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

Abstract v

Acknowledgments vii

List of publications ix

Summary of contributions xi

I Introductory Chapters 1

1 Introduction 3

2 Reinforcement learning and multi-armed bandits 7

2.1 What is reinforcement learning? . . . 7

2.2 Multi-armed bandits . . . 8

2.3 The contextual bandit . . . 9

2.4 Lower bounds in multi-armed bandits . . . 10

2.5 Relevant algorithms . . . 11

2.5.1 REINFORCE . . . 11

2.5.2 Thompson sampling . . . 11

2.5.3 Optimism in the face of uncertainty . . . 12

2.5.4 Best-arm identification algorithms . . . 12

3 Reinforcement learning and efficient communication 15

3.1 Why do languages look the way they do? . . . 15

3.1.1 Efficient semantic categories . . . 15

3.2 Simulating language evolution . . . 19

3.2.1 Reinforcement learning and the signaling game . . . 20

3.2.2 Why reinforcement learning? . . . 21

3.2.3 Iterated learning . . . 22

4 Summary of included papers **25**
4.1 Paper 1: A reinforcement-learning approach to efficient communication 25
4.2 Paper 2: Learning approximate and exact numeral systems via reinforcement learning . . . 28
4.3 Paper 3: Pragmatic reasoning in structured signaling games . . . 30
4.4 Paper 4: Cultural evolution via iterated learning and communication explains efficient color naming systems . . . 33
4.5 Paper 5: Thompson sampling in bandits with clustered arms . . . 36
4.6 Paper 6: Pure exploration in bandits with linear constraints . . . 38
4.7 Paper 7: Active preference learning for ordering items . . . 42

5 Concluding remarks and future directions **45**
5.1 Future directions . . . 46

Bibliography **47**

II Appended Papers **59**

1 A reinforcement-learning approach to efficient communication **61**
1 Introduction . . . 63
1.1 Linguistic background on color identification . . . 65
1.2 Approach and contributions . . . 68
2 Efficient communication: A theoretical framework . . . 69
2.1 Information-theoretic communication loss . . . 69
2.2 Well–formedness . . . 71
2.3 Reinforcement learning framework for communication over a noisy channel . . . 72
2.4 Discrete policies . . . 74
2.5 Reward . . . 75
2.6 Training . . . 76
2.7 Generate partitioning . . . 76
3 Efficiency analysis . . . 76
3.1 Discrete vs continuous RL training . . . 77
3.2 KL loss evaluation . . . 78
3.3 Expected surprise evaluation . . . 78
3.4 Well–formedness evaluation . . . 79
3.5 Quantitative similarity using adjusted Rand index . . . 80
3.6 Analysis of consensus color partitions . . . 82
3.7 Developing an artificial language . . . 84
3.8 Modulating the vocabulary size by varying environmental noise 84
3.9 Modulating the vocabulary size by varying communication noise 87
4 Materials and methods . . . 87
4.1 CIELAB correlation clustering . . . 87
4.2 Consensus maps by correlation clustering . . . 88
4.3 REINFORCE . . . 88

4.4 Adjusted Rand Index . . . 89
4.5 The World Color Survey . . . 89
5 Discussion . . . 90
6 Conclusion . . . 91
References . . . 91

2 Learning approximate and exact numeral systems via reinforcement learning 95
1 Introduction . . . 97
2 Learning to communicate: Signalling games . . . 98
2.1 Reinforcement learning for efficient communication . . . 99
3 Numeral systems . . . 101
3.1 Artificial numeral systems . . . 101
3.2 Complexity and communication cost . . . 102
4 Experiments . . . 103
5 Conclusions and future work . . . 107
6 Acknowledgments . . . 108
References . . . 109

3 Pragmatic reasoning in structured signaling games 111
1 Introduction . . . 113
2 Structured signaling games and sRSA . . . 115
2.1 Similarity-sensitive utility and sRSA . . . 115
3 Color domain: Efficiency and well-formedness . . . 116
3.1 Human representations . . . 117
3.2 Artificial agents . . . 120
4 Conclusions . . . 122
5 Acknowledgements . . . 123
References . . . 124

4 Cultural evolution via iterated learning and communication explains efficient color naming systems 127
1 Introduction . . . 129
2 Not all efficient systems are human-like . . . 132
3 Iterated learning and communication . . . 134
4 Analyses and results . . . 137
4.1 Iterated learning and communication operating together . . . 137
4.2 Iterated learning alone, and communication alone . . . 140
4.3 The distribution of systems produced by IL+C . . . 141
4.4 Learnability and convexity . . . 143
5 Discussion . . . 146
A The framework of Zaslavsky et al. (2018) . . . 149
References . . . 151

5 Thompson sampling for bandits with clustered arms 157
1 Introduction . . . 159
2 Stochastic multi-armed bandit with clustered arms . . . 160
2.1 Thompson sampling for MABC . . . 160
2.2 Regret analysis TSC . . . 161
2.3 Lower bounds for disjoint clustering . . . 163
2.4 Regret analysis HTS . . . 164
3 Contextual bandit with linear rewards and clustered arms . . . 165
4 Experimental results . . . 165
4.1 Stochastic multi-armed bandit . . . 165
4.2 Contextual bandit . . . 168
5 Related work . . . 169
6 Conclusions . . . 169
A Proofs . . . 170
A.1 Lemma 2.2 . . . 170
A.2 Theorem 2.3 . . . 172
A.3 Theorem 2.4 . . . 173
A.4 Theorem 2.5 . . . 174
A.5 Theorem 2.6 . . . 174
A.6 Theorem 2.7 . . . 174
B Empirical evaluation MABC . . . 175
References . . . 176

6 Pure exploration in bandits with linear constraints 179
1 Introduction . . . 181
1.1 Related work . . . 183
2 Problem formulation . . . 184
3 Lower bound . . . 186
3.1 Lower bound for Gaussian distributions . . . 188
4 Algorithms . . . 190
5 Experimental analysis . . . 192
6 Conclusions and future directions . . . 195
A Notations . . . 197
B Lower bound on sample complexity . . . 199
B.1 Proof of Lemma 3.1 . . . 200
B.2 Proof of Theorem 3.2 . . . 201
B.3 Proof of Theorem 3.3 . . . 201
B.4 Proof of Corollary 3.4 . . . 203
B.5 Proof of Corollary 3.5 . . . 204
B.6 Theorem 3.3 reduces to the standard BAI bounds with simplex constraints . . . 206
C Upper bounds on sample complexity . . . 208
C.1 Stopping criterion . . . 208
C.2 Upper bound for CTnS . . . 209
C.3 Upper bound for CGE . . . 211

D Finding ϵ-good policies under linear constraints . . . 217
E Additional experimental analysis . . . 218
E.1 Running times . . . 219
E.2 IMDB environment . . . 221
F On the sub-optimality of PTnS . . . 222
G Useful definitions and results . . . 224
References . . . 225

7 Active preference learning for ordering items in- and out-of-sample 229
1 Introduction . . . 231
2 Ordering items with active preference learning . . . 233
3 Related work . . . 234
4 Which comparisons result in a good ordering? . . . 235
5 Greedy uncertainty reduction for ordering (GURO) . . . 237
5.1 Preference models for in- and out-of-sample ordering . . . 239
6 Experiments . . . 240
6.1 Ordering X-ray images under the logistic model . . . 241
6.2 Ordering items with human preference data . . . 242
7 Conclusion . . . 244
A Notation . . . 246
A Algorithms . . . 247
A.1 MLE estimator for logistic regression . . . 247
A.2 Bayesian estimator for logistic regression . . . 247
A.3 Stochastic Bayesian uncertainty reduction (BayesGURO) . . . 248
A.4 Uniform sampling . . . 248
A.5 BALD . . . 249
A Proofs of Lemma 4.1 and Theorem 4.2 . . . 251
A.1 Proof of Lemma 4.1 . . . 251
A.2 Proof of Theorem 4.2 . . . 255
A.3 Extensions of current theory . . . 256
A Comparison with regret minimization . . . 258
A Experiment details . . . 259
A.1 Datasets . . . 259
A.2 Additional figures . . . 260
References . . . 263

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].

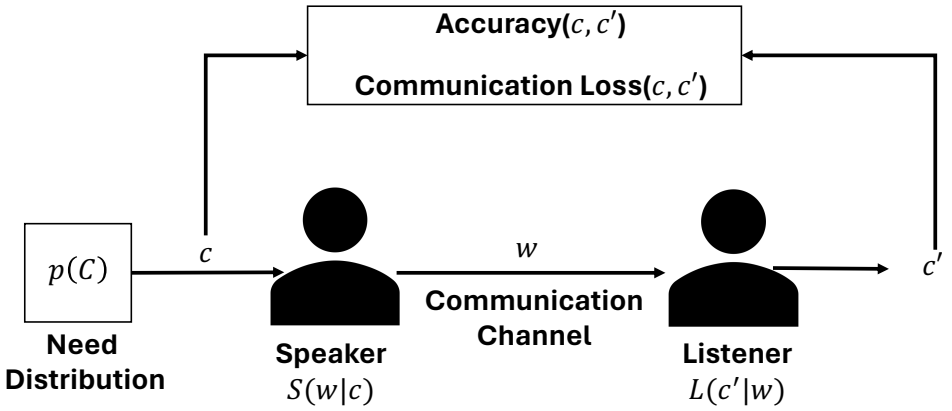


Figure 3.1: The efficiency of semantic categories, or naming systems, is usually studied in a communication setup grounded in Shannon's information theory. A concept is drawn from a need distribution over possible concepts and given to a speaker. The speaker acts as an encoder and encodes the concept 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.

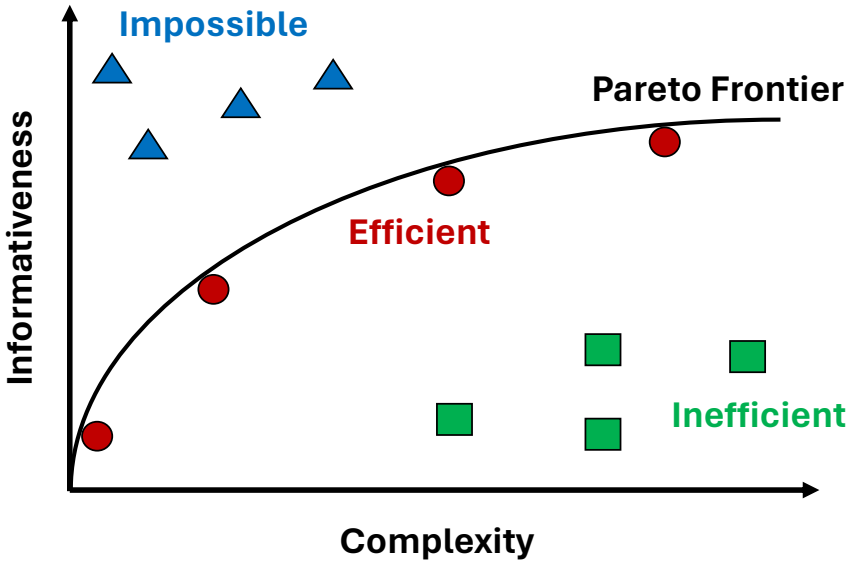


Figure 3.2: Illustration of the trade-off between complexity and accuracy studied in, e.g., Kemp, Xu, et al. (2018) and Zaslavsky et al. (2018). The Pareto frontier corresponds to the languages that achieves highest possible informativeness given a fixed level of complexity. Thus, it is not possible 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

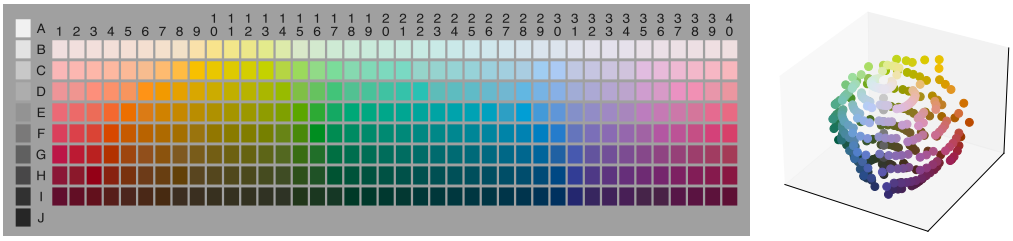


Figure 3.3: (**left**) The Munsell chart used to collect the WCS data. (**right**) Color chips from the Munsell chart represented in 3 dimensional CIELAB space.

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' },

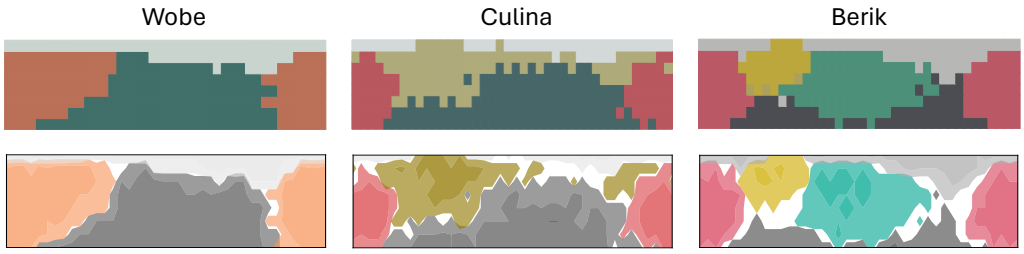


Figure 3.4: Wobe (Ivory Coast), Culina (South America), and Berik (Indonesia) are languages in the WCS data with different numbers of color terms. The top row illustrate the mode map of each language relative to the Munsell chart. That is, each color chip in the Munsell chart is assigned the word most 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)).

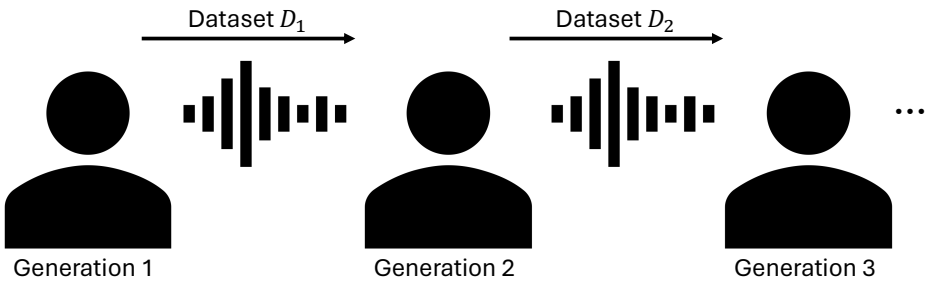


Figure 3.5: In iterated learning, one generation of agents learn their language from a finite dataset generated from the previous generation. This generation then produces a new dataset that is passed to the next generation.

[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.

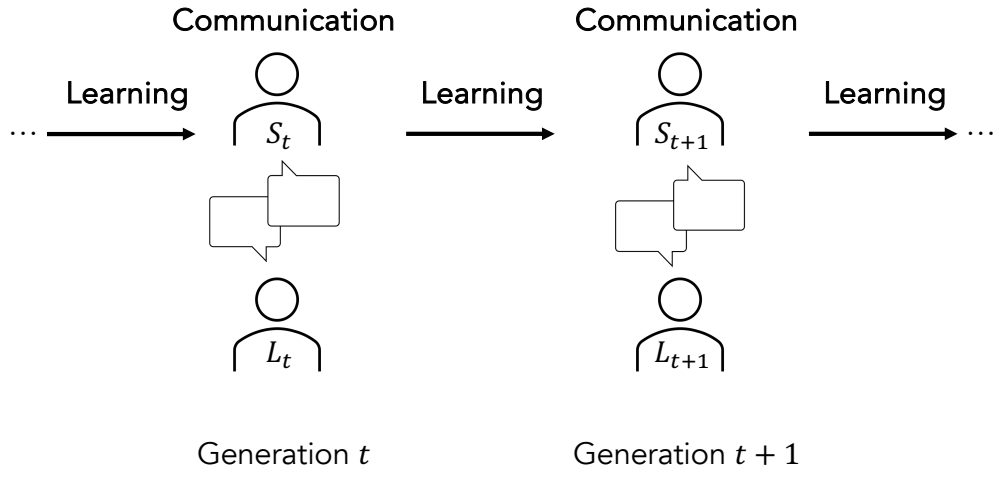


Figure 3.6: Illustration of the NIL algorithm (Ren et al. 2020). The algorithm alternates between communication within a generation, and learning across generations.

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.

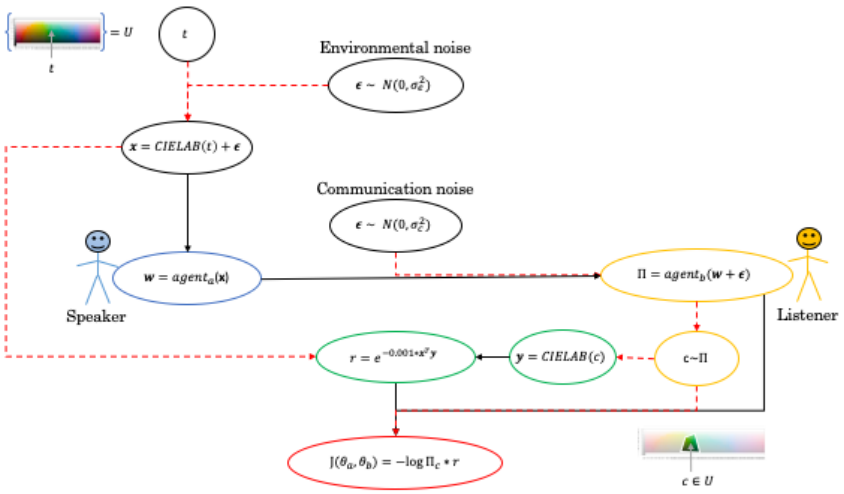


Figure 4.1: The communication setup considered in Kågebäck et al. (2020).

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.

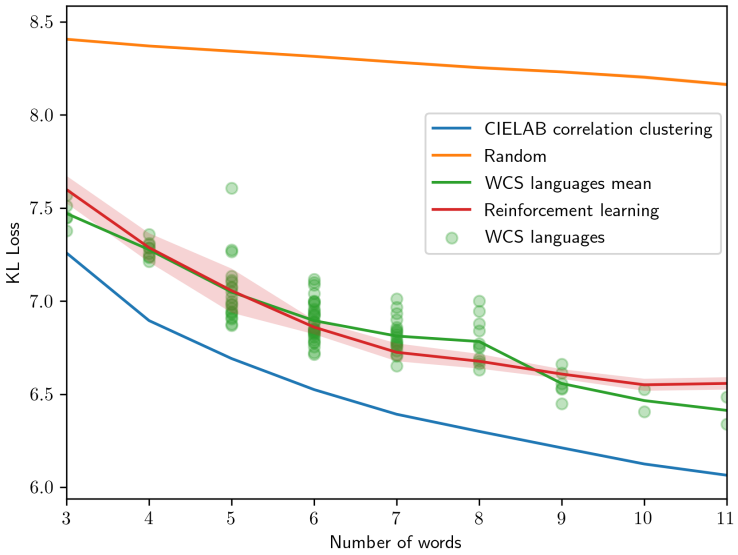


Figure 4.2: Trade-off between communication loss and vocabulary size. The Pareto frontier is estimated using correlation clustering in CIELAB space. We observe that our agents (the line corresponding to reinforcement learning) are able to develop a color naming system, from just maximizing reward, that 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

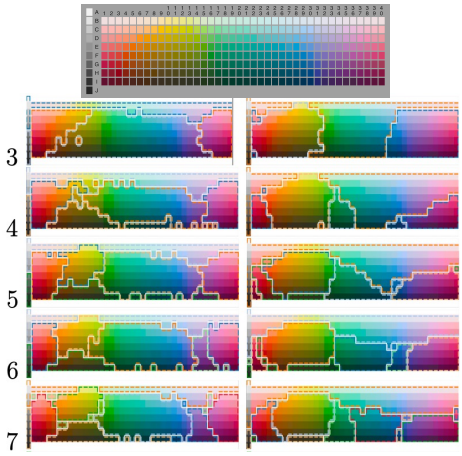


Figure 4.3: The top grid is the Munsell chart used to collect the WCS data. The left column corresponds to human languages with different number of color words while the right column corresponds to artificial naming systems produced by our reinforcement learning agents. Each colored line in a grid corresponds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

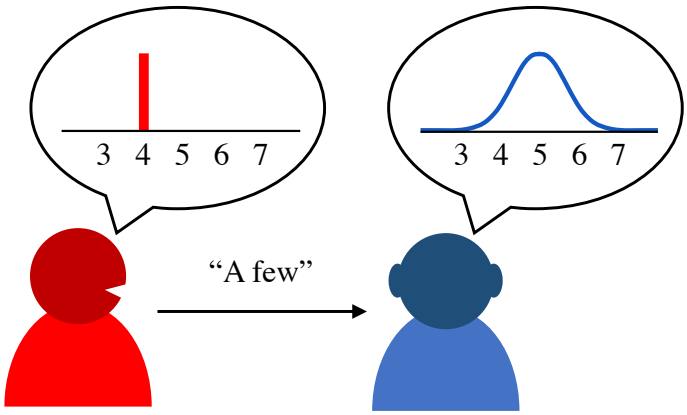


Figure 4.4: The communication model considered here and also by Xu, Liu, et al. (2020). The sender wants to convey the numeral concept 4 and utters "a few". The listener is unsure of which numeral the sender is referring to and produces a probability distribution over possible numerals.

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

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

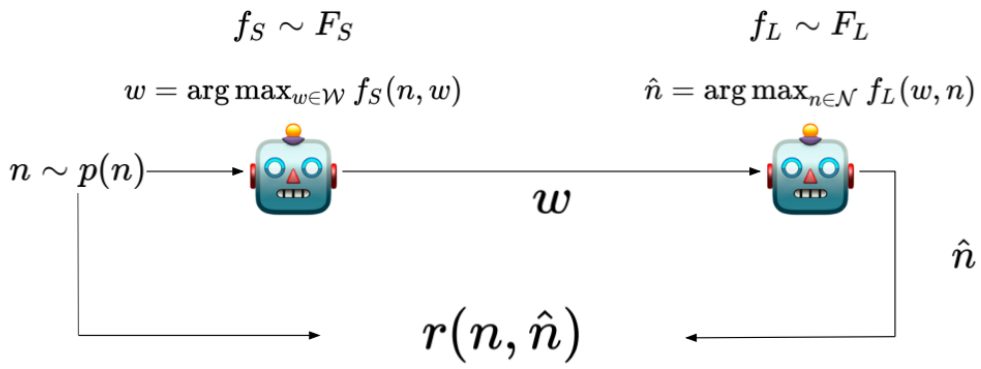
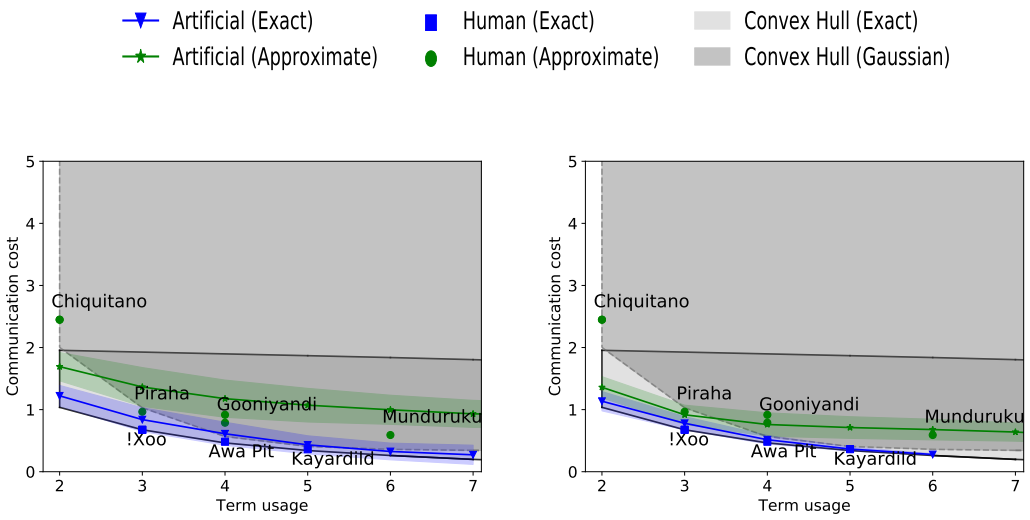


Figure 4.5: At each round of the game, the agents sample smaller networks, f_S and f_L, and using dropout, i.e. some neurons in the larger networks are ignored with a certain probability. This can be viewed as sampling from a belief distribution (Gal and Ghahramani 2016). After this , The speaker is 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).

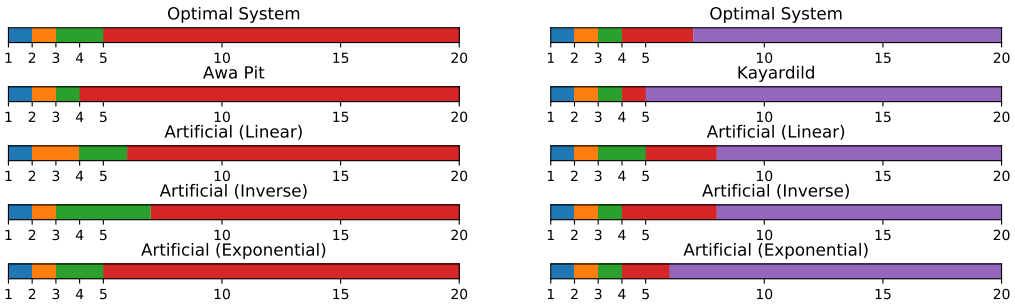


Figure 4.7: Comparison between the optimal numeral systems w.r.t. communication cost, human systems and the artificial systems produced by our agents. Each color represents a numeral word and the corresponding interval on the number line that the word represents.

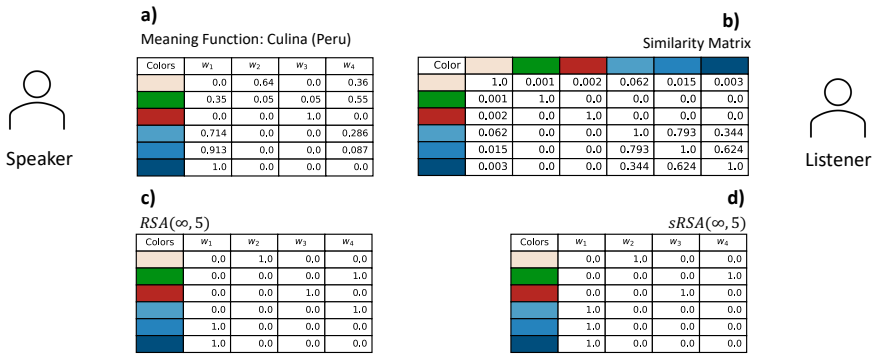
4.3 Paper 3: Pragmatic reasoning in structured signaling games

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.

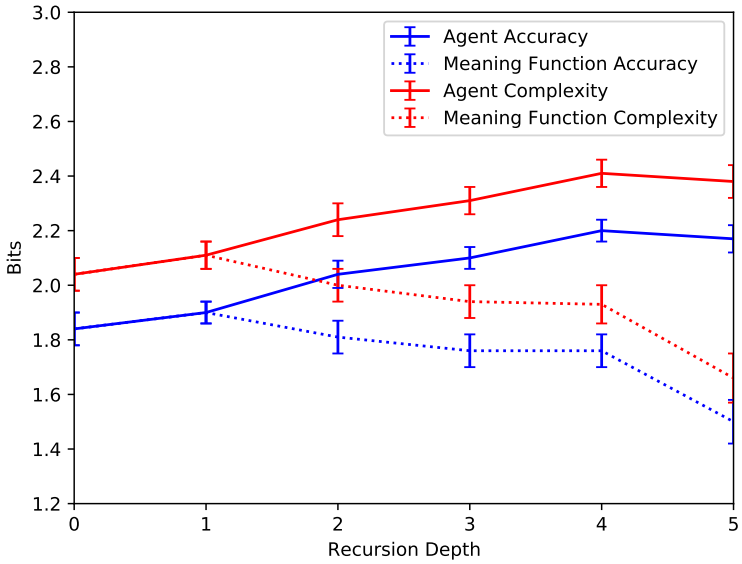
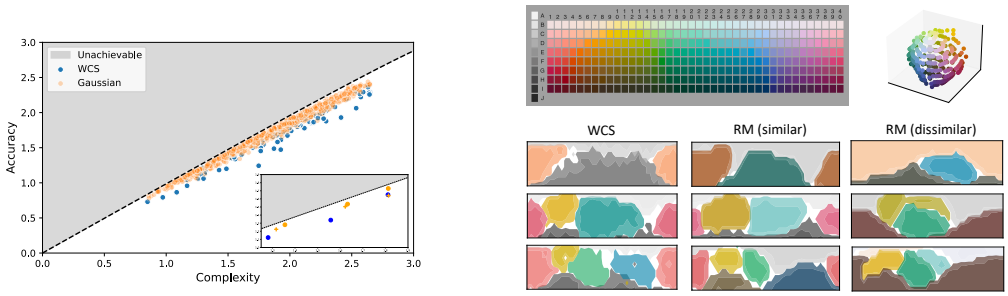


Figure 4.9: The complexity and accuracy of the sRSA agents increase with recursion depth, while the complexity and accuracy of the corresponding meaning functions decrease. Hence, as the reasoning depth increases, the ambiguity of the learned meaning function increases. Depth indicates the level of the 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.

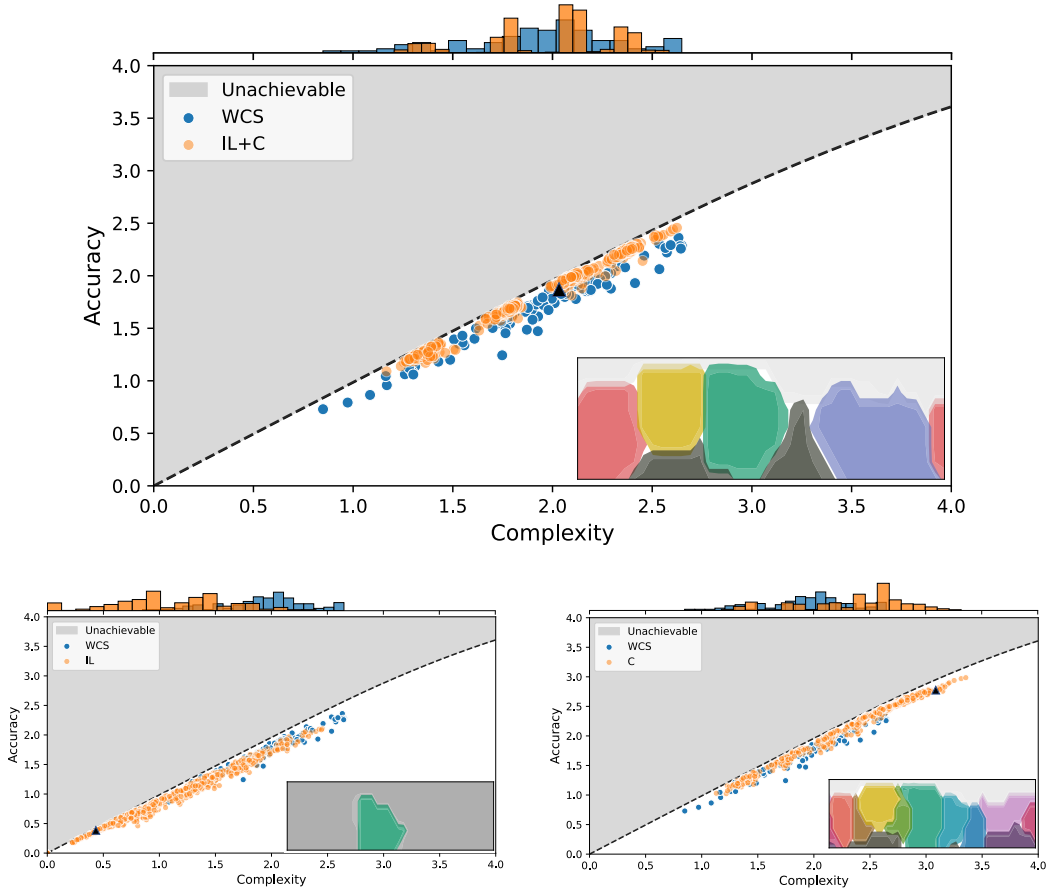


Figure 4.11: Efficiency of the (top) IL+C, (bottom left) IL, and (bottom right) C evolved color naming systems (orange dots), in each case compared with the natural systems of the WCS (blue dots). The black triangle indicates the end state of one run, shown in the inset color map. The histograms above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

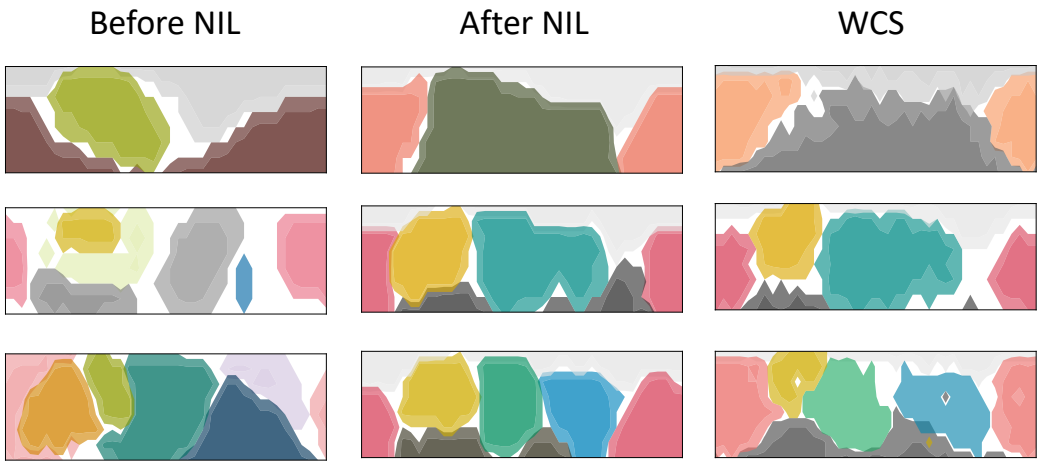


Figure 4.12: IL+C transforms efficient color naming systems to become more similar to the WCS. In each row, the left column shows a randomly generated efficient system that was used to initialize the first generation, the middle column shows the result of running NIL from that initialization state, and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

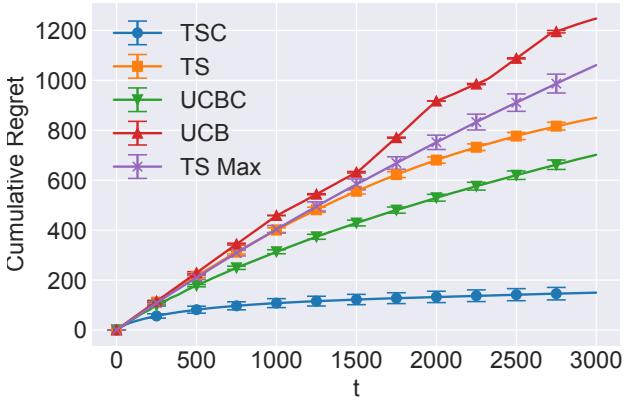


Figure 4.13: An instance with 1000 arms, 32 clusters and 32 arms in each cluster. TSC is our approach. TS (Thompson sampling) and UCB (upper confidence bounds) are algorithms suited for the standard multi-armed bandit. UCBC (Pandey et al. 2007; Bouneffouf et al. 2019) and TSMax (Zhao et al. 2019) are 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.

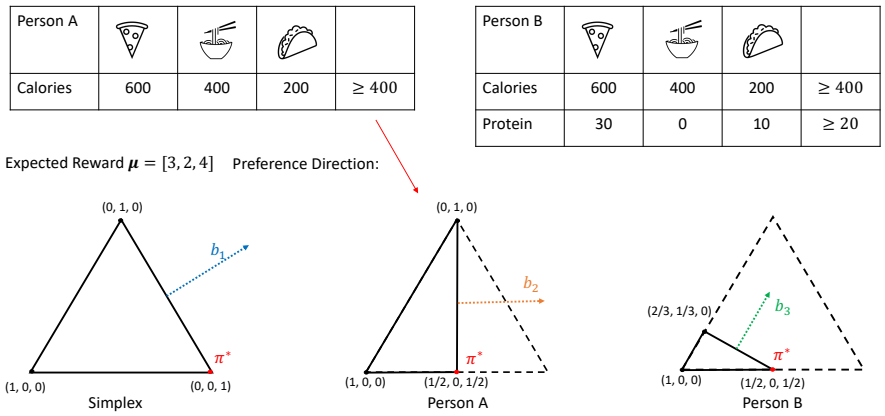


Figure 4.14: Two people, A and B, are searching for a meal plan π that maximizes taste, i.e., expected reward $\mu^\top \pi$, while satisfying some nutrition constraints. Without any constraints this setting reduces to BAI and can be viewed as searching for the optimal policy over the probability simplex. 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.

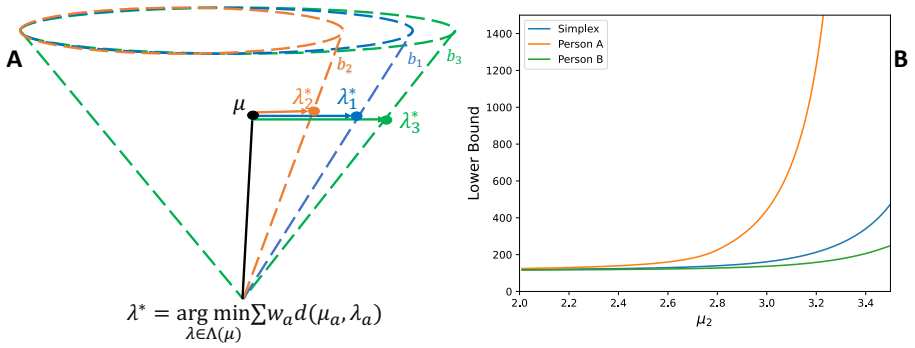


Figure 4.15: Computing the λ satisfying Equation 4.6.3, i.e. the *most confusing instance*, can be viewed as an information-theoretic projection onto the boundary of the normal cone spanned by the active constraints at π_μ. In A) we see the different normal cones for the three different examples in Figure 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.

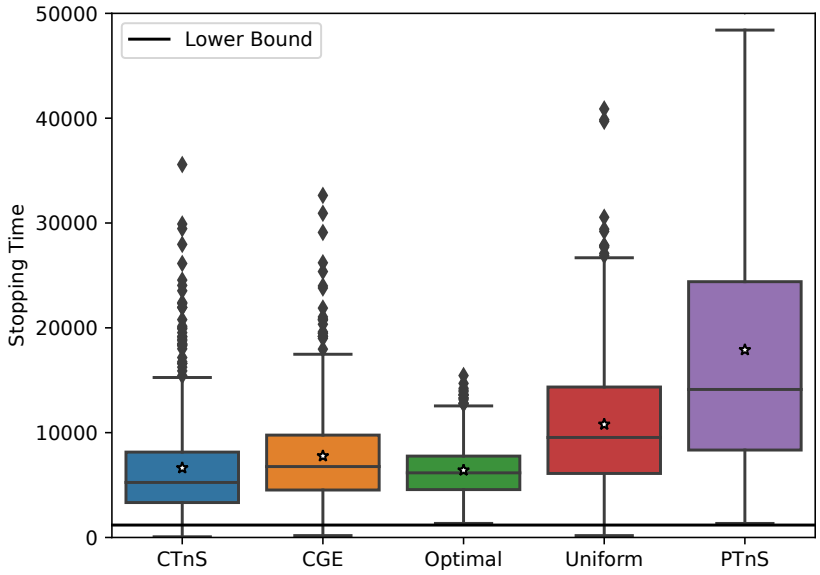


Figure 4.16: Y-axis corresponds to the time (number of samples) until an algorithm stops and outputs the best policy with confidence $1-\delta$. The figure illustrates the sample complexity of our algorithms (CTnS and CGE) against baselines on a problem where to goal is to find the optimal allocation 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

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

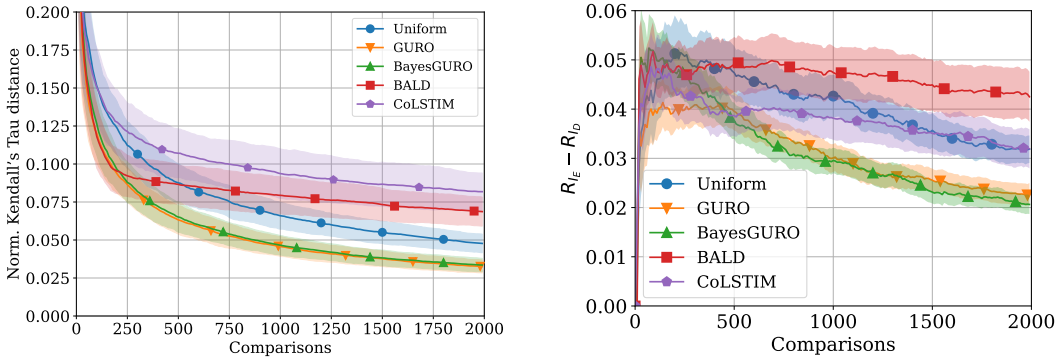
where $z_{ij} := x_i - x_j$. Our data-dependent bound suggests that the probability that the ordering error exceeds some $\epsilon > 0$ after collecting a dataset, D_T, of T comparisons is upper bounded as [3]

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

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).

Part II

Appended Papers

Paper 1

A reinforcement-learning approach to efficient communication

Mikael Kågebäck,Emil Carlsson, Devdatt Dubhashi, Asad Sayeed.

PLoS ONE, 15(7):1–26, 2020.

The paper has been reformatted for uniformity.

Paper 1. A reinforcement-learning approach to efficient communication

Mikael Kågebäck,Emil Carlsson, Devdatt Dubhashi, Asad Sayeed.

Abstract

We present a multi-agent computational approach to partitioning semantic spaces using reinforcement-learning (RL). Two agents communicate using a finite linguistic vocabulary in order to convey a concept. This is tested in the color domain, and a natural reinforcement learning mechanism is shown to converge to a scheme that achieves a near-optimal trade-off of simplicity versus communication efficiency. Results are presented both on the communication efficiency as well as on analyses of the resulting partitions of the color space. The effect of varying environmental factors such as noise is also studied. These results suggest that RL offers a powerful and flexible computational framework that can contribute to the development of communication schemes for color names that are near-optimal in an information-theoretic sense and may shape color-naming systems across languages. Our approach is not specific to color and can be used to explore cross-language variation in other semantic domains. [1]

1 Introduction

The study of word meanings across languages has traditionally served as an arena for exploring which categorical groupings of fine grained meanings tend to recur across languages, and which do not, and for deriving on that basis a set of generalizations governing cross-language semantic variation in a given domain.

There is a long history of proposals that attempt to characterize how humans manage the effort of communication and understanding (Zipf 1949) and how this management can be affected by environmental demands (Baddeley and Attewell 2009). One such increasingly influential proposal is that language is shaped by the need for *efficient communication* (Kemp et al. 2018; Regier, Kemp, et al. 2015; Gibson et al. 2017; Piantadosi et al. 2011; Jameson and D'Andrade 1997), which by its nature involves a trade-off (Kirby et al. 2015; Carr et al. under review 2018) between simplicity, which minimizes cognitive load, and informativeness which maximizes communication effectiveness. Specifically, they propose that good systems of categories have a near-optimal trade-off between these constraints. This trade-off is couched in the classic setting of Shannon information theory (Cover and Thomas 2006) which considers the fundamental laws of transmitting information over a noisy

[1] Code available at: `https://github.com/kageback/colorwords`

channel. Examples formalized in information-theoretic terms include suggestions that word frequency distributions, syllable durations, word lengths, syntactic structures, and case marking all facilitate efficient communication (see (Kemp et al. 2018; Regier, Kemp, et al. 2015) and references cited therein). The information theoretic view leads naturally to view the symbolic linguistic terms used for the communication as *codes* that create *partitions* of semantic spaces.

Given the principle of efficient communication, a fundamental challenge is to seek a concrete computational mechanism that could lead to optimal or near optimal communication schemes. Here we propose *Reinforcement learning* (RL) as a potential computational mechanism that may contribute to the development of efficient communication systems. Various systems, both artificial and in nature, can be represented in terms of the way they learn environmental interaction strategies that are near-optimal using RL techniques that employ reward/punishment schemas (Sutton and Barto 1998; Wiering and Otterlo 2012; Dayan and Niv 2008). RL's basis in operations research and mathematical psychology and ability to provide quantitative and qualitative models means it can be applied to a wide range of areas(Dayan and Niv 2008).

RL appears to be transparently implemented in neural mechanisms, for example, in dopamine neuron activity. For this reason, RL is increasingly recognized as having scientific value beyond mere computational modeling of decision-making processes (Dayan and Niv 2008; Niv 2009; Niv and Langdon 2016). That RL appears to be biologically so well-embedded implies that it can be seen as a general cognitive mechanism and used in an immediate way to make hypotheses about and interpretations of a variety of data collected in behavioral and neuroscientific studies.

The availability of a growing suite of environments (from simulated robots to Atari games), toolkits, and sites for comparing and reproducing results about RL algorithms applied to a variety of tasks (Lazaridou et al. 2016; Foerster et al. 2016; Havrylov and Titov 2017; Evtimova et al. 2017; Jorge et al. 2016) makes it possible to study cognitive science questions through a different lens using RL. Cognitive science experiments are often carried out in real life settings involving questionnaires and surveys that are costly and sometimes suffer from high variability in responses. If simple RL algorithms are indeed a good proxy for actual human learning, then insights about questions of universals in language learning could be obtained very cheaply and reliably via controlled experiments in such *in silico* settings. Our approach could be used to explore various trade-offs at the heart of efficient communication (Kemp et al. 2018). Some languages are *simple* i.e. have few color terms while others have more color terms and are hence more *informative*. There is a tradeoff between these two properties and our framework can be used to test the prediction that human semantic systems will tend to lie along or near the optimal frontier of the region of achievable efficiency in communication systems as depicted schematically in Figure 1.1, see also (Kemp et al. 2018; Carr 2019) for more discussion on this. Representing the question as an accuracy vs. complexity tradeoff specific to the domain of color terms, Zaslavsky et al. (Zaslavsky et al. 2018) demonstrate that a number of human languages, English included, come very close to that frontier. As pointed out by a referee, it is interesting to compare the approach here to Zaslavksy

et al (Zaslavsky et al. 2018) who derive efficient naming systems as solutions to a differential equation implied by an information bottleneck (IB) loss function with terms to maximize information transfer and minimize lexicon complexity (which works out to, essentially, lexicon size). In contrast, this work considers a setting where two RL agents communicate through a noisy channel about color tiles, also observed through a noisy channel, and have to eventually agree on a communication protocol. The RL agents' reward function is based on a similarity function in CIELAB space. We show that the resulting communication protocols satisfy the same efficiency measures that were used to define the information bottleneck, although the system was not explicitly optimized for these quantities. The environmental and communication noise rate ends up playing a similar role to the complexity penalty in the IB formulation (although with different dynamics over time), by reducing lexicon size. Thus, the two approaches are complementary: while the IB principle offers a descriptive analysis and establishes fundamental information–theoretic limits on the efficiency and complexity of communication schemes our approach is an algorithmic prescriptive route to how such optimal or near optimal schemes could be obtained.

While there may be reason to think that RL has a deep biological basis, in this work, we do not focus on the specifics of the underlying neurocognitive mechanism. Rather we demonstrate that very simple RL mechanisms are able to generate partitions for efficient (and near optimal) communication. We demonstrate this with a focus on questions about the universality of color categories and words in language. While there has been previous work (Baronchelli et al. 2010) on computational mechanisms involving naming games for the emergence of universality of color names, our work is the first to provide a mechanism based on a fundamental machine learning paradigm (reinforcement learning) that is also biologically plausible.

1.1 Linguistic background on color identification

A theory of universals Color naming universals have a long history in linguistic research (Berlin and Kay 1969). At an individual level, color perception is subjective; it differs for biological reasons across individuals (extreme examples being colorblindness and tetrachromacy). There are commonly-observed differences in individual color-naming choices. What is "turquoise" to one person may be a variant of "blue" to another. Nevertheless, within the same linguistic milieu, there is overall agreement as to color-naming; most English-speaking people recognize the typical supermarket tomato as "red".

Berlin and Kay showed across a survey of 20 languages that there are strong consistencies in color naming and produced a set of universals: e.g., there are a maximum of eleven major color categories and, where fewer than eleven are realized for a given language, there is a standard pattern of emergence. This work came under methodological criticism (Lucy 1997; Saunders 1995), particularly the use of standardized color systems to abstract away from the interactional and cultural basis of color identification.

Given this methodological conflict, is it really the case that such universals are artifacts of methods of investigation that take color communication out of its

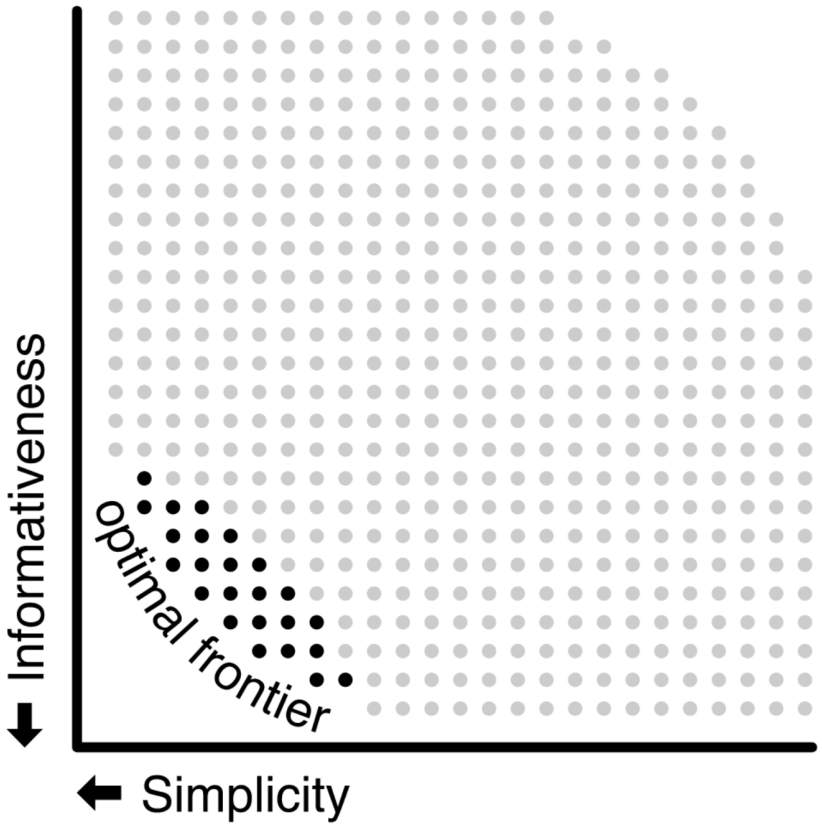


Figure 1.1: Human semantic systems will tend to lie along this optimal frontier of achievable efficiency in communication systems. Reprinted with permission from (Carr 2019).

natural context in human linguistic interaction? Accounting for patterns of wide but constrained variation that have been observed empirically is a central challenge in understanding why languages have the particular forms they do.

Color terms represent a limited semantic domain with easily manipulated parameters. By gradual changes of color value, an experimenter can manipulate red into orange, unlike other semantic domains, where the distinctions between potential referents (e.g., "car" vs. "truck") are not easily captured in explicit terms. In addition, recent work (Regier, Kemp, et al. 2015; Gibson et al. 2017) argues that color categories in language should support efficient communication.

Color naming models Developed in 1905, the Munsell color system uses three color dimensions (hue, value, and chroma) to represent colors based on an "equidistance" metric calibrated experimentally by Albert Munsell. The World Color Survey (WCS; e.g. figure 1.2) uses the Munsell color system in a matrix arranged by 40 hues, 8 values (lightness), and at maximum chroma (saturation). A color map can be developed for a particular language by asking speakers of that language to name each color. Color identification boundaries can be compared across languages using the WCS mapping.

The WCS color map technique enables the testing of automatic systems to

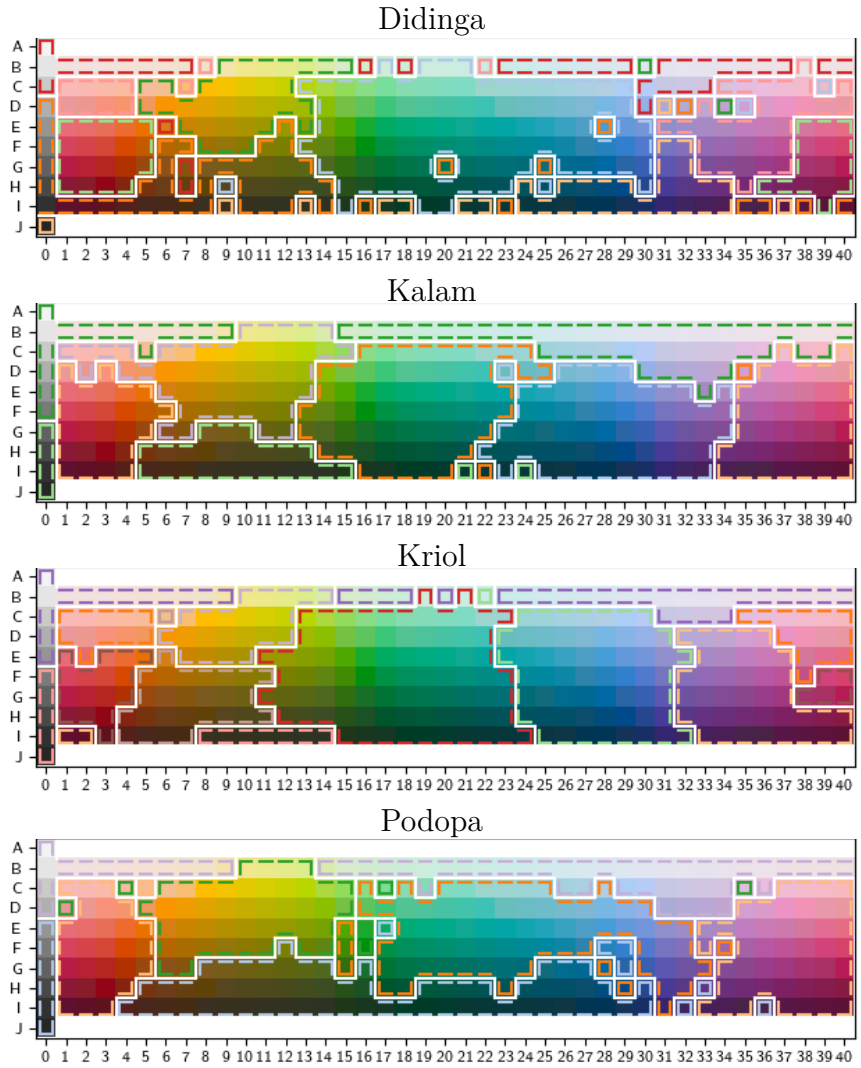


Figure 1.2: Reproduced color mode maps (Regier, Kemp, et al. 2015) corresponding to four randomly selected languages of the WCS.

partition colors. Regier et al. (Regier, Kay, et al. 2007) experiment with partitioning the color space using a distance metric as a clustering criterion. They find a good distance metric by translating the WCS color map to the CIELAB space. CIELAB enables the translation of the WCS colors to a three-dimensional space, wherein the WCS colors appear to take an irregular spherical form. Regier et al. then use a standard "well–formedness" metric, which is essentially placing similar colors together and dissimilar apart. (Technically, this is called correlation clustering, which we explain later in the paper.) This allows them to automatically construct color partitions in the CIELAB space. Regier et al. find correspondences between optimal color partitions and observed color maps from human surveys as well as determine that rotating the WCS color space for a given observed color map causes reduced well–formedness in the corresponding CIELAB space. This is preliminary evidence for the optimality of color spaces in human language in relation to a well–formedness trade-off statistic.

Following their earlier work, Regier et al. adopt an information-theoretic approach (Regier, Kemp, et al. 2015) by introducing a communication system between two agents for multiple semantic domains (including color) and the corresponding notion of reconstruction error as the relative entropy (Kullback-Leibler divergence). The relative entropy is computed between the speaker's model and the listener's model of the probability that a particular term encodes a particular color. This becomes the communication cost of a color labeling system. They then show that real-world color-naming systems not only tend to have high well–formedness, but they also have low communication cost. A similar framework is adopted by Gibson et al. (Gibson et al. 2017).

1.2 Approach and contributions

This work focuses specifically on the role of speaker-listener communication efficiency in the partitioning of color spaces. To this end, we set up a two-agent paradigm closely mirroring the information-theoretic frameworks (Kemp et al. 2018; Regier, Kemp, et al. 2015; Gibson et al. 2017) that represent a series of negotiations between speaker and listener in the context of a "game". Agent-based simulations are widely used in the study of the development of communication systems, including color communication (Steels and Belpaeme 2005; Belpaeme and Bleys 2005; Baronchelli et al. 2010; Jäger and Rooij 2007). The basic paradigm used in our work is one in which the speaker and listener both begin with a set of available words (represented as integer identifiers) associated with a map of color "tiles", where regions on the map are represented by the words. However, the speaker and the listener have different randomly-initialized maps. The speaker agent chooses a color tile and sends the listener agent the word that represents the region in which the tile is located. The listener agent then selects a tile that is in the region from its own map that most likely to be represented by that word in the speaker map. A reinforcement learning paradigm is used, as above, to update the parameters representing the shape of the maps, so that the game is run over many iterations.

This approach is a highly constrained representation of the "real-world" scenario of many speakers negotiating meaning in a speech community. Constrained simulations of communicative phenomena can allow the identification of plausible hypotheses about the factors that affect the corresponding real-world scenario, assuming that at least part of the expected behavior is reflected in the simulation.

In this work, we find that our two-agent simulation closely tracks the behavior of the languages in the World Color Survey in terms of both communication efficiency and perceptual well–formedness, relative to the number of primary color terms used. These are clearly separable from a random baseline and an idealized color map based on the CIELAB color space. Furthermore, the similarity of the color maps derived from the two-agent setting to the WCS maps remains relatively stable as the number of words are varied. We vary other metrics, such as perceptual and communication noise, to make predictions about color term convergence and demonstrate the flexibility of the model. The naturalness and stability of the model are evidence that our agent simulation paradigm is a suitable setting for

investigation and hypothesis generation about cognitive and environmental effects on color communication in linguistic settings.

Enabled by recent advances in deep reinforcement learning (Lazaridou et al. 2016; Havrylov and Titov 2017; Evtimova et al. 2017; Jorge et al. 2016), this work therefore makes a methodological contribution to the study of the development of meaning in human languages given communicative factors. Our approach can offer complementary insight to the recent approach of Zaslavsky et al. (Zaslavsky et al. 2018) who argued that languages efficiently compress ideas into words by optimizing the *information bottleneck* (IB) trade-off between the complexity and accuracy of the lexicon.

2 Efficient communication: A theoretical framework

The color game

We adopt a previously proposed (Regier, Kemp, et al. 2015) general communication framework which takes an information-theoretic perspective via a scheme involving a speaker and a listener communicating over a noisy channel. The speaker attempts to communicate a color from the domain of colors U. The speaker wishes to communicate about a specific color $c \in U$, and she represents that object in her own mind as a probability distribution

$$s = \delta(c) \tag{2.1}$$

over the universe U, with mass concentrated at c. The speaker then utters the word w using a policy corresponding to a distribution $p(w \mid c)$ to convey this mental representation to the listener. Upon hearing this word, the listener attempts to reconstruct the speaker's mental representation (s) using information conveyed in the word used by the speaker. The listener reconstruction is in turn represented by the probability distribution

$$\ell = p(c|w), c \in U \tag{2.2}$$

To enable us to later compare artificial languages to real languages, we will now define a number of efficiency measures that has previously been shown to be important for human languages (Regier, Kemp, et al. 2015; Gibson et al. 2017).

2.1 Information-theoretic communication loss

Though the goal of the communication game is to perfectly transmit information, there are several challenges (e.g., limited vocabulary, noisy limited-bandwidth communication medium, and differences in word definitions between speakers) that make this goal impossible in reality. We take a semantic system to be informative to the extent that it yields low communication cost which can be estimated using one of the following related methods.

Expected surprise based on empirical estimation

The information loss can be defined as the listener's expected surprise (Gibson et al. 2017), i.e., the surprise incurred by the listener when the actual color tile that the sender encoded as a word is revealed to the listener. The expected surprise for one color tile c is computed as

$$E_c^{ES} := -\sum_{w \in W} p(w|c) \log_2 p(c|w). \tag{2.3}$$

The probability distribution $p(w|c)$ can be obtained in several different ways. In Gibson et al. (Gibson et al. 2017), $p(w|c)$ was empirically estimated from the WCS data by computing the fraction of respondents that choose to use a particular word for a given tile c, however, when evaluating artificial languages this is not always as easy. Fortunately, we can query the artificial agents after training, in analog to the WCS interviews, to estimate $p(w|c)$. Finally, rather then separately estimating $p(c|w)$, this can be computed using Bayes theorem as

$$p(c|w) = \frac{p(w|c)p(c)}{\sum_{c' \in U} p(w|c')p(c')} \tag{2.4}$$

where $p(c)$ is taken to be uniform. In this case $p(c|w)$ can be seen as a Bayesian decoder.

KL divergence using mode map based estimation

An alternative approach, suggests the use of the KL divergence between the speaker distribution s and the listener distribution l (Regier, Kemp, et al. 2015), i.e.,

$$E_c^{KL} = D_{KL}(s(c)||l(w)), \tag{2.5}$$

as the measure of information loss. In the case of discrete distributions, where s has all its probability mass concentrated on one meaning, and $l(w) = p(c|w)$ this becomes

$$E_c^{KL} = -\log_2 p(c|w). \tag{2.6}$$

Though $p(c|w)$ can be estimated empirically for, e.g., the WCS data, it may also be computed directly from a color space partitioning (Regier, Kemp, et al. 2015). This method gives us a measure of the communication cost of using a given semantic system to communicate about this domain, i.e., the distributions are derived from a mode map over U. More specifically, $p(c|w)$ is computed as

$$p(c|w) = \frac{\sum_{j \in Cat(c)} sim(c,j)}{\sum_{i \in U} \sum_{j \in Cat(i)} sim(i,j)} \tag{2.7}$$

which is motivated by an exemplar selection argument (i.e., from a category); one tends to select the most representative exemplar. $Cat(c)$ refers to the category/partition that c belongs to, and $sim(i,j)$ measures the similarity between two

colors i and j which is standard in these studies as in Regier et al. (Regier, Kay, et al. 2007):

$$sim(i,j) := \exp\left(-c * dist(i,j)^2\right), \tag{2.8}$$

In equation 2.8, the CIELAB distance is represented as $dist(x,y)$ for colors x and y. In all the simulations we report, we set c, the scaling factor, to 0.001 as in Regier et al. (Regier, Kay, et al. 2007)[2]. When $x = y$ (identical chips), the maximum value 1 is attained. As the distance between the chips grows, the value of the function falls rapidly to 0. What does this mean in qualitative terms? It means that there is a point at which the colors look so different that no noticeable additional dissimilarity effect can be distinguished.

It is interesting to note that if $p(w|c)$ is taken to be a distribution with all its probability mass concentrated on the word that corresponds to the partition that c belongs to (which is natural given how the distribution s is constructed), then E_c^{KL} can be derived from E_c^{ES} as $E_c^{ES} - \sum_{w \in W} p(w|c) \log_2 p(c|w) = -\log_2 p(c|w) = E_c^{KL}$. Hence, the main difference between the two is how the distributions are estimated.

Aggregate measure of the communication cost

To get an aggregate measure of the reconstruction error over all colors in the domain universe of colors, we compute the expected communication cost[3] incurred when transferring color information between two agents over a linguistic communication channel as

$$E := \sum_{c \in U} p(c) E_c. \tag{2.9}$$

Where E_c corresponds to either E_c^{KL} or E_c^{ES} and the need probability $p(c)$ may be taken to be uniform(Regier, Kemp, et al. 2015; Gibson et al. 2017) or more informed(Zaslavsky et al. 2018). However, all experiments in this paper use a uniform need probability.

2.2 Well–formedness

A different criterion for evaluating the quality of a partition of the color space is the so-called *well–formedness* criterion (Regier, Kay, et al. 2007). In fact this criterion is exactly the same as the maximizing agreements objective of the *correlation clustering* problem discussed extensively in the theoretical computer science literature (Bansal et al. 2004; Demaine et al. 2006). Given the CIELAB similarity measure, we consider a graph G on the tiles and assign the weight $sim(x,y) - \frac{1}{2}$ on the edge connecting tiles x and y. Thus similar tiles (with similarity exceeding 1/2) will have a positive weight while dissimilar tiles (with similarity less than 1/2) will carry negative weights on the corresponding edges. The objective is then to find a clustering to maximize the weights on edges within clusters. For a given partition, we can compute this sum over all intra-cluster edges and compare it to the optimum over all partitions. While

[2]As pointed out by a reviewer, the similarity function (sim) may be interpreted as a Gaussian likelihood in CIELAB space with variance defined by s.

[3]It was noted by a reviewer that this measure is equivalent to the conditional entropy $H[C|W]$.

this optimum may be approximated using an heuristic approach (Regier, Kay, et al. 2007), we have used an algorithm with guaranteed convergence to optima.

2.3 Reinforcement learning framework for communication over a noisy channel

We develop a version of the general communication setup, i.e. The color game, as two automated agents trained via reinforcement learning. Our framework offers two different training approaches.

In the first training approach the agents are allowed to use continuous real valued messages during training in order to enable faster training. After training the agents are however evaulated using discrete messages. In the second approach the agents are both trained and evaluated using discrete messages.

An overview of the model trained with continuous real valued messages is shown in Fig 1.3, the model trained with discrete messages is shown in Fig 1.4. Note that the main difference between the training approaches is whether the communication channel is differentiable, black solid arrows, or not, red dashed arrows.

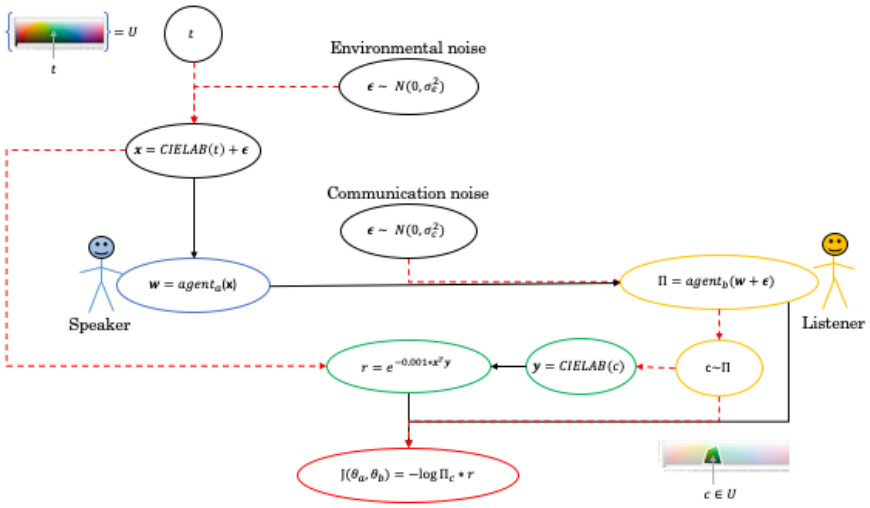


Figure 1.3: An overview showing each computation step in the model, while using continuous real valued messages during training, for one instance of the color game. Black solid arrows indicate a differentiable relation while red dashed arrows indicates a non-differentiable relation. The color of the ovals are used to highlight the different parts of the model where black is the model input, blue and yellow the agents, green the reward system, and red the reinforce cost function.

It turns out that training with discrete messages is more time consuming and it becomes harder for the agents to converge and agree on a certain color partition. Most our analysis will therefore be with respect to agents trained with continuous real valued messages and it can be assumed that continuous real valued messages was used during training if nothing else is stated. However, we also provide a section

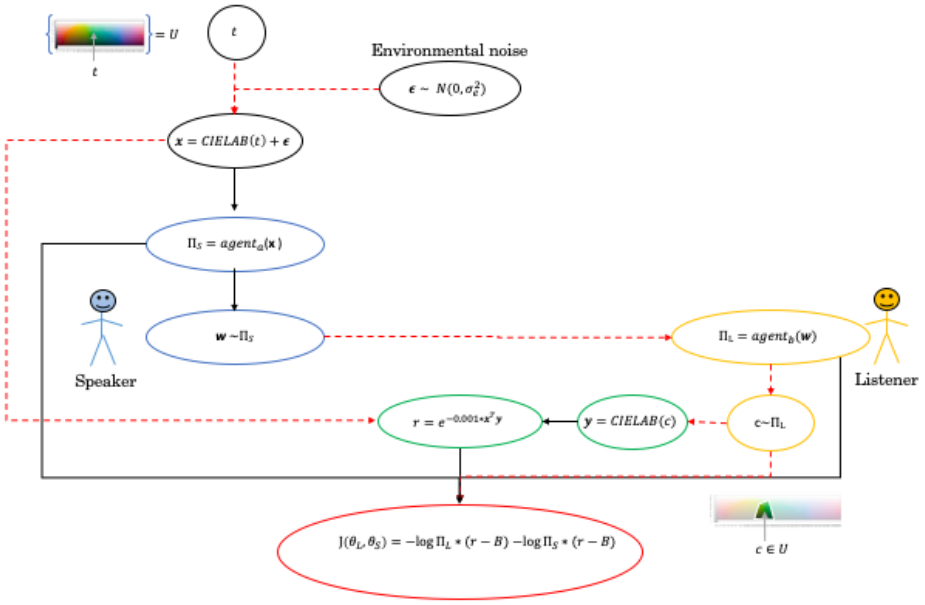


Figure 1.4: An overview showing each computation step in the model, while using discrete messages during training, for one instance of the color game. Black solid arrows indicate a differentiable relation while red dashed arrows indicates a non-differentiable relation. The color of the ovals are used to highlight the different parts of the model where black is the model input, blue and yellow the agents, green the reward system, and red the reinforce cost function.

where we compare a limited number of experiments ran with discrete messages to their corresponding continuous real valued counterpart.

Continuous policy

The sender trying to communicate the target color $t \in U$ creates a word vector

$$\mathbf{w} = \text{softmax}(\phi_s^T \text{ReLU}(\theta_s^T [\text{CIELAB}(t) + \epsilon_\mathbf{e}])), \epsilon_\mathbf{e} \sim N(0, \sigma_e^2). \tag{2.10}$$

where $\text{softmax}_j(\mathbf{z}) = e^{z_j} / \sum_i^{|\mathbf{z}|} e^{z_i}$, $\text{ReLU}(z) = \max(0, z)$, $\{\phi_s, \theta_s\}$ are the parameters of the sender agent, and $\epsilon_\mathbf{e}$ model environment noise. $\mathbf{w}$ is subsequently sent to the listener agent over a noisy communication channel as

$$\mathbf{m} = \mathbf{w} + \epsilon_\mathbf{c}, \epsilon_\mathbf{c} \sim N(0, \sigma_c^2). \tag{2.11}$$

Please note that, though this message will start out as a continuous real valued message the noise will make it converge, as training goes on, to a peaked distribution with almost all probability mass concentrated to one dimension for each color (Foerster et al. 2016). Further, when we extract the final resulting language we use discrete m vectors as, i.e. where all dimensions but one is zero, to ensure that no extra information is encoded.

The receiver interprets the message received ($\mathbf{m}$) and computes a distribution over all colors in U as

$$p(U|\mathbf{m}) = \text{softmax}(\phi_r^T \text{ReLU}(\theta_r^T \mathbf{m})). \tag{2.12}$$

By now merging Equation (2.10), (2.11), and (2.12) we get the final policy

$$\Pi_\Omega(U|t) = p(U|t) \tag{2.13}$$

where $\Omega = \{\theta_s \in \mathbb{R}^{k\times 3}, \phi_s \in \mathbb{R}^{d\times k}, \theta_r \in \mathbb{R}^{k\times d}, \phi_r \in \mathbb{R}^{|U|\times k}\}$ parameterizes the entire model. The sender and receiver agents are modeled using *multilayer perceptrons*, bias terms have been omitted for brevity, with one hidden layer of $k = 20$ units, and the size of the message vector is set to $d = 50$ for all experiments. Note that d will set the maximum number of color terms that the system can use to 50; however, this is far above what is used in practice and not what will determine the number of terms actually used by the system.

Cost function for continuous policy

Finally, plugging the policy and reward into REINFORCE (Williams 1992) we get the cost function

$$J(\Omega) = -\frac{1}{N_b}\sum_{n}^{N_b} \log \Pi_\Omega(U = c_n|t) * r_n. \tag{2.14}$$

where N_b corresponds to the number of games over which the cost is computed and r_n is the reward detailed in section Reward. For more on REINFORCE please see the section describing Materials and methods.

2.4 Discrete policies

In order to use discrete communication during training a message $\mathbf{m}$, represented as a discrete vector where all but one dimension is equal to zero, is sampled from the categorical distribution over the set of possible color terms

$$\begin{gathered}\mathbf{m} \sim p(W|t) = \text{softmax}(\phi_s^T \text{ReLU}(\theta_s^T[\text{CIELAB}(t) + \epsilon_\mathbf{e}])) \\ \epsilon_\mathbf{e} \sim N(0, \sigma_e^2).\end{gathered} \tag{2.15}$$

Equation (2.15) gives us the policy of the sender

$$\Pi_{\Omega_s}(W|t) = p(W|t) \tag{2.16}$$

where $\Omega_s = \{\theta_s \in \mathbb{R}^{k\times 3}, \phi_s \in \mathbb{R}^{d\times k}\}$, under discrete communication.

Further, the receiver interprets the received message ($\mathbf{m}$) and computes a distribution over all colors U as described in equation (2.12). Hence, the receiver policy becomes

$$\Pi_{\Omega_r}(U|\mathbf{m}) = p(U|\mathbf{m}) \tag{2.17}$$

where $\Omega_r = \{\theta_r \in \mathbb{R}^{k\times d}, \phi_r \in \mathbb{R}^{|U|\times k}\}$.

As in the case with the continuous policy, the sender and reciever will be modelled using *multilayer perceptions* with one hidden layer consisting of $k = 20$ units. The size of the message vector is set to $d = 50$.

Cost functions for discrete policies

Furthermore, due to the non-existence of a gradient over the communication channel we end up with two distinct policies, one for the sender and one for the receiver, which require us to have two cost functions that will be optimized simultaneously.

As a result, the cost function for the sender becomes

$$J(\Omega_s) = -\frac{1}{N_b}\sum_{n}^{N_b} \log \Pi_{\Omega_s}(\mathbf{m}_n|t) * (r_n - B_n) \tag{2.18}$$

and one for the receiver it becomes

$$J(\Omega_r) = -\frac{1}{N_b}\sum_{n}^{N_b} \log \Pi_{\Omega_r}(U = c_n|t) * (r_n - B_n). \tag{2.19}$$

Here the term B_n is the running mean of the rewards acquired so far and is used as a baseline. Introducing a baseline to the cost function is a standard procedure used to reduce the inherent high variance in the REINFORCE algorithm (Sutton and Barto 2018) and we add this baseline to cope with the difficulties induced by using discrete messages.

Since there is no gradient over the communication channel the policy update of one agent will be independent of the policy update of the other agent. Thus, the environment will be non-stationary and it will be harder for the agents to agree on a certain color partition and converge.

2.5 Reward

When training the model the computed policy is used to sample a guess

$$c \sim \Pi_{\Omega}(U|t) \tag{2.20}$$

which is in turn used to compute a reward r that reflects the quality of the guess in respect to the target color t.

$$r := \text{sim}(c, t) \tag{2.21}$$

, where sim is the color similarity function defined in Equation 2.8.

Comment: One could think of the reward in the setting of the sender and the receiver attempting to solve a task co-operatively. Suppose that in the process, they need to communicate the color. Then, presumably, their success in carrying out the task is related to how well the color decoded by the receiver approximates the color the sender intended to transmit. Thus, it is reasonable to assume that the reward corresponding to how well they succeed in carrying out the task is proportional to the similarity of the decoded color to the one the sender intended to convey. One could argue the reward above is a good proxy for the reward corresponding to successfully carrying out the task co-operatively.

2.6 Training

All parameters are initialized to random values and trained using stochastic gradient decent with ADAM (Kingma and Ba 2014). The batch size when training with continuous real valued messages is set to $N_b = 100$ games, and the model is trained for a total of $N = 20000$ episodes.

Moreover, when using discrete communication in the training step we set the batch size to $N_b = 256$ and the two models are trained for $N = 25000$ episodes. We have to increase the number of episodes and the batch size, compared to the case with a continuous real valued, in order to handle the increased difficulty induced by the discrete communication. All other parameters are set to the same value used for training with continuous real valued messages.

2.7 Generate partitioning

After training the agents a color-map, characterising the emerged communication schema, is constructed. This is accomplished, in analog to the WCS, by asking the speaking agent to name (or categorize) each color-tile as

$$cat(t) = \arg\max_i w_i(t), \tag{2.22}$$

where $w_i(t)$ is the ith element of the message vector $\mathbf{w}$, defined in Equation (2.10), as a function of the color-tile (t) shown to the agent.

3 Efficiency analysis

Based on recent results (Regier, Kemp, et al. 2015; Gibson et al. 2017) showing that communication tends to be efficient, we would like to investigate whether the communication schema that emerges between reinforcement learning agents exhibits similar traits. In order to evaluate this, we compare the reinforcement learning agent languages to the languages of the WCS in terms of the communication cost, defined in Equation (2.9), and the related criterion described under Well–formedness in the Materials and methods. This comparison is done in buckets of the number of color terms used, where a higher number of words is expected to result in lower communication cost. To provide the reader with a sense of scale, we compliment this picture with results using (1) a random partitioning with a uniform distribution of tiles per color word and (2) the correlation clustering of the tiles in CIELAB space; for more details, see CIELAB correlation clustering in Materials and methods. These baselines are not to be interpreted as competing models but rather an upper and lower bound on the achievable efficiency. We have left for future work another relevant baseline to which we could have compared our systems and which may set a higher bar for the comparison, as suggested by a reviewer: the rotational baseline(Regier, Kay, et al. 2007), i.e., a communication schema derived by rotating the partitioning of a real language.

3.1 Discrete vs continuous RL training

In order to justify the use of continuous real valued messages during training, we perform a comparison between training with continuous real valued and discrete messages by computing the adjusted Rand index for the resulting partitions; see Table 1.1. (See the Materials and methods section for a short explanation of adjusted Rand index.)

We observe a high adjusted Rand index between training with the two different message types (DM-RVM), which indicates that the two training approaches result in partitions with a fair amount of similarity. In addition, their corresponding internal consistency, (DM-DM) and (RVM-RVM), seems to be on the same level as the internal consistency of human partitions (H-H). The only major difference seems to be for 3 and 10 color terms, but as previously stated, these color terms are outliers when it comes to human partitions. The main difference between the two different training models is that the discrete model takes much longer to train. Hence, in most of the rest of the paper, we report results based on the continuous training model; as indicated above, the results are quite robust to the two different modes of training. In section Quantitative similarity using adjusted Rand index, we again give an explicit comparison of results using the two different methods of training.

Terms	H-H	DM-DM	RVM-RVM	DM-RVM
3	.701(±.051)	.334(±.026)	.273(±.034)	.303(±.026)
4	.452(±.031)	.397(±.023)	.337(±.028)	.323(±.024)
5	.476(±.018)	.459(±.018)	.373(±.023)	.376(±.015)
6	.528(±.011)	.524(±.006)	.537(±.033)	.485(±.009)
7	.472(±.016)	.549(±.003)	.593(±.028)	.544(±.006)
8	.471(±.041)	.505(±.007)	.518(±.017)	.484(±.007)
9	.584(±.057)	.457(±.023)	.510(±.007)	.472(±.009)
10	.718	.443	.549(±.008)	.505(±.015)

Table 1.1: Comparison between continuous real valued messages during training and discrete messages during training. Abbreviations used in table column headers: H=human, RVM= reinforcement learning training with continuous real valued messages and DM= reinforcement learning training with discrete messages. Value within parenthesis indicate a 95% confidence interval. The row corresponding to 11 color terms was excluded since no such partition was generated when training with discrete messages.

The RL agents are trained while applying a varying amount of environmental noise $\sigma_e^2 \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512\}$, i.e. Gaussian noise added to the color chips in CIELAB space, and the results are averaged over 250 experiments (25 for each level of noise). The variation in environmental noise encourages the model to find solutions with varying numbers of color terms used, see Fig 1.5, an approach that stands in stark contrast to modeling the language giving a static number of color terms, e.g. (Regier, Kemp, et al. 2015), and allows us to investigate what environmental properties affect the size of the color vocabulary. The level of communication noise was kept constant at $\sigma_c^2 = 0.1$ for all experiments.

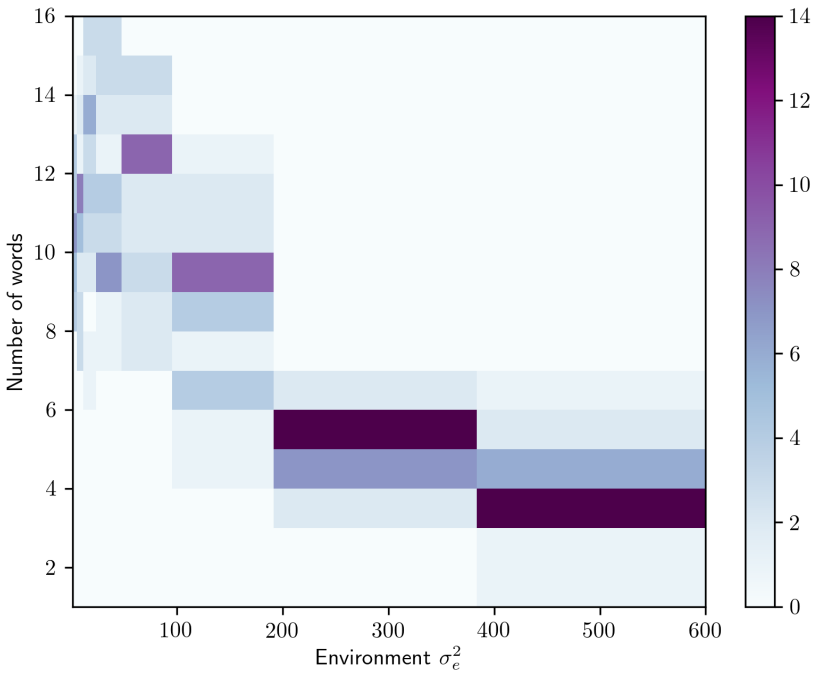


Figure 1.5: 2D histogram showing the number of emerged communication systems that end up using a particular number of color terms when trained using a particular amount of environmental noise. 25 experiments are run for each level of noise $\sigma_e^2 \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512\}$. Hence, each bin on the x axis shows the distribution over number of words resulting from that level of noise.

3.2 KL loss evaluation

The results in terms of KL loss, defined in Equation (2.6), can be seen in Fig 1.6. The WCS language data are shown both as individual languages, shown as rings, and the mean of all languages. The other results are presented as means with a 95% confidence interval indicated as a colored region. As previously shown, human languages are significantly more efficient than chance but do not reach perfect efficiency (Regier, Kemp, et al. 2015), here approximated by CIELAB CC. Further, the partitions produced by the reinforcement learning agents closely follow the efficiency of the human languages of the WCS.

3.3 Expected surprise evaluation

Fig 1.7 show the expected surprise, defined in Equation (2.3), resulting from the same experiment. These results are consistent with previously reported results in experiments with human subjects (Gibson et al. 2017).

3.4 Well–formedness evaluation

In Figure 1.8 we show the value of the well–formedness objective, for each number of color terms. The top line represents the optimal value corresponding to the optimal partition computed by correlation clustering. The remaining lines show the value attained by partitions produced by our reinforcement learning algorithm and by WCS languages. We observe that the RL partition is close to the optimal partition, and several human languages are clustered around this. Most of these are significantly better than the value for a random partition. These results are consistent with results from experiments with human subjects (Regier, Kay, et al. 2007).

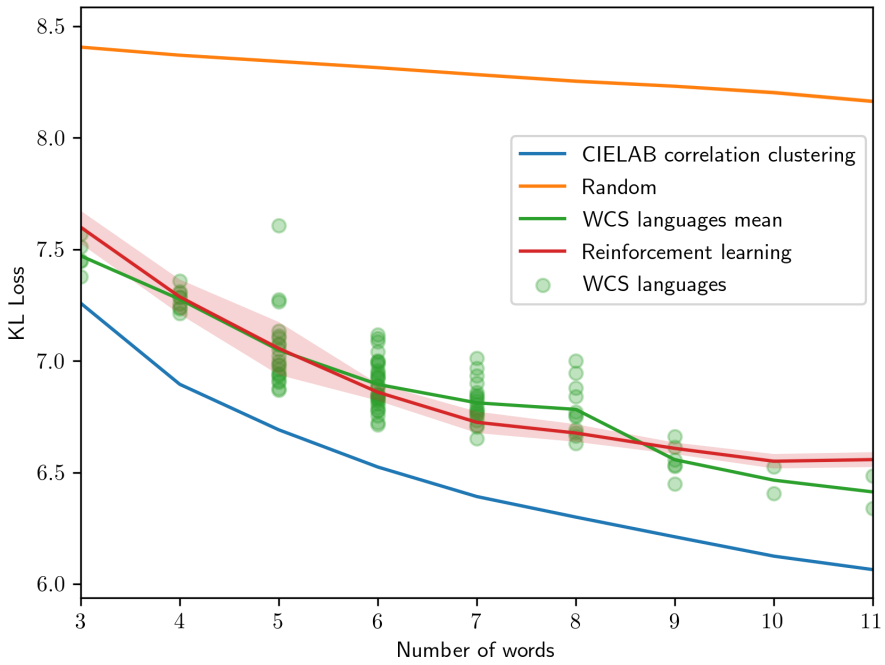


Figure 1.6: KL loss for varying number of color words used. The circles indicate the KL loss of individual WCS languages sorted based on the number of color words used. The shaded regions indicate a 95% confidence interval. Note, the WCS language data points is a reproduction from (Regier, Kemp, et al. 2015).

Partitioning characteristics

In order to further evaluate the human resemblance of our artificially-produced color space partitions, we compare a range of color maps both qualitatively and quantitatively. The quantitative comparison is done using adjusted Rand index.

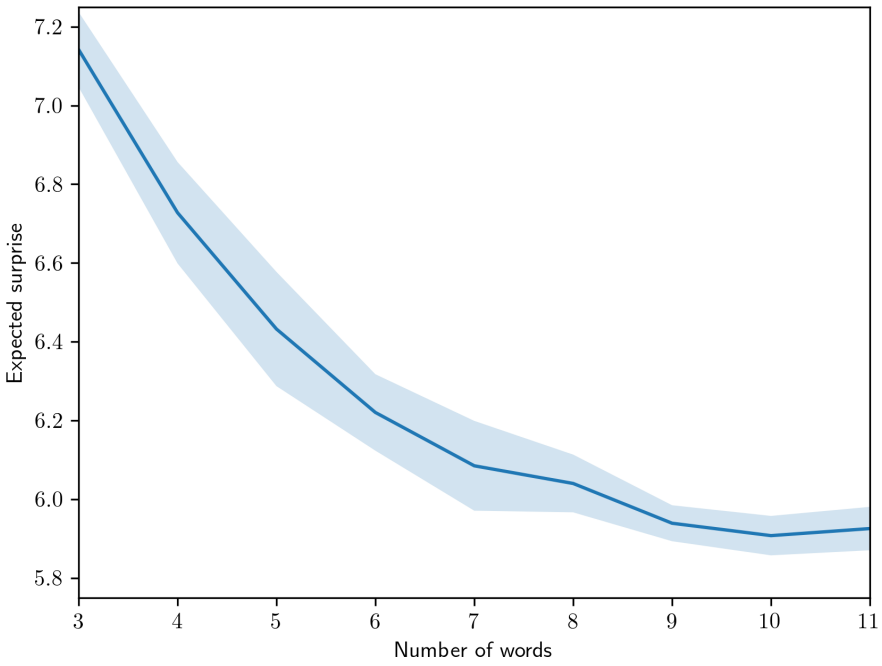


Figure 1.7: Expected surprise for varying number of color terms used. The shaded regions indicate a 95% confidence interval

3.5 Quantitative similarity using adjusted Rand index

In order to get a sense of scale, we start by computing the internal Rand index for the reinforcement learning agents and the WCS languages; see Table 1.2. This is accomplished by averaging the Rand index between all objects within the group. Comparing the internal consistency of human and RL partitionings, it seems to be on a similar level for most numbers of terms but differs for the 3 color term and 10 color term levels where the human languages yield a higher index. However, it should be noted that there are very few samples behind the human figures for those groups (i.e., 4 languages with 3 color terms and 2 with 10), and that they are outliers compared to the others. Subsequently, we compute the average Rand index across different groups, and by comparing these numbers, we can get a sense of their level of similarity; see Table 1.2. We observe fair amounts of similarity, and the human partitions are more similar to the CIELAB partitions than to the RL partitions, but the RL partitions are more similar to the CIELAB partitions.

Again, the indices for the lower number of color terms are conspicuous, but this time it has to do with the RL agents that exhibit a much lower similarity for 3 and 4 terms. A possible reason for this is connected to the way we modulate the number of color words in the RL model, i.e., by adding noise to the color chips, which may have drowned out much of the CIELAB information for the very low number of color terms, which requires a large amount of noise to appear. This would explain why RL is less similar to CCC for low terms as well. This observation suggests that other

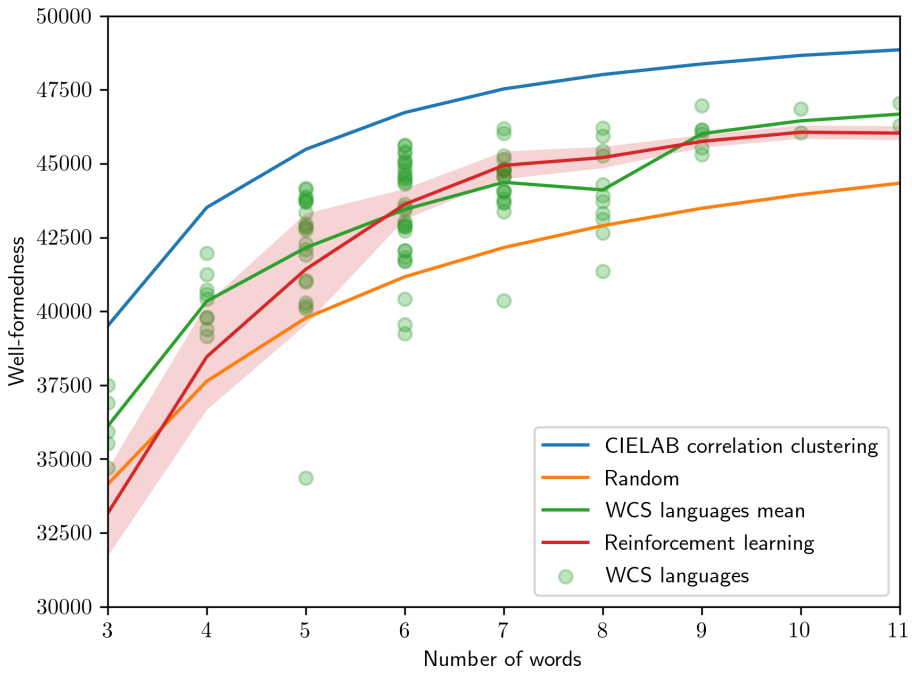


Figure 1.8: Well–formedness for varying number of color words used. The circles indicate the well–formedness of individual WCS languages sorted based on the number of words. The shaded region indicates a 95% confidence interval

mechanisms, apart from environmental noise, might influence the number of words used in human languages.

Terms	H-H	RL-RL	H-RL	H-CCC	RL-CCC	H-R
3	.701(±.051)	.273(±.034)	.173(±.028)	.385(±.038)	.192(±.020)	0(±.000)
4	.452(±.031)	.337(±.028)	.167(±.019)	.273(±.020)	.319(±.023)	0(±.000)
5	.476(±.018)	.373(±.023)	.223(±.015)	.356(±.018)	.359(±.026)	0(±.000)
6	.528(±.011)	.537(±.033)	.277(±.009)	.396(±.013)	.433(±.029)	0(±.000)
7	.472(±.016)	.593(±.028)	.292(±.008)	.409(±.016)	.456(±.007)	0(±.000)
8	.471(±.041)	.518(±.017)	.281(±.010)	.330(±.018)	.419(±.011)	0(±.000)
9	.584(±.057)	.510(±.007)	.321(±.006)	.399(±.021)	.426(±.008)	0(±.000)
10	.718	.549(±.008)	.316(±.012)	.416(±.050)	.412(±.009)	0(±.001)
11	.472	.543(±.009)	.309(±.010)	.371(±.022)	.402(±.005)	0(±.001)

Table 1.2: Comparison of the human languages in WCS to generated languages using Rand index. Abbreviations used in table column headers: H=human, RL=reinforcement learning, CCC=CIELAB correlation clustering and R=random. Value within parenthesis indicate a 95% confidence interval.

Furthermore, in Table 1.3, we compare the resulting partitions from the two different training approaches with color partitions from human language, (H-DM) and (H-RVM). We observe that both approaches seem to produce solutions which

have the same level of similarity towards human partitions, and their corresponding 95% confidence intervals overlap for all but 5, 6 and 7 color terms. However, for this number of color terms, the corresponding adjusted Rand indices for the different training approaches are still close to each other.

In our setting, we have observed a fair amount of similarity between the resulting color partitions when training with discrete and continuous real valued messages. The resulting partitions also shows same level of similarity towards human partitions. Since it is easier and faster to train with continuous real valued messages, downstream analysis will be performed using only the training approach with continuous real valued messages.

Terms	H-DM	H-RVM
3	.168(±.019)	.173(±.028)
4	.184(±.011)	.167(±.019)
5	.265(±.010)	.223(±.015)
6	.301(±.004)	.277(±.009)
7	.312(±.003)	.292(±.008)
8	.286(±.006)	.281(±.010)
9	.327(±.014)	.321(±.006)
10	.286(±.111)	.316(±.012)

Table 1.3: Comparison between continuous real valued messages during training and discrete messages during training. Abbreviations used in table column headers: H=human, RVM=reinforcement learning training with continuous real valued messages and DM=reinforcement learning training with discrete messages. Value within parenthesis indicate a 95% confidence interval. The row corresponding to 11 color terms was excluded since no such partition was generated when training with discrete messages.

3.6 Analysis of consensus color partitions

Color partitioning across multiple human languages

To enable qualitative comparison of human and artificial color maps, we produce one consensus color map for each number of color words where each color map is based on all the human languages in WCS with the given number of color words. The consensus map is computed using correlation clustering, described under Consensus maps by correlation clustering in Materials and methods. This process results in the 9 color maps shown to the left in Fig 1.9. Each of them represents a consensus color partitioning of all languages using the respective number of color words; e.g., all languages using three color terms form one consensus map.

Reinforcement learning consensus partitions

The same procedure, as described above, is subsequently performed for the artificial languages produced in the Efficiency analysis experiment and presented in the middle column of Fig 1.9. The main motivation for creating consensus maps over many

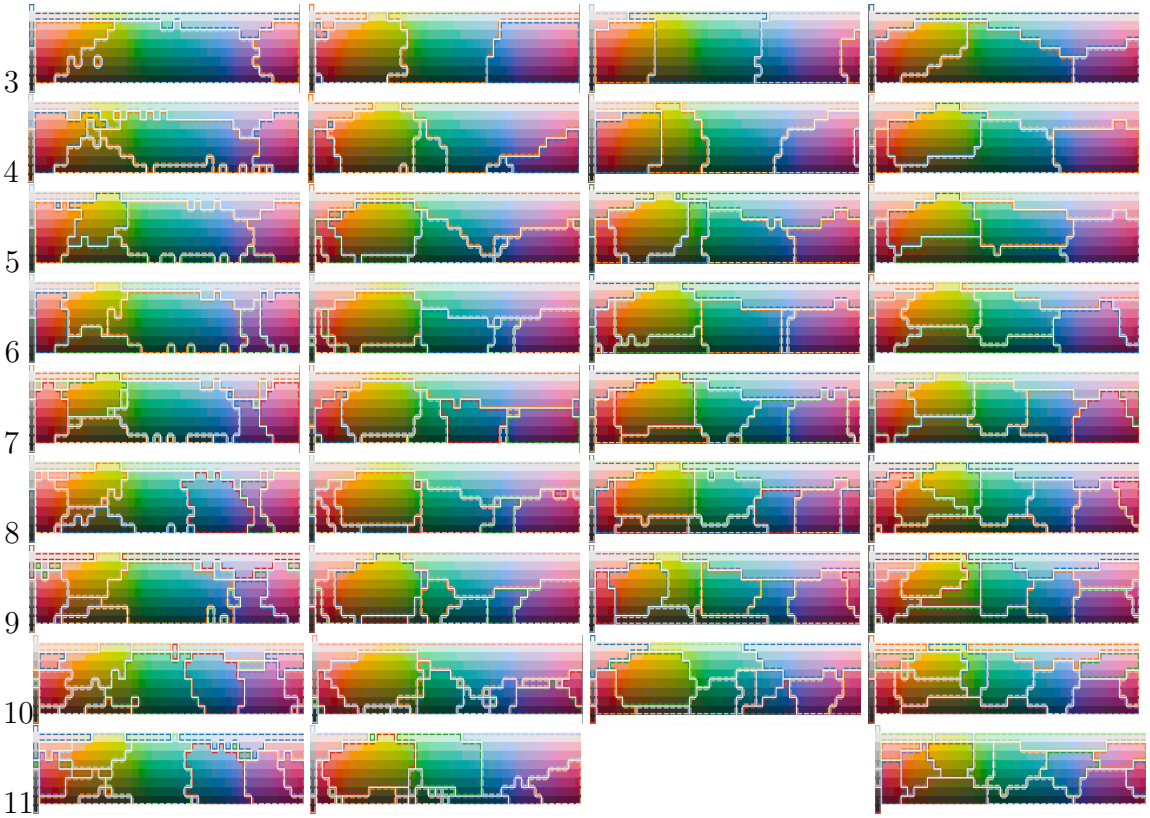


Figure 1.9: Color space partitions for different number of color terms from four different sources. The first column show human language consensus maps, i.e., consensus cross all languages in the WCS that uses the particular number of color term indicated to the left. The second column corresponds to the consensus maps over reinforcement learning partitionings using continuous real valued messages during training. The third column corresponds to the consensus maps over reinforcement learning partitionings using discrete messages during training. Finally, the fourth column of partitions is constructed using correlation clustering directly on the graph defined by the CIELAB distance between each color tile. Notice that, for 11 color terms, no partition was generated using discrete messages.

experiments is to make the result more robust to variations between experiments. That said, as shown in a Table 1.2, the consistency between reinforcement learning experiments (RL-RL) are at a level similar to human language variation (H-H). Comparing the consensus maps of the RL model to the human consensus maps, there are many similarities, especially for the languages with many color terms. One exception is however the lack of light/dark gray separation for languages with few color terms, which is not captured in the RL maps. It is however captured in the maps with higher number of color terms, which might indicate that it has to do with the type of noise that is applied to the environment during training, which is uniform in all dimensions, something that might not be true in a natural environment. In fact, analyzing the WCS color chips, the light/dark dimension has the lowest standard deviation of the 3 dimensions, i.e., 23.3 compared to 29.0 and 32.9.

CIELAB correlation clustering partitions

Finally, to the right in Fig 1.9, we show the partitions produced by applying correlation clustering to CIELAB similarities produced in the Efficiency analysis experiment.

3.7 Developing an artificial language

As a language develops over time, concepts tend to get refined into sub-categories; e.g., when a new color term comes into use, it tends to represent a subset of a region previously represented by another color term. It was suggested in Berlin and Kay (Berlin and Kay 1969) that there is an evolutionary order on the partitioning of the color space. In this proposal, the existing partitions are updated in a specific order, with the initial distinction being light/dark, followed by red, green, yellow, blue, and then other colors. The update occurs on the emergence of new color words.

To investigate whether similar patterns emerge while the languages developed between reinforcement learning agents, we show snapshots of the color partitionings as they develop during one training episode in Fig 1.10. To complement this picture, we show how the number of terms develops on a timeline in Fig 1.11 and how the KL loss falls as the number of terms used goes up on the same timeline in Fig 1.12. The color partition snapshots were captured on the last episode using that number of color terms. As seen in Fig 1.10, the order in which colors emerge in human languages is not very well replicated in the artificial language while the subdivision of partitions is captured to a greater extent. Further examining Fig 1.11, it is interesting to note that the number of color terms used tend to steadily go up during training—this resembles how the vocabulary of human speakers tends to grow when a community communicates frequently regarding a specific subject; e.g., people working with color tend to use a larger-than-average color vocabulary, especially when talking to each other.

Environmental impact on partitioning

In this section we describe the results of controlling environmental factors such as the noise level in the various channels over which the agents communicate.

3.8 Modulating the vocabulary size by varying environmental noise

Environmental noise is noise added to the color chips before shown to the agent. In information-theoretic terms, this channel refers to the conditional probability $p(w \mid c)$. This emulates the fact that when referring to an object in the real world it may vary in color. This is especially true in a natural environment where, for instance, a tree may vary considerably in color over time; hence, when referring to specific trees using color, it may not be useful to develop very exact color terms. In contrast, in an industrialized society, exact color information may carry more

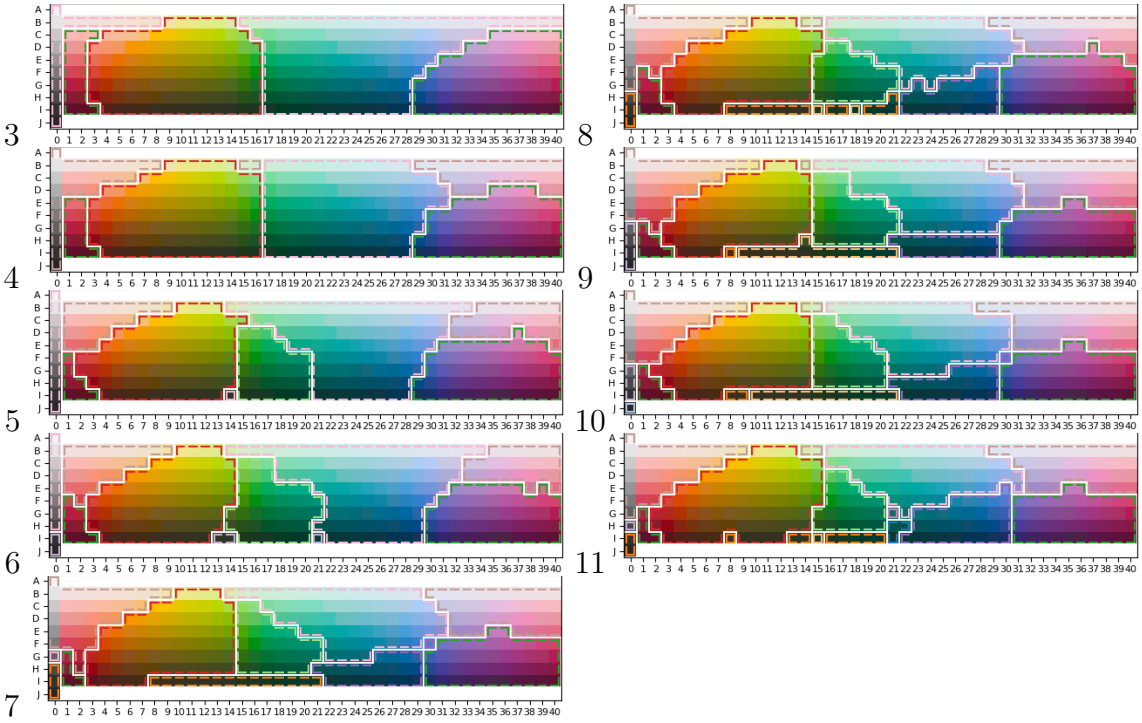


Figure 1.10: Color maps captured during one training session as the emerging language progress towards an increasing number of terms.

information, which could be one reason for why they tend to use more color terms. To show the effect of varying the environmental noise on our artificially synthesized languages, two experiments are conducted:

The first experiment investigates the effect on the number of terms used as a function of environmental noise. As can be seen in Fig 1.13, this has the effect of lowering the number of color words of the resulting language. Though we cannot say that this effect is the main driving force behind language complexity in real languages, it is clear that it can have a significant effect in a setting like ours. An interesting effect that we have seen consistently is that low levels of environmental noise increase the size of the vocabulary in the resulting language.

The second experiment measures to what extent the noise affects how the space is partitioned, apart from the number of terms used. The experiment is conducted by computing, for each number of color terms used, the internal consistency between all partitionings that resulted in that number of terms regardless of the level of noise and the average internal consistency between partitionings created using the same level of noise. The environmental noise levels used in this experiment are $\epsilon_{\mathbf{e}} \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512\}$ and the results are presented in Table 1.4. From the numbers, we can conclude that partitions resulting from other noise groups are as similar as within the same noise group for most levels of terms used. However, we again see that for the small vocabulary groups (induced with a high level of noise) there seems to be more discrepancy, especially when 3 terms are used, which might help to further explain the lower performance in previous experiments on those groups.

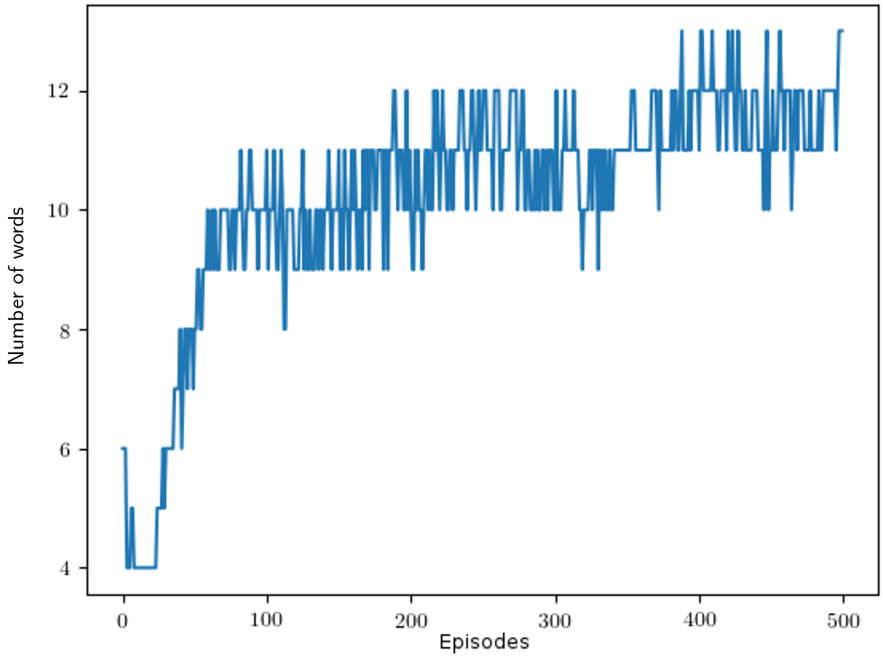


Figure 1.11: Change in the number of words used by the agents during training. The X-axis represents the number of episodes trained and the Y-axis the number of words used at that point.

Terms used	All	Within noise group
3	0.273(±0.034)	0.324(±0.039)
4	0.337(±0.028)	0.377(±0.069)
5	0.373(±0.023)	0.275(±0.021)
6	0.537(±0.033)	0.486(±0.099)
7	0.593(±0.028)	0.632(±0.095)
8	0.518(±0.017)	0.573(±0.146)
9	0.510(±0.007)	0.541(±0.048)
10	0.549(±0.008)	0.521(±0.178)
11	0.543(±0.009)	0.538(±0.096)

Table 1.4: Estimating the secondary effects of environmental noise, i.e., other than the number of terms used. For each number of terms used (column one) the table shows the internal consistency, measured using adjusted Rand index, for all generated maps (column two), and the mean internal consistencies computed for each noise level (column three), e.g. if the maps that ended up using 6 terms all where generated using 100 and 200 $\epsilon_{\mathbf{e}}$ noise then this column would be the average of the internal consistency of those two groups. 95% confidence interval indicated within parenthesis.

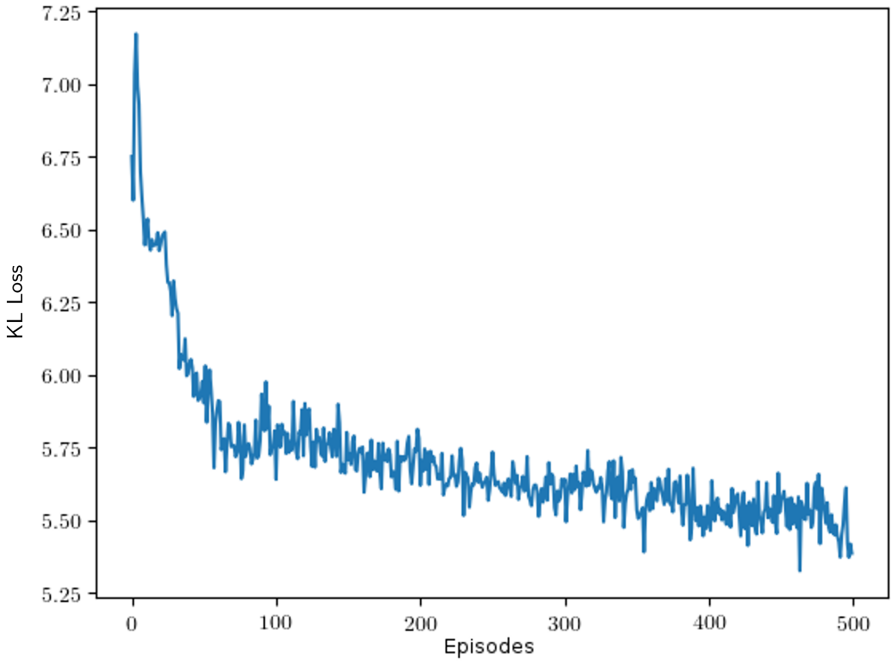


Figure 1.12: KL loss change during training. The X-axis shows the number of episodes.

3.9 Modulating the vocabulary size by varying communication noise

In order to investigate the effect of noise further, we turn to the noise on the communication channel over which words are transmitted. In Fig 1.14, we show how the number of words is affected when noise is introduced to the communication channel. In similarity with environmental noise, we see a decline in the number of terms used as we increase the noise in the communication channel. However, the characteristics seem to differ slightly where communication noise has a greater initial effect and then levels out.

4 Materials and methods

4.1 CIELAB correlation clustering

The CIELAB clustering is the partitioning obtained by applying correlation clustering (Demaine et al. 2006; Bansal et al. 2004), a technique to obtain clusterings when there are both similarity and dissimilarity judgments on objects. This is applied to a graph with vertices corresponding to color tiles and where the edge (u, v) has weight $sim(u, v) - \frac{1}{2}$ where sim is the similarity metric defined in Equation (2.8). Thus, there are both positive weights corresponding to similar tiles ($sim > 1/2$) and negative weights corresponding to dissimilar tiles ($sim < 1/2$). Correlation

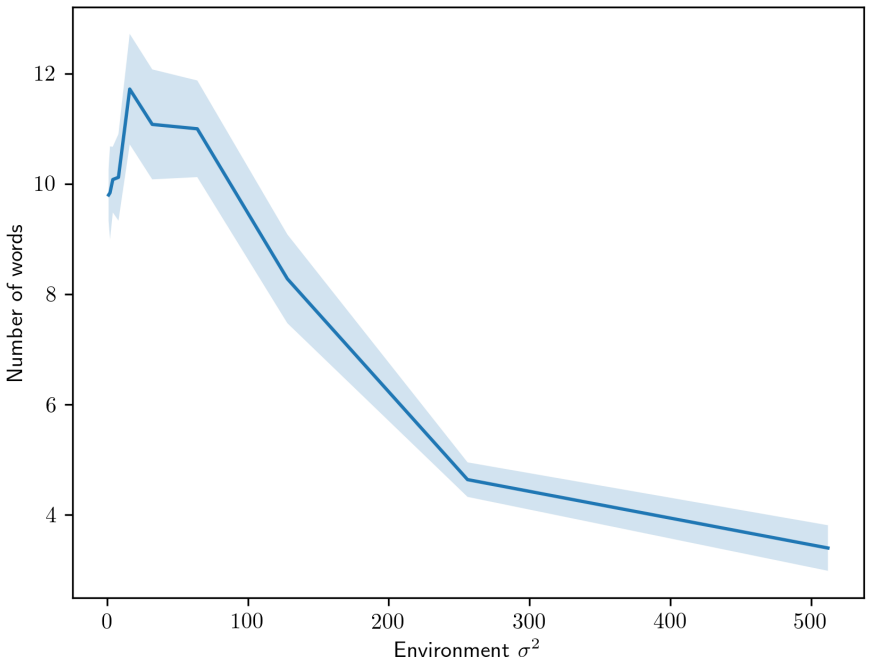


Figure 1.13: The number of color terms used by the agents when different amounts of noise are applied to their environments.

clustering is a NP-hard problem, so we have used a new method that we developed based on a non-convex relaxation that is guaranteed to converge to a local optimum (forthcoming).

4.2 Consensus maps by correlation clustering

In order to obtain a consensus maps of several different runs of our RL algorithm, we again use correlation clustering. Each run of our algorithm provides a similarity judgment between two tiles if they are placed in the same color partition and a dissimilarity judgment otherwise. We use these judgments as input to the correlation clustering algorithm to produce the consensus partition that aggregates all these judgments together.

4.3 REINFORCE

REINFORCE (Williams 1992) is a well known reinforcement learning algorithm in the policy gradient family, meaning that the policy is directly parameterized by a differentiable model, in our case a neural network. The model is trained to maximize expected reward by updating the neural network that suggests what actions to take to increase the probability of actions that have previously led to rewards in similar situations.

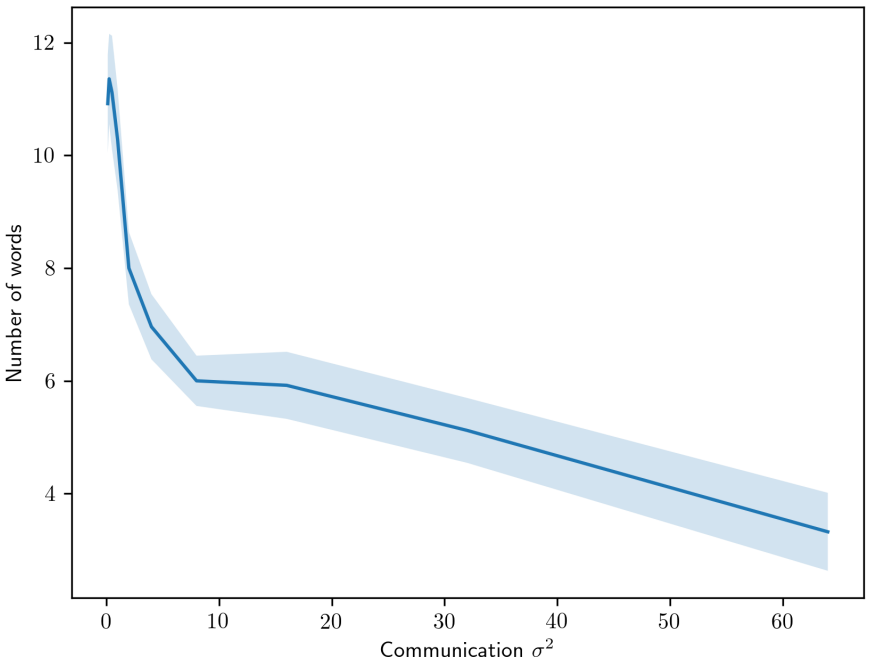


Figure 1.14: The number of color terms used by the agents when different amounts of noise are applied to their communication.

4.4 Adjusted Rand Index

The Adjusted Rand index (Rand 1971) is a method of computing the similarity between two data clusterings or partitions that was introduced by William M. Rand. Essentially it computes the relative number of pairs of objects that appear together in the same class in both partitions.

4.5 The World Color Survey

The World Color Survey (WCS) (Kay and Cook 2014) is a project that compiled color naming data from 110 unwritten languages and made it publicly available at `http://www1.icsi.berkeley.edu/wcs/data.html`. For each language, an average of 25 speakers were asked to name each color in a matrix of 330 color chips (see Fig 1.9) sampled from the Munsell color system to uniformly cover the human visual color spectra.

5 Discussion

We see in figs. 1.6 and 1.7 that the RL results track the human results very closely in the KL loss and well-formedness characteristics. In terms of Rand index similarity (table 1.2), the overall similarity of human languages to other human languages and RL mode maps to other RL mode maps is much greater than human language to RL mode maps at each number of words used. The human-to-RL similarity, however, is consistently greater than the human-to-random similarity, which is zero. Taken together, the reinforcement learning process produces mode maps that take into account some factor of human color space partitioning, and it also produces well-formedness and efficiency outcomes that represent a model significantly closer to the human behavior relative to these latter criteria.

One explanation for this difference may be found by looking at the Rand index similarity of the human-generated mode maps to the CIELAB maps. The latter is an idealized partitioning of the space based on color distances taken from CIELAB's perceptually uniform (relative to human vision) color space. The similarity is consistently, but not hugely greater between the human maps and CIELAB than between the human maps and the RL maps. Given the success of the RL maps at modeling the communication characteristics of human color maps, this difference likely reflects biological and environmental aspects of human color perception that the simulated agents, due to their simplicity, do not represent. The RL-based maps also show similarly high Rand index similarity to the CIELAB maps, possibly due to the influence of the CIELAB distances on the reward function in the RL process. Our RL model therefore successfully separates communicative factors from the details of human perception, and gives space for experimentation on the influence of biological and environmental detail in arriving at a color term consensus within a simulated speech community.

Looking at the color maps in Fig 1.9, we perceive qualitatively *some* similarity in overall partitioning between humans and RL agents for a given number of color words, but the RL agents still do not closely replicate the human partitions—unsurprising, given the Rand index differences as above. The principal difficulty that the RL agents seem to have is in replicating human light/dark distinctions, which are under-emphasized in the RL partitions. We hypothesize that the light/dark distinction needs a different treatment, for reasons posed by the human perceptual architecture (for example, non-uniform need probabilities (Baddeley and Attewell 2009)), than the other components of the CIELAB or WCS color spaces.

On the other hand, the RL maps do share the behavior of the human maps with regard to how partitions of the color space are refined as we increase the number of colors used: the resulting partition tends to constitute a sub–partition rather than producing a completely different partitioning. Thus, the RL results appear to confirm the behavior observed by Berlin and Kay (Berlin and Kay 1969).

As argued in (Kemp et al. 2018), there are trade-offs between cognitive and communication costs which could change over time in response to various evolutionary forces. Such changes may be quite difficult to study in real languages, but our framework provides a very powerful and flexible tool for studying such changes under

carefully controlled conditions where we can adjust one parameter (say noise) while keeping the rest fixed.

6 Conclusion

In this work, we successfully demonstrated the value of a reinforcement learning approach to simulating the conditions under which speakers might come to an agreement on how to partition a semantic space. Color provided a convenient domain of experiment because of the extent of real-world data collection and analysis that has already been performed and also due to the ability to represent the color space as evenly-selected samples from a continuous space, as with the WCS. Our RL agents replicate important aspects of human color communication, even though they lack the full perceptual and linguistic architecture of human language users. However, the RL paradigm will enable us in future work to represent more detailed aspects of the environment and biological architecture *in silico*, allowing our system to be used as a platform for hypothesis generation and cognitive modeling.

As for hypothesis generation, the behavior of our model suggests that greater communication and environmental noise produces an overall drop in the number of color words. This outcome provides further clues as to where to look for environmental factors that may account for differences in color vocabulary across real-world speaker groups.

Our approach can offer complementary insight to the recent approach of (Zaslavsky et al. 2018) who argued that languages efficiently compress ideas into words by optimizing the *information bottleneck*. Additional future work includes expanding from a two-agent paradigm to a multi-agent and even a large-population paradigm, which are areas under active development in the field of agent simulation. A key long-term goal for this work is to expand from the domain of color to other semantic domains, such as culture-specific partitions of approximate number (e.g., "few" vs. "many") and even "general-domain" semantic relatedness hierarchies, such as WordNet.

References

Baddeley, Roland and David Attewell (2009). "The Relationship Between Language and the Environment: Information Theory Shows Why We Have Only Three Lightness Terms". In: *Psychological Science* 20.9, pp. 1100–1107 (cit. on pp. 63, 90).

Bansal, Nikhil, Avrim Blum, and Shuchi Chawla (2004). "Correlation Clustering". In: *Machine Learning* 56.1-3, pp. 89–113 (cit. on pp. 71, 87).

Baronchelli, Andrea, Tao Gong, Andrea Puglisi, and Vittorio Loreto (2010). "Modeling the emergence of universality in color naming patterns". In: *Proceedings of the National Academy of Sciences* 107.6, pp. 2403–2407 (cit. on pp. 65, 68).

Belpaeme, Tony and Joris Bleys (2005). "Explaining universal color categories through a constrained acquisition process". In: *Adaptive Behavior* 13.4, pp. 293–310 (cit. on p. 68).

Berlin, Brent and Paul Kay (1969). *Basic Color term. Their Universality and Evolution*. 2010. Berlin, Boston: De Gruyter Mouton (cit. on pp. 65, 84, 90).

Carr, Jon W, Kenny Smith, Jennifer Culbertson, and Simon Kirby (under review 2018). *Simplicity and informativeness in semantic category systems*. URL: psyarxiv.com/jkfyx (cit. on p. 63).

Carr, Jon William (2019). "Induction and interaction in the evolution of language and conceptual structure". PhD thesis. University of Edinburgh (cit. on pp. 64, 66).

Cover, Thomas M. and Joy A. Thomas (2006). *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience (cit. on p. 63).

Dayan, P. and Y. Niv (2008). "Reinforcement learning: The Good, The Bad and The Ugly". In: *Current Opinion in Neurobiology* 18.3, pp. 1–12 (cit. on p. 64).

Demaine, Erik D., Dotan Emanuel, Amos Fiat, and Nicole Immorlica (2006). "Correlation clustering in general weighted graphs". In: *Theor. Comput. Sci.* 361.2-3, pp. 172–187 (cit. on pp. 71, 87).

Evtimova, Katrina, Andrew Drozdov, Douwe Kiela, and Kyunghyun Cho (2017). "Emergent Language in a Multi-Modal, Multi-Step Referential Game". In: *CoRR* abs/1705.10369. arXiv: 1705.10369 (cit. on pp. 64, 69).

Foerster, Jakob, Yannis 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*. Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. Curran Associates, Inc., pp. 2137–2145 (cit. on pp. 64, 73).

Gibson, E., R. Futrell, J. Jara-Ettinger, K. Mahowald, L. Bergen, S. Ratnasingam, M. Gibson, S.T. Piantadosi, and B. R. Conway (2017). "Color naming across languages reflects color use". In: *Proc Natl Acad Sci USA* 114.40, pp. 10785–10790 (cit. on pp. 63, 66, 68–71, 76, 78).

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 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., pp. 2146–2156 (cit. on pp. 64, 69).

Jäger, Gerhard and Robert van Rooij (Nov. 2007). "Language Structure: Psychological and Social Constraints". In: *Synthese* 159, pp. 99–130. DOI: 10.1007/s11229-006-9073-5 (cit. on p. 68).

Jameson, Kimberly and Roy G. D'Andrade (1997). "It's not really red, green, yellow, blue: an inquiry into perceptual color space". In: *Color Categories in Thought and Language*. Ed. by C. L. Hardin and Luisa Maffi. Cambridge University Press, pp. 295–319 (cit. on p. 63).

Jorge, Emilio, Mikael Kågebäck, and Emil Gustavsson (2016). "Learning to Play Guess Who? and Inventing a Grounded Language as a Consequence". In: *CoRR* abs/1611.03218. arXiv: 1611.03218 (cit. on pp. 64, 69).

Kay, Paul and Richard S Cook (2014). "World Color Survey". In: *Encyclopedia of Color Science and Technology*, pp. 1–8 (cit. on p. 89).

Kemp, Charles, Yang Xu, and Terry Regier (2018). "Semantic Typology and Efficient Communication". In: *Annual Review of Linguistics* 4.1, pp. 109–128 (cit. on pp. 63, 64, 68, 90).

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (cit. on p. 76).

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. ISSN: 0010-0277 (cit. on p. 63).

Lazaridou, Angeliki, Alexander Peysakhovich, and Marco Baroni (2016). "Multi-Agent Cooperation and the Emergence of (Natural) Language". In: *CoRR* abs/1612.07182. arXiv: 1612.07182 (cit. on pp. 64, 69).

Lucy, John A. (1997). "The linguistics of "color"". In: *Color Categories in Thought and Language*. Ed. by C. L. Hardin and LuisaEditors Maffi. Cambridge University Press. Chap. 15, pp. 320–346 (cit. on p. 65).

Niv, Y. (2009). "Reinforcement learning in the brain". In: *The Journal of Mathematical Psychology* 53.3, pp. 139–154 (cit. on p. 64).

Niv, Y. and A. Langdon (2016). "Reinforcement Learning with Marr". In: *Current Opinion in Behavioral Sciences* 11.3 (cit. on p. 64).

Piantadosi, Steven T., Harry Tily, and Edward Gibson (2011). "Word lengths are optimized for efficient communication". In: *Proceedings of the National Academy of Sciences* 108.9, pp. 3526–3529 (cit. on p. 63).

Rand, William M (1971). "Objective criteria for the evaluation of clustering methods". In: *Journal of the American Statistical association* 66.336, pp. 846–850 (cit. on p. 89).

Regier, T., P. Kay, and N. Khetrapal (2007). "Color naming reflects optimal partitions of color space". In: *Proc Natl Acad Sci USA* 104.3, pp. 1436–1441 (cit. on pp. 67, 71, 72, 76, 79).

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. 63, 64, 66–71, 76–79).

Saunders, Barbara (1995). "Disinterring Basic Color Terms : a study in the mystique of cognitivism". In: *History of the Human Sciences* 8.4, pp. 19–38 (cit. on p. 65).

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. 68).

Sutton, Richard S and Andrew G Barto (2018). *Reinforcement learning: An introduction*. MIT press (cit. on p. 75).

Sutton, Richard S. and Andrew G. Barto (1998). *Reinforcement Learning An Introduction*. MIT Press (cit. on p. 64).

Wiering, M. and M. van Otterlo, eds. (2012). *Reinforcement Learning: State-of-the-Art*. Springer (cit. on p. 64).

Williams, Ronald J (1992). "Simple statistical gradient-following algorithms for connectionist reinforcement learning". In: *Reinforcement Learning*. Springer, pp. 5–32 (cit. on pp. 74, 88).

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*. DOI: 10.1073/pnas.1800521115 (cit. on pp. 64, 65, 69, 71, 91).

Zipf, George K. (1949). *Human Behaviour and the Principle of Least Effort*. Addison-Wesley (cit. on p. 63).

Paper 2

Learning approximate and exact numeral systems via reinforcement learning

Emil Carlsson, Fredrik D. Johansson., Devdatt Dubhashi.

Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci), 43, 2021.

The paper has been reformatted for uniformity.

Paper 2. Learning approximate and exact numeral systems via reinforcement learning

Emil Carlsson, Fredrik D. Johansson., Devdatt Dubhashi.

Abstract

Recent work (Xu et al. 2020) has suggested that numeral systems in different languages are shaped by a functional need for efficient communication in an information-theoretic sense. Here we take a learning-theoretic approach and show how efficient communication emerges via reinforcement learning. In our framework, two artificial agents play a Lewis signaling game where the goal is to convey a numeral concept. The agents gradually learn to communicate using reinforcement learning and the resulting numeral systems are shown to be efficient in the information-theoretic framework of Regier et al. (2015) and Gibson, Futrell, Jara-Ettinger, et al. (2017). They are also shown to be similar to human numeral systems of same type. Our results thus provide a mechanistic explanation via reinforcement learning of the recent results in Xu et al. (2020) and can potentially be generalized to other semantic domains.

Keywords: efficient communication; reinforcement learning; numeral systems

1 Introduction

Why do languages partition mental concepts into words the ways they do? A recent influential body of work suggests language is shaped by a pressure for efficient communication which involves an information-theoretic tradeoff between cognitive load and informativeness (Kemp and Regier 2012; Gibson, Futrell, Jara-Ettinger, et al. 2017; Zaslavsky, Kemp, Tishby, et al. 2019). This means that language is under pressure to be simultaneously informative, to support effective communication, while also being simple, in order to minimize the cognitive load.

While the information-theoretic framework is insightful and has broad explanatory power across a variety of domains, see the reviews by Kemp, Xu, et al. (2018) and Gibson, Futrell, Piantadosi, et al. (2019), a fundamental question that is left unaddressed is if there is *mechanistic explanation* for how such efficient communication schemes could arise. We address this question here from a learning-theoretic viewpoint: *is there a computational learning mechanism that leads to efficient communication?*

We can relate our approach to previous work using the influential "three levels of analysis" framework posited by David Marr (Marr 1982) which has been described as one of the most enduring constructs of twentieth century cognitive science and computational neuroscience. While the previous work such as Kemp and Regier

(2012), Kemp, Xu, et al. (2018), and Gibson, Futrell, Piantadosi, et al. (2019) is situated at the first or "theory" level of Marr, our analysis is at the *representation and algorithmic* level. In particular, we propose very natural reinforcement learning mechanisms that are able to learn such efficient communication schemes. The learning aspect is emphasised by Tomaso Poggio (Poggio 2012) in an update of Marr:

> it is ... important to understand how an individual organism, and in fact a whole species, learns and evolves [the computations and the representations used by the brain] from experience of the natural world ... a description of the learning algorithms and their a priori assumptions is deeper, more constructive, and more useful than a description of the details of what is actually learned ... the problem of learning is at the core of the problem of intelligence and of understanding the brain ... learning should be added to the list of levels of understanding ...

Recent research gives evidence that the style of learning algorithms we consider here seem to be centrally implicated in exploration strategies used by humans (Schulz and Gershman 2019).

Reinforcement learning has been proposed recently as a mechanistic explanation for how efficient communication arises in the colour domain (Kågebäck et al. 2020; Chaabouni et al. 2021) and it was observed that this approach could potentially be applied to other domains. Here we investigate the reinforcement learning approach in the domain of numeral systems. It has been shown recently that numeral systems across languages reflect a need for efficient communication (Xu et al. 2020). Numeral systems come in many shapes, some are recursive like English and can express any numeriosity while other non-recursive systems only consists of a small set of words (Comrie 2013). These non-recursive systems could be either *exact restricted* - in the sense that exact numerosities can only be expressed on a restricted range, or *approximate* like in the language Mundurukú where most numeral words have an imprecise meaning (Pica et al. 2004). Here we only consider non-recursive systems.

We show that reinforcement learning mechanisms can indeed be used to learn exact and approximate numeral systems which are near-optimal in an information-theoretic sense and similar in structure to human numeral systems of the same complexity. Unlike Kågebäck et al. (2020), who use a policy-gradient method, we use a Q-learning algorithm with an implicit Thompson Sampling exploration scheme (Sutton and Barto 1998).

2 Learning to communicate: Signalling games

We consider the communication framework developed in Regier et al. (2015) and Xu et al. (2020) which consists of a sender and a listener. The sender has a concept in mind and wishes to convey this to a listener over a discrete communication channel. The listener then tries to reconstruct the concept. This is illustrated schematically in Figure 2.1.

We extend this setup to a *Lewis signaling game* (Lewis 1969), by considering two artificial agents starting *tabula rasa* and gradually learning to communicate efficiently

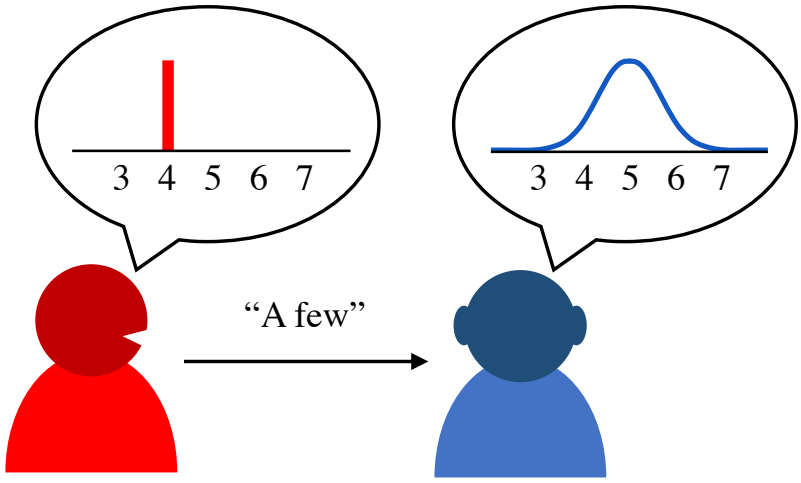


Figure 2.1: Illustration of the communication setup presented in Xu et al. (2020). The sender wants to convey the numeral concept 4 and utters "a few". The listener is unsure of which numeral the sender is referring to and produces a probability distribution over possible numerals.

via a reinforcement learning algorithm (introduced in detail in later sections) by playing several rounds of the game. In each round of the game, a number $n \in \mathcal{N}$ from the interval $\mathcal{N}$ is sampled according to a need probability of the environment, $p(n)$, which represent how often a numeral concept has to be referred to in the environment. The sampled number n is then given to the sender which has to pick a word w from the vocabulary $\mathcal{W}$ and utter to the listener. Having received a word w, the listener guesses a number $\hat{n} \in \mathcal{N}$ and a shared reward, $r(n, \hat{n})$, is given to both agents based on the distance between the guess $\hat{n}$ and the true number n. Here we explore three different reward functions, one linear, one inverse and one exponential

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

One round of the signaling game is visualized in Figure 2.2 and one could interpret it as follows: the agents are playing a *cooperative game* which involves solving a common task in which success depends on how well the listener reconstructed the number the sender had in mind. The reward functions considered were chosen in order to model different pressure for how precise the listener's reconstruction has to be.

2.1 Reinforcement learning for efficient communication

Reinforcement learning is an area of machine learning which studies how agents in an environment can learn to pick actions given states as to maximize a reward signal (Sutton and Barto 1998) and recent studies suggests that reinforcement learning may be an component in neural mechanisms such as the phasic activity of dopamine

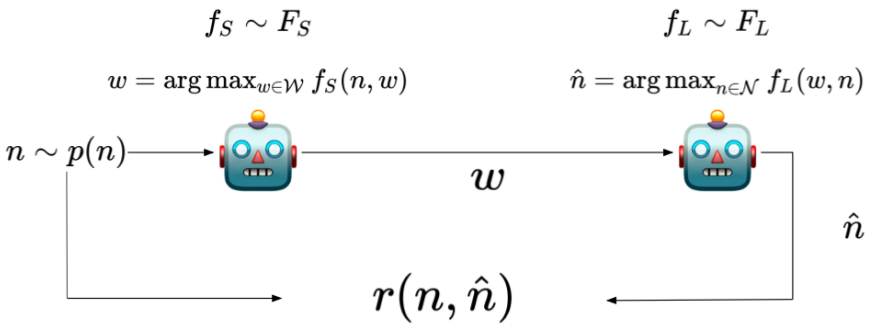


Figure 2.2: Illustration of one round of our Lewis signaling game, which will be formally introduced in later sections. The sender is given a number n and samples a model f_S from F_S using dropout and conveys the word w giving highest reward according to f_S. The listener proceeds in similar fashion, given w it samples a model f_L from F_L and guesses the number $\hat{n}$ that yields most reward according to f_L. A shared reward is given to both agent based on how close $\hat{n}$ is to n.

neurons (Niv et al. 2005; Dabney et al. 2020). In this work our agents will learn to communicate efficiently using reinforcement learning by maximizing the reward in the Lewis signaling game, Figure 2.2. For the sender this translates into conveying the word w which yields highest expected reward given the number n and for the listener to guess the number $\hat{n}$ yielding highest expected reward given the word w.

Inherent in this setup is an exploration-exploitation tradeoff—the agents have to balance between exploring uncertain actions in order to gain new insights about the environment and exploiting it current knowledge in order to maximize the reward signal. Recent work in neuroscience suggests that classical machine learning strategies, such as Thompson sampling (Thompson 1933), seem to mechanistically correspond to exploration strategies used by humans (Schulz and Gershman 2019).

In this work we will use the Bayesian approach and Thompson sampling in order to handle the exploration-exploitation tradeoff. This means that each agent keeps a belief, or posterior distribution, over possible models of the environment and at each time step it samples a plausible model from the belief and acts optimal according to the sampled model. After getting feedback from the real environment an agent updates its belief over possible models accordingly. We will use an implicit form of Thompson sampling presented in Gal and Ghahramani (2016) where each agent will be represented as a feedforward neural network[1] that maps input and action to expected reward

$$F_S : \mathcal{N} \times \mathcal{W} \longrightarrow [0, 1]$$
$$F_L : \mathcal{W} \times \mathcal{N} \longrightarrow [0, 1].$$

Given a new round of our signaling game each agent samples a smaller network $f_S \sim F_S$ and $f_L \sim F_L$ from its neural network using the regularization technique dropout (Srivastava et al. 2014) which means that the activation at each neuron in the network is randomly set to 0 with probability p. In this way the agents sample,

[1]From now on we will use the subscript S for the sender and the subscript L for the listener.

via dropout, one out of all possible models of the expected rewards spanned by F_S and F_L. Hence, the networks f_S and f_L become the current internal models of the expected reward of the speaker and listener. Given an input, each agent acts greedily w.r.t. the smaller networks f_S and f_L; given the number n, the sender conveys the word $\hat{w}$ yielding highest expected reward according the sampled model

$$\hat{w} = \arg\max_{w \in \mathcal{W}} f_S(n, w)$$

Similarly, given the word $\hat{w}$, the listener guesses the number $\hat{n}$ satisfying

$$\hat{n} = \arg\max_{n' \in \mathcal{N}} f_L(\hat{w}, n').$$

After playing the game for m rounds, each agent update the weights in F_S (or respectively F_L) by finding the values which minimize the mean-squared error (MSE)

$$\text{MSE}_S = \frac{1}{m} \sum_{i}^{m} (f_S(\hat{w}_i, n_i) - r_i)^2,$$
$$\text{MSE}_L = \frac{1}{m} \sum_{i}^{m} (f_L(\hat{n}_i, \hat{w}_i) - r_i)^2.$$

It should be noted that this game is only partially observable—in each round of the game the sender observes the tuple $(n, \hat{w}, r)$ while the listener observes $(\hat{w}, \hat{n}, r)$.

3 Numeral systems

We study two of the three types of numeral systems presented in Xu et al. (2020). First, we consider the *exact restricted* systems, or simply *exact* systems, where exact numerosities can only be expressed on a restricted range. An example of this is the numeral system *one, two, three* and *more than three*. With this system precise communication can only be achieved for the first three numerals and it is clear which part of the number line each numeral word corresponds to.

The second type is *approximate* numeral systems where the meaning of numerals are approximate. Example of such inexact numerals are *a few* and *many* which do not cover a precise restricted range.

We do not address recursive numeral systems in this work since it require a different way of modelling the agents and we leave it for future work.

3.1 Artificial numeral systems

Given that a sender-listener pair has played the signaling game in Figure 2.2 for a certain number of rounds we would like to compute the resulting numeral system. We do this by first estimating the conditional probability $p(w|n)$, i.e the probability that the sender refers to the number n with the word w, by running $m = 1000$ rounds of the game, without updating the agents, with the number n given to the

sender and count how many times each word is used. Hence, we do the following Monte-Carlo estimation

$$p(w|n) \approx \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}(w = \arg\max_{\hat{w}} f_{S,i}(\hat{w}, n))$$

where $\mathbf{1}(\cdot)$ is the indicator function. We check if the resulting conditional distribution is peaked, i.e if it for each n assigns more than 0.90 probability mass to one token w, if not we interpret it as an approximate numeral system. Moreover, we consider the mode of $p(w|n)$ to be an exact numeral system.

3.2 Complexity and communication cost

We measure complexity of a numeral system simply as the number of words used in the system. In Xu et al. (2020) a grammar based complexity measure was used. This is not needed here since we do not consider recursive numeral systems and for exact and approximate systems there is no pressure for systematicity.

Given a sender distribution S and a listener distribution L_w we measure the communicative cost of conveying a number n as the information lost in the listener's reconstruction of the sender distribution given the numeral w. As has been done in previous studies (Xu et al. 2020), we model this as the Kullback-Leibler divergence (KL) between S and L_w. Under sender certainty, $S(n) = 1$, this reduces to the surprisal

$$\mathbb{KL}(S||L_w) = \sum_i S(i) \log \frac{S(i)}{L_w(i)} = -\log L_w(n),$$

which can be viewed as how surprised the listener would be by the fact that the sender uttered w if they knew the true number n.

In order to measure the full communication cost of a numeral system we compute the expected surprisal as

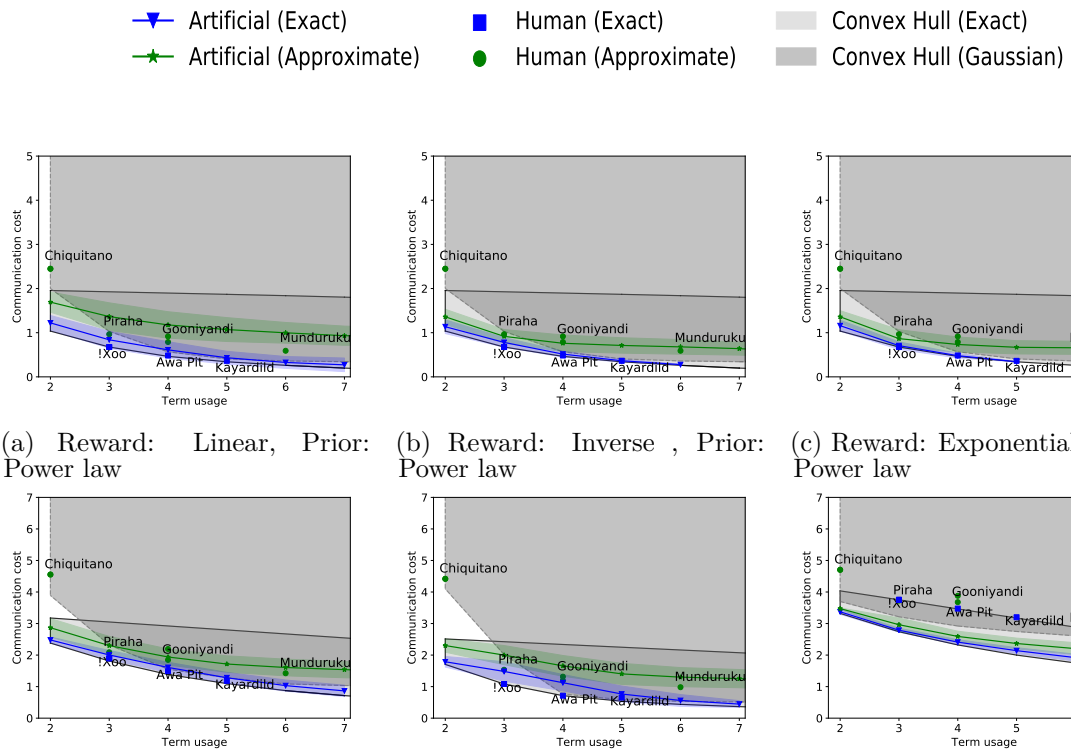
$$C = -\sum_{n,w} p(w|n)p(n) \log L_w(n),$$

where $L_w(n)$ is computed using Bayes rule

$$L_w(n) = \frac{p(w|n)p(n)}{\sum_{n'} p(w|n')p(n')}.$$

Here $p(w|n)$ denotes the sender partition of the number line and $p(n)$ the need probability of the environment. The measure of the total communication cost of a numeral system used here is exactly the measure of communication cost used in Gibson, Futrell, Jara-Ettinger, et al. (2017) and by taking a deterministic sender, i.e a sender which for each n assigns all probability mass to a single word w, we get the measure of communication cost used in Xu et al. (2020).

Note that we use the speaker model to compute the listener distribution, instead of the listener model, because given a number the sender is forced to assign positive



(a) Reward: Linear, Prior: Power law

(b) Reward: Inverse , Prior: Power law

(c) Reward: Exponential, Prior: Power law

(d) Reward: Linear, Prior: CAP

(e) Reward: Linear, Prior: Max-Ent

(f) Reward: Linear, Prior: Uniform

Figure 2.3: Term usage vs communication cost. 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. We plot the numeral systems from the human languages presented Table 2.1 and since many of them are very similar we only get a few distinct points for human languages in the figure.

probability to at least one word while the listener can choose to never guess on a number no matter which word is conveyed from the sender. For example the word "many" might refer to a large, or possible infinite, of numbers while the listener may choose to only guess on small subset of these numbers given that "many" has been uttered. Another argument for computing the listener distribution using Bayes rule is because, given a sender distribution, it minimizes the communication cost in the information bottleneck framework presented in Zaslavsky, Kemp, Regier, et al. (2018). The proof of this is presented in the supplementary files of Zaslavsky, Kemp, Regier, et al. (2018).

4 Experiments

We consider the interval $\mathcal{N} = [1, 20]$ and each agent is modelled as a feed-forward neural network with one hidden layer consisting of 50 hidden neurons with a dropout

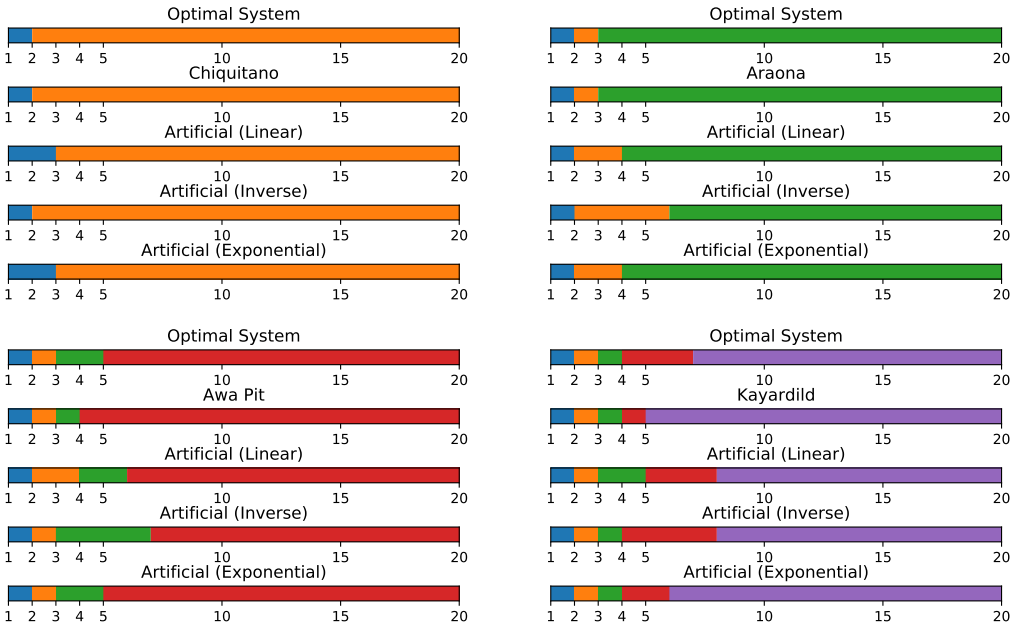


Figure 2.4: Comparison between the optimal numeral systems w.r.t communication cost, human systems and the artificial consensus systems produced by our agents under the different reward functions. We considered the experiments using the power-law prior and the optimal systems are computed under this prior. Each color represents a numeral word and the corresponding interval on the number line that the word represents.

rate of $p = 0.3$ and with ReLu activation [2]. The agents starts with a vocabulary $\mathcal{W}$ [3] of size 10 and is trained for 10 000 updates where each update is over a batch of 100 rounds of the signaling game. The weights in the neural networks are updated using a version of stochastic gradient descent called Adam (Kingma and Ba 2014) with an initial learning-rate of 0.001. The dropout rate, learning rate and batch size are in the range of what is commonly used in machine learning. However, we also performed experiments varying these parameters and found the downstream results to be robust.

We estimate the need probability in four different ways and the priors are shown in Figure 2.5a. The power-law prior is computed by first taking the normalized frequencies of English numerals in the Google ngram corpus English 2000 (Michel et al. 2011) and smoothing the frequencies using a power-law distribution as done in Xu et al. (2020). We also derive need probabilities using the capacity-achieving prior (CAP) method (Zaslavsky, Kemp, Regier, et al. 2018), which infer a prior directly from naming data, and by using the maximum-entropy (MaxEnt) method (Zaslavsky, Kemp, Tishby, et al. 2019), which given a naming distribution $p(w|n)$ and word frequencies $p(w)$ computes the maximum-entropy achieving prior $p(n)$ given these constraints. We obtain a universal CAP by first computing a CAP for each exact

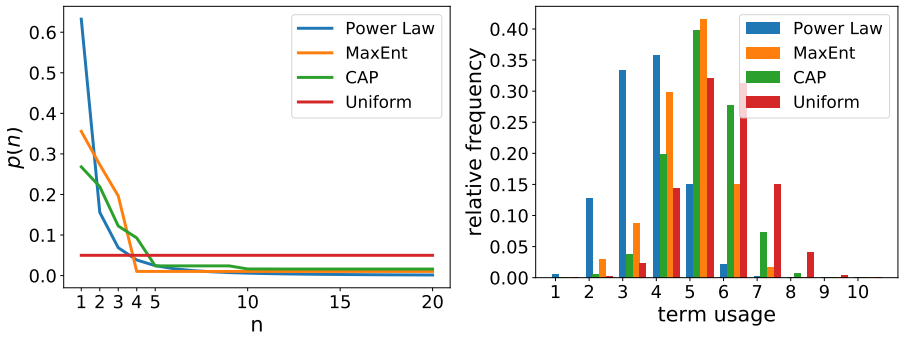
[2]This interval was chosen since the need distributions are exponentially decaying and very little probability mass lies beyond 20, see Figure 2.5a.

[3]The size of the vocabulary $\mathcal{W}$ was taken to be equal to the largest number of terms among the human systems analyzed in Xu et al. (2020), which are presented in Table 2.1.

numeral system presented in Table 2.1 and then averaging them together. Further, to compute a MaxEnt prior we consider the language Gooniyandi, which has four number terms translated to *one, two, three, many*, and the corpus data available for the language Gooniyandi[p. 204] (McGregor 2004). When computing the MaxEnt prior the fourth term, *many*, is modelled as a Gaussian distribution with mean $\mu = 5$ and standard deviation $\sigma = 0.31 \times \mu$. Lastly, we consider an uniform prior which was also done in Xu et al. (2020) and the authors showed that human systems are less optimal under this prior compare to the more skewed power-law prior, illustrating that the near-optimality patterns found in human numeral systems depend critically on the need probability.

We start by training 6000 independent sender-listener pairs under the power-law prior, for each reward function. We then fix the reward function to be linear and train 6000 independent sender-listener pairs for each of the priors CAP, MaxEnt and Uniform. Note that the agents are free to decide how many terms from the vocabulary that are actually used during communication and it is possible for the agents to converge to a numeral system with less than 10 terms. Thus, the actual number of terms in the final numeral system will vary over sender-listener pairs due to randomness in the initialization of the neural networks and the sampling from the need probability.

Following Xu et al. (2020), we compute the convex hull of hypothetical approximate and exact numeral systems to use as baselines. For exact systems this is done using an approach where we start from a random numeral system and greedily updates the system until a local optima is encountered w.r.t communication cost. For approximate systems we proceed in similar fashion but model a numeral word as Gaussian with a mean μ_w and a standard deviation $\sigma = 0.31 \times \mu_w$ following Xu et al. (2020). We start from randomly chosen means and perform greedy updates until a local optima is reached. For both types of systems we solve for both the best and worst performing numeral system and the optimization procedures are repeated 1000 times for each number of terms.



(a) (b)

Figure 2.5: a) The need probabilities, or priors, used. b) Relative frequency of term uses over sender-listener pairs using the linear reward function and varying the need probability. The more left-skewed the need probability is, the fewer terms are generally used by the agents.

Further, we compare the numeral systems developed by our agents to the human

approximate and exact restricted numeral systems considered in Xu et al. (2020) which are presented in Table 2.1. Most of this data was collected from Comrie (2013) except for Chiquitano, Fuyuge, Krenák which comes from Hammarström (2010) and Mundurukú which comes Pica et al. (2004).

Approximate systems:
Chiquitano, Fuyuge, Gooniyandi, Mundurukú, Pirahã, Wari
Exact restricted systems:
Achagua, Araona, Awa Pit, Barasano, Baré, Hixkaryana, Imonda, Kayardild, Krenák, Mangarrayi, Martuthunira, Pitjantjatjara, Rama, Yidiny, !Xóõ

Table 2.1: Human numeral systems considered in Figure 2.3.

In Figure 2.3 we present the performance of our agents, w.r.t communication cost, relative to numeral systems found in human languages and the convex hull of hypothetically possible numeral systems, for the different need probabilities and various reward functions. We observe that our agents produce numeral systems that are near-optimal for all need probabilities and reward functions. For the left-skewed priors we observe that the communication cost of our agents are close to the communication cost of human systems.

Furthermore, in Figure 2.5b we plot the relative frequency of term usages between the sender-listener pairs when using the linear reward function and varying the need probability. As expected, we observe that a more skewed distribution generally results in fewer terms used by the agents which indicates that numeral systems with few terms can be sufficient to achieve a near-optimal reward while we observe a pressure for using more terms under the uniformed need probability.

We use Correlation Clustering (Bansal et al. 2004) to find the consensus numeral system for each number of terms over all experiments. Correlation Clustering is a method for finding the optimal clustering, w.r.t. a similarity measure. We create a 20×20 matrix and each time two numbers i and j belongs to the same partition, or word, over two different sender-listener pairs we increase the element (i, j) of the matrix by 1 otherwise we decrease it with 1. We apply Correlation Clustering to the final matrix to get a consensus system and this will be an exact numeral system. The resulting systems for the experiments using the power-law prior are presented in Figure 2.4 and we observe some similarities between the consensus systems and human systems with the same number of terms. The main difference seems to be that our agents produce systems that tends to be slightly less precise for smaller numbers, especially for the linear reward function, and this could be a result of having reward functions that gives a fair amount of reward for imprecise reconstruction of the number the sender had in mind.

In addition, we compare the representation of numbers developed by our agents to the Gaussian model used in Xu et al. (2020), which is inspired by the the formalization of the approximate number line presented in Pica et al. (2004). The model assumes that a numeral word, w, is represented as a Gaussian distribution with some mean μ_w and standard deviation $\sigma = \nu \times \mu_w$ where ν is the *Weber fraction*. We fit this

model to the distributions produced by our agents by first computing, for each sender-listener pair i, the expected number μ_w^i given a word w under the listener distribution $\mu_w^i = \mathbb{E}_{L_w^i}[n|w]$. We then compute a distribution according to

$$p_\nu^i(n|w) \propto e^{-(\frac{|n-\mu_w^i|}{2\nu \times \mu_w^i})^2}$$

and search for $\nu \in [0.05, 2]$, with a granularity of 0.01, that minimizes the the MSE w.r.t the listener distribution of pair i. The best fitting Weber fractions along with the corresponding MSEs are presented in Table 2.2 and the Gaussian model fits the listener distribution well with an average MSE in the interval $[0.0032, 0076]$ over all the sender-listener pairs. These errors are of the same magnitude as the error reported between the Gaussian model and the numeral system of Mundurukú in Xu et al. (2020) and with similar Weber fraction as reported for Mundurukú adults in Piazza et al. (2013). Hence, our agents produce approximate numeral systems via reinforcement learning which exhibit similar behavior as the Gaussian models used in Xu et al. (2020) and Pica et al. (2004) without being explicitly programmed to do so.

Reward	Best ν	MSE
Linear	0.31	0.0042 ± 0.0036
Inverse	0.31	0.0032 ± 0.0042
Exponential	0.44	0.0076 ± 0.0063

Table 2.2: The Weber fractions corresponding the Gaussian model that on average fits the listener distribution best along with the average MSE ± 1 standard deviation for that Weber fraction, averaged over all sender-listener pairs trained using the particular reward function.

5 Conclusions and future work

We have shown that artificial agents can develop exact and approximate numeral systems, via interaction and reinforcement learning, which are near-optimal in an information-theoretic sense and similar to human systems. Our work offers a mechanistic explanation via reinforcement learning of the results in Xu et al. (2020). More generally, it offers a powerful framework to address fundamental questions of cognition across a wide range of semantic domains using a learning theoretic approach that complements the normative approaches summarized in Kemp, Xu, et al. (2018) and Gibson, Futrell, Piantadosi, et al. (2019).

In the numerals domain, there are still several questions that remain to be explored: Would the results be the same if we increase the range of numbers? Can approximate arithmetic be learned in the same way? Could the recursive systems described in Xu et al. (2020) be learned via interaction? An interesting topic for future work is to establish a rigorous connection between reward function and communication cost in our setup.

In this work our artificial agents have been completely driven by the reward signal. In the future we would like to add a pragmatic reasoning scheme to our model, similar to RSA (Frank and Goodman 2012), and explore what effect this has on the emergent behavior.

6 Acknowledgments

We thank Terry Regier for extremely detailed and valuable feedback on an early draft of this paper. We thank the anonymous reviewers for providing valuable feedback and comments that really improved the final version of the paper. We also thank Meng Liu for clarifying results from Xu et al. (2020), people at CLASP for providing valuable feedback on an early draft of this paper and Harald Hammarström for valuable comments on an early draft and for providing the reference to the corpus data for Goonyandi.

This work was supported by funding from CHAIR (Chalmers AI Research Center) and from the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The computations in this work were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC).

References

Bansal, Nikhil, Avrim Blum, and Shuchi Chawla (2004). "Correlation Clustering". In: *Machine Learning* 56.1, pp. 89–113 (cit. on p. 106).

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. 98).

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 pp. 98, 106).

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. 100).

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. 108).

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 p. 100).

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. 97, 102).

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. ISSN: 1364-6613 (cit. on pp. 97, 98, 107).

Hammarström, H. (Jan. 2010). "Rarities in Numeral Systems". In: *Business Communication Quarterly - Bus Comm Q* (cit. on p. 106).

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 p. 98).

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 p. 97).

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. 97, 98, 107).

Kingma, Diederik P. and Jimmy Ba (2014). *Adam: A Method for Stochastic Optimization*. arXiv: 1412.6980 [cs.LG] (cit. on p. 104).

Lewis, David K. (1969). *Convention: A Philosophical Study*. Wiley-Blackwell (cit. on p. 98).

Marr, D. (1982). *Vision: A Computational Approach*. San Francisco, Freeman & Co. (cit. on p. 97).

McGregor, William B. (2004). *The Languages of the Kimberley, Western Australia*. RoutledgeCurzon (cit. on p. 105).

Michel, Jean-Baptiste et al. (2011). "Quantitative Analysis of Culture Using Millions of Digitized Books". In: *Science* 331.6014, pp. 176–182. ISSN: 0036-8075 (cit. on p. 104).

Niv, Yael, Michael Duff, and Peter Dayan (June 2005). "Dopamine, uncertainty and TD learning". In: *Behavioral and brain functions : BBF* 1, p. 6 (cit. on p. 100).

Piazza, M., P. Pica, V. Izard, E. Spelke, and S. Dehaene (2013). "Education Enhances the Acuity of the Nonverbal Approximate Number System". In: *Psychological Science* 24, pp. 1037–1043 (cit. on p. 107).

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 pp. 98, 106, 107).

Poggio, Tomaso (2012). "The Levels of Understanding framework, revised". In: *Perception* 41, pp. 1017–1023 (cit. on p. 98).

Regier, Terry, Charles Kemp, and Paul Kay (2015). "Word Meanings across Languages Support Efficient Communication". In: *The Handbook of Language Emergence* January 2015, pp. 237–263 (cit. on pp. 97, 98).

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 pp. 98, 100).

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. 100).

Sutton, Richard S. and Andrew G. Barto (1998). *Reinforcement Learning: An Introduction*. Second. The MIT Press (cit. on pp. 98, 99).

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. 100).

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. 97–99, 101, 102, 104–108).

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. 103, 104).

Zaslavsky, Noga, Charles Kemp, Naftali Tishby, and Terry Regier (2019). "Color Naming Reflects Both Perceptual Structure and Communicative Need". In: *Topics in Cognitive Science* 11.1, pp. 207–219 (cit. on pp. 97, 104).

Paper 3

Pragmatic reasoning in structured signaling games

Emil Carlsson, Devdatt Dubhashi.

Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci), 44, 2022.

The paper has been reformatted for uniformity.

Paper 3. Pragmatic reasoning in structured signaling games

Emil Carlsson, Devdatt Dubhashi.

Abstract

In this work we introduce a structured signaling game, an extension of the classical signaling game with a similarity structure between meanings in the context, along with a variant of the Rational Speech Act (RSA) framework which we call structured-RSA (sRSA) for pragmatic reasoning in structured domains. We explore the behavior of the sRSA in the domain of color and show that pragmatic agents using sRSA on top of semantic representations, derived from the World Color Survey, attain efficiency very close to the information theoretic limit after only 1 or 2 levels of recursion. We also explore the interaction between pragmatic reasoning and learning in multi-agent reinforcement learning framework. Our results illustrate that artificial agents using sRSA develop communication closer to the information theoretic frontier compared to agents using RSA and just reinforcement learning. We also find that the ambiguity of the semantic representation increases as the pragmatic agents are allowed to perform deeper reasoning about each other during learning.

Keywords: efficient communication; multi-agent reinforcement learning; pragmatic reasoning

1 Introduction

The Rational Speech Act (RSA) framework (Frank and Goodman 2012; Goodman and Frank 2016) has emerged as a leading probabilistic model of pragmatic communication formalizing the Gricean view on pragmatics (Grice 1975). In RSA models, each agent reasons about the other agent's belief, in a game-theoretic fashion, in order to infer the context dependent meaning of an utterance. Models of this type have been used to make accurate predictions about human behavior over a wide range of different and complex tasks (Goodman and Frank 2016).

It was recently shown by Peloquin et al. (2020) that efficient language use and structure emerge as pragmatic agents interact with each other in a signaling game. In their framework the efficiency was measured as the expected cross-entropy between the speaker and listener distributions.

However, in certain settings, the meaning space may have special structure which needs to be exploited to develop efficient communication. A good example is the domain of colors where it is possible to quantify the similarity between different colors. Hence, in a context where agents are talking about different colors an error

might be quantified differently depending on whether the listener confused the color the speaker was referring to with a very similar color or with a completely different color. This is something that is not captured by a purely entropy-based efficiency measure.

Here we take a new approach to the basic question addressed in Peloquin et al. (2020) about how efficient communication arises via the interaction of pragmatic agents. First, to take structure into account, we introduce a notion of a *structured signaling game*, an extension of the standard signaling game, commonly used in work regarding pragmatic reasoning. For this type of signaling game we introduce an extension of the standard RSA which we call *structured-RSA* (sRSA) where an agent accounts for the structure in the meaning space during the reasoning process. We explore the differences between RSA and sRSA in the color domain, a domain commonly used in cognitive science to explore various linguistic phenomena (Regier, Kemp, et al. 2015; Gibson et al. 2017). Second, we quantify the efficiency of the resulting communication schemes using the information theoretic notions of efficiency from Zaslavsky, Kemp, et al. (2018) and the well-formedness measure from Regier, Kay, et al. (2007).

We first investigate the use of human representations such as the color naming systems found in the World Color Survey (Cook et al. 2005) as a basis for reasoning by pragmatic agents. We show that efficiency of communication increases much more when agents reason using sRSA compared to agents using RSA and base policies. The most striking result is that sRSA agents initialized with human representations only need a recursion depth of 1 or 2 in order to come very close to the optimal frontier.

Next, we consider computational learning agents interacting with each other in a multi-agent reinforcement learning framework similar to those considered in Kågebäck et al. (2020), Chaabouni et al. (2021), Carlsson et al. (2021), and Ohmer et al. (2022). Our results in this learning framework suggest that pragmatic agents equipped with sRSA learn more efficient color naming systems compared to agents using RSA or pure reinforcement learning. We also find that ambiguity arises to a greater extent in the semantic representation as the computational agents are allowed to perform deeper reasoning about each other. Even though the ambiguity increases, the computational agents using sRSA still develop efficient and accurate communication. Compared to previous works (Monroe et al. 2017; Kågebäck et al. 2020; Chaabouni et al. 2021; Hu et al. 2021), which only account for the structure of the color space in the non-contextual meaning function. Our approach extends this and explicitly accounts for structure in the RSA recursion.

The work of Zaslavsky, Hu, et al. (2021) is also related to our work. They use the fact that the softmax operator maximizes a trade-off between utility and entropy (Fudenberg and Levine 1998) to argue that the RSA recursion can be viewed as an alternating maximization of a least-effort objective. They ground the recursion in Rate-Distortion theory and derive a new update of the sender based on the mutual information between meaning and utterance. In contrast to their work, our sRSA is based on the standard RSA recursion, with the difference that our utility function leverages the pair-wise similarity, or distortion, between meanings in the context.

2 Structured signaling games and sRSA

In our signaling game, two agents, one sender and one listener, observe a context of n meanings $\mathcal{C} = \{m_i\}$ where each m_i lies in some meaning space $\mathcal{M}$. The goal of the sender is to describe one of the meanings to the listener. In the standard setup of a signaling game, the agents share a semantic representation, or meaning function, $\mathcal{L}(m, w)$, which describes how well the utterance w describes the object m. In our structured version we also assume that the agents share a similarity matrix Z where element Z_{ij} describes how similar meanings m_i and m_j are. We assume $Z_{ij} \in [0, 1]$ with $Z_{ii} = 1$. An example of a structured signaling game in the domain of colors is presented in Figure 3.1.

2.1 Similarity-sensitive utility and sRSA

Following Degen et al. (2020), we consider agents equipped with a continuous meaning function, or semantic representation, $\mathcal{L}(m, w) \in [0, 1]$ which describes how well a meaning m can be mapped to an utterance w. On top of the meaning function, our agents use the RSA in order to reason about each other's behavior given the context C. Given a literal listener proportional to the meaning function, $L_0(m|w) \propto \mathcal{L}(m, w)$, the following recursion is applied in the RSA

$$S_t(w|m, C) \propto e^{\alpha U_t(m,w,C)} \tag{2.1}$$

$$L_t(m|w, C) \propto S_t(w|m, C)p(m|C) \tag{2.2}$$

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. In RSA the utility of the sender is usually based on reducing the epistemic uncertainty the listener carries about the true meaning, and is taken to be the negative surprisal of the listener $U_t(w, m, C) = \log L_{t-1}(m|w, C)$. We will denote an agent using RSA at a recursion depth of t with parameter α as $RSA(t, \alpha)$.

Similarity-sensitive surprisal

Leinster (2021) recently introduced extensions of entropy and other information theoretic concepts in the context of structured domains, where one has a matrix of similarities Z. Inspired by this, we define the *similarity-sensitive surprisal* of a listener, L, as

$$I^Z(m, w, C) = -\log \sum_{m'} Z_{mm'} L(m'|w, C). \tag{2.3}$$

Here $Z(m, m')$ is the similarity between the two meanings m and m'. This measure captures the desirable property that a listener shouldn't be as surprised if a speaker used the same word for two similar colors compared to if the speaker used the same word for two very different colors.

Defining the utility as $U(m, w) = -I^Z(m, w)$ we arrive at structured version of RSA (sRSA) with similarity-sensitive sender. Note that this utility yields a sender proportional to the power α of the expected similarity

$$S_t(w|m, \mathcal{C}) \propto (\sum_{m' \in \mathcal{C}} Z_{mm'} L_{t-1}(m'|w))^\alpha. \tag{2.4}$$

In next section 3 and in Figure 3.1 we give a simple example in the color domain to illustrate the difference between RSA and sRSA.

In the special case where Z is the identity matrix, i.e. where meanings in the context share no similarity, (2.3) reduces to the standard surprisal and the sender in (2.4) reduces to the standard RSA sender. We will denote an agent using sRSA at a recursion depth of t with parameter α as $sRSA(t, \alpha)$.

In general, given a distortion measure on the meaning space $d : \mathcal{M} \times \mathcal{M} \to \mathbb{R}^+$, we can construct a natural similarity measure as $Z_{mm'} := e^{-\beta d(m,m')}$, $\beta > 0$.

3 Color domain: Efficiency and well-formedness

We will use colors as our testbed for pragmatic reasoning in structured signaling games. The seminal work of Zaslavsky, Kemp, et al. (2018) showed that color naming systems in the World Color Survey (WCS) (Cook et al. 2005) optimize an information-theoretic trade-off between complexity and accuracy of the meaning function. Following Zaslavsky, Kemp, et al. (2018) we will take the complexity of a color naming system as the mutual information between word and meaning

$$\text{Complexity} = I(M; W)$$

and the accuracy as

$$\text{Accuracy} = I(W; U).$$

As in Zaslavsky, Kemp, et al. (2018) we assume a meaning m to be a distribution over color chips proportional to a isotropic Gaussian, $m(u) \propto e^{-\frac{1}{64}||x_m - x_u||^2}$ where x_m is the CIELAB vector corresponding to color chip m.

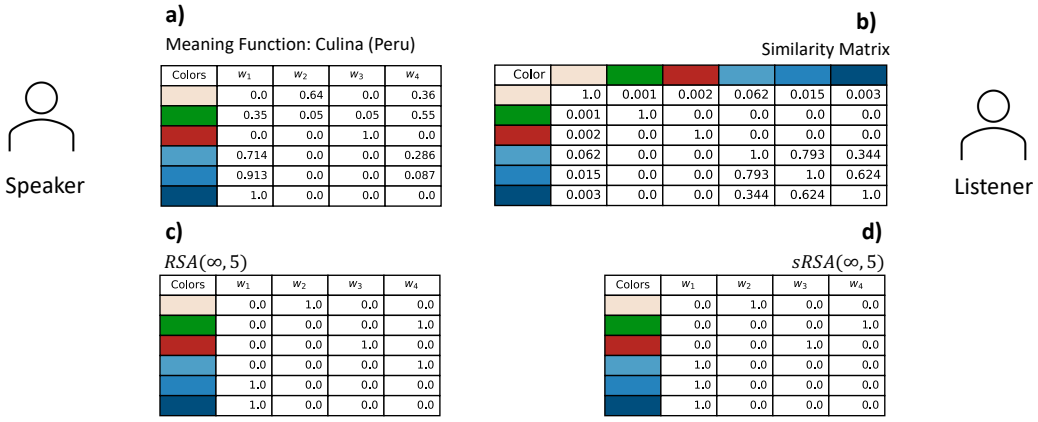
Regier, Kay, et al. (2007) showed also that human color naming reflects optimal partitions of the color space w.r.t. to a measure of *well-formedness*. The well-formedness criterion was based on the following measure of perceptual similarity between colors

$$\text{sim}(m, m') = e^{-0.001||x_m - x_{m'}||^2} \tag{3.1}$$

This similarity measure will be used in our sRSA model in the downstream analysis.

sRSA vs RSA

Figure 3.1 gives a simple example of a structured signaling game where the context consists of 6 different colors. The meaning function mapping color to word is based



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 3.1: An example of a structured signaling game in the color domain.

on the naming data found in the World Color Survey for the language Culina is shown in Figure 1a. The similarity matrix, which describes how similar two colors are w.r.t. the similarity measure defined in (3.1), is shown in Figure 1b. We use $RSA(t, \alpha)$ to denote the result of applying depth t RSA and $RSA(\infty, \alpha)$ to denote the limit as $t \to \infty$, and similarly for sRSA. Figure 1c and Figure 1d show the limit points for RSA and sRSA (with $\alpha = 5$). 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.

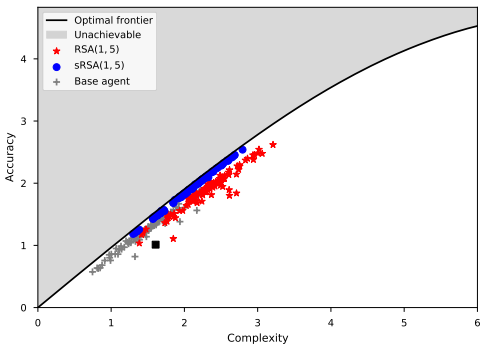
3.1 Human representations

The WCS data consist of naming data from 110 languages, with an average of 25 speakers for each language. Since the WCS data contain data from speakers, we believe it is more appropriate to consider a slightly different version of the RSA recursion, where the agents start reasoning from a literal sender proportional to the naming data from WCS[1]. For a language l in the WCS study and corresponding naming data $D^l(w, m)$ we consider the following recursion

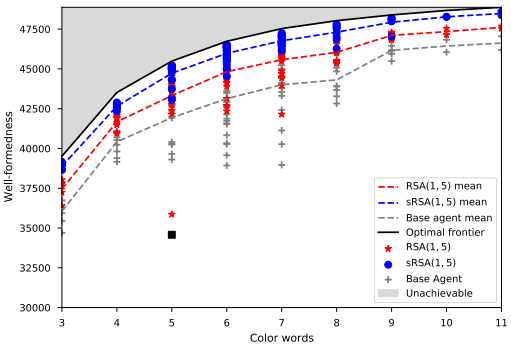
$$\begin{aligned} S_0^l(w, m, C) &\propto D^l(w, m) \\ L_t^l(m|w, C) &\propto S_{t-1}^l(w|m, C)p(m|C) \\ S_t^l(w|m) &\propto e^{U_t(w,m,C)}. \end{aligned}$$

We consider a structured signaling game with the context, $\mathcal{C}$, being the entire Munsell chart. Hence, a sender is given a certain color chip from the Munsell chart

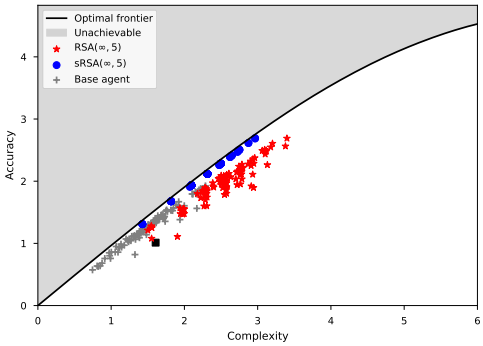
[1]As in Regier, Kemp, et al. (2015), we only consider major color terms. We say that a color term is major if it is the mode category for at least 10 chips in the Munsell Chart.



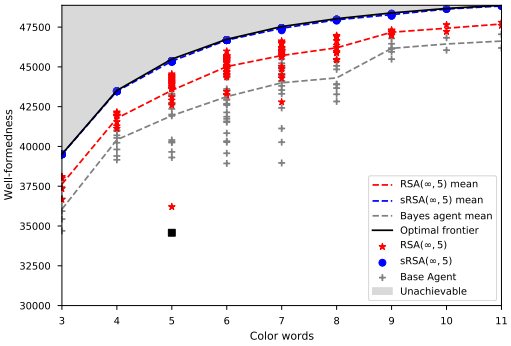
(a) Information-theoretic trade-off between complexity and accuracy after one recursion.



(b) Well-formedness of the agents after one recursion.



(c) Information-theoretic trade-off between complexity and accuracy in the limit.

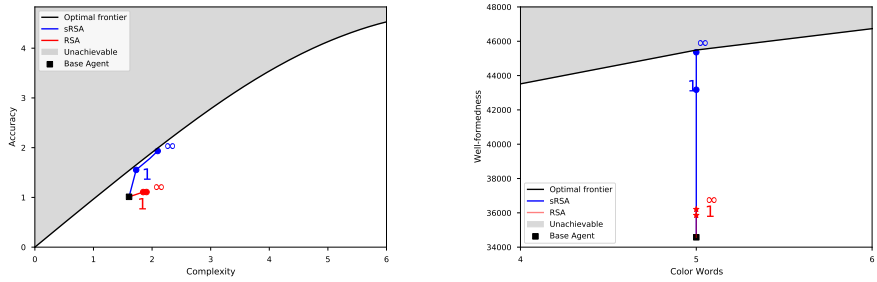


(d) Well-formedness as the recursions goes to infinity.

Figure 3.2: Results for applying pragmatic reasoning on-top of the color naming data in WCS. Here depth of recursion indicates the depth of the final sender in the recursion. We use $\alpha = 5$ in the recursions. The black square indicates the position of the base agent of the language Karajá.

and should describe this to the listener, which then produces a distribution over the color chips in the chart. The context we consider here is much larger compared to the ones considered in, for example, Monroe et al. (2017). The reason is that we are interested in larger contexts where the number of meanings is much larger than the number of utterances and exact communication is impossible. We will consider a uniform need distribution over the chart and leave it for future work to study skewed priors like the one used in Zaslavsky, Kemp, et al. (2018). As a baseline we will consider the base agents from the recursion, i.e. a sender proportional to the naming data and the corresponding Bayesian listener. The information-theoretic frontier is computed using the Blahut-Arimoto algorithm with the annealing scheme outlined in Zaslavsky, Kemp, et al. (2018) and a uniform prior. The well-formedness frontier is computed using the Correlation Clustering approach described in Kågebäck et al. (2020).

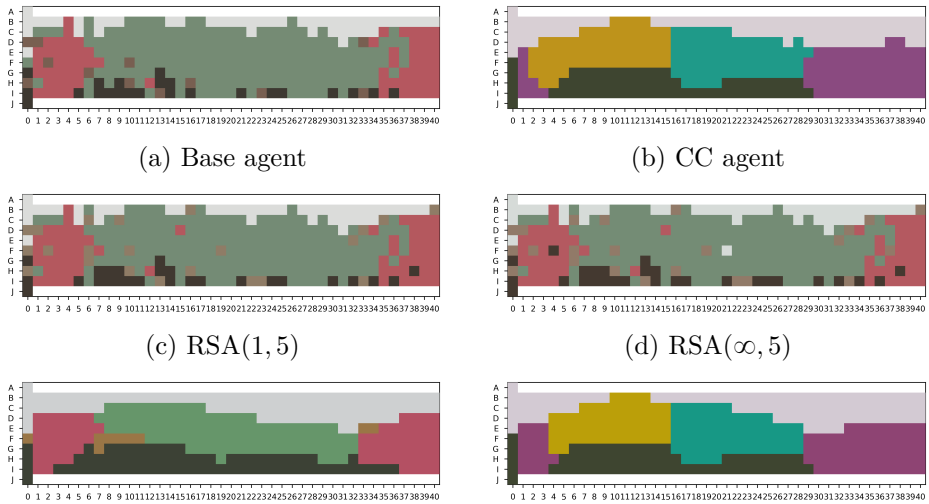
In Figure 3.2a we compare the efficiency of the base agents to the efficiency of the pragmatic agents after performing one recursion in the respective reasoning model. We observe that *pragmatic reasoning leads to more complex and accurate behavior*



(a) Efficiency trade-off (b) Well-formedness

Figure 3.3: Trajectories of RSA and sRSA for Karajá.

for both RSA and sRSA compared to the base agents. However, we also observe that the RSA agents have not moved closer to the optimal frontier while the sRSA agents are very close to the frontier *after only one recursion*. Interestingly, when the recursions are allowed to go the limit, Figure 3.2c, the RSA agents seem to move away from the optimal frontier while the sRSA converges to naming distributions very close to the optimal frontier.



(a) Base agent (b) CC agent

(c) RSA$(1, 5)$ (d) RSA$(\infty, 5)$

(e) sRSA$(1, 5)$ (f) sRSA$(\infty, 5)$

Figure 3.4: Karajá, Brazil. The sRSA model refine and smooth the colormap in only one recursion. In the limit, we observe that the sRSA approaches the true optimal agent w.r.t. well-formedness (CC Agent). Each color term is colored with the average color mapped to the term.

Further, Figure 3.2b illustrates the well-formedness of the agents after one recursion. The pragmatic agents greatly improve the well-formedness of the base agents *after only one recursion*. As observed for efficiency as well, we see that sRSA, which takes the structure into account, improves the well-formedness to a greater extent. In the limit, see Figure 3.2d, the sRSA agents converge to optimal naming distributions w.r.t. the well-formedness criterion.

Many studies, including the recent one in Frank, Emilsson, et al. (2021), have

reported that humans rarely use more than 1 or 2 levels of recursion in signaling games. It is therefore intriguing that the sRSA only needs only 1 or 2 recursions to reach the information-theoretic frontier. We believe this is something worth exploring further in the future.

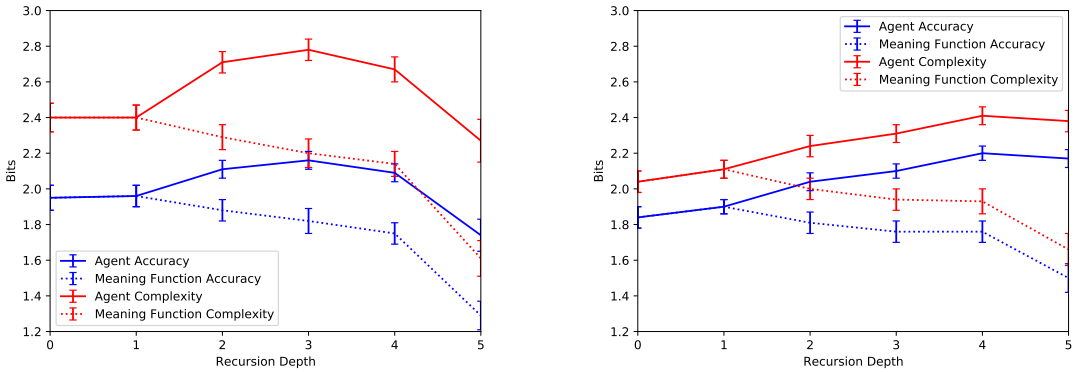
An outlier, when it comes to both efficiency and well-formedness, is the base agent of the language Karajá, highlighted by the black square in Figures 3.2a and 3.2b. In Figure 3.3 we illustrate the efficiency and well-formedness of the corresponding RSA and sRSA agents as we increase the recursion depth. Interestingly, applying a few steps of sRSA, see Figure 3.3, yields a near-optimal agent, both when it comes to well-formedness and efficiency. This suggests that even though the naming distribution of Karajá is not efficient and well-formed in itself, it serves as a good initialization for a pragmatic and rational agent - but for an agent that takes domain structure into account. Without taking the structure into account, the RSA agent doesn't lead to a more efficient behavior; instead the RSA agent seems to be moving away from the optimal frontier.

In Figure 3.4, we see the corresponding mode-maps for the different RSA versions at depth 1 and in the limit. We clearly see that taking the structure into account in the reasoning process produces agents that have very smooth mode-maps already at depth 1, see Figure 3.4e. Here we also see that the standard RSA objective, see Figures 3.4c and 3.4d, fails to produce smooth mode-maps since it does not account for the structure of the domain space. Worth highlighting is that the sRSA, Figure 3.4f, seems to converge to a mode-map very close to the optimal mode-map w.r.t. the well-formedness measure, see Figure 3.4b. This is perhaps expected since the sRSA utility considers perceptual similarity.

3.2 Artificial agents

In our multi-agent reinforcement learning framework, two agents will play a structured signaling game about colors. In the beginning of each game, one agent is randomly assigned to be the speaker agent and the other one acts as a listener. Each agent will keep their own parameterization of the meaning function $\mathcal{L}_\theta$ using a neural network with parameters θ and ϕ. Given a context, both agents will apply either RSA or sRSA on the meaning function for t iterations to get their corresponding policies $S_{t,\theta}(w|m,\mathcal{C})$ and $L_{t,\phi}(m|w,\mathcal{C})$. The speaker agent then samples an utterance given the target according to $S_{t,\theta}(w|m,\mathcal{C})$, and upon receiving the utterance, the listener samples a guess according to the distribution $L_{t,\phi}(m|w,\mathcal{C})$. A binary reward is given to both agents depending on whether the listener produced a correct guess and both agents will update their respective meaning function using the REINFORCE objective (Williams 1992), which for the sender agent corresponds to taking the gradient of $r \log S_{t,\theta}(w|m,\mathcal{C})$ and for the listener gradient of $r \log L_\theta(m|w,\mathcal{C})$. A similar computational setup was recently considered in Ohmer et al. (2022).

We take each neural network to have one hidden layer of 25 neurons with ReLU activation for the hidden layer and sigmoid activation in the output layer. We train the agents on contexts consisting of 5 colors sampled from the Munsell chart and represented as a vector in CIELAB space. We vary the depth of the agent from 0



(a) RSA (b) sRSA

Figure 3.5: In the following plots, depth indicates the level of the final listener in the recursion, and the error bars correspond to the width of the 95% confidence interval. We observe that, as the depth of recursion increases, the accuracy and complexity of the agent differs more compared to the accuracy and complexity of the corresponding meaning function. Noteworthy is that 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. The efficiency and accuracy of the agents and meaning functions should be the same at depth 0 and 1 since both correspond to the sender $S_1(w|m)$.

to 5, where depth 0 indicates a sender interacting with a literal listener, and we set $\alpha = 5$. During the evaluation, the context given to the agents will be the entire Munsell chart, as was done for human representations. Each configuration of agents is averaged over 100 different random seeds. We update the neural networks using standard stochastic gradient descent, with the learning rate set to 0.001. The agents were trained for 10 000 updates using a batch size of 100. We compare the results to a pure reinforcement learning baseline (RL) with the meaning function of the same size as that of the pragmatic agents, but with linear activation in the output layer. The RL sender performs a softmax operation over words given a color, and the RL listener performs a softmax operation over colors given a word. This color game is similar to the ones considered in Kågebäck et al. (2020) and Chaabouni et al. (2021) with the difference that the sender observes the context in our setup.

In Figure 3.6, we observe the efficiency of the agents when performing 2 recursions. The RSA agents develop less efficient communication compared to the sRSA agents and the RL baseline. The sRSA agents develop communication closer to the optimal frontier compared to the RL and RSA agents, illustrating that pragmatic agents with appropriate utility functions develop efficient communication. It is worth highlighting that the RSA and RL agents account for the structure of the color space in their non-contextual meaning functions, i.e. in their neural networks. The results in Figure 3.6 thus suggest that the efficiency of the sRSA agents cannot be mimicked by just a graded, or fuzzy, meaning function, but is due to explicitly accounting for the structure in the recursion. We also note that the non-pragmatic RL baseline learns color naming systems which are more efficient than the pragmatic RSA agents, and that these systems are also close to the information-theoretic frontier (the efficiency

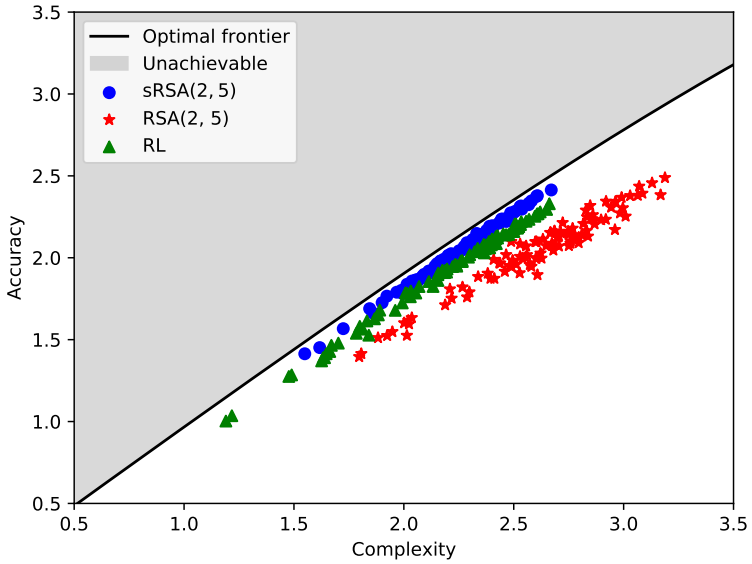


Figure 3.6: The efficiency of the RSA and sRSA agents trained using a recursion depth of 2 compared to the RL baseline.

of RL agents w.r.t. this objective was first reported in Chaabouni et al. (2021)).

In Figure 3.5 we see how the complexity and accuracy of the agents and the meaning function changes as the agents are allowed to perform deeper reasoning during learning. As the recursion depth increases, the sRSA agents develop more complex and accurate behavior while ambiguity emerges to a higher extent in the corresponding meaning functions, see Figure 3.5b. Hence, the sRSA agents are able to use ambiguity as a tool to reach greater communicative efficiency. This is consistent with the observations in Peloquin et al. (2020) and the claims in Piantadosi et al. (2012) that ambiguity is associated with efficient communication. The ambiguity of the meaning function increases with recursion depth, also for the RSA agents, which can be seen in Figure 3.5a. However, for the RSA agents we also observe that the accuracy and complexity of the agent decreases after a few recursions, which seems to indicate that a small number of recursions is better for developing accurate behavior compared to higher recursion depth when using RSA.

4 Conclusions

In this work we have explored pragmatic reasoning in a structured signaling game in the color domain. We explored human representations from the World Color Survey, as well as representations learned by artificial agents using reinforcement learning that incorporate pragmatic reasoning. We have seen that, in both cases, incorporating the domain structure in the reasoning process greatly improves the efficiency in the standard information-theoretic sense, compared to using the standard RSA recursion.

We believe that an interesting future direction is to extend the idea of a structured signaling game and sRSA to more complex environments. An example is a scenario where meanings constitute several different features, and not just one, as considered here. Another interesting future direction, pointed out by one of the reviewers, is to explore scenarios where agents do not share the exact same notion of similarity.

5 Acknowledgements

We thank Terry Regier and the reviewers for providing valuable input on this work. We also want to thank Fredrik D. Johansson, Emilio Jorge and Niklas Åkerblom for providing valuable comments on a previous draft of this paper.

This work was supported by funding from Chalmers AI Research Center (CHAIR) and the computations in this work were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC).

References

Carlsson, Emil, Devdatt Dubhashi, and Fredrik D. Johansson (2021). "Learning Approximate and Exact Numeral Systems via Reinforcement Learning". In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 43 (cit. on p. 114).

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 pp. 114, 121, 122).

Cook, Richard S., Paul Kay, and Terry Regier (2005). "Chapter 9 - THE WORLD COLOR SURVEY DATABASE". In: *Handbook of Categorization in Cognitive Science*. Ed. by Henri Cohen and Claire Lefebvre. Oxford: Elsevier Science Ltd, pp. 223–241 (cit. on pp. 114, 116).

Degen, Judith, Robert Hawkins, Caroline Graf, Elisa Kreiss, and Noah Goodman (Apr. 2020). "When redundancy is useful: A Bayesian approach to "overinformative" referring expressions". In: *Psychological Review* 127 (cit. on p. 115).

Frank, M., A.G. Emilsson, B. Peloquin, N. Goodman, and C. Potts (2021). "Rational speech act models of pragmatic reasoning in reference games". In: *psyarxiv* (cit. on p. 119).

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. 113).

Fudenberg, Drew and David Levine (1998). *The Theory of Learning in Games*. 1st ed. Vol. 1. The MIT Press (cit. on p. 114).

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 p. 114).

Goodman, Noah D. and Michael C. Frank (2016). "Pragmatic Language Interpretation as Probabilistic Inference". In: *Trends in Cognitive Sciences* 20.11, pp. 818–829 (cit. on p. 113).

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. 113).

Hu, Jennifer, Roger Levy, and Noga Zaslavsky (2021). "Scalable pragmatic communication via self-supervision". In: *ICML Workshop on Self-Supervised Learning for Reasoning and Perception* (cit. on p. 114).

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. 114, 118, 121).

Leinster, Tom (2021). *Entropy and Diversity The Axiomatic Approach*. Cambridge University Press (cit. on p. 115).

Monroe, Will, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts (2017). "Colors in Context: A Pragmatic Neural Model for Grounded Language

Understanding". In: *Transactions of the Association for Computational Linguistics* 5, pp. 325–338 (cit. on pp. 114, 118).

Ohmer, Xenia, Michael Franke, and Peter König (2022). "Mutual Exclusivity in Pragmatic Agents". In: *Cognitive Science* 46 (cit. on pp. 114, 120).

Peloquin, Benjamin, Noah Goodman, and Michael Frank (Jan. 2020). "The Interactions of Rational, Pragmatic Agents Lead to Efficient Language Structure and Use". In: *Topics in Cognitive Science* 12, pp. 433–445 (cit. on pp. 113, 114, 122).

Piantadosi, Steven T., Harry Tily, and Edward Gibson (2012). "The communicative function of ambiguity in language". In: *Cognition* 122.3, pp. 280–291 (cit. on p. 122).

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. 114, 116).

Regier, Terry, Charles Kemp, and Paul Kay (2015). "Word Meanings across Languages Support Efficient Communication". In: *The Handbook of Language Emergence* January 2015, pp. 237–263 (cit. on pp. 114, 117).

Williams, Ronald J. (1992). "Simple statistical gradient-following algorithms for connectionist reinforcement learning". In: *Machine Learning* 8.3, pp. 229–256 (cit. on p. 120).

Zaslavsky, Noga, Jennifer Hu, and Roger Levy (2021). "A Rate–Distortion view of human pragmatic reasoning". In: *Proceedings of the Society for Computation in Linguistics* 4. DOI: doi.org/10.7275/gc1z-ck09 (cit. on p. 114).

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. 114, 116, 118).

Paper 4

Cultural evolution via iterated learning and communication explains efficient color naming systems

Emil Carlsson, Devdatt Dubhashi, Terry Regier.

To appear in the Journal of Language Evolution. Earlier version in Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci) 45, 2023.

The paper has been reformatted for uniformity.

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

Emil Carlsson, Devdatt Dubhashi, Terry Regier.

Abstract

It has been argued that semantic systems reflect pressure for efficiency, and a current debate concerns the cultural evolutionary process that produces this pattern. We consider efficiency as instantiated in the Information Bottleneck (IB) principle, and a model of cultural evolution that combines iterated learning and communication. We show that this model, instantiated in neural networks, converges to color naming systems that are efficient in the IB sense and similar to human color naming systems. We also show that some other proposals such as iterated learning alone, communication alone, or the greater learnability of convex categories, do not yield the same outcome as clearly. We conclude that the combination of iterated learning and communication provides a plausible means by which human semantic systems become efficient.

Keywords: cultural evolution; iterated learning; efficient communication; semantic categories; color naming

1 Introduction

Semantic categories vary across languages, and it has been proposed that this variation can be explained by functional pressure for efficiency. On this view, systems of categories are under pressure to be both simple and informative (e.g. Rosch (1978)), and different languages arrive at different ways of solving this problem, yielding wide yet constrained cross-language variation. There is evidence for this view from semantic domains such as kinship (Kemp and Regier 2012), container names (Xu, Regier, et al. 2016), names for seasons (Kemp, Gaby, et al. 2019), indefinite pronouns (Denić, Steinert-Threlkeld, et al. 2022), modals (Imel and Steinert-Threlkeld 2022), and numeral systems (Xu, Liu, et al. (2020), and relatedly Denić and Szymanik (2024)). Zaslavsky, Kemp, et al. (2018) gave this proposal an independent theoretical foundation by grounding it in an information-theoretic principle of efficiency, the Information Bottleneck (IB) principle (Tishby et al. 1999); they also showed: (1) that color naming systems across languages are efficient in the IB sense, (2) that optimally IB-efficient systems resemble those found in human languages, and (3) that the IB principle accounts for important aspects of the data that had eluded earlier explanations. Subsequent work has shown that container naming (Zaslavsky, Regier, et al. 2019), grammatical categories of number, tense, and evidentiality (Mollica et al.

2021), and person systems (Zaslavsky, Maldonado, et al. 2021) are also efficient in the IB sense.

In a commentary on this line of research, Levinson (2012) asked how semantic systems evolve to become efficient, and suggested that an important role may be played by iterated learning (e.g. Scott-Phillips and Kirby (2010)). In iterated learning, a cultural convention is learned by one generation of agents, who then provide training data from which the next generation learns, and so on. The convention changes as it passes through generations, yielding a cultural evolutionary process. The idea that such a process could eventually lead to efficient semantic systems has since been explored and broadly supported. Xu, Dowman, et al. (2013) showed that chains of human learners who were originally given a randomly generated color category system eventually produced systems that were similar to those of the World Color Survey (Cook et al. (2005)), a large dataset of color naming systems from 110 unwritten languages. Although this study did not directly address efficiency, Carstensen et al. (2015) drew that link explicitly: they reanalzyed the data of Xu, Dowman, et al. (2013) and showed that the color naming systems produced by iterated learning not only became more similar to those of human languages – they also became more informative; the same paper also presented analogous findings for semantic systems of spatial relations. In response, Carr et al. (2020), building on earlier work by Kirby et al. (2015) and others, argued that iterated learning primarily contributes simplicity rather than informativeness — but that a bias for simplicity can nonetheless sometimes result in an increase in informativeness. Overall, there is support for the idea that iterated learning can lead to efficient semantic systems, with continuing debate over how and why. There are also recent proposals that non-iterated learning – e.g. in the context of a dyad of communicating agents (e.g. Kågebäck et al. (2020), Chaabouni et al. (2021), and Tucker et al. (2022)), or in a single agent without communication (e.g. Steinert-Threlkeld and Szymanik (2020) and Gyevnar et al. (2022)) – can explain efficient color naming systems. In particular, Steinert-Threlkeld and Szymanik (2020) argued that "[e]ase of learning explains semantic universals" (see also Gentner and Bowerman (2009)). To support this claim, Steinert-Threlkeld and Szymanik (2020) first noted that earlier proposals (e.g. Gärdenfors (2000) and Jäger (2010)) had argued for the importance of convexity in conceptual space as an important constraint on human semantic categories; they then demonstrated the greater learnability, in a neural network, of convex as opposed to non-convex color categories. These recent contributions, and the present one, build on an important line of earlier work using agent-based simulations cast as evolutionary models, without explicitly addressing efficiency (e.g. Steels and Belpaeme (2005), Belpaeme and Bleys (2005), Dowman (2007), Jameson and Komarova (2009), and Baronchelli et al. (2010)).

Several of these prior studies have engaged efficiency in the IB sense, and two are of particular relevance to our own work. Chaabouni et al. (2021) showed that a dyad of neural network agents, trained to discriminate colors via communication, eventually arrived at color naming systems that were highly efficient in the IB sense. However, these systems did not always resemble those of human languages: their categories "depart to some extent from those typically defined by human color naming"

(Chaabouni et al. (2021), p. 11 of SI). Tucker et al. (2022) explored a similar color communication game, and found that their neural agents gravitated to color naming systems that are both essentially optimally efficient in the IB sense, and similar to human color naming systems from the WCS. They achieved this by optimizing an objective function that is based on the IB objective. To our knowledge, earlier work leaves open whether both high IB efficiency and similarity to human languages can be achieved through processes and principles that are independent of IB. We explore that question here. We also wish to establish here whether such independent principles may address the one case in which IB-optimal color naming systems deviate to some extent from empirical observation: the case of 3-term systems (Zaslavsky, Kemp, et al. 2018, p. 7941). Overall, we wished to ascertain whether a natural model of cultural evolution might account both for the many cases in which IB matches the data, and for the one case in which it deviates from the data to some extent.

A natural candidate model of cultural evolution was advanced by Kirby et al. (2015), and the ideas we pursue here build on that general model. Specifically, Kirby et al. (2015) proposed a model of cultural evolution that interleaves two kinds of learning touched on above: (1) learning that occurs during transmission of a linguistic system from one generation to the next, and (2) learning that occurs during communication among agents within a single generation. That formulation allowed them to isolate the effect of each of the two kinds of learning, and to examine their combination. They were interested in particular in what evolutionary forces could give rise to compositional structure of the sort found in human language. In computational simulations and experiments with human participants, they found that transmission from one generation to the next exerted pressure for simplicity, that within-generation communication exerted pressure for informativeness — and that only the two forces operating together gave rise to compositional structure. Here, we apply the same general cultural evolutionary model to a different question, that of color naming systems in human languages.

In what follows, we first demonstrate that there exist many possible color naming systems that are highly efficient in the IB sense, but do not closely resemble human systems. The fact that there exist such efficient-yet-not-human-like systems is not surprising given that IB is a non-convex optimization problem (Tishby et al. 1999; Zaslavsky, Kemp, et al. 2018), but appreciating the prevalence of such systems may be helpful in understanding how Chaabouni et al. (2021) achieved high IB efficiency with systems that deviate from human ones. We then show that the general cultural evolutionary model of Kirby et al. (2015), instantiated in neural networks (Ren et al. 2020), gravitates toward efficiency and, within the class of efficient systems, gravitates more toward human color naming systems than toward others. Finally, we show that iterated learning alone, communication alone, and convexity alone, do not yield that outcome as clearly. We conclude that iterated learning and communication jointly provide a plausible explanation of how human color naming systems become efficient.

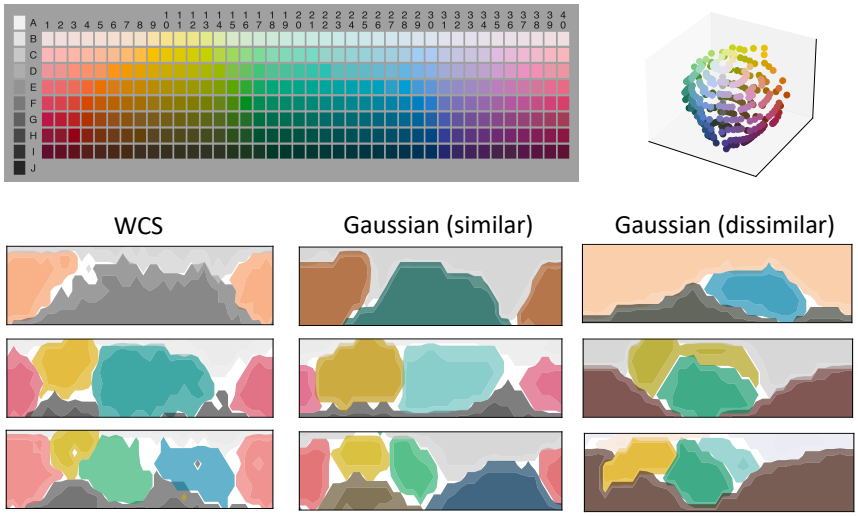


Figure 4.1: Top: Color naming stimulus grid (left), and stimuli plotted in CIELAB space (right). Bottom: 9 color naming systems displayed relative to the grid. The left column contains color naming systems from 3 languages in the World Color Survey (WCS). Colored regions indicate category extensions, 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 Gaussian systems of complexity comparable to that of the WCS system in the same row. The middle column shows random Gaussian systems that are similar to the WCS system in the same row. The right column shows random Gaussian systems that are dissimilar to the WCS system in the same row; at the same time, there is no other WCS system that is more similar to this Gaussian system.

2 Not all efficient systems are human-like

We considered a natural class of artificial color naming systems (see e.g. Abbott et al. (2016) and Zaslavsky, Garvin, et al. (2022)). In this class, each named category w is modeled as a spherical Gaussian-shaped kernel with mean (prototype) x_w in 3-dimensional CIELAB color space (Figure 4.1, top right panel), such that the distribution over words w given a color chip c at location x_c in CIELAB space is:

$$S(w|c) \propto e^{-\eta||x_c - x_w||_2^2} \tag{2.1}$$

where $\eta > 0$ is a parameter controlling the precision of the Gaussian kernel. We then generated artificial color category systems with $K = 3 \ldots 10$ categories each, by first sampling η randomly from a uniform distribution over the interval $[0.001, 0.005]$ for each system, using the same η for all categories in a given system, and then sampling the prototype x_w of each category w randomly, without replacement, from a uniform distribution over the cells of the color naming grid shown in the top left panel of Figure 4.1. This figure shows the same set of colors as in the top right panel, but now in a 2-D array. In analyzing these systems, we draw on four quantities from the IB framework as presented by Zaslavsky, Kemp, et al. (2018) and reviewed below in Appendix A: the complexity of a category system, the accuracy of a category system, ϵ (a measure of the inefficiency of a category system, or its deviation from

the theoretical limit of efficiency), and gNID (a measure of dissimilarity between two category systems). We noted that the range of complexity (in the IB sense) for systems in the World Color Survey (WCS) was $[0.84, 2.65]$, and also noted that our random model sometimes generated systems outside this range; we only considered artificial systems with complexity within this range, and generated 100 such systems for each K; we refer to these randomly-generated systems as Gaussian systems.

The lower panels of Figure 4.1 compare natural color naming systems to artificial Gaussian systems. The leftmost column shows three attested color naming systems from the World Color Survey (WCS), from top to bottom: Bété (iso: bev, Côte d'Ivoire), Colorado / Tsafiki (iso: cof, Ecuador), and Dyimini (iso: dyi, Côte d'Ivoire). The middle column shows randomly-generated Gaussian systems that are similar to the WCS system in the same row, and the rightmost column shows Gaussian systems that are dissimilar to the WCS system in the same row but of about the same complexity. In each row, the rightmost system, which is dissimilar to the WCS system in that row, is nonetheless more similar to that WCS system than to any other WCS system; this means it is dissimilar to all WCS systems. Thus, there exist Gaussian systems that are quite similar to naturally occurring systems, and other Gaussian systems that are quite dissimilar to naturally occurring systems. To quantify this pattern, we separated the Gaussian systems into two groups, based on whether their gNID to the closest WCS system exceeded a threshold. We set this threshold to the smallest gNID between systems in the left (WCS) and right (Gaussian dissimilar) columns of Figure 4.1, which is 0.29. We then grouped all Gaussian systems with gNID to the closest WCS system below this threshold into one group, Gaussian[S] (for similar to WCS), and the other Gaussian systems into another group, Gaussian[D] (for dissimilar to WCS). We found that 38% of the Gaussian systems fell in Gaussian[D] and they spanned the complexity range $[0.86, 2.26]$. Thus, a substantial proportion of the randomly-generated Gaussian systems are at least as dissimilar to WCS systems as are those in the right column of Figure 4.1.

Figure 4.2 shows the results of an IB efficiency analysis of the WCS systems (replicating Zaslavsky, Kemp, et al. (2018), and assuming their least-informative prior), and also of our Gaussian systems. It can be seen that all Gaussian systems are highly efficient in the IB sense – i.e. they are close to the IB curve that defines the theoretical limit of efficiency in this domain. Mann-Whitney U tests revealed (1) that the Gaussian systems tend to exhibit greater efficiency (lower inefficiency ϵ) than do the WCS systems in the same complexity range ($P \ll .001$), and (2) that the Gaussian[D] systems, which are dissimilar to WCS systems, are also more efficient than WCS systems ($P \ll .001$, one-sided), and slightly to marginally more efficient than Gaussian[S] systems ($P = .019$ one-sided; Bonferroni corrections do not change the qualitative outcome).[1] These findings suggest that there is a substantial number of color naming systems that are dissimilar to those of human languages, yet more efficient than them. This in turn may help to make sense of Chaabouni et al.'s 2021

[1]Throughout the paper we use one-sided tests when comparing different sets of color naming systems to the Gaussian systems. The reason for this is that we are interested in knowing whether various systems generated by an evolutionary process exceed a random baseline when it comes to either efficiency or similarity to human systems.

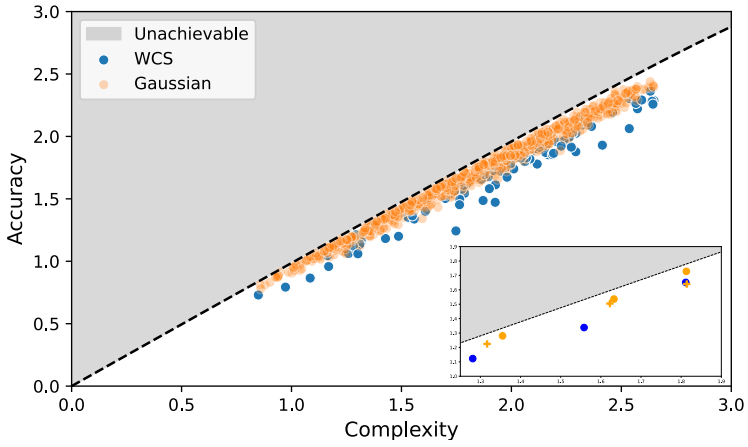


Figure 4.2: Efficiency of color naming, following Zaslavsky et al., 2018. The dashed line is the IB theoretical limit of efficiency for color naming, indicating the greatest possible accuracy for each level of complexity. The color naming systems of the WCS are shown in blue, replicating the findings of Zaslavsky et al., 2018. Our randomly-generated Gaussian systems are shown in orange. The Gaussian systems are often closer to the IB curve than the WCS systems are. The inset shows the 9 color systems of Figure 4.1, with the dissimilar Gaussian systems shown as +.

finding that their evolutionary process yielded systems that were highly efficient but not particularly similar to human ones: our analysis illustrates that there are many such systems. Given this, we wished to determine whether a natural evolutionary process would yield both efficiency in the IB sense, and similarity to human systems.

3 Iterated learning and communication

As noted above, iterated learning (e.g. Kirby (2001) and Smith et al. (2003)) is a cultural evolutionary process in which a cultural convention is learned first by one generation of agents, who then pass that convention on to another generation, and so on — and the convention changes during inter-generational transmission. Some of the work we have reviewed above addresses iterated learning (e.g. Levinson (2012) and Carstensen et al. (2015)). However other work we have reviewed instead addresses cultural evolution through communication within a single generation (e.g. Kågebäck et al. (2020), Chaabouni et al. (2021), and Tucker et al. (2022)). We wished to explore the roles of both iterated learning and communication, and so we adopted the general approach of Kirby et al. (2015), which involves both in a way that allows the role of each to be highlighted. Specifically, we adopted the recently proposed *neural iterated learning* (NIL) algorithm (Ren et al. 2020), which can be seen as a neural network implementation of the approach of Kirby et al. (2015). In the NIL algorithm, illustrated in overview in Figure 4.3, artificial agents are implemented as neural networks that communicate with each other within a generation, and transmit information across generations. Cultural convention (in our

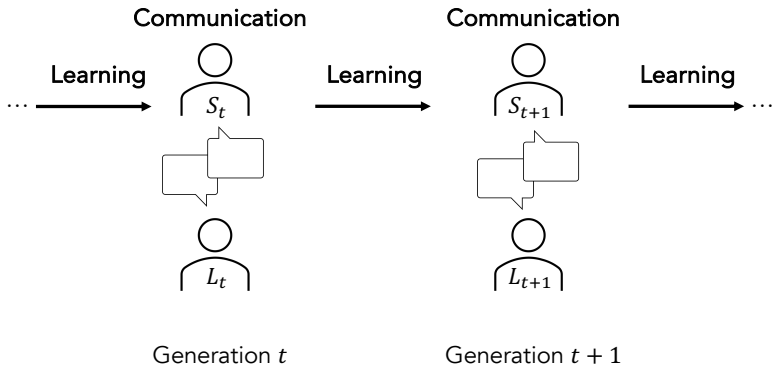


Figure 4.3: Illustration of the neural iterated learning (NIL) algorithm (Ren et al., 2020). The algorithm alternates between communication within a generation, and learning that is iterated across generations. The speaker (S) in each generation learns from the speaker in the previous generation, and communicates with the listener (L) in their own generation.

case, a color naming system) evolves both from within-generation communication and from inter-generational transmission, as the convention is iteratively passed down through generations of artificial agents, with each new generation learning from the previous one.[2]

In the NIL algorithm, each generation t (for time step) consists of two artificial agents, a speaker S_t and a listener L_t. The NIL algorithm operates in three phases. (1) In the first phase, the *learning phase*, both agents are exposed to the naming convention of the previous generation. This is done by first training the speaker S_t, using cross-entropy loss, on color-name pairs generated by the speaker of the previous generation. The listener L_t is then trained via reinforcement learning in a few rounds of a signaling game while keeping S_t fixed: that is, the speaker learns from the previous generation, and the listener then learns from the speaker. We had the agents play the signaling game used by Kågebäck et al. (2020), in which the speaker is given a color chip c, sampled from a prior distribution over color chips, and produces a category name describing that color. The listener then attempts to identify the speaker's intended color based on the name produced, by selecting a color chip $\hat{c}$ from among those of the naming grid shown in Figure 4.1. A reward is given to the listener depending on how perceptually similar the selected chip is to the original color, following Equation 3.1 below. (2) In the second phase, the *interaction phase*, the agents play the same signaling game but this time both agents receive a joint reward and update their parameters during communicative interactions. (3) In the third phase, the *transmission phase*, color-name pairs are generated by sampling colors from the prior distribution and obtaining names for them from the speaker S_t. These color-name pairs are then passed on to the next generation of agents. In all three phases, color chips are sampled according to the least-informative prior of

[2]NIL, or neural iterated learning, is therefore not an entirely informative name for this process, as it does not explicitly label the important element of within-generation communication.

Zaslavsky, Kemp, et al. (2018). Algorithm 4.1 presents a schematic overview of the NIL algorithm, and Ren et al. (2020) present a detailed description. Both the NIL algorithm and the setting explored in Kågebäck et al. (2020) build on important earlier work exploring the emergence of communication in neural network models (e.g. Foerster et al. (2016), Havrylov and Titov (2017), Lazaridou et al. (2017), and Mordatch and Abbeel (2018)).

In our experiments, we represent both the speaker and listener as neural networks with one hidden layer consisting of 25 units with a sigmoidal activation function followed by a softmax output layer. Individual colors are represented in 3-dimensional CIELAB space when supplied as input to the speaker, and category names as one-hot encoded vectors. The speaker's network parameterizes a conditional distribution over categories given a color. To produce an utterance during communication, the speaker samples a category from this distribution and conveys it to the listener. The input to the listener is the category uttered by the speaker, represented as a one-hot encoded vector. The output of the listener's network is a probability distribution over the stimulus set, and the listener produces a guess by sampling from this distribution. For the reinforcement learning parts of NIL we use the classical algorithm REINFORCE (Williams 1992). For the transmission phase we sample 300 color-name pairs with replacement, out of the 330 chips in the entire stimulus set; this ensures that the new generation will have seen examples from most of color space but it is impossible for them to have seen all color-name pairs. To optimize the neural networks, we use the optimizer Adam (Kingma and Ba 2014), both in the learning and interaction phase, with learning rate 0.005 and batch size 50. For each phase in the NIL algorithm we take 1000 gradient steps. We stop the NIL algorithm either after 250 generations or once the maximum difference in IB complexity and accuracy over the ten latest generations is smaller than 0.1 bit, i.e. when the last ten generations are all within a small region of the IB plane. Note that NIL is not guaranteed to converge in the IB plane and might oscillate back and forth. This is because the transmission dataset is finite and randomly sampled, so the next generation might only be able to approximately reconstruct the naming system of the previous generation.

The reward function: The reward function of Kågebäck et al. (2020), which we use here, takes the form:

$$r(c, \hat{c}) = e^{-\gamma ||x_c - x_{\hat{c}}||_2^2} \tag{3.1}$$

where c is the chip sampled by the speaker, $\hat{c}$ is the chip chosen by the listener as their interpretation of the chip intended by the speaker, x_c is the location in CIELAB space of chip c, and γ is a parameter that controls how precise the listener's choice $\hat{c}$ has to be. As $\gamma \to \infty$ the above reduces to a binary reward function, i.e. the listener has to perfectly reconstruct the color to get any reward. On the other hand, if $\gamma = 0$ the reward function is vacuous in the sense that any possible reconstruction yields a reward of 1. We use $\gamma = 0.001$ which was originally used by Kågebäck et al. (2020) and motivated by the analysis in Regier et al. (2007).

Algorithm 4.1 Neural Iterated Learning

Initialize dataset D_1 uniformly at random
for $t = 1...$ **do**
  Learning Phase
  Randomly initialize S_t and L_t.
  Train S_t on D_t using stochastic gradient descent and cross-entropy loss.
  Play signaling game between S_t and L_t and update parameters of only L_t using the rewards.
  Interaction Phase
  Play signaling game between S_t and L_t and update parameters of **both** agents using the rewards.
  Transmission Phase
  Create transmission dataset D_{t+1} consisting of color-name pairs, (c, w) by sampling colors from the prior $p(c)$ and providing them as input to S_t.
end for

4 Analyses and results

4.1 Iterated learning and communication operating together

For each vocabulary size $K = 3 \ldots 10$ and $K = 100$ we ran 100 independent instances of the NIL algorithm. For each instance, we considered the color naming system of the last speaker to be the result of that instance — we call these systems IL+C, as they are the result of iterated learning plus communication, and we evaluated the IL+C systems in the IB framework. As can be seen in Figure 4.4 (top panel), the IL+C systems are highly efficient in the IB sense: they lie near the theoretical efficiency limit (median inefficiency $\epsilon = 0.07$), and they are no less efficient than the random Gaussian systems we considered above (median inefficiency $\epsilon = 0.09$), which in turn are more efficient than the human systems of the WCS (see above). Thus, iterated learning plus communication as formalized in the NIL algorithm leads to semantic systems that are efficient in the IB sense. This is consistent with existing proposals: the reward during the signaling game favors informativeness (higher reward for similar colors, following Kågebäck et al. (2020)), and it has been argued that iterated learning favors simplicity (e.g. Kirby et al. (2015) and Carr et al. (2020)). Interestingly, all the resulting systems lie within the complexity range of the WCS systems even though NIL could theoretically produce much more complex systems, especially when initialized with $K = 100$.

Xu, Dowman, et al. (2013) examined how color naming systems evolved through chains of iterated human learners without within-generation communication, but with the number of categories constrained. They found that these lab-evolved systems tended to gravitate toward color naming systems that were similar to those of the WCS, and we wished to know whether the same was true of computational agents in the NIL framework. For each IL+C system, we determined the dissimilarity (gNID) between that system and the most similar (lowest gNID) WCS system. We also determined the analogous quantity (dissimilarity to the most similar WCS system)

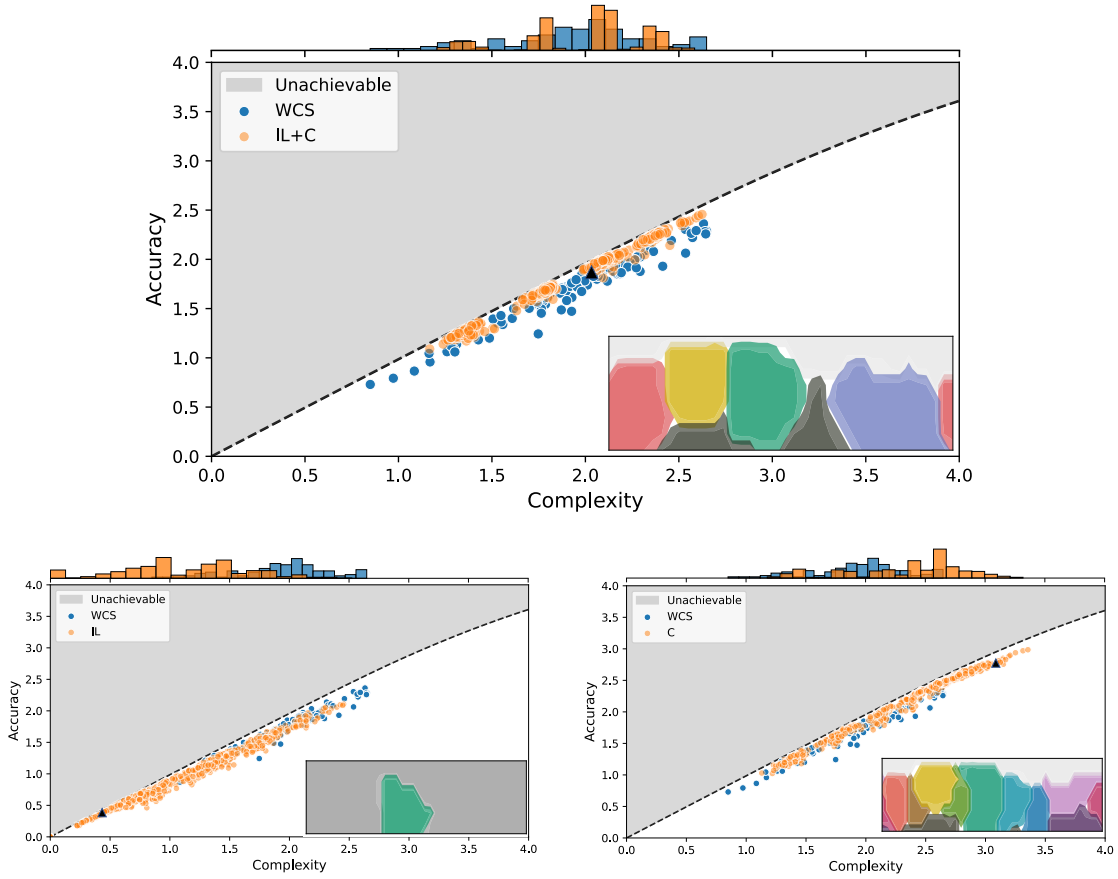


Figure 4.4: Efficiency of the (top) IL+C, (bottom left) IL, and (bottom right) C evolved color naming systems (orange dots), in each case compared with the natural systems of the WCS (blue dots). The black triangle indicates the end state of one run, shown in the inset color map. The histograms above each figure indicate the proportion of systems at the corresponding complexity level.

for each random Gaussian system. Figure 4.5 shows that IL+C systems tend to be similar to WCS systems to a greater extent than Gaussian systems do, and this was confirmed by a one-sided Mann-Whitney U test ($P \ll .001$). Thus, the NIL process tends to gravitate toward human (WCS) systems to a greater extent than a random but efficient baseline, the Gaussian systems.[3]

We also asked whether NIL would transform efficient systems that were dissimilar to those of the WCS (namely those of Gaussian[D]) into comparably efficient systems that were more similar to the WCS. To test this, we initialized the NIL algorithm with a Gaussian[D] system, ran the NIL algorithm, and compared the initial system to the one that resulted from NIL. Figure 4.6 illustrates the beginning and end points of this process for a small set of systems, and shows that NIL transforms systems that are efficient but unlike the WCS into systems that are similar to particular WCS

[3]We found that 14% of the IL+C experiments ran for the maximum number of generations without converging in the IB plane. Excluding these systems from the analysis and only considering the IL+C runs that did converge does not change the qualitative outcome of the analysis above.

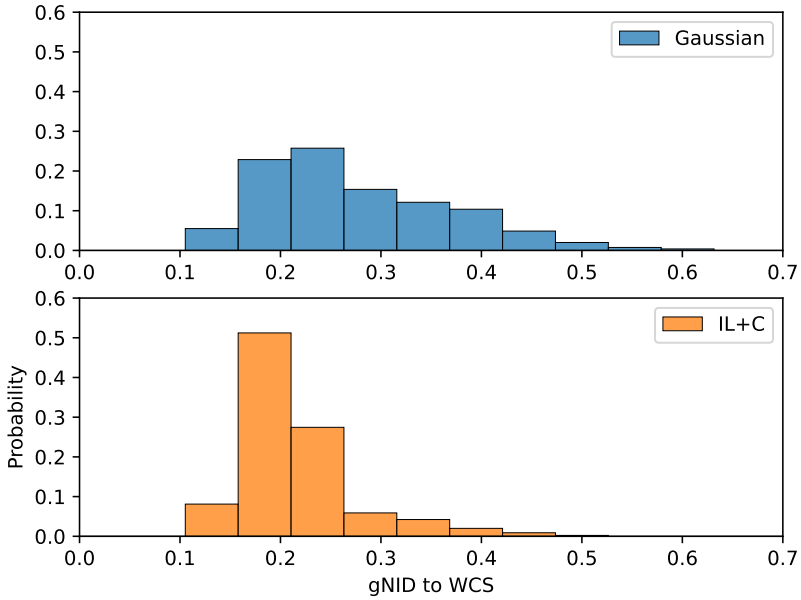


Figure 4.5: Distribution of dissimilarity to WCS systems (minimum gNID to any WCS system), shown for IL+C and Gaussian systems. The Gaussian systems include both Gaussian[S] and Gaussian[D]. Evolved IL+C systems tend to be more similar to attested WCS systems than are random but highly efficient Gaussian systems.

systems. Figure 4.7 shows that the same general pattern also holds over Gaussian[D] systems taken as a whole. For each Gaussian[D] system, we created an NIL chain, and initialized the chain with that Gaussian[D] system. For each such NIL chain, we measured the dissimilarity (gNID) of its initial Gaussian[D] system to the most similar WCS system, and the gNID of the end result of NIL to its most similar WCS system. We found that NIL tends to transform Gaussian[D] systems into systems that are more similar to the human systems of the WCS. The mean gNID to WCS was 0.38 before NIL and 0.25 after, and the reduction in dissimilarity to WCS after applying NIL was significant (one-sided (paired) Wilcoxon signed-rank test, $n = 302$, $T = 1113$, $P \ll .001$). The median inefficiency of Gaussian[D] systems is $\epsilon = 0.09$ and the median inefficiency of the results of NIL is slightly lower at $\epsilon = 0.07$, meaning that NIL made the already-efficient Gaussian[D] systems slightly more efficient (one-sided (paired) Wilcoxon signed-rank test, $n = 302$, $T = 7716$, $P \ll .001$). Thus, NIL moves already-efficient systems closer to the attested systems of the WCS, while maintaining and even slightly improving efficiency. Finally, it is noteworthy that NIL with 3 terms converges to a system that is similar to a 3-term WCS system (see the top row of Figure 4.6), because 3-term systems are the one case in which IB optimal systems qualitatively diverge from human data (Zaslavsky, Kemp, et al. (2018), p. 7941; see also Figure 4.8 below and accompanying text). Thus, this is a case in which NIL appears to provide a better qualitative fit to the data than IB does.

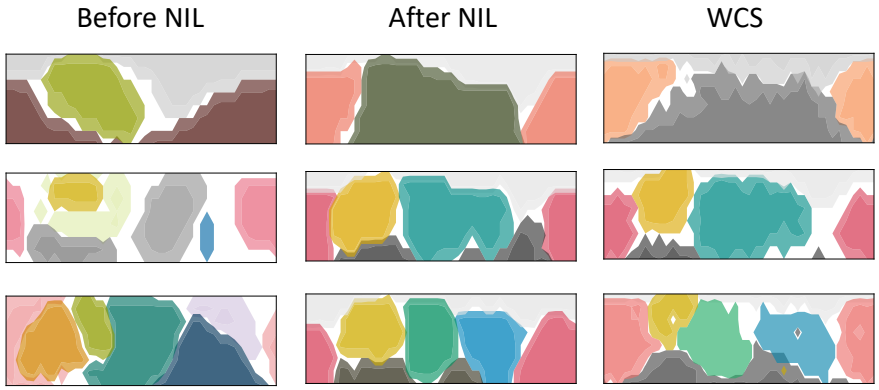


Figure 4.6: NIL transforms efficient color naming systems to become more similar to the WCS. In each row, the left column shows a Gaussian[D] system that was used to initialize NIL, the middle column shows the result of running NIL from that initialization state, and the right column shows a WCS system (from top to bottom: Bété, Colorado, Dyimini) that is similar to the NIL result.

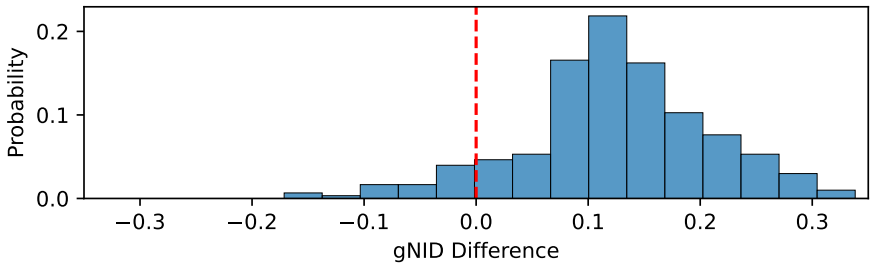


Figure 4.7: NIL tends to transform efficient Gaussian[D] color naming systems to become more similar to the WCS. The difference score is dissimilarity to WCS (minimum gNID to any WCS system) before NIL, minus the same quantity after NIL. Values above zero (marked by the dashed vertical red line) indicate that NIL has brought a system closer to the systems of the WCS. There is a clear trend towards positive values, indicating that NIL tends to transform already-efficient systems into systems that are more human-like.

4.2 Iterated learning alone, and communication alone

So far, we have seen evidence that the Kirby et al. (2015) model of cultural evolution, as implemented in the NIL algorithm, may provide a plausible model of the cultural evolutionary process by which human color naming systems become efficient. We have referred to the result of the full NIL algorithm as IL+C systems, because these systems result from both iterated learning (IL) and communication (C). This raises the question whether iterating learning alone, or communication alone, would yield comparable results.

To find out, following Kirby et al. (2015), we ran two variants of this cultural evolutionary algorithm. One variant included only iterated learning but no communication (i.e. lines 6-8 of Algorithm 4.1 were omitted). The other variant included

communication but no iterated learning (i.e. there was only one pass through the main loop, which stopped at line 9); this is exactly the experiment that was performed by Kågebäck et al. (2020). We refer to the results of the iterated-learning-only algorithm as IL (for iterated learning), and the results of the communication-only algorithm as C (for communication). For the C experiments, we trained each dyad of agents for at most $250,000$ batches but stopped the training once the agents satisfied the stopping criterion used for IL+C. Note that Kågebäck et al. (2020) only trained each dyad for $50,000$ steps without any early stopping criterion. We found that 99.6% of the C experiments converged before reaching the maximum number of batches. All the IL experiments converged in the IB plane before reaching 250 generations.

Comparison of the three panels of Figure 4.4 reveals that there are qualitative differences in the profiles of the systems produced by the 3 variants of the NIL algorithm (IL+C, IL, and C). We have already seen that IL+C systems (top panel) are both efficient and similar to human systems; we also note that they lie within roughly the same complexity range as the human systems of the WCS. In contrast, the IL systems (bottom left panel) skew toward lower complexity than is seen in human systems, and in fact about 6% of the IL systems lie at the degenerate point $(0, 0)$ in the IB plane, at which there is a single category covering the entire color domain. This skew toward simplicity is compatible with the view (e.g. Kirby et al. (2015) and Carr et al. (2020)) that iterated learning provides a bias toward simplicity. At the same time, the IL systems are not only simple but also quite efficient (i.e. informative for their level of complexity), which is in turn compatible with Carstensen et al.'s 2015 claim that iterated learning can produce informativeness, and with Carr et al.'s 2020 proposal that a process that primarily drives toward simplicity can sometimes also result in greater informativeness. Finally, the C systems (bottom right panel) show the opposite pattern: a bias toward higher informativeness, at the price of higher complexity, extending well above the complexity range observed in the human systems of the WCS.

Taken together, these results suggest that iterated learning alone over-emphasizes simplicity, communication alone over-emphasizes informativeness, and iterated learning with communication provides a balance between the two that aligns reasonably well with what is observed in human color naming systems. Overall, these results suggest that iterated learning plus communication is a more plausible model of the cultural evolutionary process that leads to efficient human color naming systems than is either iterated learning alone, or communication alone. These findings echo those of Kirby et al. (2015), who found that compositional structure evolved in a communicative system only under the combination of iterated learning and within-generation communication, and not under either process taken alone.

4.3 The distribution of systems produced by IL+C

To further explore the distribution of systems produced by IL+C we grouped all IL+C systems from the main experiment based on the number of color terms, K, in the systems. For each number of color terms, we clustered the systems using spectral clustering (Luxburg 2007) with gNID as the dissimilarity measure. To find the

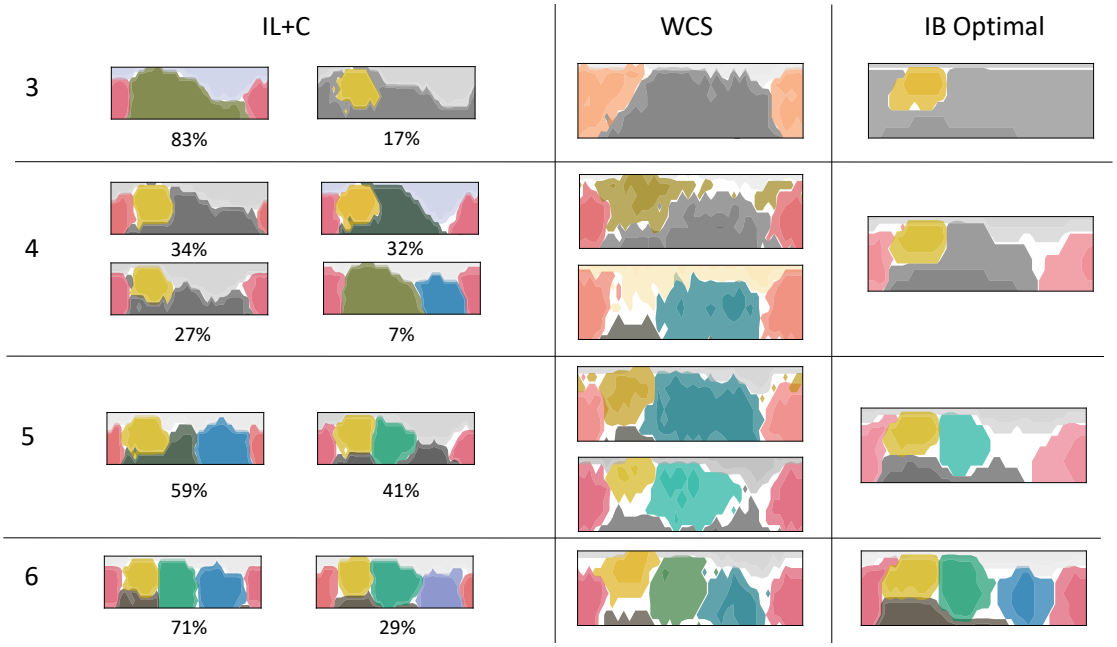


Figure 4.8: Representative IL+C systems (left column), WCS systems (middle column) and IB optimal systems (right column), with 3, 4, 5, and 6 color terms (rows). The % under each IL+C system indicates the percentage of IL+C systems in the corresponding cluster. The WCS systems are, from top to bottom: Nafaanra (iso: nfr, Ghana), Culina (iso: cul, Peru, Brazil), Waorani (iso: auc, Ecuador), Jicaque (iso: jic, Honduras), Berik (iso: bkl, Indonesia), and Kalam (iso: kmh, Papua New Guinea).

appropriate number of clusters for each number of color terms, we performed spectral clustering with $C = 2, 3, 4$ clusters and reported the clustering with the highest silhouette score (Rousseeuw 1987) which is standard in clustering. Since spectral clustering does not return cluster centers, we take the system that minimizes the average pairwise gNID to all other systems in the cluster as a representative sample of that cluster. The resulting systems, for $K = 3...6$, are presented in Figure 4.8 along with some WCS systems and the optimal IB systems. The number under each representative IL+C system indicates the percentage of systems contained in the corresponding cluster.

Interestingly, we see that the IL+C systems with three color terms appear in two clusters: a larger cluster that corresponds reasonably well to 3-term systems observed in the WCS, and a smaller cluster that is similar to the unattested IB optimal system. This suggests that there are two different optima that IL+C converges to: one human-like and the other corresponding to the IB optimal solution. The fact that the cluster corresponding to the IB solution is much smaller suggests that IL+C has a bias toward systems that are more similar to the WCS systems. These results are compatible with the idea that the attested 3-term systems represent a local optimum that is easier to reach through a process of cultural evolution than is the IB optimal solution. Related ideas have also been proposed in connection with kin terminologies, e.g. Epling et al. (1973), Kemp and Regier (2012).

For the four term systems we observe that 93% of the IL+C systems end up in

Figure 4.9: Hue-based artificial systems, with 3 (left) and 10 (right) categories.

clusters that correspond fairly well with the optimal IB system and one of the WCS systems shown in Figure 4.8. The last 7% of the systems end up in a cluster that does not map clearly onto the WCS data. For both $K = 5$ and $K = 6$ we observe that at least one of the IL+C clusters seems to correspond fairly well with systems in the WCS and with IB optimal systems.

4.4 Learnability and convexity

As mentioned above in our review of relevant literature, an influential idea holds that human categories form convex regions in a given conceptual space (Gärdenfors 2000). In the case of color, a natural space for testing this claim is CIELAB space (Figure 4.1, top right panel), and Jäger (2010) has shown that the natural color categories found in the WCS are convex sets in CIELAB space — supporting the convexity claim of Gärdenfors (2000) in the domain of color. More recently, Steinert-Threlkeld and Szymanik (2020) have extended this line of thought by arguing that convex color categories are easier to learn than are non-convex ones, and that this greater learnability helps to explain why human color categories tend to be convex.

We sought to situate this argument relative to the one we have been advancing here. Intuitively, it seems plausible that the artificial Gaussian systems we have considered above should also be convex, because they are based on spherical Gaussian-shaped kernels — but as we have seen, many of these Gaussian systems are quite dissimilar to the human systems of the WCS. This suggests that convexity may be a necessary but not sufficient criterion for characterizing human-like semantic categories, a suggestion with which proponents of the convexity argument are comfortable (P. Gärdenfors, G. Jäger, personal communication; see also Gärdenfors (2024)). To probe this possibility further, we assessed the convexity, the (non-iterated) learnability, and the efficiency of the WCS systems, the randomly-generated Gaussian systems, and an additional set of baseline systems that draw category distinctions based only on hue. These hue-based systems were designed to be convex but not similar to human systems. Specifically, for vocabulary sizes $K = 3...10$ we divided the Munsell chart into equally sized categories by grouping together color chips based on their hue only; in case equally sized categories were not possible we created $K - 1$ equally sized categories and added the remaining color chips to the last category. Example hue-based systems are shown in Figure 4.9: these are deterministic systems in which hue column fully determines the category to which a given chip belongs.

To assess the convexity of a color naming system, we adopted the measure of Steinert-Threlkeld and Szymanik (2020). They took the degree of convexity of a single category, named by a word w, to be:

$$\text{dcc}(w) := \frac{|w|}{|\text{ConvHull}(w)|}$$

where $|\cdot|$ is the size of a set, i.e. the number of color chips in that set, and ConvHull(w) is the convex hull, in CIELAB space, of those chips in category w. Thus, dcc(w) gives us the proportion of those chips in the convex hull of category w that are also in the category w itself. For a perfectly convex category, this proportion will be 1. Steinert-Threlkeld and Szymanik (2020) then defined the degree of convexity of an entire system S of categories to be the average, weighted by category size, of dcc(w) across categories w in S:

$$\text{dc}(S) := \frac{\sum_{w \in S} |w| \cdot \text{dcc}(w)}{\sum_{w \in S} |w|}$$

A dc(S) value of 1 corresponds to a system of perfectly convex color categories.[4]

To assess the (non-iterated) learnability of a color naming system, we took a system to be easily learned to the extent that a neural network learner *generalizes* the system well — that is, to the extent that the learned system matches the one from which training data was sampled. We assessed this by considering only the learning phase of the NIL algorithm, and considering only the speaker's learning (specifically lines 3-5 of Algorithm 4.1), leaving all parameters unchanged. We then measured the gNID between the learned system and the system from which training data was drawn. During training, the agent sees only part of the entire system, so this gNID is a measure of how well the agent generalizes from the data it receives. To mitigate possible effects caused by sampling the training dataset, we performed each experiment over 10 independent runs and averaged.

We assessed the convexity, the learnability, and the IB efficiency of the (natural) WCS, (artificial) Gaussian, and (artificial) hue-based systems. Convexity results are shown in Figure 4.10 (left panel), and learnability results are shown in Figure 4.10 (right panel). All three types of system are highly convex, with the artificial Gaussian and hue-based systems being slightly more convex than the natural WCS systems – perhaps because the natural systems include noise. Moreover, in line with the expectation that convex systems will be learnable, all three types of system show good generalization, with no advantage for the natural WCS systems over the artificial Gaussian and hue-based systems. These results confirm that convex systems tend to be highly learnable, and also highlight that something beyond convexity and (non-iterated) learnability must play a role in differentiating human systems from artificial semantic systems that do not resemble them. Finally, Figure 4.11 shows that artificial hue-based systems are not especially efficient — in contrast with artificial Gaussian systems and natural WCS systems. We take these results to

[4]This method assumes deterministic rather than probabilistic category membership. When applying this method to probabilistic systems, we first converted the probabilistic system to a deterministic one by assigning each chip to the modal category for that chip; we then applied this convexity measure to the resulting deterministic system.

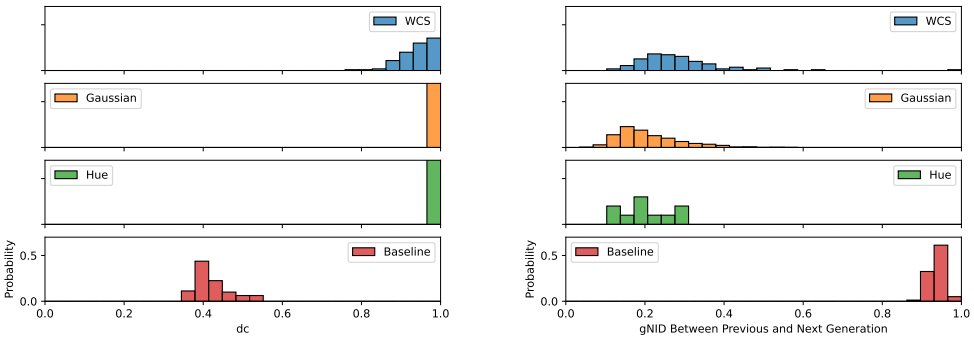


Figure 4.10: **Left panel: Convexity.** Convexity for different types of category systems. The natural systems of the WCS, artificial Gaussian systems, and artificial hue-based systems, are all highly convex when compared with a baseline of randomly generated systems in which each color chip is assigned to a category selected uniformly at random (labeled "Baseline"). We generated such baseline random systems with $k = 3...10$ color categories and for each k we drew 10 random systems. **Right panel: Learnability.** Ease of learning is assessed by how well a learner generalizes, and generalization is measured by gNID between a learned system and the system from which training data was drawn. Artificial Gaussian and hue-based systems show generalization that is no worse than that of natural WCS systems.

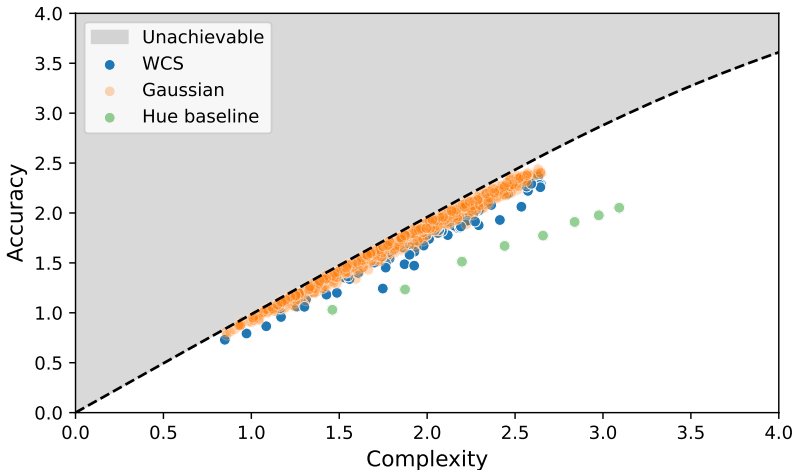


Figure 4.11: Some convex and learnable category systems are not efficient. Efficiency of the artificial hue-based systems (green dots), compared with that of the artificial Gaussian (orange dots) and natural WCS (blue dots) systems.

suggest that convexity and learnability provide a partial answer to the question of what characterizes human semantic categories — and that a fuller answer may be provided by iterated learning and communication operating together, as a model of cultural evolution that leads toward efficient and human-like systems of semantic categories.

5 Discussion

We have shown (1) that there exists a reasonably sized class of color naming systems that are highly efficient in the IB sense but dissimilar from human systems; (2) that iterated learning plus communication, as captured in the NIL algorithm, leads to color naming systems that are both efficient in the IB sense and similar to human systems, and (3) that iterated learning alone, communication alone, and convexity alone, do not yield that result as clearly. These findings help to answer some questions, and also open up others.

As we have noted, the existence of highly efficient systems that do not align with human ones is not in itself surprising. IB is a non-convex optimization problem (Tishby et al. 1999; Zaslavsky, Kemp, et al. 2018), so multiple optima and near-optima are to be expected. However we feel that our identification of such systems may nonetheless be helpful, because it highlights just how many such systems exist, and just how dissimilar from human systems they sometimes are. In particular, this helps to make sense of Chaabouni et al.'s 2021 finding that simulations of cultural evolution can lead to color naming systems that exhibit high IB efficiency but deviate to some extent from human systems — something that we also sometimes find, as seen above in Figure 4.8. This in turn highlights the importance of identifying cultural evolutionary processes that tend to avoid these outcomes and instead converge toward systems we find in human languages.

We have argued that iterated learning plus communication, as proposed by Kirby et al. (2015) and implemented in the NIL algorithm (Ren et al. 2020), is such a process, and that it provides a better account of cross-language color naming data than either iterated learning alone, or communication alone. Our findings supporting this idea thus generalize Kirby et al.'s 2015 argument, which concerned compositionality in language, to a different aspect of language. Our findings also confirm a proposed resolution to a tension in the literature. As we have noted, Carstensen et al. (2015) argued that iterated learning alone can lead to informative semantic systems, whereas Carr et al. (2020) argued that iterated learning provides a bias for simplicity, and communication provides a bias for informativeness. However Carr et al. (2020) also found that a bias for simplicity — such as that provided through iterated learning — can produce systems that are informative as well as simple, and they suggested that this helps to resolve the tension. Specifically, they suggested that an increase in informativeness through iterated learning, as observed by Carstensen et al. (2015), can result from a process (iterated learning) the primary outcome of which is a drive toward simplicity. Our finding that both forces are needed to account for the data aligns with Carr et al.'s 2020 central position. In addition, our finding that iterated learning alone also converges to efficient and thus informative systems — although often to overly simple ones — qualitatively replicates the findings of Carstensen et al. (2015), in a way that confirms Carr et al.'s 2020 proposed resolution of the tension: iterated learning does lead to simplicity, as suggested, but it also leads to informativeness to some extent.

It is natural to think of cultural evolution as a means by which the abstract computational goal of optimal efficiency might be attained or approximated. The

optimally efficient color naming systems on the IB curve closely resemble those in human languages (Zaslavsky, Kemp, et al. 2018), and the IL+C systems are likewise highly efficient and similar to those in human languages. However, as noted above, there is an exception to this pattern. In the case of 3-term systems, the IB optimal system qualitatively differs from the color naming patterns found in the WCS (Zaslavsky, Kemp, et al. (2018), p. 7941), whereas IL+C systems often qualitatively match attested 3-term systems (recall the top rows of Figures 4.6 and 4.8). Thus, in this one case, it appears that human languages do not attain the optimal solution or something similar to it, and instead attain a somewhat different near-optimal solution that is apparently more easily reached by a process of cultural evolution.

A major question left open by our findings is exactly why we obtain the results we do. The general model of Kirby et al. (2015), as implemented in the NIL algorithm, is just one possible cultural evolutionary process, and we have seen that that process accounts for existing data reasonably well. It makes sense intuitively that NIL strikes a balance between the simplicity bias of iterated learning and the informativeness bias of communication — but what is still missing is a finer-grained sense for exactly which features of this detailed process are critical, vs. replaceable by others, and what the broader class of such processes is that would account well for the data (e.g. Tucker et al. (2022)). A related direction for future research concerns the fact that the evolutionary process we have explored here is somewhat abstract and idealized, in that agents communicate with little context or pragmatic inference. Actual linguistic communication is highly context-dependent, and supported by rich pragmatic inference — it seems important to understand whether our results would still hold in a more realistic and richer environment for learning and interaction. Our agents also have designated roles: an agent acts either as a speaker or as a listener, and a direction for future research is to extend our setting to a more realistic model in which agents can alternate between the two roles. In addition, in our idealized setup a given agent interacts with only one other agent, whereas in human social interaction, communication within a generation happens in social networks such that an agent interacts with many other agents throughout their lifetime. An interesting direction for future research would be to explore what biases are introduced by certain population structures and whether varying the population structure can account for the variance observed in human color naming data.

Another important issue concerns the situating of this evolutionary account relative to the classic account of Berlin and Kay (1969). Our work here inherits, from the work of Zaslavsky, Kemp, et al. (2018) on which we build, an important connection to that earlier classic account: a trace along the IB curve reveals a sequence of color naming systems that gradually increase in complexity and that recapitulate the Berlin and Kay (1969) hierarchy, while also capturing aspects of competing accounts (MacLaury 1997; Levinson 2000). However, the mapping of that connection to fine-grained empirical data concerning language change over historical time has only recently begun (Zaslavsky, Garvin, et al. 2022), and a connection to the evolutionary model we explore here has not to our knowledge been attempted. Finally, we have focused on the semantic domain of color, but the ideas we have pursued are not specific to color, so another open question is the extent to which our

results generalize to other semantic domains.

Acknowledgments

We thank the 2 anonymous reviewers and the editor for their valuable comments on this paper. An earlier version of this paper appeared in the Proceedings of the 45th Annual Meeting of the Cognitive Science Society (2023). We thank Noga Zaslavsky and 3 anonymous reviewers for helpful comments on that earlier version of the paper. We also thank Gerhard Jäger and Peter Gärdenfors for helpful discussion of category convexity. Author contributions: EC, DD, and TR designed the research; EC performed the research; EC analyzed the data; and EC, DD, and TR wrote the paper. EC was funded by Chalmers AI Research (CHAIR) and the Sweden-America Foundation (SweAm). Computational resources were provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725. We thank the library of Chalmers University of Technology for covering publication costs.

Data Availability

The WCS data can be found in the WCS Data Archive `https://www1.icsi.berkeley.edu/wcs/data.html`. We used the original code of Zaslavsky, Kemp, et al. (2018), available at `https://github.com/nogazs/ib-color-naming`, to compute the IB objective, the inefficiency of the languages, and the gNID between languages. Our code is available here:
`https://github.com/e-carlsson/iterated-learning-color-naming`

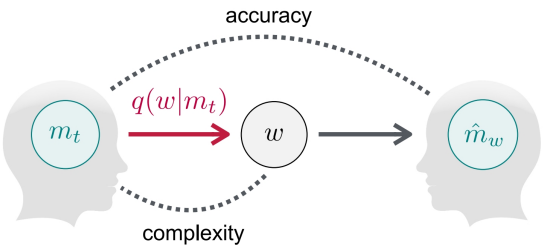


Figure 4.12: The framework of Zaslavsky et al. (2018). A speaker communicates a specific referent to a listener by producing a word. The IB principle provides formal specifications of various quantities associated with this communicative act; see text for details. The figure is from Zaslavsky et al. (2021).

A The framework of Zaslavsky et al. (2018)

Zaslavsky, Kemp, et al. (2018) cast the notion of efficiency in terms of an independent information-theoretic principle, the Information Bottleneck (IB) principle (Tishby et al. 1999). In the framework of Zaslavsky, Kemp, et al. (2018), a semantic system is considered efficient to the extent that it achieves an optimal tradeoff between the complexity of a system, and the accuracy of communication that that system supports. These notions are grounded in the communicative scenario illustrated in Figure 4.12, in which a speaker attempts to communicate with a listener about referents in a given domain universe U, in our case the domain of color. Here, the speaker considers a specific target color $t \in U$ and holds it in mind in the form of a mental representation m_t, which is a probability distribution over color space (CIELAB; recall Figure 4.1), centered at t. To communicate that mental representation, the speaker utters a word w, drawn from a language-specific probabilistic encoder $q(w|m_t)$ that maps from meanings m_t to words w; this encoder $q(w|m_t)$ is the semantic system by which the speaker and listener communicate. The listener then produces, on the basis of the uttered word w, a mental representation $\hat{m}_w$ that is the listener's reconstruction of the speaker's original representation m_t. Casting this simple communicative scenario in terms of the IB principle results in formal definitions of four quantities that are central to the IB formalization of efficiency, and on which we rely in our work: *complexity, accuracy,* ϵ, and *gNID*.

The complexity of a semantic system q is given by $I_q(M_t; W)$, i.e. the mutual information between the speaker's mental representation m_t and the word w used to express it. The greater the complexity of the system, the more information the word w carries about the speaker's mental representation m_t. The accuracy of a semantic system is given by $I_q(W; U)$, which can be shown to capture the similarity of the speaker's and listener's mental representations (see Zaslavsky, Kemp, et al. (2018)). The core idea of efficiency in this framework is to obtain the greatest accuracy possible for a given level of complexity — i.e. to communicate as precisely as possible for a given amount of information sent. An optimally efficient semantic system q is thus one that minimizes the IB objective function:

$$\mathcal{F}_\beta[q] = I_q(M_t; W) - \beta I_q(W; U)$$

where $\beta \geqslant 0$ is a tradeoff parameter that controls the relative weight given to complexity and accuracy. Those systems q^* that minimize this objective function for different values of β yield the IB theoretical limit of efficiency; that is, these are the systems with the greatest possible accuracy for each level of complexity. Zaslavsky, Kemp, et al. (2018) showed that human color naming systems achieve near-optimal efficiency in the IB sense, and that fully IB-optimal systems often closely correspond to color naming systems in human languages.

In our analyses, we also make use of two other quantities from the framework of Zaslavsky, Kemp, et al. (2018). First, ϵ_q measures the inefficiency of a semantic system, or its deviation from optimal efficiency, as described on p. 7939 of their article:

$$\epsilon_q = \frac{1}{\beta}\left(\mathcal{F}_\beta[q] - \mathcal{F}^*_\beta\right)$$

Here $\mathcal{F}^*_\beta$ is the optimal value of the IB objective for a given value of β, and β is chosen to minimize the difference $\mathcal{F}_\beta[q] - \mathcal{F}^*_\beta$ for a given semantic system q. Finally, we follow Zaslavsky, Kemp, et al. (2018) in using their gNID measure to measure the dissimilarity between two semantic systems, as described on p. 7942 of their article. This measure assumes that a single meaning m is assigned a name by each of two semantic systems q_1 and q_2: $W_1 \sim q_1(w_1|m)$ and $W_2 \sim q_2(w_2|m)$. Then the dissimilarity between q_1 and q_2 is given by:

$$\mathrm{gNID}(W_1, W_2) = 1 - \frac{I(W_1; W_2)}{\max\left\{I(W_1; W_1'), I(W_2; W_2')\right\}}.$$

Here, W_i' corresponds to another independent speaker using the system q_i.

References

Abbott, Joshua T., Thomas L. Griffiths, and Terry Regier (2016). "Focal colors across languages are representative members of color categories". In: *Proceedings of the National Academy of Sciences* 113, pp. 11178–11183 (cit. on p. 132).

Baronchelli, Andrea, Tao Gong, Andrea Puglisi, and Vittorio Loreto (2010). "Modeling the emergence of universality in color naming patterns". In: *Proceedings of the National Academy of Sciences* 107.6, pp. 2403–2407 (cit. on p. 130).

Belpaeme, Tony and Joris Bleys (2005). "Explaining Universal Color Categories Through a Constrained Acquisition Process". In: *Adaptive Behavior* 13.4, pp. 293–310 (cit. on p. 130).

Berlin, Brent and Paul Kay (1969). *Basic Color term. Their Universality and Evolution*. 2010. Berlin, Boston: De Gruyter Mouton (cit. on p. 147).

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. 130, 137, 141, 146).

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 pp. 130, 134, 141, 146).

Chaabouni, Rahma, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni (2021). "Communicating artificial neural networks develop efficient color-naming systems". In: *Proceedings of the National Academy of Sciences* 118, e2016569118 (cit. on pp. 130, 131, 133, 134, 146).

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. 130).

Denić, Milica, Shane Steinert-Threlkeld, and Jakub Szymanik (2022). "Indefinite Pronouns Optimize the Simplicity/Informativeness Trade-Off". In: *Cognitive Science* 46.5, e13142 (cit. on p. 129).

Denić, Milica and Jakub Szymanik (2024). "Recursive Numeral Systems Optimize the Trade-off Between Lexicon Size and Average Morphosyntactic Complexity". In: *Cognitive Science* 48.3, e13424 (cit. on p. 129).

Dowman, Mike (2007). "Explaining Color Term Typology With an Evolutionary Model". In: *Cognitive Science* 31.1, pp. 99–132 (cit. on p. 130).

Epling, P.J., Jerome Kirk, and John Paul Boyd (1973). "Genetic Relations of Polynesian Sibling Terminologies". In: *American Anthropologist* 75.5, pp. 1596–1625 (cit. on p. 142).

Foerster, Jakob N., Yannis M. Assael, Nando de Freitas, and Shimon Whiteson (2016). "Learning to Communicate with Deep Multi-Agent Reinforcement Learning". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., pp. 2145–2153. ISBN: 9781510838819 (cit. on p. 136).

Gärdenfors, Peter (2000). "Conceptual spaces: The geometry of thought". In: *MIT Press* 3, p. 16 (cit. on pp. 130, 143).

Gärdenfors, Peter (2024). "Natural Concepts and the Economics of Cognition and Communication". In: *Philosophia*, pp. 1–18 (cit. on p. 143).

Gentner, Dedre and Melissa Bowerman (2009). "Why some spatial semantic categories are harder to learn than others: The typological prevalence hypothesis". In: *Crosslinguistic approaches to the psychology of language: Research in the tradition of Dan Isaac Slobin*. Hove, UK: Psychology Press, pp. 465–480 (cit. on p. 130).

Gyevnar, Balint, Gautier Dagan, Coleman Haley, Shangmin Guo, and Frank Mollica (2022). "Communicative Efficiency or Iconic Learning: Do Acquisition and Communicative Pressures Interact to Shape Colour-Naming Systems?" In: *Entropy* 24.11. DOI: 10.3390/e24111542 (cit. on p. 130).

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 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., pp. 2146–2156 (cit. on p. 136).

Hunter, J. D. (2007). "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3, pp. 90–95. DOI: 10.1109/MCSE.2007.55.

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. 129).

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 pp. 130, 143).

Jameson, Kimberly A. and Natalia Komarova (July 2009). "Evolutionary models of color categorization I Population categorization systems based on normal and dichromat observers". In: *Journal of the Optical Society of America. A, Optics, image science, and vision* 26, pp. 1414–23 (cit. on p. 130).

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. 130, 134–137, 141).

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. 129).

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. 129, 142).

Kingma, Diederik P. and Jimmy Ba (2014). *Adam: A Method for Stochastic Optimization*. arXiv: 1412.6980 [cs.LG] (cit. on p. 136).

Kirby, S. (2001). "Spontaneous evolution of linguistic structure - an iterated learning model of the emergence of regularity and irregularity". In: *IEEE Transactions on Evolutionary Computation* 5, pp. 102–110 (cit. on p. 134).

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. 130, 131, 134, 137, 140, 141, 146, 147).

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 p. 136).

Levinson, Stephen C. (2000). "Yélı Dnye and the Theory of Basic Color Terms". In: *Journal of Linguistic Anthropology* 10.1, pp. 3–55 (cit. on p. 147).

Levinson, Stephen C. (2012). "Kinship and Human Thought". In: *Science* 336.6084, pp. 988–989 (cit. on pp. 130, 134).

Luxburg, Ulrike von (2007). "A tutorial on spectral clustering". In: *Statistics and Computing* 17.4, pp. 395–416 (cit. on p. 141).

MacLaury, Robert E. (1997). *Color and cognition in Mesoamerica: Constructing categories as vantages*. University of Texas Press (cit. on p. 147).

McKinney, Wes et al. (2010). "Data structures for statistical computing in Python." In: *SciPy*. Vol. 445. 1, pp. 51–56.

Mollica, Francis, Geoff Bacon, Noga Zaslavsky, Yang Xu, Terry Regier, and Charles Kemp (2021). "The forms and meanings of grammatical markers support efficient communication". In: *Proceedings of the National Academy of Sciences* 118.49 (cit. on p. 129).

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 p. 136).

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. (2019). "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

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 p. 136).

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. 131, 134, 136, 146).

Rosch, Eleanor (1978). "Principles of categorization". In: *Cognition and categorization*. Ed. by E. Rosch and B. B. Lloyd. New York: Lawrence Erlbaum Associates, pp. 27–48 (cit. on p. 129).

Rousseeuw, Peter J. (1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20, pp. 53–65. ISSN: 0377-0427. DOI: https://doi.org/10.1016/0377-0427(87)90125-7 (cit. on p. 142).

Scott-Phillips, Thomas C. and Simon Kirby (2010). "Language evolution in the laboratory". In: *Trends in Cognitive Sciences* 14.9, pp. 411–417 (cit. on p. 130).

Smith, Kenny, Simon Kirby, and Henry Brighton (Feb. 2003). "Iterated Learning: A Framework for the Emergence of Language". In: *Artificial life* 9, pp. 371–86 (cit. on p. 134).

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. 130).

Steinert-Threlkeld, Shane and Jakub Szymanik (2020). "Ease of learning explains semantic universals". In: *Cognition* 195, p. 104076 (cit. on pp. 130, 143, 144).

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. 129, 131, 146, 149).

Tucker, Mycal, Roger P. Levy, Julie Shah, and Noga Zaslavsky (2022). "Trading off Utility, Informativeness, and Complexity in Emergent Communication". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (cit. on pp. 130, 131, 134, 147).

Virtanen, Pauli et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17, pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

Waskom, Michael L. (2021). "seaborn: statistical data visualization". In: *Journal of Open Source Software* 6.60, p. 3021. DOI: 10.21105/joss.03021.

Williams, Ronald J. (1992). "Simple statistical gradient-following algorithms for connectionist reinforcement learning". In: *Machine Learning* 8.3, pp. 229–256 (cit. on p. 136).

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 pp. 130, 137).

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 p. 129).

Xu, Yang, Terry Regier, and Barbara C. Malt (2016). "Historical semantic chaining and efficient communication: The case of container names". In: *Cognitive Science* 40, pp. 2081–2094 (cit. on p. 129).

Zaslavsky, Noga, Karee Garvin, Charles Kemp, Naftali Tishby, and Terry Regier (2022). "The evolution of color naming reflects pressure for efficiency: Evidence from the recent past". In: *Journal of Language Evolution* (cit. on pp. 132, 147).

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. 129, 131–133, 136, 139, 146–150).

Zaslavsky, Noga, Mora Maldonado, and Jennifer Culbertson (2021). "Let's talk (efficiently) about us: Person systems achieve near-optimal compression". In:

Proceedings of the 43rd Annual Meeting of the Cognitive Science Society (cit. on p. 130).

Zaslavsky, Noga, Terry Regier, Naftali Tishby, and Charles Kemp (2019). "Semantic categories of artifacts and animals reflect efficient coding". In: *41st Annual Conference of the Cognitive Science Society* (cit. on p. 129).

Paper 5

Thompson sampling for bandits with clustered arms

Emil Carlsson , Fredrik D. Johansson, Devdatt Dubhashi.

Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI), 2021.

The paper has been reformatted for uniformity.

Paper 5. Thompson sampling for bandits with clustered arms

Emil Carlsson , Fredrik D. Johansson, Devdatt Dubhashi.

Abstract

We propose algorithms based on a multi-level Thompson sampling scheme, for the stochastic multi-armed bandit and its contextual variant with linear expected rewards, in the setting where arms are clustered. We show, both theoretically and empirically, how exploiting a given cluster structure can significantly improve the regret and computational cost compared to using standard Thompson sampling. In the case of the stochastic multi-armed bandit we give upper bounds on the expected cumulative regret showing how it depends on the quality of the clustering. Finally, we perform an empirical evaluation showing that our algorithms perform well compared to previously proposed algorithms for bandits with clustered arms.

1 Introduction

In a bandit problem, a learner must iteratively choose from a set of N actions, also known as arms, in a sequence of T steps as to minimize the expected cumulative regret over the horizon T (Lai and Robbins 1985). Inherent in this setup is an exploration-exploitation tradeoff where the learner has to balance between exploring actions she is uncertain about in order to gain more information and exploiting current knowledge to pick actions that appears to be optimal.

In this work, we consider versions of the standard multi-armed bandit problem (MAB) and the contextual bandit with linear rewards (CB) where there is a clustering of the arms known to the learner. In the standard versions of these problems the cumulative regret scales with number of arms, N, which becomes problematic when the number of arms grows large (Bubeck and Cesa-Bianchi 2012). Given a clustering structure one would like the exploit it to remove the explicit dependence on N and replace it with a dependence on the given clustering instead. A motivating example is recommender systems in e-commerce where there may be a vast amount of products organized into a much smaller set of categorizes. Users my have strong preferences for certain categorizes which yields similar expected rewards for recommending products from the same category.

Our Contributions. We propose algorithms based on a multi-level Thompson sampling (Thompson 1933) scheme for the stochastic multi-armed bandit with clustered arms (MABC) and its contextual variant with linear expected rewards and clustered arms (CBC). For the MABC, we provide regret bounds for our algorithms which completely removes the explicit dependence on N in favor for a dependence on properties of the given clustering. We perform an extensive empirical evaluation showing both how the quality of the clustering affects the regret and that our algorithms are very competitive with recent algorithms proposed for MABC and CBC. Noteworthy is that the empirical evaluation shows that our algorithms still performs well even in the case where our theoretical assumptions are violated.

2 Stochastic multi-armed bandit with clustered arms

We consider the MABC. As in the standard MAB problem we have a set of arms $\mathcal{A}$ of cardinality N. At each time step $t > 0$ the learner must pick an arm $a_t \in \mathcal{A}$ after which an instant stochastic reward, $r_t(a_t)$, drawn from some distribution, $r_t \sim \mathcal{D}_{a_t}$, with an unknown mean $\mathbb{E}_{\mathcal{D}_{a_t}}[r_t] = \mu_{a_t}$. The goal of the learner is to maximize its expected cumulative reward over a sequence of T time steps or equivalently, to minimize its expected cumulative regret $\mathbb{E}[R_T]$ w.r.t the optimal arm $a^* = \arg\max_{a \in \mathcal{A}} \mu_a$ in hindsight, $R_T := \sum_{t=1}^{T} r_t(a^*) - r_t(a_t)$.

In the MABC, the learner has, in addition, access to a clustering of the N arms which may be used to guide exploration. We will consider two types of clustering:

Disjoint Clusters The N arms are partitioned into a a set of clusters $\mathcal{K}$ such that each arm $a \in \mathcal{A}$ is associated to exactly one cluster.

Hierarchical Clustering The N arms are organized into a tree $\mathcal{T}$ of depth L such that each arm is associated with a unique leaf of the tree.

We will show in Section 2.2 and 2.4 that when rewards are drawn from Bernoulli distributions, $r_t \sim \mathcal{B}(\mu_a)$, with unknown parameters μ_a, the learner can exploit the known clustering to greatly improve the expected cumulative regret compared to the regret achievable with no knowledge of the cluster structure (under certain assumptions on the quality of the clustering).

2.1 Thompson sampling for MABC

In the celebrated Thompson sampling (TS) algorithm for MAB with Bernoulli distributed rewards a learner starts at time $t = 0$ with a prior belief $\text{Beta}(1, 1)$ over possible expected rewards, $\theta_a \in [0, 1]$, for each $a \in A$. At time t, having observed $S_t(a)$ number of successful ($r = 1$) plays and $F_t(a)$ the number of unsuccessful ($r = 0$) plays of arm a, the learner's posterior belief over possible expected rewards for arm a is $\text{Beta}(S_t(a), F_t(a))$, where $S_0(a) = F_0(a) = 1$. At each time step t, the learner samples an expected reward for each arm $\theta_a \sim \text{Beta}(S_t(a), F_t(a))$ and then acts greedily w.r.t.

Algorithm 5.1 TSC

Require: $\mathcal{A}$, $\mathcal{K}$
  Set $S_0 = F_0 = 1$ 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

the sample means, i.e. the learner plays the arm $a_t = \arg\max_{a \in \mathcal{A}} \theta_a$. Given a reward r_t the learner updates the posterior of the played arm a_t as $S_{t+1}(a_t) = S_t + r_t$ and $F_{t+1}(a_t) = F_t(a_t) + (1 - r_t)$,. The posteriors of the arms not played are not updated.

Given a clustering of the arms into a set of clusters $\mathcal{K}$, we introduce a natural two-level bandit policy based on TS, Algorithm 5.1. In addition to the belief for each arm a, $\text{Beta}(S_t(a), F_t(a))$, the learner also keeps a belief over possible expected rewards $\text{Beta}(S_t(C), F_t(C))$ for each cluster $C \in \mathcal{K}$. At each t, the learner first use TS to pick a cluster - that is, it samples $\theta_C \sim \text{Beta}(S_t(C), F_t(C))$ for each cluster $C \in \mathcal{K}$ and then considers the cluster $C_t = \arg\max_{C \in \mathcal{K}} \theta_C$. The learner then samples $\theta_a \sim \text{Beta}(S_t(a), F_t(a))$ for each $a \in C_t$ and plays the arm $a_t = \arg\max_{a \in C_t} \theta_a$. Given a reward r_t the learner updates the beliefs for a_t and C_t as follows $S_{t+1}(a_t) = S_t(a_t) + r_t$, $F_{t+1}(a_t) = F_t(a_t) + (1 - r_t)$, $S_{t+1}(C_t) = S_t(C_t) + r_t$ and $F_{t+1}(C_t) = F_t(C_t) + (1 - r_t)$.

We extended this two-level scheme to hierarchical clustering of depth L, by recursively applying TS at each level of the tree, in Algorithm 5.2. The learner starts at the root of the hierarchical clustering, $\mathcal{T}$, and samples an expected reward for each of the sub-trees, $\mathcal{T}_1^i$ spanned by its children, $i = 1, ...,$ from $\text{Beta}(S_t(\mathcal{T}_1^i), F_t(\mathcal{T}_1^i))$. The learner now traverses down to the root of the sub-tree satisfying $\mathcal{T}_{1,t}^i = \arg\max_{\mathcal{T}_1^i} \theta_{\mathcal{T}_1^i}$. This scheme is recursively applied until the learner reaches a leaf, i.e. an arm a_t, which is played. Given a reward r_t, each belief along the path from the root to a_t is updated using a standard TS update.

Algorithm 5.1 and 5.2 are not restricted to Bernoulli distributed rewards and can be used for any reward distribution with support $[0, 1]$ or for unbounded rewards by using Gaussian prior and likelihood in TS, as done for the standard MAB in Agrawal and Goyal (2017).

2.2 Regret analysis TSC

Assume that we have a clustering of N Bernoulli arms, into a set of clusters $\mathcal{K}$. For each arm a, let μ_a denote the expected reward and let a^* be the unique optimal arm with expected reward μ^*. We denote the cluster containing a^* as C^*. We denote the expected regret for each a as $\Delta_a := \mu^* - \mu_a$ and for each cluster $C \in \mathcal{K}$, we define $\overline{\mu}_C = \max_{a \in C} \mu_a$, $\underline{\mu}_C = \min_{a \in C} \mu_a$ and $\Delta_C = \mu^* - \overline{\mu}_C$.

Algorithm 5.2 HTS

Require: $\mathcal{A}$, $\mathcal{T}$
  Set $S_0(\mathcal{T}_l^i) = F_0(\mathcal{T}_l^i) = 1$ for each sub-tree $\mathcal{T}_l^i$.
  for $t = 1, ..., T$ **do**
    Set $\mathcal{T}_t = \mathcal{T}$.
    while $\mathcal{T}_t$ is not a leaf **do**
      For each sub-tree $\mathcal{T}_l^i$ spanned by the children of $\mathcal{T}_t$ sample $\theta_{\mathcal{T}_l^i} \sim$ Beta$(S_t(\mathcal{T}_l^i), F_t(\mathcal{T}_l^i))$.
      Set $\mathcal{T}_t = \arg\max \theta_{\mathcal{T}_l^i}$.
    end while
    Play the arm a_t corresponding to the leaf $\mathcal{T}_t$ and collect the reward r_t.
    Perform a TS update on each $S_t(\mathcal{T}_l^i), F_t(\mathcal{T}_l^i)$ on the path to a_t.
  end for

For each cluster $C \in \mathcal{K}$ we define distance d_C to the optimal cluster C^* as $d_C = \min_{a \in C^*, \hat{a} \in C} \mu_a - \mu_{\hat{a}}$ and the width w_C as $w_C = \overline{\mu}_C - \underline{\mu}_C$, let w^* denote the width of the optimal cluster.

Assumption 2.1 (Strong Dominance)**.** For $C \neq C^*, d_C > 0$.

This assumption is equivalent to what is referred to as *tight clustering* in Bouneffouf et al. (2019) and *strong dominance* in Jedor et al. (2019). In words, we assume that, in expectation, every arm in the optimal cluster is better than every arm in any suboptimal cluster.

In order to bound the regret of TSC we will repeatedly use the following seminal result for the standard MAB case (without clustering) from Kaufmann et al. (2012). Here, we denote the Kullback-Leibler divergence between two Bernoulli distributions with means μ_1 and μ_2 as $\mathbb{KL}(\mu_1, \mu_2)$ and the natural logarithm of T as $\log T$.

Theorem 2.1 ((Kaufmann et al. 2012))**.** *In the standard multi-arm bandit case with optimal arm reward μ^*, the number of plays of a sub–optimal arm a using TS is bounded from above, for any $\epsilon > 0$, by*

$$(1 + \epsilon)\frac{1}{\mathbb{KL}(\mu_a, \mu^*)}(\log T + \log\log T) + O(1).$$

Our plan is to apply Theorem 2.1 in two different cases: to bound the number of times a sub-optimal cluster is played and to bound the number of plays of a sub-optimal arm in the optimal cluster. However, the theorem not directly applicable to the number of plays of a sub-optimal cluster, $N_{C,T}$, since the reward distribution is drifting as the policy is learning about the arms within C. Nevertheless, we can use a comparison argument to bound the number of plays of a sub-optimal cluster by plays in an auxiliary problem with stationary reward distributions and get the following lemma.

Lemma 2.2. *For any $\epsilon > 0$ and assuming strong dominance, the expected number of plays of a sub-optimal cluster C at time T using TSC is bounded from above by*

$$E[N_{C,T}] \leqslant \frac{1+\epsilon}{\mathbb{KL}(\overline{\mu}_C, \underline{\mu}_{C^*})}(\log T + \log\log T) + O(1).$$

We can use Lemma 2.2 to derive the following instance-dependent regret bound for TSC.

Theorem 2.3. *For any $\epsilon > 0$, the expected regret of TSC under the assumption of strong dominance is bounded from above by*

$$(1+\epsilon)\left(\sum_{C \neq C^*} \frac{\Delta_C}{\mathbb{KL}(\overline{\mu}_C, \underline{\mu}_{C^*})} + \sum_{a \in C^*} \frac{\Delta_a}{\mathbb{KL}(\mu_a, \mu^*)}\right)\log T + o(\log T).$$

We can derive an instance-independent upper bound from Theorem 2.3 which only depends on number of clusters, number of arms in the optimal cluster and the quality of the clustering . Now, define γ_C as the ratio between width of the optimal cluster and the distance of C to the optimal cluster:

$$\gamma_C := \begin{cases} w^*/d_C, & C \neq C^* \\ 0, & \text{otherwise} \end{cases}$$

and let $\gamma := \sum_C \gamma_C / K$. We arrive at the following result.

Theorem 2.4. *Assume strong dominance and let A^* be the number of arms in the optimal cluster and K the number of sub-optimal clusters. The expected regret of TSC is bounded from above by $\mathbb{E}[R_T] \leqslant O\left(\sqrt{(A^* + K(1+\gamma))T\log T}\right)$.*

Clustering Quality and Regret. As a sanity check, we note that if the expected rewards of all arms in the optimal cluster are equal we have $\gamma = 0$ and the bound in Theorem 2.4 reduces to the bound for the standard MAB in (Agrawal and Goyal 2017) with $K + 1$ arms. On the other hand, if the optimal cluster has a large width along with many sub-optimal clusters with a small distance to the optimal cluster γ becomes large and little is gained from the clustering. Two standard measures of cluster quality are the (a) the maximum diameter/width of a cluster and (b) inter-cluster separation. We see that for our upper bound, *only the width of the optimal cluster and the separation of other clusters from the optimal cluster* are important. These dependencies are consistent with the observations in Pandey et al. (2007), which suggest that high cohesiveness within the optimal cluster and large separation are crucial for achieving low regret. However our analysis is more precise than their observations and we also provide rigorous regret bounds.

2.3 Lower bounds for disjoint clustering

In the case of Bernoulli distributed rewards we can derive the following lower bound for the instance dependent case using the pioneering works of Lai and Robbins (1985).

Theorem 2.5. *The expected regret for any policy, on the class of bandit problems with Bernoulli distributed arms clustered such that strong dominance holds, is bounded from below by*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[R_T]}{\log T} \geqslant \sum_{a\in C^*} \frac{\Delta_a}{\mathbb{KL}(\mu_a, \mu^*)} + \sum_{C\neq C^*} \frac{\Delta_C}{\mathbb{KL}(\underline{\mu}_C, \mu^*)}$$

We compare the lower bound in Theorem 2.5 to our instance-dependent upper bound in Theorem 2.3 and we see that the regret suffered in TSC from playing sub-optimal clusters asymptotically differs from the corresponding term in the lower bound by a factor depending on the width of the clusters since

$$\mathbb{KL}(\overline{\mu}_C, \underline{\mu}_{C^*}) = \mathbb{KL}(\underline{\mu}_C + w_C, \mu^* - w^*) \leqslant \mathbb{KL}(\underline{\mu}_C, \mu^*).$$

Thus, as the width of the clusters goes to zero, the regret of TSC approaches the lower bound. However, as also discussed in Jedor et al. (2019) it is unclear whether the lower bound derived in Theorem 2.5 can be matched by any algorithm since it doesn't depend at all on the quality on the given clustering and assumes the optimal policy to always play the worst action in sub-optimal clusters.

The following minimax lower bound follows trivially from the $\Omega(\sqrt{NT})$ minimax lower bound for standard MAB (Auer, Cesa-Bianchi, Freund, et al. 1998) by considering the two cases: where all clusters are singletons and all arms are in one cluster.

Theorem 2.6. *Let K be the number of sub-optimal clusters and let A^* be the number of arms in the optimal cluster. The expected regret for any policy, on the class of bandit problems with Bernoulli distributed arms clustered such that strong dominance holds, satisfies $\mathbb{E}[R_T] \geqslant \Omega(\sqrt{(A^* + K)T})$.*

Let $d > 0$ be the smallest distance between any sub-optimal and the optimal cluster. We compare Theorem 2.6 to the upper bound in Theorem 2.4 and observe that $\sqrt{(A^* + K)T} \leqslant \sqrt{(A^* + (1+\gamma)K)T} \leqslant \sqrt{\left(1 + \frac{1}{d}\right)}\sqrt{(A^* + K)T}$. Hence, our upper bound in Theorem 2.4 matches the lower bound up to logarithmic factors and a constant depending on the separation of the clusters.

2.4 Regret analysis HTS

Assume we have N Bernoulli arms clustered into a tree $\mathcal{T}$ and for simplicity we assume it to be perfectly height-balanced. We denote the sub-tree corresponding to node j on level i as $\mathcal{T}_i^j$ and on each level i we denote the sub-tree containing the optimal arm as $\mathcal{T}_i^*$. Let $\mathcal{T}_{i+1}^j$, $j \in [1, K_i^*]$, denote sub-trees spanned by the child nodes of the root in $\mathcal{T}_i^*$, where K_i^* is the number of children of the root in $\mathcal{T}_i^*$. W.l.o.g let $j = 1$ be the sub-tree, $\mathcal{T}_{i+1}^1$, that contains the optimal action. For each sub-tree $\mathcal{T}_i^j$ we define $\Delta_i^j := \mu^* - \max_{a\in\mathcal{T}_i^j} \mu_a$ and $d_j^i := \min_{a\in\mathcal{T}_i^*} \mu_a - \max_{a\in\mathcal{T}_i^j} \mu_a$.

Assumption 2.2 (Hierarchical Strong Dominance)**.** We assume $d_i^j > 0$, $\forall i, j$ except for $\mathcal{T}_i^*$.

Under this assumption the results in Theorem 2.3 can be naturally extended to HTS by recursively applying Theorem 2.3.

Theorem 2.7. *Assuming hierarchical strong dominance. For any $\epsilon > 0$, the expected regret of HTS is upper bounded by*

$$(1+\epsilon)\left(\sum_{i=0}^{L-1}\sum_{j=2}^{K_i^*}\frac{\Delta_i^j}{(d_j^i)^2}+\sum_{a\in\mathcal{T}_L^*}\frac{1}{\Delta_a}\right)\log T + o(\log T).$$

For $L = 0$ Theorem 2.4 reduces to the instance-dependent bound for standard TS and for $L = 1$ it reduces to the bound for TSC presented in Theorem 2.3. Hierarchical structures and bandits have previously been studied in the prominent works Coquelin and Munos (2007) and Bubeck, Munos, et al. (2011) which assumes there is a known smoothness. Here we do not make such assumptions and Theorem 2.7 instead relies on an assumption regarding the ordering of the tree.

Plausibility of Hierarchical Strong Dominance. The hierarchical strong dominance assumption is perhaps too strong for a general hierarchical clustering but it might be reasonable for shallow trees. One example is in e-commerce where products can be organized into sub-categories and later categories. A user might have a strong preference for the sub-category "Football" in the category "Sports".

3 Contextual bandit with linear rewards and clustered arms

In this section, we consider the MABC problem in its contextual variant with linear expected rewards (CBC). As in the classic CB, there is for each arm $a \in \mathcal{A}$ an, a priori, unknown vector $\theta_a \in \mathbf{R}^d$. At each time t, the learner observes a context vector $x_t \in \mathbf{R}^d$ and the expected reward for each arm a at time t, given that the learner has observed x_t, is $\mathbb{E}[r_t(a)|x_t] = x_t^\top \theta_a$. Similar to MABC, the learner has, in addition, access to a clustering of the N arms and for CBC we assume the arms to be clustered into a set of $\mathcal{K}$ disjoint clusters.

For the CBC we extend TSC, Algorithm 5.1, to LinTSC, as defined in Algorithm 5.3. At each level of LinTSC, we use the Thompson sampling scheme developed for standard CB in Agrawal and Goyal (2012).

4 Experimental results

4.1 Stochastic multi-armed bandit

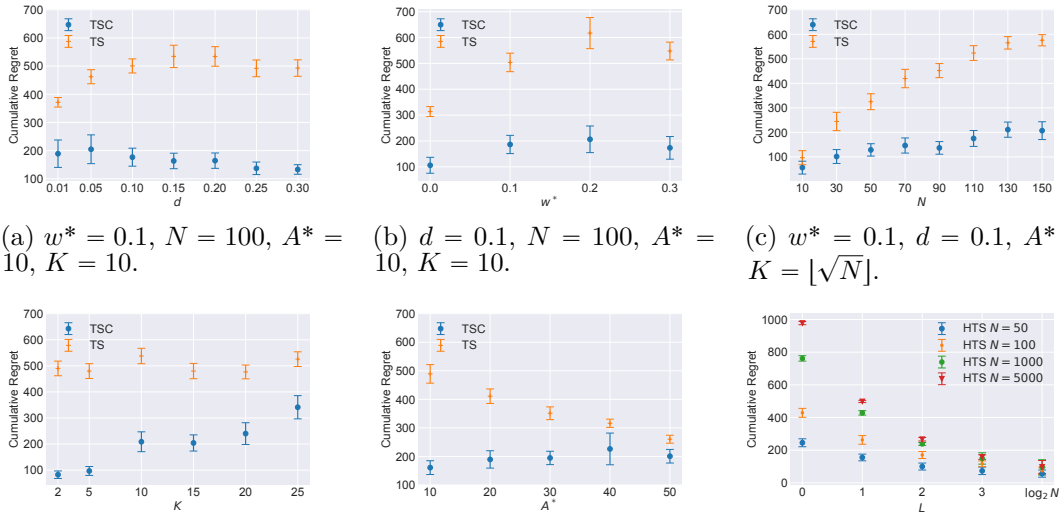
Strong dominance. We generate synthetic data, for which strong dominance holds, in the following way: We have N arms and each arm i is Bernoulli distributed with reward probability p_i. The arms are clustered into K clusters and we have A^* arms in the optimal cluster. For the remaining $N - A^*$ arms we assign each

Algorithm 5.3 LinTSC

Require: $v > 0$
Set $B_c = \mathbf{1}_d$, $f_c = \mathbf{0}_d$, $\mu_c = \mathbf{0}_d$, $B_{c,i} = \mathbf{1}_d$, $f_{c,i} = \mathbf{0}_d$, $\mu_{c,i} = \mathbf{0}_d$
for $t = 1, ..., T$ **do**
Observe context x_t
Sample $\theta_c \sim \mathcal{N}(\mu_c^\top x_t, v x_t^\top B_c^{-1} x_t)$
Consider cluster $k = \arg\max_c \theta_c$
Sample $\theta_{k,i} \sim \mathcal{N}(\mu_{k,i}^\top x_t, v x_t^\top B_{k,i}^{-1} x_t)$
Play arm $a = \arg\max_i \theta_{k,i}$
Observe reward r_t and update $B_k = B_k + x_t x_t^\top$, $B_{k,a} = B_{k,a} + x_t x_t^\top$, $f_k = f_k + r x_t$, $f_{k,i} = f_{k,i} + r x_t$, $\mu_k = B_k^{-1} f_k$ and $\mu_{k,i} = B_{k,i}^{-1} f_{k,i}$.
end for

arm to one of the sub-optimal clusters with uniform probability. We set the reward probability of the best arm in the optimal cluster to be 0.6 and for the worst arm in the optimal cluster we set it to be $0.6 - w^*$. For the remaining $A^* - 2$ arms in the optimal cluster we draw the reward probability from $\mathcal{U}(0.6 - w^*, 0.6)$ for each arm. In each sub-optimal cluster we set the probability of the best arm to be $0.6 - w^* - d$ and for the worst arm to be $0.5 - w^* - d$, the probability for the remaining arms are drawn from $\mathcal{U}(0.5 - w^* - d, 0.6 - w^* - d)$. The optimal cluster will then have a width of w^* and the distance from each sub-optimal cluster to the optimal cluster will be d. In Figures 5.1a–5.1e, we run TS and TSC on the same instances for $T = 3000$ time steps, varying the different instance parameters and plotting the cumulative regret of each algorithm at the final time step T. For each set of parameters we evaluate the algorithms using 50 different random seeds and the error bars corresponds to ± 1 standard deviation. In Figures 5.1a and 5.1b, we observe that the cumulative regret scales depending on the clustering quality parameters d and w^* as suggested by our bounds in Section 2.2—that is, the cumulative regret of TSC decreases as d increases and increases as w^* increases. In Figure 5.1c, we observe that the linear dependence in N for TS is changed to a linear dependence in K and A^*, Figures 5.1d and 5.1e, which greatly reduces the regret of TSC compared to TS as the size of the problem instance increases. In Figure 5.1e we also see that as the number of arms in the optimal cluster, A^*, increases to be a substantial amount of the total number of arms, the gain from using TSC compared to TS vanishes.

Hierarchical strong dominance. We generate a bandit problem by first uniformly sample N Bernoulli arms from $\mathcal{U}(0.1, 0.8)$ followed by recursively sorting and merging the arms into a balanced binary tree, which has the hierarchical strong dominance property. In Figure 5.1f, we ran the algorithms for $T = 3000$ over 50 random seeds and illustrated how the cumulative regret at time T of HTS changes as we alter the depth L of the given tree and the total number of arms N. Note that $L = 0$ corresponds to TS and $L = 1$ corresponds to TSC. We observe that as the size of the problem instance grows, i.e increasing N, using more levels in the tree becomes more beneficial due to aggressive exploration scheme of HTS. Hence, once we realize



(a) $w^* = 0.1$, $N = 100$, $A^* = 10$, $K = 10$.

(b) $d = 0.1$, $N = 100$, $A^* = 10$, $K = 10$.

(c) $w^* = 0.1$, $d = 0.1$, $A^* = K = \lfloor\sqrt{N}\rfloor$.

(d) $w^* = d = 0.1$, $N = 100$, $A^* = 10$.

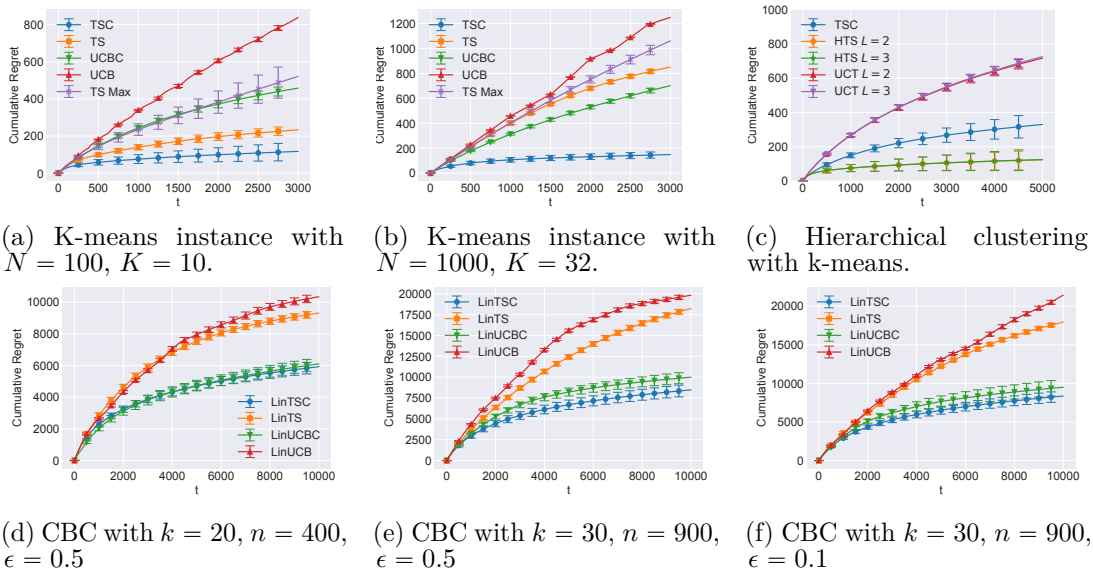
(e) $w^* = d = 0.1$, $N = 100$, $K = 10$.

(f) Dependence on the number of levels L.

Figure 5.1: Strong and Hierarchical Strong Dominance.

that one sub-tree is better than the other we discard all arms in the corresponding sub-optimal sub-tree. Connecting back to Theorem 2.7 we see that HTS gets only a dependence $O(\log_2 N)$ in the number of arms when using the full hierarchical tree in Figure 5.1f.

Violation of assumptions. In a real world setting, assuming that strong dominance and especially hierarchical strong dominance holds completely is often too strong. We thus evaluate our algorithms on instances for which these assumptions are violated. We generate N arms by for each arm i we sample a value $x_i \sim \mathcal{U}(0, 1)$. We cluster the arms into K clusters, based on the values $\{x_i\}$, using K-means. The reward distribution of each arm i is a Bernoulli distribution with mean $f(x_i)$ where $f(x) = \frac{1}{2}(\sin 13x \sin 27x + 1)$. This function is illustrated in the supplementary material, Appendix A, and has previously been used to evaluate bandit algorithms in Bubeck, Munos, et al. (2011), the smoothness of the function ensures arms within the same cluster to have similar expected rewards, on the other hand the periodicity of sin yields many local optima and the optimal cluster won't strongly dominate the other clusters. On these instances, we benchmark TSC against two another algorithms proposed for MABC, UCBC (Pandey et al. 2007; Bouneffouf et al. 2019) and TSMax (Zhao et al. 2019). We also benchmark against UCB1 (Auer, Cesa-Bianchi, and Fischer 2002) and TS which both considers the problem as a standard MAB, making no use of the clustering. We run the algorithms on two different instances, one with $N = 100$ and $K = 10$ and the other one with $N = 1000$ and $K = 32$. For each instance we run the algorithms on 100 different random seeds and we present the results in Figure: 5.2a and 5.2b, the error bars corresponds to ± 1 standard deviation. TSC outperforms the other algorithms on both instances and especially on the larger instance where there is a big gap between the regret of TSC and the regret of the other algorithms. In order to test HTS we generate an instance, as above, with



(a) K-means instance with $N = 100$, $K = 10$.

(b) K-means instance with $N = 1000$, $K = 32$.

(c) Hierarchical clustering with k-means.

(d) CBC with $k = 20$, $n = 400$, $\epsilon = 0.5$

(e) CBC with $k = 30$, $n = 900$, $\epsilon = 0.5$

(f) CBC with $k = 30$, $n = 900$, $\epsilon = 0.1$

Figure 5.2: CBC and violation of assumptions in MABC.

$N = 5000$ and $K = 15$ and construct a tree by recursively breaking each cluster up into 15 smaller clusters using k-means. In Figure 5.2c we show the performance of HTS for two different levels, $L = 2, 3$, compared to TSC using the clusters at level $L = 1$ in the tree and also compared to the UCT-algorithm (Kocsis and Szepesvári 2006) using the same levels of the tree as HTS. We averaged over 100 random seeds. The HTS performs well on this problem and is slightly better than TSC while both HTS and TSC outperforms UCT. We present more empirical results for MABC in the supplementary material.

4.2 Contextual bandit

We generate contextual data in the same way as in Bouneffouf et al. (2019). We have K clusters and N arms. Each arm j is randomly assigned to a cluster i. For each cluster i we sample a centroid $\theta_i^c \sim \mathcal{N}(0, \mathbf{1}_5)$ and define a coefficient for each arm j in the cluster as $\theta_j = \theta_i^c + \epsilon v$, $v \sim \mathcal{N}(0, \mathbf{1}_5)$. We take the reward of an arm to be $\mathcal{U}(0, 2\theta_j^\intercal x)$ where x is the given context. The reward becomes linear and we can control the expected diameter of a cluster by varying ϵ.

We benchmark LinTSC against the UCB-based counterpart LinUCBC (Bouneffouf et al. 2019) and the standard algorithms LinTS (Agrawal and Goyal 2012) and LinUCB (Li et al. 2010), which treats the problem as a standard CB. We ran the algorithms on three different instances presented in Figures 5.2d, 5.2e and 5.2f, over 25 different random seeds and the error bars corresponds to ± 1 standard deviation. We run all algorithms with there corresponding standard parameter ($v = 1$ for LinTS and LinTSC, $c = 2$ for LinUCB and LinUCBC). We see a clear improvement between not using the clustering (TS) and using the clustering (TSC). LinTSC performs slightly better than LinUCBC as the problem becomes larger w.r.t number of arms and clusters, Figures 5.2e and 5.2f.

5 Related work

Bandits are now a classical subject in machine learning and recent textbook treatments are Bubeck and Cesa-Bianchi (2012), Slivkins (2019), and Lattimore and Szepesvári (2020). The MABC and CBC can be considered as natural special cases of the more general finite-armed structured bandit which is studied in (Lattimore and Munos 2014; Combes et al. 2017; Gupta, Joshi, et al. 2018; Gupta, Chaudhari, et al. 2019). To the best of our knowledge, the idea of clustered arms was first studied in Pandey et al. (2007) and the MABC corresponds to their undiscounted MDP setup for which the authors propose a general two-level bandit policy and gives theoretical justifications on how the regret scales depending on the characteristics of the clustering, but without stating rigorous regret bounds. Bandits with clustered arms were also recently studied in Bouneffouf et al. (2019) and Jedor et al. (2019) and both papers prove regret bounds for UCB-styled algorithms in the MABC under various assumptions on the clustering. Bouneffouf et al. (2019) is the work most related to ours since they consider a two-level UCB scheme and regret bounds that exhibits similar dependence on the clustering quality as our bounds. In Zhao et al. (2019) the authors propose a two-level TS algorithm where the belief of a cluster is set to the belief of the best performing arm in the cluster so far and the authors give no theoretical analysis of its regret. Clustered arms also appear in the regional bandit model (Wang et al. 2018; Singh et al. 2020) under the assumption that all arms in one cluster share the same underlying parameter. Another model related to our work is the latent bandit (Maillard and Mannor 2014; Hong et al. 2020) where the reward distributions depends on a latent state and the goal of the learner is to identify this state.

Bandits and tree structures are studied using a UCB-styled algorithm for Monte-Carlo-based planning in the influential work Kocsis and Szepesvári (2006) and later studied for various bandit problems with smoothness in the seminal works Coquelin and Munos (2007) and Bubeck, Munos, et al. (2011).

We have based our bandit algorithms on the classical method Thompson sampling (Thompson 1933) which has been shown to perform well in practise (Chapelle and Li 2011) and for which rigorous regret analyses recently have been established for the standard MAB in Kaufmann et al. (2012) and Agrawal and Goyal (2017). The contextual version of Thompson sampling we use in our two-level scheme for CBC was originally proposed and analyzed for standard CB in Agrawal and Goyal (2012) and recently revisited in Abeille and Lazaric (2017).

6 Conclusions

In this paper, we have addressed the stochastic multi-armed bandit problem and the contextual bandit with clustered arms and proposed algorithms based on multi-level Thompson sampling. We have shown that our algorithms can be used to drastically reduce the regret when a clustering of the arms is known and that these algorithms are competitive to its UCB-based counterparts. We think that the simplicity of our algorithms and the fact that one can easily incorporate prior knowledge makes them

well-suited options for bandit problems with a known clustering structure of the arms. In the future we would like to explore how the regret of TSC behaves under weaker assumptions on the clustering. We want to determine what are sufficient properties of the clustering to ensure sub-linear regret of LinTSC.

Acknowledgments

This work was supported by funding from CHAIR (Chalmers AI Research Center) and from the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The computations in this work were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC).

A Proofs

A.1 Lemma 2.2

Assume $c^* = 1$ is the cluster containing the optimal arm. We want to bound $\mathbb{E}[N_c]$ for some sub-optimal cluster c. Let $\underline{\mu}_c$ be the smallest mean in cluster c and let $\theta_{c,t}$ be the sample drawn from the belief of TSC for c at time t.

If cluster c is played at time t, i.e. $C_t = c$, then one of the two events need to happen

- The sample, $\theta_{1,t}$ for cluster 1 satisfy
$$\theta_{1,t} \leqslant \underline{\mu}_1 - \sqrt{\frac{6 \log t}{N_1}}$$
- Or $\theta_{1,t} > \underline{\mu}_1 - \sqrt{\frac{6 \log t}{N_1}}$ but $C_t = c$ anyway.

Thus the expected number of pulls, N_c, of cluster c can be decomposed as

$$\mathbb{E}[N_{c,T}] \leqslant \sum_{t=1}^{T} P\left(\theta_{1,t} \leqslant \underline{\mu}_1 - \sqrt{\frac{6 \log t}{N_{1,t}}}\right) \tag{A.1}$$
$$+ \sum_{t=1}^{T} P\left(\theta_{c,t} > \underline{\mu}_1 - \sqrt{\frac{6 \log t}{N_{c,t}}}\right). \tag{A.2}$$

Let $\leqslant$ denote stochastic domination, i.e. $X \leqslant Y$ iff $P(X \geqslant x) \leqslant P(Y \geqslant x)\, \forall x$. Let $S_{c,t}$ and $F_{c,t}$ be the corresponding number of success and fail observations from cluster c at time t, as in Algorithm 5.1. That is,

$$S_{c,t} = 1 + \sum_{i}^{N_{c,t}} r_i \tag{A.3}$$

$$F_{c,t} = 1 + \sum_{i}^{N_{c,t}} 1 - r_i \tag{A.4}$$

where r_i is a reward drawn from some arm in cluster c.

To bound the first term in the inequality, $\sum_{t=1}^{T} P\left(\theta_{1,t} \leqslant \underline{\mu}_1 - \sqrt{\frac{6\log t}{N_1}}\right)$, we consider an auxiliary sample $\theta'_{1,t} \sim \text{Beta}(S'_{1,t}, F'_{1,t})$ such that

$$S'_{1,t} = 1 + \sum_{i}^{N_{1,t}} r'_i \tag{A.5}$$

$$F'_{1,t} = 1 + \sum_{i}^{N_{1,t}} 1 - r'_i \tag{A.6}$$

where r'_i corresponds to sample drawn from the worse arm in cluster 1 with mean $\underline{\mu}_1$. It is easy to verify that $S'_{c,t} \leqslant S_{c,t}$ and $F_{c,t} \leqslant F'_{c,t}$ and thus $\theta'_{1,t} \leqslant \theta_{1,t}$ [1]. Thus we have

$$P\left(\theta_{1,t} \leqslant \underline{\mu}_1 - \sqrt{\frac{6\log t}{N_{1,t}}}\right) \leqslant P\left(\theta'_{1,t} \leqslant \underline{\mu}_1 - \sqrt{\frac{6\log t}{N_{1,t}}}\right)$$

and using Lemma 1 from Kaufmann et al. (2012) we can conclude that

$$\sum_{t=1}^{\infty} P\left(\theta'_{1,t} \leqslant \underline{\mu}_1 - \sqrt{\frac{6\log t}{N_{1,t}}}\right) \leqslant Q < \infty \tag{A.7}$$

where Q is some constant.

We proceed in similar fashion to bound

$$\sum_{t=1}^{T} P\left(\theta_{c,t} > \underline{\mu}_1 - \sqrt{\frac{6\log t}{N_c}}\right). \tag{A.8}$$

We note that

$$P\left(\theta_{c,t} > \underline{\mu}_1 - \sqrt{\frac{6\log t}{N_{c,t}}}\right) \leqslant P\left(\theta''_{c,t} > \underline{\mu}_1 - \sqrt{\frac{6\log t}{N_{c,t}}}\right) \tag{A.9}$$

where $\theta''_{c,t} \sim \text{Beta}(S''_{c,t}, F''_{c,t})$ with

$$S''_{c,t} = 1 + \sum_{i}^{N_{c,t}} r''_i \tag{A.10}$$

$$F''_{c,t} = 1 + \sum_{i}^{N_{c,t}} 1 - r''_i \tag{A.11}$$

where r''_i are observations from the best arm in cluster c with mean $\overline{\mu}_c$. By the same reasoning as previously we get $\theta_{c,t} \leqslant \theta''_{c,t}$. Applying Theorem 2.1 yields

$$\mathbb{E}[N_{c,T}] \leqslant (1+\epsilon)\frac{\log T + \log\log T}{\mathbb{KL}(\overline{\mu}_c, \underline{\mu}_{c*})} + O(1) \tag{A.12}$$

for $\epsilon > 0$.

[1] To see this, we use the beta-binomial trick $F^{\text{Beta}}_{a,b} = 1 - F^{\text{Binomial}}_{a+b-1,x}(a-1)$ and note that if $a+b = c+d = q$ and $a \geqslant c$ then $F^{\text{Binomial}}_{q-1,x}(c-1) \leqslant F^{\text{Binomial}}_{q-1,x}(a-1)$ which gives, using the trick, $F^{\text{Beta}}_{a,b}(x) \leqslant F^{\text{Beta}}_{c,d}(x)$, $\forall x \in [0,1]$, which implies stochastic dominance.

A.2 Theorem 2.3

We can decompose the regret into

$$E[R_T] = \sum_{C \neq C^*} \sum_{a \in C} \Delta_a \mathbb{E}[N_{a,T}] + \sum_{a \in C^*} \Delta_a \mathbb{E}[N_{a,T}]$$

where the first term consider the regret suffered from playing sub-optimal clusters and the second term regret suffered from playing sub-optimal arms within the optimal cluster. The second term can be bounded by just applying Theorem 2.1 for $\epsilon > 0$

$$\sum_{a \in C^*} \Delta_a \mathbb{E}[N_{a,T}] \leqslant (1+\epsilon) \sum_{a \in C^*} \frac{1}{\Delta_a} \log T + o(\log T).$$

To bound the first term, consider sub-optimal cluster C and let $N_{C,T}$ denote the number of times we play C. Let a_C^* be the action with highest expected reward in C. Then for any other $a \in C$, $a \neq a_C^*$ we can bound the number of plays, $N_{a,T_{C,T}}$, by Theorem 2.1

$$\begin{aligned}\mathbb{E}[N_{a_C,N_{C,T}}] &\leqslant (1+\epsilon)\frac{1}{\mathbb{KL}(\mu_a, \mu_{a^*})}(\log N_{C,T} + \log\log N_{C,T}) \\ &+ O(1)\end{aligned}$$

and for a_C^* we have

$$\mathbb{E}[N_{a_C^*,N_{C,T}}] \leqslant \mathbb{E}[N_{C,T}].$$

From Lemma 2.2 we know that for $\epsilon > 0$

$$\mathbb{E}[N_{C,T}] \leqslant (1+\epsilon)\frac{1}{\mathbb{KL}(\overline{\mu}_C, \underline{\mu}_{C^*})}(\log T + \log\log T) + O(1)$$

and we thus get a $\log\log T$ dependence on all arms in C except the one with highest expected reward

$$\begin{aligned}\mathbb{E}[N_{a,N_{C,T}}] &\leqslant (1+\epsilon)\frac{1}{\mathbb{KL}(\mu_a, \mu_{a^*})}\log\log T + o(\log\log T) \\ \mathbb{E}[N_{a_C^*,N_{C,T}}] &\leqslant (1+\epsilon)\frac{1}{\mathbb{KL}(\overline{\mu}_C, \underline{\mu}_{C^*})}\log T + o(\log T).\end{aligned}$$

Therefore we can bound the regret suffered from sub-optimal clusters for any $\epsilon > 0$ as

$$\begin{aligned}&\sum_{C \neq C^*} \sum_{a \in C} \Delta_a \mathbb{E}[N_{a,T}] \\ &\leqslant (1+\epsilon)\Big(\sum_{C \neq C^*} \frac{\Delta_C}{\mathbb{KL}(\overline{\mu}_C, \underline{\mu}_{C^*})}\log T + \\ &+ \sum_{a \in C, a \neq a^*} \frac{\Delta_a}{\mathbb{KL}(\mu_a, \mu_{a^*})}\log\log T\Big) + o(\log T) \\ &\leqslant (1+\epsilon)\sum_{C \neq C^*} \frac{\Delta_C}{\mathbb{KL}(\overline{\mu}_C, \underline{\mu}_{C^*})}\log T + o(\log T).\end{aligned}$$

Combining with the bound on regret within the optimal cluster C^* yields the instance-dependent regret bound

$$\begin{aligned}&\mathbb{E}[R_T] \leqslant \\ &\leqslant (1+\epsilon)\left(\sum_{C\neq C^*}\frac{\Delta_C}{\mathbb{KL}(\overline{\mu}_C, \underline{\mu}_{C^*})} + \sum_{a\in C^*}\frac{\Delta_a}{\mathbb{KL}(\mu_a,\mu^*)}\right)\log T \\ &+ o(\log T).\end{aligned}$$

A.3 Theorem 2.4

We rewrite $\Delta_C = d_C + w^*$ where w^* is the width of the optimal cluster and hence by the definition of γ_C we have

$$\Delta_C = (1+\gamma_C)d_C.$$

By Pinsker's inequality we have

$$\mathbb{KL}(\overline{\mu}_C, \underline{\mu}_{C^*}) \geqslant 2d_C^2$$

and for arms in the optimal cluster we have

$$\mathbb{KL}(\mu_a, \mu^*) \geqslant 2\Delta_a^2$$

Thus, the instance-dependent regret bound can be upper-bounded by

$$\frac{1+\epsilon}{2}\left(\sum_{C\neq C^*}\frac{1+\gamma_C}{d_C} + \sum_{a\in C^*}\frac{1}{\Delta_a}\right)\log T + o(\log T).$$

Let $\Delta > 0$.

- For all clusters C and arms $a \in C^*$ such that $d_C, \Delta_a < \Delta$, the cumulative regret from these are upper-bounded by ΔT.
- For each cluster C such that $d_C \geqslant \Delta$ the amount of regret suffered from playing C is $O(\frac{1+\gamma_C}{\Delta}\log T)$ and for each $a \in C^*$ the regret suffered is $O(\frac{1}{\Delta}\log T)$. In total this is $O(\frac{A^*+K(1+\gamma)}{\Delta}\log T)$.

Combining this yields

$$\mathbb{E}[R_T] \leqslant O(\Delta T + \frac{A^* + K(1+\gamma)}{\Delta}\log T).$$

Since this holds $\forall\Delta > 0$ we pick $\Delta = \sqrt{\frac{(A+K(1+\gamma))\log T}{T}}$ and hence,

$$\mathbb{E}[R_T] \leqslant O\left(\sqrt{(A^* + K(1+\gamma))T\log T}\right).$$

A.4 Theorem 2.5

We make use of the pioneering work of (Lai and Robbins 1985) which gives that

$$\lim_{T\to\infty} \inf \frac{\mathbb{E}[R_T]}{\log T} \geqslant \sum_a \frac{\Delta_a}{\mathbb{KL}(\mu_a, \mu^*)} \tag{A.13}$$

for a standard multi-armed bandit with Bernoulli rewards. We can decompose the regret over sub-optimal clusters and sub-optimal arms in the optimal cluster

$$\mathbb{E}[R_T] = \sum_{C\neq C^*} \sum_{a\in C} \Delta_a \mathbb{E}[N_{a,T}] + \sum_{a\in C^*} \Delta_a \mathbb{E}[N_{a,T}],$$

and using the fact that the regret suffered within a sub-optimal cluster is bounded from below by the smallest regret in the cluster

$$\sum_{a\in C} \Delta_a \mathbb{E}[N_{a,T}] \geqslant \Delta_C \sum_{a\in C} \mathbb{E}[N_{a,T}].$$

Now we get the proposed bound by independently bounding each term from below by Equation:A.13 and using the fact that for any cluster C and any arm $a \in C$ we have

$$\mathbb{KL}(\mu_a, \mu^*) \geqslant \mathbb{KL}(\underline{\mu}_C, \mu^*).$$

A.5 Theorem 2.6

First consider the case where all arms are assigned to the same cluster. Any algorithm needs to at least have a $\sqrt{A^*T}$ dependence in the regret otherwise the lower bound $\Omega(\sqrt{NT})$ would be violated.

Secondly, consider the case where all clusters only contain one arm each. We have that any algorithm needs at least a $\sqrt{KT}$ dependence otherwise $\Omega(\sqrt{NT})$ would be violated.

Since $\sqrt{K + A^*} \leqslant \sqrt{K} + \sqrt{A^*}$ it follows that for any algorithm we have

$$\mathbb{E}[R_T] \geqslant \Omega(\sqrt{(A^* + K)T}).$$

A.6 Theorem 2.7

We decompose the cumulative regret into

$$R_T := \sum_{\mathcal{T}_1^j \neq \mathcal{T}_1^*} \sum_{a\in\mathcal{T}_1^j} \Delta_a \mathbb{E}[N_{a,T}] + \sum_{a\in\mathcal{T}_1^*} \Delta_a \mathbb{E}[N_{a,T}].$$

Since strong dominance holds on each level we bound the first sum by $\sum_{j=2}^{K_0^*} \frac{\Delta_1^j}{(2d_1^j)^2} \log T +$ $o(\log T)$ using Theorem 2.3, where $(2d_1^j)^2$ follows from Pinsker's inequality for Bernoulli distributions. We are left with bounding the regret from

$$\sum_{a\in\mathcal{T}_1^*} \Delta_a \mathbb{E}[N_{a,T}] = \sum_{j=2}^{K_1^*} \sum_{a\in\mathcal{T}_2^j} \Delta_a \mathbb{E}[N_{a,T}] + \sum_{a\in\mathcal{T}_2^*} \Delta_a \mathbb{E}[N_{a,T}].$$

And we recursively apply Theorem 2.3 to bound the first time like above, until we reach level L for which we use Theorem 2.1 along with Pinsker's inequality to get

$$\sum_{a \in \mathcal{T}_L^*} \Delta_a \mathbb{E}[N_{a,T}] \leqslant (1+\epsilon) \sum \frac{1}{\Delta_a} \log T + o(\log T)$$

B Empirical evaluation MABC

To give an example where HTS achieves linear regret while TSC exhibits sub-linear regret we have $N = 500$ arms and for each arm a_i we draw a vector x_i from $x_i \sim \mathcal{U}([0,1]^2)$. We cluster the arms into $K = 20$ clusters using k-means and use that clustering in TSC. We also cluster the arms using agglomerative clustering and use the resulting tree for HTS and UCT. We take the reward for each arm a_i to be Bernoulli distributed with mean reward

$$f(x_1, x_2) = \frac{1}{2} e^{-100(0.2-x_1)^2} + \frac{1}{5} e^{-100(0.7-x_1)^2} + \frac{1}{5} e^{-100(0.7-x_2)^2},$$

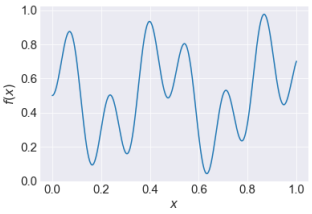
this function is illustrated in Figure 5.3c. This function is chosen such that there is a similarity between close arms but as we go higher up in the tree arms in the same sub-tree may have very different rewards. We run the algorithms for $T = 20\,000$ and over 25 random seeds and in Figure 5.4a we see that both UCT and HTS exhibits linear cumulative regret curve while TSC is still sub-linear since arms clustered together tends to have similar reward. Hence, using the full tree in this case is a too aggressive exploration scheme and we see that care has to be taken in HTS when deciding how deep the hierarchical clustering should be.

We also generated a bandit instance using the function

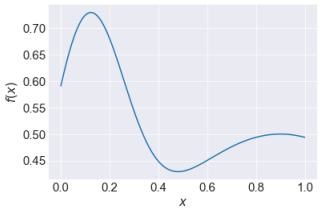
$$f(x) = \frac{1}{2}(e^{-\frac{1}{0.05}(0.1-x)^2} + e^{-\frac{1}{0.8}(0.9-x)^2}),$$

illustrated in Figure 5.3b. This function is considered since it is very smooth and one may assume similar rewards for arms in the same sub-tree of a hierarchical clustering. We generate $N = 50$ arms as before and for TSC we cluster them using k-means with $K = 5$. For HTS and UCT we use agglomerative clustering and consider the full tree. We run the algorithms for $T = 25\,000$ and over 25 random seeds and present the results in Figure 5.4b. We see that for this instance HTS exhibits sub-linear regret and performs better than TSC, for this clustering. This illustrate that the quality of the clustering is very important for the regret, especially for HTS.

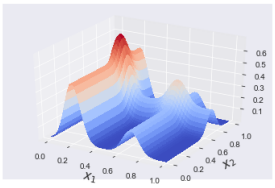
We also compare TS and TSC on an instance where there is no correlation between rewards in a cluster. We take $N = 50$ arms and divide them into $K = 10$ clusters. The reward of each arm is Bernoulli distributed and we draw the expected reward of each arm from $\mathcal{U}(0,1)$. The average over 25 random seeds is presented in Figure 5.4c and as expected we see that TSC has a cumulative regret which is worse then TS, since the quality of the clustering is bad.



(a) The function $f(x) = \frac{1}{2}(\sin 13x \sin 27x + 1)$ used for expected rewards in the evaluation of TSC and HTS with violated assumptions in Section 4.

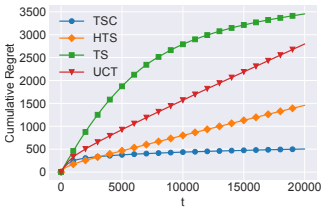


(b) $f(x) = \frac{1}{2}(e^{-\frac{1}{0.05}(0.1-x)^2} + e^{-\frac{1}{0.8}(0.9-x)^2})$ used for expected rewards in Figure 5.4b.

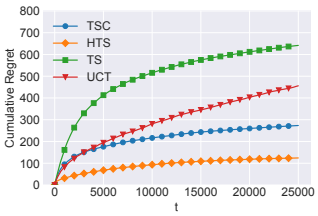


(c) 2 dimensional function used for expected rewards in Figure 5.4a.

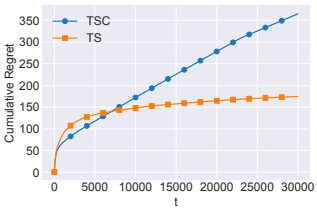
Figure 5.3: Functions used for evaluating TSC and HTS when theoretical assumptions are violated.



(a) Cumulative regret over the 2-d instance with $N = 500$ and $K = 20$.



(b) Cumulative regret over the instance with $N = 50$ arms with expected rewards as $f(x)$ in Figure 5.3b.



(c) Cumulative regret over uniformly assigned expected rewards. $N = 50$ and $K = 10$.

Figure 5.4

References

Abeille, Marc and Alessandro Lazaric (20–22 Apr 2017). "Linear Thompson Sampling Revisited". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Vol. 54. Proceedings of Machine Learning Research, pp. 176–184 (cit. on p. 169).

Agrawal, Shipra and Navin Goyal (Sept. 2012). "Thompson Sampling for Contextual Bandits with Linear Payoffs". In: *30th International Conference on Machine Learning, ICML 2013* (cit. on pp. 165, 168, 169).

Agrawal, Shipra and Navin Goyal (2017). "Near-Optimal Regret Bounds for Thompson Sampling". In: *J. ACM* 64.5, 30:1–30:24 (cit. on pp. 161, 163, 169).

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. 167).

Auer, Peter, Nicolò Cesa-Bianchi, Yoav Freund, and Robert Schapire (July 1998). "Gambling in a rigged casino: The adversarial multi-armed bandit problem". In: *Foundations of Computer Science, 1975., 16th Annual Symposium on* (cit. on p. 164).

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 pp. 162, 167–169).

Bubeck, Sébastien and Nicolò Cesa-Bianchi (2012). "Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems". In: *Foundations and Trends® in Machine Learning* 5.1, pp. 1–122. ISSN: 1935-8237 (cit. on pp. 159, 169).

Bubeck, Sébastien, Remi Munos, Gilles Stoltz, and Csaba Szepesvári (May 2011). "X-Armed Bandits". In: *Journal of Machine Learning Research* 12 (cit. on pp. 165, 167, 169).

Chapelle, Olivier and Lihong Li (2011). "An Empirical Evaluation of Thompson Sampling". In: *Advances in Neural Information Processing Systems 24*, pp. 2249–2257 (cit. on p. 169).

Combes, Richard, Stefan Magureanu, and Alexandre Proutiere (2017). "Minimal Exploration in Structured Stochastic Bandits". In: *Advances in Neural Information Processing Systems 30*, pp. 1763–1771 (cit. on p. 169).

Coquelin, Pierre-Arnaud and Rémi Munos (2007). "Bandit Algorithms for Tree Search". In: *Uncertainty in Artificial Intelligence* (cit. on pp. 165, 169).

Gupta, S., S. Chaudhari, Gauri Joshi, and O. Yağan (2019). "Multi-Armed Bandits with Correlated Arms". In: *ArXiv* abs/1911.03959 (cit. on p. 169).

Gupta, S., Gauri Joshi, and O. Yağan (2018). "Exploiting Correlation in Finite-Armed Structured Bandits". In: *ArXiv* abs/1810.08164 (cit. on p. 169).

Hong, Joey, Branislav Kveton, Manzil Zaheer, Yinlam Chow, Amr Ahmed, and Craig Boutilier (2020). "Latent Bandits Revisited". In: *Advances in Neural Information Processing Systems 33* (cit. on p. 169).

Jedor, Matthieu, Vianney Perchet, and Jonathan Louedec (2019). "Categorized Bandits". In: *Advances in Neural Information Processing Systems*. Vol. 32, pp. 14422–14432 (cit. on pp. 162, 164, 169).

Kaufmann, Emilie, Nathaniel Korda, and Rémi Munos (2012). "Thompson Sampling: An Asymptotically Optimal Finite-Time Analysis". In: *Algorithmic Learning Theory - 23rd International Conference, ALT 2012. Proceedings*. Vol. 7568, pp. 199–213 (cit. on pp. 162, 169, 171).

Kocsis, Levente and Csaba Szepesvári (2006). "Bandit based Monte-Carlo Planning". In: *In: ECML-06. Number 4212 in LNCS*, pp. 282–293 (cit. on pp. 168, 169).

Lai, T.L and H Robbins (1985). "Asymptotically efficient adaptive allocation rules". In: *Advances in Applied Mathematics* 6, pp. 4–22 (cit. on pp. 159, 163, 174).

Lattimore, Tor and Remi Munos (2014). "Bounded Regret for Finite-Armed Structured Bandits". In: *Advances in Neural Information Processing Systems*. Vol. 27, pp. 550–558 (cit. on p. 169).

Lattimore, Tor and Csaba Szepesvári (2020). *Bandit Algorithms*. Cambridge University Press. DOI: 10.1017/9781108571401 (cit. on p. 169).

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 - WWW '10* (cit. on p. 168).

Maillard, Odalric-Ambrym and Shie Mannor (May 2014). "Latent Bandits". In: *31st International Conference on Machine Learning, ICML 2014* 1 (cit. on p. 169).

Pandey, Sandeep, Deepayan Chakrabarti, and Deepak Agarwal (2007). "Multi-armed bandit problems with dependent arms". In: *ICML*, pp. 721–728 (cit. on pp. 163, 167, 169).

Singh, Rahul, Fang Liu, Yin Sun, and Ness Shroff (2020). "Multi-Armed Bandits with Dependent Arms". In: *arXiv*. eprint: 2010.09478 (cs.LG) (cit. on p. 169).

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. 169).

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 pp. 160, 169).

Wang, Zhiyang, Ruida Zhou, and Cong Shen (2018). "Regional Multi-Armed Bandits". In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. Vol. 84, pp. 510–518 (cit. on p. 169).

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 pp. 167, 169).

Paper 6

Pure exploration in bandits with linear constraints

Emil Carlsson, Debabrota Basu, Fredrik D. Johansson, Devdatt Dubhashi.

International Conference on Artificial Intelligence and Statistics (AISTATS), 2024.

The paper has been reformatted for uniformity.

Paper 6. Pure exploration in bandits with linear constraints

Emil Carlsson, Debabrota Basu, Fredrik D. Johansson, Devdatt Dubhashi.

Abstract

We address the problem of identifying the optimal policy with a fixed confidence level in a multi-armed bandit setup, when *the arms are subject to linear constraints*. Unlike the standard best-arm identification problem which is well studied, the optimal policy in this case may not be deterministic and could mix between several arms. This changes the geometry of the problem which we characterize via an information-theoretic lower bound. We introduce two asymptotically optimal algorithms for this setting, one based on the Track-and-Stop method and the other based on a game-theoretic approach. Both these algorithms try to track an optimal allocation based on the lower bound and computed by a weighted projection onto the boundary of a normal cone. Finally, we provide empirical results that validate our bounds and visualize how constraints change the hardness of the problem. [1]

1 Introduction

A classical problem in the multi-armed bandit framework is *pure exploration* (Lattimore and Szepesvári 2020), where the task of a learner is to answer some query about a set of actions, also known as arms, by iteratively choosing between the actions and receiving an immediate reward sampled from a distribution associated with the action. A very well-studied problem in this context is Best-Arm Identification (BAI), where a learner is trying to identify the arm with the highest expected reward (Even-Dar et al. 2002; Bubeck et al. 2009; Kalyanakrishnan et al. 2012). The BAI problem has many applications such as hyper-parameter tuning (Li et al. 2017), clinical trials (Aziz et al. 2021), communication networks (Lindståhl et al. 2022) and user studies (Losada et al. 2022). However, many real-world scenarios often involve *constraints on the arms* that must be satisfied. For example, in recommender systems, one may need to ensure diversity and genre constraints (Kunaver and Požrl 2017), or fairness of exposure (Wang, Bai, et al. 2021). In clinical trials, one may need to account for toxicity constraints of the available treatments (Brannath et al. 2009; Chen 2021; Demirel et al. 2022). As a result, standard BAI algorithms are not perfectly fitted in these settings and might have large sample complexity as we show empirically later on in Section 5.

[1] Code available at: `https://github.com/e-carlsson/constraint-pure-exploration`

In this paper, we introduce the problem of pure exploration in bandits with linear constraints where the goal is to identify, with a fixed confidence, a policy that maximizes the expected rewards over arms while satisfying some given constraints. A set of constraints may change the nature of the pure exploration problem fundamentally. In particular, the optimal policy *may not be deterministic*, and *finding the best arm may not be sufficient*. Let us consider the following example.

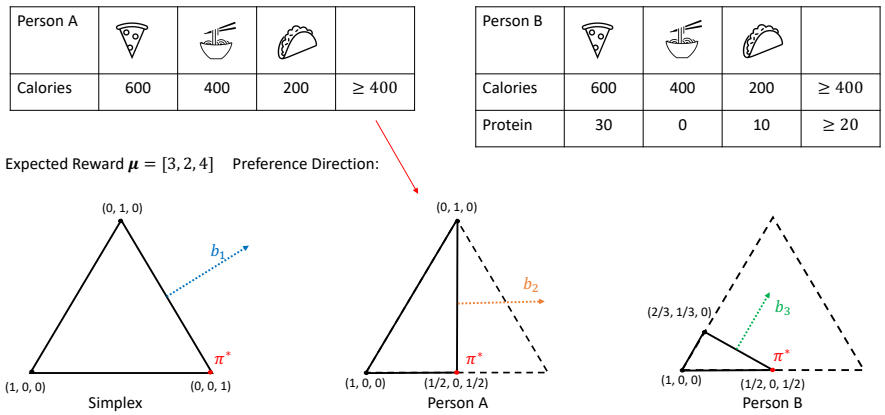


Figure 6.1: A Visual Representation of Example 1.1. Left figure with the full simplex represents the unconstrained problem. While the constraints of person A (middle) and person B (right) modify the problem to be harder and easier than the unconstrained one.

Example 1.1 (Optimal meal plan)**.** *Two people, A and B, are searching for a meal plan* $\boldsymbol{\pi}$ *that maximizes taste, i.e. expected reward* $\boldsymbol{\mu}^\top \boldsymbol{\pi}$*, 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, as illustrated in Figure 6.1, 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. In Figure 6.1, 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 6.2, the distance between* $\boldsymbol{\mu}$ *and* $\boldsymbol{b}_i$ *controls the hardness of the problem. For person A, the distance between* $\boldsymbol{b}_2$ *and* $\boldsymbol{\mu}$ *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 (ref. Fig. 6.2).*

As illustrated in Figure 6.1, a learner may need to search for a *stochastic policy* that allocates positive probabilities to multiple arms and this influences how an efficient learner should explore. Depending on the constraints, the learner's task

may become easier or harder, e.g. because the learner may need to explore several arms more extensively, or the constraints may remove several near-optimal policies, which makes the problem easier. These observations yield the following fundamental questions:

How do a specific set of constraints impact a pure exploration problem in terms of the minimum number samples required to identify the optimal policy?

Our Contributions. We define the problem of pure exploration in bandits with linear constraints and derive a corresponding lower bound on the sample complexity of any algorithm. We further derive an explicit lower bound for arms corresponding to Gaussian distributions, which shows that the hardness depends on the projection of $\boldsymbol{\mu}$ onto boundary of a normal cone, and that the lower bound diminishes with the increasing condition number of the constraints defining the optimal policy. Our results show that the lower bound can be thought of as a zero-sum game where the learner plays an exploration strategy and the adversary plays a constraint that is not active at the optimal policy. These insights allow us to modify the standard BAI algorithms, such as Track-and-Stop (Garivier and Kaufmann 2016) and the game-theoretic algorithm (Degenne, Koolen, and Ménard 2019), and extend them to the constraint setting. We prove that our proposed algorithms are optimal in the asymptotic regime for the pure exploration problem with known linear constraints. Finally, we empirically evaluate the algorithms, both on synthetic and realistic data.

1.1 Related work

Now, we review some works on policy learning, a classical problem in decision-making (Bechhofer 1958), that deal with known or learned constraints on decisions and/or constraint exploration due to safety, fairness, or other preferences.

Adapting To Known Constraints. Constraints are often used to ensure safety in reinforcement learning, online learning and control (Moldovan and Abbeel 2012; Gillulay and Tomlin 2011; Wan et al. 2022; Vaswani et al. 2022). In the bandit literature, some variants of the best-arm identification (BAI) problem impose constraints on the chosen arm, or on the exploration process. Wang, Wagenmaker, et al. (2022) and Camilleri et al. (2022) studies the setting with unknown linear rewards under known safety constraints but only allow single coordinate actions. Faizal and Nair (2022) consider BAI under fixed budget with known constraints on the arms. Their setting differs from ours in that *we look for a best "policy" over arms with linear constraints rather than a single best arm.*

Learning Unknown Constraints. Sui, Gotovos, et al. (2015) and Sui, Zhuang, et al. (2018) study online optimization of an unknown function f with constraints on f, but without formal analysis. In the bandit literature, constraints are mostly studied in the regret-minimization setting. (Moradipari et al. 2021) and (Pacchiano et al. 2021) consider regret minimization in linear bandits under linear constraints from Bayesian and Frequentist perspectives, respectively. Amani et al. (2019) study regret minimization in linear contextual bandits with unknown and unobserved linear constraints. Wang, Bai, et al. (2021) aims to minimize the fairness regret to ensure proportional exposure for each arm, which implies a known structure for the policies.

Unlike these works, we focus on the pure exploration setting. Lindner et al. (2022) considers constrained linear best-arm identification arm are vectors with *known* rewards and a single *unknown* constraint (representing preferences) on the actions.

Pure Exploration Algorithms. Our Constrained Track-and-Stop algorithm, (CTnS, Section 4), follows the Track-and-Stop TnS) meta-scheme proposed by Garivier and Kaufmann (2016). In TnS, one tracks an optimal allocation with respect to a lower bound and assumes that the current estimate is the true environment. This approach has been applied to various bandits, e.g., linear bandits (Jedra and Proutiere 2020), spectral bandits (Kocák and Garivier 2021), heavy-tailed bandits (Agrawal, Juneja, et al. 2020), bandits with multiple correct answers (Degenne and Koolen 2019), and latent bandits (Kinyanjui et al. 2023). The Constrained Game Explorer, (CGE, Section 4), follows the gamification approach to pure-exploration, which treats the lower bound as a zero-sum game between an allocation player and instance player. This approach was first introduced by Degenne, Koolen, and Ménard (2019), and later used for best-arm identification in linear bandits (Degenne, Ménard, et al. 2020) and combinatorial bandits (Degenne, Ménard, et al. 2020). In particular, CGE is an extension of the sampling rule of (Degenne, Koolen, and Ménard 2019) to the case of known linear constraints.

Transductive Linear Bandit. Another related setup is the transductive linear bandit (Fiez et al. 2019), where one set of arms, $\mathcal{A}$, are played during exploration while the goal is to detect the best arm in some other known set, $\mathcal{Z}$. This is related to our setting since we want to learn the best policy but only have access to arms. Hence, our model can be viewed as a natural special case of the transductive linear bandit where $\mathcal{A}$ is the standard basis and $\mathcal{Z}$ is the set of policies. However, the existing literature on transductive bandits does not study the impact of linear constraints that we explicitly study here and the resulting algorithms are different.

Bandits With Knapsacks. Our work is also related to the bandit with knapsack (Badanidiyuru et al. 2018; Agrawal and Devanur 2016; Immorlica et al. 2022). In this model, there are upper bounds on the total amount of resources a learner can consume while interacting with the bandit and each arm has its own resource consumption. The goal is to minimize the cumulative regret and the learner has to stop once the resources are depleted. This is different from our setting since we consider the problem of finding the best policy and not regret minimization. Our constraints are also not budget constraints but constraints in the policy space.

2 Problem formulation

We consider a multi-armed bandit problem with K arms that corresponds to reward distributions, $\{P_a\}_{a=1}^K$, with unknown means $\{\mu_a\}_{a=1}^K$ and support $\mathbb{R}$. At each time step t, a learner chooses to play one of the arms, $A_t \in [K]$, and observes an immediate reward R_t, drawn from the reward distribution P_{A_t}. The learner has access to a non-empty and compact set of feasible policies

$$\mathcal{F} \triangleq \{\boldsymbol{\pi} \in \Delta_{K-1} : B\boldsymbol{\pi} \leqslant c\}\,, \tag{2.1}$$

where Δ_{K-1} is the K-simplex and $B \in \mathbb{R}^{N\times K}$ and $\boldsymbol{c} \in \mathbb{R}^N$, are known parameters of the linear constraints. For the ease of the presentation, we absorb the simplex constraints in B and $\boldsymbol{c}$. Hereafter, these variables refer to both the simplex constraints, and the additional linear constraints of the problem. *The goal of the learner is to recommend, with probability at least* $1-\delta$*, the* unique *optimal policy* $\boldsymbol{\pi}^*_{\boldsymbol{\mu},\mathcal{F}}$ satisfying

$$\boldsymbol{\pi}^*_{\boldsymbol{\mu},\mathcal{F}} \triangleq \arg\max_{\boldsymbol{\pi}\in\mathcal{F}} \quad \boldsymbol{\mu}^\top \boldsymbol{\pi}. \tag{2.2}$$

When it is clear from the context, we denote $\boldsymbol{\pi}^*_{\boldsymbol{\mu},\mathcal{F}}$ as $\boldsymbol{\pi}^*$. We refer to such a learner as a *δ-PAC learner*. As $1-\delta$ quantifies the correctness of the learner, we also want it to be efficient, i.e. to detect the optimal policy *fast*. Let τ_δ denote the random *stopping time* at which the learner stops interacting with the bandit and makes a recommendation with confidence $1-\delta$. *We aim to design a δ-PAC learner that minimizes the expected stopping time $\mathbb{E}[\tau_\delta]$, a.k.a. sample complexity, needed to find the optimal policy.*

Depending on the application, a learner can abide by the constraints of Equation (2.1) in two ways:

- **Scenario 1: End-of-time constraint:** The learner does not have to take the constraints into account during exploration. Only the final recommended policy needs to satisfy the constraints.
- **Scenario 2: Anytime constraint:** The exploration policy needs to satisfy the constraints *in expectation* during exploration, i.e. the exploration policy $\boldsymbol{w}_t$ needs to satisfy $\boldsymbol{w}_t \in \mathcal{F}$.

For example, Scenario 1 arises while using a more sophisticated hardware to search for an optimal policy, that should satisfy some energy-constraints, before deploying it on a low-energy hardware. In contrast, Scenario 2 can be thought of as performing the search directly on the low-energy hardware. Now, we explicitly state the assumptions used in this study:

- **Assumption 1:** The reward of each arm $i \in [K]$ is distributed according to a sub-Gaussian single-parameter exponential family parameterized by its unknown mean μ_i.
- **Assumption 2:** The vector of arm means, $\boldsymbol{\mu}$, lies in a bounded domain $\mathcal{D} = [\mu_{\min}, \mu_{\max}]^K$.
- **Assumption 3:** The optimal solution $\boldsymbol{\pi}^*_{\boldsymbol{\mu},\mathcal{F}}$ to the linear program in Equation (2.2) is unique.

Assumptions 1 and 2 are standard in the literature (Degenne and Koolen 2019; Degenne, Ménard, et al. 2020). Assumption 3 is the analogue of assuming a unique best arm in the BAI problem, and it ensures that the optimum of Equation (2.2) is an extreme point. Hence, the optimal policy $\boldsymbol{\pi}^*_{\boldsymbol{\mu},\mathcal{F}}$ always corresponds to an extreme point in the polytope $\mathcal{F}$. In Appendix D, we discuss the relaxation to ϵ-good policies.

Notations. Let Π denote the set of feasible exploration policies. Thus, for Scenario 1, $\Pi = \Delta_{K-1}$, and $\Pi = \mathcal{F}$ for Scenario 2. We denote the KL-divergence between two single-parameter exponential family distributions with mean x and y as $\mathbb{KL}(x, y)$. Additionally, if the random variables are Bernoulli, we denote the KL-divergence as $\mathbf{kl}(x||y)$.

3 Lower bound

Lower bounds on the sample complexity of a δ-correct algorithm, i.e. $\mathbb{E}[\tau_\delta]$, is a driving force in designing good algorithms in the BAI literature (Garivier and Kaufmann 2016; Degenne and Koolen 2019; Agrawal, Juneja, et al. 2020).

Given a problem instance $\boldsymbol{\mu}$, a learner needs to collect enough information about the problem to be able to rule out all alternative instances, $\boldsymbol{\lambda}$, for which we have $\max_{\boldsymbol{\pi}\in\mathcal{F}} \boldsymbol{\lambda}^\top \boldsymbol{\pi} > \boldsymbol{\lambda}^\top \boldsymbol{\pi}^*$ with confidence at least $1-\delta$. We refer to this set of instances as the Alt-set and denote it as

$$\Lambda_{\mathcal{F}}(\boldsymbol{\mu}) \triangleq \{\boldsymbol{\lambda} \in \mathcal{D} : \max_{\boldsymbol{\pi}\in\mathcal{F}} \boldsymbol{\lambda}^\top \boldsymbol{\pi} > \boldsymbol{\lambda}^\top \boldsymbol{\pi}^*\}. \tag{3.1}$$

Garivier and Kaufmann (2016) introduced general techniques for deriving lower bounds on the sample complexity of any δ-PAC learner, which depends on the the distance from $\boldsymbol{\mu}$ to the closest $\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})$ in an information-theoretic sense.

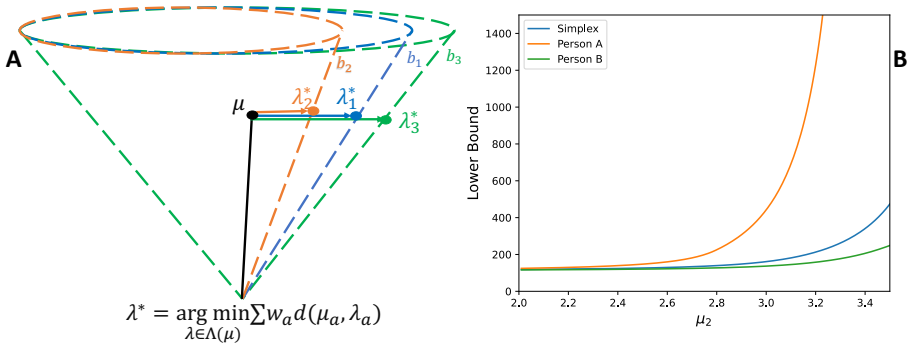


Figure 6.2: Computing the $\boldsymbol{\lambda}$ satisfying Equation 3.4, i.e. the *most confusing instance*, can be viewed as an information-theoretic projection onto the boundary of the normal cone spanned by the active constraints at $\boldsymbol{\pi}_{\boldsymbol{\mu}}$. In A) we see the different normal cones for the three different examples in Figure 6.1. In B) we have fixed μ_1 and μ_3, as in Figure 6.1, 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 Corollary 3.4.

We extend these general proof techniques and show that the expected stopping time of any δ-PAC algorithm ϕ for BAI with linear constraints satisfies

$$\mathbb{E}_{\boldsymbol{\mu},\phi}[\tau_\delta] \geqslant T_{\mathcal{F}}(\boldsymbol{\mu})\mathbf{kl}(\delta||1-\delta). \tag{3.2}$$

where $T_{\mathcal{F}}(\boldsymbol{\mu})$ is the *characteristic time*, defined as

$$T_{\mathcal{F}}^{-1}(\boldsymbol{\mu}) = \sup_{\boldsymbol{w}\in\Pi} \inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a \mathbb{KL}(\mu_a, \lambda_a). \tag{3.3}$$

The supremum in Equation (3.3) hints towards the existence of some optimal exploration policy $\boldsymbol{w}$, which any optimal algorithm should try to track. This is exactly the idea behind the Track-and-Stop meta-scheme (Garivier and Kaufmann 2016) (details in Section 4). In order to design algorithms achieving the lower bound in Equation (3.2), we need to solve the optimization problem in Equation (3.3). This requires a more explicit characterization of $\Lambda_{\mathcal{F}}(\mu)$, continuity properties of the function $D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F}) \triangleq \inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a \mathbb{KL}(\mu_a, \lambda_a)$, and the set of optimal allocations $w^*(\boldsymbol{\mu})$.

To derive an explicit expression for $\Lambda_{\mathcal{F}}(\mu)$, let M be the number of active constraints for $\boldsymbol{\pi}^*$, $B_{\boldsymbol{\pi}^*} \in \mathbb{R}^{M\times K}$ be a submatrix of B consisting of all these active constraints, and $\boldsymbol{c}_{\boldsymbol{\pi}^*} \in \mathbb{R}^M$ the corresponding bounds in $\boldsymbol{c}$. Hence, there exists *at least* K linearly independent rows in $B_{\boldsymbol{\pi}^*}$, i.e. a matrix $\hat{B}_{\boldsymbol{\pi}^*} \in \mathbb{R}^{K\times K}$ and vector $\hat{\boldsymbol{c}}_{\boldsymbol{\pi}^*} \in \mathbb{R}^K$, such that $\boldsymbol{\pi}^* = \hat{B}_{\boldsymbol{\pi}^*}^{-1} \hat{\boldsymbol{c}}_{\boldsymbol{\pi}^*}$. Since our objective (Equation (2.2)) is a linear program, we can leverage the optimality condition stating that $\boldsymbol{\mu}$ must be in the normal cone of the optimal solution (Boyd and Vandenberghe 2004). Hence, we express the Alt-set as $\Lambda_{\mathcal{F}}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} : \boldsymbol{\lambda} \notin \mathcal{N}(\boldsymbol{\pi}^*)\}$. Here, $\mathcal{N}(\boldsymbol{\pi}^*) := \left\{\boldsymbol{\lambda} : \boldsymbol{\lambda} = B_{\boldsymbol{\pi}^*}^{\top} \boldsymbol{v}, \boldsymbol{v} \in \mathbb{R}^M_{\geqslant 0}\right\}$ is the normal cone spanned by the active constraints for $\boldsymbol{\pi}^*$.

Further, we say that $\boldsymbol{\pi}'$ is a neighbor of $\boldsymbol{\pi}^*$ if it is an extreme point in $\mathcal{F}$ and shares $K-1$ active constraints with $\boldsymbol{\pi}^*$. We denote the set of all neighbors of $\boldsymbol{\pi}^*$ as $\mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)$. Hence, we can decompose the Alt-set into a union of a finite number of half-spaces $\Lambda_{\mathcal{F}}(\boldsymbol{\mu}) = \bigcup_{\boldsymbol{\pi}'\in\mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \left\{\boldsymbol{\lambda} : \boldsymbol{\lambda}^{\top}(\boldsymbol{\pi}^* - \boldsymbol{\pi}') < 0\right\}$. This formulation implies that if $\boldsymbol{\pi}^*$ is not an optimal policy for the instance $\boldsymbol{\lambda}$, there must exist an direction for the simplex algorithm to follow to increase the expected reward, i.e. $\exists \boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)$: $\boldsymbol{\lambda}^{\top}(\boldsymbol{\pi}^* - \boldsymbol{\pi}') < 0$. This formulation of Alt-sets lead us to the observation that the most confusing instances in the Alt-set w.r.t. $\boldsymbol{\mu}$ lay on the boundary of the normal cone.

Specifically, Lemma 3.1 shows that the function $D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F})$ is a weighted projection onto the plane $\boldsymbol{\lambda}^{\top}(\boldsymbol{\pi}' - \boldsymbol{\pi}^*) = 0$ for some $\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)$, as shown in Figure 6.2.

Lemma 3.1 (Projection Lemma)**.** *For any $\boldsymbol{w} \in \Pi$ and $\boldsymbol{\mu}$ it holds that*

$$D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F}) = \min_{\boldsymbol{\pi}'\in\mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \min_{\boldsymbol{\lambda}:\boldsymbol{\lambda}^{\top}(\boldsymbol{\pi}^* - \boldsymbol{\pi}')=0} \sum_{a=1}^{K} w_a \mathbb{KL}(\mu_a, \lambda_a) \tag{3.4}$$

To compute $D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F})$ from Equation (3.4), we need to have access to the true instance $\boldsymbol{\mu}$, which we do not have in reality. Rather, we sequentially obtain samples from the arms yielding an estimate $\hat{\boldsymbol{\mu}}_t$. Thus, we need $D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F})$ and $\boldsymbol{w}^*(\boldsymbol{\mu})$ to satisfy continuity properties (Theorem 3.2) w.r.t $\boldsymbol{\mu}$, that ensures as the estimates $\hat{\boldsymbol{\mu}}_t$ converge to $\boldsymbol{\mu}$, $D(\boldsymbol{w}, \hat{\boldsymbol{\mu}}_t, \mathcal{F}) \to D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F})$ and our empirical distribution of plays gets closer to some $\boldsymbol{w} \in w^*(\mu)$.

Theorem 3.2. *Following properties are true for all $\boldsymbol{\mu}$ and $\mathcal{F} = \{\boldsymbol{\pi} \in \Delta_{K-1} : B\boldsymbol{\pi} \leqslant c\}$ such that the problem $\max_{\pi \in \mathcal{F}} \boldsymbol{\mu}^\top \boldsymbol{\pi}$ has a unique solution.*

- *The function $(\boldsymbol{w}, \boldsymbol{\mu}) \mapsto D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F})$ is continuous.*
- *The function $\boldsymbol{\mu} \mapsto T_{\mathcal{F}}(\boldsymbol{\mu})$ is continuous.*
- *The set-valued function $\boldsymbol{\mu} \mapsto w^*(\boldsymbol{\mu})$ is upper hemicontinuous (definition in Appendix G).*
- *The set $w^*(\boldsymbol{\mu})$ is convex.*

3.1 Lower bound for Gaussian distributions

To gain further insights on how the constraints alter the lower bound in Equation (3.2), we consider the special case where all arms are Gaussian distributions with equal variance σ^2. This leads us to a close-form of the projection in Lemma 3.1 as in Theorem 3.3.

Theorem 3.3. *If the arms follow Gaussian distributions with identical variance σ^2 and $w_a > 0 \ \forall a$, we have that the projection $\min_{\boldsymbol{\lambda} \in \mathcal{D}: \boldsymbol{\lambda}^\top (\boldsymbol{\pi}^* - \boldsymbol{\pi}') \leqslant 0} \sum_{a=1}^K w_a \mathbb{KL}(\mu_a, \lambda_a)$ for any $\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)$ is satisfied by $\lambda_{a,\boldsymbol{\pi}'} = \mu_a - \gamma \frac{(\boldsymbol{\pi}^* - \boldsymbol{\pi}')_a}{w_a}$, for $\gamma = \frac{\boldsymbol{\mu}^\top (\boldsymbol{\pi}^* - \boldsymbol{\pi}')}{\sum_a \frac{(\boldsymbol{\pi}^* - \boldsymbol{\pi}')^2}{w_a}}$, and the characteristic time is*

$$
\begin{aligned}
T_{\mathcal{F}}(\boldsymbol{\mu})^{-1} &= \max_{\boldsymbol{w} \in \Pi} \min_{\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \frac{1}{2\sigma^2} \frac{\left(\boldsymbol{\mu}^\top (\boldsymbol{\pi}^* - \boldsymbol{\pi}')\right)^2}{\sum_a \frac{1}{w_a} (\boldsymbol{\pi}^* - \boldsymbol{\pi}')_a^2} \\
&= \max_{\boldsymbol{w} \in \Pi} \min_{\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \frac{1}{2\sigma^2} \frac{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^\top}}{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/w_a)}}
\end{aligned}
$$

Here, $\mathrm{Diag}(1/w_a)$ is a diagonal matrix with a-th entry of the diagonal as $1/w_a$.

In the classical BAI setting, i.e. we only have simplex constraints, the expressions in Theorem 3.3 reduces to the BAI results of Kaufmann, Cappé, et al. (2016), see Appendix B for a derivation. From Theorem 3.3, we further derive a lower and an upper bound on the characteristic time. Let us define $d_{\boldsymbol{\pi}'} \triangleq \min_{\boldsymbol{\lambda}: \boldsymbol{\lambda}^\top (\boldsymbol{\pi}^* - \boldsymbol{\pi}') = 0} ||\boldsymbol{\mu} - \boldsymbol{\lambda}||_2$ and note that this is the distance between $\boldsymbol{\mu}$ and the hyperplane $\boldsymbol{\pi}^* - \boldsymbol{\pi}' = 0$, see Figure 6.2 for illustration.

Corollary 3.4. *The characteristic time $T_{\mathcal{F}}(\boldsymbol{\mu})$ satisfies the following bounds:*

$$
\min_{\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \frac{2\sigma^2}{d^2_{\boldsymbol{\pi}'}} \leqslant T_{\mathcal{F}}(\boldsymbol{\mu}) \leqslant \min_{\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \frac{2\sigma^2 K}{d^2_{\boldsymbol{\pi}'}}. \tag{3.5}
$$

Corollary 3.4 implies a lower bound of

$$
\mathbb{E}[\tau] \geqslant \min_{\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \frac{2\sigma^2}{d^2_{\boldsymbol{\pi}'}} \mathbf{kl}(\delta || 1 - \delta)
$$

Impact Of Constraints: Geometric View. We first observe that, since the distance-to-projection $d_{\boldsymbol{\pi}'} = \frac{\boldsymbol{\mu}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}')}{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|_2}$, the problem becomes easier when the direction of the reward vector $\boldsymbol{\mu}$ is aligned with the deviation in policy $\boldsymbol{\pi}^* - \boldsymbol{\pi}'$. Especially, if we only consider deterministic policies, i.e. BAI problem, $d_{\boldsymbol{\pi}'} = \mu_1 - \mu_a = \Delta_a$ where μ_1 is the best arm, a is the arm played by $\boldsymbol{\pi}'$ and we retrieve the lower bound of Kaufmann, Cappé, et al. (2016).

Impact Of Constraints: Constrained Optimization View. We relate the lower bound more explicitly to the constraint matrix B by using the fact that any neighbor $\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)$ can be reached from $\boldsymbol{\pi}^*$ via an 1-rank update on a matrix $\hat{B}_{\boldsymbol{\pi}^*} \in \mathbb{R}^{K\times K}$ consisting of K active constraints at $\boldsymbol{\pi}^*$ that are linearly independent. Thus, we only need to change one row in $\hat{B}_{\boldsymbol{\pi}^*}$ and one element in the corresponding $\hat{\boldsymbol{c}}_{\boldsymbol{\pi}^*}$ to get B' and $\boldsymbol{c}'$ such that $\boldsymbol{\pi}' = B'^{-1}\boldsymbol{c}'$. This results in the lower bound on the sample complexity presented in Corollary 3.5.

Corollary 3.5. *For any $\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)$, let $\hat{B}_{\boldsymbol{\pi}^*} \in \mathbb{R}^{K\times K}$ be a set of active and linearly independent constraints at $\boldsymbol{\pi}^*$ such that the active constraints at $\boldsymbol{\pi}'$ can be achieved by a one-rank update on $\hat{B}_{\boldsymbol{\pi}^*}$. Let r' be the row in $\hat{B}_{\boldsymbol{\pi}^*}$ that is changed during this one-rank update.*

Part (a): Let $\boldsymbol{\Delta} \in \mathbb{R}^K$ denote the vector of the sub-optimality gaps, i.e. $\Delta_a = \mu_1 - \mu_a$, of each arm, then

$$T_{\mathcal{F}}(\boldsymbol{\mu})^{-1} = \max_{\boldsymbol{w}\in\Pi} \min_{\boldsymbol{\pi}'\in\mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \frac{1}{2\sigma^2} \frac{\left(\boldsymbol{\Delta}^\top \hat{B}^{-1}_{\boldsymbol{\pi}^*} \boldsymbol{e}_{r'}\right)^2}{\|\hat{B}^{-1}_{\boldsymbol{\pi}^*} \boldsymbol{e}_{r'}\|^2_{\mathrm{Diag}(1/w_a)}} \tag{3.6}$$

Part (b): Let κ^2 be the condition number of a matrix $\hat{B}_{\boldsymbol{\pi}^*} \in \mathbb{R}^{K\times K}$ consisting of K linearly independent active constraints at $\boldsymbol{\pi}^*$, then the sample complexity of any δ-PAC learner is lower bounded as

$$\mathbb{E}[\tau] = \Omega\left(\frac{H}{\kappa^2}\mathbf{kl}(\delta||1-\delta)\right) \tag{3.7}$$

with $H = \frac{2\sigma^2}{\sum_{a\neq a^*} \Delta_a^2}$.

Corollary 3.5 relates constraints, arm sub-optimality, and sample complexity. Equation (3.6) links sample complexity to perturbations of the optimal policy. Naturally, if a large perturbation of the optimal policy is only slightly sub-optimal, the sample complexity will be large. In contrast, if a small pertubation is bound to cause the resulting policy to be highly sub-optimal it is easier to detect the optimal policy. Equation (3.6) also reinterprets the lower bound as a zero-sum game where the agent plays an allocation and an adversary switches an active constraint at $\boldsymbol{\pi}^*$ to a non-active one.

Equation (3.7) provides a looser bound based on a *suboptimality gap based complexity measure* H, and *the condition number*, κ^2 of the active-constraint matrix, which measures sensitivity of the optimal policy to perturbations. A high κ^2 implies that small perturbations of the optimal policy will cause a large change of the slack corresponding to the active constraints, making exploration easier. A low κ^2 means policy perturbations have a smaller impact on the slack making neighboring policies less distinguishable from the optimal one.

Algorithm 6.1 Constrained Track-and-Stop (CTnS)

Require: Confidence level δ, constraints (B, c), exploration set Π
Play each arm once.
while $c(t,\delta) > D(\boldsymbol{N}/t, \hat{\boldsymbol{\mu}}_t, \mathcal{F})$ **do** *Weighted projection via Lemma 3.1*
Compute $\boldsymbol{w}_t^* \in \arg\max_{\boldsymbol{w}\in\Pi} D(\boldsymbol{w}, \hat{\boldsymbol{\mu}}_t, \mathcal{F})$ *Solve for optimal $\boldsymbol{w}$ w.r.t. the constraints*
Play $A_t \in \arg\min_a N_{a,t} - \sum_{s=1}^t w^*_{a,s,\epsilon_s}$ and observe reward R_t
end while
Recommend $\boldsymbol{\pi}^*_{\hat{\boldsymbol{\mu}}_t} = \arg\max_{\boldsymbol{\pi}\in\mathcal{F}} \hat{\boldsymbol{\mu}}_t^\top \boldsymbol{\pi}$

4 Algorithms

In this section, we focus on extending the classical pure exploration algorithms to the setting of pure exploration with linear constraints.

Algorithm Design. We begin by observing that any pure exploration algorithm consists of three components: *A Stopping Rule*, *a recommendation rule*, and *a sampling strategy*. The stopping rule consists of a condition deciding when to halt sampling further. The recommendation rule decides what policy to recommend as the optimal policy. The sampling rule decides which arm to sample next given the history of arms sampled and intermediate policies computed.

Component 1: Chernoff's stopping rule with constraints. As a stopping rule, we extend the Chernoff's stopping rule (Garivier and Kaufmann 2016). We first introduce the *confidence set* $\mathcal{C}_t(\delta) := \left\{\boldsymbol{\lambda} : \sum_{a=1}^K N_{a,t}\mathbb{KL}(\hat{\mu}_{a,t}, \lambda_a) \leqslant c(t,\delta)\right\}$, where $c(t,\delta)$ is a threshold defined in Lemma 4.1.

Lemma 4.1 (Garivier and Kaufmann (2016))**.** *For any $\alpha > 1$ there exists a constant $C(\alpha, K)$ such that for $c(t,\delta) = \log \frac{t^\alpha C(\alpha,K)}{\delta}$ we have for any $t \in \mathbb{N}$ $P\left(\boldsymbol{\mu} \notin \mathcal{C}_t(\delta)\right) \leqslant \delta$.*

Lemma 4.1 implies that Chernoff's stopping rule is a δ-PAC stopping rule, and we stop when

$$\inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\hat{\boldsymbol{\mu}}_t)} \sum_{a=1}^K N_{a,t}\mathbb{KL}(\hat{\mu}_{a,t}, \lambda_a) > c(t,\delta). \tag{4.1}$$

This means that the confidence set is a subset of the normal cone spanned by the active constraints at $\boldsymbol{\pi}^*_{\hat{\boldsymbol{\mu}}_t}$. The details of the constant in Lemma 4.1 are deferred to Appendix C. Note that one can also derive a stopping rule via the concentration results of Kaufmann and Koolen (2021).

Component 2: Recommendation rule. We recommend the solution of the linear programming (Equation (2.2)) with the empirical means of the arms at the stopping time, $\boldsymbol{\pi}^*_{\hat{\boldsymbol{\mu}}_t} = \arg\max_{\boldsymbol{\pi}\in\mathcal{F}} \hat{\boldsymbol{\mu}}_t^\top \boldsymbol{\pi}$. Since the empirical means might not always be within the pre-specified range $\mathcal{D}$, we let $\hat{\boldsymbol{\mu}}_t$ denote the Euclidean projection of the empirical means onto $\mathcal{D}$.

Component 3a: CTnS.

Algorithm 6.2 Constrained Game Explorer (CGE)

Require: Confidence level δ, constraints (B, c), exploration set Π
  while $c(t,\delta) > D(\boldsymbol{N}/t, \hat{\boldsymbol{\mu}}_t, \mathcal{F})$ **do** *Weighted projection via Lemma 3.1*
    Get allocation $\boldsymbol{w}_t$ from regret minimizer *Running Adagrad over* Π
    Compute best-response $\boldsymbol{\lambda}_t$ w.r.t. $\boldsymbol{w}_t$ and $\hat{\boldsymbol{\mu}}_t$ *Weighted projection via Lemma 3.1*
    Compute confidence intervals $\forall a\ [\alpha_{t,a}, \beta_{t,a}] = \{\xi : N_{a,t}\mathbb{KL}(\hat{\mu}_{a,t}, \xi) \leqslant f(t)\}$
    $\forall a\ U_t^a := \max\left\{\frac{f(t)}{N_{a,t}}, \max_{\xi\in\{\alpha_t^a, \beta_t^a\}} \mathbb{KL}(\xi, \lambda_{a,t})\right\}$
    Update AdaGrad with $l(w_t) = \sum_{a=1}^K w_a U_{a,t}$
    Play $A_t \in \arg\min_a N_{a,t} - \sum_{s=1}^t w^*_{a,s,\epsilon_s}$ and observe reward R_t
  end while
  Recommend $\boldsymbol{\pi}^*_{\hat{\boldsymbol{\mu}}_t} = \arg\max_{\boldsymbol{\pi}\in\mathcal{F}} \hat{\boldsymbol{\mu}}_t^\top \boldsymbol{\pi}$

First, we present our Constrained Track-and-Stop Algorithm (CTnS, Algorithm 6.1), which is an adaptation of the Track-and-Stop (TnS) framework (Garivier and Kaufmann 2016) to the linear constraint setting with aforementioned stopping and recommendation rules. In Algorithm 6.1, we highlight, in red, the computations that we modify from the original schematic to account for the linear constraints. The algorithm starts by playing each arm once. Then, until the stopping rule in Equation (4.1) fires, it performs *C-tracking* (Garivier and Kaufmann 2016). This means that we perform a $\max - \min$ oracle call (Line 3), and solve the problem in Equation (3.3) w.r.t our current estimate of the means $\hat{\boldsymbol{\mu}}_t$ to get an optimal allocation $\boldsymbol{w}_t^*$. This step leverage our novel projection result in Lemma 3.1. We track the optimal allocation via $A_t \in \arg\min_a N_{a,t} - \sum_{s=1}^t w^*_{a,s,\epsilon_s}$, where w^*_{a,t,ϵ_t} is the projection of $\boldsymbol{w}_t^*$ onto $\Pi \bigcap \{\boldsymbol{w} : w_a > \epsilon_t \forall a\}$, and $\epsilon_t = \frac{1}{2\sqrt{K^2+t}}$. Note that $\frac{1}{t}\sum_{s=1}^t w^*_{a,s,\epsilon_s} \in \Pi$ due to the convexity of the set of feasible exploration policies/allocations.

Theorem 4.2 (Upper bound for CTnS)**.** *For any $\alpha > 1$ and $c(t,\delta)$ be defined as in Lemma 4.1, we have that the expected stopping time of CTnS satisfies*

$$\lim_{\delta\to 0} \frac{\mathbb{E}[\tau]}{\log\frac{1}{\delta}} \leqslant T_{\mathcal{F}}(\boldsymbol{\mu}),\ \forall \boldsymbol{\mu} \in \mathcal{D}.$$

The proof of Theorem 4.2 can be found in Appendix C.2 and follows the same structure as the sample complexity proof of the original TnS in Garivier and Kaufmann (2016). However, the optimal allocation does not necessarily have to be unique. rather, we use the upper hemicontinuity and convexity of $w^*(\boldsymbol{\mu})$, while modifying the tracking lemma originally used by Garivier and Kaufmann (2016) with the tracking result of Degenne and Koolen (2019). This change allows to track a set of optimal solutions in absence of a unique optimum.

Component 3b: CGE. Track-and-Stop algorithms, like CTnS, tend to be computationally inefficient for larger problems since they requires a $\max - \min$ call at each iteration. To mitigate this, we adopt the approach of Degenne, Koolen, and Ménard (2019), and treat the optimization problem in Equation (3.3) as a two player zero-sum game. This results in the Constrained Game Explorer (CGE), in Algorithm 6.2. Instead of solving for an optimal $\boldsymbol{w}_t$ at each t, as in CTnS, we play one game

between an allocation player, who plays $\boldsymbol{w}$ to maximize $\sum_{a=1}^{K} w_a \mathbb{KL}(\hat{\mu}_{a,t}, \lambda_a)$, and an instance player, who plays the confusing instance $\boldsymbol{\lambda}$ to minimize $\sum_{a=1}^{K} w_a \mathbb{KL}(\hat{\mu}_{a,t}, \lambda_a)$. We deploy an instance of AdaGrad (Duchi et al. 2011) as the allocation player is taken to be, which enjoys sub-linear regret over any bounded domain when losses are convex, and the instance player is taken to be a best-response w.r.t. the allocation $\boldsymbol{w}_t$. The best-response is computed via Lemma 3.1. The loss provided to AdaGrad at each time step is $\sum_{a=1}^{K} w_{a,t} U_{a,t}$, where $U_{a,t}$ induces optimism as $U_{a,t} \triangleq \max_{\xi \in \{\alpha_{a,t}, \beta_{a,t}\}} N_{a,t} \mathbb{KL}(\xi, \lambda_{a,t})$. Here, $(\alpha_{a,t}, \beta_{a,t})$ are the endpoints of the confidence interval around $\hat{\mu}_{a,t}$, i.e. $[\alpha_{t,a}, \beta_{t,a}] = \{\xi : N_{a,t} \mathbb{KL}(\mu_{a,t}, \xi) < f(t)\}$, and $f(t) = 3 \log t + \log \log t$. We apply the same tracking as in CTnS.

Theorem 4.3 (Upper bound for CGE)**.** *The expected sample complexity of CGE satisfies*

$$\mathbb{E}[\tau] \leqslant T_0(\delta) + CK,$$

where

$$T_0(\delta) := \max \left\{ t \in \mathbb{N} : t \leqslant T_{\mathcal{F}}(\boldsymbol{\mu}) c(t, \delta) + O(\sqrt{tQ}) + O(\sqrt{t \log t}) \right\}.$$

C_μ is problem-dependent constant, C is a universal constant and Q is an upper bound on the losses provided to Adagrad.

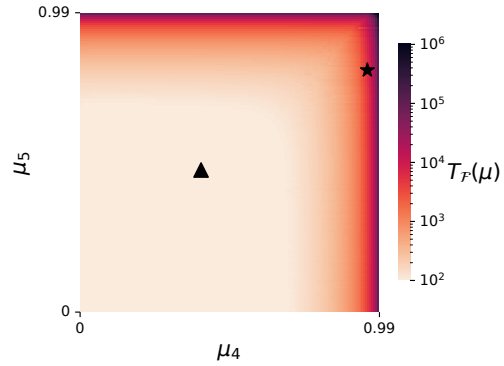
The full proof of Theorem 4.3 can be found in Appendix C.3. We simply follow the steps of the proof of Theorem 2 in Degenne, Koolen, and Ménard (2019) and apply specifics of our setting when applicable.

Theorem 4.2 and 4.3 show that CTnS and CGE are asymptotically optimally, i.e. upper bound on their sample complexities match the lower bound of constrained pure exploration for small enough δ.

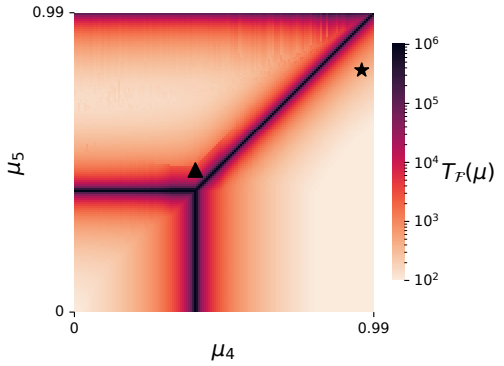
5 Experimental analysis

We evaluate our algorithms using the threshold $c(t, \delta) = \log \frac{1 + \log \log t}{\delta}$, which is commonly done in the literature (Garivier and Kaufmann 2016), and we set $f(t) = \log t$ in CGE. As benchmarks we will use the lower bound, Equation 3.2, as well as a learner that samples from the optimal allocation, given by the lower bound, at all time steps. We also consider a learner that draws arms from the uniform distribution and in scenarios where the uniform distribution is not in the set of feasible exploration policies we project it onto the set and sample from the resulting distribution.

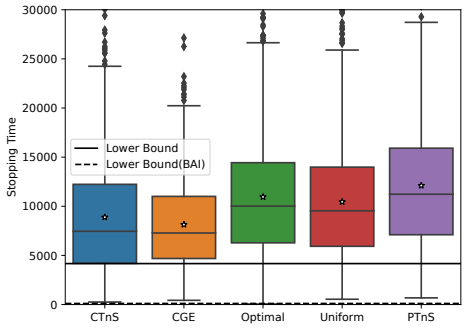
In addition, we consider a naïve adaptation of Track-and-Stop (Kaufmann, Cappé, et al. 2016), which we call the *Projected-Track-and-Stop* (PTnS). The *PTnS algorithm computes the allocation as if it was solving the classical BAI problem and projects the allocation back to the feasible set when necessary.* Comparing CGE and CTnS with PTnS demonstrates (a) the importance of tracking the constrained lower bound to design an efficient algorithm, and, (b) the desired efficiency cannot be achieved



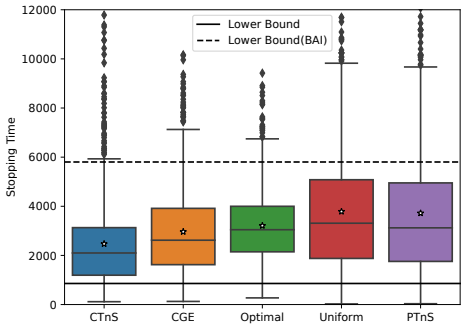
(a) Characteristic time of the BAI problem as we vary μ_4 and μ_5.



(b) Characteristic time of the constraint pure-exploration problem as we vary μ_4 and μ_5



(c) Results (for 1000 random seeds) on the instance highlighted as a triangle in Figure 6.3a and Figure 6.3b, $\boldsymbol{\mu} = (1, 0.5, 0.4, 0.4, 0.5)$, with constraints and $\delta = 0.1$.

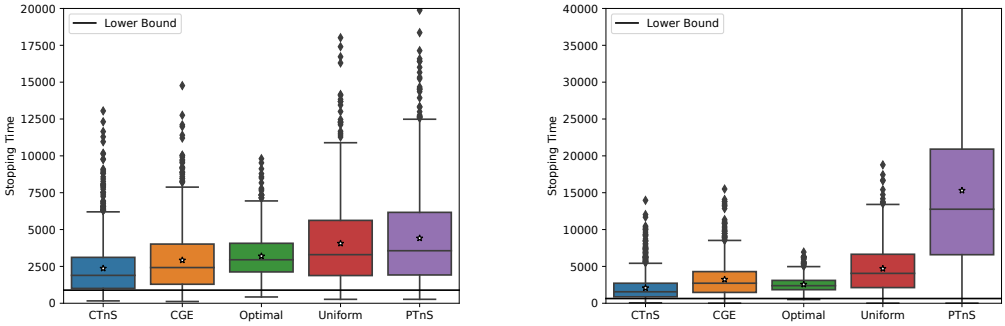


(d) Results (for 1000 random seeds) on the instance highlighted as a star in Figure 6.3a and Figure 6.3b, $\boldsymbol{\mu} = (1, 0.5, 0.4, 0.95, 0.8)$, with constraints and $\delta = 0.1$.

Figure 6.3: Figure 6.3a and 6.3b illustrate the hardness of the problem, i.e. the Characteristic time, changes in the 5 arm instance $\boldsymbol{\mu} = (1.0, 0.5, 0.4, \mu_4, \mu_5)$ as we vary μ_4 and μ_5. Figure 6.3a corresponds to the hardness in the BAI while Figure 6.3b is the constraint setting with constraints $\boldsymbol{\pi}_1 + \boldsymbol{\pi}_2 \leqslant 0.5$ and $\boldsymbol{\pi}_3 + \boldsymbol{\pi}_4 \leqslant 0.5$. We clip the characteristic time at 10^6 for visual purposes.

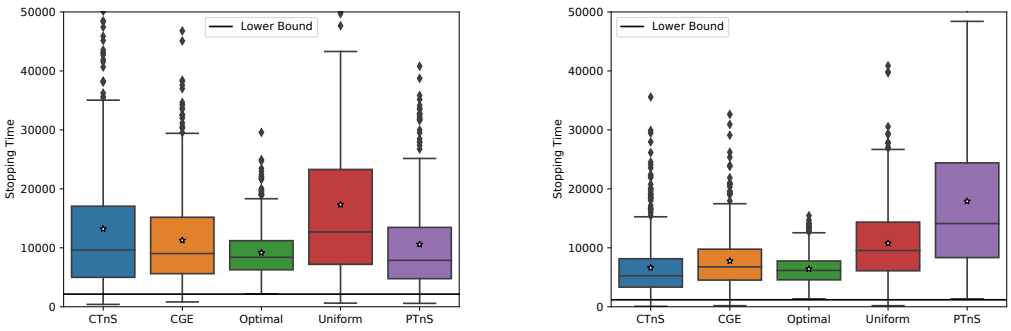
just by tracking the unconstrained lower bound and projecting the corresponding allocation policy to the constrained set. Appendix E contains additional experiments.

Observation 1: Constraints alter the hardness ff the problem. In Figures 6.3a and 6.3b we illustrate how the hardness of a bandit instance $\boldsymbol{\mu}$ may differ once we introduce constraints, assuming anytime constraints. We consider the instance $\boldsymbol{\mu} = (1.0, 0.5, 0.4, \mu_4, \mu_5)$ and plot how the characteristic time $T_{\mathcal{F}}(\boldsymbol{\mu})$ changes as we vary μ_4 and μ_5, Figure 6.3a corresponds to the classical BAI, i.e. no constraints, and in Figure 6.3b we have introduced the two constraints $\boldsymbol{\pi}_1 + \boldsymbol{\pi}_2 \leqslant 0.5$ and $\boldsymbol{\pi}_3 + \boldsymbol{\pi}_4 \leqslant 0.5$. We have highlighted two instances, one where the BAI problem is easy but the constraint problem is hard (black triangle) and one where the reverse is true (black star). We run the algorithms on these two instances in Figure 6.3c and 6.3d, assuming anytime constraints, and observe that both algorithms operate close to the lower bound and outperforms the uniform allocation strategy. We also observe that the algorithms perform equally or better than the optimal learner, this is an interesting phenomena and have been observed earlier in other pure exploration



(a) Anytime Constraints (b) End-of-Time Constraints

Figure 6.4: Problem instance with 8 Gaussian arms with $\sigma = 1$. The arm means are $\mu = [1.0, 0.7, 0.3, 0.0, -0.5, -1.0, -2.0, -3.0]$ and we have one constraint $7\pi_1 + 7\pi_2 + \pi_3 \leqslant 0.5$. The optimal policy is $\pi_3 = \pi_4 = 0.5$. Results for $\delta = 0.1$ and 1000 random seeds.



(a) Anytime Constraints (b) End-of-Time Constraints

Figure 6.5: Experiments on IMDB dataset with 12 movies and $\delta = 0.1$. Each experiment was performed over 500 random seeds.

scenarios (Degenne, Koolen, and Ménard 2019). The PTnS does not account for the constraints, as well as CTnS and CGE, and has a sample complexity on par with uniform sampling.

Observation 2: Naïve projection cause high sample complexity. In Figure 6.4, we consider an eight-armed bandit with Gaussian reward distributions. We observe that PTnS performs the worst on this instance, specially in the end-of-time setting where it is outperformed by uniform sampling. This because in a BAI problem with the same $\boldsymbol{\mu}$ the hardness of the problem lies separating arm 1 and 2 but this doesn't have to be the case in the constraint bandit. The sub-optimality of PTnS in Figure 6.4a, the anytime scenario, illustrates that naïvely projecting the allocation onto the feasible set won't account for the constraints in a meaningful way. In Appendix F we further discuss these examples and compute the optimal allocations and the allocations PTnS converge to for each scenario.

IMDB movie recommendation environment. We construct a semi-synthetic task based on the widely used IMDB 50K Movie Dataset (Maas et al. 2011) which

contains metadata on $k_0 = 50000$ movies including association with one or more of $d = 23$ genres, as indicated by a binary matrix $X \in \{0,1\}^{m \times d}$. In our setting, actions correspond to recommending one out of a subset of $k \leqslant k_0$ movies. To create reward distributions for each movie, we simulate a population of $n_u = 600$ users, each assigned $n_f = 5$ favorite genres f_i with weights $w_{if_i} = [20, 10, 5, 2, 2]$ and let $w_{ia} = 0$ for $a \notin f_i$. A score s_{ia} for user i and movie a is created as follows, $s_{ij} = \text{clip}(\lfloor \tilde{s}_{ia} / \sum_{a \in f_i} w_{ia} \cdot \sigma_0 + \sigma_1 \epsilon_{ia} \rceil \, ; 1, 5)$ where $\tilde{s}_{i\cdot} = w_i X^\top + w_0$, $\epsilon_{ia} \sim U(0,1)$, $\sigma_0 = 5, \sigma_1 = 3$, and $\lfloor x \rceil$ indicates rounding of x to the nearest integer. We construct the bandit environment by letting each movie a be represented by an arm with reward $R_a \sim \mathcal{N}(\hat{\mu}_{s\cdot a}, \hat{\sigma}^2_{s\cdot a})$ determined by the mean and standard deviation of user reviews for the movie. We sample a subset of movies and search for the optimal policy that allocates at most 0.3 to action movies, at least 0.3 to drama movies and at least 0.3 on family movies. Note that one movie might belong to more than one category. We present the result in Figure 6.5 for both the anytime scenario and the end of time scenario. We observe that CTnS and CGE outperform the uniform allocation strategy, which has a very high variance. We also observe a bigger difference between the algorithms under end of time constraints, this is reasonable since the set of plausible exploration policies is larger for that scenario. If the set of exploration policies is limited, there is little room for an algorithm to be adaptive. This is also captured in the fact that the lower bound for anytime constraints is always higher or equal to the bound for end-of-time constraints.

6 Conclusions and future directions

In this paper, we study the problem of pure exploration in bandits with linear constraints. We provide a generic lower bound for this setting that depends on an information-theoretic projection onto the boundary of the normal cone spanned by the active constraints at the optimal policy. We derive a closed-form lower bound for the case of Gaussian distributions and provide geometric insights into how constraints can make a problem easier or harder. Furthermore, we leverage the projection-based computation of the confusing instances to modify TnS (Garivier and Kaufmann 2016) and GE (Degenne, Koolen, and Ménard 2019) to corresponding CTnS and CGE versions for pure exploration in constraint bandits. We empirically evaluate the algorithms on synthetic and real data to assess the impact of constraints on the hardness of the problem.

One interesting future direction is learning when reward and constraints are unknown or partially unknown. Another future direction we deem very interesting is bandit with non-linear constraints as this would change this structure of the normal cone and the resulting projection.

Acknowledgements

Emil Carlsson is funded by Chalmers AI Research Centre (CHAIR) and the Sweden-America foundation (SweAm). Fredrik D. Johansson is funded in part by the

Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Debabrota Basu acknowledges the Inria-Kyoto University Associate Team "RELIANT", the ANR young researcher (JCJC) award for the REPUBLIC project (ANR-22-CE23-0003-01), and the CHIST-ERA project CausalXRL (ANR-21-CHR4-0007).

The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

A Notations

Table 6.1: Notations

K	:	Number of arms.
δ	:	Confidence parameter.
$\mathbb{KL}(x, y)$	:	KL-divergence between two random variables with means x and y.
$\mathbf{kl}(x\|\|y)$	:	KL-divergence between two Bernoulli random variables with means x and y.
$\mathcal{D}$	$\triangleq$	$[\boldsymbol{\mu}_{\min}, \boldsymbol{\mu}_{\max}]^K$, i.e. the range of expected rewards
$\boldsymbol{\mu}$	:	True reward vector, $\boldsymbol{\mu} \in \mathcal{D}$.
$\hat{\boldsymbol{\mu}}_t$	:	Empirical means at time t projected onto $\mathcal{D}$.
B	:	Matrix defining the linear constraints, i.e. $B\boldsymbol{\pi} \leqslant \boldsymbol{c}$.
$\boldsymbol{c}$	:	Vector defining the upper bound in the linear constraints, $B\boldsymbol{\pi} \leqslant \boldsymbol{c}$.
Δ_{K-1}	:	Simplex in K dimensions.
$\mathcal{F}$	$\triangleq$	$\{\boldsymbol{\pi} \in \Delta_{K-1} : B\boldsymbol{\pi} \leqslant \boldsymbol{c}\}$, i.e. the constrained policy space.
$\boldsymbol{\pi}$	:	A feasible policy over K arms, i.e. $\boldsymbol{\pi} \in \mathcal{F}$.
$\boldsymbol{\pi}^*$ or $\boldsymbol{\pi}^*_{\boldsymbol{\mu}}$	:	Unique optimal policy for bandit instance $\boldsymbol{\mu}$, defined as $\boldsymbol{\pi}^*_{\boldsymbol{\mu}} \triangleq \boldsymbol{\pi}^* \triangleq \arg\max_{\boldsymbol{\pi} \in \mathcal{F}} \boldsymbol{\mu}^\top \boldsymbol{\pi}$.
$\mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)$	:	Set of extreme points for $\boldsymbol{\pi}'$, which share $K-1$ linearly independent constraints with $\boldsymbol{\pi}^*$.
$\mathcal{N}(\boldsymbol{\pi}^*)$	:	Normal cone spanned by the active constraints at $\boldsymbol{\pi}^*$.
$\Lambda_{\mathcal{F}}(\boldsymbol{\mu})$	$\triangleq$	$\{\boldsymbol{\lambda} \in \mathcal{D} : \max_{\boldsymbol{\pi} \in \mathcal{F}} \boldsymbol{\lambda}^\top \boldsymbol{\pi} > \boldsymbol{\lambda}^\top \boldsymbol{\pi}^*_{\boldsymbol{\mu}}\}$, i.e. the set of alternative bandit instances.
τ	:	Random stopping time of a pure exploration algorithm.
Π	:	Set of possible exploration policies/allocations.
$T_{\mathcal{F}}(\boldsymbol{\mu})^{-1}$	$\triangleq$	$\sup_{\boldsymbol{w} \in \Pi} \inf_{\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{a=1}^K w_a \mathbb{KL}(\mu_a, \lambda_a)$, the characteristic time for the constrained policy space
$D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F})$	:	Shorthand for $\inf_{\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{a=1}^K w_a \mathbb{KL}(\mu_a, \lambda_a)$.
$D(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\lambda})$	:	Shorthand for $\sum_{a=1}^K w_a \mathbb{KL}(\mu_a, \lambda_a)$.

$w^*(\boldsymbol{\mu})$	:	Set of optimal allocations for bandit instance $\boldsymbol{\mu}$.
H	$\triangleq$	$\frac{2\sigma^2}{\|\boldsymbol{\Delta}\|_2^2}$ quantifies complexity of bandit instance $\boldsymbol{\mu}$

B Lower bound on sample complexity

The following lemma by Kaufmann, Cappé, et al. (2016) provides a general information-theoretic inequality that applies to any bandit model.

Lemma B.1 (Kaufmann, Cappé, et al. (2016))**.** *Let $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ be two bandit models with K arms such that μ_a and λ_a are mutually continuous. For any almost surely finite stopping time τ we have*

$$\sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_{a,\tau}]\mathbb{KL}(\mu_a, \lambda_a) \geqslant \mathbf{kl}(P_{\boldsymbol{\mu}}(\mathcal{E})||P_{\boldsymbol{\lambda}}(\mathcal{E})) \tag{B.1}$$

where $\mathcal{E}$ is any measurable event with respect to the filtration generated by the observed history.

From Lemma B.1 we can directly derive a lower bound on the expected stopping time of any δ-PAC algorithm in the constraint multi-armed bandit setting. We present this lower bound in Theorem B.2 and the proof is virtually the same as the proof for the lower bound in Garivier and Kaufmann (2016). We present it here for completeness.

Theorem B.2 (Lower bound on sample complexity under constraints)**.** *The stopping time τ of any δ-PAC learner satisfy*

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geqslant T_{\mathcal{F}}(\boldsymbol{\mu})\mathbf{kl}(\delta||1-\delta). \tag{B.2}$$

Proof. Let $\boldsymbol{\mu}$ and $\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})$ be two bandit models with K arms such that they do not share optimal policy, i.e. $\boldsymbol{\pi}^*_{\boldsymbol{\mu}} \neq \boldsymbol{\pi}^*_{\boldsymbol{\lambda}}$.

Let $\mathcal{E}$ denote the event of recommending $\boldsymbol{\pi}^*_{\boldsymbol{\mu}}$ for any bandit instance at stopping using some δ-PAC algorithm. Then using Lemma B.1, and δ-correctness of $\boldsymbol{\pi}^*_{\boldsymbol{\mu}}$ for $\boldsymbol{\mu}$, we have

$$\sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_{a,\tau}]\mathbb{KL}(\mu_a, \lambda_a) \geqslant \mathbf{kl}(1-\delta||\delta) = \mathbf{kl}(\delta||1-\delta).$$

Further, we multiple and divide by $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$ which yields

$$\begin{aligned}\sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_{a,\tau}]\mathbb{KL}(\mu_a, \lambda_a) &= \mathbb{E}_{\boldsymbol{\mu}}[\tau] \sum_{a=1}^{K} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_{a,\tau}]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}\mathbb{KL}(\mu_a, \lambda_a) \\ &= \mathbb{E}_{\boldsymbol{\mu}}[\tau] \sum_{a=1}^{K} w_a\mathbb{KL}(\mu_a, \lambda_a) \geqslant \mathbf{kl}(\delta||1-\delta)\,,\end{aligned}$$

where $w_a \triangleq \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_{a,\tau}]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}$, and $\sum_{a=1}^{K} w_a = 1$.

Since the above inequality is true for any $\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})$, we have

$$\inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \mathbb{E}_{\boldsymbol{\mu}}[\tau] \sum_{a=1}^{K} w_a\mathbb{KL}(\mu_a, \lambda_a) = \mathbb{E}_{\boldsymbol{\mu}}[\tau] \inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a\mathbb{KL}(\mu_a, \lambda_a) \geqslant \mathbf{kl}(\delta||1-\delta)\,.$$

The equality is due to the fact that $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$ is independent of $\boldsymbol{\lambda}$.

Now, we further maximise over w_a to get

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \sup_{\boldsymbol{w}\in\Pi} \inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a \mathbb{KL}(\mu_a, \lambda_a) \geqslant \mathbf{kl}(\delta||1-\delta).$$

Finally, using the definition of the characteristic time $T_{\mathcal{F}}(\boldsymbol{\mu})$ yields

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geqslant T_{\mathcal{F}}(\boldsymbol{\mu}) \mathbf{kl}(\delta||1-\delta).$$

□

B.1 Proof of Lemma 3.1

To derive the key properties of the optimal solution and the set of optimal allocations, as presented in Lemma 3.2, we first explicate the set of optimal solutions, and then, use Berge's theorem (Theorem G.1).

Step 1: Recall that

$$\Lambda_{\mathcal{F}}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} \in \mathcal{D} : \boldsymbol{\lambda} \notin \mathcal{N}(\boldsymbol{\pi}^*)\} \ ,$$

where the normal cone is expressed as

$$\mathcal{N}(\boldsymbol{\pi}^*) = \bigcap_{\boldsymbol{\pi}'\in\mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \left\{\boldsymbol{\lambda} \in \mathcal{D} : \boldsymbol{\lambda}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}') \geqslant 0\right\} \ .$$

This is due to the fact that if $\boldsymbol{\pi}^*$ is not the optimal policy under the environment $\boldsymbol{\lambda}$, there exists an improving direction in the simplex algorithm, i.e. a neighbor $\boldsymbol{\pi}'$, such that $\boldsymbol{\lambda}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}') < 0$.

Now, since the set of alternative hypotheses is the compliment of the normal cone, we write

$$\Lambda_{\mathcal{F}}(\boldsymbol{\mu}) = \bigcup_{\boldsymbol{\pi}'\in\mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \left\{\boldsymbol{\lambda} : \boldsymbol{\lambda}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}') < 0\right\}. \tag{B.3}$$

Applying Equation (B.3) in $D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F})$ leads to,

$$D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F}) = \inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a \mathbb{KL}(\mu_a, \lambda_a) = \min_{\boldsymbol{\pi}'\in\mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \inf_{\boldsymbol{\lambda}:\boldsymbol{\lambda}^\top(\boldsymbol{\pi}'-\boldsymbol{\pi}^*)<0} \sum_{a=1}^{K} w_a \mathbb{KL}(\mu_a, \lambda_a) \ .$$

Step 2: What remains to be shown is that the inf is attained by some λ on $\boldsymbol{\lambda}^\top(\boldsymbol{\pi}' - \boldsymbol{\pi}^*) = 0$.

For some $\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)$ take an arbitrary $\boldsymbol{\lambda}' \in \left\{\boldsymbol{\lambda} : \boldsymbol{\lambda}^\top(\boldsymbol{\pi}' - \boldsymbol{\pi}^*) < 0\right\}$. There exists an $\boldsymbol{\lambda}'' \in \left\{\boldsymbol{\lambda} \in \mathcal{D} : \boldsymbol{\lambda}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}') = 0\right\}$ such that $|\mu_a - \lambda'_a| \geqslant |\mu_a - \lambda''_a| \ \forall a$ due to the convexity of $\mathcal{D}$. The mapping $y \to \mathbb{KL}(x, y)$ is an increasing function on the domain $y > x$ and a decreasing function on $y < x$ which implies that

$$\sum_{a=1}^{K} w_a \mathbb{KL}(\mu_a, \lambda'_a) \geqslant \sum_{a=1}^{K} w_a \mathbb{KL}(\mu_a, \lambda''_a). \tag{B.4}$$

There exists a sequence $\{\boldsymbol{\lambda}_t\}_{t=1}^{\infty} \subset \left\{\boldsymbol{\lambda} : \boldsymbol{\lambda}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}') < 0\right\}$ such that $\boldsymbol{\lambda}_0 = \boldsymbol{\lambda}'$ and $\lim_{t\to\infty} \boldsymbol{\lambda}_t = \boldsymbol{\lambda}''$. Hence, we can for any $\boldsymbol{\lambda}'$ get arbitrary close to some $\boldsymbol{\lambda}''$ such that Equation (B.4) holds.

Due to continuity of $\mathbb{KL}(x, .)$, the inf is attained by some $\boldsymbol{\lambda}'' \in \left\{\boldsymbol{\lambda} \in \mathcal{D} : \boldsymbol{\lambda}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}') = 0\right\}$. Hence, we conclude the proof.

B.2 Proof of Theorem 3.2

Property (a-b). We first note that the function $D(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \triangleq \sum_{a=1}^{K} w_a \mathbb{KL}(\mu_a, \lambda_a)$ is continuous in all elements. Take any $(\boldsymbol{w}, \boldsymbol{\mu})$ such that the optimal policy in $\mathcal{F}$ is unique. Let $(\boldsymbol{w}_t, \boldsymbol{\mu}_t)_{t\geqslant 1}$ be a sequence in $\Pi \times \mathcal{D}$ such that

$$(\boldsymbol{w}_t, \boldsymbol{\mu}_t) \xrightarrow{t\to\infty} (\boldsymbol{w}, \boldsymbol{\mu}).$$

Further, for any $\epsilon > 0$ there exists a $t' \geqslant 1$ such that $||(\boldsymbol{w}, \boldsymbol{\mu}) - (\boldsymbol{w}_t, \boldsymbol{\mu}_t)||_2 < \epsilon$ and $\Lambda_{\mathcal{F}}(\mu) = \Lambda_{\mathcal{F}}(\mu_t)\ \forall t \geqslant t'$. By continuity of $D(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ we have that for any $\epsilon' > 0$ there exists exists an $t'' \geqslant 1$ such that for $t \geqslant t''$, we have

$$|D(\boldsymbol{w}_t, \boldsymbol{\mu}_t, \boldsymbol{\lambda}) - D(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\lambda})| \leqslant \epsilon', \forall \lambda \in \mathbb{R}^K.$$

Thus, by taking $t \geqslant t', t''$ leads to

$$\begin{aligned}
|D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F}) - D(\boldsymbol{w}_t, \boldsymbol{\mu}_t, \mathcal{F})| &= \left| \inf_{\lambda\in\Lambda_{\mathcal{F}}(\mu)} D(\boldsymbol{w}, \boldsymbol{\mu}, \lambda) - \inf_{\lambda\in\Lambda_{\mathcal{F}}(\mu_t)} D(\boldsymbol{w}_t, \boldsymbol{\mu}_t, \lambda) \right| \\
&\leqslant \left| \inf_{\lambda\in\Lambda_{\mathcal{F}}(\mu)} (D(\boldsymbol{w}, \boldsymbol{\mu}, \lambda) - D(\boldsymbol{w}_t, \boldsymbol{\mu}_t, \lambda)) \right| \\
&\leqslant \epsilon' ,
\end{aligned}$$

which establishes the continuity properties.

Property (c). The upper hemicontinuity of $w^*(\boldsymbol{\mu})$ and continuity of $D(\boldsymbol{\mu}, \mathcal{F})$ follows from Berge's maximum theorem, see Theorem G.1, by letting $f(x, \theta) = D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F})$ and $C(\theta) = \Pi$. As a consequence of Berge's theorem (Theorem G.1), we substitute the $\sup_w$ with $\max_w$.

Property (d). The convexity of the set $w^*(\boldsymbol{\mu})$ follows from the fact that it is the set of optimal solutions to $\max_{w\in\Pi} D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F})$ and $D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F})$ is concave (Specifically, it is linear in $\boldsymbol{w}$).

B.3 Proof of Theorem 3.3

For two bandit instances $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ consisting of Gaussian distributions with same variance σ^2, we have

$$D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F}) = \min_{\boldsymbol{\lambda}:\boldsymbol{\lambda}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}')=0} \sum_{a=1}^{K} w_a \frac{1}{2\sigma^2}(\mu_a - \lambda_a)^2 .$$

Now, by introducing the Lagrange multiplier γ, we obtain

$$L(\gamma, \boldsymbol{\lambda}) \triangleq \frac{1}{2\sigma^2} \sum_{a=1}^{K} w_a(\mu_a - \lambda_a)^2 - \gamma \boldsymbol{\lambda}^\top (\boldsymbol{\pi}^* - \boldsymbol{\pi}'). \tag{B.5}$$

For brevity, we denote $v \triangleq (\boldsymbol{\pi}^* - \boldsymbol{\pi}')$.

Computing the gradient $\nabla_\lambda L(\gamma, \boldsymbol{\lambda})$ and equating it to 0 yields

$$\lambda_a = \mu_a + \frac{\gamma\sigma^2}{w_a} v_a.$$

Substituting λ_a in Equation (B.5) yields

$$\begin{aligned} L(\gamma) = \min_{\lambda} L(\gamma, \boldsymbol{\lambda}) &= \frac{\sigma^2\gamma^2}{2} \sum_{a=1}^{K} \frac{v_a^2}{w_a} - \gamma \boldsymbol{\mu}^\top v - \sum_{a=1}^{K} \frac{\gamma^2\sigma^2}{w_a} v_a^2 \\ &= -\frac{\sigma^2\gamma^2}{2} \sum_{a=1}^{K} \frac{v_a^2}{w_a} - \gamma \boldsymbol{\mu}^\top v \,. \end{aligned} \tag{B.6}$$

Maximizing over γ yields

$$\gamma = \frac{-\boldsymbol{\mu}^\top v}{\sigma^2 \sum_a \frac{v_a^2}{w_a}} \,,$$

and putting it back in Equation (B.6) gives the final expression of λ_a

$$\lambda_a = \mu_a - \frac{v_a}{w_a} \left(\frac{\boldsymbol{\mu}^\top v}{\sum_a \frac{v_a^2}{w_a}} \right) \,. \tag{B.7}$$

B.4 Proof of Corollary 3.4

Lower bound on the characteristic time: To lower bound $T_{\mathcal{F}}(\boldsymbol{\mu})$, we need to upper bound the RHS in Equation (3.3), i.e. $T_{\mathcal{F}}(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{w}} \min_{\lambda} \sum_a w_a \mathbb{KL}(\mu_a, \lambda_a)$.

Step 1: We first observe that

$$\sup_{\boldsymbol{w}} \min_{\lambda} \sum_a w_a \mathbb{KL}(\mu_a, \lambda_a) = \max_{\boldsymbol{w}} \min_{\lambda} \sum_a w_a \mathbb{KL}(\mu_a, \lambda_a),$$

due to Berge's theorem. Further, the max-min inequality gives

$$\max_{\boldsymbol{w}} \min_{\lambda} \sum_a w_a \mathbb{KL}(\mu_a, \lambda_a) \leqslant \min_{\lambda} \max_{\boldsymbol{w}} \sum_a w_a \mathbb{KL}(\mu_a, \lambda_a).$$

Step 2: We proceed to upper bound $\max_{\boldsymbol{w}} \sum_a w_a \mathbb{KL}(\mu_a, \lambda_a)$ for each neighbor $\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)$ independently.

For a fixed $\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)$, Theorem 3.3 tells us that

$$\begin{aligned}
\min_{\lambda:\lambda^\top(\boldsymbol{\pi}^*-\boldsymbol{\pi}')=0} \sum_{a=1}^{K} w_a \mathbb{KL}(\mu_a, \lambda_a) &= \frac{\gamma^2}{2\sigma^2} \sum_{a=1}^{K} \frac{(\boldsymbol{\pi}^* - \boldsymbol{\pi}')_a^2}{w_a} \\
&= \left(\frac{\boldsymbol{\mu}^\top (\boldsymbol{\pi}^* - \boldsymbol{\pi}')}{\sum_a \frac{(\boldsymbol{\pi}^*-\boldsymbol{\pi}')^2}{w_a}} \right)^2 \frac{1}{2\sigma^2} \sum_{a=1}^{K} \frac{(\boldsymbol{\pi}^* - \boldsymbol{\pi}')_a^2}{w_a} \\
&= \frac{1}{2\sigma^2} \frac{\left(\boldsymbol{\mu}^\top (\boldsymbol{\pi}^* - \boldsymbol{\pi}')\right)^2}{\sum_{a=1}^{K} \frac{(\boldsymbol{\pi}^*-\boldsymbol{\pi}')_a^2}{w_a}}.
\end{aligned}$$

Step 3: We further minimize the expression $\frac{(\boldsymbol{\pi}^*-\boldsymbol{\pi}')_a^2}{w_a}$ under the constraint $\sum_a w_a = 1$.

Using Langrange multiplier technique, we get

$$w_a = \frac{|(\boldsymbol{\pi}^* - \boldsymbol{\pi}')|_a}{\sum_{a=1}^{K} |(\boldsymbol{\pi}^* - \boldsymbol{\pi}')|_a}$$

which yields that $\frac{(\boldsymbol{\pi}^*-\boldsymbol{\pi}')_a^2}{w_a} \geqslant \|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|_1^2$. Hence,

$$\frac{1}{2\sigma^2} \frac{\left(\boldsymbol{\mu}^\top (\boldsymbol{\pi}^* - \boldsymbol{\pi}')\right)^2}{\sum_{a=1}^{K} \frac{(\boldsymbol{\pi}^*-\boldsymbol{\pi}')_a^2}{w_a}} \leqslant \frac{1}{2\sigma^2} \frac{\left(\boldsymbol{\mu}^\top (\boldsymbol{\pi}^* - \boldsymbol{\pi}')\right)^2}{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|_1^2} \leqslant \frac{1}{2\sigma^2} \frac{\left(\boldsymbol{\mu}^\top (\boldsymbol{\pi}^* - \boldsymbol{\pi}')\right)^2}{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|_2^2}.$$

Here, the last part is exactly , $\frac{1}{2} d_{\boldsymbol{\pi}'}^2$, i.e. the squared distance between $\boldsymbol{\mu}$ and the hyperplane $\boldsymbol{\pi}^* - \boldsymbol{\pi} = 0$.

Thus, we conclude the lower bound.

Upper bound on the characteristic time: To obtain the upper bound, we aim to lower bound the inverse $T_{\mathcal{F}}(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{w}} \min_{\lambda} \sum_a w_a \mathbb{KL}(\mu_a, \lambda_a)$.

We let $w_a = \frac{1}{K}, \forall a$, and observe that

$$\max_{w} \min_{\lambda} \sum_a w_a \mathbb{KL}(\mu_a, \lambda_a) \geqslant \min_{\lambda} \frac{1}{K} \sum_a \mathbb{KL}(\mu_a, \lambda_a).$$

For some $\boldsymbol{\pi}' \in \mathcal{V}(\boldsymbol{\pi}^*)$ and using Theorem 3.3 with $w_a = \frac{1}{K}, \forall a$, we get

$$\frac{1}{K}\sum_a \mathbb{KL}(\mu_a, \lambda_a) = \frac{1}{2\sigma^2 K}\frac{\left(\boldsymbol{\mu}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}')\right)^2}{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|_2^2} = d_{\boldsymbol{\pi}'}^2 \frac{1}{2\sigma^2 K}$$

This concludes the upper bound on the characteristic time.

B.5 Proof of Corollary 3.5

Step 1: Neighboring policies and rank-1 update. let $\hat{B} \in \mathbb{R}^{K\times K}$ be a set of linearly independent constraints at $\boldsymbol{\pi}^*$ and $\hat{\boldsymbol{c}}$ be the corresponding values in $\boldsymbol{c}$ such that $\boldsymbol{\pi}^* = \hat{B}^{-1}\hat{\boldsymbol{c}}$. For any $\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)$ we let B'^{-1} and $\boldsymbol{c}'$ be the constraints such that $\boldsymbol{\pi}' = B'^{-1}\boldsymbol{c}'$.

Specifically, B' and $\boldsymbol{c}'$ can be retrieved from the following rank-1 updates

$$B' = \hat{B} + \boldsymbol{e}_r(\boldsymbol{b}'_r - \hat{\boldsymbol{b}}_r)^\top ,$$
$$\boldsymbol{c}' = \hat{\boldsymbol{c}} + (c'_r - c_r)\boldsymbol{e}_r ,$$

where $\hat{\boldsymbol{b}}_r$ a column vector corresponding to the constraint on the r-th row of $\hat{B}$ that we swap with $\boldsymbol{b}'_r$ in order to get B' and $\boldsymbol{e}_r$ a column vector with all elements equal to 0 except the r-th element which is equal to 0. Similarly, $(c'_r - c_r) \neq 0$ is the change that we perform on the r-th element in $\hat{\boldsymbol{c}}$ to get $\boldsymbol{c}'$.

Step 2: From perturbation in constraints to perturbations in policies. Now, we observe that

$$B'\boldsymbol{\pi}' - \hat{B}\boldsymbol{\pi}^* = (c'_r - c_r)\boldsymbol{e_r} .$$

Since $\hat{B}$ is invertible, further rearrangement yields

$$\begin{aligned}
\boldsymbol{\pi}' - \boldsymbol{\pi}^* &= \hat{B}^{-1}\left((c'_r - c_r)\boldsymbol{e}_r + \boldsymbol{e}_r(\hat{\boldsymbol{b}}_r - \boldsymbol{b}'_r)^\top\boldsymbol{\pi}'\right) \\
&= \hat{B}^{-1}\left((c'_r - c_r)\boldsymbol{e}_r + \boldsymbol{e}_r\hat{\boldsymbol{b}}_r^\top\boldsymbol{\pi}' - \boldsymbol{e}_r\boldsymbol{b}_r'^\top\boldsymbol{\pi}'\right) \\
&= \hat{B}^{-1}\left((c'_r - c_r)\boldsymbol{e}_r + \boldsymbol{e}_r\hat{\boldsymbol{b}}_r^\top\boldsymbol{\pi}' - c'_r\boldsymbol{e}_r\right) \\
&= \hat{B}^{-1}\left((\hat{\boldsymbol{b}}_r^\top\boldsymbol{\pi}' - c_r)\boldsymbol{e}_r\right)
\end{aligned}$$

The last part is the slack of $\boldsymbol{\pi}'$ at the r-th constraint in $\hat{B}$, hereby referred to as ξ.

We bound the norm of $\hat{B}^{-1}\boldsymbol{e}_r$ as follows

$$\sigma_{\min}(\hat{B}^{-1}) = \inf_{\boldsymbol{v}:\|v\|_2=1}\|\hat{B}^{-1}v\|_2 \leqslant \|\hat{B}^{-1}\boldsymbol{e}_r\|_2 \leqslant \sup_{\boldsymbol{v}:\|v\|_2=1}\|\hat{B}^{-1}v\|_2 = \sigma_{\max}(\hat{B}^{-1})$$

where $\sigma_{\min}(\hat{B}^{-1})$ and $\sigma_{\max}(\hat{B}^{-1})$ denote the smallest and largest singular value of $\hat{B}^{-1}$. From the properties of the inverse, we get

$$\frac{1}{\sigma_{\max(\hat{B})}} \leqslant \|\hat{B}^{-1}\boldsymbol{e}_r\|_2 \leqslant \frac{1}{\sigma_{\min(\hat{B})}}.$$

Thus, we obtain a lower and upper bound on the perturbation in policies

$$\frac{|\xi|}{\sigma_{\max(\hat{B})}} \leqslant \|\boldsymbol{\pi}' - \boldsymbol{\pi}^*\|_2 \leqslant \frac{|\xi|}{\sigma_{\min(\hat{B})}}. \tag{B.8}$$

Now, using this new representation of change in policy in terms of the slacks in the constraints, we derive our two results.

Step 3 for Part (a): A perspective of the zero-sum game. To get the expression in Equation (3.6) we simply take the expression for $\boldsymbol{\pi}^* - \boldsymbol{\pi}'$, developed in the previous step, and plug into the expression of the characteristic time in Theorem 3.3. Hence,

$$\begin{aligned}
\frac{1}{2\sigma^2}\frac{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^\top}}{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/w_a)}} &= \frac{1}{2\sigma^2}\frac{\|\hat{B}^{-1}\left((\hat{\boldsymbol{b}}_r^\top \boldsymbol{\pi}' - c_r)\boldsymbol{e}_r\right)\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^\top}}{\|\hat{B}^{-1}\left((\hat{\boldsymbol{b}}_r^\top \boldsymbol{\pi}' - c_r)\boldsymbol{e}_r\right)\|^2_{\mathrm{Diag}(1/w_a)}} \\
&= \frac{1}{2\sigma^2}\frac{\|\hat{B}^{-1}(\xi \boldsymbol{e}_r)\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^\top}}{\|\hat{B}^{-1}(\xi \boldsymbol{e}_r)\|^2_{\mathrm{Diag}(1/w_a)}} \\
&= \frac{1}{2\sigma^2}\frac{\|\hat{B}^{-1}(\boldsymbol{e}_r)\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^\top}}{\|\hat{B}^{-1}(\boldsymbol{e}_r)\|^2_{\mathrm{Diag}(1/w_a)}} \\
&= \frac{1}{2\sigma^2}\frac{(\boldsymbol{\Delta}^\top \hat{B}^{-1}(\boldsymbol{e}_r))^2}{\|\hat{B}^{-1}(\boldsymbol{e}_r)\|^2_{\mathrm{Diag}(1/w_a)}}.
\end{aligned}$$

This gives the following expression for the characteristic time

$$T_{\mathcal{F}}(\boldsymbol{\mu})^{-1} = \max_{\boldsymbol{w}\in\Pi} \min_{\boldsymbol{\pi}'\in\mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \frac{1}{2\sigma^2}\frac{\left(\boldsymbol{\Delta}^\top \hat{B}^{-1}_{\boldsymbol{\pi}^*}\boldsymbol{e}_{r'}\right)^2}{\|\hat{B}^{-1}_{\boldsymbol{\pi}^*}\boldsymbol{e}_{r'}\|^2_{\mathrm{Diag}(1/w_a)}}.$$

This formulation of the inverse characteristic time allows us to perceive it as a zero-sum max − min game, where the max-player chooses an exploration allocation and the min-player swaps one of the active constraints, at the optimal policy, with one inactive constraint.

Step 3 for Part (b): Bounds on characteristic time from perturbation in policies. From Corollary 3.4 we have

$$\begin{aligned}
\frac{1}{2\sigma^2}\frac{\left(\boldsymbol{\mu}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}')\right)^2}{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|_2^2} &= \frac{1}{2\sigma^2}\frac{\left(\boldsymbol{\mu}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}') - \mu^*\mathbf{1}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}')\right)^2}{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|_2^2} \\
&= \frac{1}{2\sigma^2}\frac{(\boldsymbol{\mu} - \mu^*\mathbf{1})^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}')^2}{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|_2^2} \\
&= \frac{1}{2\sigma^2}\frac{\left(\boldsymbol{\Delta}^\top \hat{B}^{-1}\boldsymbol{e}_r\right)^2}{\|\hat{B}^{-1}\boldsymbol{e}_r\|_2^2} \\
&\leqslant \frac{\|\boldsymbol{\Delta}\|_2^2}{2\sigma^2}\frac{\sigma^2_{\max}(\hat{B})}{\sigma^2_{\min}(\hat{B})}.
\end{aligned}$$

Step 4 for Part (b): Concluding with complexity of bandit instance and constraints. By referring to $\kappa(\hat{B}) \triangleq \frac{\sigma_{\max}(\hat{B})}{\sigma_{\min}(\hat{B})}$ as the condition number of $\hat{B}$, and $H \triangleq \frac{2\sigma^2}{\|\boldsymbol{\Delta}\|_2^2}$ as the quantifier complexity of bandit instance $\boldsymbol{\mu}$, we get

$$T_{\mathcal{F}}(\boldsymbol{\mu})^{-1} \leqslant \min_{\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \frac{\kappa^2(\hat{B})}{H} .$$

Hence, for any $\boldsymbol{\mu}$, we have that $T_{\mathcal{F}}(\boldsymbol{\mu}) \geqslant \frac{H}{\kappa^2}$, where κ^2 is the minimum condition number of any sub-matrix $\hat{B} \in \mathbb{R}^{K \times K}$ of B consisting of K linearly independent active constraints at $\boldsymbol{\pi}^*$. This leads to a lower bound

$$\mathbb{E}[\tau] \geqslant \Omega\left(\frac{H}{\kappa^2}\mathbf{kl}(\delta||1-\delta)\right).$$

B.6 Theorem 3.3 reduces to the standard BAI bounds with simplex constraints

Recall the theorem statement: If the arms follow Gaussian distributions with identical variance σ^2 and $w_a > 0 \ \forall a$,
we have that the projection $\min_{\boldsymbol{\lambda} \in \mathcal{D}: \boldsymbol{\lambda}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}') \leqslant 0} \sum_{a=1}^K w_a \mathbb{KL}(\mu_a, \lambda_a)$ for any $\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)$ is satisfied by

$$\lambda_{a,\boldsymbol{\pi}'} = \mu_a - \gamma \frac{(\boldsymbol{\pi}^* - \boldsymbol{\pi}')_a}{w_a}, \tag{B.9}$$

for $\gamma = \frac{\boldsymbol{\mu}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}')}{\sum_a \frac{(\boldsymbol{\pi}^* - \boldsymbol{\pi}')^2}{w_a}}$, and the characteristic time is

$$\begin{aligned} T_{\mathcal{F}}(\boldsymbol{\mu})^{-1} &= \max_{\boldsymbol{w} \in \Pi} \min_{\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \frac{1}{2\sigma^2} \frac{\left(\boldsymbol{\mu}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}')\right)^2}{\sum_a \frac{1}{w_a}(\boldsymbol{\pi}^* - \boldsymbol{\pi}')_a^2} \\ &= \max_{\boldsymbol{w} \in \Pi} \min_{\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \frac{1}{2\sigma^2} \frac{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^\top}}{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/w_a)}} \end{aligned}$$

Here, $\mathrm{Diag}(1/w_a)$ is a diagonal matrix with a-th entry of the diagonal as $1/w_a$.

In the case of simplex constraints all extreme points corresponds to deterministic policies and we let $\boldsymbol{\pi}_a$ corresponds to the policy that only plays arm a and let $\boldsymbol{\pi}^* = \boldsymbol{\pi}_1$. For some $\boldsymbol{\pi}_a$ we have, due to Equation (B.9),

$$\lambda_{a', \boldsymbol{\pi}_a} = \mu_{a'}, \forall a' \neq 1, a$$

we further have $\gamma = \frac{\Delta_a}{\frac{1}{w_1} + \frac{1}{w_a}}$ and

$$\begin{aligned} \lambda_{1,\boldsymbol{\pi}_a} &= \mu_1 - \frac{\mu_1 - \mu_a}{\frac{1}{w_1} + \frac{1}{w_a}} \frac{1}{w_1} = \mu_1 - w_a \frac{\mu_1 - \mu_a}{w_1 + w_a} = \frac{1}{w_1 + w_a}(w_1\mu_1 + w_a\mu_a) \\ \lambda_{a,\boldsymbol{\pi}_a} &= \mu_a + \frac{\mu_1 - \mu_a}{\frac{1}{w_1} + \frac{1}{w_a}} \frac{1}{w_a} = \mu_a + w_1 \frac{\mu_1 - \mu_a}{w_1 + w_a} = \frac{1}{w_1 + w_a}(w_1\mu_1 + w_a\mu_a) . \end{aligned}$$

Hence, $\lambda_{1,\boldsymbol{\pi}_a} = \lambda_{a,\boldsymbol{\pi}_a}$ and these are exactly the confusing instance one gets, for each arm a, in the BAI setting (Kaufmann, Cappé, et al. 2016). Plugging back into the expression for the characteristic time yields

$$T_{\mathcal{F}}(\boldsymbol{\mu})^{-1} = \max_{w} \min_{a} \frac{w_1 w_a}{w_1 + w_a} \Delta_a^2.$$

C Upper bounds on sample complexity

C.1 Stopping criterion

Lemma C.1 (Magureanu et al. (2014)). *$\forall \gamma > K+1$ and $t \in \mathbb{N}$ it holds*

$$P\left(\sum_{a=1}^{K} N_{a,t}\mathbb{KL}(\hat{\mu}_a, \mu_a) \geqslant \gamma\right) \leqslant e^{-\gamma}\left(\frac{\lceil \gamma \log t \rceil \gamma}{K}\right)^K e^{K+1}$$

The correctness of our stopping rule in Equation (4.1) follows easily from Lemma C.1. Let $\boldsymbol{\pi}_\tau$ be our recommendation at stopping

$$\begin{aligned} P(\boldsymbol{\pi}_\tau \neq \boldsymbol{\pi}^*) &\leqslant P\left(\exists t \in \mathbb{N} : \sum_{a=1}^{K} N_{a,t}\mathbb{KL}(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\mu}_a) \geqslant c(t,\delta)\right) \\ &\leqslant \sum_{t=1}^{\infty} e^{-c(t,\delta)}\left(\frac{\lceil c(t,\delta)\log t \rceil c(t,\delta)}{K}\right)^K e^{K+1}. \end{aligned}$$

We plug in $c(t,\delta) = \log \frac{t^\alpha C}{\delta}$ and choose C such that

$$\sum_{t=1}^{\infty}\left(\frac{\lceil c(t,\delta)\log t \rceil c(t,\delta)}{K}\right)^K e^{K+1} \leqslant C$$

which yields

$$P(\boldsymbol{\pi}_\tau \neq \boldsymbol{\pi}^*) \leqslant \delta.$$

C.2 Upper bound for CTnS

Proof of Theorem 4.2.
Step 1: Defining Good Event. Let $T \in \mathbb{N}$. For $\epsilon > 0$ and $h(T) = \sqrt{T}$, let $\mathcal{E}_T$ be the event

$$\mathcal{E}_T \triangleq \bigcap_{t=h(T)}^{T} \left\{ \|\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}\|_\infty \leqslant \xi(\epsilon) \right\},$$

where $\xi(\epsilon) < \max_{\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \frac{1}{4\sqrt{K}} \boldsymbol{\mu}^\top (\boldsymbol{\pi}^* - \boldsymbol{\pi}')$ is such that

$$\|\boldsymbol{\mu}' - \boldsymbol{\mu}\|_\infty \leqslant \xi(\epsilon) \implies \sup_{\boldsymbol{w}' \in w^*(\boldsymbol{\mu}')} \sup_{\boldsymbol{w} \in w^*(\boldsymbol{\mu})} \|\boldsymbol{w}' - \boldsymbol{w}\| \leqslant \epsilon$$

This $\xi(\epsilon)$ exists due to the upper hemicontinuity of $w^*(\boldsymbol{\mu})$, Theorem 3.2.

Step 2: Concentrating to Good Event. We will make use of the following Lemma from Garivier and Kaufmann (2016) which bounds the probability of the compliment $\mathcal{E}_T^c$.

Lemma C.2 (Concentration around means (Garivier and Kaufmann 2016))**.** *There exist two constants B, C such that*

$$P(\mathcal{E}_T^c) \leqslant BT \exp\left(-CT^{\frac{1}{8}}\right)$$

This Lemma is due to the fact that C-tracking ensure that each arm has been played at least $\sqrt{t}$ times at each time t, see next Lemma.

Lemma C.3 (Garivier and Kaufmann (2016))**.** *For all $t \geqslant 1$ and $\forall a$, C-Tracking ensures $N_{a,t} \geqslant \sqrt{t + K^2} - K$ and*

$$\max_a \left| N_{a,t} - \sum_{s=1}^{t} \boldsymbol{w}_{a,s} \right| \leqslant K(1 + \sqrt{t}) \tag{C.1}$$

We now leverage to following tracking Lemma of Degenne and Koolen (2019) which holds whenever we are tracking a set of optimal weights.

Lemma C.4 (Concentration in allocations (Degenne and Koolen 2019))**.** *Under $\mathcal{E}_T$, there exists a T_ϵ such that for T where $h(T) \geqslant T_\epsilon$ C-tracking will satisfy*

$$\inf_{\boldsymbol{w} \in w^*(\boldsymbol{\mu})} \|\frac{N_t}{t} - \boldsymbol{w}\|_\infty \leqslant 3\epsilon, \forall t \geqslant 4\frac{K^2}{\epsilon^2} + 3\frac{h(T)}{\epsilon}$$

This shows that C-tracking is eventually going to produce an empirical distribution of plays that is close to an optimal allocation and the empirical distribution will converge to a point in $w^*(\boldsymbol{\mu})$ as $t \to \infty$. We need Lemma C.4 instead of the original tracking result in Garivier and Kaufmann (2016) since the optimal allocation does not need to be unique. However, we know from Theorem 3.2 that the set of optimal allocations $w^*(\boldsymbol{\mu})$ is convex and we can thus apply Lemma C.4.

There exists a T_ϵ such that under $\mathcal{E}_T$ and $t \geqslant \max(T_\epsilon, h(T))$ we have

$$|(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_t)^\top \boldsymbol{\pi}^*| \leqslant \sqrt{K}\xi < \frac{1}{4} \max_{\boldsymbol{\pi}' \in \mathcal{V}_\mathcal{F}(\boldsymbol{\pi}^*)} \boldsymbol{\mu}^\top (\boldsymbol{\pi}^* - \boldsymbol{\pi}')$$

which implies that $\boldsymbol{\pi}^* = \arg\max_{\boldsymbol{\pi} \in \mathcal{F}} \hat{\boldsymbol{\mu}}_t^\top \boldsymbol{\pi}$. This ensures that we will be computing the stopping criterion w.r.t. to the correct Alt-set $\Lambda_\mathcal{F}(\boldsymbol{\mu})$.

Step 3: Complexity given the Good Event. Assume $T \geqslant T_\epsilon$ and let

$$C_{\epsilon,\mathcal{F}}(\boldsymbol{\mu}) \triangleq \inf_{\substack{\boldsymbol{\mu}' : \|\boldsymbol{\mu}' - \boldsymbol{\mu}\|_\infty \leqslant \xi(\epsilon) \\ \boldsymbol{w}' : \|\boldsymbol{w}' - \boldsymbol{w}\|_\infty \leqslant 3\epsilon, \forall \boldsymbol{w} \in w^*(\boldsymbol{\mu})}} D(\boldsymbol{w}', \boldsymbol{\mu}', \mathcal{F}).$$

This $C_{\epsilon,\mathcal{F}}(\boldsymbol{\mu})$ gives the worst-case characteristic time we might compute in the algorithm due to the fact that our estimates are not exact.

Assuming $\mathcal{E}_T$, Lemma C.4 gives for $t \geqslant T_\epsilon$

$$D(\boldsymbol{N}_t, \hat{\boldsymbol{\mu}}_t, \mathcal{F}) \geqslant t C_{\epsilon,\mathcal{F}}(\boldsymbol{\mu}).$$

Step 4: Bounding the Stopping Time for Good and Bad Events. Let τ_δ be the stopping time, then

$$\min(\tau_\delta, T) \leqslant \sqrt{T} + \sum_{t=T_\epsilon}^{T} \mathbb{I}_{\tau_\delta > t}$$

and plugging in our stopping rule, i.e. $D(\boldsymbol{N}_t, \hat{\boldsymbol{\mu}}_t, \mathcal{F}) > c(t, \delta)$ yields

$$\begin{aligned} T_\epsilon + \sum_{t=T_\epsilon}^{T} \mathbb{I}(D(\boldsymbol{N}_t, \hat{\boldsymbol{\mu}}_t, \mathcal{F}) \leqslant c(t,\delta)) &\leqslant \sqrt{T} + \sum_{t=T_\epsilon}^{T} \mathbb{I}(t C_{\epsilon,\mathcal{F}}(\boldsymbol{\mu}) \leqslant c(T,\delta)) \\ &\leqslant \sqrt{T} + \frac{c(T,\delta)}{C_{\epsilon,\mathcal{F}}(\boldsymbol{\mu})}. \end{aligned}$$

We define $T_\delta := \inf\left\{T \in \mathbb{N} : \sqrt{T} + \frac{c(T,\delta)}{C_{\epsilon,\mathcal{F}}(\boldsymbol{\mu})} \leqslant T\right\}$. Hence,

$$\mathbb{E}[\tau_\delta] \leqslant T_\epsilon + T_\delta + \sum_{T=1}^{\infty} BT \exp\left(-CT^{\frac{1}{8}}\right) \leqslant T_\epsilon + T_\delta + T'$$

where $\sum_{t=1}^{\infty} BT \exp\left(-Ct^{\frac{1}{8}}\right) \leqslant T' < \infty$. We bound T_δ in the same way as Garivier and Kaufmann (2016). Let $C(\eta) = \inf\{T : T - \sqrt{T} \geqslant T\frac{1}{1+\eta}\}$ for some $\eta > 0$. Then

$$T_\delta \leqslant C(\eta) + \inf\left\{T \in \mathbb{N} : T\frac{C_{\epsilon,\mathcal{F}}(\boldsymbol{\mu})}{1+\eta} \geqslant c(T,\delta)\right\}.$$

Step 5: Obtaining the Asymptotic Bound. Dividing Equation C.2 with $\log\frac{1}{\delta}$ and taking the limit yields

$$\liminf_{\delta \to 0} \frac{\mathbb{E}[\tau_\delta]}{\log\frac{1}{\delta}} \leqslant \frac{\alpha(1+\eta)}{C_{\epsilon,\mathcal{F}}(\boldsymbol{\mu})}.$$

$C_{\epsilon,\mathcal{F}}(\boldsymbol{\mu})$ is continuous due to Theorem 3.2 and taking the limits $\eta, \epsilon \to 0$ yields

$$\liminf_{\delta \to 0} \frac{\mathbb{E}[\tau_\delta]}{\log\frac{1}{\delta}} \leqslant \alpha T_\mathcal{F}(\boldsymbol{\mu}), \forall \alpha > 1.$$

□

C.3 Upper bound for CGE

The proof follows the same structure as the proof of Theorem 2 in Degenne, Koolen, and Ménard (2019) and we use the same concentration analysis. The main difference is that we have to adjust the definition of approximate optimistic saddle point algorithm.

Proof of Theorem 4.3.
Step 1: Defining Good Event. We start by defining the good event

$$\mathcal{E}_T \triangleq \{\forall t \leqslant T\, \forall a, N_{a,t}\mathbb{KL}(\hat{\mu}_{a,t}, \mu_t) \leqslant f(t)\}$$

where $f(t) = 3\log t + \log\log t$.

Step 2: Concentration of Good Event.

We can bound $\sum_{t=1}^{\infty} P(\mathcal{E}_T^c)$ using Lemma C.1. Hence, for any $t \in \mathbb{N}$ and arm a

$$\begin{aligned} P(N_{a,t}\mathbb{KL}(\hat{\mu}_{a,t}, \mu_t) \geqslant f(t)) &\leqslant e^{-f(t)}(1 + f(t)\log t)f(t) \\ &= \frac{e^2}{t^3 \log t}(f(t) + f(t)^2 \log t). \end{aligned}$$

Summing yields

$$\sum_{a=1}^{K}\sum_{t=1}^{\infty} P(\mathcal{E}_T^c) \leqslant K + K\sum_{t=2}^{\infty} \frac{e^2}{t^3 \log t}\left(f(t) + f(t)^2 \log t\right) \leqslant KC < \infty. \tag{C.2}$$

Here a constant $C = 21$ is sufficient.

Step 3: Starting from the Stopping Criterion The main idea of the proof is to work with the stopping criterion

$$c(t,\delta) \geqslant \inf_{\lambda \in \Lambda_{\mathcal{F}}(\hat{\boldsymbol{\mu}}_t)} \sum_{a=1}^{K} N_{a,t}\mathbb{KL}(\hat{\mu}_{a,t}, \lambda_a)$$

and show that if we have the event $\mathcal{E}_T$, our current recommendation at some t is the correct policy $\boldsymbol{\pi}^*$ and we haven't stopped yet, we can lower bound $c(t,\delta)$ in a way that depends on the characteristic time and properties of the no-regret learners. We start with assuming our current recommendation at some t is the correct policy $\boldsymbol{\pi}^*$ and we have the event $\mathcal{E}_T$,

$$c(t,\delta) \geqslant \inf_{\lambda \in \Lambda_{\mathcal{F}}(\hat{\boldsymbol{\mu}}_t)} \sum_{s=1}^{t}\sum_{a=1}^{K} w_{a,s}\mathbb{KL}(\hat{\mu}_{a,t}, \lambda_a) - (1 + \sqrt{t})K$$

which follows from Tracking Lemma C.3. We now use a concentration result, originally in Appendix D.1 of Degenne, Koolen, and Ménard (2019),

$$c(t,\delta) \geqslant \inf_{\lambda \in \Lambda_{\mathcal{F}}(\hat{\boldsymbol{\mu}}_t)} \sum_{s=1}^{t}\sum_{a=1}^{K} w_{a,s}\mathbb{KL}(\hat{\mu}_{a,s}, \lambda_a) - (1 + \sqrt{t})K - O(\sqrt{t\log t}). \tag{C.3}$$

This steps follows from the Lipschitz property of the KL and the fact we have conditioned on $\mathcal{E}_T$ (see Step 8 for further details). Hence,

$$|\mathbb{KL}(\mu_a, \lambda_a) - \mathbb{KL}(\hat{\mu}_{a,s}, \lambda_a)| \leqslant L\sqrt{2\sigma^2 \frac{f(s)}{N_{a,s}}}$$

which implies that

$$\sum_{s=1}^{t}\sum_{a=1}^{K} w_{a,s}\mathbb{KL}(\hat{\mu}_{a,t}, \lambda_a) \geqslant \sum_{s=1}^{t}\sum_{a=1}^{K} w_{a,s}\mathbb{KL}(\mu_a, \lambda_a) - L\sqrt{2\sigma^2 K t f(t)}.$$

Using the same result one more time yields

$$\sum_{s=1}^{t}\sum_{a=1}^{K} w_{a,s}\mathbb{KL}(\hat{\mu}_{a,t}, \lambda_a) \geqslant$$
$$\sum_{s=1}^{t}\sum_{a=1}^{K} w_{a,s}\mathbb{KL}(\hat{\mu}_{a,s}, \lambda_a) - L\sqrt{2\sigma^2 K t f(t)} - 2L\sqrt{2\sigma^2 f(t)}\left(K^2 + 2\sqrt{2Kt}\right)$$

which gives the result in Equation (C.3).

Step 4: Defining Approximate Optimistic Saddle Point under Constraints. We now introduce concepts and properties that will help us to further lower bound the RHS in Equation (C.3). We extend the definition of an *approximate optimistic saddle point algorithm* from Degenne, Koolen, and Ménard (2019) to the constraint setting.

Definition C.1. An algorithm playing sequences of $(\boldsymbol{w}_s, \boldsymbol{\lambda}_s)_{s\leqslant t} \in (\Pi \times \Lambda_{\mathcal{F}})^t$ is said to be an *approximate optimistic saddle point algorithm* with slack x_t if

$$\inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{s=1}^{t}\sum_{a=1}^{K} w_{s,a}\mathbb{KL}(\hat{\mu}_{a,s}, \lambda_a) \geqslant \max_{\boldsymbol{w}\in\Pi} \sum_{a=1}^{K}\sum_{s=1}^{t} w_a U_{a,s} - x_t, \tag{C.4}$$

where x_t is defined in Eq. (C.7) and the confidence bound

$$U_{a,s} = \max\left\{\frac{f(t)}{N_{a,s}}, \max_{\xi\in[\alpha_{a,s},\beta_{a,s}]} \mathbb{KL}(\xi, , \lambda_{a,s})\right\}.$$

The difference in Definition C.1 compared to the definition of an approximate optimistic saddle point algorithm in Degenne, Koolen, and Ménard (2019) is that we in Equation C.4 take the maximum over Π and instead of arms as in Degenne, Koolen, and Ménard (2019). This is due to the fact that maximum over arms might not be in the set of feasible exploration policies Π.

Step 5: Definition of Regret of the Two Players. We define the regret of the allocation player, i.e. AdaGrad, as

$$R_t^{\boldsymbol{w}} = \max_{\boldsymbol{w}\in\Pi} \sum_{s=1}^{t}\sum_{a=1}^{K} w_a U_{a,s} - \sum_{s=1}^{t}\sum_{a=1}^{K} w_{a,t} U_{a,s} \tag{C.5}$$

and note that AdaGrad has an regret scaling of $R_{\boldsymbol{w}}^t \leqslant O(\sqrt{Qt})$ where Q is an upper bound on the losses such that $Q \geqslant \max_{x,y\in[\mu_{\min},\mu_{\max}]} \mathbb{KL}(x,y)$. For the instance player we define the regret as

$$R_t^{\boldsymbol{\lambda}} = \sum_{s=1}^{t}\sum_{a=1}^{K} w_{a,s}\mathbb{KL}(\hat{\mu}_{a,s},\lambda_{a,s}) - \inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{s=1}^{t}\sum_{a=1}^{K} w_{a,s}\mathbb{KL}(\hat{\mu}_{a,s},\lambda_a) \tag{C.6}$$

and note that $R_{\boldsymbol{\lambda}}^t \leqslant 0$ since the instance player is performing a best-response against $\boldsymbol{w}_s$ at each s.

Step 6: CGE is an Approximate Optimistic Saddle Point Algorithm We now show that the CGE is an approximate optimistic saddle point algorithm. From the regret properties of $\boldsymbol{\lambda}$ player we have

$$\inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{s=1}^{t}\sum_{a=1}^{K} w_{s,a}\mathbb{KL}(\hat{\mu}_{a,s},\lambda_a) \geqslant \inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{s=1}^{t}\sum_{a=1}^{K} w_{s,a}\mathbb{KL}(\hat{\mu}_{a,s},\lambda_{a,s})$$

since $R_{\boldsymbol{\lambda}}^t \leqslant 0$.

Let $C_{a,s} = U_{a,s} - \mathbb{KL}(\hat{\mu}_{a,s},\lambda_{a,s})$. We have

$$\inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{s=1}^{t}\sum_{a=1}^{K} w_{s,a}\mathbb{KL}(\hat{\mu}_{a,s},\lambda_a) \geqslant \inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{s=1}^{t}\sum_{a=1}^{K} w_{s,a}U_{a,s}(\boldsymbol{\lambda}) - \sum_{s=1}^{t}\sum_{a=1}^{K} w_{s,a}C_{a,s}.$$

Now, we can combine Eq. (C.3) and (C.5) to get

$$c(t,\delta) \geqslant \inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}\boldsymbol{\mu}} \sum_{s=1}^{t}\sum_{a=1}^{K} w_{s,a}U_{a,s}(\boldsymbol{\lambda}) - \sum_{s=1}^{t}\sum_{a=1}^{K} w_{s,a}C_{a,s} - (1+\sqrt{t})K - O(\sqrt{t\log t})$$

Now we use the properties of $R_t^{\boldsymbol{w}}$ to get

$$\inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{s=1}^{t}\sum_{a=1}^{K} w_{s,a}\mathbb{KL}(\hat{\mu}_{a,s},\lambda_s) \geqslant \max_{\boldsymbol{w}\in\Pi} \sum_{a=1}^{K}\sum_{s=1}^{t} w_a U_{a,s} - R_{\boldsymbol{w}}^t - \sum_{s=1}^{t}\sum_{a=1}^{K} w_{a,s}C_{a,s}$$

which shows that CGE is an approximate optimistic saddle point algorithm with slack

$$x_t = R_t^{\boldsymbol{w}} + \sum_{s=1}^{t}\sum_{a=1}^{K} w_{s,a}C_{a,t}. \tag{C.7}$$

Step 7: Plug slack x_t into Equation (C.3)**.** We now use the fact that CGE is an approximate optimistic saddle point algorithm in Equation (C.3)

$$c(t,\delta) \geqslant \max_{\boldsymbol{w}\in\Pi} \sum_{a=1}^{K}\sum_{s=1}^{t} w_a U_{a,s} - R_t^{\boldsymbol{w}} - \sum_{s=1}^{t}\sum_{a=1}^{K} w_{s,a}C_{a,t} - (1+\sqrt{t})K - O(\sqrt{t\log t}) \tag{C.8}$$

Step 8: Concentration of $\sum_{a=1}^{K} w_{s,a}C_{a,t}$

Assume the event $\mathcal{E}_T$. We have

$$|\mathbb{KL}(\mu_a, \lambda_a) - \mathbb{KL}(\hat{\mu}_{a,s}, \lambda_a)| \leqslant L\mathbb{KL}(\hat{\mu}_{a,s}, \mu_a)$$

due to the Lipschitz property of the KL-divergence and under the event $\mathcal{E}_T$ we have

$$|\mathbb{KL}(\mu_a, \lambda_a) - \mathbb{KL}(\hat{\mu}_{a,s}, \lambda_a)| \leqslant L\sqrt{2\sigma^2 \frac{f(s)}{N_{a,s}}}.$$

This implies that

$$\sup_{\xi \in [\alpha_{a,s}, \beta_{a,s}]} U_{a,s} - \mathbb{KL}(\xi, \lambda_{a,s}) \leqslant \max\left\{2L\sqrt{2\sigma^2 \frac{f(s)}{N_{a,s}}}, \frac{f(s)}{N_{a,s}}\right\}$$

since either $U_{a,s} = \max_{\xi \in [\alpha_{a,s}, \beta_{a,s}]} \mathbb{KL}(\xi, , \lambda_{a,s})$ and the above is equal to the width of the confidence interval, or $U_{a,s} = \frac{f(s)}{N_{a,s}}$ and the above is trivially bounded $\frac{f(s)}{N_{a,s}}$ since the KL divergence is non-negative. Hence,

$$\begin{aligned}
\sum_{s=K+1}^{t} \sum_{a=1}^{K} w_{s,a} C_{a,s} &\leqslant \sum_{s=K+1}^{t} \sum_{a=1}^{K} w_{s,a} \left(\frac{f(s)}{N_{a,s}} + 2L\sqrt{2\sigma^2 \frac{f(s)}{N_{a,s}}}\right) \\
&\leqslant f(t) \sum_{s=K+1}^{t} \sum_{a=1}^{K} \frac{w_{s,a}}{N_{a,s}} + 2L\sqrt{2\sigma^2 f(t)} \sum_{s=K+1}^{t} \sum_{a=1}^{K} \frac{w_{s,a}}{\sqrt{N_{a,s}}} \\
&\leqslant f(t)\left(K^2 + 2K \log \frac{t}{K}\right) + 2L\sqrt{2\sigma^2 f(t)}\left(K^2 + 2\sqrt{2Kt}\right) \\
&\leqslant O(\sqrt{t \log t}).
\end{aligned}$$

We have

$$c(t, \delta) \geqslant \max_{\boldsymbol{w} \in \Pi} \sum_{a=1}^{K} \sum_{s=1}^{t} w_a U_{a,s} - R_t^{\boldsymbol{w}} - O(\sqrt{t \log t}) - (1 + \sqrt{t})K - O(\sqrt{t \log t}).$$

Step 9: Optimism
We now use the fact that $U_{a,s} \geqslant \mathbb{KL}(\mu_a, \lambda_a)$ under the event $\mathcal{E}_T$. Hence,

$$c(t, \delta) \geqslant \max_{\boldsymbol{w} \in \Pi} \sum_{a=1}^{K} \sum_{s=1}^{t} w_a \mathbb{KL}(\mu_a, \lambda_{a,s}) - R_t^{\boldsymbol{w}} - O(\sqrt{t \log t}) - (1 + \sqrt{t})K - O(\sqrt{t \log t}).$$

Step 10: Get the Characteristic Time We note that

$$\begin{aligned}
\max_{\boldsymbol{w} \in \Pi} \sum_{a=1}^{K} \sum_{s=1}^{t} w_a \mathbb{KL}(\mu_a, \lambda_{a,s}) &\geqslant t \inf_{\lambda \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \max_{\boldsymbol{w} \in \Pi} \sum_{a=1}^{K} w_a \mathbb{KL}(\mu_a, \lambda_a) \\
&\geqslant \max_{\boldsymbol{w} \in \Pi} \inf_{\lambda \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a \mathbb{KL}(\mu_a, \lambda_a) = t T_{\mathcal{F}}^{-1}(\boldsymbol{\mu}).
\end{aligned}$$

Rearanging yields

$$t \leqslant T_{\mathcal{F}}(\boldsymbol{\mu})c(t,\delta) + R_t^{\boldsymbol{w}} + O(\sqrt{t \log t})$$

Step 11: Current Recommendation is the Wrong Policy. The above result is conditioned on the fact that our current recommendation is correct. We now bound the number of time steps where the current recommendation is wrong, using similar argument as in Degenne, Koolen, and Ménard (2019).

We define the Chernoff information as $\mathrm{ch}(x,y) \triangleq \inf_{u \in \mathcal{D}} : \mathbb{KL}(u,x) + \mathbb{KL}(u,y)$. Assumption 1 gives that there $\exists\ \epsilon > 0$ such that $\forall \boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})$, $\exists a'$ such that $\mathrm{ch}(\lambda_{a'}, \mu_{a'}) > \epsilon$.

Assume that $\boldsymbol{\pi}^* \neq \arg\max_{\boldsymbol{\pi} \in \mathcal{F}} \hat{\boldsymbol{\mu}}_t^\top \boldsymbol{\pi}$, i.e. if we stop we would recommend the wrong policy. This implies that $\hat{\boldsymbol{\mu}}_t \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})$ and $\mathrm{ch}(\hat{\mu}_{a,t}, \mu_a) \geqslant \epsilon$ for some arm a. Under the good event $\mathcal{E}_T$ we have $N_{a,t}\mathbb{KL}(\hat{\mu}_{a,t}, \mu_a) \leqslant f(t)$ which implies that $\frac{f(t)}{N_{a,t}} \geqslant \epsilon$, since $\mathrm{ch}(\hat{\mu}_{a,t}, \mu_a) \leqslant \mathbb{KL}(\hat{\mu}_{a,t}, \mu_a)$.

Let $\boldsymbol{\pi}_s \triangleq \arg\max_{\boldsymbol{\pi} \in \mathcal{F}} \hat{\boldsymbol{\mu}}_s^\top \boldsymbol{\pi}$, let $n_{\boldsymbol{\pi}'}(t)$ be the number of stages where $\boldsymbol{\pi}_s = \boldsymbol{\pi}'$. Our goal is to upper bound $n_{\boldsymbol{\pi}'}(t)$ for all extreme points $\boldsymbol{\pi}' \in \mathcal{F}$ such that $'\boldsymbol{\pi}' \neq \boldsymbol{\pi}^*$. For any $\boldsymbol{\lambda}$ such that $\boldsymbol{\pi}' = \arg\max_{\boldsymbol{\pi} \in \mathcal{F}} \boldsymbol{\lambda}^\top \boldsymbol{\pi}$ we have that $\boldsymbol{\mu} \in \Lambda_{\mathcal{F}}(\boldsymbol{\lambda})$ which gives

$$\epsilon_t = \sum_{s=1, \boldsymbol{\pi}_s \neq \boldsymbol{\pi}^*}^{t} \sum_{a=1}^{K} w_{a,s}\mathbb{KL}(\hat{\mu}_{a,s}, \mu_a) \geqslant \sum_{\boldsymbol{\pi}' \neq \boldsymbol{\pi}^*} \inf_{\boldsymbol{\lambda}: \boldsymbol{\pi}' \neq \arg\max_{\boldsymbol{\pi}} \boldsymbol{\lambda}^\top \boldsymbol{\pi}} \sum_{s=1, \boldsymbol{\pi}_s = \boldsymbol{\pi}'}^{t} \sum_{a=1}^{K} w_{a,s}\mathbb{KL}(\hat{\mu}_{a,s}, \lambda_a).$$

We use the fact that on the time steps where $\boldsymbol{\pi}_s = \boldsymbol{\pi}'$ CGE is a optimistic saddle point algorithm with slack $x = R^{\boldsymbol{w}}_{n_{\boldsymbol{\pi}'}(t)} + \sum_{s=1, \boldsymbol{\pi}_s = \boldsymbol{\pi}'}^{t} \sum_{a=1}^{K} w_{s,a} C_{a,t}$. Hence,

$$\inf_{\boldsymbol{\lambda}: \boldsymbol{\pi}' \neq \arg\max_{\boldsymbol{\pi}} \boldsymbol{\lambda}^\top \boldsymbol{\pi}} \sum_{s=1, \boldsymbol{\pi}_s = \boldsymbol{\pi}'}^{t} \sum_{a=1}^{K} w_{a,s}\mathbb{KL}(\hat{\mu}_{a,s}, \lambda_a) \geqslant$$

$$\max_{\boldsymbol{\pi} \in \Pi} \sum_{s=1, \boldsymbol{\pi}_s = \boldsymbol{\pi}'}^{t} \sum_{a=1}^{K} w_a U_{a,s} - R^{\boldsymbol{w}}_{n_{\boldsymbol{\pi}'}(t)} - \sum_{s=1, \boldsymbol{\pi}_s = \boldsymbol{\pi}'}^{t} \sum_{a=1}^{K} w_{a,s} C_{a,s}.$$

Under the event $\mathcal{E}_T$, and $s \leqslant t$ such that $\boldsymbol{\pi}_s = \boldsymbol{\pi}'$ there is an arm a_s such that $U_{a_s,s} \geqslant \epsilon$. This implies that the sum $\max_{\boldsymbol{\pi} \in \Pi} \sum_{s=1, \boldsymbol{\pi}_s = \boldsymbol{\pi}'}^{t} \sum_{a=1}^{K} w_a U_{a,s}$ is increasing linearly in $n_{\boldsymbol{\pi}'}(t)$ since it is at least $\epsilon n_{\boldsymbol{\pi}'}(t)$ under the concentration event $\mathcal{E}_T$. Thus,

$$\inf_{\boldsymbol{\lambda}: \boldsymbol{\pi}' \neq \arg\max_{\boldsymbol{\pi}} \boldsymbol{\lambda}^\top \boldsymbol{\pi}} \sum_{s=1, \boldsymbol{\pi}_s = \boldsymbol{\pi}'}^{t} \sum_{a=1}^{K} w_{a,s}\mathbb{KL}(\hat{\mu}_{a,s}, \lambda_a) \geqslant \epsilon n_{\boldsymbol{\pi}'}(t) - R^{\boldsymbol{w}}_{n_{\boldsymbol{\pi}'}(t)} - \sum_{s=1, \boldsymbol{\pi}_s = \boldsymbol{\pi}'}^{t} \sum_{a=1}^{K} w_{a,s} C_{a,s}$$

and we know that $R^{\boldsymbol{w}}_{n_{\boldsymbol{\pi}'}(t)} = O(\sqrt{Q n_{\boldsymbol{\pi}'}(t)})$ and $\sum_{s=1, \boldsymbol{\pi}_s = \boldsymbol{\pi}'}^{t} \sum_{a=1}^{K} w_{a,s} C_{a,s} = O(\sqrt{n_{\boldsymbol{\pi}'}(t) \log n_{\boldsymbol{\pi}'}(t)})$. This shows that ϵ_T increases at least linear in $n_{\boldsymbol{\pi}'}(t)$ and thus also linearly in the number of time steps for whitch $\boldsymbol{\pi}_s \neq \boldsymbol{\pi}^*$. However, we have

$$\begin{aligned}\epsilon_t &= \sum_{s=1, \boldsymbol{\pi}_s \neq \boldsymbol{\pi}^*}^{t} \sum_{a=1}^{K} w_{a,s}\mathbb{KL}(\hat{\mu}_{a,s}, \mu_a) \leqslant \sum_{s=1}^{t} \sum_{a=1}^{K} w_{a,s} \frac{f(s)}{N_{a,s}} \\ &\leqslant f(t)(K^2 + 2K \log \frac{t}{K}).\end{aligned}$$

This implies that the current recommendation $\boldsymbol{\pi}_s = \arg\max \hat{\boldsymbol{\mu}}_t^\top \boldsymbol{\pi}$ differs from $\boldsymbol{\pi}^*$ at most $O(\sqrt{t \log t})$ number of times.

Step 12: Final Bound. We know from the concentration of $\mathcal{E}_T$ that the number of times the compliment happens is upper bounded by CK where C is some problem independent constant. Putting it all together, we get that $\mathbb{E}[\tau] \leqslant T_0(\delta) + CK$, where

$$T_0(\delta) := \max \left\{ t \in \mathbb{N} : t \leqslant T_{\mathcal{F}}(\boldsymbol{\mu}) c(t, \delta) + O(\sqrt{tQ}) + O(\sqrt{t \log t}) \right\}.$$

□

D Finding ϵ-good policies under linear constraints

In some cases one might be more interested in finding a policy that is ϵ-close to the optimal one, i.e. finding $\boldsymbol{\pi}'$ such that $\boldsymbol{\mu}^\top(\boldsymbol{\pi}^*_{\boldsymbol{\mu}} - \boldsymbol{\pi}') \leqslant \epsilon$, since this might have a much smaller sample complexity compared to searching for the optimal policy, see for example (Garivier and Kaufmann 2021) and (Kocák and Garivier 2021). Both CTnS and CGE can in principle be extended to this case by changing the definition of the Alt-set. Given an instance $\boldsymbol{\mu}$ let $\Omega_{\mathcal{F},\epsilon}(\boldsymbol{\mu}) := \{\boldsymbol{\pi} \in \mathcal{N}_{\mathcal{F}} : \mu^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}) \leqslant \epsilon\}$ be the set of ϵ-good policies where $\mathcal{N}_{\mathcal{F}}$ is the set of all extreme points in the polytope $\mathcal{F}$. For each $\boldsymbol{\pi} \in \Omega_{\mathcal{F},\epsilon}(\boldsymbol{\mu})$ we get the following Alt-set

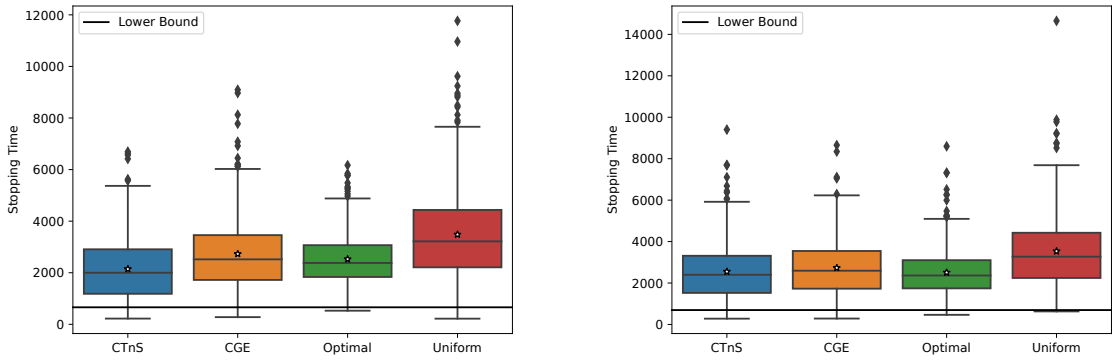
$$\Lambda_{\mathcal{F},\epsilon}(\boldsymbol{\mu}, \boldsymbol{\pi}) := \left\{\boldsymbol{\lambda} : \boldsymbol{\lambda}^\top \left(\boldsymbol{\pi}^*_{\boldsymbol{\lambda}} - \boldsymbol{\pi}\right) > \epsilon\right\}.$$

Hence, the sample complexity might be different depending on which near-optimal policy the learner is considering. To handle this we would have to augment CTnS and CGE with the "sticky" approach developed in (Degenne and Koolen 2019), where the learner commits to a recommendation since otherwise the learner might oscillate between near-optimal policies and a mixture of their optimal allocations might not be optimal since $w^*(\boldsymbol{\mu})$ is no longer ensured to be convex. Furthermore, due to $\epsilon > 0$ it is no longer sufficient to project onto the normal cone and a naive implementation would have to optimize over $|\mathcal{N}_{\mathcal{F}}|$ convex sets which might only be tractable for a small set of constraints and/or arms.

E Additional experimental analysis

In Figure 6.6 and 6.7 we present results for arms with Bernoulli distributions and in Figure 6.8 and 6.9 we present additional results for arms with Gaussian distributions. CTnS and CGE outperforms the uniform baseline in all cases and are usually on par with or better than the learner that always sample according to the asymptotically optimal allocation. We also see that the algorithms tend to be close to the lower bound in all cases. An interesting observation, which we commented on already in the main text, is that there tend to be a larger difference between all sampling rules for end-of-time constraints compared to anytime constraints. This is due to the fact that anytime constraints can be very restrictive on which sampling allocations are allowed and there might not be less room for an adaptive learner.

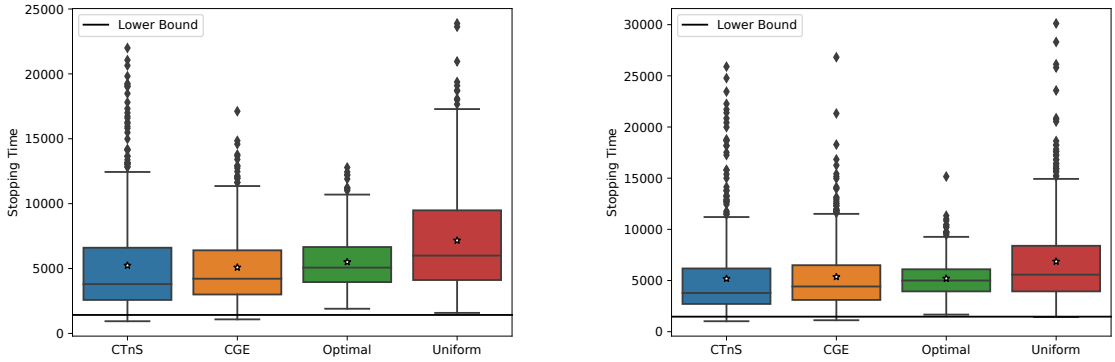
In the case of arms with Bernoulli distributions we did not use a close-form projection, as for Gaussian distributions, and instead computed the projection numerically by minimizing the KL-divergence subject to $\boldsymbol{\lambda}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}') = 0$, which is a convex problem. We discuss the effect of this in Section E.1



(a) End-of-time constraints

(b) Anytime constraints

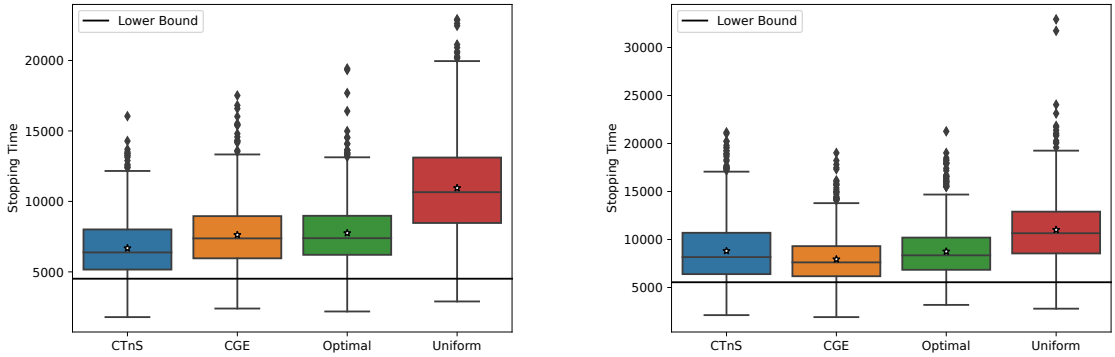
Figure 6.6: End-of-time and Anytime constraints with *Bernoulli* arms. The reward vector is $\boldsymbol{\mu} = (0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2)$ and the constraints are $\boldsymbol{\pi}_1 + \boldsymbol{\pi}_2 \leqslant 0.5$ and $\boldsymbol{\pi}_3 + \boldsymbol{\pi}_4 \leqslant 0.5$. Average over 500 seeds and $\delta = 0.1$. Optimal policy is $\boldsymbol{\pi}_1 = 0.5$ and $\boldsymbol{\pi}_3 = 0.5$.



(a) End-of-time constraints

(b) Anytime constraints

Figure 6.7: End-of-time and Anytime constraints with Bernoulli arms. The reward vector is $\boldsymbol{\mu} = (0.8, 0.7, 0.6, 0.5, 0.4)$ and the constraints are $4\boldsymbol{\pi}_1 - \boldsymbol{\pi}_5 \leqslant 1$ and $3\boldsymbol{\pi}_2 - \boldsymbol{\pi}_4 \leqslant 1$. Average over 500 seeds and $\delta = 0.1$. Optimal policy is $\boldsymbol{\pi}_1 = 0.25$, $\boldsymbol{\pi}_2 = 0.33$ and $\boldsymbol{\pi}_3 = 0.42$.



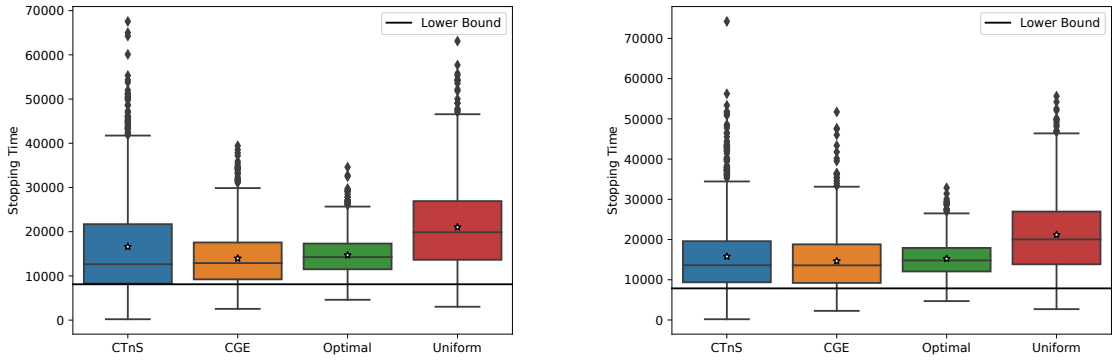
(a) End-of-time constraints

(b) Anytime constraints

Figure 6.8: End-of-time and Anytime constraints with *Gaussian* arms $\sigma^2 = 1$. The reward vector is $\boldsymbol{\mu} = (2.0, 1.5, 1.45, 0.5, 0.3, -1.0, -1.0)$ and the constraints are $4\boldsymbol{\pi}_1 + \boldsymbol{\pi}_2 \leqslant 0.7$ and $\boldsymbol{\pi}_2 + 2\boldsymbol{\pi}_3 \leqslant 0.5$. Average over 1000 seeds and $\delta = 10^{-4}$. Optimal policy is $\boldsymbol{\pi}_1 = 0.05$, $\boldsymbol{\pi}_2 = 0.5$ and $\boldsymbol{\pi}_4 = 0.45$.

E.1 Running times

In Table 6.2 we present the average time it take for the algorithms to check the stopping criterion and select a new arm to play. The test was performed on 1 core of a Intel Xeon Gold 6130 CPU with 2.1 GHz. Gaussian indicates the experiments in Figure 6.8a, Bernoulli the experiments in Figure 6.6a and IMDB the experiments in Figure 6.5b. As expected CTnS is the algorithm requiring most computational time and the excessive running time it has on the experiment with Bernoulli distributions is due to the fact that we numerically solve the projection instead of relying on a close-form expression as in the case of Gaussian distributions. In contrast, we see that CGE has a relatively light computational footprint in all cases. Another



(a) End-of-time constraints

(b) Anytime constraints

Figure 6.9: End-of-time and Anytime constraints with Gaussian arms $\sigma^2 = 1$. The reward vector is $\boldsymbol{\mu} = [1.0, 0.5, 0.4, 0.3, 0.2, 0.1]$ and the constraints are $\boldsymbol{\pi}_1 - \boldsymbol{\pi}_4 - \boldsymbol{\pi}_5 - \boldsymbol{\pi}_6 \leqslant 0.3$ and $\boldsymbol{\pi}_2 \leqslant 0.7$. Average over 1000 seeds and $\delta = 10^{-3}$. Optimal policy is $\boldsymbol{\pi}_1 = 0.65$ and $\boldsymbol{\pi}_4 = 0.35$.

advantage of CGE is that it performs a finite number of max calls at each iteration which can easily be parallelized for larger bandit instances with many constraints.

Algorithm	Bernoulli	Gaussian	IMDB
CTnS	1.00 ± 0.244	0.030 ± 0.006	0.033 ± 0.015
CGE	0.02 ± 0.001	0.005 ± 0.001	0.008 ± 0.001
Uniform	$0.009 \pm 3 \times 10^{-4}$	$0.001 \pm 1 \times 10^{-4}$	$0.002 \pm 2 \times 10^{-4}$

Table 6.2: Average time, in seconds, it takes to check the stopping criterion and select a new arm for the different algorithms. The$\pm$ indicates one standard deviation. We omitted the optimal sampler since this one has the same running time as the uniform sampler.

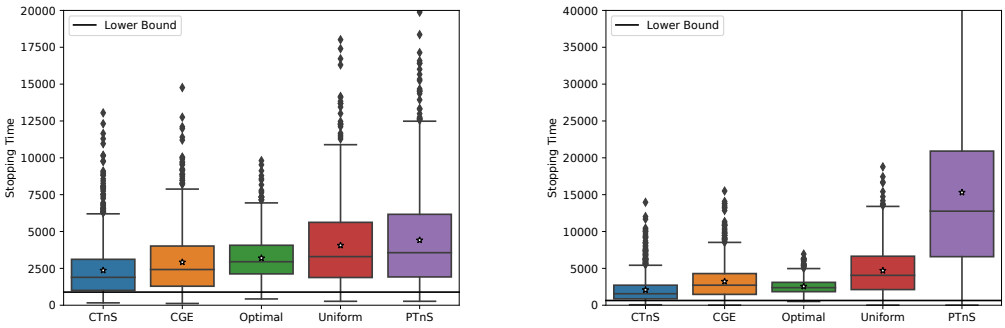
E.2 IMDB environment

For resproducibility, here we provide the specifics of the IMDB data in the Table 6.3 as used in the experiments (Figure 6.5).

Movie	Average Rating	σ	Action	Drama	Family
The Net	3.67	1.26	1	1	0
Happily N'Ever After	2.97	1.30	0	0	1
Tomorrowland	2.94	1.31	1	0	1
American Hero	3.52	1.33	1	1	0
Das Boot	3.18	1.30	0	1	0
Final Destination 3	2.02	0.93	0	0	0
Licence to Kill	2.79	1.22	1	0	0
The Hundred-Foot Journey	2.97	1.31	0	1	0
The Matrix	2.32	1.14	1	0	0
Creature	2.53	1.20	0	0	0
The Basket	2.55	1.19	0	1	0
Star Trek: The Motion Picture	2.54	1.16	0	0	0

Table 6.3: Movies used in the experiments presented in Figure 6.5. The optimal policy is $\boldsymbol{\pi}_1^* = 0.3$, $\boldsymbol{\pi}_2^* = 0.3$ and $\boldsymbol{\pi}_5^* = 0.4$. We used the maximum σ in the algorithms. This means that the algorithms didn't have access to the true σ of each arm and instead modelled them all as Gaussian distributions with $\sigma = 1.33$ but the rewards were sampled from the environment using the true σ.

F On the sub-optimality of PTnS



(a) Anytime Constraints

(b) End-of-Time Constraints

Figure 6.10: Problem instance with 8 Gaussian arms with $\sigma = 1$. The arm means are $\mu = [1.0, 0.7, 0.3, 0.0, -0.5, -1.0, -2.0, -3.0]$ and we have one constraint $7\pi_1 + 7\pi_2 + \pi_3 \leqslant 0.5$. The optimal policy is $\pi_3 = \pi_4 = 0.5$. Results for $\delta = 0.1$ and 1000 random seeds.

In Figure 6.10, we consider an eight-armed bandit with Gaussian reward distributions with means

$$\boldsymbol{\mu} = [1.0, 0.7, 0.3, 0.0, -0.5, -1.0, -2.0, -3.0],$$

variance 1, and the constraint $7\pi_1 + 7\pi_2 + \pi_3 \leqslant 0.5$.

We observe that PTnS performs the worst on this instance, specially in the end-of-time setting. This reflects the fact that *the optimal allocation w.r.t. classical BAI bound does not have to be close to the optimal allocation given by the constraint version of the lower bound.*

In Figure 6.10b, the optimal allocation for the constraint problem is

$$\boldsymbol{w}^* = [0.09, 0.02, \mathbf{0.43}, \mathbf{0.36}, 0.03, 0.02, 0.02, 0.02],$$

while the unconstrained optimal BAI allocation with the same $\boldsymbol{\mu}$ is

$$\hat{\boldsymbol{w}} = [\mathbf{0.43}, \mathbf{0.42}, 0.05, 0.03, 0.02, 0.02, 0.02, 0.02].$$

Hence, PTnS focuses on exploring arm 1 and 2 the most, which makes sense without any constraints. In contrast, the optimal allocation under constraint, i.e. w^*, suggests that one should focus on arm 3 and 4 as the constraint puts a disproportional cost on arm 1 and 2.

In the anytime scenario, Figure 6.10a, the optimal allocation is

$$w^* = [0.02, 0.01, \mathbf{0.32}, \mathbf{0.54}, 0.03, 0.03, 0.03, 0.03].$$

In this scenario, the allocation $\hat{\boldsymbol{w}}$, computed by PTnS, is no longer feasible and PTnS instead converges to the projected version

$$w' = [0.03, 0.02, 0.12, 0.18, \mathbf{0.16}, \mathbf{0.16}, \mathbf{0.16}, \mathbf{0.16}].$$

We observe that the previous issue is now mitigated by the projection, PTnS is no longer overly obsessed with arm 1 and 2. However, another issue arises as the projection distributes a substantial probability to the arms $5 - 8$, which are highly sub-optimal. These phenomena lead to worse performance of PTnS w.r.t. CTnS and CGE, as shown in Figure 6.10a.

G Useful definitions and results

Definition G.1 (Upper hemicontinuity)**.** We say that a set-valued function $C : \Theta \to \Omega$ is upper hemicontinuous at the point $\theta \in \Theta$ if for any open set $S \subset \Omega$ with $C(\theta) \subset S$ there exists a neighborhood U around θ, such that $\forall x \in U$, $C(x)$ is a subset of S.

Theorem G.1 (Berge's maximum theorem (Berge 1963))**.** *Let X and Θ be topological spaces. Let $f : X \times \Theta \to \mathbb{R}$ be a continuous function and let $C : \Theta \to X$ be a compact-valued correspondence such that $C(\theta) \neq \varnothing$ $\forall \theta \in \Theta$. If C is continuous at θ then $f^*(\theta) = \sup_{x \in C(\theta)} f(x, \theta)$ is continuous and $C^* = \{x \in C(\theta) : f(x, \theta) = f^*(\theta)\}$ is upper hemicontinuous.*

Below we restate the upper bound on the sample complexity of the Gamified Explorer (GE) of Degenne, Koolen, and Ménard (2019).

Theorem G.2 (Theorem 2 in Degenne, Koolen, and Ménard (2019))**.** *The sample complexity of GE is*

$$\mathbb{E}[\tau] \leqslant T_0(\delta) + \frac{eK}{a}$$

where

$$T_0(\delta) = \max\{t \in \mathbb{N} : t \leqslant T(\boldsymbol{\mu}) c(t, \delta) + C_{\boldsymbol{\mu}}(R_t^{\boldsymbol{\lambda}} + R_t^{\boldsymbol{w}} + O(\sqrt{t \log t}))\}$$

where $R_t^{\boldsymbol{\lambda}}$ is the regret of the instance player, $R_t^{\boldsymbol{w}}$ the regret of the allocation player and $C_{\boldsymbol{\mu}}$ an instance-dependent constant.

References

Agrawal, Shipra and Nikhil Devanur (2016). "Linear contextual bandits with knapsacks". In: *Advances in Neural Information Processing Systems* 29 (cit. on p. 184).

Agrawal, Shubhada, Sandeep Juneja, and Peter Glynn (2020). "Optimal δ-Correct Best-Arm Selection for Heavy-Tailed Distributions". In: *Algorithmic Learning Theory*. PMLR, pp. 61–110 (cit. on pp. 184, 186).

Amani, Sanae, Mahnoosh Alizadeh, and Christos Thrampoulidis (2019). "Linear stochastic bandits under safety constraints". In: *Advances in Neural Information Processing Systems* 32 (cit. on p. 183).

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. 181).

Badanidiyuru, Ashwinkumar, Robert Kleinberg, and Aleksandrs Slivkins (Mar. 2018). "Bandits with Knapsacks". In: *J. ACM* 65.3. ISSN: 0004-5411. DOI: 10.1145/3164539 (cit. on p. 184).

Bechhofer, Robert E (1958). "A sequential multiple-decision procedure for selecting the best one of several normal populations with a common unknown variance, and its use with various experimental designs". In: *Biometrics* 14.3, pp. 408–429 (cit. on p. 183).

Berge, C. (1963). *Topological Spaces: Including a Treatment of Multi-valued Functions, Vector Spaces and Convexity*. Macmillan (cit. on p. 224).

Boyd, Stephen and Lieven Vandenberghe (Mar. 2004). *Convex Optimization*. Cambridge University Press (cit. on p. 187).

Brannath, Werner, Emmanuel Zuber, Michael Branson, Frank Bretz, Paul Gallo, Martin Posch, and Amy Racine-Poon (2009). "Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology". In: *Statistics in medicine* 28.10, pp. 1445–1463 (cit. on p. 181).

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. 181).

Camilleri, Romain, Andrew Wagenmaker, Jamie H Morgenstern, Lalit Jain, and Kevin G Jamieson (2022). "Active Learning with Safety Constraints". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 33201–33214 (cit. on p. 183).

Chen, Ningyuan (2021). "Multi-armed bandit requiring monotone arm sequences". In: *Advances in Neural Information Processing Systems* 34, pp. 16093–16103 (cit. on p. 181).

Degenne, Rémy and Wouter M Koolen (2019). "Pure Exploration with Multiple Correct Answers". 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 pp. 184–186, 191, 209, 217).

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 pp. 183, 184, 191, 192, 194, 195, 211, 212, 215, 224).

Degenne, Rémy, Pierre Ménard, Xuedong Shang, and Michal Valko (2020). "Gamification of pure exploration for linear bandits". In: *International Conference on Machine Learning*. PMLR, pp. 2432–2442 (cit. on pp. 184, 185).

Demirel, Ilker, Ahmet Alparslan Celik, and Cem Tekin (2022). "Escada: Efficient safety and context aware dose allocation for precision medicine". In: *Advances in Neural Information Processing Systems* 35, pp. 27441–27454 (cit. on p. 181).

Duchi, John, Elad Hazan, and Yoram Singer (2011). "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization". In: *Journal of Machine Learning Research* 12.61, pp. 2121–2159 (cit. on p. 192).

Even-Dar, Eyal, Shie Mannor, and Y. Mansour (2002). "PAC Bounds for Multi-armed Bandit and Markov Decision Processes". In: *Annual Conference Computational Learning Theory* (cit. on p. 181).

Faizal, Fathima Zarin and Jayakrishnan Nair (2022). "Constrained Pure Exploration Multi-Armed Bandits with a Fixed Budget". In: *arXiv preprint arXiv:2211.14768* (cit. on p. 183).

Fiez, Tanner, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff (2019). "Sequential experimental design for transductive linear bandits". In: *Advances in neural information processing systems* 32 (cit. on p. 184).

Garivier, Aurélien and Emilie Kaufmann (2016). "Optimal best arm identification with fixed confidence". In: *Conference on Learning Theory*. PMLR, pp. 998–1027 (cit. on pp. 183, 184, 186, 187, 190–192, 195, 199, 209, 210).

Garivier, Aurélien and Emilie Kaufmann (2021). "Nonasymptotic sequential tests for overlapping hypotheses applied to near-optimal arm identification in bandit models". In: *Sequential Analysis* 40.1, pp. 61–96. DOI: 10.1080/07474946.2021.1847965. eprint: https://doi.org/10.1080/07474946.2021.1847965 (cit. on p. 217).

Gillulay, Jeremy H and Claire J Tomlin (2011). "Guaranteed safe online learning of a bounded system". In: *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, pp. 2979–2984 (cit. on p. 183).

Immorlica, Nicole, Karthik Sankararaman, Robert Schapire, and Aleksandrs Slivkins (2022). "Adversarial bandits with knapsacks". In: *Journal of the ACM* 69.6, pp. 1–47 (cit. on p. 184).

Jedra, Yassir and Alexandre Proutiere (2020). "Optimal best-arm identification in linear bandits". In: *Advances in Neural Information Processing Systems* 33, pp. 10007–10017 (cit. on p. 184).

Kalyanakrishnan, Shivaram, Ambuj Tewari, Peter Auer, and Peter Stone (Jan. 2012). "PAC Subset Selection in Stochastic Multi-armed Bandits". In: *Proceedings of the 29th International Conference on Machine Learning, ICML 2012* 1 (cit. on p. 181).

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. 188, 189, 192, 199, 207).

Kaufmann, Emilie and Wouter M. Koolen (2021). "Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals". In: *Journal of Machine Learning Research* 22.246, pp. 1–44 (cit. on p. 190).

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. 184).

Kocák, Tomáš and Aurélien Garivier (Aug. 2021). "Epsilon Best Arm Identification in Spectral Bandits". 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. 2636–2642. DOI: 10.24963/ijcai.2021/363 (cit. on pp. 184, 217).

Kunaver, Matevž and Tomaž Požrl (2017). "Diversity in recommender systems – A survey". In: *Knowledge-Based Systems* 123, pp. 154–162 (cit. on p. 181).

Lattimore, Tor and Csaba Szepesvári (2020). *Bandit Algorithms*. Cambridge University Press. DOI: 10.1017/9781108571401 (cit. on p. 181).

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. 181).

Lindner, David, Sebastian Tschiatschek, Katja Hofmann, and Andreas Krause (2022). "Interactively Learning Preference Constraints in Linear Bandits". In: *International Conference on Machine Learning*. PMLR, pp. 13505–13527 (cit. on p. 184).

Lindståhl, Simon, Alexandre Proutiere, and Andreas Johnsson (2022). "Measurement-based admission control in sliced networks: A best arm identification approach". In: *GLOBECOM 2022-2022 IEEE Global Communications Conference*. IEEE, pp. 1484–1490 (cit. on p. 181).

Losada, David E, David Elsweiler, Morgan Harvey, and Christoph Trattner (2022). "A day at the races: using best arm identification algorithms to reduce the cost of information retrieval user studies". In: *Applied Intelligence* 52.5, pp. 5617–5632 (cit. on p. 181).

Maas, Andrew L., Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts (June 2011). "Learning Word Vectors for Sentiment Analysis". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 142–150 (cit. on p. 194).

Magureanu, Stefan, Richard Combes, and Alexandre Proutiere (13–15 Jun 2014). "Lipschitz Bandits: Regret Lower Bound and Optimal Algorithms". In: *Proceedings of The 27th Conference on Learning Theory*. Ed. by Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári. Vol. 35. Proceedings of Machine Learning Research. Barcelona, Spain: PMLR, pp. 975–999 (cit. on p. 208).

Moldovan, Teodor Mihai and Pieter Abbeel (2012). "Safe exploration in markov decision processes". In: *arXiv preprint arXiv:1205.4810* (cit. on p. 183).

Moradipari, Ahmadreza, Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis (2021). "Safe linear thompson sampling with side information". In: *IEEE Transactions on Signal Processing* 69, pp. 3755–3767 (cit. on p. 183).

Pacchiano, Aldo, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang (2021). "Stochastic bandits with linear constraints". In: *International conference on artificial intelligence and statistics*. PMLR, pp. 2827–2835 (cit. on p. 183).

Sui, Yanan, Alkis Gotovos, Joel Burdick, and Andreas Krause (2015). "Safe exploration for optimization with Gaussian processes". In: *International conference on machine learning*. PMLR, pp. 997–1005 (cit. on p. 183).

Sui, Yanan, Vincent Zhuang, Joel Burdick, and Yisong Yue (2018). "Stagewise safe bayesian optimization with gaussian processes". In: *International conference on machine learning*. PMLR, pp. 4781–4789 (cit. on p. 183).

Vaswani, Sharan, Lin F. Yang, and Csaba Szepesvári (Nov. 2022). "Near-Optimal Sample Complexity Bounds for Constrained MDPs". In: *NeurIPS* (cit. on p. 183).

Wan, Runzhe, Branislav Kveton, and Rui Song (17–23 Jul 2022). "Safe Exploration for Efficient Policy Evaluation and Comparison". In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 22491–22511 (cit. on p. 183).

Wang, Lequn, Yiwei Bai, Wen Sun, and Thorsten Joachims (2021). "Fairness of exposure in stochastic bandits". In: *International Conference on Machine Learning*. PMLR, pp. 10686–10696 (cit. on pp. 181, 183).

Wang, Zhenlin, Andrew J Wagenmaker, and Kevin Jamieson (2022). "Best arm identification with safety constraints". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 9114–9146 (cit. on p. 183).

Paper 7

Active preference learning for ordering items in- and out-of-sample

Herman Bergström*, Emil Carlsson*, Devdatt Dubhashi, Fredrik D. Johansson.

To appear in the Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS), 2024.
** indicates equal contribution.*

The paper has been reformatted for uniformity.

Paper 7. Active preference learning for ordering items in- and out-of-sample

Herman Bergström*, Emil Carlsson*, Devdatt Dubhashi, Fredrik D. Johansson.

Abstract

Learning an ordering of items based on pairwise comparisons is useful when items are difficult to rate consistently on an absolute scale, for example, when annotators have to make subjective assessments. When exhaustive comparison is infeasible, actively sampling item pairs can reduce the number of annotations necessary for learning an accurate ordering. However, many algorithms ignore shared structure between items, limiting their sample efficiency and precluding generalization to new items. It is also common to disregard how noise in comparisons varies between item pairs, despite it being informative of item similarity. In this work, we study active preference learning for ordering items with contextual attributes, both in- and out-of-sample. We give an upper bound on the expected ordering error of a logistic preference model as a function of which items have been compared. Next, we propose an active learning strategy that samples items to minimize this bound by accounting for aleatoric and epistemic uncertainty in comparisons. We evaluate the resulting algorithm, and a variant aimed at reducing model misspecification, in multiple realistic ordering tasks with comparisons made by human annotators. Our results demonstrate superior sample efficiency and generalization compared to non-contextual ranking approaches and active preference learning baselines.

1 Introduction

The success of supervised learning is built on annotating items at great volumes with small error. For subjective assessments, however, assigning a value from an arbitrary rating scale can be difficult and prone to inconsistencies, causing many to favor *preference feedback* from pairwise comparisons (Yannakakis and Martínez 2015; Christiano et al. 2017; Ouyang et al. 2022; Zhu et al. 2023). Preference feedback is sufficient to learn an *ordering* of items (Fürnkranz and Hüllermeier 2003), but for n items, there are $O(n^2)$ possible pairs of items to compare. A common solution is to use crowd-sourcing (Chen, Bennett, et al. 2013; Yang et al. 2021; Larkin et al. 2022), but many tasks require domain *expertise*, making annotations *expensive* to collect. This is the case in the field of medical imaging, where annotations require trained radiologists (Phelps et al. 2015; Jang et al. 2022; Lidén et al. 2024; Tärnåsen

and Bergström 2023). So, how can we learn the best ordering possible from a limited number of comparisons?

Classically, this problem is solved by active learning, sampling comparisons based on preference feedback and estimated item scores (Herbrich et al. 2006; Maystre and Grossglauser 2017; Heckel et al. 2018). However, consider a radiologist who wants to quantify the expression of a disease in a collection of X-ray images. Purely preference-based algorithms utilize only the outcomes of comparisons but ignore the contents of the X-rays, which can reveal similarities between items and inform an ordering strategy. Moreover, the set we want to order is often larger than the set of items observed during training—we may want to rank new X-rays in relation to previous ones. This cannot be solved by learning per-item scores alone. As an alternative, active learning for classification can be used to fit a map from pairs of item contexts x_i, x_j (e.g., the contents of images) to the comparison $i \succ_? j$, that can be applied to old and new items alike (Houlsby et al. 2011; Qian et al. 2015). However, as we show in Section 4, learning this map to recover a *complete ordering* is distinct from the tasks preference learning is commonly used for, and existing algorithms lack theoretical justification for this application. Moreover, formal results for related problems, such as contextual bandits or reinforcement learning (Das et al. 2024; Filippi et al. 2010; Zhu et al. 2023; Bengs, Saha, et al. 2022), do not translate directly to effective active sampling criteria for ordering. There is a small body of work on learning a contextual model to recover the complete ordering (Jamieson and Nowak 2011; Ailon 2011) but these either assume noiseless preference feedback or that the noise is unrelated to the similarity of items, which is unrealistic for subjective assessments.

Contributions. We propose using a contextual logistic preference model to support efficient in-sample ordering and generalization to new items. Our analysis yields the first bound on the expected ordering error achievable given a collected set of comparisons (Section 4). This result justifies an active sampling principle that accounts for both epistemic and aleatoric uncertainty which we implement in a greedy deterministic algorithm called GURO (Section 5). We further propose a hybrid variant of the contextual preference model, compatible with GURO as well as existing sampling strategies, that overcomes model misspecification by adding per-item parameters (Section 5.1). We evaluate GURO and baseline algorithms in four diverse ordering tasks, three of which utilize comparisons performed by human annotators (Section 6). Our sampling strategy compares favorably to active preference learning baselines, and our hybrid model benefits both GURO and other sampling criteria, achieving the low variance of contextual models and the low bias of fitting per-item parameters. This results in faster convergence in-sample, better generalization to new items, and efficient continual learning when new items are added.

2 Ordering items with active preference learning

Our goal is to learn an ordering of items $\mathcal{I}$ according to an unobserved score $y_i \in \mathbb{R}$, defined for each item $i \in \mathcal{I}$. The ground-truth ordering of $\mathcal{I}$ is determined by a comparison function $\pi_{ij} := \mathbf{1}[y_i > y_j]$, where $\pi_{ij} = 1$ indicates that i ranks higher than j. We assume there are no ties.

We define the *ordering error* $R_{\mathcal{I}}(h)$ of a learned comparison function $h : \mathcal{I} \times \mathcal{I} \to \{0, 1\}$ as the frequency of pairwise inversions under a uniform distribution of item pairs,

$$R_{\mathcal{I}}(h) = \frac{2}{n(n-1)} \sum_{i \neq j \in \mathcal{I}} \mathbf{1}[h(i,j) \neq \pi_{ij}] \ , \tag{2.1}$$

where $n = |\mathcal{I}|$. This error is equivalent to the normalized Kendall's Tau distance (Kendall 1948).

Hypotheses h are learned from *preference feedback*—noisy pairwise comparisons $C_{ij} \in \{0, 1\}$ for items (i, j) related to their score, for example, provided by human annotators. $C_{ij} = 1$ indicates that an annotator perceived that item i has a higher score than j, i.e., that they prefer i over j. *Our goal is to minimize the ordering error $R_{\mathcal{I}}(h)$ for a fixed budget $T \geqslant 1$ of adaptively chosen comparisons.*

We are interested in contextual problems, where each item $i \in \mathcal{I}$ is endowed with item-specific attributes $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$. As we will see, this permits more sample-efficient ordering and learning algorithms that can order items out-of-sample, trained on comparisons of a subset of items $\mathcal{I}_D \subseteq \mathcal{I}$ and generalizing to $\mathcal{I} \backslash \mathcal{I}_D$. Ordering algorithms based *only* on preference feedback cannot solve this problem since observed comparisons are uninformative of new items.

Our *active preference learning* scenario proceeds as follows:

- A learner is given an annotation budget T, a pool of items $\mathcal{I}_D \subseteq \mathcal{I}$ and item attributes x_i for $i \in \mathcal{I}_D$.
- Over rounds $t = 1, ..., T$, the learner requests a comparison of two items $i_t, j_t \in \mathcal{I}_D$ according to a sampling criterion and receives noisy binary preference feedback $c_t \sim p(C_{ij})$, independently of previous comparisons.
- After T rounds, the learner returns a comparison function $h : \mathcal{I} \times \mathcal{I} \to \{0, 1\}$.

We denote the history of accumulated observations until and including time t by $D_t = ((i_1, j_1, c_1), ..., (i_t, j_t, c_t))$.

We assume that comparisons C_{ij} follow a logistic model applied to the difference between item scores, $p(C_{ij} = 1) = \sigma(y_i - y_j)$, the so-called Bradley-Terry model (Bradley and Terry 1952), which assumes linear stochastic transitivity (Oliveira et al. 2018). Throughout, $\sigma(z) = 1/(1 + e^{-z})$ and $\dot{\sigma}(z)$ its derivative at z. Specifically, we study the case where y_i is a linear function of item attributes, $y_i = \theta_*^\top x_i$, with $\theta_* \in \mathbb{R}^d$ the ground-truth coefficients. Thus, comparisons are determined by a logistic regression model applied to the attribute difference vector $z_{ij} := x_i - x_j$,

$$p(C_{ij} = 1) = \sigma(\theta_*^\top z_{ij}) \ . \tag{2.2}$$

We face two kinds of uncertainty when actively learning the model in (2.2): *epistemic* and *aleatoric*. Epistemic uncertainty, or model uncertainty, is the uncertainty about the true parameter θ_*, while aleatoric uncertainty is the irreducible uncertainty about labels due to noisy annotation.

3 Related work

Active Preference Learning: *Preference learning* (Fürnkranz and Hüllermeier 2003; Chu and Ghahramani 2005) is related to the problem of *learning to rank* (Burges et al. 2005; Busse et al. 2012). When using adaptively chosen comparisons it may be posed as an *active learning* or *bandit* problem (Brinker 2004; Long et al. 2010; Silva et al. 2014; Ling et al. 2020). Non-contextual active learners, such as TrueSkill (Herbrich et al. 2006; Minka et al. 2018), Hamming-LUCB (Heckel et al. 2018), and ProbeRank (Lou et al. 2022) produce in-sample preference orderings, but must be updated if new items are to be ranked. Contextual algorithms, such as BALD (Houlsby et al. 2011), mitigate this by exploiting item structure and Kirsch and Gal (2022) show that many recently proposed contextual active learning strategies may be unified in a framework based on Fisher information. Similarly, methods have been proposed to recover a linear preference model by adaptively sampling paired comparisons (Qian et al. 2015; Massimino and Davenport 2021; Canal et al. 2019). Still, this setting differs from ours in that we emphasize recovering the full ordering, not perfectly estimating the parameters. While it is true that knowing the parameters is sufficient to order the list, reducing uncertainty for all parameters equally will likely be wasteful (see Section 4). Ailon (2011) offer guarantees for ordering using contextual features in the noiseless setting, while Jamieson and Nowak (2011) analyze the setting where noise is unrelated to item similarity.

Bandits: Bandit algorithms with *relative* or *dueling* feedback (Yue and Joachims 2009; Bengs, Busa-Fekete, et al. 2021; Yan et al. 2022) also learn from pairwise comparisons, and have been proposed both in contextual (Dudík et al. 2015) and non-contextual settings (Yue, Broder, et al. 2012) to minimize regret or identify top-k items. Bengs, Saha, et al. (2022) proposed CoLSTIM, a contextual dueling bandit for regret minimization under linear stochastic transitivity, matching (2.2), and Di et al. (2023) gave variance-aware regret bounds for this setting. However, algorithms that find the top-k items, such as pure exploration bandits (Fang 2022; Jun et al. 2021), can be arbitrarily bad at learning a full ordering (see Appendix A). Related are also George and Dimitrakakis (2023) who learn Kemeny rankings in non-contextual dueling bandits, and Wu, Jin, et al. (2023) who minimize Borda regret. Zhu et al. (2023) studies the problem of estimating a preference model from offline data. Our analysis uses techniques from logistic bandits (Filippi et al. 2010; Li et al. 2017; Faury et al. 2020; Kveton et al. 2020).

RLHF: Preference learning is commonly used when training large language models through reinforcement learning with human feedback (RLHF) (Christiano et al. 2017;

Bai et al. 2022; Ouyang et al. 2022; Wu, Zhu, et al. 2023). In this line of work, Zhu et al. (2023) provide guarantees on the sample complexity of learning a preference model from offline data. They leverage similar tools from statistical learning and bandits as we do. In contrast to their work, we provide sampling strategies for the online setting. Mehta et al. (2023) consider active learning for RLHF in a dueling bandit framework where the goal is to optimize a contextual version of the Borda regret. Concurrent work by Mukherjee et al. (2024) and Das et al. (2024) studies a similar problem, as we do here, in the RLHF setting but with the objective to identify an optimal policy in a contextual bandit with dueling feedback. In contrast to their objective, we are interested in recovering the ordering of items. Das et al. (2024) use similar bandit techniques as we do, and their selection criterion, when adapted for ordering, corresponds to our NormMin baseline (see Section 6).

4 Which comparisons result in a good ordering?

We give an upper bound on the ordering error $R_{\mathcal{I}}(h)$ for a hypothesis h fit using a set of T comparisons. The bound is retrospective, attempting to answer the question "if we collect comparisons D_T, how good is our resulting model at ordering the items in $\mathcal{I}$?". In Section 5, we use insights from the result to design an active learning algorithm.

We restrict our analysis to the logistic model in (2.2) and denote by $R(\theta) \equiv R_{\mathcal{I}}(h_\theta)$ the risk of the hypothesis defined by $h_\theta(i,j) = \mathbf{1}[\theta^\top z_{ij} > 0]$. Recall that $z_{ij} = x_i - x_j$ for $i, j \in \mathcal{I}$, and define $z_t \equiv z_{i_t j_t}$ as the difference between attributes for the pair of items selected at round t. Let θ_t be the maximum-likelihood estimate (MLE) fit to t rounds of feedback, D_t

$$\theta_t = \arg\max_\theta \sum_{s=1}^{t} \big(c_s \log \sigma(\theta^\top z_s) + (1 - c_s)(1 - \sigma(\theta^\top z_s))\big) \ . \tag{4.1}$$

Let $\Delta_{ij} > 0$ lower bound the margin of comparison, $|\sigma(z_{ij}^\top \theta_*) - 1/2| > \Delta_{ij}$ for all $i, j \in \mathcal{I}$ and define $\Delta_* = \min_{i \neq j} \Delta_{ij}/|i - j|$. Next, let $\mathbf{H}_t(\theta) := \sum_{s=1}^{t} \dot{\sigma}(z_s^\top \theta) z_s z_s^\top$ be the Hessian of the negative log-likelihood of observations at time t under (2.2), given the parameter θ, also known as *observed Fisher information*. We define $\tilde{\mathbf{H}}_t(\theta) := \frac{1}{t}\mathbf{H}_t(\theta)$. For a square matrix V, we define $\|x\|_V = \sqrt{x^\top V x}$. We make the following assumptions for our analysis:

Assumption 4.1. θ_* satisfies $\|\theta_*\|_2 \leqslant S$ for some $S > 0$.

Assumption 4.2. $\forall i \in \mathcal{I}$, we have $\|x_i\|_2 \leqslant Q$ for $Q > 0$.

Assumption 4.3. $\mathbf{H}_T(\theta_T)$ and $\mathbf{H}_T(\theta_*)$ have full rank and minimum eigenvalues larger than $\lambda_0 > 0$.

Assumption 4.1 implies that θ_* lies in some ball with radius S and cannot have unbounded coefficients. Assumption 4.2 states that there exists an upper bound on the norm of the feature vectors. This assumption is trivially satisfied whenever we

have a finite set of data points. Both assumptions 4.1 and 4.2 are standard in the bandit literature and only required for our analysis. Assumption 4.3 is naturally satisfied for sufficiently large T by any sampling strategy with support on d linearly independent vectors, or can be ensured by allowing for a burn-in phase of d samples in the beginning of an adaptive strategy. Assumption 4.3 ensures the uniqueness of θ_t.

We start by stating the following concentration result for the deviation of $\sigma(z_{ij}^\top \theta_T)$ from the true probability $\sigma(z_{ij}^\top \theta_*)$. The proof of Lemma 4.1 is found in Appendix A and builds on results for optimistic algorithms in logistic multi-armed bandits (Filippi et al. 2010; Faury et al. 2020).

Lemma 4.1 (Concentration Lemma)**.** *Define, for all pairs of items $i, j \in \mathcal{I}$, and any $\Delta > 0$,*

$$\alpha_{ij}(\Delta) := \exp\left(\frac{-\Delta^2 T}{8dC_1(\dot\sigma(z_{ij}^\top\theta_T)\|z_{ij}\|_{\tilde{\mathbf{H}}_T^{-1}(\theta_T)})^2}\right), \quad \beta_{ij}(\Delta) := \exp\left(\frac{-\Delta T}{dC_1\|z_{ij}\|^2_{\tilde{\mathbf{H}}_T^{-1}(\theta_T)}}\right).$$

Then, if $\alpha := \alpha_{ij}(\Delta), \beta := \beta_{ij}(\Delta)$ and $\alpha, \beta \leqslant \frac{1}{4dT}$,

$$P\left(|\sigma(z_{ij}^\top\theta_*) - \sigma(z_{ij}^\top\theta_T)| > \Delta\right) \leqslant 2dT\,(\alpha + \beta)\,.$$

C_1 depends on S, λ_0, Q from Assumptions 4.1–4.3 (see Appendix A for definition and proof).

The concentration result in Lemma 4.1 is *verifiable* (given by observables) since the upper bound depends only on the maximum likelihood estimate θ_T at time T, not on θ_*. We present a sharper, *unverifiable* bound in Appendix A which instead depends on θ_* but does not suffer from the explicit scaling with d in the definitions of α and β. The bound in Lemma 4.1 can also be expressed in terms of $\mathbf{H}_T^{-1}(\theta_T)$ by using the equality $||z_{ij}||^2_{\mathbf{H}_T^{-1}(\theta_T)} = \frac{1}{T}||z_{ij}||^2_{\tilde{\mathbf{H}}_T^{-1}(\theta_T)}$. As long as our sampling strategy ensures that the minimum eigenvalue of $\tilde{\mathbf{H}}_t(\theta_t)$ does not tend to zero, i.e., the strategy is *strongly consistent* (Chen, Hu, et al. 1999), we have

$$\alpha_{ij}(\Delta_{ij}) \sim \exp[-\Delta_{ij}^2 T/(\dot\sigma(z_{ij}^\top\theta_T)^2\|z_{ij}\|^2_{\tilde{\mathbf{H}}_T^{-1}(\theta_T)})]$$

and

$$\beta_{ij}(\Delta_{ij}) \sim \exp[-\Delta_{ij} T/\|z_{ij}\|^2_{\tilde{\mathbf{H}}_T^{-1}(\theta_T)}].$$

. Since $\Delta_{ij}^2 < \Delta_{ij} < 1/2$ by definition, we can view α as the *first-order* term and β as the *second-order* term of our bound.

Lemma 4.1 formally captures the intuition that it should be easier to sort when annotations contain little noise, i.e., $\dot\sigma(z_{ij}^\top\theta_T)$ is small. Especially, we observe $\dot\sigma(z_{ij}^\top\theta_T) \approx 0$ for pairs where Δ_{ij} is sufficiently large, causing the first-order term to vanish, leaving us with the faster decaying second-order term β. Lemma 4.1 also tells us that the hardest pairs to guarantee a correct ordering for are the ones with both high *aleatoric* uncertainty under the MLE model, e.g., where annotators disagree or

labels are noisy, captured by $\dot{\sigma}(z_{ij}^\top \theta_T)$, as well as high *epistemic* uncertainty captured by $||z_{ij}||_{\tilde{\mathbf{H}}_T^{-1}(\theta_T)}$.

A direct consequence of Lemma 4.1 is the following bound on the ordering error of h_{θ_T} over $\mathcal{I}$,

$$\mathbb{E}[R(\theta_T)] \leqslant \sum_{i \neq j} \frac{2 \min\left\{2dT\left(\alpha_{ij}(\Delta_{ij}) + \beta_{ij}(\Delta_{ij})\right), 1\right\}}{n(n-1)}.$$

The right-hand side in the above inequality can be bounded further by utilizing that $\Delta_{ij} \geqslant |i-j|\Delta_*$. Together with Markov's inequality, this yields the following bound on $P(R(\theta_T) \geqslant \epsilon)$.

Theorem 4.2 (Upper bound on the ordering error)**.**
Let $\alpha_* := \max_{i \neq j} \alpha_{ij}(\Delta_*)$ and $\beta_* := \max_{i \neq j} \beta_{ij}(\Delta_*)$, with α, β from Lemma 4.1. Then, for $\alpha_*, \beta_* \leqslant \frac{1}{4dT}$ and any $\epsilon \in (0,1)$, the ordering error $R(\theta_T)$ satisfies

$$P(R(\theta_T) \geqslant \epsilon) \leqslant \frac{4dT}{\epsilon n}\left(\left(\alpha_*^{-1} - 1\right)^{-1} + \left(\beta_*^{-1} - 1\right)^{-1}\right) \approx \frac{4dT}{\epsilon n}\left(\alpha_* + \beta_*\right) \ ,$$

where α_* and β_* decay exponentially with T.

Theorem 4.2 suggests that the probability of $R(\theta_T) \geqslant \epsilon$ decays exponentially with a rate that depends on the quantities $\max_{i,j} \dot{\sigma}(z_{ij}^\top \theta_T)\|z_{ij}\|_{\tilde{\mathbf{H}}_t^{-1}(\theta_T)}$ and $\max_{i,j} \|z_{ij}\|^2_{\tilde{\mathbf{H}}_t^{-1}(\theta_T)}$. Both quantities are random variables that depend on the particular sampling strategy that yields $\mathbf{H}_t$. Focusing on the leading term, $\max_{i,j} \dot{\sigma}(z_{ij}^\top \theta_T)\|z_{ij}\|_{\tilde{\mathbf{H}}_t^{-1}(\theta_T)}$, Theorem 4.2 suggests that an active learner should gather data to minimize this quantity and obtain the smallest possible bound. The factor $\|z_{ij}\|^2_{\tilde{\mathbf{H}}_T^{-1}(\theta_T)}$ is the weighted norm of z_{ij} w.r.t. the inverse of the observed Fisher information (cf. Kirsch and Gal (2022)). It controls the shape of the confidence ellipsoid around θ_T and the width of the confidence interval around $\theta_T^\top z_{ij}$. The leading term in Theorem 4.2 re-scales this quantity with aleatoric noise under the MLE estimate θ_T. This suggests that higher epistemic (model) certainty is needed in directions with high aleatoric uncertainty—where item similarity increases noise in comparisons.

In Appendix A.3, we comment on i) generalizations to regularized preference models, ii) applications to generalized linear models with other link functions, iii) lower bounds on the ordering error, and iv) an algorithm-specific upper bound.

5 Greedy uncertainty reduction for ordering (GURO)

We present an active preference learning algorithm based on greedy minimization of the bound in Theorem 4.2, called GURO. We begin with fully contextual preference models of the form $\sigma(\theta^\top z_{ij})$ and return in Section 5.1 to parameterization variants to reduce the effects of model misspecification.

The main component of the bound in Theorem 4.2 to be controlled by an active learner is the term

$$\max_{i,j \in \mathcal{I}} \ \dot{\sigma}(z_{ij}^\top \theta_T)\|z_{ij}\|_{\mathbf{H}_T^{-1}(\theta_T)} \ , \tag{5.1}$$

Algorithm 7.1 **G**reedy **U**ncertainty **R**eduction for **O**rdering (GURO), [BayesGURO]

Require: Training items $\mathcal{I}_D$, attributes $\mathbf{X} = \{x_i\}_{i \in \mathcal{I}_d}$
1: Initialize θ_0
2: **for** $t = 1, ..., T$ **do**
3: Draw (i_t, j_t) based on θ_t according to (5.2) [or (5.4)]
4: Observe c_t from noisy comparison (annotator)
5: $D_t = D_{t-1} \cup \{i_t, j_t, c_t)\}$
6: $\theta_t = \text{MLE}(D_t)$ according to (4.1) [or $\theta_t = \text{MAP}(D_t)$ as in (A.2)] in the Appendix
7: **end for**
8: Return h_T

which represents the highest uncertainty in the comparison of any items $i, j \in \mathcal{I}$ under the model θ_T. A smaller value of (5.1) yields a smaller bound and a stronger guarantee. Recall that, for any $t = 1, ..., T$, θ_t is the MLE estimate of the ground-truth parameter θ_* with respect to the observed history D_t. Both factors in (5.1) are determined by the sampling strategy that yielded the item pairs (i_t, j_t) in D_T and, therefore, $\mathbf{H}_T$ and θ_T (the results of comparisons c_{ij} are outside the control of the algorithm, but z_{ij} are known).

Direct minimization of (5.1), for a subset $\mathcal{I}_D$, is not feasible without access to comparisons c_{ij} and their likelihood under θ_T. Instead, we adopt a greedy, alternating approach: In each round, a) a single pair is sampled for comparison by maximizing (5.1) under the current model estimate, and b) θ_t is recomputed based on D_t. Specifically, at $t = 1, ..., T$, we sample,

$$i_t, j_t = \underset{i,j \in \mathcal{I}_D, i \neq j}{\arg\max}\, \dot{\sigma}(z_{ij}^\top \theta_{t-1}) \|z_{ij}\|_{\mathbf{H}_{t-1}^{-1}(\theta_{t-1})} \,. \tag{5.2}$$

We refer to this sampling criterion as Greedy Uncertainty Reduction for Ordering (GURO), since it reduces the uncertainty of θ_t in the direction of z_{ij}. To see this, consider the change of $\mathbf{H}_t(\theta_t)$ after a single play of i_t, j_t. The Sherman-Morrison formula (Sherman and Morrison 1950) yields,

$$\mathbf{H}_t^{-1}(\theta_{t-1}) = \mathbf{H}_{t-1}^{-1}(\theta_{t-1}) - \dot{\sigma}(z_t^\top \theta_{t-1}) \frac{\mathbf{H}_{t-1}^{-1}(\theta_{t-1}) z_t z_t^\top \mathbf{H}_{t-1}^{-1}(\theta_{t-1})}{1 + \dot{\sigma}(z_t^\top \theta_{t-1}) \|z_t\|^2_{\mathbf{H}_{t-1}^{-1}(\theta_{t-1})}} \,, \tag{5.3}$$

where $z_t := z_{i_t j_t}$. With ξ as the second term in (5.3), it holds for all $i < j \in \mathcal{I}$, with $\mathbf{H}_{t-1} = \mathbf{H}_{t-1}(\theta_{t-1})$, that $||z_{ij}||^2_{\mathbf{H}_t^{-1}(\theta_{t-1})} = ||z_{ij}||^2_{\mathbf{H}_{t-1}^{-1}} - ||z_{ij}||^2_{\xi} \leqslant ||z_{ij}||^2_{\mathbf{H}_{t-1}^{-1}}$. The inequality is strict for the pair i_t, j_t in (5.2). As θ_t converges to θ_*, this pair becomes representative of the maximizer of (5.1) provided there is no major systematic discrepancy between $\mathcal{I}_D$ and $\mathcal{I}$.

Surprisingly, GURO can also be justified from a Bayesian analysis. Consider a Bayesian model of the parameter θ with $p(\theta)$ the prior belief and $p(\theta \mid D_t)$ the posterior after observing the preference feedback in D_t. A natural active learning strategy is to sample items i_t, j_t for which the model preference is highly uncertain

under the posterior distribution,

$$i_t, j_t = \underset{i,j \in \mathcal{I}_D, i<j}{\arg\max} \hat{\mathbb{V}}_{\theta|D_{t-1}}[\sigma(\theta^\top z_{ij})] \ , \tag{5.4}$$

where $\hat{\mathbb{V}}_{\theta|D_{t-1}}[\sigma(\theta^\top z_{ij})]$ is a finite-sample estimate of the variance in predictions, computed by sampling from the posterior. In Appendix A.3, we show that the first-order Taylor expansion of the true variance is equal to the GURO criterion. Hence, we refer to sampling according to (5.4) as BayesGURO. Unlike GURO, BayesGURO can incorporate prior knowledge through $p(\theta)$ and benefits from controlled stochasticity through the empirical estimate $\hat{\mathbb{V}}$, which makes it appropriate for batched algorithms—a deterministic criterion would construct batches of a single item pair. Both GURO and BayesGURO are presented in Algorithm 7.1.

Computational Complexity: Running the algorithms requires $O(n^2)$ operations each iteration to evaluate the sampling criteria (Equation 5.2 or 5.4) on all possible pairs, a problem shared by many active preference learning algorithms (Qian et al. 2015; Canal et al. 2019; Houlsby et al. 2011). A way of mitigating this computational complexity is to, at each time step, sample a fixed number of comparisons and only evaluate on these, similar to the approach taken in Canal et al. (2019). When only looking at a sample of $m \ll n^2$ pairs, the complexity is reduced to $O(m)$. While making m too small can hurt the sample complexity, we describe in Appendix A how we implemented this sub-sampling strategy to speed up computations in one of our experiments and observed no noticeable change in performance. Lastly, we want to highlight that in many realistic scenarios, the computational burden pales in comparison to the time it takes to query an annotator.

5.1 Preference models for in- and out-of-sample ordering

Our default preference model $h(i,j) = \mathbf{1}[f(i,j) > 0]$ is based on a *fully contextual* scoring function

$$f_\theta(x_i, x_j) = \theta^\top (x_i - x_j) \ , \tag{5.5}$$

fit with a logistic likelihood $\sigma(f(i,j)) \approx p(C_{ij} = 1)$. The model's strength is that the variance in its estimates grows with d, but not with $n = |\mathcal{I}|$, often resulting in quicker convergence than non-contextual methods for moderate dimension d (see, e.g., Figure 7.2c). The fully contextual model also generalizes to unseen items as long as the attributes for $\mathcal{I}_D$ span attributes observed for $\mathcal{I}$.

The limitations of a fully contextual model are model misspecification (error due to the functional form), and noise (error due to C not being fully determined by X). The former can be mitigated by applying the linear model to a representation function $\phi : \mathcal{X} \to \mathbb{R}^{d'}$, $f_\theta(x_i, x_j) = \theta^\top(\phi(x_i) - \phi(x_j))$. A good representation ϕ, e.g., from a foundation model, can mitigate misspecification and admit different input modalities. As demonstrated in Figure 7.5 in the Appendix, even a representation pre-trained for a different task can perform much better than a random initialization.[1]

[1]It is feasible to update representations during exploration (Xu et al. 2022; Singh and Chakraborty 2021), but we do not consider that here.

Noise due to insufficiencies in X cannot be mitigated by a representation $\phi(x)$; If annotators consistently compare items based on features U not included in X, no function $h(X_i, X_j)$ can perfectly order the items. However, for in-sample ordering of $\mathcal{I}_D$, adding per-item parameters $\zeta_i \in \mathbb{R}$ to the scoring function, one for each item $i \in \mathcal{I}_D$, can mitigate both misspecification and noise,

$$f_{\theta,\boldsymbol{\zeta}}(x_i, x_j) = \theta^\top(\phi(x_i) - \phi(x_j)) + (\zeta_i - \zeta_j) \ . \tag{5.6}$$

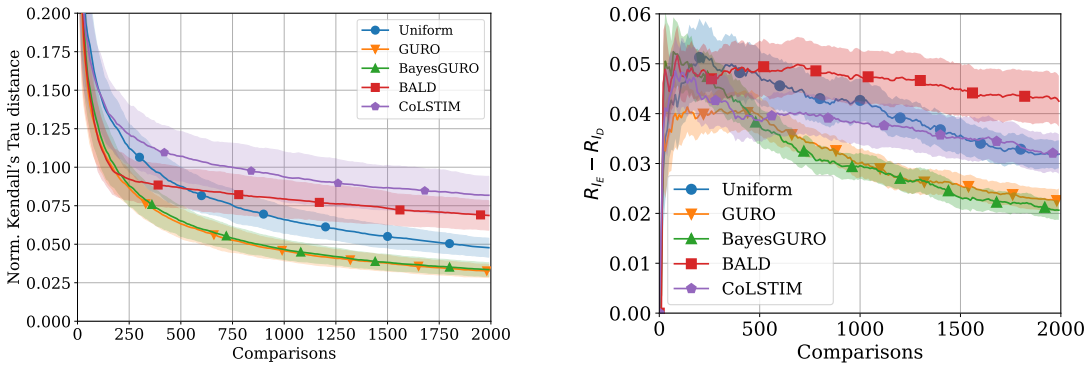
We call this a *hybrid* model and apply it in "GURO Hybrid" and baselines in experiments. The term $\zeta_i - \zeta_j$ can correct the residual of the fully contextual model, which is small if a) the context captures the most relevant information about the ordering, and b) the functional form $\theta^\top\phi(x_i)$ is nearly well-specified. Using $\zeta_i - \zeta_j$ alone is sufficient in-sample, but has high variance (the dimension is n instead of d) and poor generalization (ζ_i are unknown for items $i \notin \mathcal{I}_D$). In practice, we use L2 regularization to prevent the model from learning an arbitrary θ by using the full expressivity of ζ_i (see Appendix A for details). Empirically, our hybrid models exhibit the best of both worlds: When ϕ is poor, the model recovers and competes with non-contextual models (Figure 7.5); when ϕ is good, convergence matches fully contextual models (Figure 7.2).

6 Experiments

We evaluate GURO (Algorithm 7.1) and GURO Hybrid (see Section 5.1) in four image ordering tasks, one with logistic (synthetic) preference feedback, and three tasks based on real-world feedback from human annotators[2].We provide a synthetic experiment in Appendix A.2 that includes empirical estimates of the bound in Theorem 4.2. The experiments include five diverse baseline algorithms, described next. BALD (Houlsby et al. 2011) is *a priori* the strongest baseline since it is a contextual active learning algorithm, unlike the others. Its selection criterion greedily maximizes the decrease in posterior entropy, which amounts to reducing the epistemic uncertainty and includes a term to downplay the influence of aleatoric uncertainty. This is not always beneficial, as suggested by our analysis in Section 4, since learners may require several comparisons of high-uncertainty pairs to get the order right. CoLSTIM (Bengs, Saha, et al. 2022) is a contextual bandit algorithm, developed for regret minimization and is not expected to perform well here. It is included to illustrate the mismatch between regret minimization and our setting.

TrueSkill (Herbrich et al. 2006; Graepel 2012) is a non-contextual skill-rating system that models the score of each item as a Gaussian distribution, disregarding item attributes, and has been adopted in various works to score items based on subjective pairwise comparisons (Larkin et al. 2022; Naik et al. 2014; Sartori et al. 2015). We use the sampling rule from Hees et al. (2016), designed for ordering. Finally, we include Uniform sampling, and to illustrate the importance of accounting for aleatoric uncertainty, we use a version of GURO called NormMin that ignores the $\dot{\sigma}(z_{ij}^\top\theta_t)$ term and plays the pair maximizing $\|z_{ij}\|_{\mathbf{H}_t^{-1}(\theta_t)}$, i.e., it minimizes the

[2]Our code is available at: https://github.com/Healthy-AI/GURO



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

(b) Mean generalization error (95% CI)

Figure 7.1: **X-RayAge**. Performance of active sampling strategies when comparisons are simulated using a logistic model according to (2.2). In-sample Kendall's Tau distance 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.

second-order term in Lemma 4.1. NormMin corresponds to the selection criterion in the concurrent work Das et al. (2024), adapted to our problem of finding the correct ordering. We refer the reader to Appendix A.2 for a detailed comparison where NormMin performs significantly worse than Uniform on certain problem instances, and Appendix A for details regarding the implementation and the choice of hyperparameters for GURO, BayesGURO, and baselines.

6.1 Ordering X-ray images under the logistic model

Our first task (X-RayAge) is to order X-ray images based on perceived age (Ieki et al. 2022) where the preference feedback follows a (well-specified) logistic model. We base this experiment on the data from the Kaggle competition "X-ray Age Prediction Challenge" (Felipe Kitamura 2023) which contains more than 10 000 de-identified chest X-rays, along with the person's true age. Features were extracted using the 121-layer DenseNet in the TorchXrayVision package (Cohen et al. 2022) followed by PCA projection, resulting in 35 features. A ridge regression model, θ_*, was fit to the true age ($R^2 \approx 0.67$). During active learning, feedback is drawn from $p(C_{ij} = 1) = \sigma\left(\theta_*^\top z_{i,j} \cdot \lambda\right)$, where λ (set to 0.1) controls the noise level. We only include the fully contextual models here since they are well-specified by design, meaning $\mathcal{I}$ can be ordered using only contextual features.

In the first setting, we sub-sample 200 X-ray images uniformly at random from the full set. A ground-truth ordering of these elements is derived using the learned linear model. Figure 7.1a shows the ordering error over 2 000 iterations. GURO and BayesGURO perform similarly, both better than the baselines. BALD starts off converging about as fast as GURO, but plateaus, most likely as a result of actively avoiding comparisons with high aleatoric uncertainty—pairs where annotators disagree in their preferences. The poor performance of CoLSTIM highlights the discrepancy between regret minimization and recovering a complete ordering.

Table 7.1: Datasets with preference feedback from annotators. Pretrained models are ResNet34 (He et al. 2016), all-mpnet-base-v2 (Reimers and Gurevych 2019), and FaceNet (Schroff et al. 2015).

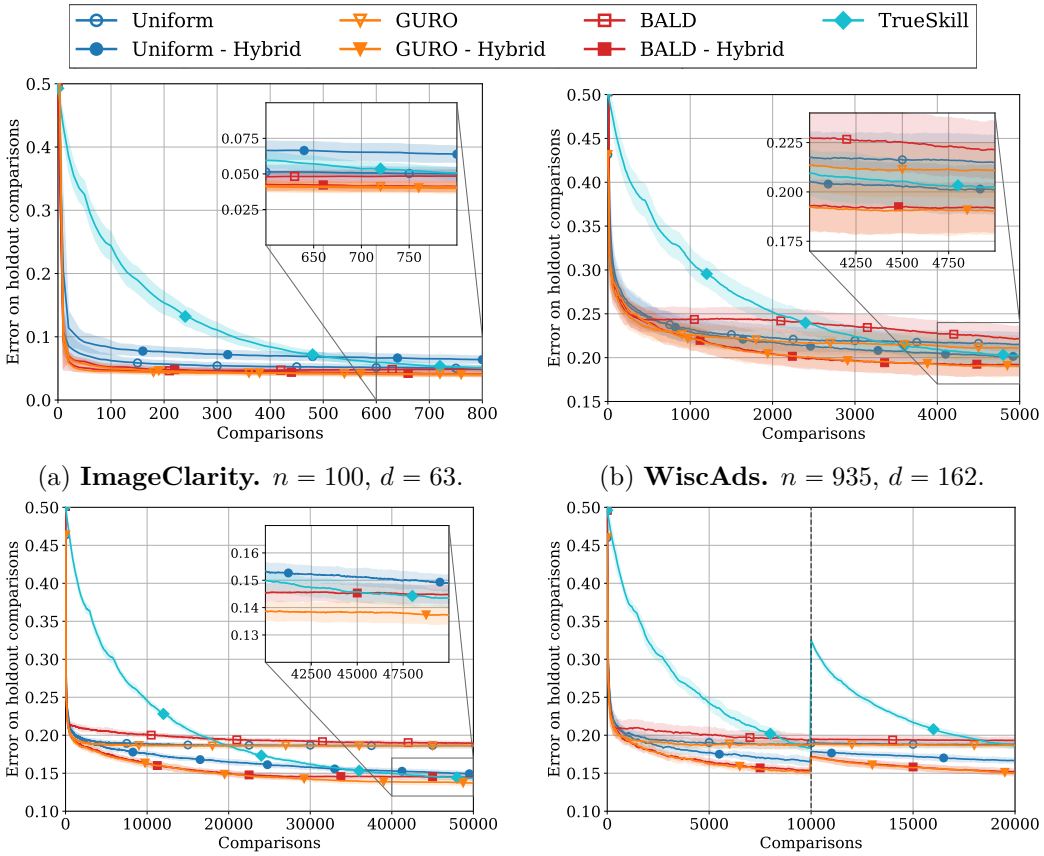
Dataset	n	d	#comparisons	Data type	Embedding Model
ImageClarity	100	63	27 730	Image	ResNet34 (Imagenet)
WiscAds	935	162	9 528	Text	all-mpnet-base-v2
IMDB-WIKI-SbS	6072	75	110 349	Image	FaceNet (CASIA-Webface)

In the second setting, we evaluate how well the algorithms generalize to new items. First, we sample 300 X-ray images from the full dataset. Next, we split these into two sets, with one (I_D) containing the youngest 50% and the other (I_E) the oldest 50%. The algorithms were then trained to order the list containing the younger subjects, but were simultaneously evaluated on how well they could sort the list containing the older subjects. The continuously measured difference in ordering error evaluated on I_E and I_D are presented in Figure 7.1b. While all algorithms are worse at ordering items in I_E, GURO and BayesGURO achieve the lowest average difference. Together with Figure 7.1a, this means that our proposed algorithms achieved the best in-sample and out-of-sample orderings. For completeness, the in-sample performance of algorithms in the generalization experiment in Figure 7.1b are included in Appendix A.2.

6.2 Ordering items with human preference data

Next, we evaluate our algorithm on three publicly available datasets to study the algorithms' performance when preference feedback comes from human annotators (see Table 7.1 for an overview, detailed information of datasets in Appendix A.1). The datasets are IMDB-WIKI-SbS (Pavlichenko and Ustalov 2021), where annotators have stated which of two people appear older, ImageClarity (Zhang et al. 2016), where modified versions of the same image have been compared according to the level of distortion, as well as the extended WiscAds dataset (Carlson and Montgomery 2017), where labels correspond to which political advertisement is perceived as more negative toward an opponent. In all datasets, pairs of items were sampled uniformly for annotation. For each experiment, we construct a feature vector $\phi(x_i) \in \mathbb{R}^d$ for all n items using a pre-trained embedding model followed by PCA, applied to reduce computational complexity. We restrict algorithms to only query pairs for which an annotation exists and remove the annotation from the pool once queried. In cases where multiple annotations exist for the same pair, the feedback is chosen randomly among these.

The images in the ImageClarity dataset have been constructed to have an objective ground truth ordering but this is not the case for WiscAds or IMDB-WIKI-SbS. As the ground-truth ordering is generally unknown also in real-world applications, we evaluate methods by the error on a held-out set of comparisons D', $\hat{R}_{D'}(h) = \frac{1}{|D'|}\sum_{(i,j,c)\in D'} \mathbf{1}[h(i,j) \neq c]$. This serves as an empirical analog of Kendall's Tau distance and a minimizer of $\hat{R}_{D'}(h)$ will minimize $R_{\mathcal{I}}(h)$ for sufficiently large D', but



(a) **ImageClarity.** $n = 100$, $d = 63$.

(b) **WiscAds.** $n = 935$, $d = 162$.

(c) **IMDB-WIKI-SbS.** $n = 6\ 072$, $d = 75$.

(d) **IMDB-WIKI-SbS.** $n = 3\ 000$ (6 072).

Figure 7.2: The empirical error $\hat{R}_{D'}(h)$ on a holdout comparison set D' when comparisons are made by human annotators. The plots are averaged over 100 (a,b) or 10 (c,d) seeds, and the shaded area represents one standard deviation above and below the mean. For every seed, 10% of comparisons were used for the holdout set. In (d) we initially order a list $\mathcal{I}_D$ of 3 000 images. After 10 000 comparisons the remaining 3 072 images, $\mathcal{I} \backslash \mathcal{I}_D$, are added.

will not converge toward 0 since there is inherent noise in annotations. This metric makes no assumptions on the ground truth ordering unlike the alternative approach of fitting an ordering to all available comparisons, see e.g., Maystre and Grossglauser (2017). In Appendix A.2, we show results for the latter that highlight the limitations of estimating a "ground-truth" ordering, as well as the similar results when measuring the distance to the objective ground-truth ordering of the ImageClarity dataset. The longest trajectory (single seed) for any algorithm took less than 35hrs to complete on one core of an Intel Xeon Gold 6130 CPU and required at most 10 GB of memory.

In all experiments, we compare fully contextual (5.5) and hybrid (5.6) versions of GURO, BALD, and Uniform, as well as TrueSkill. The results of each experiment can be seen in Figure 7.2. Figure 7.2a shows that the ImageClarity dataset is the easiest to order using contextual (non-hybrid) features. This is expected, as features relevant to the level of distortion are low-level. In this case, the choice of adaptive strategy has a modest impact on the ordering error. Figures 7.2b and 7.2c highlight the differences between modeling strategies. The fully contextual algorithms initially

improve rapidly, achieving a rough ordering of the items, before plateauing and not making any real improvements. This indicates that the features are informative enough to roughly order the list, but insufficient for retrieving a more granular ordering. The non-contextual TrueSkill converges at a much slower pace but keeps improving steadily throughout. Perhaps most interesting are the hybrid algorithms, which seemingly reap the benefits of both methods, improving as quickly as the contextual methods, but avoiding the plateau. In fact, in Figure 7.5 in the Appendix we show that the hybrid models perform comparably to TrueSkill even when features are completely uninformative.

The limitations of BALD are most noticeable in the fully contextual case, where it plateaus at a higher error compared to GURO and Uniform. This is however not as prominent when we use BALD in conjunction with our hybrid model, likely a result of the increased dimensionality of the model causing BALD Hybrid to attribute more of the observed errors to model uncertainty. While this initially causes the algorithm to avoid fewer comparisons that are subject to aleatoric uncertainty, the final iterations in Figure 7.2c suggest that BALD Hybrid can still run into this issue given enough samples. In all experiments, GURO and GURO Hybrid perform better than or similar to our baselines, never worse. Additionally, Figures 7.2b and 7.2c showcase how our hybrid model can increase performance when used with existing sampling strategies, such as BALD or Uniform.

The final experiment, visible in Figure 7.2d, is a few-shot scenario where after some time, additional images are added to the pool of items. IMDB-WIKI-SbS was used as it contained the highest number of both images and comparisons. The initial pool consists of 3 000 images sampled from the dataset. After 10 000 steps, the remaining 3 072 images were added to the pool. The results again emphasize the differences between our three types of models; the increase in error of the fully contextual model is very slight, likely a result of added samples being drawn from the same distribution. For TrueSkill, the error increases drastically as a result of the algorithm not having seen these items before and having no way of generalizing the results of previous comparisons to them. Lastly, the hybrid algorithms seem to be moderately affected. The error increases as the model has not yet tuned any of the added per-item parameters, but the extent is much smaller than for TrueSkill as the model can provide a rough ranking of the out-of-sample elements using the contextual features.

7 Conclusion

We have demonstrated the benefits of utilizing contextual features in active preference learning to efficiently order a list of items. Empirically, this leads to quicker convergence, compared to non-contextual methods, and allows algorithms to generalize out-of-sample. We derived an upper bound on the ordering error and used it to design an active sampling strategy that outperforms or matches baselines on realistic image and text ordering tasks. Both theoretical and empirical results highlight the benefit of accounting for noise in comparisons when learning from human annotators.

The optimality of our sampling strategy remains an open question. A future direction is to derive a lower bound on the ordering error, and prove an—ideally matching—algorithm-specific upper bound. However, constructing upper bounds for related fixed-budget tasks is an open problem (Qin 2022). Moreover, motivated by the annotation setting, our focus has been on reducing sample complexity and we leave it to future work to explore potential linear approximations of the sampling criteria and other trade-offs between sample complexity and computational complexity. Further, our approach can potentially be improved by performing representation learning throughout the learning process. Finally, our experiments are constrained to a limited amount of already-collected (offline) human preference data, causing different algorithms to select disproportionately similar comparisons. Future work should evaluate the strategies in an online setting.

Acknowledgements

FDJ and HB are supported by Swedish Research Council Grant 2022-04748. FDJ is also supported in part by the Wallenberg AI, Autonomous Systems and Software Program founded by the Knut and ALice Wallenberg Foundation. EC and DD are supported by Chalmers AI Research Centre (CHAIR). The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

A Notation

Table 7.2: Notation

$\mathcal{I}$	Collection of items $\mathcal{I} = \{1, ..., n\}$
n	Number of items
d	Dimension of item attributes
$x_i \in \mathbb{R}^d$	Context attributes for item $i \in \mathcal{I}$
$z_{ij} \in \mathbb{R}^d$	$z_{ij} = x_i - x_j$ for $i, j \in \mathcal{I}$
y_i	Score for item $i \in \mathcal{I}$
$c_t \in \{0, 1\}$	The outcome of the comparison at time t, 1 if i_t was preferred to j_t
D_t	$D_t = ((i_1, j_1, c_1), ..., (i_t, j_t, c_t))$
$\theta \in \mathbb{R}^d$	Model parameter
$\theta_* \in \mathbb{R}^d$	Model parameter of the environment
$\theta_t \in \mathbb{R}^d$	Estimated parameter at time t
$\sigma(.)$	Sigmoid (logistic) function
$\dot{\sigma}(.)$	derivative of $\sigma(.)$
$\mathbf{H}_t(\theta)$	Hessian of the negative log-likelihood $\mathbf{H}_t(\theta) := \sum_{s=1}^{t} \dot{\sigma}(z_s^\top \theta) z_s z_s^\top$
$\tilde{\mathbf{H}}_t(\theta)$	Hessian normalized by number of plays $\tilde{\mathbf{H}}_t(\theta) := \frac{1}{t}\mathbf{H}_t(\theta)$
$\theta_{B,t} \in \mathbb{R}^d$	The MAP estimate of θ at time t
$\mathbf{H}_{B,t}$	The Hessian in the Bayesian setting, adjusted by the prior covariance $\mathbf{H}_{B,0}^{-1}$
$\|z_{ij}\|_{\mathbf{H}_t^{-1}(\theta)}$	$\|z_{ij}\|_{\mathbf{H}_t^{-1}(\theta)} = \sqrt{z_{ij}^\top \mathbf{H}_t^{-1}(\theta) z_{ij}}$
h	Comparison model (binary output)
f	Comparison logit (typically linear), e.g., $f_\theta(i, j) = \theta^\top (x_i - x_j)$

A Algorithms

A.1 MLE estimator for logistic regression

The log-likelihood $L_t(\theta)$ of data $D_t = \{(i_s, j_s, c_s)\}_{s=1}^t$, with $z_s = x_{i_s} - x_{j_s}$, under a logistic regression model with parameters θ is defined by

$$L_t(\theta) = \sum_{s=1}^{t} \left(c_s \log \sigma(\theta^\top z_s) + (1 - c_s)(1 - \sigma(\theta^\top z_s)) \right) \ .$$

The maximum likelihood estimator (MLE) at time t is the parameters

$$\theta_t = \arg\max_\theta L_t(\theta) \ . \tag{A.1}$$

The regularized estimator with ridge/ℓ_2 penalty with parameter λ is

$$\theta_t^R = \arg\min_\theta -L_t(\theta) + \lambda \|\theta\|_2^2 \ .$$

A.2 Bayesian estimator for logistic regression

$\theta_{B,t}$ is the MAP estimate of θ at time t according to the log likelihood

$$\theta_{B,t} = \arg\max_\theta \ln p(\theta \mid D_t), \tag{A.2}$$

where

$$\begin{aligned} \ln p(\theta \mid D_t) = &- \frac{1}{2}(\theta - \theta_{B,0})^\top \mathbf{H}_{B,0}^{-1}(\theta - \theta_{B,0}) \\ &+ \sum_t c_t \ln(\sigma(z_{i_t,j_t}^\top \theta)) + (1 - c_t) \ln(1 - \sigma(z_{i_t,j_t}^\top \theta)) + const. \end{aligned}$$

The hessian at time t is defined as

$$\mathbf{H}_{B,t} = \mathbf{H}_{B,0} + \sum_{(i,j) \in D_t} \dot{\sigma}(z_{i,j}^\top \theta_{B,t}) z_{i,j} z_{i,j}^\top = \mathbf{H}_{B,0} + \mathbf{H}_t.$$

Moreover, if priors $\theta_{B,0} = \mathbf{0}$ and $\mathbf{H}_{B,0}^{-1} = I_d$ are used, the log likelihood boils down to:

$$\ln p(\theta \mid D_t) = -\frac{1}{2}||\theta||_2^2 + \sum_t c_t \ln(\sigma(z_{i_t,j_t}^\top \theta)) + (1 - c_t) \ln(1 - \sigma(z_{i_t,j_t}^\top \theta)) + const$$

which implies that the MAP estimate will be the same as the MLE estimate with ridge regularisation in the frequentist setting. Similarly, the Hessian becomes:

$$\mathbf{H}_{B,t} = \mathbf{H}_t + I_d$$

Sequential updates are also possible in the Bayesian setting by using your current estimates as the new priors. Note that this will give slightly different results, as the calculation of $\mathbf{H}_{B,t}$ depends on the current estimate of $\theta_{B,t}$.

A.3 Stochastic Bayesian uncertainty reduction (BayesGURO)

We describe BayesGURO, a Bayesian sampling criterion, closely related to GURO. Consider a Bayesian model of the parameter θ with $p(\theta)$ the prior belief and $p(\theta \mid D_t)$ the posterior after observing the preference feedback in D_t. A natural strategy for learning more about the ordering of $\mathcal{I}$ is to sample items i_t, j_t based on an estimate of the posterior variance of predictions for their comparison,

$$i_t, j_t = \underset{i,j \in \mathcal{I}_D, i<j}{\arg\max}\, \hat{\mathbb{V}}_{\theta|D_{t-1}}[\sigma(\theta^\top z_{ij})] \ . \tag{A.3}$$

Here, $\hat{\mathbb{V}}_{\theta|D_t}[\sigma(\theta^T z_{ij})]$ is an estimate of the variance of probabilities $\sigma(\theta^T z_{ij})$, computed from finite samples drawn from the posterior of θ. Estimating the variance in this way both i) allows for tractable implementation, and ii) induces controlled stochasticity in the selection of item pairs. This can be useful in batched learning settings so that multiple pairs can be sampled within the same batch. A deterministic criterion would return the same item pair every time until θ is updated. We refer to the sampling criterion in (5.4) as BayesGURO.

For the logistic model considered in Section 4, using Laplace approximation with a Normal prior $\mathcal{N}(0, \mathbf{H}_{B,0}^{-1})$ on θ, the Bayesian criterion in (5.4) is related to the GURO sampling criterion in (5.2) through the first-order Taylor expansion of the variance:

$$\mathbb{V}_{\theta|D_t}(\sigma(\theta^\top z_{ij})) \approx (\dot{\sigma}(\mathbb{E}_{\theta|D_t}[\theta^\top z_{ij}]))^2 \mathbb{V}_{\theta|D_t}[\theta^\top z_{ij}] = (\dot{\sigma}(\theta_{B,t}^\top z_{ij}) ||z_{ij}||_{\mathbf{H}_{B,t}^{-1}(\theta_{B,t})})^2 \ ,$$

where $\theta_{B,t}$ is the MAP estimate of θ at time t and $\mathbf{H}_{B,t}$ is the Hessian adjusted by the prior covariance $\mathbf{H}_{B,0}^{-1}$ (further described in Appendix A.2). Thus, to a first-order approximation, for a large number of posterior samples, the GURO and BayesGURO active learning criteria are equivalent, save for the influence of the prior. In practice, we find that the Bayesian variant lends itself well to sequential updates of the posterior. The choice of prior $p(\theta)$, which could be useful under strong domain knowledge, and the stochasticity of using few posterior samples to approximate $\mathbb{V}$ make the two criteria distinct.

A.4 Uniform sampling

The uniform sampling algorithm is given in Algorithm 7.2. The corresponding Bayesian version replaces line 5 with the MAP estimate.

Algorithm 7.2 Uniform sampling algorithm

Require: Training items $\mathcal{I}_D$, attributes $\mathbf{X} = \{x_i\}_{i \in \mathcal{I}_d}$
1: **for** $t = 1, ..., T$ **do**
2: Sample (i_t, j_t) uniformly
3: Observe c_t from noisy comparison (annotator)
4: $D_t = D_{t-1} \cup \{i_t, j_t, c_t)\}$
5: Let $\theta_t = \text{MLE}(D_t)$
6: **end for**
7: Return h_T

Algorithm 7.3 BALD bandit

Require: Training items $\mathcal{I}_D$, attributes $\mathbf{X} = \{x_i\}_{i \in \mathcal{I}_d}$
1: Initialize $\theta_{B,0} = \mathbf{0}$, $\mathbf{H}_{B,0} = \lambda^{-1} I$
2: **for** $t = 1, ..., T$ **do**
3: Draw $(i_t, j_t) = \arg\max_{i,j} H[y \mid z_{i,j}, D_{t-1}] - \mathbb{E}_{\theta \sim p(\theta|D_{t-1})}[H[y \mid z_{i,j}, \theta]]$
4: Observe c_t from noisy comparison (annotator)
5: $D_t = D_{t-1} \cup \{i_t, j_t, c_t)\}$
6: Let $\theta_t = \text{MAP}(D_t)$
7: Update $\mathbf{H}_{B,t} \leftarrow \mathbf{H}_{B,0} + \sum_{(i,j) \in D_t} \dot{\sigma}(z_{i,j}^\top \theta_t) z_{i,j} z_{i,j}^\top$
8: **end for**
9: Return h_T

A.5 BALD

Where the posterior is calculated as in Appendix A.2 and $H[y \mid z_{i,j}, D_{t-1}] - \mathbb{E}_{\theta \sim p(\theta|D_{t-1})}[H[y \mid z_{i,j}, \theta]]$ is approximated as in Appendix A.5.

Deriving the BALD sampling criterion

The BALD criteria formalized using our notation becomes

$$\arg\max_{i,j} H[y \mid z_{i,j}, D_t] - \mathbb{E}_{\theta \sim p(\theta|D_t)}[H[y \mid z_{i,j}, \theta]],$$

where H represents Shannon's entropy

$$h(p) = -p \log_2(p) - (1-p) \log_2(1-p).$$

The first term of the equation becomes

$$H[y \mid z_{ij}, D_t] = h(\mathbb{P}\left[(\right] y \mid z_{i,j}, D_t)) = h\left(\int \mathbb{P}\left[(\right] y \mid z_{i,j}, \theta) \mathbb{P}\left[(\right] \theta \mid D_t) d\theta\right).$$

Here $\mathbb{P}\left[(\right] y \mid z_{i,j}, D_t)$ is the predictive distribution for our Bayesian logistic regression model. As covered in Chapter 4 of Bishop and Nasrabadi (2006), this expectation cannot be evaluated analytically but can be approximated using the probit function Φ;

$$\mathbb{P}\left[(\right] y \mid z_{ij}, D_t) \approx \Phi\left(\frac{\theta_t^\top z_{i,j}}{\sqrt{\lambda^{-2} + ||z_{ij}||^2_{\mathbf{H}_t^{-1}}}}\right) \approx \sigma\left(\frac{\theta_t^\top z_{i,j}}{\sqrt{1 + \frac{\pi ||z_{ij}||^2_{\mathbf{H}_t^{-1}(\theta_*)}}{8}}}\right).$$

Next, the term $\mathbb{E}_{\theta \sim p(\theta|D_t)}[H[y \mid z_{i,j}, \theta]]$ must be calculated. The true definition is

$$\mathbb{E}_{\theta \sim p(\theta|D_t)}[H[y \mid z_{i,j}, \theta]] = \int h(\sigma(\theta^\top z_{i,j})) \mathcal{N}(\theta \mid \theta_t, \mathbf{H}_t^{-1}) d\theta.$$

To make this a one variable integral, let $X = \theta^\top z_{i,j}$ define a new random variable. Since $\theta \sim \mathcal{N}(\theta_t, \mathbf{H}_t^{-1})$, and $z_{i,j}$ is just a constant vector, we know that X will follow a

univariate normal distribution $X \sim \mathcal{N}(\theta_t^\top z_{i,j}, ||z_{ij}||^2_{\mathbf{H}_t^{-1}})$. This allows us to rewrite the integral as

$$\int h(\sigma(\theta^T \mathbf{z}))\mathcal{N}(\theta \mid \theta_t, \mathbf{H}_t^{-1})d\theta = \int h(\sigma(x))\mathcal{N}(\theta_t^\top z_{i,j}, ||z_{ij}||^2_{\mathbf{H}_t^{-1}})dx.$$

However, this integral has no closed form solution. Instead we perform the same strategy as in Houlsby et al. (2011) and do a Taylor expansion of $\ln h(\sigma(\theta^\top \mathbf{z}))$. The third-order Taylor expansion gives us

$$h(\sigma(x)) \approx \exp\left(-\frac{x^2}{8\ln 2}\right).$$

Inserting this, the term can be approximated as

$$\begin{aligned}\int h(\sigma(x))\mathcal{N}(x \mid \theta_t^\top z_{i,j}, ||z_{ij}||^2_{\mathbf{H}_t^{-1}})dx &\approx \int \exp\left(-\frac{x^2}{8\ln 2}\right)\mathcal{N}(x \mid \theta_t^\top z_{i,j}, ||z_{ij}||^2_{\mathbf{H}_t^{-1}})dx \\ &= \frac{C}{\sqrt{||z_{ij}||^2_{\mathbf{H}_t^{-1}} + C^2}} \exp\left(-\frac{(\theta_t^\top z_{i,j})^2}{2(||z_{ij}||^2_{\mathbf{H}_t^{-1}} + C^2)}\right),\end{aligned}$$

where $C = \sqrt{4\ln 2}$. Finally, we arrive at an estimation of the objective function we wish to maximize:

$$\begin{aligned}H[y \mid z_{i,j}, D_t] - \mathbb{E}_{\theta\sim p(\theta|D_t)}[H[y \mid z_{i,j}, \theta]] \approx\ & h\left(\sigma\left(\frac{\theta_t^\top z_{i,j}}{\sqrt{1 + \frac{\pi}{8}||z_{ij}||^2_{\mathbf{H}_t^{-1}}}}\right)\right) \\ & - \frac{C}{\sqrt{||z_{ij}||^2_{\mathbf{H}_t^{-1}} + C^2}} \exp\left(-\frac{(\theta_t^\top z_{i,j})^2}{(||z_{ij}||^2_{\mathbf{H}_t^{-1}} + C^2)}\right)\end{aligned}$$

A Proofs of Lemma 4.1 and Theorem 4.2

A.1 Proof of Lemma 4.1

Proof. We now proceed to bound

$$P\left(|\sigma(z_{ij}^\top\theta_t) - \sigma(z_{ij}^\top\theta_*)| > \Delta\right).$$

From the self-concordant property of logistic regression we have (Faury et al. 2020)

$$|\sigma(z_{ij}^\top\theta_t) - \sigma(z_{ij}^\top\theta_*)| \leqslant \dot{\sigma}(z_{ij}\top\theta_t)|z_{ij}^\top(\theta_t - \theta_*)| + \frac{1}{4}|z_{ij}^\top(\theta_t - \theta_*)|^2.$$

We will prove a high probability bound on the event

$$\dot{\sigma}(z_{ij}\top\theta_t)|z_{ij}^\top(\theta_t - \theta_*)| + \frac{1}{4}|z_{ij}^\top(\theta_t - \theta_*)|^2 \leqslant \Delta. \tag{A.1}$$

Directly trying to bound the LHS in Equation A.1 will result in a rather messy expression. Instead, we define the events

$$\mathcal{E}_1 := \left\{\dot{\sigma}(z_{ij}\top\theta_t)|z_{ij}^\top(\theta_t - \theta_*)| \leqslant \frac{\Delta}{2}\right\}$$
$$\mathcal{E}_2 := \left\{\frac{1}{4}|z_{ij}^\top(\theta_t - \theta_*)|^2 \leqslant \frac{\Delta}{2}\right\}.$$

Clearly $\mathcal{E}_1 \bigcup \mathcal{E}_2$ implies the expression in Equation A.1. Assume we have bounds on the complement of these events, $P(\mathcal{E}_1^c) \leqslant \alpha$ and $P(\mathcal{E}_2^c) \leqslant \beta$. Then

$$\begin{aligned} P\left(|\sigma(z_{ij}^\top\theta_t) - \sigma(z_{ij}^\top\theta_*)| > \Delta\right) &\leqslant \alpha + \beta + \alpha\beta \\ &\leqslant 2\alpha + 2\beta. \end{aligned}$$

We now proceed to bound the probability of these complements separately.

Step 1. Relating θ_t to θ_*: The first challenge in our analysis to is relate θ_* and θ_t. In contrast to linear regression, where we have a closed-form expression for θ_t, there is no analytical solution for θ_t given a set of observation. However, we know that θ_t is the MLE, corresponding to

$$\theta_t = \arg\max_\theta L_t(\theta)$$

where

$$L_t(\theta) = \sum_{s=1}^{t} c_s \log\sigma\left(z_s^\top\theta\right) + (1 - c_s)\log\left(1 - \sigma\left(z_s^\top\theta\right)\right).$$

We have

$$\nabla_\theta L_t(\theta) = \sum_{s=1}^{t} c_s z_s - \underbrace{\sum_{s=1}^{t}\sigma\left(z_s^\top\theta\right) z_s}_{g_t(\theta)}$$

and hence $g_t(\theta_t) = \sum_{s=1}^{t} c_s z_s$.

A standard trick in logistic bandits (Filippi et al. 2010; Faury et al. 2020; Jun et al. 2021) is to relate $\theta_* - \theta_t$ to $g_t(\theta_*) - g_t(\theta_t)$. Especially, the following equality is due to the mean-value theorem (see Filippi et al. (2010))

$$g_t(\theta_*) - g_t(\theta_t) = \mathbf{H}_t(\theta') \left(\theta_* - \theta_t\right) \tag{A.2}$$

where θ' is some convex combination of θ_*, θ_t. Note that $\mathbf{H}_t(\theta')$ has full rank.

Using Equation A.2 yields

$$\left|z_{ij}^\top \left(\theta_* - \theta_t\right)\right| = \left|z_{ij}^\top \mathbf{H}_t^{-1}(\theta') \left(g_t(\theta_*) - g_t(\theta_t)\right)\right|$$

Furthermore, since $g_t(\theta_t) = \sum_{s=1}^t c_s z_s$, due to $\nabla_\theta L_t(\theta_t) = 0$, we have

$$g_t(\theta_t) - g_t(\theta_*) = \sum_{s=1}^t \underbrace{\left(c_s - \sigma\left(z_s^\top \theta_*\right)\right)}_{\epsilon_s} z_s$$

where ϵ_s is a sub-Gaussian random variable with mean 0 and variance $\nu_s^2 := \dot{\sigma}\left(z_s^\top \theta_*\right)$. We define

$$S_t := \sum_{s=1}^t \epsilon_s z_s.$$

We now have

$$\left|z_{ij}^\top \left(\theta_* - \theta_t\right)\right| = \left|z_{ij}^\top \mathbf{H}_t^{-1}(\theta') S_t\right|$$

and Lemma 10 in Faury et al. (2020) states that $\mathbf{H}_t^{-1}(\theta') \preccurlyeq (1 + 2S)\mathbf{H}_t^{-1}(\theta_*)$ where $||\theta_*||_2 \leqslant S$. Hence,

$$\left|z_{ij}^\top \left(\theta_* - \theta_t\right)\right| \leqslant (1 + 2S) \left|z_{ij}^\top \mathbf{H}_t^{-1}(\theta_*) S_t\right|$$

Step 2. Tail bound for vector-valued martingales:

We will now prove an upper bound on the probability that $\left|z_{ij}^\top \mathbf{H}_t^{-1}(\theta_*) S_t\right|$ deviates much from a certain threshold. This step is based on the proof of Lemma 1 in Filippi et al. (2010) which itself is based on a derivation of a concentration inequality in Rusmevichientong and Tsitsiklis (2010). The difference compared to Filippi et al. (2010) is that we work with the Hessian $\mathbf{H}_t(\theta_*)$ instead of the design matrix for linear regression $V_t = \sum_s x_s x_s^\top$. This require us to construct a slightly different martingale.

Let A and B are two random variables such that

$$\mathbb{E}\left[\exp\left\{\gamma A - \frac{\gamma^2}{2} B^2\right\}\right] \leqslant 1, \forall \gamma \in \mathbb{R} \tag{A.3}$$

then due to Corollary 2.2 in Peña et al. (2004) it holds that $\forall a \geqslant \sqrt{2}$ and $b > 0$

$$P\left(|A| \geqslant a\sqrt{(B^2 + b)\left(1 + \frac{1}{2}\log\left(\frac{B^2}{b} + 1\right)\right)}\right) \leqslant \exp\left\{\frac{-a^2}{2}\right\}. \tag{A.4}$$

Let $\eta \in \mathbb{R}^d$ and consider the process

$$M_t^{\gamma}(\theta_*, \eta) := \exp\left\{\gamma \eta^\top S_t - \gamma^2 ||\eta||^2_{\mathbf{H}_t(\theta_*)}\right\}. \tag{A.5}$$

We will now proceed to prove that $M_t^{\gamma}(\theta, \eta)$ is a non-negative super martingale satisfying Equation A.3. Note that

$$\gamma \eta^\top S_t - \gamma^2 ||\eta||^2_{\mathbf{H}_t(\theta_*)} = \sum_{s=1}^{t} \underbrace{\left(\gamma \eta^\top z_s \epsilon_s - \dot{\sigma}(\theta^\top z_s)\gamma^2 \left(\eta^\top z_s\right)^2\right)}_{F_s} = \sum_{s=1}^{t} F_s.$$

Further we use the fact that ϵ_s is sub-Gaussian with parameter ν_s, .i.e,

$$\mathbb{E}\left[\exp\{\lambda \epsilon_s\}\right] \leqslant \exp\left\{\nu_s^2 \lambda^2\right\}, \forall \lambda > 0.$$

Let D_{s-1} denote the observations up until time s, then

$$\begin{aligned}\mathbb{E}\left[\exp\{F_s\} \mid D_{s-1}\right] &= \mathbb{E}\left[\exp\left\{\underbrace{\gamma \eta^\top z_s}_{\lambda} \epsilon_s\right\}\right] \exp\left\{-\underbrace{\dot{\sigma}(\theta_t^\top z_s)}_{\nu_s^2} \gamma^2 \left(\eta^\top z_s\right)^2\right\} \\ &\leqslant \exp\left\{\nu_s^2 \lambda^2\right\} \exp\left\{-\nu_s^2 \lambda^2\right\} = 1.\end{aligned}$$

This also implies

$$\mathbb{E}\left[M_t^{\gamma}(\theta_*, \eta) \mid D_{t-1}\right] \leqslant M_{t-1}^{\gamma}(\theta_*, \eta)$$

and $M_t^{\gamma}(\theta_*, \eta)$ is a super-martingale satisfying

$$\mathbb{E}\left[\exp\left\{\gamma \eta^\top S_t - \gamma^2 ||\eta||^2_{\mathbf{H}_t(\theta_*)}\right\}\right] \leqslant 1, \forall \gamma \geqslant 0$$

and we can apply the results of Peña et al. (2004).

We now follow the last step of the proof of Lemma 1 in Filippi et al. (2010). We let $a = \sqrt{2 \log \frac{1}{\delta}}$ for some $\delta \in (0, 1/e)$ and let $b = \lambda_0 \|\eta\|_2^2$. We have with probability at least $1 - \delta$

$$|\eta^\top S_t| \leqslant \sqrt{2 \log \frac{1}{\delta}} \sqrt{\|\eta\|^2_{\mathbf{H}_t(\theta_*) + \lambda_0 \|\eta\|_2^2} \left(1 + \frac{1}{2} \log\left(1 + \frac{\|\eta\|^2_{\mathbf{H}_t(\theta_*)}}{\lambda_0 \|\eta\|_2^2}\right)\right)}.$$

Rearanging and using the fact that $\lambda_0 ||\eta||_2^2 \leqslant \|\eta\|^2_{\mathbf{H}_t(\theta_*)} \leqslant t\|\eta\|_2$ yields

$$|\eta^\top S_t| \leqslant \rho(\lambda_0) ||\eta||_{\mathbf{H}_t(\theta_*)} \sqrt{2 \log \frac{t}{\delta}}. \tag{A.6}$$

where ρ is defined as

$$\rho(\lambda_0) = \sqrt{3 + 2 \log\left(1 + \frac{4Q^2}{\lambda_0}\right)}.$$

We take M_t to be a matrix such that $M_t^2 = \mathbf{H}_t(\theta_*)$ and note that for any $\tau > 0$

$$P\left(||S_t||^2_{\mathbf{H}_t^{-1}(\theta_*)} \geqslant d\tau^2\right) \leqslant \sum_{i=1}^{d} P\left(\left|S_t^\top M_t^{-1} e_i\right| \geqslant \tau\right)$$

where e_i is the i:th unit vector. Equation A.6 with $\eta = M_t^{-1}e_i$ together with $||M_t^{-1}e_i||_{\mathbf{H}_t(\theta_*)} = 1$ yield that the following holds with with probability at least $1-\delta$

$$||S_t||_{\mathbf{H}_t^{-1}(\theta_*)} \leqslant \rho(\lambda_0)\sqrt{2d\log t}\sqrt{\log\frac{d}{\delta}}. \tag{A.7}$$

Step 3. (Unverifiable) High-probability bounds on $\mathcal{E}_1$ and $\mathcal{E}_2$.

We now have enough machinery to state high-probability bounds for our two events. These bounds will be *unverifiable* in the sense that the depend on the true parameter θ_* which is not known to us during runtime. We derive verifiable bounds in the next step of the proof.

Recall that $\mathbf{H}_t^{-1}(\theta_*)$ is symmetric. We apply Equation A.6 with $\eta = \mathbf{H}_t^{-1}(\theta_*)z_{ij}$ and $\alpha > 0$ in place of δ. First, we note that $\|\mathbf{H}_t^{-1}(\theta_*)z_{ij}\|_{\mathbf{H}_t(\theta_*)} = \|z_{ij}\|_{\mathbf{H}_t^{-1}(\theta_*)}$ which implies with probability at least $1-\alpha$

$$\left|z_{ij}^\top \mathbf{H}_t^{-1}(\theta_*)S_t\right| = \left|S_t^\top \mathbf{H}_t^{-1}(\theta_*)z_{ij}\right| \leqslant \rho(\lambda_0)\|z_{ij}\|_{\mathbf{H}_t^{-1}(\theta_*)}\sqrt{2\log\frac{t}{\alpha}}. \tag{A.8}$$

We solve for smallest possible $\alpha \in (0, 1/e)$ such that

$$(1+2S)\rho(\lambda_0)\dot{\sigma}(z_{ij}^\top\theta_*)||z_{ij}||_{\mathbf{H}_t^{-1}(\theta_*)}\sqrt{2\log\frac{t}{\alpha}} \leqslant \frac{\Delta}{2}$$

Rearanging yields

$$\alpha \leqslant \exp\left\{\frac{-\Delta^2}{8\rho^2(\lambda_0)(1+2S)^2\left(\dot{\sigma}(z_{ij}^\top\theta_*)||z_{ij}||_{\mathbf{H}_t^{-1}(\theta_*))}\right)^2} + \log T\right\}. \tag{A.9}$$

For $\mathcal{E}_2$ and the bound on its probability, $\beta > 0$ we have

$$\frac{1}{4}|z_{ij}^\top(\theta_t - \theta_*)|^2 \leqslant \frac{1}{2}(1+2S)^2||z_{ij}||^2_{\mathbf{H}_t^{-1}(\theta_*)}\rho^2(\lambda_0)\log\frac{t}{\beta} \leqslant \frac{\Delta}{2}$$

and

$$\beta \leqslant \exp\left\{\frac{-\Delta}{\rho^2(\lambda_0)(1+2S)^2\left(||z_{ij}||_{\mathbf{H}_t^{-1}(\theta_*))}\right)^2} + \log T\right\}. \tag{A.10}$$

Note that both Equation A.9 and Equation A.10 are under the assumption that the RHS satisfy $< 1/e$ since this is required in order to apply the results of Peña et al. (2004). As we discuss in the main text, these quantities are approaching zero as $O(Te^{-T})$, ignoring various constants, for reasonable sampling strategies and will satisfy this condition eventually.

Step 4. (Verifiable) High-probability bounds on $\mathcal{E}_1$ and $\mathcal{E}_2$.

The bounds in the previous step depend on the true parameter θ_* which we do not have access to in practise. We again use Lemma 10 of Faury et al. (2020) together with Cauchy-Schwartz

$$\begin{aligned}|z_{ij}(\theta_* - \theta_t)| &= \left|z_{ij}^\top \mathbf{H}_t^{-1/2}(\theta')\mathbf{H}_t^{1/2}(\theta')S_t\right| \\ &\leqslant (1+2S)||z_{ij}||_{\mathbf{H}_t^{-1}(\theta_t)}||S_t||_{\mathbf{H}_t^{-1}(\theta_*)}.\end{aligned}$$

Using Equation A.7 we have with probability at last $1-\alpha$

$$(1+2S)\sigma(z_{ij}^\top\theta_*)||z_{ij}||_{\mathbf{H}_t^{-1}(\theta_t)}||S_t||_{\mathbf{H}_t^{-1}(\theta_*)} \leqslant (1+2S)\dot{\sigma}(z_{ij}^\top\theta_*)||z_{ij}||_{\mathbf{H}_t^{-1}(\theta_t)}\rho(\lambda_0)\sqrt{2d\log t}\sqrt{\log\frac{d}{\alpha}}. \tag{A.11}$$

We solve for smallest $\alpha \in (1/e)$ such that Equation A.11 is smaller than $\Delta_{ij}/2$. This yields

$$\alpha \leqslant \exp\left\{\frac{-\Delta^2}{8d\rho^2(\lambda_0)(1+2S)^2\left(\dot{\sigma}(z_{ij}^\top\theta_T)||z_{ij}||_{\mathbf{H}_t^{-1}(\theta_T))}\right)^2} + \log dT\right\}.$$

Same steps for β yields

$$\beta \leqslant \exp\left\{\frac{-\Delta}{d\rho^2(\lambda_0)(1+2S)^2\left(||z_{ij}||_{\mathbf{H}_t^{-1}(\theta_T))}\right)^2} + \log dT\right\}.$$

For brevity, define $C_1 = \rho^2(\lambda_0)(1+2S)^2$.
Using the definition of $\tilde{\mathbf{H}}_t$ yields the statement of Lemma 4.1. □

A.2 Proof of Theorem 4.2

Proof. We let $i \succ j$ denote that i is preferred to j. W.l.o.g assume $1 \succ 2 \succ ... \succ n$. The key observation is thatfor any i and j such that $i < j$ it holds that

$$\Delta_{i,j} > (j-i)\Delta_*.$$

If we get the wrong relation between i, j then $\sigma(z_{ij}^\top\theta_*) - \sigma(z_{ij}^\top\theta_T) > (j-i)\Delta_*$. Lemma 4.1 implies

$$P(\sigma(z_{ij}^\top\theta_*) - \sigma(z_{ij}^\top\theta_T) > (j-i)\Delta) \leqslant dT(\underbrace{\exp\left\{\frac{-(j-i)\Delta^2}{8d\rho^2(\lambda_0)(1+2S)^2\left(\dot{\sigma}(z_{ij}^\top\theta_T)||z_{ij}||_{H_t^{-1}(\theta_T))}\right)^2}\right\}}_{\alpha_{ij}^{j-i}}$$

$$+ \underbrace{\exp\left\{\frac{-(j-i)\Delta}{d\rho(\lambda_0)(1+2S)^2\left(||z_{ij}||_{H_t^{-1}(\theta_T))}\right)^2}\right\}}_{\beta_{ij}^{j-i}}).$$

Let $R(\theta_T)$ be the ordering error of the n items. Then, under a uniform distribution over items we have

$$\mathbb{E}[R(\theta_T)] \leqslant \frac{4dT}{n(n-1)} \left(\underbrace{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \alpha_{ij}^{j-i}}_{A} + \underbrace{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \beta_{ij}^{j-i}}_{B} \right) \tag{A.12}$$

A and B will be upper bounded using the same argument. We now upper bound sum A

Let $\alpha_* := \exp\left\{ \frac{-\Delta_*^2}{8dC_1 \max_{i,j} \dot{\sigma}(z_{ij}^\top \theta_T) ||z_{ij}||^2_{\mathbf{H}_t^{-1}(\theta_*)}} \right\}$ then

$$\begin{aligned} A &\leqslant \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \alpha_*^{j-i} \leqslant (n-1) \left(\sum_{j=0}^{n} \alpha_*^j - 1 \right) \\ &\leqslant (n-1) \left(\frac{1}{1-\alpha_*} - 1 \right). \end{aligned}$$

This follows from the definition of $\delta_{1,*}$ and properties of the geometric sum. It is easy to see that $\frac{1}{1-e^{-x}} - 1 = \frac{1}{e^x - 1}$. Hence,

$$\frac{4dT}{n(n-1)} A \leqslant \frac{4dT}{n} \left(\alpha_*^{-1} - 1 \right)^{-1}.$$

For B we perform the same steps with $\beta_* := \exp\left\{ \frac{-\Delta_*}{dC_1 \max_{i,j} ||z_{ij}||^2_{\mathbf{H}_t^{-1}(\theta_*)}} \right\}$ to get

$$\frac{4dT}{n(n-1)} A \leqslant \frac{4dT}{n} \left(\beta_*^{-1} - 1 \right)^{-1}.$$

Combing yields and

$$\mathbb{E}\left[R(\theta_T)\right] \leqslant \frac{4dT}{n} \left(\left(\alpha_*^{-1} - 1 \right)^{-1} + \left(\beta_*^{-1} - 1 \right)^{-1} \right)$$

By Markov's inequality we have

$$P(R(\theta_T) \geqslant \epsilon) \leqslant \frac{4dT}{\epsilon n} \left(\left(\alpha_*^{-1} - 1 \right)^{-1} + \left(\beta_*^{-1} - 1 \right)^{-1} \right). \tag{A.13}$$

□

A.3 Extensions of current theory

Regularized estimators. In our analysis in Section 4, we have assumed that θ_T is the maximum likelihood estimate and that $\mathbf{H}(\theta_T)$ has full rank. This can be relaxed by considering ℓ_2 (Ridge) regularization where $\theta_{\lambda_0,T}$ is the optimum of the regularized log-likelihood with regularization $\lambda_0 \mathbf{I}$ and $\mathbf{H}_{\lambda_0}(\theta_{\lambda_0,T}) = \sum_{s=1}^{T} \dot{\sigma}(z_s^\top \theta_{\lambda_0,T}) z_s z_s^\top + \lambda_0 \mathbf{I}$. The same machinery used to prove Lemma 4.1 (Filippi et al. 2010; Faury et al. 2020) can be applied to this regularized version with small changes to the final bound.

Generalized linear models. It is also possible to derive similar results for generalized linear models with other link functions, $\mu(z_{ij}^\top \theta_*)$, by using the general inequality $\mathbf{H}(\theta) \geqslant \kappa^{-1}\mathbf{V}$ with $\mathbf{V} = \sum_{s=1}^{T} z_s z_s^\top$ and $\kappa \geqslant 1/\min_{z_{ij}} \dot{\mu}(z_{ij}^\top \theta_*)$. We conjecture that this will yield a scaling of $\sim \exp(-\Delta^2 T/\kappa)$ where, unfortunately, κ might be very large. For a more thorough discussion on the dependence on κ in generalized linear bandits, see Chapter 19 of Lattimore and Szepesvári (2020).

Lower and algorithm-specific upper bounds on the ordering error. A worst-case lower bound on the ordering error can be constructed in the fixed-confidence setting, where the goal is to minimize the number of comparisons until a correct ordering is found with a given confidence, by following Garivier and Kaufmann (2016). This involves defining the set of *alternative* models $\mathrm{Alt}(\theta_*)$ which differs from θ_* in their induced ordering of $\mathcal{I}$. The bound is then constructed by optimizing the frequency of comparisons of each pair of items so that such alternative models are distinguished as much as possible from the true parameter. We have left this result out of the paper as we find it uninformative in the regime when the number of comparisons is small, (see Simchowitz et al. (2017) for a discussion on the limitations of these asymptotic results in the standard bandit setting). Constructing a lower bound for our fixed-budget setting, of learning as good an ordering as possible with a fixed number of comparisons, is much more challenging. The fixed-confidence result yields *a* bound for the fixed-budget case (Garivier and Kaufmann 2016), but constructing either a tight lower bound or a tight algorithm-specific upper bound is an open problem (Fang 2022).

A Comparison with regret minimization

Bengs, Saha, et al. (2022) considered a problem formulation where the goal is to learn a parameter θ which determines the utility $Y_{i,t}$ for a set of arms $i = 1, ..., n$ as a function of observed context vectors $x_{i,t}$ in a sequence of rounds $t = 1, ..., T$,

$$Y_{i,t} = \theta^\top X_{i,t} \ .$$

The probability that item i is preferred over j (denoted $i \succ j$) in round t is decided through a comparison function F,

$$\mathbb{P}\left[(\right] i \succ j \mid X_{i,t}, X_{j,t}) = F(Y_{i,t} - Y_{j,t}) \ .$$

The goal in their setting is to, in each round, select two items (i_t, j_t) so that their maximum (or average) utility is as close as possible to the utility of the best item. The expected regret in their average-utility setting is

$$\Re_{BSH} = \mathbb{E}[\sum_{t=1}^{T} 2Y_{i_t^*,t} - Y_{i_t,t} - Y_{j_t,t}] \ .$$

Theorem A.1 (Informal)**.** *An algorithm which achieves minimal regret in the setting of Bengs, Saha, et al. (2022) can perform arbitrarily poorly in our setting.*

Proof. The optimal choice of arm pair in the BSH setting is the optimal and next-optimal arm (i_t^*, i_t') such that $i_t^* \succ i_t' \succ j$ for any other arms j. Assume that the ordering of all other arms j is determined by a feature $X_{j,t}(k)$ but that $X_{i_t^*,t}(k) = X_{i_t',t}(k)$. Then, no knowledge will be gained about arms other than the top 2 choices under the BSH regret. As the number of arms grows larger, the error in our setting grows as well. □

Saha (2021) study the same average-utility regret setting and give a lower bound under Gumbel noise. Saha and Krishnamurthy (2022) investigated where there is a computationally efficient algorithm that achieves the derived optimality guarantee.

A Experiment details

For BayesGURO and BALD, the posterior $p(\theta \mid D_t)$ is estimated using the Laplace approximation as described in Chapter 4 of Bishop and Nasrabadi (2006). With this approximation, the covariance matrix is the same as the inverse of the Hessian of the log-likelihood. For both methods, the priors $\theta_{B,0} = \mathbf{0}^d$ and $\mathbf{H}_{B,0}^{-1} = I_d$ were used, and sequential updates were performed every iteration. The sample criterion for BALD under a logistic model is given in Appendix A.5. For BayesGURO, 50 posterior samples were used to estimate $\hat{\mathbb{V}}_{\theta|D_t}[\sigma(\theta^T z_{ij})]$ for every z_{ij}. The hybrid algorithms follow the same structure with the added constraint that each per-item parameter ζ_i is independent of other parameters. This allows for efficient updates of $\mathbf{H}_{B,t}^{-1}$ by using sparsity in the covariance.

GURO, CoLSTIM, and Uniform use LogisticRegression from Scikit-learn (Pedregosa et al. 2011) with default Ridge regularization ($C = 1$) and the lbfgs optimizer. The former two updates θ_t every iteration using the full history, D_t in all experiments except for IMDB-WIKI-SbS, where GURO updates θ_t every 25th iteration. This caused no noticeable change in performance as GURO still updates $\mathbf{H}_t^{-1}$ every iteration using the Sherman-Morrison formula. Note that when using the Sherman-Morrison formula in practice, you only get an estimate of $\mathbf{H}_t^{-1}(\theta_t)$ since previous versions have been calculated using older estimates of θ. This method for approximating the inverse hessian is covered in Chapter 5 of Bishop and Nasrabadi (2006) and when we compared it to calculating $\mathbf{H}_t^{-1}(\theta_t)$ from scratch every iteration we observed that the methods performed equally. The design matrix for CoLSTIM is updated as in Bengs, Saha, et al. (2022): the confidence width c_1 was chosen to be $\sqrt{d \log(T)}$, and the perturbed values were generated using the standard Gumbel distribution.

To increase computational efficiency for the large IMDB-WIKI-SbS dataset, the hybrid algorithms did not evaluate all $\sim$ 100 000 comparisons at every time step. Instead, a subset of 5 000 comparisons was first sampled, and the highest-scoring pair in this set was chosen. This resulted in a large speed-up and no noticeable change in performance during evaluation.

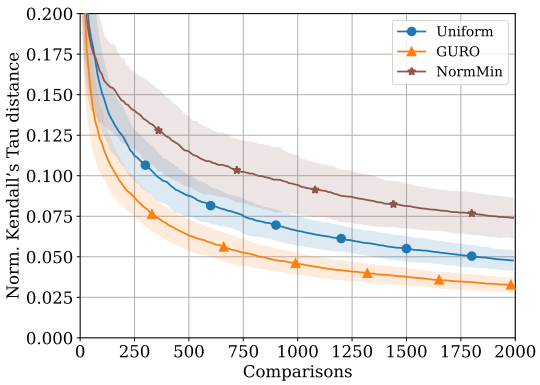
A.1 Datasets

ImageClarity Data available at:
`https://dbgroup.cs.tsinghua.edu.cn/ligl/crowdtopk`.

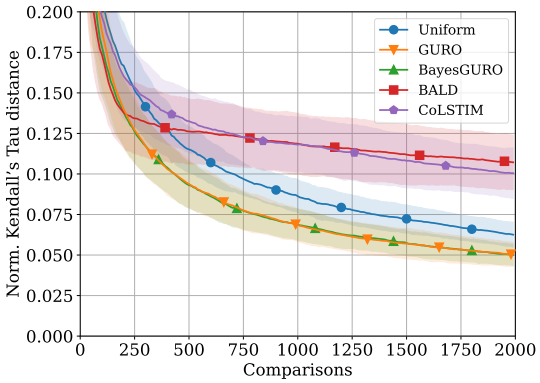
This dataset contained differently distorted versions of the same image. To extract relevant features, we used a ResNet34 model (He et al. 2016) that had been pre-trained on Imagenet (Deng et al. 2009). After PCA projection feature dimensionality was reduced to $d = 63$. The dataset consisted of 100 images and 27 730 comparisons. Since the type of distortion is the same for all images, the dataset has a true ordering with regards to the strength of the distortion applied.

WiscAdds Data available at:
`https://dataverse.harvard.edu/dataset.xhtml?persistentId= doi:10.7910/DVN/0ZRGEE`
(license: CC0 1.0).

The WiscAdds dataset, containing 935 political texts, has been extended with 9 528 pairwise comparisons by Carlson and Montgomery (2017). In comparisons, annotators have stated which of two texts has a more negative tone toward a political opponent. To extract



(a) **X-RayAge.** NormMin included in the experiment shown in Figure 7.1a.



(b) **X-RayAge.** The in-sample error R_{I_D} for the generalization experiment performed in Figure 7.1b.

Figure 7.3: Additional figures from the **X-RayAge** experiment.

general features from the text, sentences were embedded using the pre-trained all-mpnet-base-v2 model from the Sentence-Transformers library (Reimers and Gurevych 2019). After applying PCA to the sentence embeddings, each embedding had a dimensionality of $d = 162$.

IMDB-WIKI-SbS Data available at:
`https://github.com/Toloka/IMDB-WIKI-SbS` (license: CC BY).

IMDB-WIKI-SbS consists of close-up images of actors of different ages. For each comparison, the label corresponds to which of two people appears older. The complete dataset consists of 9 150 images and 250 249 comparisons, but images that were grayscale or had a resolution lower than 160×160 were removed, resulting in 6 072 images and 110 349 comparisons. We extract features from each image using the Inception-ResNet implemented in FaceNet (Schroff et al. 2015) followed by PCA, resulting in $d = 75$ features per image.

A.2 Additional figures

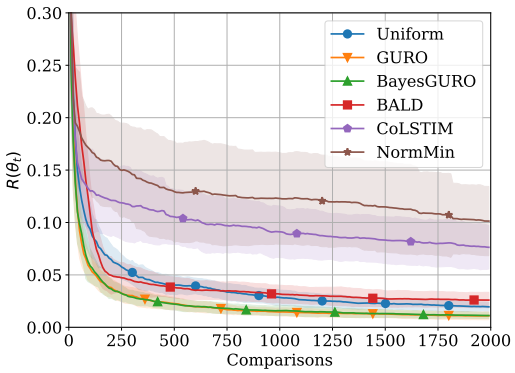
X-RayAge

To highlight the importance of the first-order term in Lemma 4.1, we evaluated NormMin on the same X-ray ordering task as in Figure 7.1a. The results, shown in Figure 7.3a, indicate that not only does the algorithm perform worse than GURO, but is seemingly also outperformed by a uniform sampling strategy. Furthermore, for completeness, we include Figure 7.3b which shows the in-sample error, R_{I_D}, during the generalization experiment.

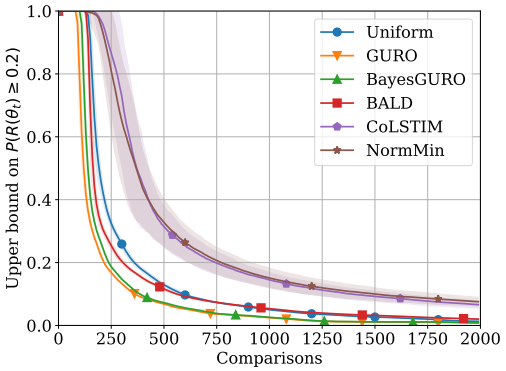
Synthetic Example and Illustration of Upper Bound

In this setting, 100 synthetic data points were generated. Each data point consisted of 10 features, where the feature values were sampled according to a standard normal distribution. The true model, θ_*, was generated by sampling each value uniformly between -3 and 3. The pairwise comparison feedback was simulated the same way as in Section 6.1, with $\lambda = 0.5$. The upper bound of the probability that $R(\theta_t) \geqslant 0.2$ was calculated

every iteration according to Theorem 4.2. Each algorithm was run for 2000 comparisons, updating every 10th, the results of which can be seen in Figure 7.4. We observe in Figure 7.4b that our greedy algorithms are seemingly the fastest at minimizing the upper bound. The order of performance follows the same trend as in the experiments of Section 6.



(a) The risk $R(\theta_t)$, defined as the normalized Kendall's tau distance between estimated and true orderings.



(b) The probability that the frequency of pairwise inversions is $\geqslant 20\%$ after every comparison, according to (4.2).

Figure 7.4: The loss (left) along with the upper bound (right) when ordering a list of size 100 in a synthetic environment. The results have been averaged over 50 seeds.

Randomly initialized representation

As discussed in Section 6.2, the performance of our contextual approach will depend on the quality of the representations. To underscore the practical usefulness of our algorithms, we have performed the same experiment as in Figure 7.2c, but this time the model used to extract image features was untrained (i.e., the weights were random). As to be expected, the results, shown in Figure 7.5, demonstrate that the fully contextual algorithms have no real way of ordering the items according to these uninformative features. However, GURO Hybrid performs similarly to TrueSkill, despite model misspecification. This is promising, since you may not know in advance how informative the extracted features will be for the target ordering task.

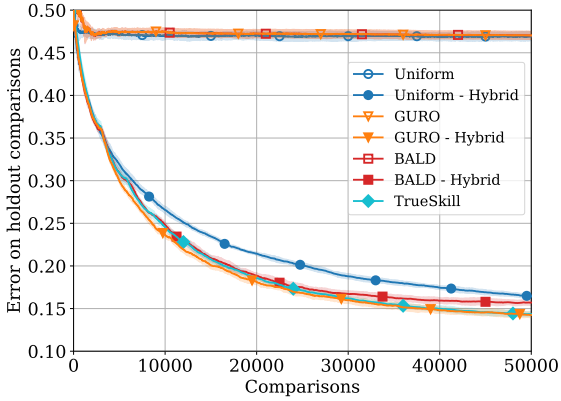


Figure 7.5: **IMDB-WIKI-SbS.** The same experiment as presented in Figure. 7.2c, but the model used for feature extraction is untrained.

ImageClarity ground truth

The ImageClarity dataset consists of multiple versions of the *same image*, with the *same distortion* applied to it to varying degrees. Due to this artificial construction, the pairwise comparisons should, given enough samples, reflect the magnitudes of the applied distortions. In Figure 7.6 we perform the same experiment as in Figure 7.2a, but instead of evaluating on a holdout comparison set, we measure the distance to the ground-truth ordering. The overall results are very similar, although we do see a slight increase in the performance of contextual algorithms compared to the non-contextual TrueSkill.

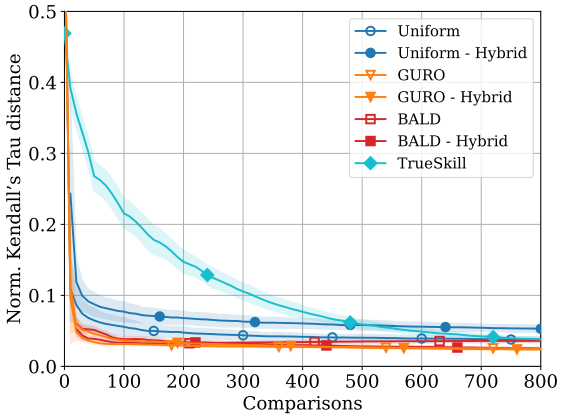


Figure 7.6: **ImageClarity.** Same experiment as in Figure 7.2a, but now measuring the distance to the ground-truth ordering. Averaged over 25 seeds along with the 1-sigma error region.

Ground truth ordering using the Bradley-Terry model

An alternate approach to evaluate ordering quality is to estimate a "ground-truth" ordering by applying the popular Bradley-Terry (BT) model (Bradley and Terry 1952) to all available comparisons. We used the CrowdKit library (Ustalov et al. 2024) to find the MLE scores for each item and ordered the elements accordingly. In Figure 7.8 we run the same experiments as in Figure 7.2, but instead measure the distance to the constructed BT ordering. The overall trends remain, but for (b) and (c) there is a slight shift for the later iterations. More specifically we see non-contextual TrueSkill eventually overtaking the contextual algorithms.

The issue is that algorithms with orderings closer to the maximum likelihood estimate of the BT model will be favored. To exemplify this we use the ImageClarity dataset since it contains the largest number of comparisons relative to the number of items. We sample 1 000 comparisons and let this be the collection that is available to the algorithms. We further construct two target orderings, one from the BT estimate using the sampled subset of comparisons, and a second, more probable ordering, from the BT estimate using all 27 730 available comparisons. Figure 7.7 shows the distance between the GURO and TS algorithms and the different target orderings, where dashed lines indicate the distance to the ordering generated using the full list of comparisons. If we only look at the distance to the ordering produced using our subset of comparisons, TrueSkill seemingly outperforms GURO after about 350 comparisons. However, if we instead measure the distance to the more probable ordering, we see that GURO converges toward a lower distance. Note that these are the same orderings, evaluated against different targets. This is likely the effect

we observe in Figure 7.8b and c, but not in Figure 7.8a as a result of the high amount of comparisons available to us.

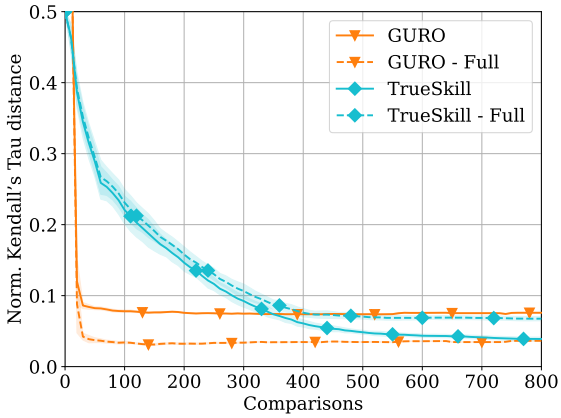


Figure 7.7: **ImageClarity.** The same experiment as presented in Figure 7.2a, but we instead measure the distance to target orderings that correspond to the maximum likelihood estimate of the BT model using different numbers of comparisons. The dashed lines show the distance to the BT estimate using all 27 730 comparisons.

References

Ailon, Nir (2011). "Active learning ranking from pairwise preferences with almost optimal query complexity". In: *Advances in Neural Information Processing Systems* 24 (cit. on pp. 232, 234).

Bai, Yuntao et al. (2022). *Constitutional AI: Harmlessness from AI Feedback*. arXiv: `2212.08073 [cs.CL]` (cit. on p. 235).

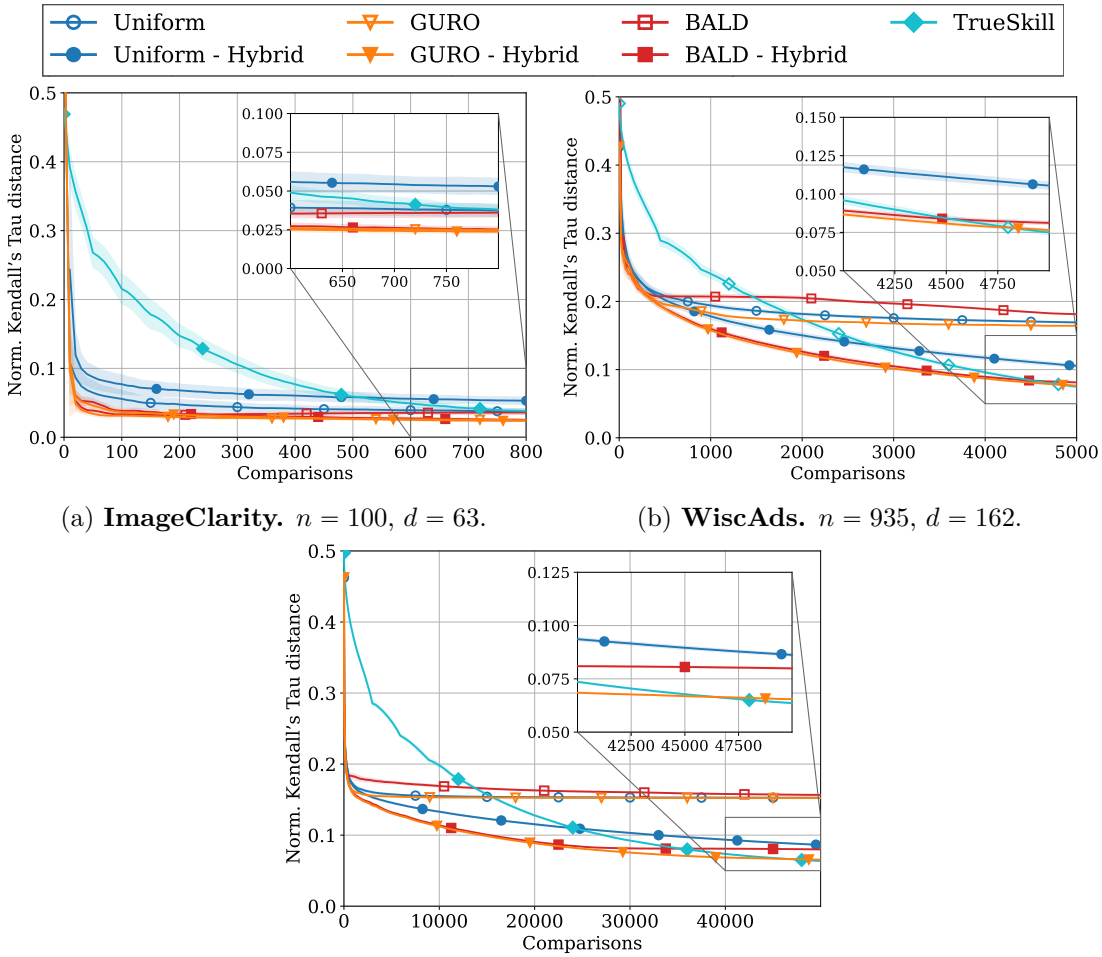
Bengs, Viktor, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier (2021). "Preference-based online learning with dueling bandits: A survey". In: *The Journal of Machine Learning Research* 22.1, pp. 278–385 (cit. on p. 234).

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 pp. 232, 234, 240, 258, 259).

Bishop, Christopher M and Nasser M Nasrabadi (2006). *Pattern recognition and machine learning*. Springer (cit. on pp. 249, 259).

Bradley, Ralph Allan and Milton E Terry (1952). "Rank analysis of incomplete block designs: I. The method of paired comparisons". In: *Biometrika* 39.3/4, pp. 324–345 (cit. on pp. 233, 262).

Brinker, Klaus (2004). "Active learning of label ranking functions". In: *Proceedings of the twenty-first international conference on Machine learning*, p. 17 (cit. on p. 234).



(a) **ImageClarity.** $n = 100$, $d = 63$.

(b) **WiscAds.** $n = 935$, $d = 162$.

(c) **IMDB-WIKI-SbS.** $n = 6\ 072$, $d = 75$.

Figure 7.8: The same experiment as presented in Figure. 7.2c, but we instead measure the distance to a ground-truth estimated using all available comparisons.

Burges, Chris, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender (2005). "Learning to rank using gradient descent". In: *Proceedings of the 22nd international conference on Machine learning*, pp. 89–96 (cit. on p. 234).

Busse, Ludwig M., Morteza Haghir Chehreghani, and Joachim M. Buhmann (2012). "The information content in sorting algorithms". In: *2012 IEEE International Symposium on Information Theory Proceedings*, pp. 2746–2750. DOI: 10.1109/ISIT.2012.6284021 (cit. on p. 234).

Canal, Gregory, Andy Massimino, Mark Davenport, and Christopher Rozell (2019). "Active embedding search via noisy paired comparisons". In: *International Conference on Machine Learning*. PMLR, pp. 902–911 (cit. on pp. 234, 239).

Carlson, David and Jacob M. Montgomery (Nov. 2017). "A Pairwise Comparison Framework for Fast, Flexible, and Reliable Human Coding of Political Texts". en. In: *American Political Science Review* 111.4, pp. 835–843. ISSN: 0003-0554, 1537-5943. (Visited on 05/14/2024) (cit. on pp. 242, 259).

Chen, Kani, Inchi Hu, and Zhiliang Ying (1999). "Strong Consistency of Maximum Quasi-Likelihood Estimators in Generalized Linear Models with Fixed and Adaptive Designs". In: *The Annals of Statistics* 27.4, pp. 1155–1163. (Visited on 01/08/2024) (cit. on p. 236).

Chen, Xi, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz (Feb. 2013). "Pairwise ranking aggregation in a crowdsourced setting". In: *Proceedings of the sixth ACM international conference on Web search and data mining*. WSDM '13. New York, NY, USA: Association for Computing Machinery, pp. 193–202. ISBN: 978-1-4503-1869-3. DOI: 10.1145/2433396.2433420. (Visited on 01/26/2024) (cit. on p. 231).

Christiano, Paul F, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei (2017). "Deep reinforcement learning from human preferences". In: *Advances in neural information processing systems* 30 (cit. on pp. 231, 234).

Chu, Wei and Zoubin Ghahramani (2005). "Preference learning with Gaussian processes". In: *Proceedings of the 22nd international conference on Machine learning*, pp. 137–144 (cit. on p. 234).

Cohen, Joseph Paul et al. (2022). "TorchXRayVision: A library of chest X-ray datasets and models". In: *Medical Imaging with Deep Learning* (cit. on p. 241).

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 pp. 232, 235, 241).

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei (2009). "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255 (cit. on p. 259).

Di, Qiwei, Tao Jin, Yue Wu, Heyang Zhao, Farzad Farnoud, and Quanquan Gu (2023). "Variance-Aware Regret Bounds for Stochastic Contextual Dueling Bandits". In: *arXiv preprint arXiv:2310.00968* (cit. on p. 234).

Dudík, Miroslav, Katja Hofmann, Robert E Schapire, Aleksandrs Slivkins, and Masrour Zoghi (2015). "Contextual dueling bandits". In: *Conference on Learning Theory*. PMLR, pp. 563–587 (cit. on p. 234).

Fang, Boli (2022). "Fixed-Budget Pure Exploration in Multinomial Logit Bandits". In: *International Joint Conference on Artificial Intelligence* (cit. on pp. 234, 257).

Faury, Louis, Marc Abeille, Clément Calauzènes, and Olivier Fercoq (2020). "Improved optimistic algorithms for logistic bandits". In: *International Conference on Machine Learning*. PMLR, pp. 3052–3060 (cit. on pp. 234, 236, 251, 252, 255, 256).

Felipe Kitamura Lilian Mallagoli, Paulo Kuriki (2023). *SPR X-Ray Age Prediction Challenge* (cit. on p. 241).

Filippi, Sarah, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári (2010). "Parametric Bandits: The Generalized Linear Case". In: *Advances in Neural Information Processing Systems*. Ed. by J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta. Vol. 23. Curran Associates, Inc. (cit. on pp. 232, 234, 236, 252, 253, 256).

Fürnkranz, Johannes and Eyke Hüllermeier (2003). "Pairwise preference learning and ranking". In: *European conference on machine learning*. Springer, pp. 145–156 (cit. on pp. 231, 234).

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. 257).

George, Anne-Marie and Christos Dimitrakakis (2023). *Eliciting Kemeny Rankings*. arXiv: 2312.11663 [cs.LG] (cit. on p. 234).

Graepel, Thore (Jan. 2012). "Score-based Bayesian Skill Learning". In: *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD-12)* (cit. on p. 240).

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 770–778 (cit. on pp. 242, 259).

Heckel, Reinhard, Max Simchowitz, Kannan Ramchandran, and Martin Wainwright (2018). "Approximate ranking from pairwise comparisons". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1057–1066 (cit. on pp. 232, 234).

Hees, Jörn, Benjamin Adrian, Ralf Biedert, Thomas Roth-Berghofer, and Andreas Dengel (2016). "Tssort: Probabilistic noise resistant sorting". In: *arXiv preprint arXiv:1606.05289* (cit. on p. 240).

Herbrich, Ralf, Tom Minka, and Thore Graepel (2006). "TrueSkill™: a Bayesian skill rating system". In: *Advances in neural information processing systems* 19 (cit. on pp. 232, 234, 240).

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 pp. 232, 234, 239, 240, 250).

Ieki, Hirotaka et al. (Dec. 2022). "Deep learning-based age estimation from chest X-rays indicates cardiovascular prognosis". en. In: *Communications Medicine* 2.1. Number: 1 Publisher: Nature Publishing Group, pp. 1–12. ISSN: 2730-664X. DOI: 10.1038/s43856-022-00220-6. (Visited on 09/21/2023) (cit. on p. 241).

Jamieson, Kevin G and Robert Nowak (2011). "Active ranking using pairwise comparisons". In: *Advances in neural information processing systems* 24 (cit. on pp. 232, 234).

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. 231).

Jun, Kwang-Sung, Lalit Jain, Blake Mason, and Houssam Nassif (18–24 Jul 2021). "Improved Confidence Bounds for the Linear Logistic Model and Applications to Bandits". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 5148–5157 (cit. on pp. 234, 252).

Kendall, Maurice George (1948). *Rank correlation methods*. Griffin (cit. on p. 233).

Kirsch, Andreas and Yarin Gal (2022). "Unifying Approaches in Active Learning and Active Sampling via Fisher Information and Information-Theoretic Quantities". In: *Transactions on Machine Learning Research*. Expert Certification. ISSN: 2835-8856 (cit. on pp. 234, 237).

Kveton, Branislav, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier (2020). "Randomized exploration in generalized linear bandits". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 2066–2076 (cit. on p. 234).

Larkin, Andrew, Ajay Krishna, Lizhong Chen, Ofer Amram, Ally R. Avery, Glen E. Duncan, and Perry Hystad (Nov. 2022). "Measuring and modelling perceptions of the built environment for epidemiological research using crowd-sourcing and image-based deep learning models". en. In: *Journal of Exposure Science & Environmental Epidemiology* 32.6. Number: 6 Publisher: Nature Publishing Group, pp. 892–899. ISSN: 1559-064X. DOI: 10.1038/s41370-022-00489-8. (Visited on 08/03/2023) (cit. on pp. 231, 240).

Lattimore, Tor and Csaba Szepesvári (2020). *Bandit Algorithms*. Cambridge University Press. DOI: 10.1017/9781108571401 (cit. on p. 257).

Li, Lihong, Yu Lu, and Dengyong Zhou (2017). "Provably optimal algorithms for generalized linear contextual bandits". In: *International Conference on Machine Learning*. PMLR, pp. 2071–2080 (cit. on p. 234).

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. 231).

Ling, Suiyi, Jing Li, Anne Flore Perrin, Zhi Li, Lukáš Krasula, and Patrick Le Callet (2020). "Strategy for boosting pair comparison and improving quality assessment accuracy". In: *arXiv preprint arXiv:2010.00370* (cit. on p. 234).

Long, Bo, Olivier Chapelle, Ya Zhang, Yi Chang, Zhaohui Zheng, and Belle Tseng (2010). "Active learning for ranking through expected loss optimization". In: *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 267–274 (cit. on p. 234).

Lou, Hao, Tao Jin, Yue Wu, Pan Xu, Quanquan Gu, and Farzad Farnoud (2022). "Active Ranking without Strong Stochastic Transitivity". In: *Advances in neural information processing systems* 35, pp. 297–309 (cit. on p. 234).

Massimino, Andrew K and Mark A Davenport (2021). "As you like it: Localization via paired comparisons". In: *Journal of Machine Learning Research* 22.186, pp. 1–39 (cit. on p. 234).

Maystre, Lucas and Matthias Grossglauser (2017). "Just sort it! A simple and effective approach to active preference learning". In: *International Conference on Machine Learning*. PMLR, pp. 2344–2353 (cit. on pp. 232, 243).

Mehta, Viraj, Vikramjeet Das, Ojash Neopane, Yijia Dai, Ilija Bogunovic, Jeff Schneider, and Willie Neiswanger (2023). "Sample Efficient Reinforcement Learning from

Human Feedback via Active Exploration". In: *arXiv preprint arXiv:2312.00267* (cit. on p. 235).

Minka, Tom, Ryan Cleven, and Yordan Zaykov (2018). "Trueskill 2: An improved bayesian skill rating system". In: *Technical Report* (cit. on p. 234).

Mukherjee, Subhojyoti, Anusha Lalitha, Kousha Kalantari, Aniket Deshmukh, Ge Liu, Yifei Ma, and Branislav Kveton (2024). *Optimal Design for Human Feedback*. arXiv: 2404.13895 [cs.LG] (cit. on p. 235).

Naik, Nikhil, Jade Philipoom, Ramesh Raskar, and César Hidalgo (2014). "Streetscore-predicting the perceived safety of one million streetscapes". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 779–785 (cit. on p. 240).

Oliveira, IFD, S Zehavi, and O Davidov (2018). "Stochastic transitivity: Axioms and models". In: *Journal of Mathematical Psychology* 85, pp. 25–35 (cit. on p. 233).

Ouyang, Long et al. (2022). *Training language models to follow instructions with human feedback*. arXiv: 2203.02155 [cs.CL] (cit. on pp. 231, 235).

Pavlichenko, Nikita and Dmitry Ustalov (2021). "IMDB-WIKI-SbS: An Evaluation Dataset for Crowdsourced Pairwise Comparisons". In: *CoRR* abs/2110.14990. arXiv: 2110.14990 (cit. on p. 242).

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830 (cit. on p. 259).

Peña, Victor H. de la, Michael J. Klass, and Tze Leung Lai (2004). "Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws". In: *The Annals of Probability* 32.3, pp. 1902–1933. DOI: 10.1214/009117904000000397 (cit. on pp. 252–254).

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. 231).

Qian, Li, Jinyang Gao, and HV Jagadish (2015). "Learning user preferences by adaptive pairwise comparison". In: *Proceedings of the VLDB Endowment* 8.11, pp. 1322–1333 (cit. on pp. 232, 234, 239).

Qin, Chao (2022). "Open problem: Optimal best arm identification with fixed-budget". In: *Conference on Learning Theory*. PMLR, pp. 5650–5654 (cit. on p. 245).

Reimers, Nils and Iryna Gurevych (Nov. 2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics (cit. on pp. 242, 260).

Rusmevichientong, Paat and John N Tsitsiklis (2010). "Linearly parameterized bandits". In: *Mathematics of Operations Research* 35.2, pp. 395–411 (cit. on p. 252).

Saha, Aadirupa (2021). "Optimal algorithms for stochastic contextual preference bandits". In: *Advances in Neural Information Processing Systems* 34, pp. 30050–30062 (cit. on p. 258).

Saha, Aadirupa and Akshay Krishnamurthy (2022). "Efficient and optimal algorithms for contextual dueling bandits under realizability". In: *International Conference on Algorithmic Learning Theory*. PMLR, pp. 968–994 (cit. on p. 258).

Sartori, Andreza, Victoria Yanulevskaya, Almila Akdag Salah, Jasper Uijlings, Elia Bruni, and Nicu Sebe (2015). "Affective analysis of professional and amateur abstract paintings using statistical analysis and art theory". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5.2, pp. 1–27 (cit. on p. 240).

Schroff, Florian, Dmitry Kalenichenko, and James Philbin (June 2015). "FaceNet: A Unified Embedding for Face Recognition and Clustering". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. arXiv:1503.03832 [cs], pp. 815–823. DOI: 10.1109/CVPR.2015.7298682. (Visited on 12/22/2023) (cit. on pp. 242, 260).

Sherman, Jack and Winifred J. Morrison (1950). "Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix". In: *The Annals of Mathematical Statistics* 21.1, pp. 124–127. DOI: 10.1214/aoms/1177729893 (cit. on p. 238).

Silva, Rodrigo M, Marcos A Gonçalves, and Adriano Veloso (2014). "A Two-stage active learning method for learning to rank". In: *Journal of the Association for Information Science and Technology* 65.1, pp. 109–128 (cit. on p. 234).

Simchowitz, Max, Kevin Jamieson, and Benjamin Recht (July 2017). "The Simulator: Understanding Adaptive Sampling in the Moderate-Confidence Regime". In: *Proceedings of the 2017 Conference on Learning Theory*. Ed. by Satyen Kale and Ohad Shamir. Vol. 65. Proceedings of Machine Learning Research. PMLR, pp. 1794–1834 (cit. on p. 257).

Singh, Ankita and Shayok Chakraborty (2021). "Deep active learning with relative label feedback: An application to facial age estimation". In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8 (cit. on p. 239).

Tärnåsen, Hanna and Herman Bergström (2023). "Rank based annotation system for supervised learning in medical imaging". Master's thesis. Chalmers University of Technology (cit. on p. 231).

Ustalov, Dmitry, Nikita Pavlichenko, and Boris Tseitlin (2024). "Learning from Crowds with Crowd-Kit". In: *Journal of Open Source Software* 9.96, p. 6227. ISSN: 2475-9066. DOI: 10.21105/joss.06227. arXiv: 2109.08584 [cs.HC] (cit. on p. 262).

Wu, Tianhao, Banghua Zhu, Ruoyu Zhang, Zhaojin Wen, Kannan Ramchandran, and Jiantao Jiao (2023). *Pairwise Proximal Policy Optimization: Harnessing Relative Feedback for LLM Alignment*. arXiv: 2310.00212 [cs.LG] (cit. on p. 235).

Wu, Yue, Tao Jin, Hao Lou, Farzad Farnoud, and Quanquan Gu (2023). "Borda Regret Minimization for Generalized Linear Dueling Bandits". In: *arXiv preprint arXiv:2303.08816* (cit. on p. 234).

Xu, Pan, Zheng Wen, Handong Zhao, and Quanquan Gu (2022). "Neural Contextual Bandits with Deep Representation and Shallow Exploration". In: *International Conference on Learning Representations* (cit. on p. 239).

Yan, Xinyi, Chengxi Luo, Charles LA Clarke, Nick Craswell, Ellen M Voorhees, and Pablo Castells (2022). "Human preferences as dueling bandits". In: *Proceedings*

of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 567–577 (cit. on p. 234).

Yang, Miao, Ge Yin, Yixiang Du, and Zhiqiang Wei (Nov. 2021). "Pair comparison based progressive subjective quality ranking for underwater images". In: *Signal Processing: Image Communication* 99, p. 116444. ISSN: 09235965. DOI: 10.1016/j.image.2021.116444 (cit. on p. 231).

Yannakakis, Georgios N. and Héctor P. Martínez (July 2015). "Ratings are Overrated!" In: *Frontiers in ICT* 2. DOI: 10.3389/fict.2015.00013 (cit. on p. 231).

Yue, Yisong, Josef Broder, Robert Kleinberg, and Thorsten Joachims (2012). "The k-armed dueling bandits problem". In: *Journal of Computer and System Sciences* 78.5, pp. 1538–1556 (cit. on p. 234).

Yue, Yisong and Thorsten Joachims (2009). "Interactively optimizing information retrieval systems as a dueling bandits problem". In: *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1201–1208 (cit. on p. 234).

Zhang, Xiaohang, Guoliang Li, and Jianhua Feng (Apr. 2016). "Crowdsourced top-k algorithms: an experimental evaluation". In: *Proceedings of the VLDB Endowment* 9.8, pp. 612–623. ISSN: 2150-8097. DOI: 10.14778/2921558.2921559. (Visited on 05/14/2024) (cit. on p. 242).

Zhu, Banghua, Michael Jordan, and Jiantao Jiao (23–29 Jul 2023). "Principled Reinforcement Learning with Human Feedback from Pairwise or K-wise Comparisons". In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 43037–43067 (cit. on pp. 231, 232, 234, 235).