Neural Iterated Learning

Shangmin Guo

Department of Computer Science and Technology, University of Cambridge

April 3, 2020

Content

Motivation

How could we learn from machines?

Learn from them by interacting with them.

- Teach them our languages:
  - Tasks: NLU and NLG
  - Problem: languages co-evolve with environments.
- Help them communicate with interpretable languages:
  - Task: Emergent Communication Protocols
  - Problem: interpretability v.s. deep learning.

How could we learn from machines?

Learn from them by interacting with them.

- Teach them our languages:
  - Tasks: NLU and NLG
  - Problem: languages co-evolve with environments.
- Help them communicate with interpretable languages:
  - Task: Emergent Communication Protocols
  - Problem: interpretability v.s. deep learning.

Compositionality -> Interpretability

Compositionality

A Hierarchy of Compositionality

- **1st Order**: Combination
  - Combine symbols for orthogonal attributes.
  - “Non-terminals -> Terminals” in Context-Free Grammar.
  - Example: ’red circle’, ’green square’ for (colour, shape).
- **2nd Order**: Recursion
  - Recursively apply same rule.
  - “Non-terminals -> Non-terminals” in CFG, e.g. (VP->Vp NP VP)
  - Example: “I know you know I don’t know something you know.”

A Hierarchy of Compositionality

- **1st Order**: Combination
  - Combine symbols for orthogonal attributes.
  - "Non-terminals -> Terminals" in Context-Free Grammar.
  - Example: 'red circle', 'green square' for (colour, shape).
- **2nd Order**: Recursion
  - Recursively apply same rule.
  - "Non-terminals -> Non-terminals" in CFG, e.g. (VP->Vp NP VP)
  - Example: "I know you know I don't know something you know."

In this work, we only focus on the 1st Order Compositionality.

Different Types of Languages (1)

Definition

A language is a mapping function from object space $\mathcal{X}$ to the message space $\mathcal{M}$, i.e., $\mathcal{L}(\cdot) : \mathcal{X} \mapsto \mathcal{M}$.

- Meaning Space
  - 2 attributes: colour, shape
  - 2 value per attribute: (blue, red), (box, circle)
- Message Space
  - length: 2
  - vocabulary: (a, b)

Different Types of Languages (2)

Group	Compositional (8)	Holistic (16)	Other (232)
Language Examples	*blue box = aa* *red box = ba* *blue circle = ab* *red circle = bb*	*blue box = ba* *red box = aa* *blue circle = ab* *red circle = bb*	*blue box = aa* *red box = bb* *blue circle = aa* *red circle = bb*

Table: Different groups of language and corresponding examples.

- **Unambiguous language**: fully 1-to-1 mappings between meanings and messages.
  - **Compositional language**: exhibits systematic compositional structure when forming messages.
  - **Holistic language**: not *fully* exhibits systematic compositional structure.

Different Types of Languages (3)

Group	Compositional (8)	Holistic (16)	Other (232)
Language Examples	*blue box = aa* *red box = ba* *blue circle = ab* *red circle = bb*	*blue box = ba* *red box = aa* *blue circle = ab* *red circle = bb*	*blue box = aa* *red box = bb* *blue circle = aa* *red circle = bb*

Table: Different groups of language and corresponding examples.

- **Ambiguous language**: not fully 1-to-1 mappings between meanings and messages.
  - **Degenerate language**: all meanings correspond to the same message.
  - **Degenerate component**: n-to-1 mappings between meanings and messages.

Evaluate Compositionality

Topological Similarity

Intuition: similar (closer) meanings should have similar (closer) messages.

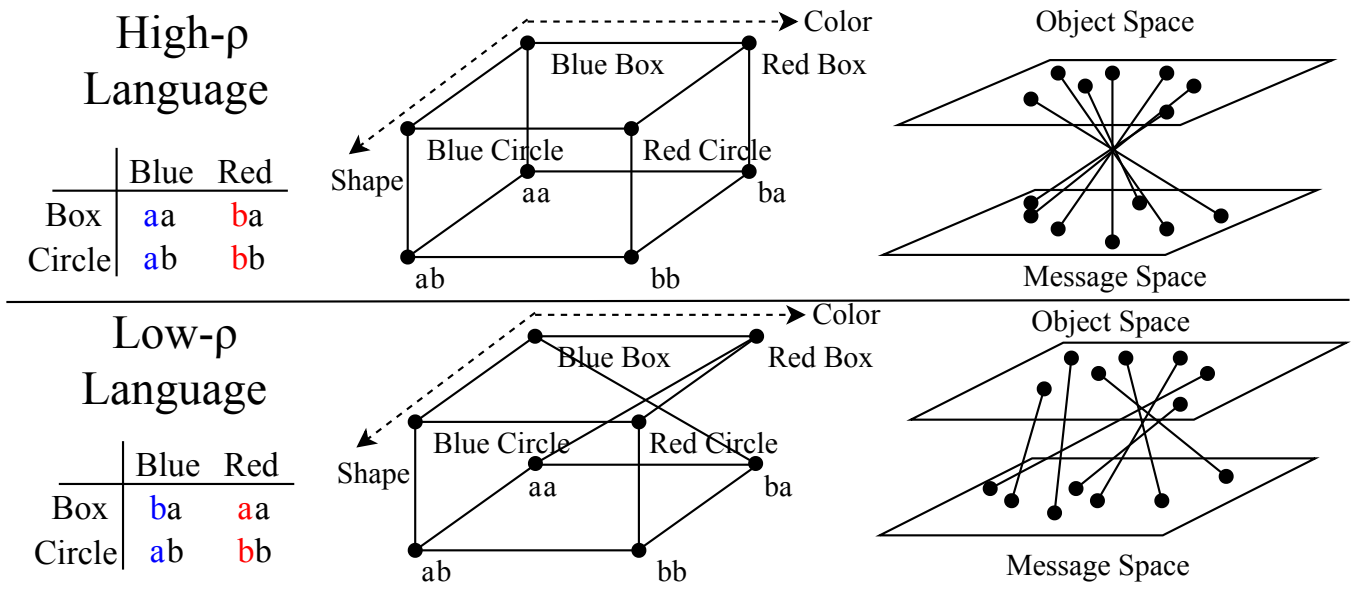


Figure: A simple representation of two languages corresponding to topological similarities of $\rho = 1$ (top) and $\rho = 0.5$ (bottom).

Number of Different Languages

N_a: # attributes, N_v: # values per attribute.
N_L: length of messages, $|V|$: vocabulary size.

$$\text{\# all possible languages} = \left(|V|^{N_L}\right)^{(N_v^{N_a})} \tag{1}$$

$$\text{\# unambiguous languages} = \frac{(|V|^{N_L})!}{(|V|^{N_L} - N_v^{N_a})!} \tag{2}$$

$$\text{\# compositional languages} = \frac{N_L!}{(N_L - N_a)!} \cdot \left(\frac{|V|!}{(|V| - N_v)!}\right)^{N_a} \tag{3}$$

$$\text{\# holistic languages} = \text{\# unambiguous} - \text{\# compositional} \tag{4}$$

Proportion of Compositional Languages

$N_a = 6, N_v = 10, N_L = 6, |V| = 10$

$$\#\text{ all possible languages} = \left(10^6\right)^{(10^6)} = 10^{6\times 10^6} \tag{5}$$

$$\#\text{ compositional languages} = \frac{6!}{(6-6)!} \cdot \left(\frac{10!}{(0)!}\right)^6 < 1.65 \times 10^{42} \tag{6}$$

$$\text{proportion of compositional} < \frac{1.65 \times 10^{42}}{10^{6\times 10^6}} < \frac{1}{10^{5999957}} \tag{7}$$

atoms in the known, observable universe: $10^{78} \sim 10^{82}$

Creating a compositional language is almost a mission impossible!

We human beings are extremely amazing!

Problem: How could agents do so?

Neural Iterated Learning

Methodology

1. Iterated Learning: how do human prefer compositional language.
2. Computationalise IL: symbolic referential game & probabilistic modelling.
3. Neural Iterated Learning
4. Performance of Neural Iterated Learning.
5. Robustness of Neural Iterated Learning.

Iterated Learning

Key idea

Language must be learned by new speakers at each generation, while also being used for communication.

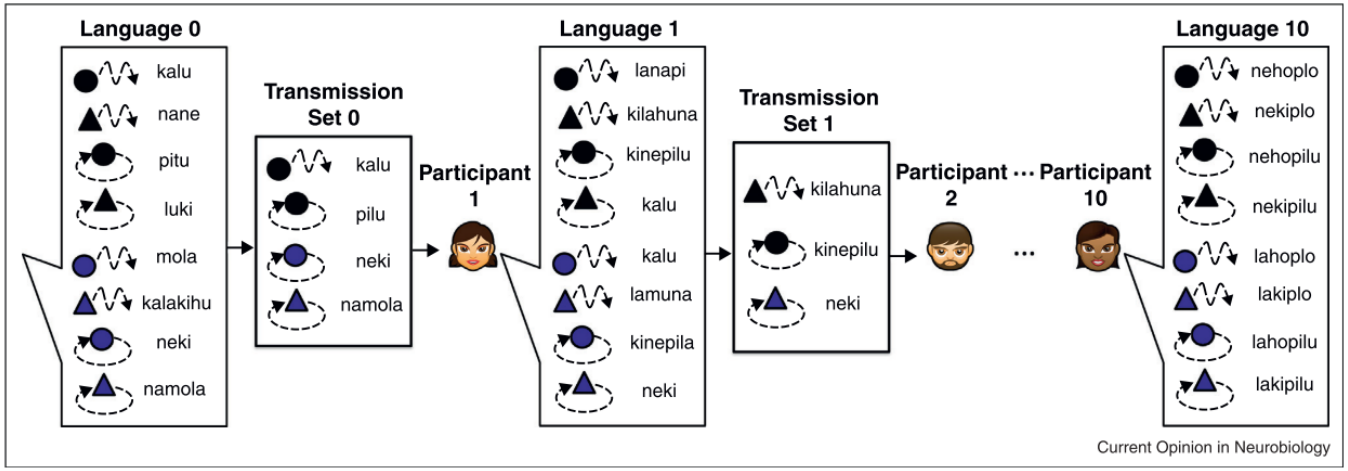


Figure: An illustration of the iterated artificial language learning method and indicative results. Phases are: i) learning; ii)interaction; iii)transmission.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. Current Opinion in Neurobiology, 28, 108–114.

Computationalise Iterated Learning

- Tasks for human -> Symbolic referential game for agents.
- Language in human brain -> Probabilistic distribution of languages in speakers.

Symbolic Referential Game

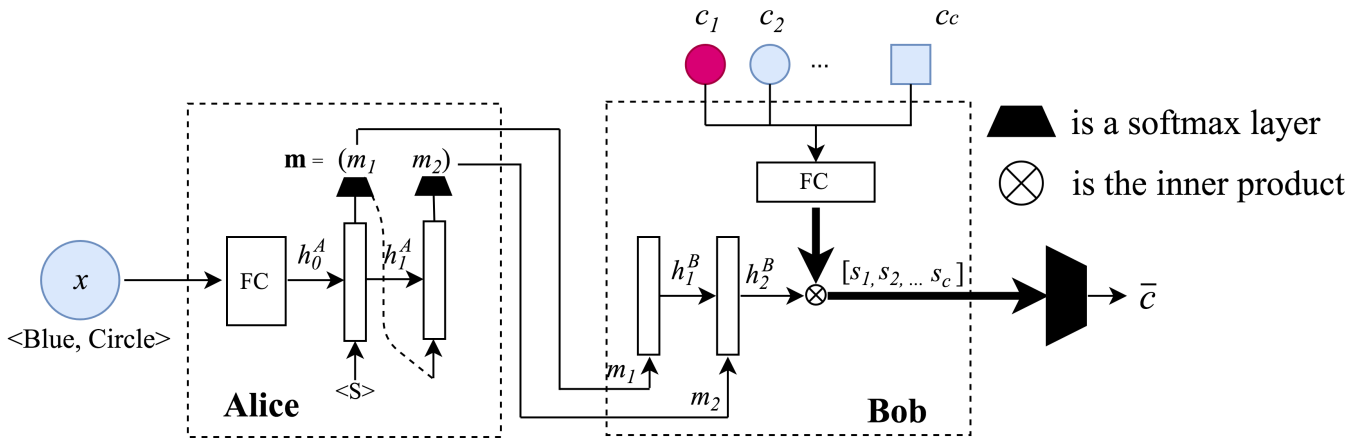


Figure: Referential communication game and architectures of the agents..

- Message generator and encoder: LSTM
- REINFORCE to update the policy to maximise the expected reward:
  - $\nabla_{\theta_A} \mathcal{J} = \mathbb{E}\left[R(\bar{c}, x) \nabla \log p_A(\mathbf{m}|x)\right] + \lambda_A \nabla H[p_A(\mathbf{m}|x)]$
  - $\nabla_{\theta_B} \mathcal{J} =$
    $\mathbb{E}\left[R(\bar{c}, x) \nabla \log p_B(\bar{c}|\mathbf{m}, c_1, ..., c_c)\right] + \lambda_B \nabla H[p_B(\bar{c}|\mathbf{m}, c_1, ..., c_c)]$
  - $R(\bar{c}, x) = \mathbb{1}(\bar{c}, x)$, θ_A and θ_B are parameters of Alice and Bob respectively.

Probabilistic Model of Emergent Languages (1)

Definition

A language is a mapping function from object space $\mathcal{X}$ to the message space $\mathcal{M}$, i.e., $\mathcal{L}(\cdot) : \mathcal{X} \mapsto \mathcal{M}$.

Assume that the messages are conditionally independent given an object x_n (where $n \in [1, 2, \ldots, N]$), then

$$
\begin{aligned}
P(\mathcal{L}) &= P(\mathbf{m_1}, \ldots, \mathbf{m_N} | x_1, \ldots, x_N) \\
&= \prod_{n=1}^{N} P(\mathbf{m_n} | x_1, \ldots, x_N) \\
&= \prod_{n=1}^{N} P(\mathbf{m_n} | x_n).
\end{aligned}
$$

Probabilistic Model of Emergent Languages (2)

Formal definition of the speaking agent (Alice) and listening agent (Bob):

- Alice: $\mathbf{m} = h(x), h : \mathcal{X} \mapsto \mathcal{M}$ (same as $\mathcal{L}$).
  - The probability of specific $\mathbf{m}$ given x for Alice:

$$P(\mathbf{m}|x) = P(m_1|x) \prod_{l=2}^{N_L} P(m_l|x, m_{l-1}, m_{l-2}, \dots)$$

- Bob: $s = f(\mathbf{m}, x), f : \mathcal{M} \times \mathcal{X} \mapsto \mathbb{R}^1$.

Neural Iterated Learning (1)

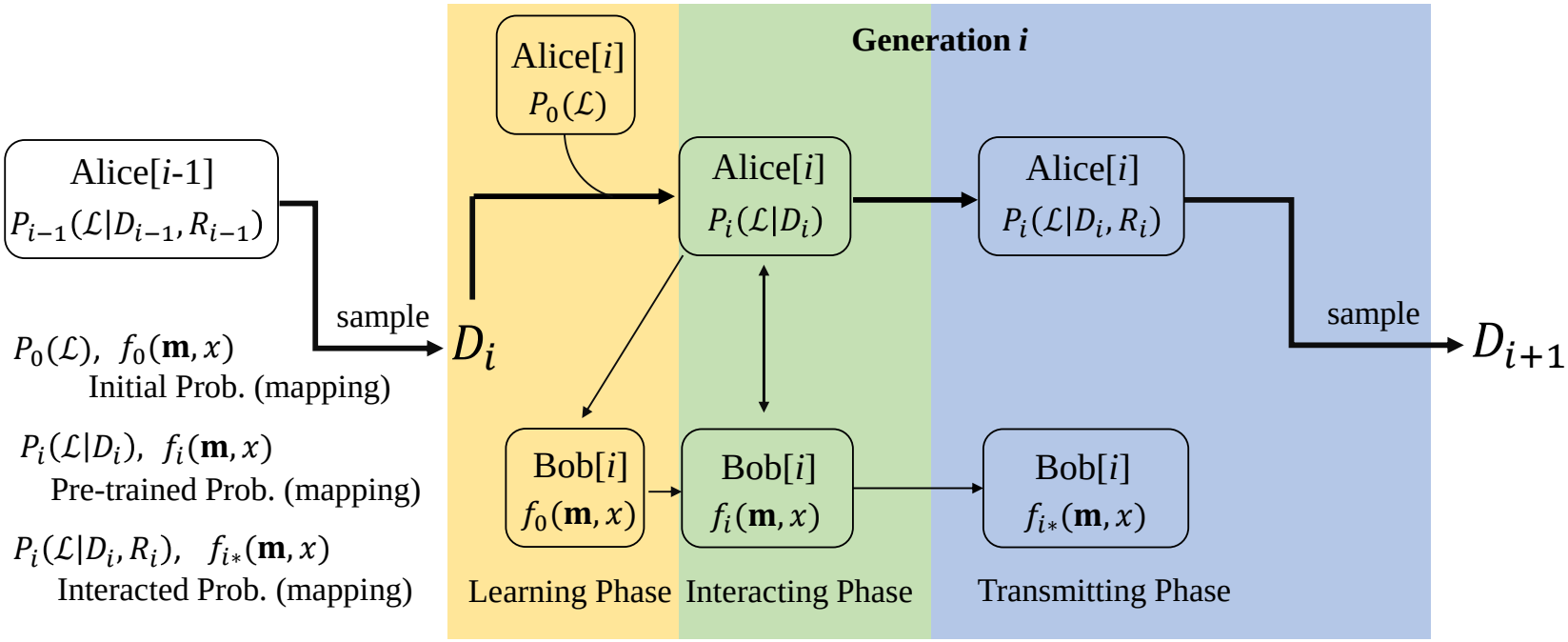


Figure: Probabilistic illustration of Neural Iterated Learning.

Neural Iterated Learning (2)

1. **Initialization:** randomly initialise Alice[0] and Bob[0] with uniform distribution. ($\sim$ Uniform prior.)
2. **Learning Phase:**
   1. pre-train Alice[i] to imitate D_i, and get $P_i(\mathcal{L}|D_i)$.
   2. pre-train Bob[i] with samples $\sim P_i(\mathcal{L}|D_i)$ and get $f_i(\mathbf{m}, x)$.
3. **Interacting Phase:** Alice[i] and Bob[i] interact with each other and update their parameters accordingly, which we argue has the same effect to
   1. Alice[i]: delete ineffective $\langle \mathbf{m}, x \rangle$ pairs in sampled data set $D_* \sim P_i(\mathcal{L}|D_i)$ -> $P_i(\mathcal{L}|D_i, R_i)$.
   2. Bob[i]: refine its parameters to form $f_{i*}(\mathbf{m}, x)$.
4. **transmitting Phase:** sample $D_{i+1} \sim P_i(\mathcal{L}|D_i, R_i)$
   1. randomly feeding x_n to Alice[i].
   2. sample a message $\mathbf{m}_n \sim P_i(\mathbf{m}|x_n, D_i, R_i)$.

Performace of NIL (1)

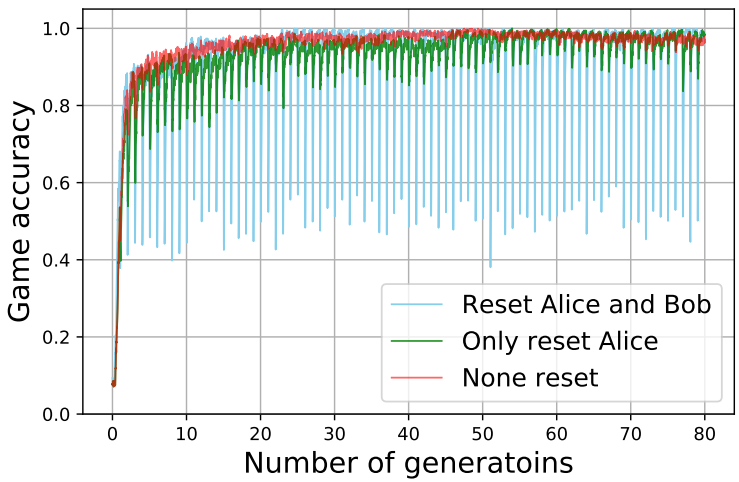
	blue	green	cyan	brown	red	black	yellow	white
box	aa	ea	ba	ga	da	ca	ha	fa
circle	ab	eb	bb	gb	db	cb	hb	fb
triangle	ae	**eb**	be	ge	de	ce	he	fe
square	af	ef	bf	gf	df	cf	hf	ff
star	ac	ec	bc	gc	dc	cc	**dh**	fc
diamond	ad	ed	bd	gd	dd	cd	hd	fd
pentagon	ag	eg	bg	gg	dg	cg	hg	fg
capsule	ah	eh	bh	gh	**hc**	ch	hh	fh

Table: Example of the converged language under NIL.

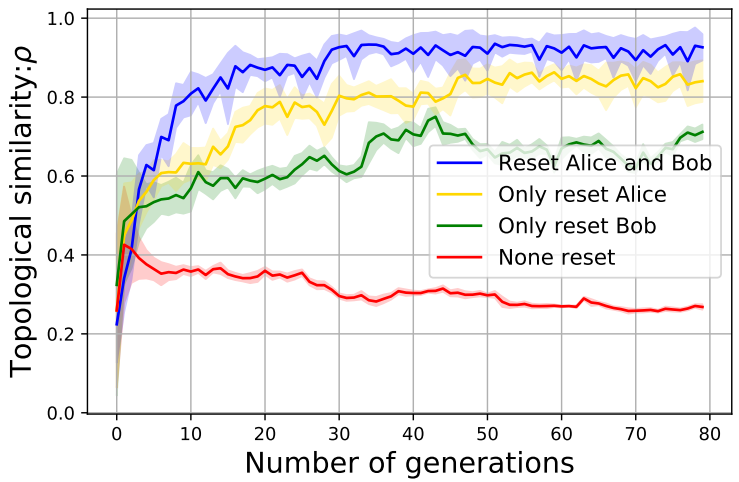
Performance of NIL (2)

Quantitative objective:

$$\mathbb{E}_{\mathcal{L}\sim D_{i+1}}[\rho(\mathcal{L})] \geq \mathbb{E}_{\mathcal{L}\sim D_i}[\rho(\mathcal{L})]. \tag{8}$$



(a) Game performance

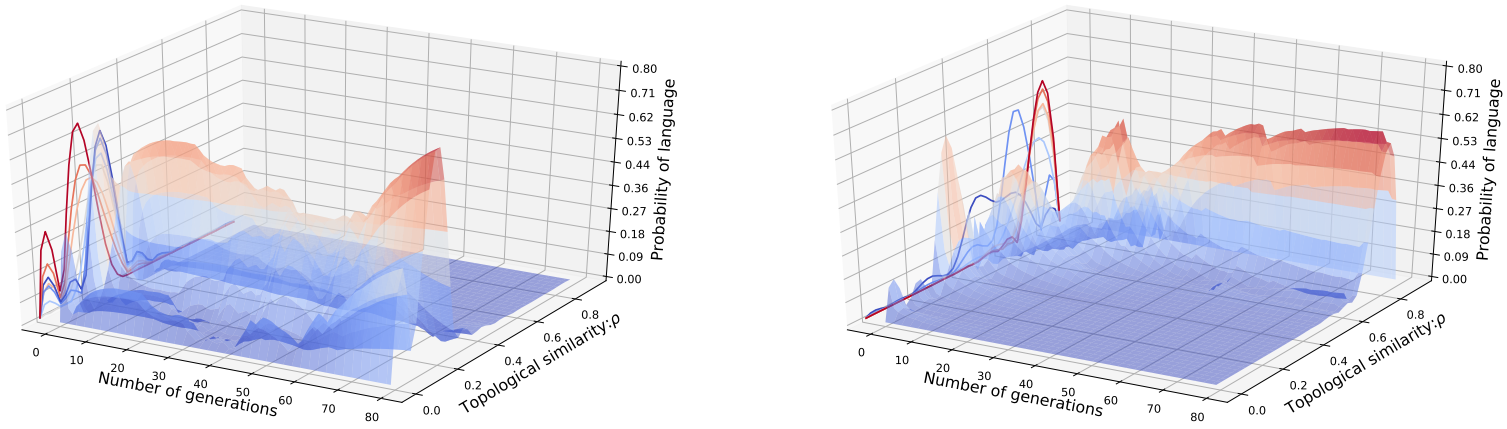


(b) Average ρ of emergent language

Figure: Game performance and average topological similarity for the possible resetting strategies of our proposed iterated learning procedure of 80 generations.

Performance of NIL (3)

Delicate (but same effective) 3D illustrations:



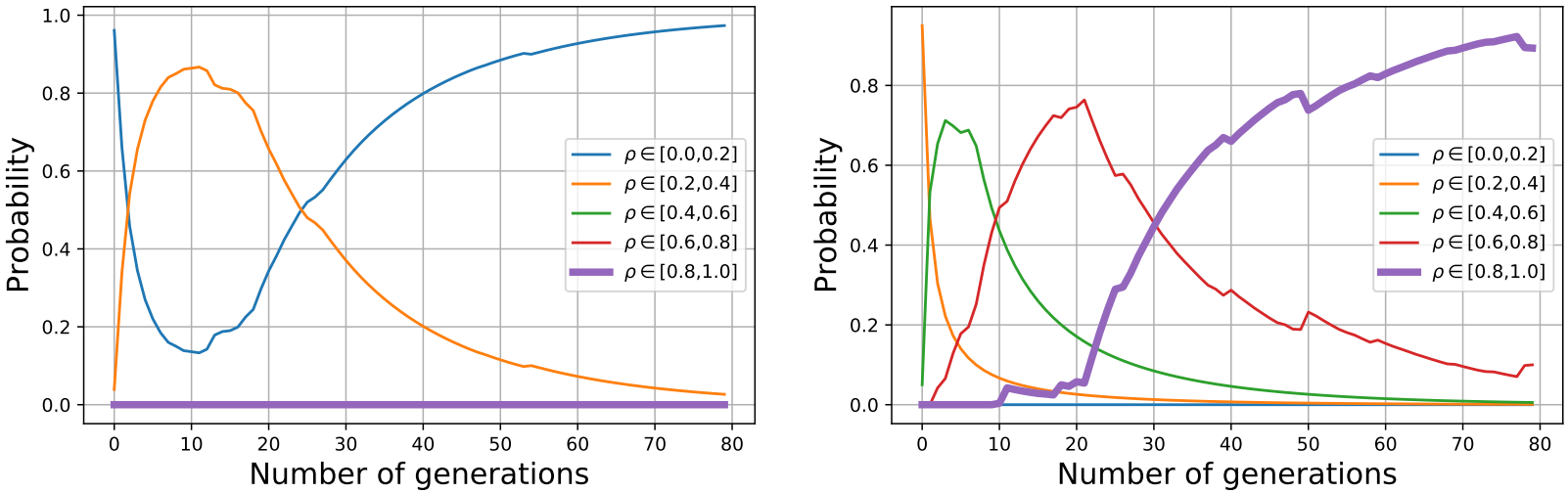
(a) None-reset

(b) Reset-both

Figure: Distribution of $\rho(\mathcal{L})$ over generations.

Performance of NIL (4)

Evolution of language with different compositionality:



(a) None-reset case.

(b) Resetting-both case.

Figure: Evolution of language with different values of ρ.

Robustness for Message Space Size

Size of message space is $|V|^{N_L}$.

	N_L	$\|V\|$= 8	$\|V\|$=12	$\|V\|$=16	$\|V\|$=24	$\|V\|$=40	$\|V\|$=72
$\mathbb{E}[\rho_{71:80}]$	2	0.986±0.01	0.937±0.02	0.933±0.01	0.854±0.02	0.830±0.02	0.793±0.02
	3	0.712±0.01	0.833±0.01	0.798±0.02	0.777±0.01	0.793±0.02	0.780±0.03
$\mathbb{E}[\rho_{1:10}]$	2	0.767±0.18	0.690±0.18	0.684±0.20	0.630±0.17	0.668±0.19	0.572±0.14
	3	0.528±0.11	0.647±0.15	0.640±0.17	0.664±0.14	0.637±0.16	0.628±0.21
$G_{0.85}$	2	9	16	10	37	68	-
	3	-	-	39	-	-	59
Valid Acc.	2	0.868±0.14	0.914±0.06	0.833±0.11	0.866±0.11	0.801±0.10	0.828±0.14
	3	0.804±0.13	0.677±0.16	0.773±0.15	0.858±0.10	0.867±0.01	0.900±0.07

Table: Values of 4 metrics when $|V|$ and N_L changes. Metric $G_{0.85}$ means the first generation that the average ρ of the previous three generations exceed 0.85. The notation "-" means the agents never satisfy the requirement.

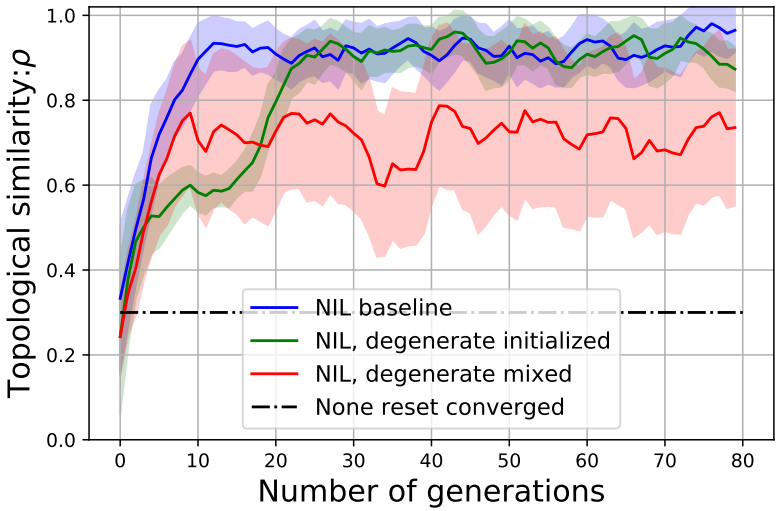
Robustness on Degenerate Components (1)

We designed 2 task to investigate the robustness on degenerate components.

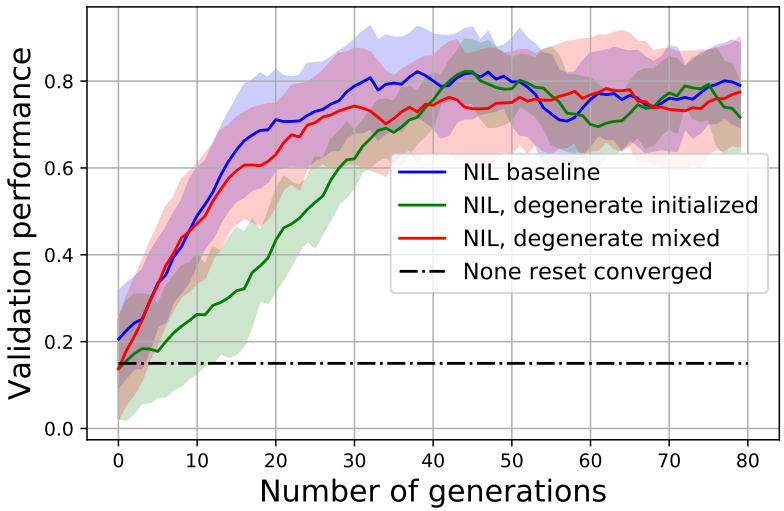
We would make the speaker (Alice)

1. **Degenerate initialized:** Speaker learn from a pure degenerate language instead of D_i.
2. **Degenerate mixed:** Mix the data pair generated by a pure degenerate language to D_i and ensures theproportion of the degenerate pairs is more than 50%.

Robustness on Degenerate Components (2)



(a) Topological similarity



(b) Validation accuracy.

Figure: NIL's robustness to degenerate components.

Effectiveness of Neural Iterated Learning

Methodology

- Learnability of Compositional Languages
- An Optimisation Perspective on the Learnability
- Expressivity of Compositional Languages
- Pre-training listeners
- Bottleneck-effect

Learnability of Compositional Languages (1)

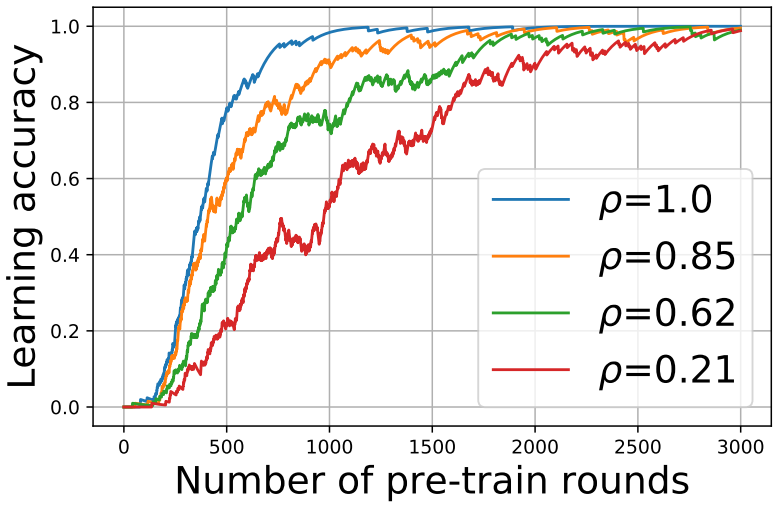
According to Smith et al.[1], 2 pressures are key for the emergence of linguistic structure:
i) Simplicity; ii) Expressivity.

Hypothesis 1

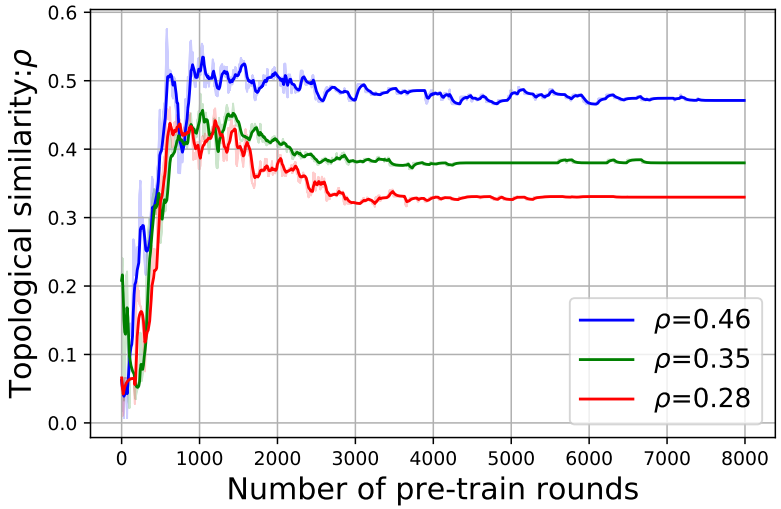
High topological similarity improves the learning speed of the speaking neural agent.

[1] Smith, K., Tamariz, M., & Kirby, S. (2013). Linguistic structure is an evolutionary trade-off between simplicity and expressivity. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 35, No. 35).

Leanability of Compositional Languages (2)



(a) Learning accuracy of Alice.



(b) Alice's ρ

Figure: Illustration of the learning speed of Alice pre-trained with languages of various topological similarities.

Optimisation Perspective

Toy Example (1)

Meaning space: Blue Box (BB), Red Box (RB), Blue Circle (BC), Red Circle (RC)
A compositional language: BB -> aa, RB -> ba, BC -> ab, RC -> bb.
A holistic language: BB -> ba, RB -> aa, BC -> ab, RC -> bb.

$$P(\mathcal{L}_{\text{cmp}}) = P(m_1 = a | BB) \cdot P(m_2 = a | BB, m_1 = a) \cdot P(m_1 = b | RB) \cdot P(m_2 = a | RB, m_1 = b) \cdot C \quad (9)$$

$$P(\mathcal{L}_{\text{hol}}) = \underbrace{P(m_1 = b | BB)}_{①} \cdot \underbrace{P(m_2 = a | BB, m_1 = b)}_{②} \cdot \underbrace{P(m_1 = a | RB)}_{③} \cdot \underbrace{P(m_2 = a | RB, m_1 = a)}_{④} \cdot C \quad (10)$$

$$C = \underbrace{P(m_1 = a | BC)}_{⑤} \cdot \underbrace{P(m_2 = b | BC, m_1 = a)}_{⑥} \cdot \underbrace{P(m_1 = b | RC)}_{⑦} \cdot \underbrace{P(m_2 = b | RC, m_1 = b)}_{⑧} \quad (11)$$

Optimisation Perspective

Toy Example (2)

Learning $\langle ab, BC \rangle$ would:

- Increase: $P(m_1 = a | BC)$; $P(m_1 = a | BB)$; $P(m_1 = a | RC)$.
- Decrease: $P(m_1 = b | BC)$; $P(m_1 = b | BB)$; $P(m_1 = b | RC)$.

Thus,

- For $P(\mathcal{L}_{\text{cmp}})$, ⑤ $\uparrow$ and ① $\uparrow$.
- For $P(\mathcal{L}_{\text{hol}})$, ⑤ $\uparrow$ and ① $\downarrow$.
- $\Delta P(\mathcal{L}_{\text{cmp}}) > \Delta P(\mathcal{L}_{\text{hol}})$

Optimisation Perspective

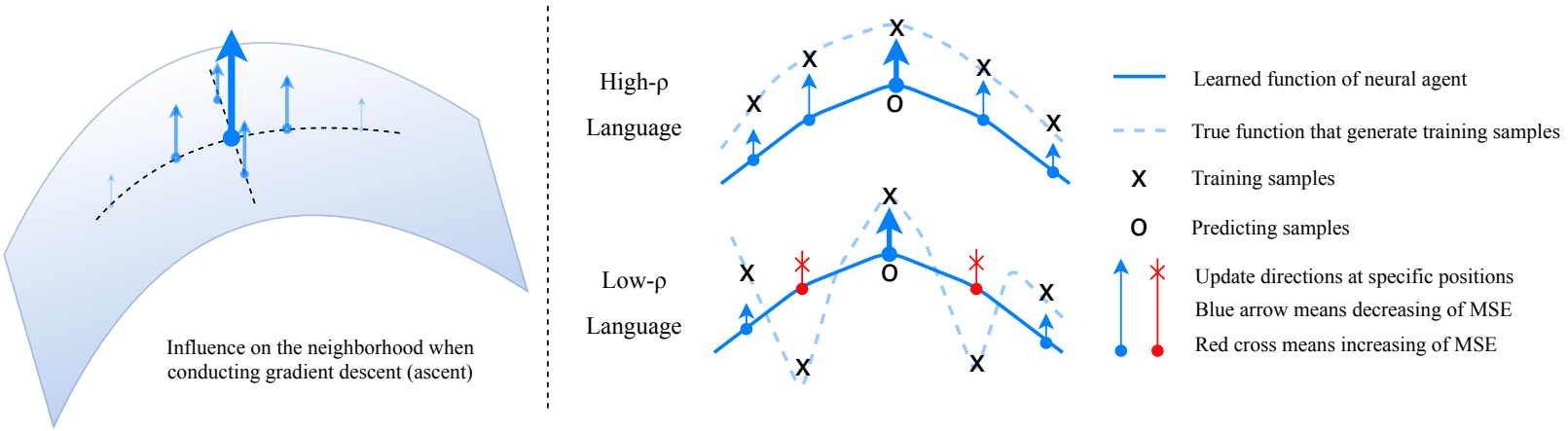
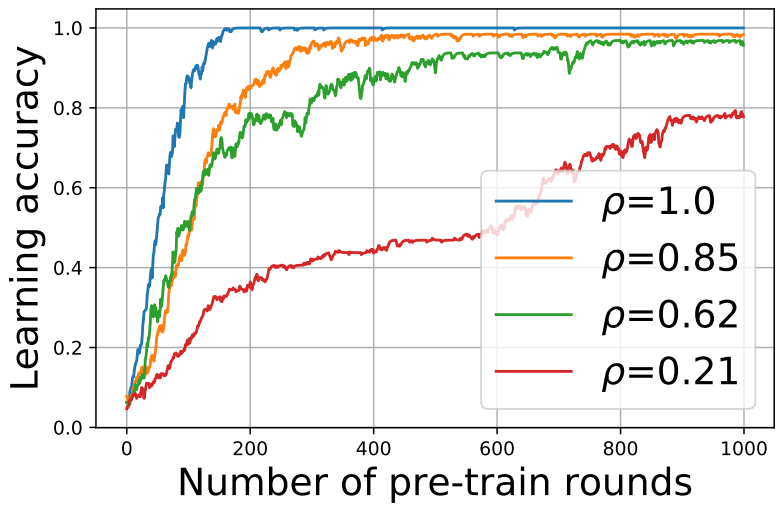


Figure: Illustration of learning a high-ρ language and low-ρ language.

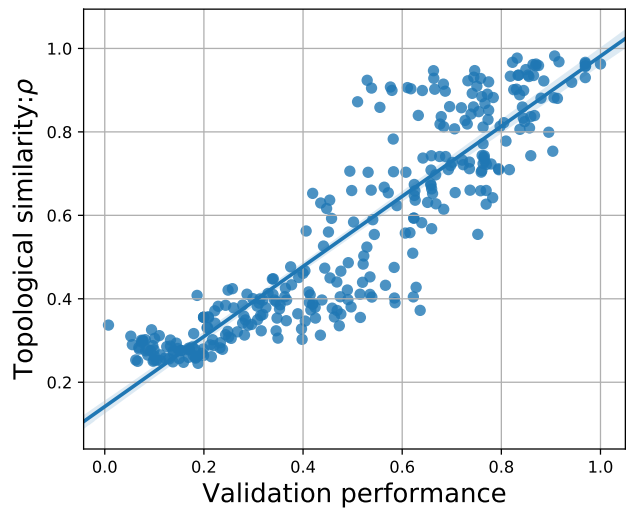
Expressivity of Compositional Languages (1)

Hypothesis 2

High topological similarity allows the listening agent to successfully recognize more concepts, using less samples.



(a) Learning accuracy of Bob.



(b) Validation performance and ρ.

Expressivity of Compositional Languages (2)

NIL improves generalisation performance

Valid set size	0		8		16		32	
	Train	Valid	Train	Valid	Train	Valid	Train	Valid
No-reset	0.985	-	0.986	0.136	0.990	0.132	0.995	0.102
Bob-reset	0.967	-	0.943	0.094	0.962	0.152	0.947	0.116
Alice-reset	0.981	-	0.976	0.598	0.979	0.280	0.947	0.210
Both-reset	0.988	-	0.986	0.847	0.984	0.755	0.973	0.558

Table: Validation performance under different validation set sizes.

When compositional language has no advantage

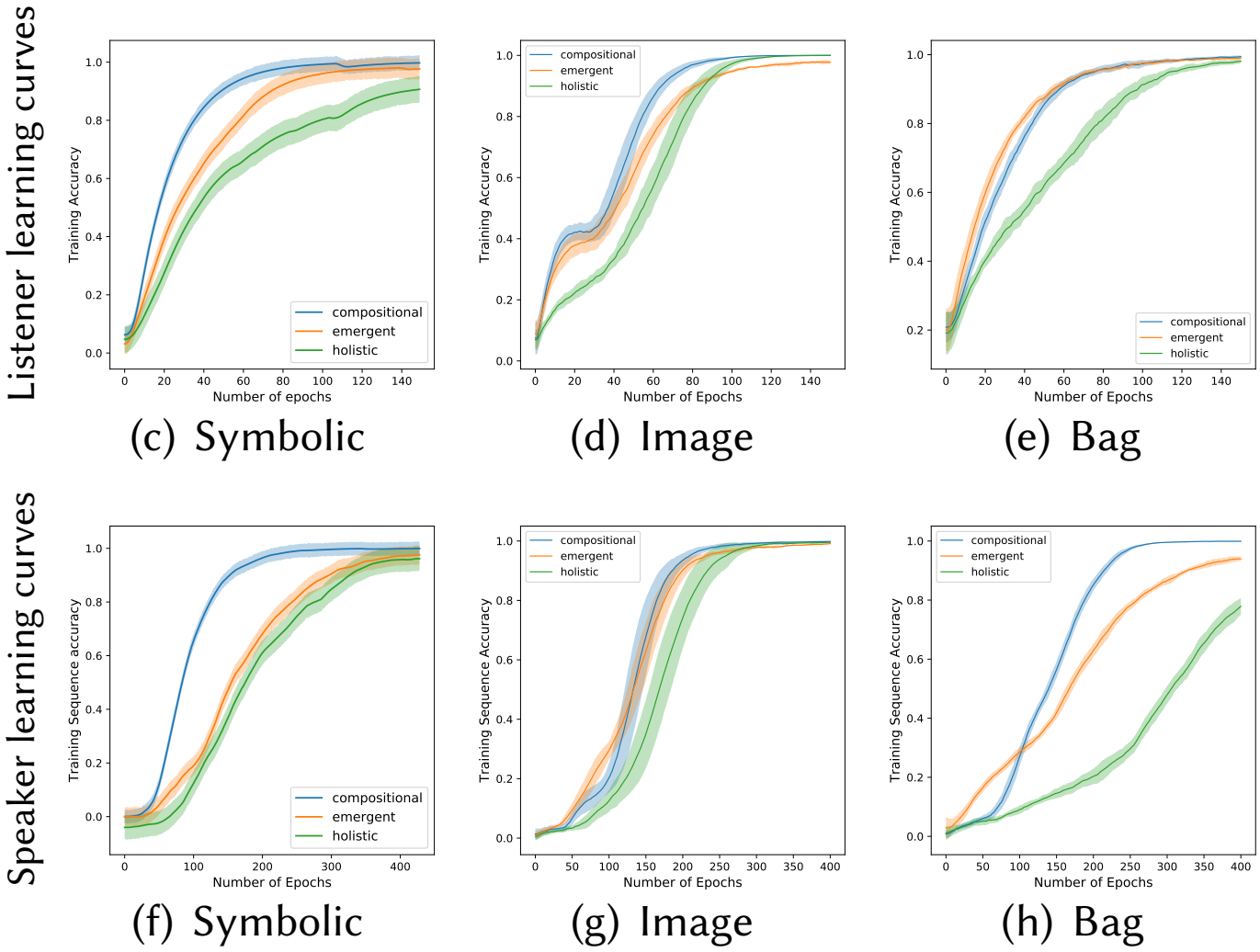
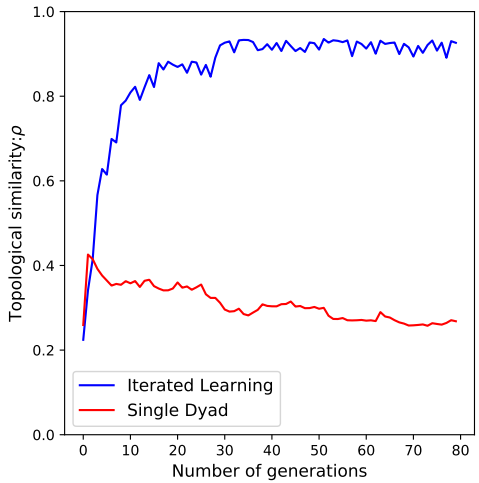
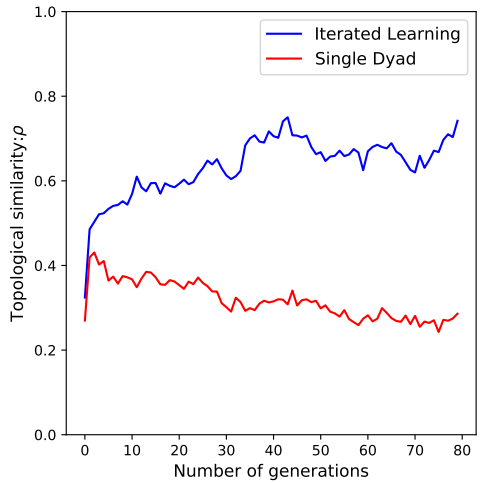


Figure: Experiments on different input formats.

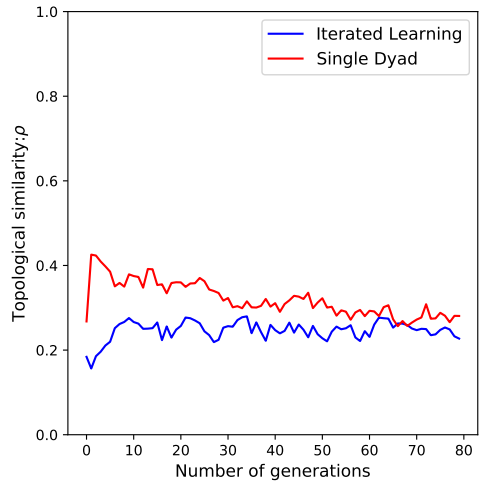
When compositional language has no advantage



(a) Symbolic



(b) Image



(c) Set

Figure: Experiments results on different input representations.

Role of Pre-training Listener

Overall, pre-training listeners makes NIL more robust, but why?

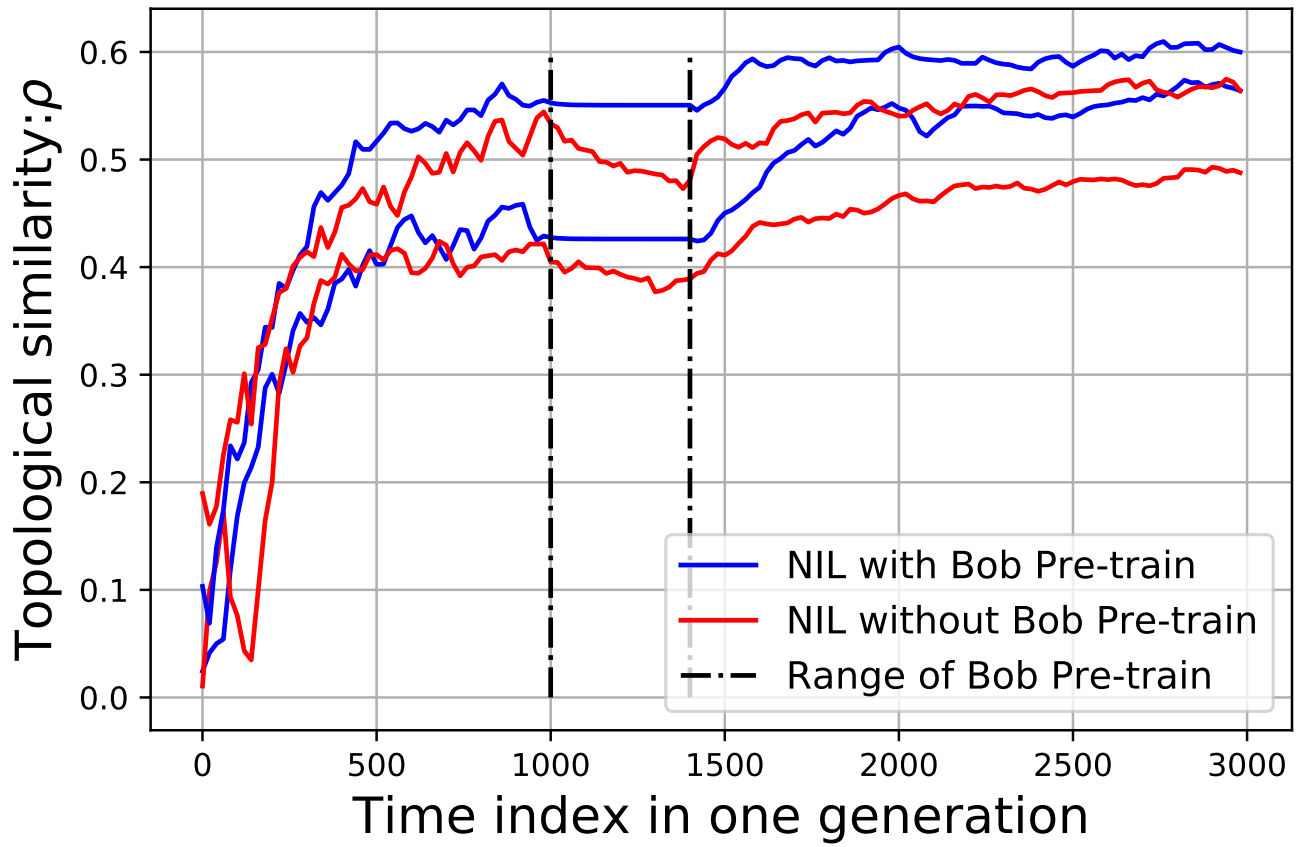


Figure: The change of $\mathbb{E}_{\mathcal{L}\sim P(\mathcal{L}|\text{NIL})}[\rho(\mathcal{L})]$ in generation 3 and 6.

Bottle-neck Effect

Original IL:
The amounts of data available in transmission phase.

NIL: Limiting the number of pre-training rounds of the agents to the "interval of advantage".

Conclusions

Conclusions

- The learning speed advantages of highly compositional languages.
- Neural iterated learning model to utilise the advantages.
- Performance and robustness of NIL.
- Probabilistic model of emergent languages.

Questions?

Any question about our work is welcome to be sent to sg955@cam.ac.uk or renyi.joshua@gmail.com.

Thank you