scores: JS-divergence of between-agent signal-action mappings (Figure 1b and 2b), within-agent signal-action mappings, and talking distributions of different agents (Figure 1c and 2c). Supervision rate also negatively affected behavioral predictability exhibited by all the groups in our experiments (Figure 1e and 2e).

**Self-play**

The self-play condition made the task more challenging, which is indicated by the lower reward scores than in the control, no self-play, condition (Figure 1a). However, agents in the self-play group demonstrated much lower between- and within-agent signal-action mapping divergence (Figure 1b and 1c), as well as lower divergence of talking distributions (Figure 1c) than the control group. Agents in self-play condition had more consistent responses to signals in both listener and speaker roles than the agents learning with no self-play (Figure 1d). Agents in both self-play and the alternative conditions achieved low behavioral (action) predictability. In the self-play group, however, the signaling distribution was coordinated with the action distribution, whereas for the alternative group these distributions were...
highly disparate (Figure 1e).

**Population size**

Learning in bigger populations was more challenging, but it was greatly leveraged by the supervising feedback (Figure 2a). All communication asymmetry scores were much lower for populations of 3-6 agents than for the dyad learning condition (Figure 2b and 2c). The 3-5 agents’ groups exhibited higher speaker and listener consistencies than the 2 and 6 agents’ groups (Figure 2d), the listener consistency was lower than the speaker consistency in all the groups. We explain the decreased consistency scores of the 6 agents’ population as a result of smaller amount of training updates per agent (10000 in this case).

Agents learning in 2-agent condition demonstrated low action predictability, which was not aligned with the signaling distribution. The agents learning in bigger groups exhibited higher action predictability solutions, which decreased with the supervision rate. The signaling and acting predictability scores were coordinated in the 3-6 agents’ populations (Figure 2e).

**Discussion**

First of all, our simulations replicate the result that naive reinforcement learning feedback is not sufficient to ensure that independently adapting communicative agents converge on effective strategies, even in a simple coordination game with 4 actions.

Motivated by examples of naturalistic communication learning, we investigated whether such variables as population size, limited supervision, and self-play (inner speech), can foster the development of efficient communication protocols. To isolate such effects, we followed the general principles advocated for in Lowe et al. (2019), significantly expanding, however, the set of proposed metrics.

We found that independent learning in a group of two agents leads to the development of internally and externally asymmetric communication protocols. These between- and within-agent asymmetries are ameliorated by either introducing self-play or increasing the agent population. Adding supervising feedback positively affected agents’ performance across all metrics, but it was not as efficient in solving the asymmetry issues as introducing self-play or increasing the population size.

We believe that one of the main merits of our contribution is proposing a number of metrics that allowed us to isolate different components of communication protocols learned by agents. In line with Lowe et al. (2019), we found that using the average reward metric alone does not provide enough information about the success of communication learning (even in our simple case), and that it can even be misleading.

For example, the agents in no self-play condition achieved higher reward scores than the agents in the alternative, self-play, condition (Figure 1). However, further examination reveals that this result is mainly associated with a more uniform action distribution exhibited by the agents with no self-play (and hence, a smaller diversity penalty), and not by a development of a more symmetric and consistent communication system. While the non-uniform penalty is specific to our setup, most contemporary reinforcement learning studies include additional “shaping rewards” of similar nature to avoid trivial solutions. This further supports the idea of using a set of more specific metrics in addition to the average reward.

Our extended set of metrics also allowed us to achieve a more nuanced understanding of the population size effects. The connection between the population size and the success of communication learning was recently reported in Graesser et al. (2020), but the authors only examined the “self-play success” of the agents as a measure of communication symmetry. Our results provide a new insight on this improvement: perhaps surprisingly, increasing the population size mostly affects the within-agent asymmetry of the developed communicative patterns (similar to “self-play success” used in Graesser et al. (2020)), while the between-agent analysis demonstrates that the agents in a group still converge to dissimilar “meanings” (see Figure 2 b and c).

The relative simplicity of our experimental setup presents a sharp contrast with a number of recent studies. Our results demonstrate, however, that such a setting is sufficiently rich to capture some of the key challenging aspects of multi-agent reinforcement learning. We believe that it motivates a partial shift in research practices from necessarily aiming to expand the set of currently solvable problems to focusing on a more nuanced analysis of limitations and properties of already available solutions. In our view, such a shift may yield numerous insights for the studies of language evolution and we hope to inspire more works in this direction.

**Data Availability**

The code and data for this work are available online at the project’s github repository.

**Acknowledgements**

Authors thank Robert Goldstone, Denizhan Pak, Jack Avery, Mahi Luthra, Brian Dahlberg, all the attendees of the PCL lab meetings, and two anonymous reviewers for their valuable feedback on this project. This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute. https://kb.iu.edu/d/anwt#carbonate

**References**


