Thesis for the degree of Doctor of Philosophy

Reinforcement Learning: Efficient Communication and Sample Efficient Learning

Emil Carlsson

Division of Data Science and AI
Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2024

Reinforcement Learning: Efficient Communication and Sample Efficient Learning
Emil Carlsson
ISBN: 978-91-8103-128-7

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 5586
ISSN 0346-718X

Division of Data Science and AI
Department of Computer Science and Engineering
Chalmers University of Technology
SE–412 96 Göteborg, Sweden
Telephone + 46 (0) 31 – 772 1000

Typeset by the author using LaTeX.

Printed by Chalmers Reproservice
Göteborg, Sweden 2024

till Francis

Abstract

Life is full of decision-making problems where only partial information is available to the decision-maker and where the outcomes are uncertain. Whether choosing a restaurant for dinner, selecting a movie on a streaming service, or conveying concepts during a lecture, the decision-maker observes only the results of their choices without knowing what would have happened if it had acted differently. Because of this, the decision-maker needs to carefully balance between using its current knowledge, to make good decisions, and exploring the unknown to gather new information that might lead to even better decisions in the future.

In this thesis, we explore several topics in reinforcement learning - a computational approach to sequential decision-making under uncertainty. The first part investigates how efficient communication emerges between reinforcement learning agents in signaling games. The support for efficient communication, in an information-theoretic sense, is an important characteristic of human languages. Our agents create artificial languages that are as efficient as human languages as well as similar to human ones. We also combine reinforcement learning with iterated learning and find that this combination accounts better for human color naming systems than what any of the models do individually.

The second part focuses on sample-efficient algorithms for multi-armed bandits. We propose Thompson sampling-based methods for regret minimization in multi-armed bandits with clustered arms. Additionally, we address finding optimal policies with fixed confidence in bandits with linear constraints. For this problem, we characterize a lower bound and illustrate how it depends on a non-convex projection onto the normal cone spanned by the constraints. We leverage these insights to derive asymptotically optimal algorithms for pure exploration in bandits with linear constraints. Finally, we apply techniques from multi-armed bandits to develop active learning strategies for ordering items based on noisy preference feedback.

Keywords: Reinforcement Learning, Multi-armed Bandits, Contextual Bandits, Efficient Communication, Emergent Communication, Iterated Learning, Pure Exploration, Color Naming, Numeral Systems, Preference Learning.

Acknowledgments

Throughout this journey, I've been privileged to have had great and inspiring people around me, and I have made many friends. First, I want to thank my supervisor, Devdatt Dubhashi. This thesis would not have been possible without Devdatt's unwavering support and knowledge. He has always encouraged me to go the extra mile and set the bar high. Together, we have generated countless research ideas, some of which made it to actual papers and are part of this thesis. I would also like to thank my co-supervisor, Fredrik D. Johansson, for all the support during these years. Fredrik's door has always been open, whether I wanted to discuss research or other matters like the latest football games. I am very grateful to Terry Regier for hosting me at UC Berkeley. Terry has always been a source of inspiration, and our two semesters together helped me mature as a researcher. Thanks to my examiner, Dag Wedelin, for all his support and interesting questions during my follow-up meetings.

My PhD years wouldn't have been half as fun if it weren't for Niklas, Tobias, Edvin, and Emilio. We have shared many enjoyable moments, and they have always been there supporting me during stressful times. I would also like to thank my co-authors, Debabrota, Herman, Ahmet, Newton, Jonathan, Andrea, Mikael, Asad, and Moa. Working with you all has been inspiring and a pleasure. I am also very grateful to all the other PhD students at the division whose presence has made the visits to the office way more fun.

My friends outside the department should not be forgotten, especially those from my undergraduate years: Carl, Wille, Erik, Jerry, Alfred, Garcia, and Filip. They have been a continuous support throughout my entire PhD. I would also like to send a warm thanks to my friends from back home, Anton, Pontus, and Oscar, for their support all these years.

I thank my family for their never-ending support and love. My mother, father, and two sisters have always asked interesting questions about my research and supported me. My aunt Lotta encouraged me to be curious and pursue research from a very young age. My soon-to-be wife, Emelie, has always supported me, putting up with my crazy ideas and sometimes bad planning. Our son, Francis, always greets me with the biggest and brightest smile when I come home from work.

Lastly, I thank Chalmers AI Research Centre (CHAIR) for enabling my research via their generous grant and the Sweden-America Foundation (SweAm) for funding my research visit to UC Berkeley.

Emil Carlsson
Göteborg, October 2024

List of publications

This thesis is based on the following appended papers:

Paper 1. Mikael Kågebäck, Emil Carlsson, Devdatt Dubhashi, Asad Sayeed. *A reinforcement learning approach to efficient communication.* PLoS ONE, 15(7):1–26, 2020.

Paper 2. Emil Carlsson, Fredrik D. Johansson, Devdatt Dubhashi. *Learning approximate and exact numeral systems via reinforcement learning.* Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci) , 43 2021.

Paper 3. Emil Carlsson, Devdatt Dubhashi. *Pragmatic reasoning in structured signaling games.* Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci) , 44, 2022.

Paper 4. Emil Carlsson, Devdatt Dubhashi, Terry Regier. *Cultural evolution via iterated learning and communication explains efficient color naming systems.* To appear in the Journal of Language Evolution. An earlier version of this paper appeared in Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci) 45, 2023.

Paper 5. Emil Carlsson, Fredrik D. Johansson, Devdatt Dubhashi. *Thompson sampling for bandits with clustered arms.* Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI), 2021.

Paper 6. Emil Carlsson, Debabrota Basu, Fredrik D. Johansson, Devdatt Dubhashi. *Pure exploration in bandits with linear constraints.* International Conference on Artificial Intelligence and Statistics (AISTATS), 2024.

Paper 7. Herman Bergström*, Emil Carlsson*, Devdatt Dubhashi, Fredrik D. Johansson. *Active preference learning for ordering items in- and out-of-sample.* To appear in the Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS), 2024. $*$ indicates equal contribution.

The following publications have been made during the author's time as a PhD student but are not part of this thesis:

Paper 8. Erik Jergéus, Leo Karlsson Oinonen, Emil Carlsson, and Moa Johansson. *Towards Learning Abstractions via Reinforcement Learning.* 8th International Workshop on Artificial Intelligence and Cognition (AIC), 2022.

Paper 9. Newton Mwai Kinyanjui, Emil Carlsson, and Fredrik D. Johansson. *Fast Treatment Personalization with Latent Bandits in Fixed-Confidence Pure Exploration* Transactions on Machine Learning Research (TMLR), 2023.

Paper 10. Emil Carlsson, Devdatt Dubhashi, Terry Regier. *Iterated learning and communication jointly explain efficient color naming systems.* Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci) 45, 2023.

Paper 11. Jonathan David Thomas, Andrea Silvi, Devdatt Dubhashi, Emil Carlsson, and Moa Johansson. *Learning Efficient Recursive Numeral Systems via Reinforcement Learning.* AI for Math Workshop @ ICML, 2024.

Paper 12. Ahmet Zahid Balcıoğlu, Emil Carlsson, and Fredrik D. Johansson. *Identifiable latent bandits: Combining observational data and exploration for personalized healthcare.* ICML Workshop: Foundations of Reinforcement Learning and Control – Connections and Perspectives, 2024.

Summary of contributions

The contributions to the appended papers by the author of this thesis are listed below.

Paper 1. Contributed to the code, contributed to the experiments, and contributed with visualization and writing after receiving initial reviews.

Paper 2. Co-designed the study, wrote the code, performed the experiments, analysed the results, and wrote most of the manuscript.

Paper 3. Co-designed the study, wrote the code, performed the experiments, analysed the results, and wrote most of the manuscript.

Paper 4. Co-designed the study, wrote the code, performed the experiments, analysed the results, and contributed to the writing of the manuscript.

Paper 5. Co-designed the study, wrote the code, proved the theoretical statements, performed the experiments, analysed the results, and wrote most of the manuscript.

Paper 6. Co-designed the study, wrote the code, proved the theoretical statements, performed the experiments, analysed the results, and wrote most of the manuscript.

Paper 7. Co-designed the study, proved the theoretical statements, had a supporting role in writing the code, and contributed to the writing of the manuscript. The first two authors contributed equally to the paper.

Contents

Part I

Introductory Chapters

Chapter 1

Introduction

Life is full of decision-making problems where only partial information is available to the decision-maker and where the outcomes are uncertain. Whether choosing a restaurant for dinner, selecting a movie on a streaming service, or conveying concepts during a lecture, the decision-maker observes only the results of their choices without knowing what would have happened if it had acted differently. Because of this, the decision-maker needs to carefully balance between using its current knowledge, to make good decisions, and exploring the unknown to gather new information that might lead to even better decisions in the future. This trade-off is known as the *exploration-exploitation trade-off* and is a central challenge faced by both human and artificial decision-makers in any sequential decision-making problem with uncertain outcomes.

A computational approach to decision-making under uncertainty is *reinforcement learning* (Sutton and Barto 1998) which has grown in popularity in recent years. In this framework, an artificial agent interacts with its environment (and potentially other agents) and receives feedback in the form of rewards. The goal of the agent is to learn a policy, i.e., a way of acting given a certain state of the environment, that maximizes the agent's rewards over time. Reinforcement learning has been successfully applied in a wide range of domains such as recommender systems (Li, Chu, et al. 2010), navigation (Åkerblom et al. 2023), healthcare (Yu et al. 2021), games (Mnih et al. 2015; Silver et al. 2016), and robotics (Kober et al. 2013). In addition, due to its emphasis on learning from interactions with the environment, something that is a fundamental aspect of both animal and human intelligence (Thorndike 1898; Rovee and Rovee 1969; Piaget 2013), reinforcement learning has also been used as a model in neuroscience and psychology (Niv 2009; O'Doherty et al. 2015; Gershman and Daw 2017).

A decision-making problem that will be central to this thesis, and which is often studied in cognitive science, is how to communicate certain concepts to others. *Why are concepts mapped to words the way they are? What processes lead to patterns found in human languages?* These are all central questions in cognitive science and a prominent proposal suggests that human languages are shaped to support efficient communication in an information-theoretic sense (Kemp, Xu, et al. 2018; Gibson, Futrell, Piantadosi, et al. 2019). This means that human languages are

simultaneously optimized to be simple, to ease learnability and reduce cognitive load, and to be informative, to support accurate communication.

The main contribution of this thesis is connecting concepts from reinforcement learning with results regarding efficient communication in human languages. We will study how reinforcement learning agents that communicate with each other in various signaling games (Lewis 1969) develop joint artificial languages. In the basic version of these games, a speaker observes a concept and tries to communicate this concept to a listener. Upon hearing the message, the listener guesses which concept the speaker refers to from a set of available concepts. A reward is provided to both the speaker and listener depending on how well they communicated. The agents start as *tabula rasa* and develop an artificial language by maximizing their joint reward function. We find that reinforcement learning leads to artificial languages with similar levels of efficiency as their human counterparts and these artificial languages tend to be human-like. Our results open up the question of whether similar mechanisms could be involved in shaping human languages toward efficiency and suggest that reinforcement learning may be a useful building block for studying language evolution *in silico*.

The aforementioned signaling game falls into a class of reinforcement learning problems known as *multi-armed bandit* problems (Lattimore and Szepesvári 2020). In a bandit problem, a reinforcement learner sequentially interacts with the environment by executing actions, also known as arms, and then obtains, potentially noisy, rewards associated with the arms that were played. An extension of this model is the *contextual bandit* where contextual cues are revealed to the learner to help guide it towards arms with high rewards. In contrast to the general reinforcement learning problem, temporal dependencies between actions and contexts are not modeled in a bandit problem. This means that the current context and potential rewards are assumed to be independent of previously observed actions and contexts. As a result, bandit models are simpler and more tractable models for studying decision-making under uncertainty compared to general reinforcement learning.

The signaling game can be viewed as a multi-agent contextual bandit. From the speaker's perspective, the observed concept provides contextual information and the set of possible messages can be viewed as the set of arms in a bandit problem. The message sent serves as a contextual cue for the listener who then has to decide what concept, or arm, to play from the set of available concepts. This view was recently leveraged to study how humans use language (Sumers et al. 2023) and we will make use of it throughout this thesis.

In addition to studying the emergence of efficient communication via reinforcement learning, a second contribution of this thesis is sample-efficient algorithms designed for various multi-armed bandit tasks. In practice, there are often structures and various constraints imposed on the set of arms available to the learner. These structures might be exploited for faster learning while constraints can make the learning problem both easier and harder. One example of such a structure studied in this thesis is when a clustering of the arms is available to the learner. We also study the effect of constraints on the arms and characterize how this changes the hardness of the problem.

The papers forming this thesis are listed below. They have been categorized depending on whether they study the emergence of efficient communication or if they study efficient learning in the multi-armed bandit framework.

Efficient communication

- Paper 1 (Kågebäck et al. 2020) proposes a multi-agent reinforcement learning approach to the partitioning of semantic spaces. This is explored in the domain of colors where the reinforcement learning agents develop color naming systems that achieve a near-optimal trade-off between communicative efficiency and complexity. The efficiency of the artificial naming systems is on the same level of efficiency as color naming systems found in human languages.

- Paper 2 (Carlsson, Dubhashi, and Johansson 2021a) explores how efficient numeral systems emerge in a communicative dyad of reinforcement learning agents. The agents develop efficient exact and approximate numeral systems that are similar to those found in human languages. These results give a learning-theoretic account of how these systems might have emerged to be efficient.

- Paper 3 (Carlsson and Dubhashi 2022) studies what impact coupling reinforcement learning with pragmatic reasoning has on the efficiency of the resulting languages. The paper also introduces a pragmatic reasoning model that better accounts for the structure of the domain and the current context the agents communicate in. The model is evaluated in the domain of colors and the results suggest that the emerging vocabulary becomes less complex when the agent's reasoning capabilities grow stronger.

- Paper 4 (Carlsson, Dubhashi, and Regier 2024) revisits the color experiments from Paper 1 and couples reinforcement learning with iterated learning, a model for how language is shaped over generations of agents. The resulting color naming systems better match human systems than the systems produced in Paper 1 and the systems produced by exclusively applying iterated learning. The paper also introduces a simple random model that generates highly efficient naming systems that share very little similarity with human systems. This highlights the importance of studying plausible evolutionary models that result in efficient and human-like languages. Note that this paper is an extended version of our conference contribution Carlsson, Dubhashi, and Regier (2023).

Efficient learning in the multi-armed bandit framework

- Paper 5 (Carlsson, Dubhashi, and Johansson 2021b) introduces Thompson sampling algorithms for multi-armed bandits with clustered arms. Clusterings appear naturally in many decision-making tasks and we show, both theoretically and empirically, that our proposed algorithms outperform baselines.

- Paper 6 (Carlsson, Basu, et al. 2024) introduces algorithms for finding the optimal policy in multi-armed bandits where arms are subject to linear constraints. We prove that our proposed algorithms have optimal sample complexity in an asymptotic sense. The algorithms also outperform baselines in our empirical evaluation.

- Paper 7 (Bergström et al. 2024) introduces an active sampling strategy, based on multi-armed bandits, for ordering items under noisy comparison feedback. Our proposed sampling strategy outperforms the baseline in both synthetic and real-world experiments.

During the time as a PhD student, the following publications have been made by the author but are not part of the thesis: Jergéus et al. (2022), Kinyanjui et al. (2023), Thomas, Silvi, et al. (2024), and Balcıoğlu et al. (2024).

The rest of the thesis is structured as follows. In Chapter 2 we introduce relevant concepts from reinforcement learning and multi-armed bandits. In Chapter 3 we discuss relevant concepts and results from cognitive science, regarding human languages, and how reinforcement learning fits into this picture. This chapter is mostly relevant for Paper 1 to Paper 4. Chapter 4 summarizes the papers that this thesis is based on, and in Chapter 5 we discuss our conclusions and potential future directions. The papers are appended in the second part of this thesis and have been reformatted for uniformity, but are otherwise unchanged.

Chapter 2

Reinforcement learning and multi-armed bandits

This chapter gives a brief introduction to reinforcement learning and bandit problems. For a more comprehensive introduction to reinforcement learning see Sutton and Barto (1998) and for some recent textbooks on multi-armed bandits see Slivkins (2019) and Lattimore and Szepesvári (2020).

2.1 What is reinforcement learning?

The goal of reinforcement learning is to design computational agents that seek to maximize a notion of reward in their corresponding environments (Sutton and Barto 1998). In contrast to supervised learning, where the agent is provided a dataset of input-output pairs, the reinforcement learning agent gathers its data by interacting with the environment. This gives rise to the famous exploration-exploitation trade-off, where the agent must balance between exploiting its current knowledge about the environment, to achieve high reward, and exploring new actions that might lead to even higher rewards in the future.

Algorithm 1 The Markov decision making process.

Require: A set of states $\mathcal{X}$, a set of actions $\mathcal{A}$, a transition kernel P, a reward function R, initial state x_1, a policy π.
for t=1,... **do**
Take action $a_t \in \mathcal{A}$ by sampling from the policy $a_t \sim \pi(x_t)$.
The environment samples a new state $x_{t+1} \sim P(x_t, a_t)$ and reveals a reward $r_t \sim R(x_t, x_{t+1}, a_t)$.
end for

In reinforcement learning, a learner sequentially interacts with the environment: It observes the current state of the environment, takes an action, and observes a reward and the new state. The core challenge is to design a policy π that maximizes the cumulative reward the agent achieves in the environment. The interaction with the environment is often modeled as a *Markov decision process* (MDP) (Bellman

1957). This model assumes the *Markov property* which says that the state-transition only depends on the current state and the action taken in this state. The MDP model is not central to this thesis but we illustrate it in Algorithm 1 so that the reader can more easily see how the bandit models, introduced in later sections, are simplifications of this more general framing of reinforcement learning.

2.2 Multi-armed bandits

In a multi-armed bandit, a reinforcement learner iteratively interacts with the environment by playing an action, also known as arm, a_t at every time step t and observes a reward, r_t, drawn from a probability distribution, with unknown mean, associated with the chosen arm. In contrast to the general reinforcement learning problem, there is either no state or the state is constant in the multi-armed bandit and as a result, the learner doesn't need to model any temporal dependencies or relations between state and reward. Hence, the learner only needs to model the relationship between arms and rewards. The problems one considers in the bandit model can often be categorized into either *regret minimization* or *best-arm identification*, also known as *pure exploration*.

Algorithm 2 The multi-armed bandit.

Require: A set of arms $\mathcal{A}$, a reward distribution for each arm R, and a policy π.
for t=1,... **do**
Play arm according to learner's policy $a_t \sim \pi_t$.
Observe reward $r_t \sim R(a_t)$ drawn from a probability distribution associated with a_t.
Update learner's policy to π_{t+1}.
end for

Regret minimization: In regret minimization for multi-armed bandits, the goal of the learner is to maximize its cumulative reward over a time horizon T (Lai and Robbins 1985). Maximizing the cumulative reward is equivalent to minimizing the cumulative regret, defined as

$$\text{Regret}_T = \sum_{t=1}^{T} r^* - r_t,$$

where r^* denotes the reward drawn from the arm with the highest expected reward, a^*. In this regime, the goal is often to design algorithms with good guarantees on their expected cumulative regret, $\mathbb{E}[\text{Regret}_T]$. We study regret minimization for bandits with clustered arms in Paper 5.

Fixed-confidence best-arm identification: In this regime, the goal of the learner is to interact with the bandit until they are sufficiently confident in which

arm is the one with the largest mean (Chernoff 1959). More formally, the learner interacts with the bandit and stops at some random time, τ, and recommends some arm, $\hat{a}_\tau$, which should be equal to the best arm, a^*, with probability at least $1-\delta$, for some predefined $\delta \in (0,1)$, i.e.,

$$P(\hat{a} \neq a) \leqslant \delta.$$

In this setting, one would like to design learning algorithms that minimize the expected sample complexity, $\mathbb{E}[\tau]$, while still ensuring that the fixed confidence level δ is reached. The property that the learner stops and outputs the correct arm with probability at least $1-\delta$ is referred to as δ-*PAC*. Fixed confidence best-arm identification is relevant for Paper 6.

Fixed-budget best-arm identification: Here the learner is given a fixed budget T and needs to play arms such that the probability of recommending the wrong arm, once the budget is depleted, is minimized (Audibert and Bubeck 2010). This problem is, at least conceptually, the dual of the fixed confidence setting even though some open problems for the fixed budget are closed in the fixed confidence version (Qin 2022). The reason there is a gap between the settings is because many theoretical results in the fixed-confidence regime are in an asymptotic sense, e.g., when $\delta \to 0$ and thus not easy to translate to the fixed-budget setting since this setting is inherently non-asymptotic. In Paper 7 we study active learning for ordering and our algorithm builds on results from fixed-budget best-arm identification.

Remark: Even though regret minimization and best-arm identification are related, algorithms for regret minimization are not suitable for best-arm identification and vice versa (Bubeck et al. 2009; Russo 2016). The main reason is that regret minimization algorithms focus on quickly identifying good arms, to minimize regret, while best-arm identification algorithms often need to allocate more plays to sub-optimal arms to gather enough statistical evidence.

2.3 The contextual bandit

Algorithm 3 The contextual bandit.

Require: A set of arms $\mathcal{A}$, a set of contexts $\mathcal{X}$, a reward function R, and a policy π .
for t=1,... **do**
  Observe context $x_t \in \mathcal{X}$.
  Play arm according to learner's policy $a_t \sim \pi_t(x_t)$.
  Observe reward $r_t \sim R(x_t, a_t)$.
  Update learner's policy to π_{t+1}.
end for

In the contextual bandit, the learner observes, at every time step, a context x_t before deciding which arm to play. The reward for an arm a at time t is assumed to

be an unknown and stochastic function of both the arm and the context, $r(x, a)$. The key distinction between the contextual bandit and the general reinforcement learning problem is that the context x_t is assumed to be independent of previous contexts and actions. Thus, the learner does not need to model any temporal dependences, in contrast to general reinforcement learning. The contextual bandit model is mostly relevant for the appended papers related to the emergence of artificial languages (Paper 1 to Paper 4). In these papers, we consider various signaling games, properly introduced in Section 3.2.1, that can be viewed as instances of the contextual bandit. We also study a contextual bandit in Paper 5.

2.4 Lower bounds in multi-armed bandits

In multi-armed bandit work, an important task is to characterize what is theoretically possible under some given assumptions. This is done by deriving information-theoretic lower bounds, on either the cumulative regret or the sample complexity, that holds true for any learning algorithm from some family of algorithms.

Let $\mathcal{M}$ be the set of all possible bandit environments. Let $\mu \in \mathcal{M}$ be a particular bandit environment and let μ_a denote the mean reward of arm a. In the case when the reward distributions are parameterized only by their mean, we let $\mathcal{M} = \mathbb{R}^K$. We assume the best arm to be unique and define the set of *alternative instances* w.r.t. μ as

$$\Lambda(\mu) := \left\{ \lambda \in \mathcal{M} : \arg\max_a \lambda_a \neq \arg\max_a \mu_a \right\}.$$

The set $\Lambda(\mu)$ contains all possible bandit environments where the best arm *differs* from the best arm in the environment parameterized by μ[1]. If the true environment is μ but we, given the data we observe so far, think it is some $\lambda \in \Lambda(\mu)$, we will make the wrong decision. Thus, bandit problems can be viewed as sequential hypothesis testing where to goal is to sample arms in a way that ensures, with high probability, that our estimate $\hat{\mu}_t$ of the true environment μ satisfies $\hat{\mu}_t \notin \Lambda(\mu)$. Exactly how the sampling should be done is dictated by whether we are performing regret minimization or best-arm identification.

In the fixed confidence best-arm identification setting, mention in Section 2.2, Kaufmann et al. (2016) derived the following generic lower bound on the expected stopping time, $\mathbb{E}[\tau]$, of any δ-PAC learner and for any $\mathcal{M}$

$$\mathbb{E}[\tau] \geqslant \mathcal{T}(\mu) \log \frac{1}{2.4\delta} \tag{2.4.1}$$

where $\mathcal{T}(\mu)$ is the solution to

$$\mathcal{T}^{-1}(\mu) = \sup_{w:\sum_a w_a = 1} \inf_{\lambda \in \Lambda(\mu)} \sum_a w_a \mathbb{KL}(\mu_a || \lambda_a). \tag{2.4.2}$$

[1]This definition of the alternative set only works for the multi-armed bandit and not the contextual version. However, it is possible to extend this to the contextual case (Magureanu et al. 2014; Kato and Ariu 2024)

Here, w is the fraction of plays the learner allocates to the different arms and λ is some instance from $\Lambda(\mu)$. Equation (2.4.2) can be interpreted as a zero-sum game where the learner plays an exploration strategy, w, and an adversary plays an instance λ that will be hard to reject given the strategy of the learner. Note that this bound doesn't make any assumptions on the structure of the model class and is thus a generic bound. However, the exact value of $\mathcal{T}(\mu)$ depends on the specific model class considered since the model class dictates the structure of $\Lambda(\mu)$ and thus controls the set over which the infimum is taken over. This lower bound result serves as a starting point for our work in Paper 6.

Moreover, in Chapter 5 we briefly discuss how these types of results might open up interesting research directions when it comes to language evolution and learnability of language. In short, one could let μ be the language a learner is trying to learn and let $\Lambda(\mu)$ be the set of languages that differs distinctly from μ. One could then ask whether the language μ is fundamentally easy to learn, measured by whether the lower bound on the sample complexity is relatively small.

2.5 Relevant algorithms

This section introduces some of the bandit algorithms relevant for this thesis.

2.5.1 REINFORCE

The REINFORCE algorithm (Williams 1992) is an algorithm used in reinforcement learning when the policy is parameterized by some θ. In the case of contextual bandits, the update rule of REINFORCE is

$$\theta_{t+1} = \theta_t + \eta(r_t - \bar{r}_t)\nabla \log \pi_{\theta_t}(a_t|x_t),$$

where η denotes the learning rate and $\bar{r}_t$ the average reward achieved so far. In practice, the update rule above is often performed over a batch of interactions with the environment to make training more stable. The subtraction by $\bar{r}_t$ is not necessary but often introduced to reduce variance and make the algorithm more stable (Sutton and Barto 1998).

2.5.2 Thompson sampling

Thompson sampling is probably the oldest bandit algorithm for regret minimization and was introduced in 1933 by William R. Thompson (Thompson 1933). It is a Bayesian approach to bandits that is very simple and intuitive. Given a set of observations so far, H_t, Thompson sampling keeps a posterior distribution over possible bandit models, $p(\mu|H_t)$, acts by sampling one model from the posterior and then plays the arm that is optimal in the sampled model. In Algorithm 4 we show Thompson sampling for a generic multi-armed bandit task.

Thompson sampling is not just limited to the multi-armed bandit but can be applied to contextual bandits (Agrawal and Goyal 2013; Riquelme et al. 2018)

Algorithm 4 Thompson sampling for multi-armed bandit

Require: A set of arms $\mathcal{A}$ and a prior distribution p_0 over bandit models μ.
  Initialize history $H_1 = \{\}$.
  for $t = 1, ...$ **do**
    Sample model from posterior $\hat{\mu} \sim p(\mu|H_t)$.
    Play arm $a_t = \arg\max_a \hat{\mu}_a$.
    Observe reward r_t and update history $H_{t+1} = H_t \bigcup \{(a_t, r_t)\}$.
  end for

and more general reinforcement learning tasks (Strens 2000). It has also been shown to work well in practice (Chapelle and Li 2011). For cases where precise Bayesian inference is not possible, e.g., when the model is a neural network, there are approximate versions of Thompson sampling (Gal and Ghahramani 2016; Riquelme et al. 2018).

2.5.3 Optimism in the face of uncertainty

Optimism in the face of uncertainty (OFUL) is a general approach decision-making under uncertainty that is often applied to bandits (Auer et al. 2002; Abbasi-Yadkori et al. 2011). The core idea is to compute confidence intervals for the expected reward of each arm and then always play the arm with the highest upper confidence bound on the reward. Hence, the learner is always optimistic about the environment and plays the arm with the highest *plausible* expected reward. In Algorithm 5, we show the UCB1 algorithm (Auer et al. 2002) which is used as a baseline in Paper 5. In the algorithm $\hat{\mu}_{a,t}$ denotes the average reward of arm a and $N_t(a)$ the number of times the arm has been played.

Algorithm 5 UCB1

Require: A set of arms $\mathcal{A}$ of size K.
  Play each arm once.
  for $t = K, ...$ **do**
    for each $a \in \mathcal{A}$ **do**
      $I_t(a) := \hat{\mu}_{a,t} + \sqrt{\frac{2 \log t}{N_t(a)}}$.
    end for
    Play arm $a_t = \arg\max_a I_t(a)$.
    Observe reward r_t and update $\hat{\mu}_{a,t}$ and $N_t(a)$.
  end for

2.5.4 Best-arm identification algorithms

In the case of fixed-confidence best-arm identification, a standard design pattern in the literature is to solve the lower bound in Equation 2.4.2, using one's current estimate of the environment, and then track the exploration policy suggested by

the lower bound. The idea is that our estimate of the environment will eventually be close to the true environment, which will result in our exploration policy being close to the optimal one suggested by the lower bound. There are mainly two ways of approaching the optimization problem in Equation 2.4.2. In the *Track-and-Stop* algorithm (Garivier and Kaufmann 2016) the optimization problem is solved at every time step to get a new exploration policy to track. Degenne et al. (2019) proposed an alternative approach and instead view Equation 2.4.2 as a zero-sum game and apply game-strategies to solve the lower bound. This results, in a strategy that never solves the optimization problem to convergence and is thus *computationally* much cheaper. Both these approaches are used in Paper 6.

Chapter 3

Reinforcement learning and efficient communication

In this chapter, we introduce relevant results and concepts from cognitive science and language evolution and discuss how reinforcement learning is connected to these things.

3.1 Why do languages look the way they do?

Why do languages look the way they do? This intriguing question lies at the very heart of linguistics and cognitive science (Zipf 1949; Chomsky 1986; Pinker and Bloom 1990). Surprisingly, there is a large variation between human languages across the globe (Evans and Levinson 2009). For example, some languages completely lack recursive numeral systems (Pica et al. 2004); color naming systems vary both in size and structure between different languages (Berlin and Kay 1969); spatial systems vary between languages both w.r.t. frame of reference (Majid et al. 2004) and in lexicalized concepts (Levinson et al. 2003). Still, there are recurring patterns that are found in many languages (Dryer 1998; Von Fintel and Matthewson 2008).

It is suggested that at least some of these observations can be explained by the interaction between the cognitive constraints of the agents and the properties of the environment in which they communicate (Rosch 1978; Gärdenfors 2014; Gibson, Futrell, Piantadosi, et al. 2019). Especially, it is suggested that languages are shaped by the need to efficiently communicate information (Kemp, Xu, et al. 2018; Gibson, Futrell, Piantadosi, et al. 2019). That is, languages are under pressure to be both informative, to convey the intended meaning as accurately as possible, and simple, to minimize cognitive load.

3.1.1 Efficient semantic categories

In this chapter, we are mostly concerned with the efficiency of semantic categories, i.e., how well a set of words can be used to convey a set of meanings, or concepts. It has been shown that category systems found in human languages support efficient communication across a wide range of domains, e.g., color naming (Regier, Kay,

et al. 2007; Zaslavsky et al. 2018), kinship terms (Kemp and Regier 2012), spatial relations (Khetarpal et al. 2013; Chen et al. 2023), modals (Imel and Steinert-Threlkeld 2022), season naming (Kemp, Gaby, et al. 2019), and numeral systems (Xu, Liu, et al. 2020)[1].

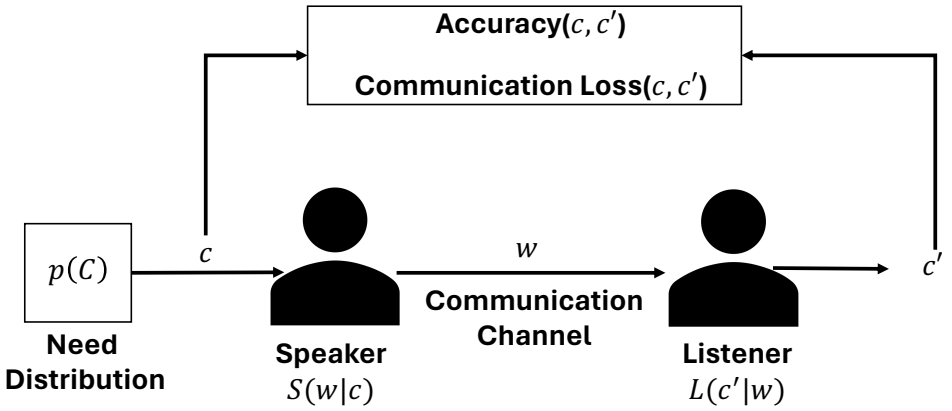


Figure 3.1: The efficiency of semantic categories, or naming systems, is usually studied in a communication setup grounded in Shannon's information theory. A concept is drawn from a need distribution over possible concepts and given to a speaker. The speaker acts as an encoder and encodes the concept into a word. The word is communicated over a, possibly noisy, channel to a listener. The listener then decodes the message into a concept. The informativeness of the speaker is measured in how well the listener's reconstruction matches the original concept in expectation over the need distribution.

These works all ground their notion of efficiency in the classical communication setup of Claude Shannon (Shannon 1948), see Figure 3.1. In this setup, a speaker tries to communicate a certain concept c, from a set of concepts $\mathcal{C}$, to a listener by uttering a certain word w drawn from a set of words $\mathcal{W}$ according to the speaker's distribution $S(w|c)$. Upon hearing the word, the listener decodes the message into a concept using the distribution $L(c|w)$, and the communication accuracy, or loss, is measured based on how well the listener's reconstruction matches the original concept the speaker had in mind. These concepts are assumed to be drawn from a *need distribution*, $p(c)$ that controls how often the speaker has to refer to various concepts. The need distribution is often skewed and puts more emphasis on certain concepts, e.g., in the numeral domain the quantities 1 and 2 are more frequently communicated than the quantity 78 (Xu, Liu, et al. 2020). A language is said to be *efficient*, under a certain need distribution, if it finds a *near-optimal* trade-off between language complexity and expected accuracy. That is, the language is near the *Pareto frontier* between informativeness and complexity, see Figure 3.2.

There are various ways of measuring the complexity and informativeness, or communication loss, of a naming system. One way of measuring the loss of information

[1]Note that some of these works consider the minimization of communication loss, rather than maximization of accuracy/informativeness, given a certain level of complexity. However, these problems are essentially duals of each other.

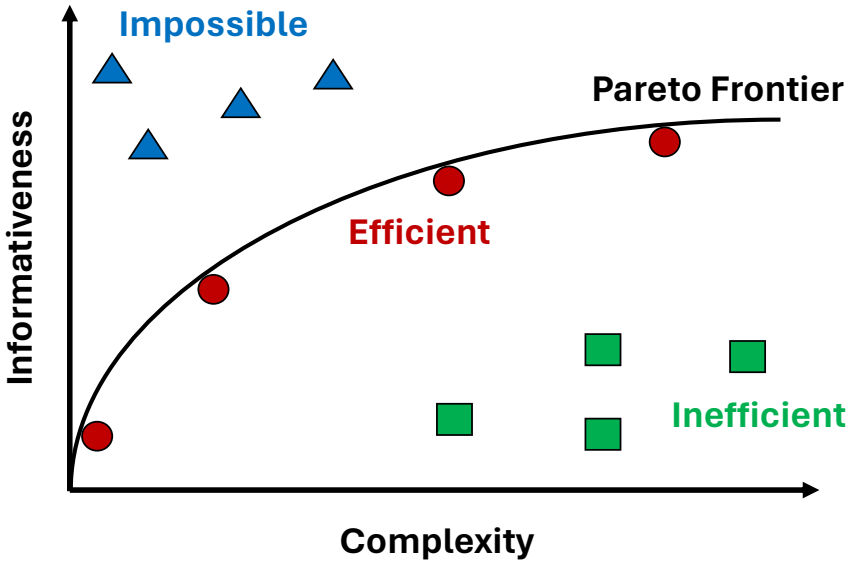


Figure 3.2: Illustration of the trade-off between complexity and accuracy studied in, e.g., Kemp, Xu, et al. (2018) and Zaslavsky et al. (2018). The Pareto frontier corresponds to the languages that achieves highest possible informativeness given a fixed level of complexity. Thus, it is not possible to improve the informativeness of these languages without increasing their complexity as well. As a result, the blue triangles correspond to impossible languages that cannot exist. The green boxes corresponds to highly inefficient languages since the have a high complexity, and induces a high cognitive load on the user, while they do not support accurate communication. It is suggested that human languages find a near-optimal balance between these two forces and populate the region close to the Pareto frontier, like the red circles.

during communication is the expected *surprisal* (Gibson, Futrell, Jara-Ettinger, et al. 2017)

$$\mathrm{E}^S := -\sum_{c,w} p(c)S(w|c)L(c|w).$$

Another approach measures the expected KL-divergence between the speakers uncertainty about the concept, $S(c)$, and the listener distribution (Kemp, Xu, et al. 2018; Xu, Liu, et al. 2020)

$$\mathrm{E}^{\mathbb{KL}} := \sum_{c,w} p(c)S(w|c)\mathbb{KL}(S(c)||L(c|w)).$$

Recall that the KL-divergence is defined as $\mathbb{KL}(S(c)||L(c|w)) = \sum_c S(c) \log \frac{S(c)}{L(c|w)}$. The complexity of a language can for example be measured by number of words used by the speaker (Regier, Kay, et al. 2007) or by the number of rules needed to define the naming system of the speaker (Kemp and Regier 2012; Xu, Liu, et al. 2020).

Another approach for measuring complexity and informativeness is given by Zaslavsky et al. (2018) who recently gave the efficiency hypothesis a firm theoretical foundation by grounding it in the independent Information-Bottleneck (IB) principle (Tishby et al. 1999). In short, the IB framework suggests that the complexity of

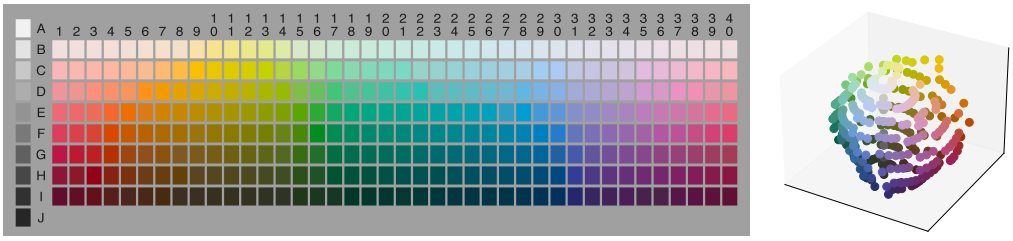


Figure 3.3: (**left**) The Munsell chart used to collect the WCS data. (**right**) Color chips from the Munsell chart represented in 3 dimensional CIELAB space.

a language should be measured by the mutual information between the speaker's mental representation of a concept and words, $I_S(M;W)$. The accuracy is measured as the mutual information between actual concepts and words $I_S(C;W)$ and this can be shown to measure the similarity between the speaker's and listener's mental representations. The framework of Zaslavsky et al. (2018) is further summarized in the Appendix of Paper 4.

In Paper 1 and Paper 2, we use number of words as the complexity measure, and the relevant measures of informativeness are E^S and $\mathrm{E}^{\mathbb{KL}}$. The IB framework of Zaslavsky et al. (2018) is relevant for Paper 3 and Paper 4.

Efficient color naming systems

In Paper 1, Paper 3, and Paper 4 we study how efficient communication emerges in the domain of colors and compare to how human languages partition the color space. These papers rely on the data from the World Color Survey (WCS) (Cook et al. 2005) which contains color naming data from 110 non-industrial languages, with approximately 25 speakers of each language participating in the survey. The speakers were asked to name each of the 330 color chips presented in the Munsell chart in Figure 3.3. The resulting data shows a large variation in color naming between languages, see Figure 3.4, but patterns between languages are also observed (Berlin and Kay 1969). As mentioned earlier, recent work suggests that the languages in the WCS support efficient communication (Regier, Kay, et al. 2007; Zaslavsky et al. 2018).

Efficient numeral systems

Numeral systems vary between languages, both in terms of structure and number of terms, (Hurford 1987; Hammarström 2010; Comrie 2013). Some languages, like Swedish or English, have recursive numeral systems and thus an infinite set of numeral terms generated from a finite set of rules. However, there are languages without any recursive numeral systems, where precise description of a numeral can only be done in an restricted range, referred to as *exact restricted* numeral systems, or where numeral terms only have an approximate meaning, referred to as *approximate* numeral systems. In an exact restricted system, each term refers to a precise interval of the numberline, with one such example being {'one,' 'two', 'three', 'larger than three' },

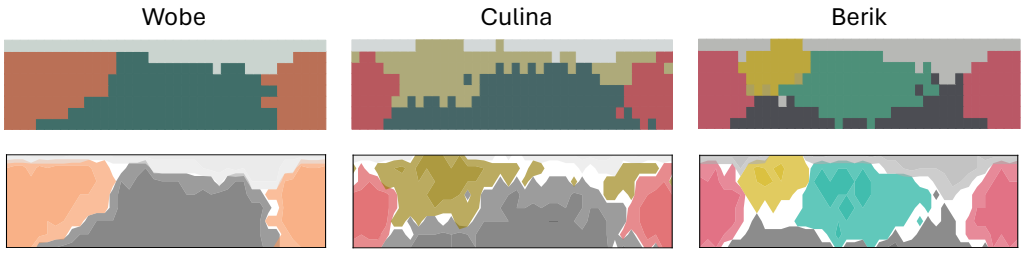


Figure 3.4: Wobe (Ivory Coast), Culina (South America), and Berik (Indonesia) are languages in the WCS data with different numbers of color terms. The top row illustrate the mode map of each language relative to the Munsell chart. That is, each color chip in the Munsell chart is assigned the word most frequently used by speakers of that language and colorized by its average color in the CIELAB space. Thus, each colorized region corresponds to a color term and indicates the region of the color space covered by that particular term. Since speakers of the same languages are inconsistent with each other, the color terms can also be viewed as soft clusters or distributions. This is illustrated in the bottom row where we instead highlight level sets of the color terms. Here, unfaded area indicates the level sets between $0.75 - 1.0$ while the faded area indicates the sets between $0.3 - 0.75$.

while the terms in an approximate system have a fuzzy meaning, e.g., 'a few' or 'many'. Xu, Liu, et al. (2020) recently argued that numeral systems support efficient communication and these results are relevant for Paper 2.

3.2 Simulating language evolution

If we accept the hypothesis that language is, at least partially, shaped by efficiency, a natural question is:

How does language become efficient?

In Paper 1, Paper 2, Paper 3 and Paper 4 we explore this question by simulating language evolution using reinforcement learning.

The idea of simulating language evolution with artificial agents was pioneered by Steels (1995) which sparked interest in studying how language can emerge in artificial systems (e.g. , Shennan (2001), Kirby (2002b), Wagner et al. (2003), Smith and Hurford (2003), Steels and Belpaeme (2005), Griffiths and Kalish (2007), Skyrms (2010), Jäger et al. (2011), and Dale and Lupyan (2012)). Recent developments in deep learning have rekindled this interest in the emergence of language in artificial systems (e.g. , Foerster et al. (2016), Lazaridou, Peysakhovich, et al. (2017), and Havrylov and Titov (2017)) since it is now feasible to conduct more complex experiments, compared to what was previously possible. These recent works often study the emergence of language in a communicative dyad consisting of deep reinforcement learning agents. In these works, agents often start as *tabula rasa* and develop a grounded language solely from maximizing a joint reward, see Section 3.2.1 below for a detailed description.

3.2.1 Reinforcement learning and the signaling game

There is a growing body of work that explore the emergence of communication in collaborative multi-agent reinforcement learning (Jorge et al. 2016; Foerster et al. 2016; Lazaridou, Peysakhovich, et al. 2017; Havrylov and Titov 2017; Mordatch and Abbeel 2018; Chaabouni et al. 2021; Downey et al. 2022; Lian et al. 2023; Thomas, Santos-Rodriguez, et al. 2022; Guo, Hao, et al. 2024). A central concept in this line of work, as well as in this thesis, is the Lewis signaling game (Lewis 1969), which is shown in Algorithm 6 and resembles the communicative setup in Figure 3.1.

Algorithm 6 Lewis signaling game.

for t=1,..., T **do**
Speaker observes $c_t \sim p(c)$ and samples a signal w_t from the policy $S(w|c_t)$.
Listener observes w_t and samples a state c'_t from the policy $L(c'|w_t)$.
Both speaker and listener observes the reward $r(c_t, c'_t)$ and update their policies using some reinforcement learning algorithm.
end for

This game proceeds as follows: The speaker observes a concept c drawn from a set of possible concepts $\mathcal{C}$ according to the probability distribution p. After observing c, the speaker samples a word w from a set of words $\mathcal{W}$ according to its distribution $S(w|c)$. The word is observed by a listener who must infer the concept c based on the word w. This is done by sampling from the distribution $L(c'|w)$. A joint reward, $r(c, c')$, is given to both agents based on how well the listener's reconstruction of the concept, c', matches the original concept c. The core idea is that the agents will start as *tabula rasa*, the words in $\mathcal{W}$ carry no meaning and the agents will converge to a joint language by maximizing the reward. Hence, they develop a language that is grounded in the current environment and the reward function.

Note that the speaker and listener are solving contextual bandit problems. The speaker is solving a contextual bandit task where concept c is the context and the action is uttering a word w. The listener is solving a contextual bandit where the context is the word w and the action is choosing a concept c'. In Paper 1, Paper 3 and Paper 4 we apply the REINFORCE algorithm (Williams 1992) to these contextual bandit problems while we in Paper 2 apply a randomized approach that mimics Thompson sampling (Gal and Ghahramani 2016).

There is also recent work exploring emergent communication using the evolutionary model *replicator dynamics* (Imel, Futrell, et al. 2023; Imel 2023). This model is tightly connected to reinforcement learning, see Börgers and Sarin (1997). In fact, a particular version of the bandit algorithm *follow-the-regularized-leader* (Cesa-Bianchi and Lugosi 2006) is equivalent to a finite-time version of the replicator dynamics (Mertikopoulos and Sandholm 2016; Hennes et al. 2020).

A reader interested in knowing more of about the current state of emergent communication in reinforcement learning might find the following two surveys useful, Lazaridou and Baroni (2020) and Boldt and Mortensen (2024).

3.2.2 Why reinforcement learning?

The fact that the reinforcement agents develop their language from scratch makes the setup described in the earlier section a powerful tool for simulating language evolution and exploring the question of what mechanisms lead to the emergence of *efficient communication*.

We can further motivate the use of reinforcement learning for simulating language evolution by viewing it through the lens of Marr's famous three levels of analysis (Marr 1982), a decomposition that offers both functional and mechanistic views on information processing systems. Marr proposed that any such system can be understood by studying it on three different levels, the *computational*, the *algorithmic*, and the *implementation* level. At the computational level, the goal of the system, or agent, is defined, i.e., what type of computational problem is the agent trying to solve. At the algorithmic level, we ask what algorithm the agent is deploying to solve the computational problem. At the implementation, or hardware, level, the focus is on how the algorithm is realised, or implemented.

Further, as argued by Niv and Langdon (2016), reinforcement learning spans all three of Marr's levels. At the computational level, the problems a reinforcement learning agent usually tries to solve consist of maximizing and/or predicting future rewards. To connect this to the functional view on language offered by Kemp, Xu, et al. (2018) and Gibson, Futrell, Piantadosi, et al. (2019), we note that in a collaborative setting where agents have to coordinate, being informative is often be a prerequisite for reward maximization. The more informative a message is, the better the agents can coordinate, which in the end yields higher rewards for the agents. In this way, we can view informativeness as a sub-goal the agents need to achieve to solve the problem of maximizing rewards. This is in line with the goal-driven paradigm for language learning in neural models explored by e.g., Lazaridou, Peysakhovich, et al. (2017), Havrylov and Titov (2017), and Mordatch and Abbeel (2018).

At Marr's algorithmic level, reinforcement learning offers several algorithmic solutions to the problem of maximizing reward, e.g., policy optimization, temporal-difference learning, Thompson sampling, and optimistic principles. Some of these algorithmic solutions have been used in neuroscience and psychology to model learning in both single-agent tasks (Niv 2009; Ludvig et al. 2011; Tomov et al. 2021) as well as social tasks (Jones et al. 2014). It is also worth mentioning that there are intriguing connections between classical reinforcement learning techniques for handling the exploration-exploitation trade-off, like Thompson sampling, and how humans seem to approach this trade-off (Gershman 2018; Schulz and Gershman 2019). Going back to language evolution and the emergence of efficient communication, we argue that reinforcement learning introduces a natural bias towards simplicity at the algorithmic level. This is because multiple agents need to converge to a joint language by interacting with each other, which results in a bias towards solutions that are easily accessible for their learning algorithms, and simple languages should be easier to learn than complex ones (Kirby, Cornish, et al. 2008; Kirby, Tamariz, et al. 2015; Carr et al. 2020). One could potentially challenge the various notions of complexity in the efficient communication literature and simply ask whether or not

learnability itself serves as a sufficient measure of simplicity (Steinert-Threlkeld and Szymanik 2019; Steinert-Threlkeld and Szymanik 2020).

Furthermore, there are connections between certain neurons in the brain and reward predictions (Schultz et al. 1997; Niv 2009; Dabney et al. 2020) which suggest that reinforcement learning might also be present at the hardware level in the brain. However, we want to highlight that these results from neuroscience, regarding the hardware level, are not relevant to this thesis. The papers summarized later in this chapter all consider agents with simple neural networks, updated using gradient descent, as "hardware", and it is unclear whether this mimics the architecture of the brain in any sensible way.

Hence, in the context of this thesis, reinforcement learning is primarily relevant at Marr's computational and algorithmic levels.

3.2.3 Iterated learning

A very influential model for cultural evolution is *iterated learning* (Kirby 2001; Smith, Kirby, et al. 2003). Iterated learning models how language evolves over generations of agents, see Figure 3.5, and has similarities to the children's game *telephone* where a message is whispered from person to person. In iterated learning, a generation of agents will learn their language from data generated from the previous generation and then generate data that the next generation will learn from[2]. This model has been implemented in the lab, with real humans, to show how various language structures emerge (e.g., Kirby, Cornish, et al. (2008), Smith and Wonnacott (2010), Xu, Dowman, et al. (2013), and Verhoef et al. (2014)), as well as with artificial agents (e.g., Thompson et al. (2016), Carcassi et al. (2021), and Kirby and Tamariz (2022)).

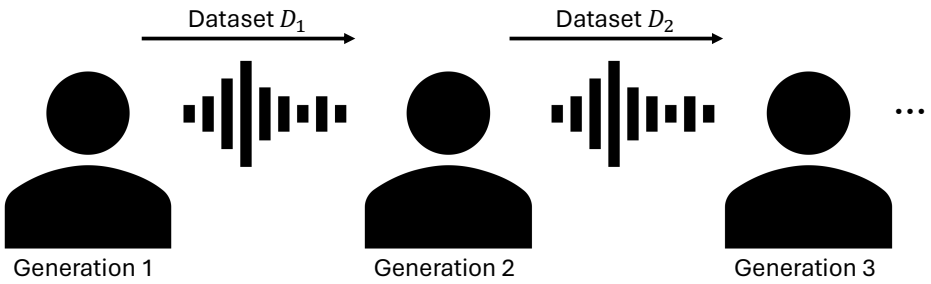


Figure 3.5: In iterated learning, one generation of agents learn their language from a finite dataset generated from the previous generation. This generation then produces a new dataset that is passed to the next generation.

[2]Note that the iterated learning process can be applied to any scenario where one agent learns its behavior from other agents, not just language. However, we are only interested in the application to language evolution in this thesis.

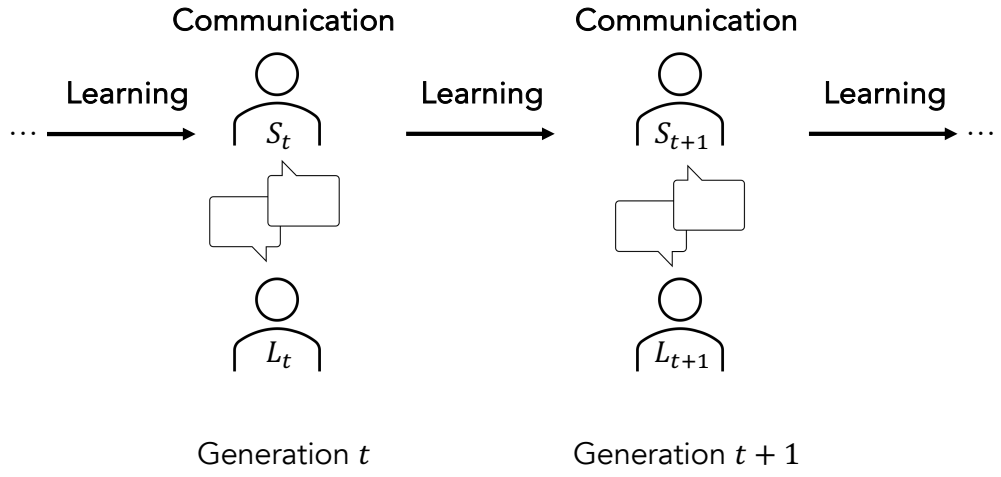


Figure 3.6: Illustration of the NIL algorithm (Ren et al. 2020). The algorithm alternates between communication within a generation, and learning across generations.

The transmission between generations forms a Markov chain, and it has been shown that iterated learning with Bayesian agents, that use the language with the highest posterior probability, converges to a stationary distribution that is an exaggeration of the agents' prior distribution (Griffiths and Kalish 2007). This suggests that cultural evolution, over generations, amplifies learning biases and results in languages that are easy to learn for the agents. This is not hard to imagine, even outside the Bayesian framework, since learning language from a *finite set* of samples creates a *bottleneck* (Zuidema 2002; Kirby 2002a; Kirby, Tamariz, et al. 2015) that restricts what type of languages can emerge and induces a bias towards languages that are simple and easy to learn from a small set of samples. This simplicity bias has been observed in iterated learning experiments with humans (Kirby, Cornish, et al. 2008) and a possible explanation is that learners apply Occam's razor and, given several possible languages that fits the data, choose the simplest one. To connect to Marr's levels of analysis, iterated learning tends to amplify the biases in the algorithmic level of the agent, i.e., the biases in the specific learning algorithm used by the agent.

The fact that iterated learning has a clear bias towards simplicity suggests that it plays a part in the emergence of efficient communication. Interestingly, Carstensen et al. (2015) showed, in a series of human simulations, that iterated learning not only leads to simpler systems but also gravitates towards more informative ones. One way these findings can be interpreted is that iterated learning provides a bias towards both simplicity and informativeness and thus provides an account for the emergence of efficient communication. This is also in line with previous findings that language learners are biased towards efficient languages (Fedzechkina et al. 2012). However, as noted by Carr et al. (2020), these results are in contrast with other works which suggest that (iterated) learners have a bias towards simple and

uninformative languages and that an informativeness bias only arises in the presence of a communicative task (Kirby, Tamariz, et al. 2015; Motamedi et al. 2019; Kirby and Tamariz 2022). See also Rafferty et al. (2011) for evidence that learnability does not fully account for the presence of linguistic universals.

The argument that learning needs to be coupled with communicative tasks for efficient communication to arise suggests that one could combine iterated learning with goal-driven learning approaches, such as reinforcement learning, to simulate language evolution. Such a model has been proposed by Kirby, Tamariz, et al. (2015) and recently explored in the context of deep learning by Ren et al. (2020) who introduced the *neural iterated learning* (NIL) algorithm, see Figure 3.6. Ren et al. (2020) showed that this algorithm leads to the emergence of compositional language in deep learning models (see also Guo, Ren, et al. (2020)). The NIL model alternates between cultural evolution over generations of artificial agents, using iterated learning, and intra-generational communication using reinforcement learning. This type of model is interesting since it models language evolution on two different time scales, the slow cultural evolution over generations and the fast, goal-driven, learning within a generation, as well as having very clear biases at every stage of the model. In Paper 4, we use NIL to argue that iterated learning and communication together account for *efficient* and *human-like* color naming systems, see the summary in Section 4.4.

Chapter 4

Summary of included papers

This chapter provides brief summaries of the appended papers.

4.1 Paper 1: A reinforcement-learning approach to efficient communication

In Paper 1 we present a multi-agent computational approach to partitioning semantic spaces using reinforcement learning. Two agents communicate about colors in a noisy environment using a finite vocabulary, see Figure 4.1. Our two-agent paradigm closely mirrors the information-theoretic frameworks of Regier, Kemp, et al. (2015) and Gibson, Futrell, Jara-Ettinger, et al. (2017) and our main contribution is the insight that an, independently motivated, computational learning mechanism accounts for the emergence of efficient color naming systems.

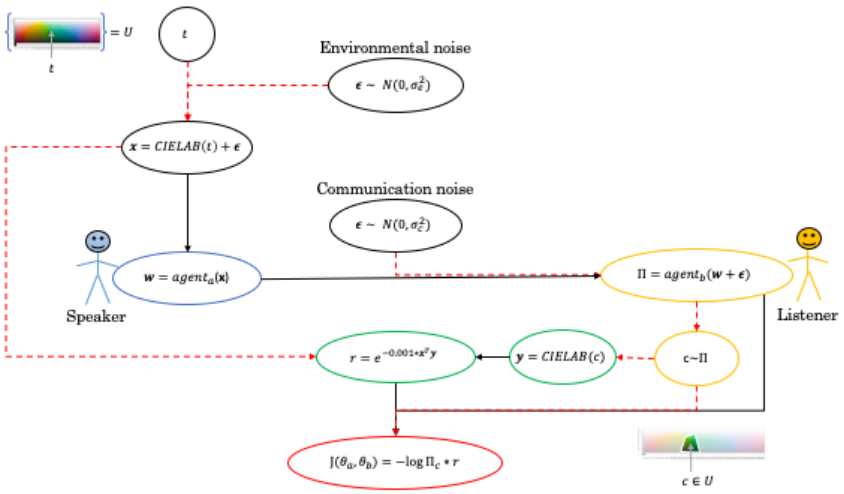


Figure 4.1: The communication setup considered in Kågebäck et al. (2020).

In our model, a speaker observes a color, represented in CIELAB space, and has to communicate this color to a listener. A joint reward, that measures the similarity between the color the speaker intended to communicate and the listener's reconstruction, is given to both agents. The agents are implemented as neural networks with one hidden layer and are updated using REINFORCE over a sequence of rounds of the signaling game. We consider two different versions of this game: one variant where the communication channel between agents is continuous, and thus differentiable, and where the presence of channel noise makes the agents gravitate towards discrete communication, as well as a variant where the communication channel is discrete and non-differentiable. In the continuous setting, we only compute the listener's loss and backpropagate this information through the communication channel to the speaker, while in the discrete setting, we update both the speaker and listener separately.

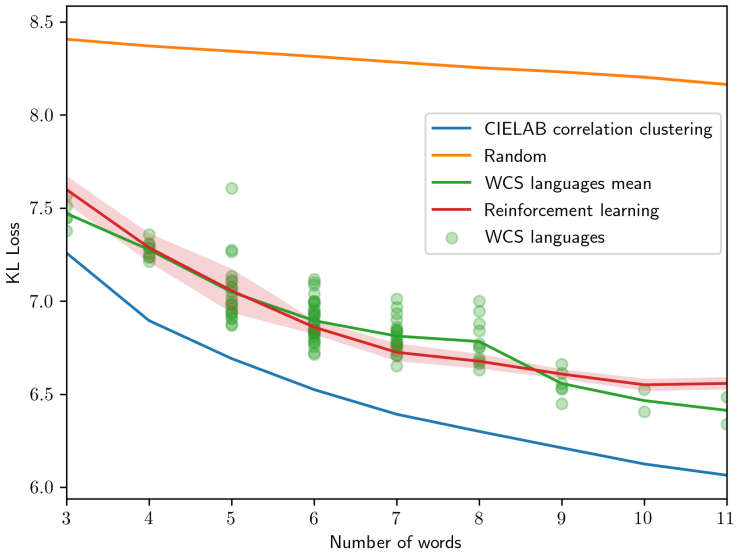


Figure 4.2: Trade-off between communication loss and vocabulary size. The Pareto frontier is estimated using correlation clustering in CIELAB space. We observe that our agents (the line corresponding to reinforcement learning) are able to develop a color naming system, from just maximizing reward, that matches the efficiency of human color naming systems (the line corresponding to WCS). The Pareto frontier is estimated using correlation clustering. Note, the WCS language data points is a reproduction from Regier, Kemp, et al. (2015). The error bars around the red line corresponds to a 95% confidence interval.

In Figure 4.2 we show the efficiency, measured as expected communication loss vs vocabulary size, of our artificial agents, human systems in the WCS data, and random agents. The communication loss is measured as the KL-divergence between the speaker and listener, as by Regier, Kemp, et al. (2015). We observe that reinforcement learning can replicate the efficiency of human color naming systems solely by maximizing reward. We also observe that both the artificial agents and human systems are close to the Pareto frontier and much more efficient compared to

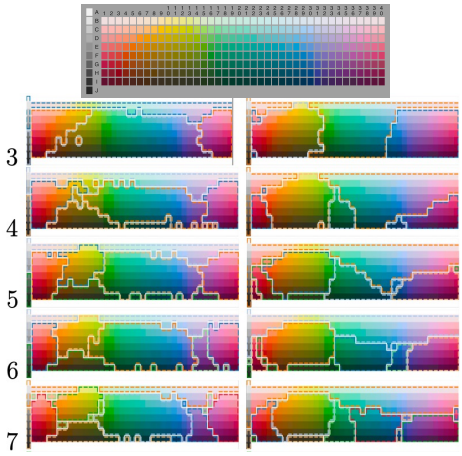


Figure 4.3: The top grid is the Munsell chart used to collect the WCS data. The left column corresponds to human languages with different number of color words while the right column corresponds to artificial naming systems produced by our reinforcement learning agents. Each colored line in a grid corresponds to a color word in the language and the region encapsulated by a word corresponds to the colors for which this word is used.

a random baseline.

Some of the color maps produced by reinforcement learning are presented in Figure 4.3 along with human color maps derived from the WCS data. We observe that reinforcement learning produces color maps that have a fair amount of similarity to human ones, without ever being exposed to human systems. This result is further examined in the paper using quantitative approaches.

Beyond the aforementioned results showing that reinforcement learning leads to efficient color naming systems with some similarities to human systems, we also explore how the amount of noise in the environment affects the resulting color language. Our results indicate that there is a strong negative correlation between environmental noise and the resulting complexity of the produced color naming system. This can potentially be explained by the fact that there is an implicit pressure towards simple solutions in our reinforcement learning model. The higher the noise is, the harder it is for the agents to learn a joint language, and they are thus more likely to converge to simple solutions where very few color words are used.

4.2 Paper 2: Learning approximate and exact numeral systems via reinforcement learning

Xu, Liu, et al. (2020) recently suggested that numeral systems found in human languages are optimized for efficient communication. In Paper 2 we study how efficient approximate and exact numeral systems emerge in a signaling game played by two reinforcement learning agents. Our main contribution is showing that reinforcement learning leads to efficient numeral systems that are similar to those found in human language. A motivation for using reinforcement learning in the context of numeral systems is the work of O'Shaughnessy et al. (2021) which highlights the influence that social and economic factors have on the emergent numeral system.

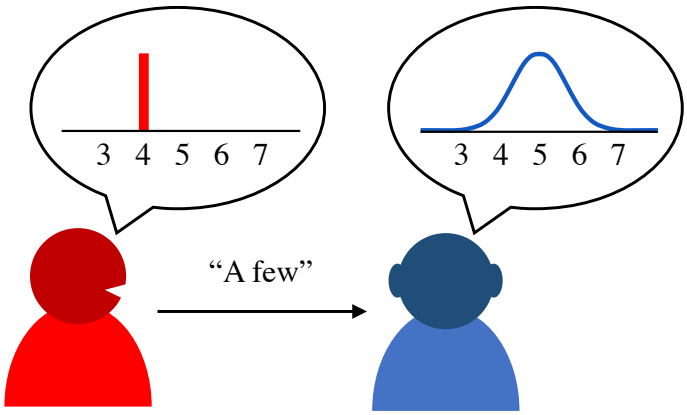


Figure 4.4: The communication model considered here and also by Xu, Liu, et al. (2020). The sender wants to convey the numeral concept 4 and utters "a few". The listener is unsure of which numeral the sender is referring to and produces a probability distribution over possible numerals.

In contrast to Kågebäck et al. (2020), we instead consider a bandit approach with an implicit Thompson sampling scheme (Gal and Ghahramani 2016). Each agent keeps a neural network that models the expected reward for each number-word pair (n, w). At each round of the game, the agents sample a smaller network from the larger one using dropout (Srivastava et al. 2014). This smaller network is later used during the next round of the signaling game. Gal and Ghahramani (2016) showed that this scheme can be viewed as approximate Bayesian inference and we can thus think of the larger networks as belief distributions that we sample from using dropout. Figure 4.5 offers a schematic view of our signaling game with this approach.

In this work, we consider three need distributions inferred from human data and three different reward functions

$$
\begin{aligned}
r_{\text{linear}}(n, \hat{n}) &= 1 - \frac{|n - \hat{n}|}{|\mathcal{N}|}, \\
r_{\text{inverse}}(n, \hat{n}) &= (1 + |n - \hat{n}|)^{-1}, \\
r_{\text{exp}}(n, \hat{n}) &= e^{-|n-\hat{n}|}.
\end{aligned}
$$

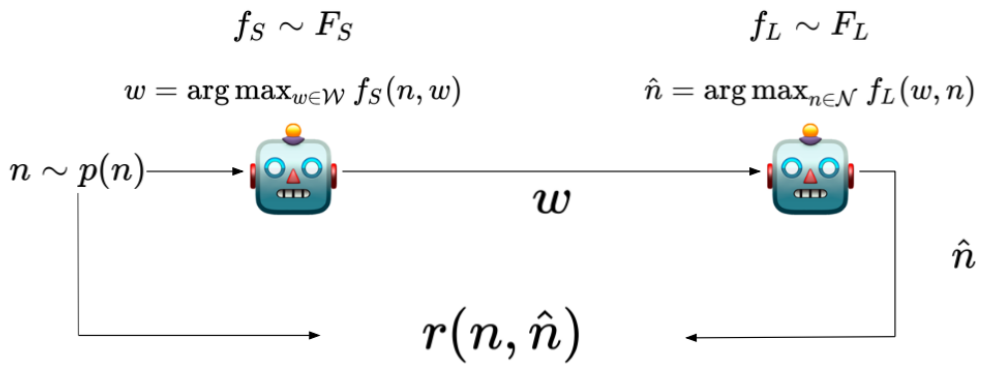
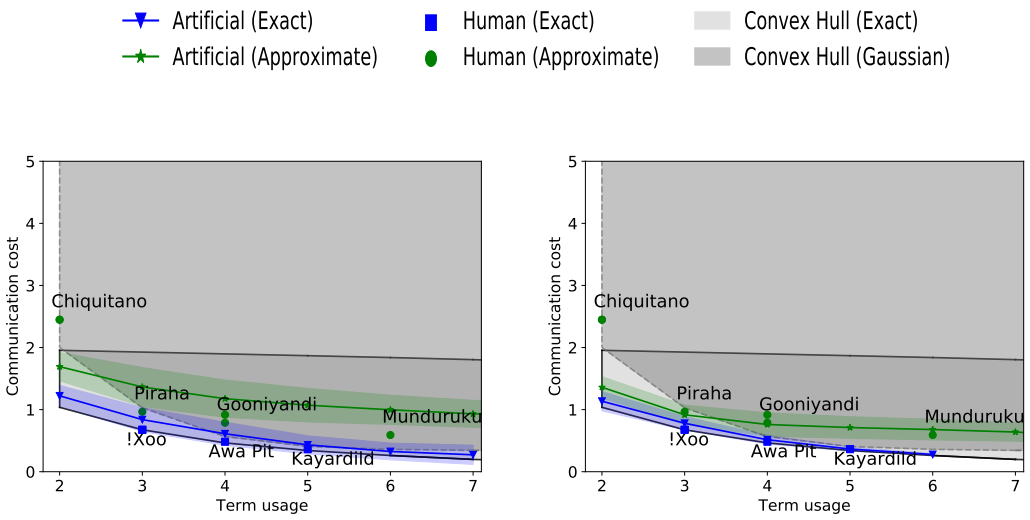


Figure 4.5: At each round of the game, the agents sample smaller networks, f_S and f_L, and using dropout, i.e. some neurons in the larger networks are ignored with a certain probability. This can be viewed as sampling from a belief distribution (Gal and Ghahramani 2016). After this , The speaker is given a number n, drawn form a need distribution n, and conveys the word with the highest expected reward according to f_s. The listener proceeds in similar fashion, given w it produces the guess, $\hat{n}$, that has the highest expected reward according to f_L. A shared reward is given to both agent based on how close $\hat{n}$ is to n, Th networks are updated by minimizing the MSE between predicted reward and observed reward.

We do not suggest that humans explicitly optimize any of these reward functions, the reward functions should merely be thought of as a way to model different amounts of pressure toward informativeness. That is, the quicker the reward decays in terms of $|n - n'|$, the more precise must the listener's reconstruction be to achieve high reward. This results in a higher bias towards informativeness.

After training the reinforcement learning agents, we estimated whether their produced numeral system was exact or approximate by estimating the speaker's distribution over 1000 rounds of the signaling game. If the speaker, for each n, assigned more than 0.90 probability mass to a single word w, we interpreted that as being an exact numeral system, otherwise, we took it to be approximate. Figure 4.6 shows the efficiency of these agents under one of the need distributions considered. Here, both the convex hulls and efficiency were computed as in Xu, Liu, et al. (2020). Further, Xu, Liu, et al. (2020) modeled the human approximate systems as Gaussians while our agents are not restricted to this assumption. This explains why they are below the Pareto frontier for 2-term approximate systems. We observe that the reinforcement learning agents have numeral systems close to the Pareto frontier and populate the same part of the region as the human systems studied by Xu, Liu, et al. (2020). We further observed that these systems are similar to their human counterparts, see Figure 4.7.

An important question that is left open in our work is how these approximate and exact systems evolve into (efficient) recursive numeral systems, like the ones in English or Swedish. Answering this question would probably require a combination of neuro-symbolic methods and reinforcement learning.



(a) Reward: Linear

(b) Reward: Inverse

Figure 4.6: Term usage vs communication cost. This plot shows the result when numbers are drawn according to the need distribution derived by Xu, Liu, et al. (2020). Note that our agents are not restricted to model the words as Gaussian distributions and can create other probability distributions. This explains why the line goes below the convex hull, for 2 terms, which was computed assuming Gaussian distributions for tractability reasons. Our results for human systems matches the ones originally reported by Xu, Liu, et al. (2020).

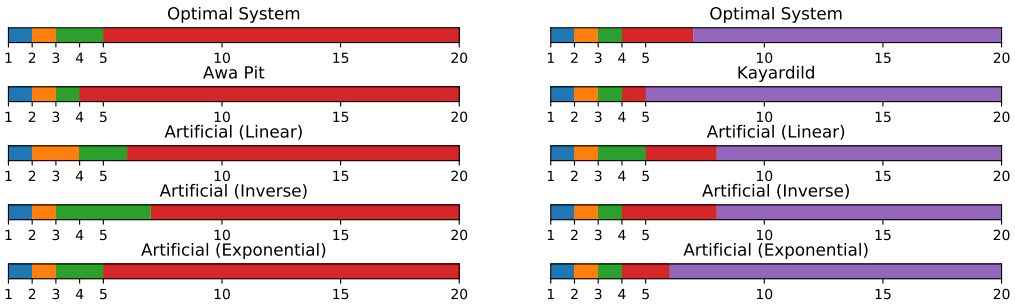


Figure 4.7: Comparison between the optimal numeral systems w.r.t. communication cost, human systems and the artificial systems produced by our agents. Each color represents a numeral word and the corresponding interval on the number line that the word represents.

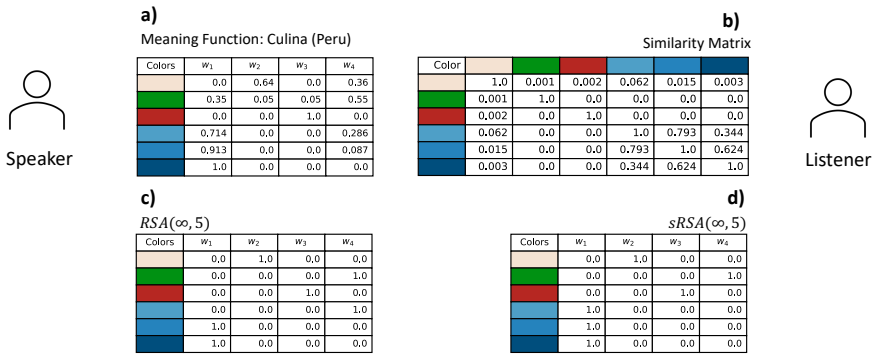
4.3 Paper 3: Pragmatic reasoning in structured signaling games

In Paper 3 we extend our two-agent framework to include agents able to do pragmatic reasoning (Grice 1975). Here, both the speaker and listener observe a set of meanings, also known as a context, and the speaker chooses one of these meanings as the target to communicate to the listener. The language of the agents does not need to be precise in scenarios where the contextual information helps the listener to decode the utterance from the speaker. We introduce the notion of a structured signaling game,

where there is a similarity measure between meanings, and explore how efficient communication emerges between pragmatic agents in this game in the domain of colors. We also introduce a version of the Rational Speech Act (RSA) (Frank and Goodman 2012), tailored for our structured signaling game, that we call structured-RSA (sRSA). In RSA the speaker and listener reason about each others behavior using the following recursion

$$\begin{aligned} L_0(m|w) &\propto \mathcal{L}(m,w) \\ S_t(w|m,C) &\propto e^{\alpha U_t(m,w,C)} \\ L_t(m|w,C) &\propto S_t(w|m,C)p(m|C) \end{aligned}$$

where $U_t(w,m,C)$ is the expected utility, of conveying message w given the meaning m in the context C, and $p(m|C)$ is the prior probability of m given C. Here, $\mathcal{L}(m,w) \in [0,1]$ is a meaning function, or semantic representation, that states to what extent the meaning m can be described by the utterance w. We can think if this function as the lexicon of the agents. In our sRSA, the utility function is defined



Colors	w_1	w_2	w_3	w_4
	0.0	0.64	0.0	0.36
	0.35	0.05	0.05	0.55
	0.0	0.0	1.0	0.0
	0.714	0.0	0.0	0.286
	0.913	0.0	0.0	0.087
	1.0	0.0	0.0	0.0

Color						
	1.0	0.001	0.002	0.062	0.015	0.003
	0.001	1.0	0.0	0.0	0.0	0.0
	0.002	0.0	1.0	0.0	0.0	0.0
	0.062	0.0	0.0	1.0	0.793	0.344
	0.015	0.0	0.0	0.793	1.0	0.624
	0.003	0.0	0.0	0.344	0.624	1.0

Colors	w_1	w_2	w_3	w_4
	0.0	1.0	0.0	0.0
	0.0	0.0	0.0	1.0
	0.0	0.0	1.0	0.0
	0.0	0.0	0.0	1.0
	1.0	0.0	0.0	0.0
	1.0	0.0	0.0	0.0

Colors	w_1	w_2	w_3	w_4
	0.0	1.0	0.0	0.0
	0.0	0.0	0.0	1.0
	0.0	0.0	1.0	0.0
	1.0	0.0	0.0	0.0
	1.0	0.0	0.0	0.0
	1.0	0.0	0.0	0.0

Figure 4.8: An example of a structured signaling game in the color domain. **a)** Shows the meaning function of the agents derived from the language Culina found in the WCS. **b)** The similarity matrix between the colors. **c)** The limit point of RSA as $t \to \infty$ **d)** The limit point of sRSA, as $t \to \infty$. Since RSA minimizes only the surprisal of the listener and does not account for the similarity structure we observe that the lighter blue color and green color are mapped to the same word. Unlike RSA, the sRSA takes the similarity matrix into account and converges to a solution where the first 3 colors can be uniquely determined, while the last 3, all variants of blue, are mapped to the same word.

as the *similarity-sensitive surprisal* (Leinster 2021) of the listener, L,

$$U_t(w,m,C) = -\log \sum_{m'} Z_{mm'} L_{t-1}(m'|w,C)$$

where $Z_{mm'}$ is a similarity measure between the target meaning and some other meaning m' in the context. This measure captures the desirable property that a listener shouldn't be as surprised if a speaker uses the same word for two similar meanings compared to if the speaker used the same word for two very different meanings. Recall that the standard RSA uses the classical surprisal $U_t(w,m,C) =$

$\log L_{t-1}(m|w, C)$ which doesn't explicitly account for the structure in the context. Figure 4.8 shows how RSA and sRSA produces different behavior in the case of colors.

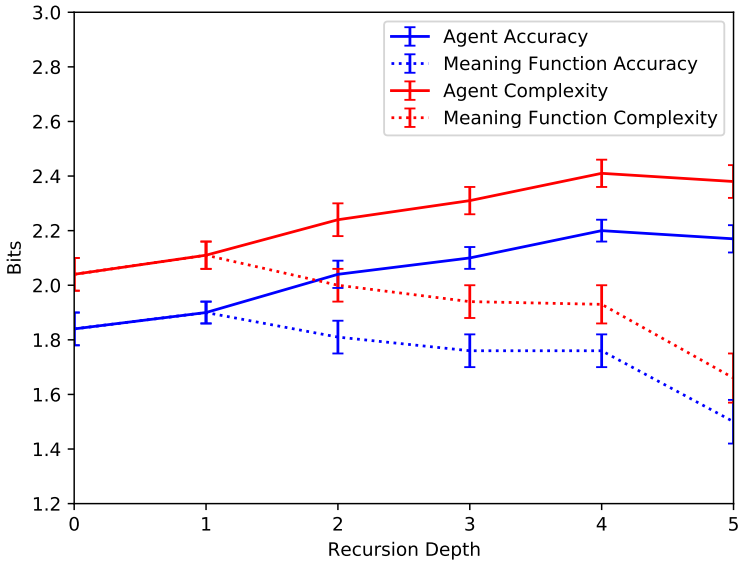
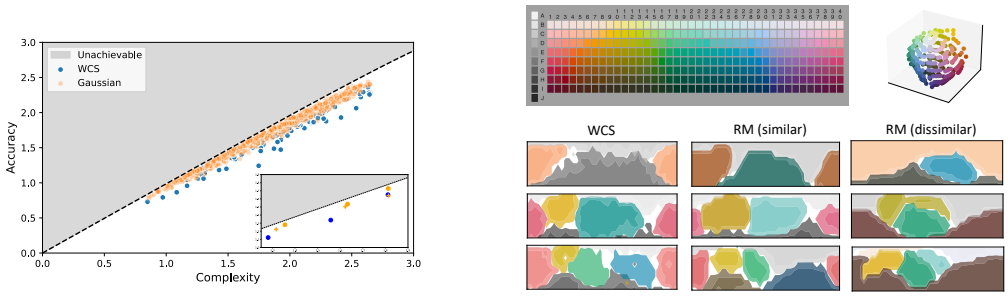


Figure 4.9: The complexity and accuracy of the sRSA agents increase with recursion depth, while the complexity and accuracy of the corresponding meaning functions decrease. Hence, as the reasoning depth increases, the ambiguity of the learned meaning function increases. Depth indicates the level of the final listener in the recursion, and the error bars correspond to the width of the 95% confidence interval.

In the paper, we show that pragmatic agents with semantic representations derived from the WCS data attain efficiency close to the information-theoretic limit after only 1 or 2 levels of recursion. We also show that reinforcement learning agents equipped with sRSA develop highly efficient representations. Especially, our results indicate that as the reasoning power of the agents increases i.e., the number of recursions in sRSA increases, the emergent semantic representation becomes more ambiguous, see Figure 4.9. Hence, our pragmatic agents seem to obey principles of least effort (Zipf 1949). If the agents can perform deep and contextual reasoning there is no need to develop a very precise lexicon. On the other hand, if the agents cannot reason about how the context influences the meaning of an utterance, the resulting lexicon has the be very precise to support efficient communication. These results suggest that there might be an additional trade-off, than the one between informativeness and complexity, between different notions of complexity. Namely, a trade-off between semantic complexity (the complexity of the meaning function) and reasoning complexity (recursion depth) which might be interesting to explore in future work.

4.4 Paper 4: Cultural evolution via iterated learning and communication explains efficient color naming systems

In Paper 4 we consider efficiency using the Information Bottleneck (IB) principle (Tishby et al. 1999; Zaslavsky et al. 2018), and a model of cultural evolution that combines iterated learning and communication (Kirby, Tamariz, et al. 2015). We show that this model converges to color naming systems that are efficient in the IB sense and similar to human systems. We show that some other proposals, such as iterated learning alone, communication alone (like the model in Paper 1), or the greater learnability of convex categories, do not yield the same outcome as clearly. We also highlight the importance of an evolutionary process that leads to *human-like* and efficient systems, since there exists a large set of color naming systems that are highly efficient in the IB sense but not similar to any human systems, see Figure 4.10.



(a) (b)

Figure 4.10: **a)** Efficiency of color naming, following Zaslavsky et al., 2018. The color naming systems of the WCS are shown in blue, replicating the findings of Zaslavsky et al., 2018. We introduce a simple Gaussian random model, shown in orange, that generates highly efficient color naming systems. It can be seen that the RM systems are often closer to the IB curve than the WCS systems are. The inset shows the 9 color systems in **b)**, with the dissimilar random systems shown as + . **b)** The left column contains color naming systems from 3 languages in the WCS. Colored regions indicate category extensions, and the color code used for each category is the mean of that category in CIELAB color space. The named color categories are distributions, and for each category we highlight the level sets between $0.75 - 1.0$ (unfaded area) and $0.3 - 0.75$ (faded area). The middle and right columns contain randomly-generated systems of complexity comparable to that of the WCS system in the same row. The middle column shows random systems that are similar to the WCS system in the same row while the right column shows random systems that are dissimilar to any WCS system.

Our evolutionary model is based on the NIL algorithm (Ren et al. 2020) which alternates between a communicative phase, where agents within a generation interact with each other, and a learning phase, where a new generation learns from the previous generation. Here the learning phase is done by training, using supervised learning, the new generation on data generated from the previous generation. The communication is the same signaling game as Kågebäck et al. (2020) and the agents

are updated using reinforcement learning. For more details about the algorithm and various hyperparameters, see the full paper.

In Figure 4.11 we show the efficiency of the color naming systems that emerge during learning and communication (IL+C), as well as the efficiency of the systems that emerge under learning only (L) and communication[1] only (C). We observe that IL+C produces efficient systems that all end up in the same region as the WCS, even though the agents could in principle produce more complex systems. We also observe that just learning is skewed towards less complex systems than observed in human languages, which is in line with the claims of Carr et al. (2020) that iterated learning induces a bias towards simplicity. On the other end, we see that just communication results in naming systems more complex than what is observed in human systems. To conclude, iterated learning with intra generation communication provides a balance between these forces that corresponds well with what is observed in human color naming systems.

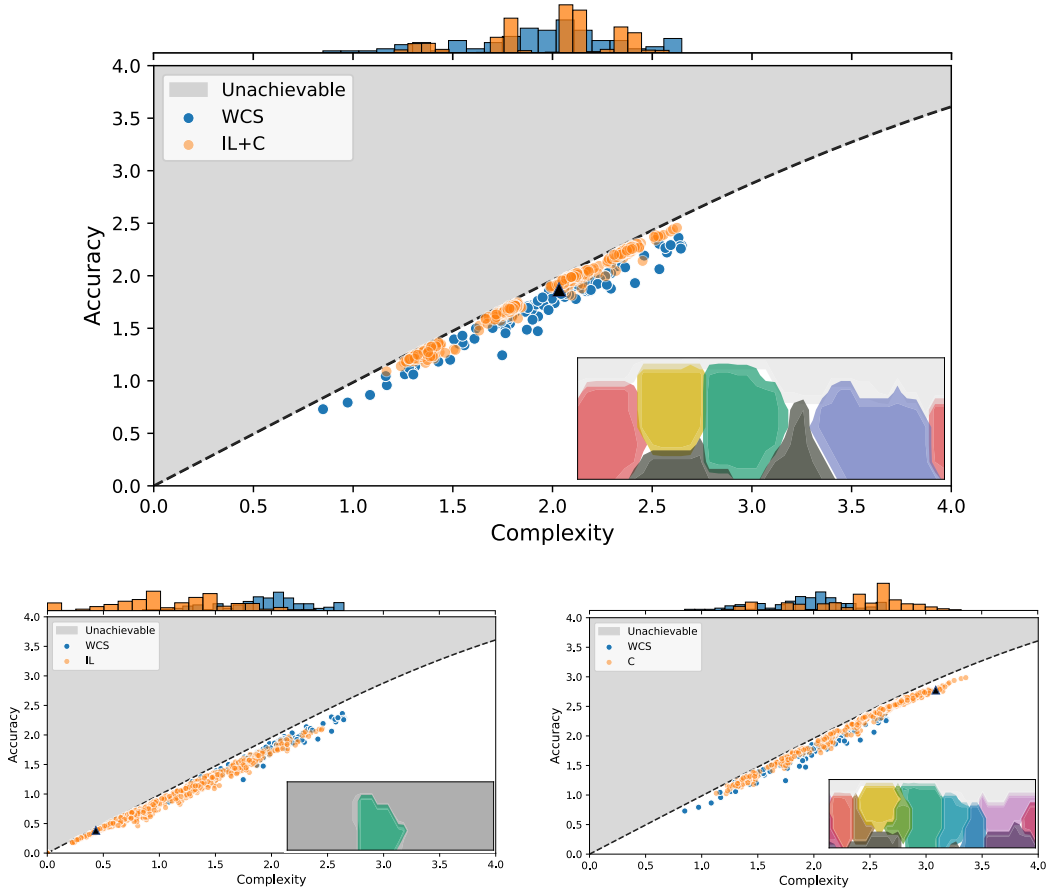


Figure 4.11: Efficiency of the (top) IL+C, (bottom left) IL, and (bottom right) C evolved color naming systems (orange dots), in each case compared with the natural systems of the WCS (blue dots). The black triangle indicates the end state of one run, shown in the inset color map. The histograms above each figure indicate the proportion of systems at the corresponding complexity level.

[1]Note that this is exactly the model in Paper 1, evaluated in the IB framework.

However, as highlighted in Figure 4.10, efficiency does not equal human-like systems. In the paper, we both qualitatively and quantitatively show that IL+C leads to both human-like and efficient systems. For example, Figure 4.12 shows an experiment where we initialized the first generation with a color naming language, generated by our random model, that was efficient but dissimilar to any human systems. We observe that IL+C transforms already efficient systems to become more similar to human systems. In the paper, we further explore what types of

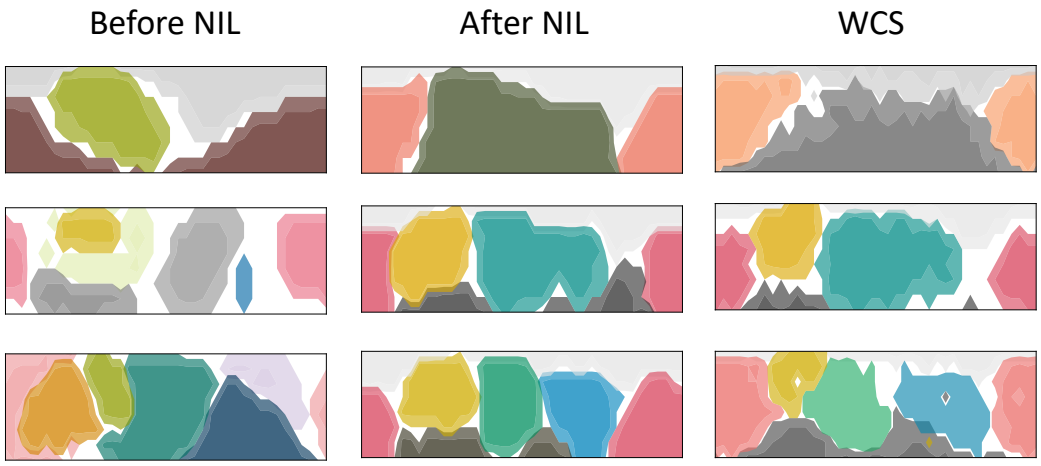


Figure 4.12: IL+C transforms efficient color naming systems to become more similar to the WCS. In each row, the left column shows a randomly generated efficient system that was used to initialize the first generation, the middle column shows the result of running NIL from that initialization state, and the right column shows a WCS system.

systems are produced by the model and connect our results to ideas regarding learnability (Steinert-Threlkeld and Szymanik 2020) and convexity of semantic categories (Gärdenfors 2000).

4.5 Paper 5: Thompson sampling in bandits with clustered arms

In Paper 5, we study a version of the multi-armed bandit problems where the learner has been given a pre-defined clustering of the arms. This could either be a disjoint clustering or a hierarchical clustering of the arms. One motivating example for this model is recommender systems where a user may have strong preferences for certain categories. Our main contribution is proposing a multi-level Thompson sampling algorithm (TSC) for the stochastic multi-armed bandit with clustered arms (MABC), see Algorithm 7, and for a contextual version of the problem, where the expected reward of each arm is linear in the context vector.

Algorithm 7 TSC

Require: $\mathcal{A}$, $\mathcal{K}$
  Set $S_1 = F_1 = 1$ for all a and C.
  for $t = 1..., T$ **do**
    For each cluster C sample $\theta_C \sim \text{Beta}(S_t(C), F_t(C))$ and pick $C_t = \arg\max_{C \in \mathcal{K}} \theta_C$
    For each $a \in C_t$ sample $\theta_a \sim \text{Beta}(S_t(a), F_t(a))$.
    Play arm $a_t = \arg\max_{a \in C_t} \theta_a$ and collect reward r_t.
    Update $S_{t+1}(a_t) = S_t(a_t) + r_t$, $F_{t+1}(a_t) = F_t(a_t) + (1 - r_t)$.
    Update $S_{t+1}(C_t) = S_t(C_t) + r_t$ and $F_{t+1}(C_t) = F_t(C_t) + (1 - r_t)$.
  end for

For the MABC, we provide a regret bound for our algorithm under the assumption that the clusters are well-separated in terms of reward. We show an instance-dependent regret bound, that scales with the gap between sub-optimal clusters and the cluster containing the optimal arm, as well as the gaps between arms in the optimal cluster, informally stated below

$$\mathbb{E}[\text{Regret}_T] \leqslant \left(\sum_{C \neq C^*} \frac{\Delta_C}{\mathbb{KL}(\overline{\mu}_C || \underline{\mu}_{C^*})} + \sum_{a \in C^*} \frac{\Delta_a}{\mathbb{KL}(\mu_a || \mu^*)} \right) \log T + o(\log T).$$

Here, $\overline{\mu}_C$ is the largest achievable expected reward in cluster C, C^* denotes the cluster containing the optimal arm, $\underline{\mu}_{C^*}$ the smallest expected reward for any arm in the optimal cluster, and μ^* the optimal reward. Δ_a is the regret suffered by playing arm a and Δ_C is the regret suffered from playing the arm with the highest reward in cluster C.

We do also prove an instance-independent regret bound on the form

$$\tilde{O}\left(\sqrt{A^* + K(1+\gamma)T}\right) \tag{4.5.1}$$

where A^* is the number of arms in the same cluster as the optimal arm, K is the number of clusters, and γ a parameter that measures the quality of the clustering (lower is better), see the paper for more details. Here $\tilde{O}(.)$ hides logarithmic factors. Recall that standard bandit algorithms have a regret scaling as $\tilde{O}\left(\sqrt{NT}\right)$ where

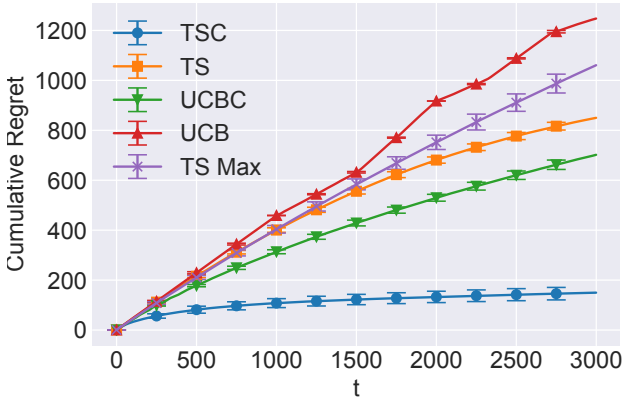


Figure 4.13: An instance with 1000 arms, 32 clusters and 32 arms in each cluster. TSC is our approach. TS (Thompson sampling) and UCB (upper confidence bounds) are algorithms suited for the standard multi-armed bandit. UCBC (Pandey et al. 2007; Bouneffouf et al. 2019) and TSMax (Zhao et al. 2019) are previously suggested algorithms for the MABC. We observe that TSC outperforms all algorithms. The cumulative regret is averaged over 50 random seeds and the error bars corresponds to $\pm$ the standard deviation.

N is the total number of arms. Thus, our bounds suggest that our TSC algorithm should improve over classical approaches when either there are few clusters, small K, or when the optimal arm belongs to a cluster containing few arms (small A^*). Since A^* is not *a priori* known, the bound in (4.5.1) suggests that our algorithm reaps the most benefit over standard approaches when $K = \sqrt{N}$ and each cluster contains $\sqrt{N}$ arms. In addition, our empirical evaluation shows that our approach has an advantage over both classical approaches and other algorithms introduced for the MABC, see Figure 4.13. or more empirical results see the paper. In the paper we also provide regret bounds for hierarchical clusterings as well as an extensive empirical evaluation of the contextual version of TSC.

4.6 Paper 6: Pure exploration in bandits with linear constraints

The best-arm identification (BAI) in the bandit framework has many applications such as hyper-parameter tuning (Li, Jamieson, et al. 2017) and clinical trials (Aziz et al. 2021). However, in practice, many decision-making problems involve constraints on the available actions that need to be satisfied. For clinical trials, this could be certain safety constraints w.r.t. toxicity or in a recommender system one might have constraints that require a certain level of diversity in the recommendations. As a result, standard BAI algorithms are not perfectly appropriate for these settings since the constraints might force the learner to output a stochastic policy instead of one best arm, see the example in Figure 4.14.

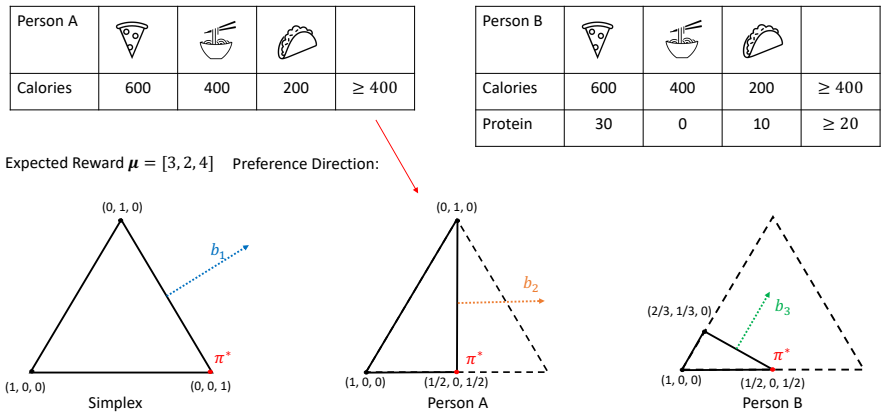


Figure 4.14: Two people, A and B, are searching for a meal plan π that maximizes taste, i.e., expected reward $\mu^\top \pi$, while satisfying some nutrition constraints. Without any constraints this setting reduces to BAI and can be viewed as searching for the optimal policy over the probability simplex. However, the nutrition constraints alter the set of feasible sets and a person might have to mix between several dishes to satisfy the constraints while maximizing the reward. The red arrow indicates the preference direction and the red dot corresponds to the optimal policy for each case. The dotted arrows, $\boldsymbol{b}_i$, corresponds to the normal of that boundary, i.e. the constraint causing the boundary, and as we will see later, in Figure 4.15, the distance between μ and $\boldsymbol{b}_i$ controls the hardness of the problem. For person A, the distance between $\boldsymbol{b}_2$ and μ decreases compared to the unconstrained case, while it increases for person B. Thus, the problem of finding the optimal pure exploration policy gets easier for person B while harder for person A. This is quantified by the minimum number of samples required to identify the optimal policies for person A, B, and the unconstrained case, see Figure 4.15.

In Paper 6 we study the problem of finding the best option when arms are subject to a set of linear constraints. We consider this problem in the *fixed confidence regime* where the goal is, with as few collected samples as possible, to output the optimal

solution π^* to the following problem

$$\arg\max_{\pi \in \mathcal{F}} \pi^\top \mu \tag{4.6.1}$$

with probability at least $1-\delta$ for some given $\delta \in (0,1)$. Here, $\mu \in \mathbb{R}^K$ is the *unknown* reward vector where an entry μ_i corresponds to the expected reward of arm $i \in [K]$ and $\mathcal{F}$ is the set of policies that satisfy our constraints. Thus, our goal is to query entries of μ until we can output the optimal solution to (4.6.1) with probability $1-\delta$. We further assume that the noise in the observations follows some sub-Gaussian distribution.

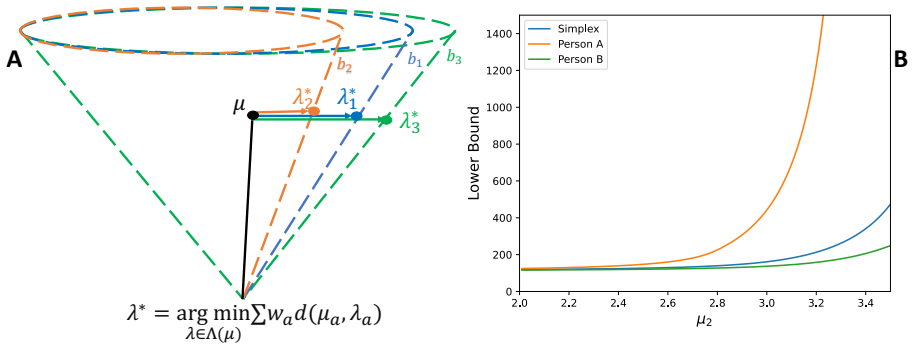


Figure 4.15: Computing the λ satisfying Equation 4.6.3, i.e. the *most confusing instance*, can be viewed as an information-theoretic projection onto the boundary of the normal cone spanned by the active constraints at π_μ. In A) we see the different normal cones for the three different examples in Figure 4.14. In B) we have fixed μ_1 and μ_3, as in Figure 4.14, and plot the lower bound, assuming $N(0,1)$ noise and with $\delta = 0.1$, for increasing μ_2 which mean that we are moving $\boldsymbol{\mu}$ closer to the boundaries in A). We observe an inverse relationship between the distance to the boundary and the lower bound, properly characterized in Paper 6.

Recall, from Section 2.4, that lower bounds in multi-armed bandits can be written on the form

$$\mathbb{E}_{\boldsymbol{\mu},\phi}[\tau_\delta] \geqslant T_{\mathcal{F}}(\boldsymbol{\mu}) \log \frac{1}{2.4\delta}$$

where $T_{\mathcal{F}}$ is the solution to a zero-sum game between a learner, that samples arms according to w, and an adversary that outputs a confusing instance λ where the optimal policy is different form the one under μ [2]

$$T_{\mathcal{F}}^{-1}(\mu) = \sup_{w} \inf_{\lambda \in \Lambda_{\mathcal{F}}(\mu)} \sum_{a=1}^{K} w_a \mathbb{KL}(\mu_a, \lambda_a) \tag{4.6.2}$$

here $\Lambda_{\mathcal{F}}(\mu)$ is the set of alternative instances

$$\Lambda_{\mathcal{F}}(\mu) = \{\lambda \in \mathbb{R}^K : \max_{\pi \in \mathcal{F}} \lambda^\top \pi > \lambda^\top \pi^*\}.$$

[2]Here, $\mathbb{KL}(\mu_a, \lambda_a) = \mathbb{KL}(\mu_a, ||\lambda_a)$ and the different notation, compared to Chapter 2, is due to the notion $\mathbb{KL}(.,.)$ being used in Paper 6.

One of our contributions is to show that the lower bound in the constraint setting depends on a non-convex projection onto the boundary normal cone spanned by the active constraints at the optimal policy, see Figure 4.15. Especially, given an allocation w, the adversary will output a problem instance that satisfies

$$\min_{\lambda:\lambda\in\partial\mathcal{N}(\pi*)} \sum_{a=1}^{K} w_a \mathbb{KL}(\mu_a, \lambda_a). \tag{4.6.3}$$

Here, $\partial\mathcal{N}(\pi*)$ denotes the boundary of the normal cone spanned by the active constraints at the optimal policy. A formal version of this result, with an explicit expression of the boundary of the cone, is given in Lemma 3.1 in the main paper. We also leverage properties of set-valued functions to show that this projection satisfies certain continuity properties in w and μ, which in turn enables us to compute it with standard optimization techniques.

The lower bound in (4.6.2) is implicit and doesn't reveal how the hardness of the problem depends on the constraints and the reward vector μ. We address this in the paper by deriving more explicit lower bounds for Gaussian reward distributions.

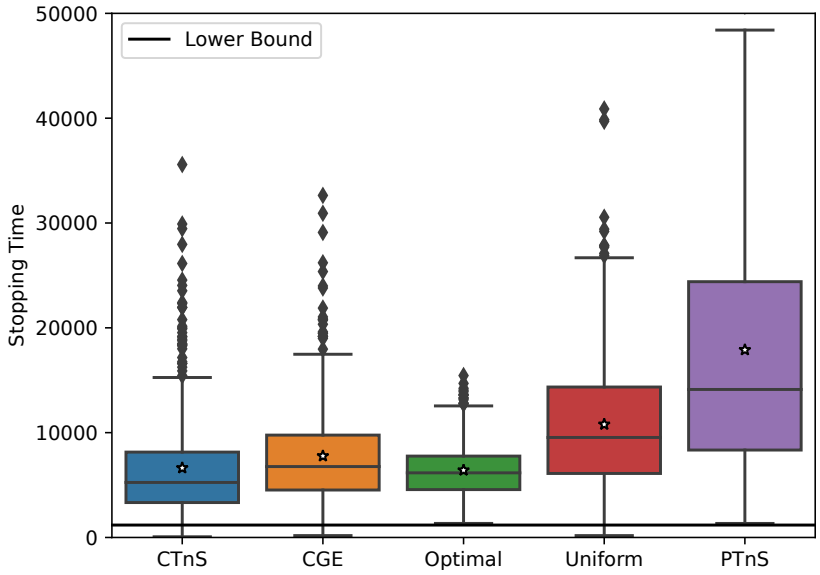


Figure 4.16: Y-axis corresponds to the time (number of samples) until an algorithm stops and outputs the best policy with confidence $1-\delta$. The figure illustrates the sample complexity of our algorithms (CTnS and CGE) against baselines on a problem where to goal is to find the optimal allocation of movies, w.r.t. genre constraints, in the IMDB dataset. For each algorithm, we performed the experiments over 1000 different random seeds.

On the algorithmic side, we introduce two algorithms, CTnS and CGE, which are adaptations of standard BAI algorithms, to the constraint setting. We prove that both these algorithms are *asymptotically* optimal in δ. That is, their expected

sample complexity τ satisfy

$$\limsup_{\delta \to 0} \mathbb{E}[\tau] / \log \frac{1}{\delta} \leqslant T_{\mathcal{F}}.$$

Our empirical evaluation shows that our algorithms have an advantage over baselines. In Figure 4.16 we show the performance of our algorithms against three baselines: optimal, uniform, and a version of TnS (Kaufmann et al. 2016) that projects the exploration policy onto the feasible set. Note that the optimal baseline is not possible in practice since it samples from the w given by (4.6.2) which requires knowledge of the true rewards μ. We observe that our algorithms operate close to the lower bound even for moderately large δ and their performance is on par with the optimal sampling policy. Since publishing this paper, other works have extended this setting to the fixed-budget regime (Tang et al. 2024) and unknown constraints (Gangrade et al. 2024; Das and Basu 2024).

4.7 Paper 7: Active preference learning for ordering items

In Paper 7, we study the problem of ordering a set of items, $\mathcal{I}$, using active preference learning. In our model, each item, $i \in \mathcal{I}$, is associated with a known feature vector $x_i \in \mathbb{R}^d$ and an unknown score $y_i \in \mathbb{R}$. Our goal is to order the items based on their score.

We assume that we can request a comparison of any two items, $i, j \in \mathcal{I}$, and receive a noisy binary preference $c \sim p(C_{ij})$. We further assume that the unknown scores satisfy a linear model,

$$y_i = \theta_*^\top x_i,$$

for some unknown $\theta_* \in \mathbb{R}^d$, and that the noisy preference feedback follows a logistic model

$$p(C_{ij}) = \sigma\left(y_i - y_j\right),$$

where $\sigma(.)$ is the sigmoid function. Hence, to order the items in $\mathcal{I}$ we need to estimate θ_* sufficiently well in the direction of the feature vectors $\{x_i\}_{i\in\mathcal{I}}$. This type of model has applications in medical imaging (Phelps et al. 2015; Jang et al. 2022; Lidén et al. 2024) as well as in *reinforcement learning with human feedback* (RLHF) (Ouyang et al. 2022; Das, Chakraborty, et al. 2024).

Our main contributions consist of deriving a data-dependent upper bound for the ordering error after T noisy comparisons, followed by two sampling strategies that, greedily, try to minimize this upper bound.

Let the ordering error of an estimate θ_T be defined as

$$R(\theta_T) := \frac{2}{n(n-1)} \sum_{i\neq j\in\mathcal{I}} \mathbf{1}[\text{sgn}\left(\theta_T^\top z_{ij}\right) \neq sgn\left(\theta_*^\top z_{ij}\right)]$$

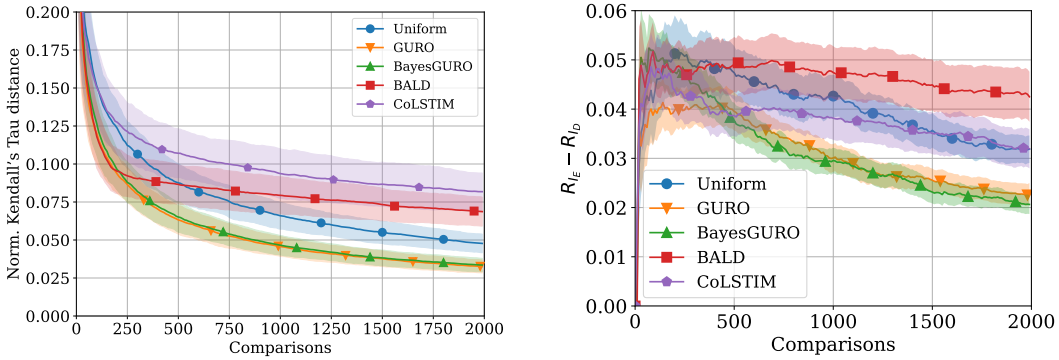
where $z_{ij} := x_i - x_j$. Our data-dependent bound suggests that the probability that the ordering error exceeds some $\epsilon > 0$ after collecting a dataset, D_T, of T comparisons is upper bounded as [3]

$$P\left(R(\theta_T) \geqslant \epsilon\right) \lessapprox \frac{4dT}{\epsilon} \exp\left[-\Delta^2 T / (\max_{i,j} \dot{\sigma}(z_{ij}^\top \theta_T)^2 \|z_{ij}\|^2_{\tilde{\mathbf{H}}_T^{-1}(\theta_T)})\right]. \tag{4.7.1}$$

Here, $\Delta = \min_{i\neq j} \Delta_{ij}/|i-j|$ where Δ_{ij} is difference in score between any i, j, $\mathbf{H}_T(\theta_T)$ is the Hessian of the negative log-likelihood around our estimated parameter θ_T

$$\mathbf{H}_T(\theta_T) := \sum_{t=1}^{T} \dot{\sigma}(z_{i_t,j_t}^\top \theta) z_{i_t,j_t} z_{i_t,j_t}^\top,$$

[3] To ease the presentation, we have ignored second-order terms here. See Theorem 4.2 in the paper for a precise upper bound.



(a) Mean R_{I_D} with 1-sigma error region.

(b) Mean generalization error (95% CI)

Figure 4.17: **X-RayAge**. Performance of active sampling strategies when comparisons are simulated using a logistic model. In-sample Kendall's Tau distance (ordering error) R_{I_D} on 200 images (left) and generalization error $R_{I_E} - R_{I_D}$ for models trained on 150 images and evaluated on 150 images from a different distribution (right). All results are averaged over 100 different random seeds.

and $\|z_{ij}\|_{\tilde{\mathbf{H}}_T^{-1}(\theta_T)} = \sqrt{z_{ij}^\top \mathbf{H}_T(\theta_T)^{-1} z_{ij}}$. The bound in (4.7.1) holds true for any sampling strategy and depends on the collected data through the estimated parameter θ_T as well as the Hessian $\mathbf{H}_T(\theta_T)$, which is also known as the *observed Fisher information*. In short, the bound suggest that a good active learning strategy should collect data such that the quantity $\max_{i,j} \dot{\sigma}(z_{ij}^\top \theta_T)\|z_{ij}\|_{\tilde{\mathbf{H}}_T^{-1}(\theta_T)}$ is minimized, as this would minimize our upper bound on the probability of error. Note that the variance in a noisy comparison between two items, i, j, under the predicted model θ_T, is equal to the derivative $\dot{\sigma}(z_{ij}^\top \theta_T)^2$ while $\|z_{ij}\|^2_{\tilde{\mathbf{H}}_T^{-1}(\theta_T)}$ is a measure of model uncertainty. Thus, (4.7.1) suggests that high model certainty is needed in directions with high variance.

In the paper, we leverage these theoretical insights and introduce the active learning algorithm GURO, short for *Greedy Uncertainty Reduction for Ordering*, which at every time t query a pair of items that satisfy

$$\max_{i,j} \dot{\sigma}(z_{ij}^\top \theta_t)\|z_{ij}\|_{\tilde{\mathbf{H}}_T^{-1}(\theta_t)}.$$

Here θ_t is the maximum-likelihood estimate given the samples seen so far. In the paper, we also present a Bayesian version of GURO, named BayesGURO, that can incorporate prior beliefs about the underlying environment.

In Section 6 of Paper 7, we compare our proposed algorithms against various baselines in both synthetic experiments as well as experiments that build on real preference feedback from human annotators. Our results indicate that our algorithms have an advantage over baselines. In Figure 4.17 we present one of our experiments where the goal is to order a set of X-ray images according to patient age. Here, the feature vectors $\{x_i\}_{i\in\mathcal{I}}$ were extracted by passing the X-ray images through a pre-trained CNN, and the unknown scores are the age of the patients. We observe that our algorithms outperform both uniform sampling as well as two other active learning algorithms, BALD (Houlsby et al. 2011) and CoLSTIM (Bengs et al. 2022).

Chapter 5

Concluding remarks and future directions

In this thesis, we have used reinforcement learning and multi-armed bandits to explore several aspects of sequential decision-making under uncertainty and how these decisions might gradually shape the behavior of the agents. We have shown that reinforcement learning agents, communicating with each other in a collaborative setting, eventually develop a shared language. The resulting artificial languages are efficient in an information-theoretic sense, an important property of human languages. Recent works have argued that a combination of a pressure for informativeness, coming from the need to solve communicative tasks, and a pressure for simplicity, stemming from learning, accounts for the efficiency found in human languages (Kirby, Tamariz, et al. 2015; Carr et al. 2020) and our results support these arguments. This is because our reinforcement learning agents have a clear bias towards informativeness, induced by their goal to maximize the joint reward, while they also have a bias towards simplicity due to the fact that they need to learn and converge on a joint language. In addition, one of our key results in this line of work was showing that a combination of reinforcement learning and iterated learning accounts for efficient color naming systems found in human languages. In this model, iterated learning reinforces the simplicity bias and our results suggest that this model account better for the data, compared to either reinforcement learning alone or iterated learning alone.

We have also explored how theoretical insights can be used to derive more sample efficient algorithms for multi-armed bandit problems. This has resulted in sample efficient algorithms for the multi-armed bandit problem with clustered arms, as well as provably optimal algorithms for the problem of identifying an optimal policy that is subject to pre-defined constraints. In Paper 7, we used theoretical results from multi-armed bandits to derive algorithms for active preference learning and showed that these outperform baselines.

5.1 Future directions

An interesting future direction is to explore whether the combination of reinforcement learning and iterated learning, used in Paper 4 , can account for efficient communication in other domains where human languages have been shown to support efficient communication. This is important because the notion of efficiency is not always sufficient to account for naming systems found in human languages and additional constraints might be induced by an evolutionary process.

Recent works have explored the learnability of various semantic universals by applying off-the-shelf machine learning methods and studying how rapidly these learn certain properties (Steinert-Threlkeld and Szymanik 2020; Douven 2023). A key finding is that many of the universals found in human languages, such like color words being convex regions in the color space (Gärdenfors 2000; Jäger 2010), are easier to learn for machine learning models. A limitation of these works is that they study learnability trough the lens of just one particular learning algorithm. Here, we think an interesting direction would be to borrow from the vast amount of theoretical results regarding sample complexity that is found in the multi-armed bandit literature. These results can potentially be used to study learnability for a whole class of learning algorithms simultaneously. To give an example, an interesting future direction is to use tools from the bandit literature, like the lower bound result described in Section 2.4, to compute lower bounds on the sample complexity of certain semantic universals. These lower bounds might give an indication for how hard certain properties are to learn for a whole family of algorithms and thus complement the already existing works on semantic universals and learnability.

Another important direction is to extend the work in Paper 2 to recursive numeral systems. Some work has already been done in this direction using either a single agent setup (Thomas, Silvi, et al. 2024) or iterated learning (Guo, Ren, et al. 2020). What is currently unknown is whether efficient recursive systems can emerge in a cooperative multi-agent setting, like the ones considered in this thesis, and whether a single model can learn approximate, exact restricted, and recursive numeral systems. The latter is interesting because such a model would account for how a numeral system evolves from one type of system to another. A potential approach is to combine iterated learning with some (neuro) symbolic mechanism. In such a model, one would expect that the presence of a communicative task dictates what type of system emerges. If the task requires a very precise communication of numbers over a large range, a recursive system should emerges, while a lower pressure towards informativeness might lead to approximate or exact restricted systems.

A limitation of our work is the one-way communication between the speaker and listener. In practice, agents are able to communicate back and forth with each other, and exploring how this impacts the efficiency of the communication is an important future direction.

When it comes to sample efficient algorithms in multi-armed bandits, an important direction is to extend the work done in Paper 6 to the case with *a priori* unknown constraints. Another interesting direction is extending the algorithms introduced in Paper 7 to be able to handle preferences along several dimensions at the same time.

Bibliography

Abbasi-Yadkori, Yasin, Dávid Pál, and Csaba Szepesvári (2011). "Improved Algorithms for Linear Stochastic Bandits". In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger. Vol. 24. Curran Associates, Inc. (cit. on p. 12).

Agrawal, Shipra and Navin Goyal (17–19 Jun 2013). "Thompson Sampling for Contextual Bandits with Linear Payoffs". In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by Sanjoy Dasgupta and David McAllester. Vol. 28. Proceedings of Machine Learning Research 3. Atlanta, Georgia, USA: PMLR, pp. 127–135 (cit. on p. 11).

Åkerblom, Niklas, Yuxin Chen, and Morteza Haghir Chehreghani (2023). "Online learning of energy consumption for navigation of electric vehicles". In: *Artificial Intelligence* 317, p. 103879 (cit. on p. 3).

Audibert, Jean-Yves and Sébastien Bubeck (2010). "Best arm identification in multi-armed bandits". In: *COLT-23th Conference on learning theory-2010*, 13–p (cit. on p. 9).

Auer, Peter, Nicolò Cesa-Bianchi, and Paul Fischer (2002). "Finite-time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47.2, pp. 235–256 (cit. on p. 12).

Aziz, Maryam, Emilie Kaufmann, and Marie-Karelle Riviere (2021). "On multi-armed bandit designs for dose-finding clinical trials". In: *The Journal of Machine Learning Research* 22.1, pp. 686–723 (cit. on p. 38).

Balcıoğlu, Ahmet Zahid, Emil Carlsson, and Fredrik D. Johansson (2024). "Identifiable latent bandits: Combining observational data and exploration for personalized healthcare". In: *ICML 2024 Workshop: Foundations of Reinforcement Learning and Control – Connections and Perspectives* (cit. on p. 6).

Bellman, Richard (1957). "A Markovian Decision Process". In: *Journal of Mathematics and Mechanics* 6.5, pp. 679–684. ISSN: 00959057, 19435274. (Visited on 05/09/2024) (cit. on p. 7).

Bengs, Viktor, Aadirupa Saha, and Eyke Hüllermeier (2022). "Stochastic Contextual Dueling Bandits under Linear Stochastic Transitivity Models". In: *International Conference on Machine Learning*. PMLR, pp. 1764–1786 (cit. on p. 43).

Bergström, Herman, Emil Carlsson, Devdatt Dubhashi, and Fredrik D. Johansson (2024). "Active Preference Learning for Ordering Items In- and Out-of-sample". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. Forthcoming (cit. on p. 6).

Berlin, Brent and Paul Kay (1969). *Basic Color term. Their Universality and Evolution*. 2010. Berlin, Boston: De Gruyter Mouton (cit. on pp. 15, 18).

Boldt, Brendon and David R Mortensen (2024). "A Review of the Applications of Deep Learning-Based Emergent Communication". In: *Transactions on Machine Learning Research*. ISSN: 2835-8856 (cit. on p. 20).

Börgers, Tilman and Rajiv Sarin (1997). "Learning through reinforcement and replicator dynamics". In: *Journal of economic theory* 77.1, pp. 1–14 (cit. on p. 20).

Bouneffouf, Djallel, Srinivasan Parthasarathy, Horst Samulowitz, and Martin Wistuba (July 2019). "Optimal Exploitation of Clustering and History Information in Multi-armed Bandit". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 2016–2022 (cit. on p. 37).

Bubeck, Sébastien, Rémi Munos, and Gilles Stoltz (2009). "Pure exploration in multi-armed bandits problems". In: *Algorithmic Learning Theory: 20th International Conference, ALT 2009, Porto, Portugal, October 3-5, 2009. Proceedings 20*. Springer, pp. 23–37 (cit. on p. 9).

Carcassi, Fausto, Shane Steinert-Threlkeld, and Jakub Szymanik (2021). "Monotone Quantifiers Emerge via Iterated Learning". In: *Cognitive Science* 45.8, e13027 (cit. on p. 22).

Carlsson, Emil, Debabrota Basu, Fredrik Johansson, and Devdatt Dubhashi (Feb. 2024). "Pure Exploration in Bandits with Linear Constraints". In: *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. Ed. by Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li. Vol. 238. Proceedings of Machine Learning Research. PMLR, pp. 334–342 (cit. on p. 6).

Carlsson, Emil and Devdatt Dubhashi (2022). "Pragmatic Reasoning in Strcutured Signalling Games". In: *Proceedings of the Annual Meeting of the Cognitive Science Society 44* (cit. on p. 5).

Carlsson, Emil, Devdatt Dubhashi, and Fredrik D. Johansson (2021a). "Learning Approximate and Exact Numeral Systems via Reinforcement Learning". In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 43 (cit. on p. 5).

Carlsson, Emil, Devdatt Dubhashi, and Fredrik D. Johansson (Aug. 2021b). "Thompson Sampling for Bandits with Clustered Arms". In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. Ed. by Zhi-Hua Zhou. Main Track. International Joint Conferences on Artificial Intelligence Organization, pp. 2212–2218. DOI: 10.24963/ijcai.2021/305 (cit. on p. 6).

Carlsson, Emil, Devdatt Dubhashi, and Terry Regier (2023). "Iterated learning and communication jointly explain efficient color naming systems". In: *Proceedings of the annual meeting of the cognitive science society*. Vol. 45. 45 (cit. on p. 5).

Carlsson, Emil, Devdatt Dubhashi, and Terry Regier (2024). "Cultural evolution via iterated learning and communication explains efficient color naming systems". In: *Journal of Language Evolution*. DOI: 10.1093/jole/lzae010. Forthcoming (cit. on p. 5).

Carr, Jon W., Kenny Smith, Jennifer Culbertson, and Simon Kirby (2020). "Simplicity and informativeness in semantic category systems". In: *Cognition* 202, p. 104289 (cit. on pp. 21, 23, 34, 45).

Carstensen, Alexandra, Jing Xu, Cameron T. Smith, and Terry Regier (2015). "Language evolution in the lab tends toward informative communication." In: *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (cit. on p. 23).

Cesa-Bianchi, Nicolo and Gábor Lugosi (2006). *Prediction, learning, and games*. Cambridge university press (cit. on p. 20).

Chaabouni, Rahma, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni (Mar. 2021). "Communicating artificial neural networks develop efficient color-naming systems". In: *Proceedings of the National Academy of Sciences* 118, e2016569118. DOI: 10.1073/pnas.2016569118 (cit. on p. 20).

Chapelle, Olivier and Lihong Li (2011). "An Empirical Evaluation of Thompson Sampling". In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger. Vol. 24. Curran Associates, Inc. (cit. on p. 12).

Chen, Sihan, Richard Futrell, and Kyle Mahowald (2023). "An information-theoretic approach to the typology of spatial demonstratives". In: *Cognition* 240 (cit. on p. 16).

Chernoff, Herman (1959). "Sequential Design of Experiments". In: *The Annals of Mathematical Statistics* 30.3, pp. 755–770. ISSN: 00034851 (cit. on p. 9).

Chomsky, Noam (1986). *Knowledge of language: Its nature, origin, and use*. New York (cit. on p. 15).

Comrie, Bernard (2013). "Numeral Bases". In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology (cit. on p. 18).

Cook, Richard S., Paul Kay, and Terry Regier (2005). "The World Color Survey Database: History and use". In: *Handbook of Categorization in Cognitive Science*. Ed. by Henri Cohen and Claire Lefebvre. Amsterdam: Elsevier, pp. 223–241 (cit. on p. 18).

Dabney, Will, Zeb Kurth-Nelson, Naoshige Uchida, Clara Starkweather, Demis Hassabis, Remi Munos, and Matthew Botvinick (Jan. 2020). "A distributional code for value in dopamine-based reinforcement learning". In: *Nature* 577, pp. 1–5 (cit. on p. 22).

Dale, Rick and Gary Lupyan (2012). "Understanding the Origins of Morphological Diversity: the Linguistic Niche Hypothesis". In: *Adv. Complex Syst.* 15 (cit. on p. 19).

Das, Nirjhar, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury (2024). "Provably Sample Efficient RLHF via Active Preference Optimization". In: *arXiv preprint arXiv:2402.10500* (cit. on p. 42).

Das, Udvas and Debabrota Basu (2024). "Learning to Explore with Lagrangians for Bandits under Unknown Constraints". In: *ICML 2024 Workshop: Foundations of Reinforcement Learning and Control–Connections and Perspectives* (cit. on p. 41).

Degenne, Rémy, Wouter M Koolen, and Pierre Ménard (2019). "Non-Asymptotic Pure Exploration by Solving Games". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32 (cit. on p. 13).

Douven, Igor (2023). "The role of naturalness in concept learning: A computational study". In: *Minds and Machines* 33.4, pp. 695–714 (cit. on p. 46).

Downey, CM, Leo Z Liu, Xuhui Zhou, and Shane Steinert-Threlkeld (2022). "Learning to translate by learning to communicate". In: *arXiv preprint arXiv:2207.07025* (cit. on p. 20).

Dryer, Matthew S (1998). "Why statistical universals are better than absolute universals". In: *Papers from the 33rd Regional Meeting of the Chicago Linguistic Society*, pp. 1–23 (cit. on p. 15).

Evans, Nicholas and Stephen C Levinson (2009). "The myth of language universals: Language diversity and its importance for cognitive science". In: *Behavioral and brain sciences* 32.5, pp. 429–448 (cit. on p. 15).

Fedzechkina, Maryia, T Florian Jaeger, and Elissa L Newport (2012). "Language learners restructure their input to facilitate efficient communication". In: *Proceedings of the National Academy of Sciences* 109.44, pp. 17897–17902 (cit. on p. 23).

Foerster, Jakob, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson (2016). "Learning to communicate with deep multi-agent reinforcement learning". In: *Advances in neural information processing systems* 29 (cit. on pp. 19, 20).

Frank, Michael C. and Noah D. Goodman (2012). "Predicting Pragmatic Reasoning in Language Games". In: *Science* 336.6084, pp. 998–998. DOI: 10.1126/science.1218633 (cit. on p. 31).

Gal, Yarin and Zoubin Ghahramani (2016). "Dropout as a bayesian approximation: Representing model uncertainty in deep learning". In: *international conference on machine learning*. PMLR, pp. 1050–1059 (cit. on pp. 12, 20, 28, 29).

Gangrade, Aditya, Tianrui Chen, and Venkatesh Saligrama (2024). "Safe Linear Bandits over Unknown Polytopes". In: *The Thirty Seventh Annual Conference on Learning Theory*. PMLR, pp. 1755–1795 (cit. on p. 41).

Gärdenfors, Peter (2000). "Conceptual spaces: The geometry of thought". In: *MIT Press* 3, p. 16 (cit. on pp. 35, 46).

Gärdenfors, Peter (2014). *The geometry of meaning: Semantics based on conceptual spaces*. MIT press (cit. on p. 15).

Garivier, Aurélien and Emilie Kaufmann (2016). "Optimal best arm identification with fixed confidence". In: *Conference on Learning Theory*. PMLR, pp. 998–1027 (cit. on p. 13).

Gershman, Samuel J (2018). "Deconstructing the human algorithms for exploration". In: *Cognition* 173, pp. 34–42 (cit. on p. 21).

Gershman, Samuel J and Nathaniel D Daw (2017). "Reinforcement learning and episodic memory in humans and animals: an integrative framework". In: *Annual review of psychology* 68, pp. 101–128 (cit. on p. 3).

Gibson, Edward, Richard Futrell, Julian Jara-Ettinger, Kyle Mahowald, Leon Bergen, Sivalogeswaran Ratnasingam, Mitchell Gibson, Steven T. Piantadosi, and Bevil R.

Conway (2017). "Color naming across languages reflects color use". In: *Proceedings of the National Academy of Sciences*. ISSN: 0027-8424 (cit. on pp. 17, 25).

Gibson, Edward, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy (2019). "How Efficiency Shapes Human Language". In: *Trends in Cognitive Sciences* 23.5, pp. 389–407 (cit. on pp. 3, 15, 21).

Grice, H. Paul (1975). "Logic and Conversation". In: *The Semantics-Pragmatics Boundary in Philosophy*. Ed. by Maite Ezcurdia and Robert J. Stainton. Broadview Press, p. 47 (cit. on p. 30).

Griffiths, T.L. and M.L. Kalish (May 2007). "Language evolution by iterated learning with Bayesian agents". In: *Cognitive Science* 31, pp. 441–480 (cit. on pp. 19, 23).

Guo, Shangmin, Yi Ren, Serhii Havrylov, Stella Frank, Ivan Titov, and Kenny Smith (2020). "The emergence of compositional languages for numeric concepts through iterated learning in neural agents". In: *Evolution of Language International Conferences* (cit. on pp. 24, 46).

Guo, Yuxuan, Yifan Hao, Rui Zhang, Enshuai Zhou, Zidong Du, Xinkai Song, Yuanbo Wen, Yongwei Zhao, Xuehai Zhou, Jiaming Guo, et al. (2024). "Emergent Communication for Rules Reasoning". In: *Advances in Neural Information Processing Systems* 36 (cit. on p. 20).

Hammarström, H. (Jan. 2010). "Rarities in Numeral Systems". In: *Business Communication Quarterly - Bus Comm Q* (cit. on p. 18).

Havrylov, Serhii and Ivan Titov (2017). "Emergence of language with multi-agent games: Learning to communicate with sequences of symbols". In: *Advances in Neural Information Processing Systems* 2017-Decem, pp. 2150–2160. ISSN: 10495258. arXiv: 1705.11192 (cit. on pp. 19–21).

Hennes, Daniel, Dustin Morrill, Shayegan Omidshafiei, Rémi Munos, Julien Perolat, Marc Lanctot, Audrunas Gruslys, Jean-Baptiste Lespiau, Paavo Parmas, Edgar Duéñez-Guzmán, et al. (2020). "Neural replicator dynamics: Multiagent learning via hedging policy gradients". In: *Proceedings of the 19th international conference on autonomous agents and multiagent systems*, pp. 492–501 (cit. on p. 20).

Houlsby, Neil, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel (Dec. 2011). *Bayesian Active Learning for Classification and Preference Learning*. arXiv:1112.5745 [cs, stat]. DOI: 10.48550/arXiv.1112.5745. (Visited on 10/20/2023) (cit. on p. 43).

Hurford, James R (1987). *Language and number: The emergence of a cognitive system* (cit. on p. 18).

Imel, Nathaniel (2023). "The evolution of efficient compression in signaling games". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 45. 45 (cit. on p. 20).

Imel, Nathaniel, Richard Futrell, Michael Franke, and Noga Zaslavsky (2023). "Noisy Population Dynamics Lead to Efficiently Compressed Semantic Systems". In: *NeurIPS 2023 workshop: Information-Theoretic Principles in Cognitive Systems* (cit. on p. 20).

Imel, Nathaniel and Shane Steinert-Threlkeld (2022). "Modal semantic universals optimize the simplicity/informativeness trade-off". In: *Proceedings of SALT 32 (Semantics and Linguistic Theory)*, pp. 227–248 (cit. on p. 16).

Jäger, Gerhard (2010). "Natural Color Categories Are Convex Sets". In: *Logic, Language and Meaning*. Ed. by Maria Aloni, Harald Bastiaanse, Tikitu de Jager, and Katrin Schulz. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 11–20 (cit. on p. 46).

Jäger, Gerhard, Lars P Metzger, and Frank Riedel (2011). "Voronoi languages: Equilibria in cheap-talk games with high-dimensional types and few signals". In: *Games and economic behavior* 73.2, pp. 517–537 (cit. on p. 19).

Jang, Ikbeom, Garrison Danley, Ken Chang, and Jayashree Kalpathy-Cramer (2022). "Decreasing annotation burden of pairwise comparisons with human-in-the-loop sorting: Application in medical image artifact rating". In: *arXiv preprint arXiv:2202.04823* (cit. on p. 42).

Jergéus, Erik, Leo Karlsson Oinonen, Emil Carlsson, and Moa Johansson (2022). "Towards Learning Abstractions via Reinforcement Learning". In: *AIC 2022, 8th International Workshop on Artificial Intelligence and Cognition* (cit. on p. 6).

Jones, Rebecca M, Leah H Somerville, Jian Li, Erika J Ruberry, Alisa Powers, Natasha Mehta, Jonathan Dyke, and BJ Casey (2014). "Adolescent-specific patterns of behavior and neural activity during social reinforcement learning". In: *Cognitive, Affective, & Behavioral Neuroscience* 14, pp. 683–697 (cit. on p. 21).

Jorge, Emilio, Mikael Kågebäck, Fredrik D. Johansson, and Emil Gustavsson (2016). "Learning to Play Guess Who? and Inventing a Grounded Language as a Consequence". In: arXiv: `1611.03218` (cit. on p. 20).

Kågebäck, Mikael, Emil Carlsson, Devdatt Dubhashi, and Asad Sayeed (2020). "A reinforcement-learning approach to efficient communication". In: *PLoS ONE* 15.7, pp. 1–26 (cit. on pp. 5, 25, 28, 33).

Kato, Masahiro and Kaito Ariu (2024). *The Role of Contextual Information in Best Arm Identification*. arXiv: `2106.14077` (cit. on p. 10).

Kaufmann, Emilie, Olivier Cappé, and Aurélien Garivier (Jan. 2016). "On the Complexity of Best-Arm Identification in Multi-Armed Bandit Models". In: *J. Mach. Learn. Res.* 17.1, pp. 1–42. ISSN: 1532-4435 (cit. on pp. 10, 41).

Kemp, Charles, Alice Gaby, and Terry Regier (2019). "Season naming and the local environment". In: *Proceedings of the 41st Annual Meeting of the Cognitive Science Society* (cit. on p. 16).

Kemp, Charles and Terry Regier (May 2012). "Kinship Categories Across Languages Reflect General Communicative Principles". In: *Science (New York, N.Y.)* 336, pp. 1049–54 (cit. on pp. 16, 17).

Kemp, Charles, Yang Xu, and Terry Regier (Jan. 2018). "Semantic Typology and Efficient Communication". In: *Annual Review of Linguistics* 4, pp. 109–128 (cit. on pp. 3, 15, 17, 21).

Khetarpal, Naveen, Lev Michael, Terry Regier, and Grace Neveu (Jan. 2013). "Spatial terms across languages support near-optimal communication: Evidence from Peruvian Amazonia, and computational analyses". In: (cit. on p. 16).

Kinyanjui, Newton Mwai, Emil Carlsson, and Fredrik D. Johansson (2023). "Fast Treatment Personalization with Latent Bandits in Fixed-Confidence Pure Exploration". In: *Transactions on Machine Learning Research*. ISSN: 2835-8856 (cit. on p. 6).

Kirby, Simon (2001). "Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity". In: *IEEE Transactions on Evolutionary Computation* 5.2, pp. 102–110 (cit. on p. 22).

Kirby, Simon (2002a). "Learning, bottlenecks and the evolution of recursive syntax". In: (cit. on p. 23).

Kirby, Simon (2002b). "Natural Language From Artificial Life". In: *Artificial Life* 8.2, pp. 185–215. DOI: 10.1162/106454602320184248 (cit. on p. 19).

Kirby, Simon, Hannah Cornish, and Kenny Smith (2008). "Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language". In: *Proceedings of the National Academy of Sciences* 105.31, pp. 10681–10686 (cit. on pp. 21–23).

Kirby, Simon and Monica Tamariz (2022). "Cumulative cultural evolution, population structure and the origin of combinatoriality in human language". In: *Philosophical Transactions of the Royal Society B* 377.1843, p. 20200319 (cit. on pp. 22, 24).

Kirby, Simon, Monica Tamariz, Hannah Cornish, and Kenny Smith (2015). "Compression and communication in the cultural evolution of linguistic structure". In: *Cognition* 141, pp. 87–102 (cit. on pp. 21, 23, 24, 33, 45).

Kober, Jens, J Andrew Bagnell, and Jan Peters (2013). "Reinforcement learning in robotics: A survey". In: *The International Journal of Robotics Research* 32.11, pp. 1238–1274 (cit. on p. 3).

Lai, T.L and H Robbins (1985). "Asymptotically efficient adaptive allocation rules". In: *Advances in Applied Mathematics* 6, pp. 4–22 (cit. on p. 8).

Lattimore, Tor and Csaba Szepesvári (2020). *Bandit Algorithms*. Cambridge University Press. DOI: 10.1017/9781108571401 (cit. on pp. 4, 7).

Lazaridou, Angeliki and Marco Baroni (2020). *Emergent Multi-Agent Communication in the Deep Learning Era*. arXiv: 2006.02419 [cs.CL] (cit. on p. 20).

Lazaridou, Angeliki, Alexander Peysakhovich, and Marco Baroni (2017). "Multi-agent cooperation and the emergence of (natural) language". In: *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pp. 1–11. arXiv: 1612.07182 (cit. on pp. 19–21).

Leinster, Tom (2021). *Entropy and Diversity The Axiomatic Approach*. Cambridge University Press (cit. on p. 31).

Levinson, Stephen, Sérgio Meira, The Language, and Cognition Group (2003). "'Natural concepts' in the spatial topological domain-Adpositional meanings in crosslinguistic perspective: An exercise in semantic typology". In: *Language*, pp. 485–516 (cit. on p. 15).

Lewis, David K. (1969). *Convention: A Philosophical Study*. Wiley-Blackwell (cit. on pp. 4, 20).

Li, Lihong, Wei Chu, John Langford, and Robert E Schapire (2010). "A contextual-bandit approach to personalized news article recommendation". In: *Proceedings of the 19th international conference on World wide web*, pp. 661–670 (cit. on p. 3).

Li, Lisha, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar (2017). "Hyperband: A novel bandit-based approach to hyperparameter optimization". In: *The Journal of Machine Learning Research* 18.1, pp. 6765–6816 (cit. on p. 38).

Lian, Yuchen, Arianna Bisazza, and Tessa Verhoef (2023). "Communication Drives the Emergence of Language Universals in Neural Agents: Evidence from the Word-order/Case-marking Trade-off". In: *Transactions of the Association for Computational Linguistics* 11, pp. 1033–1047 (cit. on p. 20).

Lidén, Mats, Antoine Spahr, Ola Hjelmgren, Simone Bendazzoli, Josefin Sundh, Magnus Sköld, Göran Bergström, Chunliang Wang, and Per Thunberg (Jan. 2024). "Machine learning slice-wise whole-lung CT emphysema score correlates with airway obstruction". en. In: *European Radiology* 34.1, pp. 39–49. ISSN: 1432-1084. DOI: 10.1007/s00330-023-09985-3. (Visited on 01/26/2024) (cit. on p. 42).

Ludvig, Elliot A, Marc G Bellemare, and Keir G Pearson (2011). "A primer on reinforcement learning in the brain: Psychological, computational, and neural perspectives". In: *Computational neuroscience for advancing artificial intelligence: Models, methods and applications*. IGI Global, pp. 111–144 (cit. on p. 21).

Magureanu, Stefan, Richard Combes, and Alexandre Proutiere (2014). "Lipschitz bandits: Regret lower bound and optimal algorithms". In: *Conference on Learning Theory*. PMLR, pp. 975–999 (cit. on p. 10).

Majid, Asifa, Melissa Bowerman, Sotaro Kita, Daniel BM Haun, and Stephen C Levinson (2004). "Can language restructure cognition? The case for space". In: *Trends in cognitive sciences* 8.3, pp. 108–114 (cit. on p. 15).

Marr, D. (1982). *Vision: A Computational Approach*. San Francisco, Freeman & Co. (cit. on p. 21).

Mertikopoulos, Panayotis and William H Sandholm (2016). "Learning in games via reinforcement and regularization". In: *Mathematics of Operations Research* 41.4, pp. 1297–1324 (cit. on p. 20).

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. (2015). "Human-level control through deep reinforcement learning". In: *nature* 518.7540, pp. 529–533 (cit. on p. 3).

Mordatch, Igor and Pieter Abbeel (2018). "Emergence of grounded compositional language in multi-agent populations". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1 (cit. on pp. 20, 21).

Motamedi, Yasamin, Marieke Schouwstra, Kenny Smith, Jennifer Culbertson, and Simon Kirby (2019). "Evolving artificial sign languages in the lab: From improvised gesture to systematic sign". In: *Cognition* 192, p. 103964 (cit. on p. 24).

Niv, Y. (2009). "Reinforcement learning in the brain". In: *The Journal of Mathematical Psychology* 53.3, pp. 139–154 (cit. on pp. 3, 21, 22).

Niv, Y. and A. Langdon (2016). "Reinforcement Learning with Marr". In: *Current Opinion in Behavioral Sciences* 11.3 (cit. on p. 21).

O'Shaughnessy, David, Edward Gibson, and Steven T. Piantadosi (2021). "The Cultural Origins of Symbolic Number". In: *Psychological Review* (cit. on p. 28).

O'Doherty, John P, Sang Wan Lee, and Daniel McNamee (2015). "The structure of reinforcement-learning mechanisms in the human brain". In: *Current Opinion in Behavioral Sciences* 1, pp. 94–100 (cit. on p. 3).

Ouyang, Long et al. (2022). *Training language models to follow instructions with human feedback*. arXiv: 2203.02155 [cs.CL] (cit. on p. 42).

Pandey, Sandeep, Deepayan Chakrabarti, and Deepak Agarwal (2007). "Multi-armed bandit problems with dependent arms". In: *ICML*, pp. 721–728 (cit. on p. 37).

Phelps, Andrew S., David M. Naeger, Jesse L. Courtier, Jack W. Lambert, Peter A. Marcovici, Javier E. Villanueva-Meyer, and John D. MacKenzie (2015). "Pairwise comparison versus Likert scale for biomedical image assessment." en. In: *AJR. American journal of roentgenology* 204.1, pp. 8–14. ISSN: 0361-803X. DOI: 10.2214/ajr.14.13022. (Visited on 01/26/2024) (cit. on p. 42).

Piaget, Jean (2013). *The construction of reality in the child*. Routledge (cit. on p. 3).

Pica, Pierre, Cathy Lemer, Véronique Izard, and Stanislas Dehaene (2004). "Exact and approximate arithmetic in an Amazonian indigene group". In: *Science* 306.5695, pp. 499–503 (cit. on p. 15).

Pinker, Steven and Paul Bloom (1990). "Natural language and natural selection". In: *Behavioral and brain sciences* 13.4, pp. 707–727 (cit. on p. 15).

Qin, Chao (Feb. 2022). "Open Problem: Optimal Best Arm Identification with Fixed-Budget". In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, pp. 5650–5654 (cit. on p. 9).

Rafferty, Anna N, Thomas L Griffiths, and Marc Ettlinger (2011). "Exploring the relationship between learnability and linguistic universals". In: *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pp. 49–57 (cit. on p. 24).

Regier, T., C. Kemp, and P. Kay (2015). "Word meanings across languages support efficient communication". In: *The handbook of language emergence*. Ed. by B. MacWhinney and W. O'Grady. Hoboken NJ: Wiley-Blackwell., pp. 237–263 (cit. on pp. 25, 26).

Regier, Terry, Paul Kay, and Naveen Khetarpal (2007). "Color naming reflects optimal partitions of color space". In: *Proceedings of the National Academy of Sciences of the United States of America* 104.4, pp. 1436–1441 (cit. on pp. 15, 17, 18).

Ren, Yi, Shangmin Guo, Matthieu Labeau, Shay B. Cohen, and Simon Kirby (2020). "Compositional languages emerge in a neural iterated learning model". In: *International Conference on Learning Representations* (cit. on pp. 23, 24, 33).

Riquelme, Carlos, George Tucker, and Jasper Snoek (2018). *Deep Bayesian Bandits Showdown: An Empirical Comparison of Bayesian Deep Networks for Thompson Sampling*. arXiv: 1802.09127 [stat.ML] (cit. on pp. 11, 12).

Rosch, Eleanor (1978). "Principles of categorization". In: *Cognition and categorization*. Routledge, pp. 27–48 (cit. on p. 15).

Rovee, Carolyn Kent and David T Rovee (1969). "Conjugate reinforcement of infant exploratory behavior". In: *Journal of experimental child psychology* 8.1, pp. 33–39 (cit. on p. 3).

Russo, Daniel (2016). "Simple bayesian algorithms for best arm identification". In: *Conference on Learning Theory*. PMLR, pp. 1417–1418 (cit. on p. 9).

Schultz, Wolfram, Peter Dayan, and P Read Montague (1997). "A neural substrate of prediction and reward". In: *Science* 275.5306, pp. 1593–1599 (cit. on p. 22).

Schulz, Eric and Samuel J. Gershman (2019). "The algorithmic architecture of exploration in the human brain". In: *Current Opinion in Neurobiology* 55. Machine Learning, Big Data, and Neuroscience, pp. 7–14 (cit. on p. 21).

Shannon, Claude Elwood (1948). "A Mathematical Theory of Communication". In: *The Bell System Technical Journal* 27, pp. 379–423 (cit. on p. 16).

Shennan, Stephen (2001). "Demography and cultural innovation: a model and its implications for the emergence of modern human culture". In: *Cambridge archaeological journal* 11.1, pp. 5–16 (cit. on p. 19).

Silver, David, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. (2016). "Mastering the game of Go with deep neural networks and tree search". In: *nature* 529.7587, pp. 484–489 (cit. on p. 3).

Skyrms, Brian (2010). *Signals: Evolution, learning, and information*. OUP Oxford (cit. on p. 19).

Slivkins, Aleksandrs (2019). "Introduction to Multi-Armed Bandits". In: *Foundations and Trends® in Machine Learning* 12.1-2, pp. 1–286. ISSN: 1935-8237 (cit. on p. 7).

Smith, Kenny and James R Hurford (2003). "Language evolution in populations: Extending the iterated learning model". In: *Advances in Artificial Life: 7th European Conference, ECAL 2003, Dortmund, Germany, September 14-17, 2003. Proceedings 7*. Springer, pp. 507–516 (cit. on p. 19).

Smith, Kenny, Simon Kirby, and Henry Brighton (2003). "Iterated learning: A framework for the emergence of language". In: *Artificial life* 9.4, pp. 371–386 (cit. on p. 22).

Smith, Kenny and Elizabeth Wonnacott (2010). "Eliminating unpredictable variation through iterated learning". In: *Cognition* 116.3, pp. 444–449 (cit. on p. 22).

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15.56, pp. 1929–1958 (cit. on p. 28).

Steels, Luc (1995). "A self-organizing spatial vocabulary". In: *Artificial life* 2.3, pp. 319–332 (cit. on p. 19).

Steels, Luc and Tony Belpaeme (2005). "Coordinating perceptually grounded categories through language: A case study for colour". In: *Behavioral and brain sciences* 28.4, pp. 469–488 (cit. on p. 19).

Steinert-Threlkeld, Shane and Jakub Szymanik (2019). "Learnability and semantic universals". In: *Semantics and Pragmatics* 12, pp. 4–1 (cit. on p. 22).

Steinert-Threlkeld, Shane and Jakub Szymanik (2020). "Ease of learning explains semantic universals". In: *Cognition* 195, p. 104076 (cit. on pp. 22, 35, 46).

Strens, Malcolm (2000). "A Bayesian framework for reinforcement learning". In: *ICML*. Vol. 2000, pp. 943–950 (cit. on p. 12).

Sumers, Theodore R, Mark K Ho, Thomas L Griffiths, and Robert D Hawkins (2023). "Reconciling truthfulness and relevance as epistemic and decision-theoretic utility." In: *Psychological Review* (cit. on p. 4).

Sutton, Richard S. and Andrew G. Barto (1998). *Reinforcement Learning: An Introduction*. Second. The MIT Press (cit. on pp. 3, 7, 11).

Tang, Dengwang, Rahul Jain, Ashutosh Nayyar, and Pierluigi Nuzzo (2024). "Pure Exploration for Constrained Best Mixed Arm Identification with a Fixed Budget". In: *arXiv preprint arXiv:2405.15090* (cit. on p. 41).

Thomas, Jonathan David, Raul Santos-Rodriguez, and Robert Piechocki (2022). "Understanding Redundancy in Discrete Multi-Agent Communication". In: *Second Workshop on Language and Reinforcement Learning* (cit. on p. 20).

Thomas, Jonathan David, Andrea Silvi, Devdatt Dubhashi, Emil Carlsson, and Moa Johansson (2024). "Learning Efficient Recursive Numeral Systems via Reinforcement Learning". In: *AI for Math Workshop @ ICML 2024* (cit. on pp. 6, 46).

Thompson, Bill, Simon Kirby, and Kenny Smith (2016). "Culture shapes the evolution of cognition". In: *Proceedings of the National Academy of Sciences* 113.16, pp. 4530–4535 (cit. on p. 22).

Thompson, William R. (1933). "On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples". In: *Biometrika* 25.3/4, pp. 285–294 (cit. on p. 11).

Thorndike, Edward L (1898). "Animal intelligence: An experimental study of the associative processes in animals." In: *The Psychological Review: Monograph Supplements* 2.4, p. i (cit. on p. 3).

Tishby, Naftali, Fernando C. Pereira, and William Bialek (1999). "The Information Bottleneck Method". In: *Proceedings of the 37th Allerton Conference on Communication, Control and Computation*, pp. 368–377 (cit. on pp. 17, 33).

Tomov, Momchil S, Eric Schulz, and Samuel J Gershman (2021). "Multi-task reinforcement learning in humans". In: *Nature Human Behaviour* 5.6, pp. 764–773 (cit. on p. 21).

Verhoef, Tessa, Simon Kirby, and Bart De Boer (2014). "Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals". In: *Journal of Phonetics* 43, pp. 57–68 (cit. on p. 22).

Von Fintel, Kai and Lisa Matthewson (2008). "Universals in semantics". In: (cit. on p. 15).

Wagner, Kyle, James A. Reggia, Juan Uriagereka, and Gerald S. Wilkinson (2003). "Progress in the Simulation of Emergent Communication and Language". In: *Adaptive Behavior* 11.1, pp. 37–69 (cit. on p. 19).

Williams, Ronald J (1992). "Simple statistical gradient-following algorithms for connectionist reinforcement learning". In: *Reinforcement Learning*. Springer, pp. 5–32 (cit. on pp. 11, 20).

Xu, Jing, Mike Dowman, and Thomas L. Griffiths (2013). "Cultural transmission results in convergence towards colour term universals". In: *Proceedings of the Royal Society B: Biological Sciences* 280.1758, p. 20123073 (cit. on p. 22).

Xu, Yang, Emmy Liu, and Terry Regier (2020). "Numeral Systems Across Languages Support Efficient Communication: From Approximate Numerosity to Recursion". In: *Open Mind* 4, pp. 57–70 (cit. on pp. 16, 17, 19, 28–30).

Yu, Chao, Jiming Liu, Shamim Nemati, and Guosheng Yin (2021). "Reinforcement learning in healthcare: A survey". In: *ACM Computing Surveys (CSUR)* 55.1, pp. 1–36 (cit. on p. 3).

Zaslavsky, Noga, Charles Kemp, Terry Regier, and Naftali Tishby (2018). "Efficient compression in color naming and its evolution". In: *Proceedings of the National Academy of Sciences of the United States of America* 115.31, pp. 7937–7942 (cit. on pp. 16–18, 33).

Zhao, T., M. Li, and M. Poloczek (2019). "Fast Reconfigurable Antenna State Selection with Hierarchical Thompson Sampling". In: *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, pp. 1–6 (cit. on p. 37).

Zipf, George K. (1949). *Human Behaviour and the Principle of Least Effort*. Addison-Wesley (cit. on pp. 15, 32).

Zuidema, Willem (2002). "How the poverty of the stimulus solves the poverty of the stimulus". In: *Advances in neural information processing systems* 15 (cit. on p. 23).