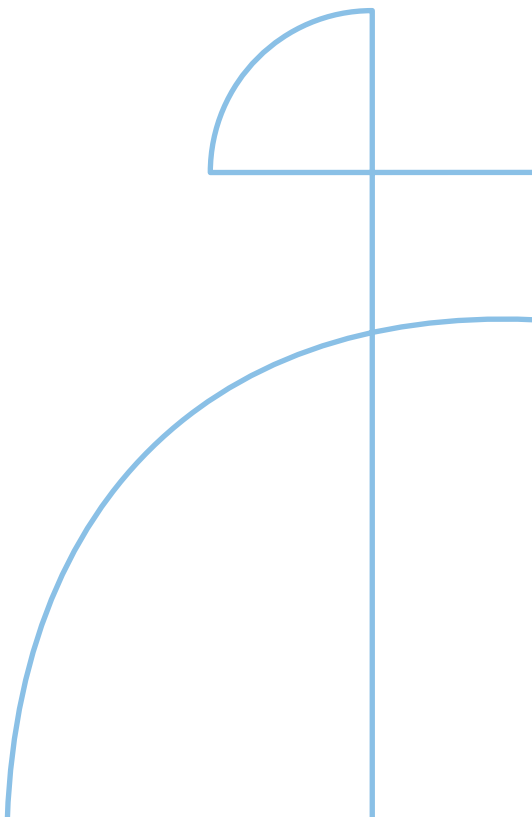
Doctoral Thesis in Electrical Engineering

An Information-Theoretic Approach to Bandits and Reinforcement Learning

AMAURY GOUVERNEUR

KTH ROYAL INSTITUTE OF TECHNOLOGY



An Information-Theoretic Approach to Bandits and Reinforcement Learning

AMAURY GOUVERNEUR

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Doctor of Philosophy on Monday the 15th of December 2025, at 13:00 in F3, Lindstedtvägen 26, Stockholm.

Doctoral Thesis in Electrical Engineering
KTH Royal Institute of Technology
Stockholm, Sweden 2025

TRITA-EECS-AVL-2025:108
ISBN 978-91-8106-474-2

Printed by: Universitetsservice US-AB, Sweden 2025

"There are years that ask questions and years that answer."

Zora Neale Hurston

Abstract

In this thesis, we develop and extend information-theoretic methods for analyzing reinforcement learning, focussing particularly on bandit problems. Our aim is to understand how learning algorithms manage the fundamental trade-off between *exploration* and *exploitation*, and how information-theoretic quantities such as entropy and mutual information can be used to characterize and bound their performance. Building on the framework of Russo and Van Roy, which relates Bayesian regret of an algorithm to its learning efficiency, measured by the *information ratio*, we derive new results that broaden the scope of this approach to more complex and structured learning settings.

The first part of the thesis revisits the analysis of linear bandits, a model in which rewards depend linearly on the selected action via an unknown parameter. While existing information-theoretic analyses provide general regret bounds, they all assume finite action sets. We close this gap by introducing refined analyses and show that near-optimal regret bounds can be obtained even for infinite or continuous action spaces.

The second part of the thesis extends the information analysis framework to more complex bandit models. In the contextual bandit setting, we build on the *lifted information ratio* and extend its application to more general reward classes. Additionally, we provide alternative proofs based on classical information-theoretic tools. This new derivation reveals simplifies the analysis and enables sharper guarantees for structured problems. We then study the Thompson Sampling algorithm for logistic bandits, a widely used model for binary rewards. We derive the *first regret bounds* that avoid exponential dependence on the logistic slope parameter and are independent of the number of actions, resolving a long standing open question in the field.

The third part of the thesis takes a step beyond bandit models and investigates how information-theoretic principles can be extended to more *general reinforcement learning* problems. We study the theoretical performance limits and learning guarantees of model-based reinforcement learning, establishing a framework for analyzing its Bayesian regret. In parallel, we derive PAC-Bayesian bounds for *offline decision-making*, including improved guarantees for offline bandits that are parameter-free and achieve near-optimal rates.

Together, the contributions of this thesis provide a coherent extension of information-theoretic analysis to a wide range of learning settings; from linear to nonlinear, from discrete to continuous, and from bandits to reinforcement learning.

Keywords: Reinforcement learning, Bandit problems, Information theory, Regret analysis, Thompson Sampling, PAC-Bayes bounds.

Sammanfattning

I denna avhandling utvecklar och utvidgar vi informationsteoretiska metoder för att analysera *förstärkningsinlärning* (eng: reinforcement learning), med särskilt fokus på *banditproblem* (eng: bandits). Vårt mål är att förstå hur inlärningsalgoritmer hanterar den grundläggande avvägningen mellan *utforskning* (eng: exploration) och *exploatering* (eng: exploitation), och hur informationsteoretiska storheter såsom entropi och ömsesidig information (eng: mutual information) kan användas för att karakterisera och begränsa deras prestanda. Med utgångspunkt i ramverket av Russo och Van Roy, som relaterar en algoritms Bayesianska *ånger* (eng: Bayesian regret) till dess inlärningseffektivitet mätt genom den så kallade *informationskvoten* (eng: information ratio), härleder vi nya resultat som utvidgar detta angreppssätt till mer komplexa och strukturerade inlärningsproblem.

Den första delen av avhandlingen återbesöker analysen av linjära *banditer*, en modell där belöningar beror linjärt på den valda handlingen via en okänd parameter. Även om befintliga informationsteoretiska analyser ger generella ångergränser, förutsätter de alltid ett ändligt handlingsutrymme. Vi fyller denna lucka genom att introducera förfinade analyser som visar att nära-optimala ångergränser kan uppnås även för oändliga eller kontinuerliga handlingsutrymmen.

Den andra delen av avhandlingen utvidgar det informationsteoretiska analysramverket till mer komplexa *bandit*-modeller. I den kontextuella *bandit*-miljön bygger vi vidare på den så kallade *upplyfta informationskvoten* (eng: lifted information ratio) och visar hur denna kan tillämpas för ett bredare spektrum av belöningsmodeller. Därtill tillhandahåller vi alternativa bevis baserade på klassiska informationsteoretiska verktyg, vilket förenklar analysen och möjliggör skarpare garantier för strukturerade problem. Vi studerar därefter Thompson Sampling-algoritmen för logistiska *banditer*, en centralt använd modell för binära belöningar. Här härleder vi de första ångergränserna som undviker det tidigare nödvändiga exponentiella beroendet av den logistiska lutningsparametern och som dessutom är oberoende av antalet handlingar, vilket löser en länge öppen fråga i området.

Den tredje delen av avhandlingen går bortom *bandit*-modeller och undersöker hur informationsteoretiska principer kan generaliseras till mer allmänna problem inom *förstärkningsinlärning*. Vi analyserar teoretiska prestandagränser och inlärningsgarantier för modellbaserad förstärkningsinlärning och etablerar ett ramverk för att studera dess Bayesianska ånger. Parallellt härleder vi PAC-Bayesianska gränser för *offline*-beslutsfattande (eng: offline decision-making), inklusive förbättrade garantier för offline *banditer* som är parameterfria och uppnår nära-optimala konvergenshastigheter.

Sammantaget presenterar avhandlingen en sammanhängande och bred utvidgning av informationsteoretisk analys till ett stort spektrum av inlärningsmiljöer: från linjära till icke-linjära modeller, från diskreta till kontinuerliga handlingsutrymmen, och från *banditer* till allmän *förstärkningsinlärning*.

Nyckelord: Förstärkningsinlärning, Banditproblem, Informationsteori, Ångeranalys, Thompson Sampling, PAC-Bayesianska gränser.

Popular Scientific Summary

"Life is like a box of chocolates. You never know what you're gonna get," said Forrest Gump to a stranger while waiting for a bus. If Forrest had had more time to elaborate, he probably would have explained that many of life's challenges are, in fact, *sequential decision problems*, situations in which we must make a series of choices without knowing in advance how they will turn out. In these situations, we constantly face a dilemma: should we explore unfamiliar options in the hope of discovering something better, or exploit what we already know and enjoy the best immediate benefit? Trying a new chocolate flavor or choosing a familiar favorite may seem trivial, but finding balance between *exploration* and *exploitation* is at the core of every system that must learn from experience while acting under uncertainty. A content platform decides which video, post, or song to recommend next, hoping to keep users engaged while trying to discover new interests. A company developing a new drug must choose which treatment variant to test on which patient group, learning from early trial results to guide later decisions while minimizing risk. A chatbot tries different reasoning steps, algebraic manipulations, geometric interpretations, numerical experiments, to figure out how to solve a math problem, gradually learning which strategies tend to work in which situations. All of these systems face the same challenge: they must make decisions repeatedly, under uncertainty, and the only way to improve is by trying out different actions and learning from experience. This is precisely the setting studied in reinforcement learning, the field of artificial intelligence that seeks to understand how an agent can learn to make good decisions through trial and error.

To study such decision problems formally, researchers in machine learning use mathematical frameworks. The bandit problem provides the simplest setting: a decision-maker repeatedly chooses among a fixed set of actions without affecting the future outcomes. The challenge is to learn which actions are best through experience, finding a balance between exploring unfamiliar options and exploiting the ones that have performed well so far. Although simple, this setting captures the essence of learning from evaluative feedback and serves as a foundation for more complex forms of sequential decision-making. Reinforcement learning extends this idea to situations where actions have long-term consequences. Here, each decision influences not only the immediate reward but also the future situations the agent will encounter. A self-driving car deciding when to change lanes, a robotic arm learning to grasp and manipulate unfamiliar objects, or a trading system executing orders while accounting for market impact are all examples of environments where actions change the future state of the system. These agents must reason about cause and effect, anticipating how their current choices will shape future opportunities for reward.

A central question in both bandit and reinforcement learning settings is: how quickly and effectively can a learning agent improve its decisions? To answer this, researchers study a quantity called *regret*, which measures how much reward is lost

by not knowing the optimal actions from the start. The lower the regret, the more efficiently an algorithm learns.

This thesis contributes to our understanding of how learning algorithms can achieve low regret by adopting an information-theoretic point of view. Originally developed in the 1940s to study the limits of communication and coding, information theory provides a rigorous mathematical language for quantifying uncertainty and information gain. Its central idea is that information measures how much uncertainty is reduced when new data is observed. In the context of learning and decision-making, information theory offers rigorous tools for describing how much knowledge agents acquire through interaction. By linking the regret of a learning algorithm to the amount of information the agent accumulates over time, information theory allows us to formalize the exploration-exploitation trade-off and derive general regret guarantees.

Building on earlier works, the main technical contributions of the thesis extend and deepen the information-theoretic analysis of reinforcement learning and bandits. First, it refines existing analyses to handle settings with continuous or infinite action spaces, where traditional approaches become ineffective. Second, the analysis is expanded to structured models such as contextual, logistic, and generalized linear bandits, where rewards may depend nonlinearly on actions or contexts. Finally, the thesis takes a step beyond bandits, exploring how these ideas can be adapted to reinforcement learning problems with evolving states and long-term dependencies, and establishing general performance guarantees, including PAC-Bayesian bounds, for model-based and offline learning.

Overall, the results presented in this thesis contribute to building a clearer theoretical foundation for reinforcement learning and provide tools that may help guide the development of future algorithms.

List of Papers

Contributions included in the thesis

[A] **A. Gouverneur**, B. Rodríguez-Gálvez, T. Oechtering, M. Skoglund, "An Information-Theoretic Analysis of Thompson Sampling with Infinite Action Spaces", in *ICASSP 2025 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Feb. 2025, pp. 1–5.

[B] **A. Gouverneur**, B. Rodríguez-Gálvez, T. J. Oechtering, M. Skoglund, *Chained Information-Theoretic bounds and Tight Regret Rate for Linear Bandit Problems*, Under revision at IEEE Transactions on Information Theory, 2024. eprint: `2403.03361` (stat.ML),

- Presented at ICML 2024 Workshop on Foundations of RL and Control.

[C] **A. Gouverneur**, B. Rodríguez-Gálvez, T. J. Oechtering, M. Skoglund, "Thompson Sampling Regret Bounds for Contextual Bandits with sub-Gaussian rewards", in *2023 IEEE International Symposium on Information Theory (ISIT)*, Taipei, Taiwan, IEEE, 2023, pp. 1306–1311.

The conceptualization of [A, B, C] was done by the candidate. The derivation of the results was done by the candidate and B. Rodríguez-Gálvez. The candidate, B. Rodríguez-Gálvez, and the supervisors contributed to writing, reviewing, and editing.

[D] **A. Gouverneur**, T. J. Oechtering, M. Skoglund, *An Information-Theoretic Analysis of Thompson Sampling for Logistic Bandit Problems*, under revision for TMLR, 2025. arXiv: `2412.02861 [stat.ML]`,

- Presented at NeurIPS 2024 Workshop on Bayesian Decision-Making under Uncertainty.

The conceptualization and derivation of the results of [D] was done by the candidate. The candidate and the supervisors contributed to writing, reviewing, and editing.

[E] **A. Gouverneur**, B. Rodríguez-Gálvez, T. J. Oechtering, M. Skoglund, "An information-theoretic analysis of Bayesian Reinforcement Learning", in *2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, IEEE, 2022, pp. 1–7.

[F] **A. Gouverneur**, T. J. Oechtering, M. Skoglund, "Refined PAC-Bayes Bounds for Offline Bandits", in *2025 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2025, pp. 1306–1311.

The conceptualization and derivation of the results of [E] was done by the candidate and B. Rodríguez-Gálvez. The candidate, B. Rodríguez-Gálvez, and the supervisors contributed to writing, reviewing, and editing. The conceptualization and derivation of the results of [F] was done by the candidate. The canditate and the supervisors contributed to writting, reviewing and editing.

Other contributions not included in the thesis

In addition to the aforementioned papers, the following papers have also been co-authored by the author of this thesis:

[I] A. Aspeel*, **A. Gouverneur***, R. M Jungers, B. Macq, "Optimal measurement budget allocation for particle filtering", in *2020 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2020, pp. 1–5,

– *These authors contributed equally.

[II] A. Aspeel, **A. Gouverneur**, Raphaël M. Jungers, B. Macq, "Optimal Intermittent Particle Filter", *Transactions on Signal Processing*, vol. 70, pp. 2814–2825, Jan. 2022.

[III] R. Bongole, **A. Gouverneur**, B. Rodríguez-Gálvez, T. J. Oechtering, M. Skoglund, "Information-Theoretic Minimax Regret Bounds for Reinforcement Learning based on Duality", in *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025, pp. 1–5.

[IV] R. Bongole, **A. Gouverneur**, T. J. Oechtering, M. Skoglund, "Information-Theoretic Minimax Regret Upper Bounds for Reinforcement Learning Problems", in *2025 IEEE Information Theory Workshop (ITW)*, IEEE, 2025, pp. 1–5.

Acknowledgements

First and foremost, I would like to thank my supervisors, *Prof. Mikael Skoglund* and *Prof. Tobias Oechtering*. Mikael, I am deeply grateful for your trust, your flexibility, and the freedom you gave me to pursue my research. Your steady support and confidence allowed me to explore ideas deeply and grow as an independent researcher. Tobias, I want to thank you for your optimism, your quick and thorough reviews, and the positive energy you bring to every discussion. Your thoughtful comments and encouragement have been invaluable throughout this journey.

I want to extend my heartfelt gratitude to my co-author, *Dr. Borja Rodríguez Gálvez*, whom I have grew to consider as my PhD big brother. Borja, by your example, you taught me academic meticulousness and a constant striving for rigor. Maybe more importantly, you showed me that it could be done while being accessible to all and generous with your time. I feel very lucky to have collaborated with you.

I am deeply thankful to everyone involved in the defence of this thesis. I am very grateful to *Prof. Gergely Neu*, University of Barcelona, for agreeing to act as my dissertation opponent. I also extend my sincere thanks to the grading committee, *Prof. Alexandre Proutière*, KTH Royal Institute of Technology, *Prof. Thomas Schön*, Uppsala University, and *Dr. Emilie Kaufmann*, Université de Lille, for investing time and effort into reading and evaluating this dissertation. I would like to extend my gratitude to *Prof. Henrik Sandberg*, KTH Royal Institute of Technology, for advance-reviewing this work.

I would like to thank *Prof. Benjamin Van Roy* at Stanford University, who very generously opened the doors of his lab and welcomed me for a quarter. Ben, I was very inspired by your sense of the big picture and your enthusiasm for tackling big research questions. Working in your lab was a blast, and I am very grateful for that experience. I also thank *Dr. Hong Jun Jeon*, *Dr. Saurabh Kumar*, *Yifan Zhu*, *Wanqiao Xu*, *Anmol Kagrecha*, and *Henrik Marklund* for making me feel so welcome in the group.

I am grateful to my colleagues at Lynx Asset Management for a rewarding internship. I thank *Ola Backman* for inviting me to join the algo team, and *Dr. Per Hallberg*, *Jonas Johansson*, and *Gustav Grudd* for setting up such an interesting research project and letting me try my ideas. I learned a lot from working with you.

I would like to thank the *WASP program*, that funded my research and made this possible. I am thankful for all the people involved in the graduate program, for the valuable curriculum and fun workshops they arranged.

I would also like to acknowledge the excellent bachelor's and master's students whose theses I had the chance to mentor, *Reza Qorbani*, *Kevin Pettersson*, *Edwin Östlund*, *Aron Malmborg*, *Tim Persson*, *Markus Palmheden*, *Zhen Tian*, *Guangze*

Shi, and *Daniel Pérez*. It was a pleasure to support you along the way.

I spent my PhD years in the *ISE division*, and I am very happy I did as I got to share the division with such amazing colleagues and researchers. *Dr. Anubhab Ghosh*, it was a real pleasure to be your office roommate, you always listened to my half thought through ideas and you made the "division's darkest office" a cheerful place. Special thanks to *Dr. Antoine Honoré*, *Martin Lindström*, and *Wendi Löffler*, whom I could always count on when I needed to distract someone for a coffee, or just a chat. I would also like to thank *Dr. Sara Saeidian*, *Dr. Baptiste Cavarec*, and *Dr. Germán Bassi*. You were among my first contacts in the division when much of our work was still online and remote, and each of you inspired me with your way of doing research and by being true examples of good researchers.

I would also like to thank all my other colleagues at the division who made these years so memorable. From the endless lunch discussions, to the banana thief mystery, ski retreats, and movie nights, working in the division has been a fun experience. I feel truly fortunate to have shared this journey with so many of you, and to be able to call many of you friends.

I am grateful to have worked alongside and learned from those who have since moved on to new adventures, including *Dr. Håkan Carlsson*, *Dr. Alireza Javid*, *Dr. Zhendong Wang*,*Dr. Lissy Pellaco*, *Dr. Prakash Borpatra Gohain*, *Dr. Sina Molavipour*, *Dr. Sandipan Das*, *Dr. Sahar Imtiaz*, *Dr. Hamid Ghourchian*, *Dr. Boules Mouris*, *Leandro Miguel Lopez*, *Dr. Samie Mostafavi*, *Dr. Linghui Zhou*, *Dr. Ramana Avula*, *Dr. Xuechun Xu*, *Dr. Movitz Lenninger*, *Dr. Aris Kanellopoulos*, and *Dr. Javier Kipen*, among others. It would like to also thank and wish good luck to those that continue doing research at ISE: *Dr. Amirreza Zamanisiboni*, *Dr. Sajad Daei Omshi*, *Raghav Bongole*, *Sangwon Seo*, *Neelabhro Roy*, *Steven Rivetti*, *Vishnu Narayanan Moothedath*, *Jeannie He*, *Leo Grosse*, *Jaume Anguera Peris*, *Yusen Wang*, *Gustav Bohlin*, *Wei Shi*, *Afroditi Letsiou*, *Ding Shi*, *Maryam Bandali*, *Mengyuan Zhao*, *Sicong Ye*, *Cheng Yang*, *Shudi Weng*, *Jingwen Fu*, *Jonas Carlsson*, *Fatemeh Ebadi*, *Cheng Yang*, *Abolfazl Changizi*, *Yanghang Zeng*, *Dario Filatrella*, *Fredrik Cumlin*, and many others. Sharing the division with you all has been a true privilege.

Throughout this time, I have also met researchers outside ISE who have helped shape my ideas and my approach to research. Some of them became collaborators, others friends. I am particularly grateful to *Frédéric Zheng* and *Dr. Alexander Fengler* for their friendship, and stimulating conversations.

To my friends and family, thank you for being present in ways big and small, for checking in when deadlines were close, for celebrating the small wins, and for reminding me to keep perspective. To my parents, your example has been my compass, your encouragement has carried me further than you know. To my wife, your patience, your curiosity, and your care make every day better, you inspire me to try harder and to be kinder, and I am endlessly grateful to share this life with you.

Amaury Gouverneur
Stockholm, November 2025

Contents

Part I

Overview and Summary

CHAPTER **1**

Introduction

Learning to solve sequential decision problems is a central challenge in artificial intelligence. Reinforcement learning addresses this challenge by studying how an agent can learn effective behavior through interaction with a system it can observe and influence. At each step, the agent perceives its current situation, takes an action, and receives feedback in the form of a reward. Over time, by acting, observing, and adjusting its behavior, the agent seeks to maximize the total reward it obtains [1–3].

To make these ideas concrete, consider a few familiar examples. A chess player chooses a move not merely for its immediate tactical gain, but for the position it may lead to several turns later. A trading execution algorithm must decide which limit order to place, weighing the immediate benefit of fast execution against the long-term cost of market impact. An online platform deciding which article to recommend today must learn from whether the user clicks, and use that experience to make better suggestions tomorrow. In each of these cases, decisions are not isolated: they are part of an ongoing dialogue between an agent and its environment. The agent must sense, act, and adapt, continually revising its behavior in light of the consequences it experiences. The history of reinforcement learning is a story of how these principles have been turned into powerful methods.

The roots of reinforcement learning trace back to optimal control and dynamic programming, where the goal was to compute optimal decision rules for engineered systems such as inventory management, sequential resource allocation, and robotic control. In the late 1950s, Bellman [4] introduced the value function and the Bellman equation, framing sequential decision making as a recursive optimization

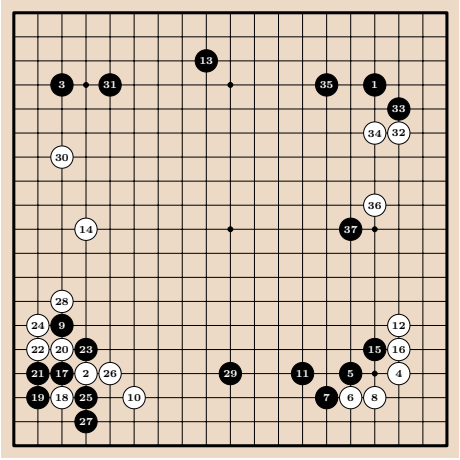


Figure 1.1: The first 37 moves from game 2 between AlphaGo and Lee Sedol in 2016, a game that AlphaGo ultimately won. The game of Go was long considered out of reach for existing learning algorithms due to its immense search space and the astronomical number of possible board positions.

over states and actions. This framework gave rise to policy and value iteration algorithms [5, 6], but also revealed the curse of dimensionality: the computational and memory requirements of dynamic programming grow exponentially with the number of state variables. Moreover, classical dynamic programming assumes full knowledge of the model, transition probabilities and reward functions must be specified exactly, which limits its applicability in complex, uncertain environments. These limitations motivated approximate and data-driven approaches [7].

A second line of work developed temporal-difference learning, where value functions are learned directly from experience [8]. Combined with off-policy updates, this led to Q-learning [9], demonstrating that effective control can be achieved without an explicit model of the environment. With the advent of neural networks as function approximators, reinforcement learning scaled to high-dimensional problems. The breakthrough occurred with Deep Q-Networks, which coupled Q-learning with convolutional networks to learn directly from pixels and achieve human-level control on many Atari games [10]. Shortly after, policy-gradient and actor–critic methods with deep function approximation broadened the toolbox for continuous control and large action spaces. Landmark systems such as AlphaGo (Figure 1.1) integrated deep learning, search, and reinforcement learning to achieve superhuman performance in Go [11].

Today, reinforcement learning extends far beyond games and research demonstrations. It is used in large-scale industrial systems to optimize content

recommendations and digital advertising, guide adaptive experimentation in drug development [12], support financial decision-making in portfolio management, and enable autonomous control in robotics [13, 14], manufacturing, and transportation [15–17]. More recently, reinforcement learning has become central to the post-training of large language models [18]. Reinforcement learning from human feedback [19] aligns models with human preferences [20], while reinforcement learning from verifiable rewards [21] uses objective signals, such as program correctness, formal proof verification, or the success of multi-step reasoning, to drive improvement. Together, these techniques are transforming large language models from static conversational systems into adaptive agents capable of solving complex tasks with minimal supervision.

Reinforcement learning can be distinguished by the type of feedback it receives and by the way learning unfolds over time. Unlike supervised learning [22], where each example specifies the correct action or label independently of what the model predicts, reinforcement learning relies on *evaluative* feedback: the agent is told only how good its actions were, not which action it should have taken. Learning therefore proceeds by trial and error, and requires the agent to explore unfamiliar actions to discover which ones lead to favorable outcomes. Moreover, the effects of actions are often delayed: what the agent does now shapes the situations it will encounter later and influences future opportunities for reward. Because its own behavior determines the data it observes, the agent must learn not only *from* experience but also *through* experience, continually balancing the need to gather new information with the desire to exploit what it already knows (Figure 1.2). Together, these features make reinforcement learning both more general and more challenging than supervised learning, and they define it as the study of how agents can acquire effective behavior through interaction with their environment.

A natural way to study this dilemma in isolation is through the multi-armed bandit problem [23]. In this simplified setting, the agent repeatedly chooses among a fixed set of actions, "arms", each yielding uncertain rewards drawn from a stationary distribution. There are no states or transitions to consider, and the environment does not change in response to the agent's choices. Despite this simplicity, the bandit framework already captures the essence of decision-making under uncertainty: the need to explore to learn which actions are good, and to exploit that knowledge to maximize cumulative reward over time.

To evaluate how well an algorithm solves this dilemma, one measures its *regret*, the expected difference between the cumulative reward of the algorithm and that of an oracle that always selects the best possible action. Regret quantifies learning efficiency: an algorithm with small regret learns quickly, while large regret indicates inefficiency or poor exploration. The goal of bandit and reinforcement learning algorithms is to minimize regret over time by efficiently balancing exploration and exploitation.

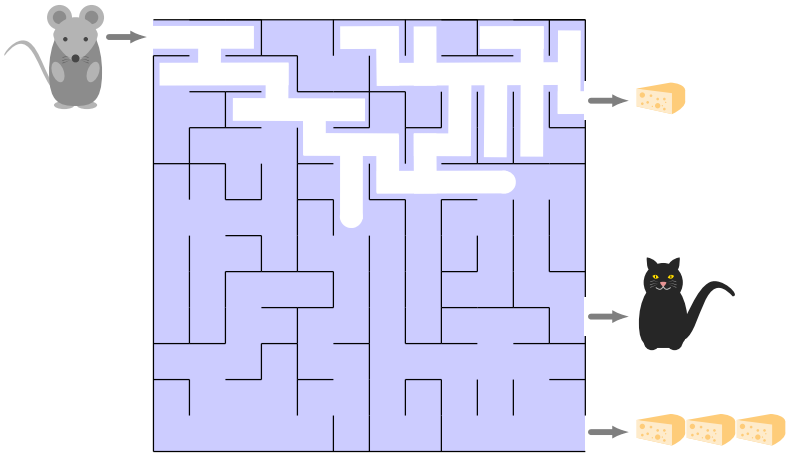


Figure 1.2: A mouse navigating a maze faces the *exploration–exploitation dilemma*: it can explore further in search of more cheese, or exploit what it has already found.

Over the last decade, information theory has emerged as a powerful tool for analyzing this trade-off. Originally developed by Claude Shannon [24] in the late 1940s, information theory was designed to quantify the transmission of messages over noisy communication channels. Its central idea is that information measures the reduction of uncertainty, the more uncertain we are before observing a signal, and the more certain we become afterward, the more information that signal conveys. Although first conceived for communication systems, information theory extends far beyond its original domain and provides rigorous tools for reasoning about knowledge, uncertainty, and learning [25–29]. In the context of reinforcement learning and bandit problems, it offers principled tools to quantify the value of information, how much useful knowledge an agent gains from each observation, and to relate this information gain to the agent's performance.

A particularly influential application of this information-theoretic perspective is the framework introduced by Russo and Van Roy [30–32]. Their analysis formalizes the exploration–exploitation trade-off through the concept of the *information ratio*, which relates the *expected instantaneous regret* to the *amount of information* an agent acquires about the optimal action. The smaller this ratio, the more efficiently the algorithm balances this trade-off. This simple but powerful idea provides general regret bounds that hold for any algorithm with bounded information ratio. One algorithm fits this framework particularly well,

the Thompson Sampling algorithm [33], for which their analysis recovers classical optimal rates for unstructured settings.

Building on this foundation, the present thesis aims to extend and refine the information-theoretic analysis of reinforcement learning. While the original framework captures the essence of information-efficient learning, some open questions remain about how these ideas can be applied to richer models and continuous domains. Our goal is to broaden the scope of the theory to settings where actions, rewards, or contexts have additional structure, and to develop tools that reveal how this structure can be exploited to achieve tighter regret guarantees.

Objective 1. *We aim to extend information-theoretic analyses of the Thompson Sampling algorithm to settings with infinite and/or continuous action and parameter spaces.*
The first objective of this thesis is to extend the information-theoretic analysis of bandit problems to infinite and continuous action spaces, where classical entropy-based analysis no longer yield meaningful bounds.

Objective 2. *We are interested in applying information-theoretic analysis to complex bandit problems.*
A second objective of this thesis is to derive *general information theoretic regret bounds* to more complex bandit settings, such as contextual, logistic, and generalized linear bandits, where rewards may be depend on the actions through a nonlinear, high-dimensional, or context-dependent function, and where direct application of previous results often leads to bounds that either fail or degrade with problem parameters.

Objective 3. *We aim to extend information-theoretic analysis beyond bandits to general reinforcement learning problems.*
A final objective of this thesis is to take a step beyond bandit models and study environments that evolve over time. We investigate how information-theoretic tools can be adapted to characterize the fundamental performance limits of such problems and to derive general learning guarantees, such as PAC-Bayesian bounds, for model-based and offline reinforcement learning.

1.1 Structure of this thesis

This thesis is organized into three chapters that move from background and motivation to core theoretical developments, and finally to conclusions and future directions.

- **Chapter 2: Background and Scope.** This chapter introduces the reinforcement learning and bandit frameworks and reviews the key information-theoretic

concepts, mutual information, entropy, and the information ratio, that form the basis of our analysis. It situates the thesis within the literature presents the information-theoretic analysis of Russo and Van Roy [31], introduces the Thompson Sampling algorithm, and formulates the research questions that guide the subsequent chapters.

- **Chapter 3: Key Contributions and Results.** This chapter presents the main theoretical contributions. It first completes the analysis of linear bandits by extending rate–distortion arguments to infinite or continuous action spaces and by introducing a chaining-based multi-resolution approximation that achieves near-optimal regret. It then develops information-theoretic bounds for more complex bandit models, including contextual, logistic, and generalized linear bandits. Finally, it connects these techniques with PAC-Bayesian methods and extends the analysis to reinforcement learning problems with state dynamics.

- **Chapter 4: Conclusions and Future Work.** The final chapter summarizes the main findings, discusses their implications for exploration strategy design, and highlights remaining limitations. It concludes with open problems and perspectives for future research on information-theoretic approaches to bandits and reinforcement learning.

CHAPTER **2**

Background and Scope

This chapter provides the conceptual and technical foundations for the results presented in this thesis. We begin in Section 2.1 by outlining the general framework of reinforcement learning, which formalizes how an *agent* interacts with an *environment*, that is, the system whose state it observes and influences through its actions, in order to learn effective behavior over time. Within this framework, we then focus on a simpler but equally fundamental setting, the multi-armed bandit problem. Bandit models capture the essence of the exploration–exploitation dilemma while abstracting away the complexities of state transitions, making them a natural starting point for theoretical analysis.

To study the performance of learning algorithms, we will use information-theoretic tools, introduced in Section 2.2, particularly the notions of entropy and mutual information, which quantify respectively uncertainty and information gain. Building on these tools, Section 2.3 reviews the information-theoretic framework introduced by D. Russo and Van Roy [31], which relates the Bayesian regret of a learning algorithm to the amount of information it gains through interaction with the environment. This framework, based on the information ratio and closely connected with the Thompson Sampling algorithm [33, 34], forms the foundation for the analyses developed in the following chapters and motivates the research questions addressed in this work.

Finally, Section 2.4 presents the scope and objectives of this thesis, summarizing how the subsequent chapters build upon these foundational concepts to advance the theoretical understanding of information-theoretic methods in reinforcement learning and bandit problems.

2.1 A Primer on Reinforcement Learning and Bandits

Reinforcement learning is a general framework for sequential decision-making under uncertainty [2]. In its most abstract form, an *agent* interacts with an *environment* over a sequence of rounds. At each time step, the agent selects an action and the environment responds with an observation and a numerical reward. The reward quantifies how desirable the outcome is in light of the agent's objectives. The goal is to choose actions over time to maximize the cumulative reward, using information gathered from past interactions. The framework makes minimal assumptions about how actions affect outcomes: the environment may be stochastic, partially observable, or even nonstationary.

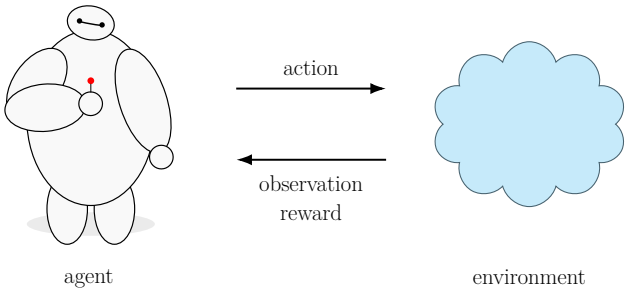


Figure 2.1: Schematic illustration of the reinforcement learning framework. The agent observes, acts, and receives feedback in the form of rewards.

This agent–environment interaction loop captures a wide range of real-world problems. In *robotics*, actions are motor commands, observations are sensor readings, and rewards measure task success [35, 36]. In *recommendation systems*, actions correspond to content suggestions, outcomes are user responses, and rewards quantify engagement [15, 37–39]. In *operations research* and *logistics*, actions represent allocation decisions, outcomes reflect realized demand, and rewards measure efficiency or profit [40, 41].

In this thesis, we focus on model-based reinforcement learning problems [1, 42]. We present in Section 2.1.1 the classical formulation of model-based reinforcement learning as a Markov decision process. We then introduce contextual bandits in Section 2.1.2, a setting in which each decision is informed by freshly sampled contextual information but where the agent's actions do not affect the environment. Finally, in Section 2.1.3, we present the multi-armed bandit problem, the simplest model in which the agent selects actions that directly generate rewards without any additional dependencies.

2.1.1 Model-Based Reinforcement Learning as an MDP

A widely studied subclass of reinforcement learning problems assumes that interaction follows a *Markov decision process* (MDP) [4]. In an MDP, the environment maintains a latent state that captures all information relevant for predicting future outcomes, so that the effects of an action depend only on the current state and not on the past history. At each time step t, the agent observes a state $S_t \in \mathcal{S}$, selects an action $A_t \in \mathcal{A}$ according to a policy π, receives a reward $R_t \in \mathbb{R}$, and the environment transitions to a new state S_{t+1}. We denote the interaction history available at the beginning of round t by $H^t := \{S_1, A_1, R_1, \ \ldots, \ S_{t-1}, A_{t-1}, R_{t-1}\}$.

The environment is characterized by an unknown parameter $\Theta \in \mathcal{O}$ that governs both the reward and transition dynamics. The process begins with an initial state drawn from the distribution $\mathbb{P}(S_1 \mid \Theta)$, produces rewards according to the reward kernel $\mathbb{P}(R_t \mid S_t, A_t, \Theta)$, and evolves through the transition kernel $\mathbb{P}(S_{t+1} \mid S_t, A_t, R_t, \Theta)$. Equivalently, one may write $R_t = R(S_t, A_t, \Theta)$, where $R : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \to \mathbb{R}$ is a stochastic process such that $R(S_t, A_t, \Theta)$ has conditional distribution $\mathbb{P}(R_t \mid S_t, A_t, \Theta)$.

In the *Bayesian* setting, the environment parameter is a random variable Θ drawn once from a known prior $\mathbb{P}(\Theta)$ and remains fixed throughout the interaction. In the *frequentist* setting, $\theta \in \mathcal{O}$ is an arbitrary but unknown element of a known parameter set. Unless stated otherwise, we will adopt the *Bayesian setting* in the following.

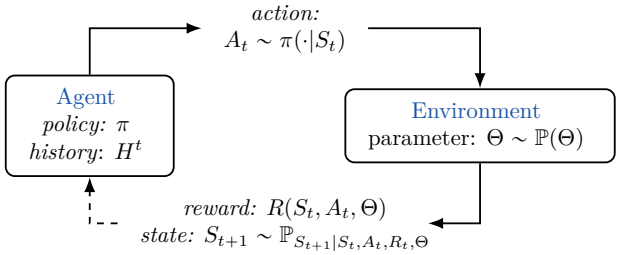


Figure 2.2: Agent–environment interaction in a parameterized MDP.

A policy π maps each state S_t to a distribution over actions and is chosen on the basis of the agent's past observations. The agent's objective is to find a policy π that maximizes the expected cumulative reward over T rounds,

$$\max_{\pi} \ \mathbb{E}\left[\sum_{t=1}^{T} R(S_t, A_t, \Theta)\right],$$

where the expectation is taken with respect to the joint distribution induced by the policy, the stochastic transitions, and (in the Bayesian case) the prior over Θ.

2.1.2 Contextual Bandits

The *contextual bandit* [43–48] model can be viewed as an MDP without state transitions: a context is sampled independently at each round. At time t, the agent observes a context $X_t \in \mathcal{X}$ drawn from a distribution $\mathbb{P}_X$, selects an action $A_t \in \mathcal{A}$, and receives a reward $R_t \in \mathbb{R}$ that depends on the context, the selected action, and on the environment parameters. The history available at round t is $H^t := \{X_1, A_1, R_1, \ \ldots, \ X_{t-1}, A_{t-1}, R_{t-1}\}$. Here and throughout, we write the reward as a stochastic function $R(X_t, A_t, \Theta)$ to emphasize its dependence on the context, the chosen action, and the unknown parameter.

Contextual bandits are useful in applications where decisions must adapt to changing conditions [40]. For example, in content recommendation, the context includes the user profile or recent activity and the agent must choose which content to display next. Learning in this setting requires generalising across contexts and identifying how actions perform in different situations.

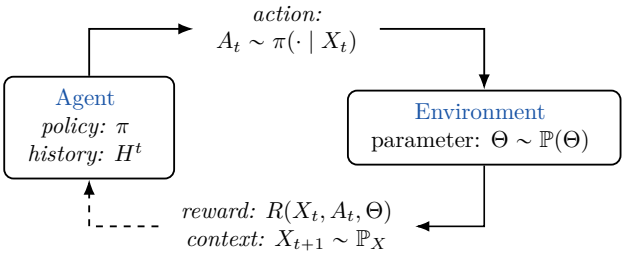


Figure 2.3: Agent–environment interaction in a contextual bandit setting.

A policy π maps each context x to a distribution over actions. The goal of the agent is to sequentially select actions that maximize the total expected reward, or equivalently, that *minimize the total expected regret* defined as:

$$\mathbb{E}[\mathrm{Regret}(T)] := \mathbb{E}\left[\sum_{t=1}^{T} R(X_t, A_t^\star, \Theta) - R(X_t, A_t, \Theta)\right],$$

where $A_t^\star$ denotes the *optimal action* for the observed context X_t and parameter Θ. We define the optimal decision rule $\pi_\star(x, \theta) \in \arg\max_{a \in \mathcal{A}} \mathbb{E}[R(x, a, \theta)]$ so that we may write $A_t^\star = \pi_\star(X_t, \Theta)$[1]. The expectation in the regret expression is taken over the randomness in the contexts, rewards, the learner's policy, and the prior over Θ.

[1] If multiple actions are optimal for a given $(x, \theta) \in (\mathcal{X} \times \mathcal{O})$, we arbitrarily fix the mapping to one of the optimal actions. The function $\pi_\star$ is therefore well defined.

2.1.3 The Bandit Problem

The *multi-armed bandit* setting [49] can be interpreted as a special case of an MDP with no state. At each round t, the agent selects an action $A_t \in \mathcal{A}$ and receives a reward $R_t \in \mathbb{R}$ that depends on the selected action and on the underlying environment parameters Θ, so that we can write the reward as a stochastic function $R(A_t, \Theta)$. The history available at round t is $H^t \coloneqq \{A_1, R_1, \ \ldots, \ A_{t-1}, R_{t-1}\}$.

The name *multi-armed bandit* originates from the classical gambling analogy by Robbins [23] in which a player faces several slot machines (the *one-armed bandits*), each associated with an unknown payoff distribution. The objective is to identify, through repeated play, the machine with the highest expected payout.

A policy π is a probability distribution over the action set $\mathcal{A}$ that depends on the observed history H^t. The goal of the agent is to sequentially select actions that maximize the total expected reward, or equivalently, that *minimize the total expected regret* defined as

$$\mathbb{E}[\mathrm{Regret}(T)] \coloneqq \mathbb{E}\left[\sum_{t=1}^{T} R(A^\star, \Theta) - R(A_t, \Theta)\right],$$

where $A^\star$ denotes the *optimal action* associated with the parameter Θ. We define the optimal decision rule $\pi_\star(\theta) \in \arg\max_{a \in \mathcal{A}} \mathbb{E}[R(a, \theta)]$ so that we may write $A^\star = \pi_\star(\Theta)$[2]. The expectation in the regret expression averages over the randomness of the rewards, the learner's policy, and the prior over Θ.

2.1.4 Notation

Throughout the thesis, we write random variables with capital letters (e.g., X), their realizations with lowercase letters (e.g., x), and their outcome spaces in calligraphic letters (e.g., $\mathcal{X}$).

Since the σ-algebras generated by the history H^t will be used frequently, we denote by

$$\mathbb{E}_t[\cdot] \coloneqq \mathbb{E}[\cdot \mid H^t] \qquad \text{and} \qquad \mathbb{P}_t(\cdot) \coloneqq \mathbb{P}(\cdot \mid H^t)$$

the conditional expectation and conditional probability given the history H^t.

We write $\mathrm{I}(X;Y)$ for mutual information between the random variables X and Y, and for each t, we let $\mathrm{I}_t(X;Y)$ denote the mutual information computed under the posterior measure $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot \mid H^t)$, so that $\mathrm{I}_t(X;Y)$ is a random variable whose expectation over the history satisfies $\mathbb{E}[\mathrm{I}_t(X;Y)] = \mathrm{I}(X;Y \mid H^t)$.

[2] Again, if multiple actions are optimal for a given parameter, we arbitrarily fix the mapping for that parameter to one of the optimal actions. The function $\pi_\star$ is therefore well defined.

2.2 Preliminaries on Information Theory

Information theory provides a mathematical language for reasoning about uncertainty, information, and learning. Originally developed by Shannon to study communication over noisy channels, it has since become a fundamental tool for analyzing statistical and learning systems. In reinforcement learning and bandit problems, information-theoretic quantities such as *entropy* and *mutual information* allow us to formalize the value of information and to quantify how difficult it is for an agent to learn an effective decision-making strategy. In this section, we review these core notions.

Definition 2.1 (Entropy)**.** *Let Z be a random variable with distribution $\mathbb{P}_Z$ and density or mass function p_Z. The* entropy *of Z is defined as*

$$\mathrm{H}(Z) := -\,\mathbb{E}\left[\log p_Z(Z)\right].$$

Entropy measures the uncertainty associated with the random variable Z. If all outcomes of Z are equally likely, the entropy is large, reflecting high uncertainty; if one outcome occurs with certainty, the entropy is zero. It can be interpreted as the expected number of bits required to describe a realization of Z, assuming an optimal code based on its distribution. In learning problems, entropy often quantifies the *agent's prior uncertainty* about a latent quantity of interest, such as the best action $A^\star$ or the model parameters Θ.

When Z takes values in a finite set $\mathcal{Z}$, its entropy is always bounded as

$$0 \leqslant \mathrm{H}(Z) \leqslant \log|\mathcal{Z}|,$$

with equality on the right if and only if Z is uniformly distributed on $\mathcal{Z}$.

Definition 2.2 (Conditional entropy)**.** *Let Z and W be random variables with joint distribution $\mathbb{P}_{Z,W}$ and conditional distribution $\mathbb{P}_{Z|W}$. The* conditional entropy *of Z given W is defined as*

$$\mathrm{H}(Z \mid W) := -\,\mathbb{E}[\log p_{Z|W}(Z \mid W)].$$

Conditional entropy measures the remaining uncertainty about Z once W is known. Two fundamental properties summarize how entropy behaves under conditioning and composition [50, Thm. 1.4 and 2.6].

Proposition 2.3 (Properties of entropy)**.** *Let $Z, W, Z_1, \ldots, Z_n$ be random variables. Then:*

- Conditioning reduces entropy: $\mathrm{H}(Z \mid W) \leqslant \mathrm{H}(Z)$.
- Chain rule: $\mathrm{H}(Z_1, \ldots, Z_n) = \sum_{i=1}^n \mathrm{H}(Z_i \mid Z_1, \ldots, Z_{i-1}) \leqslant \sum_{i=1}^n \mathrm{H}(Z_i)$

The first property expresses that knowing more information can only reduce uncertainty, while the second decomposes the total uncertainty of a collection of variables into a sum of incremental uncertainties.

Definition 2.4 (Relative entropy)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions on the same measurable space. The* relative entropy *(or* Kullback–Leibler divergence*) of $\mathbb{P}$ with respect to $\mathbb{Q}$ is defined as*

$$\mathrm{D}_{\mathrm{KL}}(\mathbb{P} \,\|\, \mathbb{Q}) := \mathbb{E}_{x\sim\mathbb{P}}\left[\log \frac{dP}{dQ}(x)\right],$$

whenever $\mathbb{P} \ll \mathbb{Q}$, and $+\infty$ otherwise.

Relative entropy provides a quantification on how different two distributions are and it plays a central role in defining other information-theoretic quantities. It is always nonnegative and equals zero if and only if $\mathbb{P} = \mathbb{Q}$ almost surely [50, Thm. 1.4]. From an operational viewpoint, it represents the expected penalty, in extra bits, incurred when coding samples from $\mathbb{P}$ using a code optimized for $\mathbb{Q}$ [51].

Definition 2.5 (Mutual information)**.** *The* mutual information *between two random variables X and Y is*

$$\mathrm{I}(X;Y) := \mathrm{H}(X) - \mathrm{H}(X \mid Y) = \mathrm{D}_{\mathrm{KL}}\big(\mathbb{P}_{X,Y} \,\big\|\, \mathbb{P}_X \otimes \mathbb{P}_Y\big).$$

Mutual information quantifies the expected reduction in uncertainty about X after observing Y, or equivalently, how different the joint distribution $\mathbb{P}_{X,Y}$ is from the product of the marginals $\mathbb{P}_X \otimes \mathbb{P}_Y$.

Proposition 2.6 (Properties of mutual information)**.** *Let $X, Y, Z, X_1, \ldots, X_n$ be random variables. Then the following hold [50, Thms. 3.2, 3.7, 4.1]:*

- Nonnegativity: $\mathrm{I}(X;Y) \geqslant 0$*, with equality iff X and Y are independent.*
- More data, more information: $\mathrm{I}(X;Y,Z) \geqslant \mathrm{I}(X;Y)$.
- Data processing inequality: *If $X \to Y \to Z$, then* $\mathrm{I}(X;Y) \geqslant \mathrm{I}(X;Z)$.
- Chain rule: $\mathrm{I}(X;Z_1,\ldots,Z_T) = \sum_{t=1}^{T} \mathrm{I}(X;Z_t \mid Z_1,\ldots,Z_{t-1})$.

Definition 2.7 (Conditional mutual information)**.** *Let X, Y, Z be random variables. The* conditional mutual information *between X and Y given Z is defined as*

$$\mathrm{I}(X;Y \mid Z) = \mathrm{H}(X \mid Z) - \mathrm{H}(X \mid Y,Z).$$

Conditional mutual information measures how much knowing Y reduces uncertainty about X when Z is already observed, typically in sequential decision problems, Z will represent past observations.

Proposition 2.8 (Properties of conditional mutual information)**.** *Let X, Y, Z be random variables. Then:*

- *If Z is independent of (X,Y), then* $\mathrm{I}(X;Y \mid Z) = \mathrm{I}(X;Y)$ *[50, Thm. 3.7].*
- $\mathrm{I}(X;Y) \geqslant \mathrm{I}(X;f(Y))$*, with equality if f is invertible [50, Thm. 4.1].*

2.3 Information-Theoretic Analysis of Bandit Problems

Russo and Van Roy [31] introduced the first information-theoretic analysis of Bayesian regret for bandit problems, which serves as the starting point for the developments in this thesis. Their framework introduces the *information ratio*, a quantity capturing the balance between exploration and exploitation. They show that if an algorithm maintains an information ratio smaller than Γ, then its Bayesian regret scales as $O\Big(\sqrt{\Gamma T \,\mathrm{H}(A^\star)}\Big)$, relating the regret to both the prior uncertainty about the optimal action and the efficiency with which information is acquired.

2.3.1 Information Ratio and General Regret Bound

The key idea of this framework is to quantify the exploration-exploitation trade-off through the *information ratio*, which measures how much squared expected regret an algorithm incurs per unit of information gained about the optimal action.

Definition 2.9 (Information Ratio)**.** *For each round $t \in \{1, \dots, T\}$, the* information ratio *is defined as the random variable*

$$\Gamma_t \;:=\; \frac{\mathbb{E}_t\big[\, R(A^\star, \Theta) - R(A_t, \Theta)\,\big]^2}{\mathrm{I}_t\big(A^\star; A_t, R_t\big)},$$

where $\mathbb{E}_t[\cdot]$ denotes conditional expectation given the history H^t, and $\mathrm{I}_t(\cdot\,;\cdot)$ is mutual information computed under the posterior measure $\mathbb{P}(\cdot|H^t)$. The numerator is the squared expected instantaneous regret, while the denominator measures the information gained about the optimal action $A^\star$ from observing the pair A_t, R_t.

The strength of this framework is that a bound on the average expected information ratio[3] directly implies a bound on the cumulative Bayesian regret.

Proposition 2.10 (Bayesian Regret Bound [31, Prop. 1])**.** *For any algorithm, if the average expected information ratio is bounded, $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\Gamma_t] \;\leqslant\; \Gamma$, then the cumulative Bayesian regret after T steps satisfies*

$$\mathbb{E}[\mathrm{Regret}(T)] \;\leqslant\; \sqrt{\Gamma\, T\, \mathrm{H}(A^\star)},$$

where $\mathrm{H}(A^\star)$ denotes the prior entropy of the optimal action.

Proof. We begin by rewriting the Bayesian regret using the definition of the information ratio:

$$\mathbb{E}[\mathrm{Regret}(T)] = \sum_{t=1}^{T} \mathbb{E}\big[\, R(A^\star, \Theta) - R(A_t, \Theta)\,\big] = \sum_{t=1}^{T} \mathbb{E}\Big[\sqrt{\Gamma_t\, \mathrm{I}_t(A^\star; (A_t, R_t))}\Big].$$

[3] The original proof of D. Russo and Van Roy [31, Proposition 1] requires a *uniform* bound, $\Gamma_t \leqslant \Gamma$, on the information ratio at every time step $t \in \{1, \dots, T\}$. Our proof in Gouverneur et al. [52] relaxes this assumption and shows that it is sufficient for the *average* expected information ratio to be bounded.

Applying Cauchy–Schwarz in expectation, $\mathbb{E}[\sqrt{XY}] \leqslant \sqrt{\mathbb{E}[X]\mathbb{E}[Y]}$, we obtain

$$\begin{aligned}
\mathbb{E}[\mathrm{Regret}(T)] &\leqslant \sum_{t=1}^{T} \sqrt{\mathbb{E}[\Gamma_t]\, \mathrm{I}(A^\star; (A_t, R_t) \mid H^t)} \\
&\overset{(i)}{\leqslant} \sqrt{\Gamma T \sum_{t=1}^{T} \mathrm{I}(A^\star; (A_t, R_t) \mid H^t)} \\
&\overset{(j)}{=} \sqrt{\Gamma T\, \mathrm{I}(A^\star; H^T)} \\
&\overset{(k)}{\leqslant} \sqrt{\Gamma T\, \mathrm{H}(A^\star)}.
\end{aligned}$$

where (i) follows from the bound $\sum_{t=1}^{T} \mathbb{E}[\Gamma_t] \leqslant \Gamma T$ together with applying Cauchy–Schwarz over the sum; (j) follows from the chain rule for mutual information, $\sum_{t=1}^{T} \mathrm{I}(A^\star; (A_t, R_t) \mid H^t) = \mathrm{I}(A^\star; H^T)$; and (k) follows from the inequality $\mathrm{I}(A^\star; H^T) \leqslant \mathrm{H}(A^\star)$.

□

This regret bound is remarkably general: it applies to any bandit algorithm and any bandit model. To obtain concrete regret guarantees for a specific algorithm, one must analyze its specific information ratio Γ_t. One Bayesian algorithm is particularly suited for such analysis, the Thompson Sampling algorithm.

2.3.2 Thompson Sampling

A natural algorithm to study within this framework is *Thompson Sampling*. It works by selecting actions randomly based on their posterior probability of being optimal. Specifically, at each time step $t \in \{1, \ldots, T\}$, the agent samples a parameter estimate $\hat{\Theta}_t$ from the posterior distribution, conditioned on the history H^t. The agent then selects the action that is optimal for the sampled parameter, $A_t = \pi_\star(\hat{\Theta}_t)$, receives a reward R_t, and updates the history to $H^{t+1} = H^t \cup \{\hat{A}_t, R_t\}$. The pseudocode for Thompson Sampling is given in Algorithm 1.

Algorithm 1 Thompson Sampling algorithm

1: **Input:** parameter prior $\mathbb{P}(\Theta)$.
2: **for** $t = 1$ **to** T **do**
3: Sample a parameter estimate $\hat{\Theta}_t \sim \mathbb{P}(\Theta | H^t)$.
4: Take the corresponding optimal action $A_t = \pi_\star(\hat{\Theta}_t)$.
5: Collect the reward $R_t = R(A_t, \Theta)$.
6: Update the history $H^{t+1} = H^t \cup \{A_t, R_t\}$.
7: **end for**

Originally proposed by Thompson [33], the method has been rediscovered multiple times over the decades and was reintroduced to the machine learning community by Chapelle and L. Li [16] and Scott [53], who demonstrated its excellent empirical performance in contextual bandit problems.

In the multi-armed bandit setting with a finite number of actions and bounded rewards, D. Russo and Van Roy [31, Proposition 3] showed that the information ratio of Thompson Sampling is uniformly bounded by the number of actions, which in turns yields a bound on the Bayesian regret of order $O(\sqrt{|\mathcal{A}|T\log(|\mathcal{A}|)})$.

Proposition 2.11 (Worst-Case Bound)**.** *In a finite multi-armed bandit with $|\mathcal{A}|$ actions and rewards in $[-1,1]$, the information ratio of Thompson Sampling satisfies*

$$\Gamma_t \leqslant 2|\mathcal{A}|.$$

Consequently,

$$\mathbb{E}[\mathrm{Regret}(T)] \leqslant \sqrt{2|\mathcal{A}|\,T\,\mathrm{H}(A^\star)}.$$

Sketch of proof *The proof proceeds in three main steps. First, under Thompson Sampling, the action A_t chosen by the algorithm and the optimal action $A^\star$ are identically distributed conditional on the history. This symmetry allows the squared instantaneous regret to be expressed as the squared expected difference between two realizations of the same stochastic reward process $R(A,\Theta)$: one where A and Θ are jointly distributed and one where they are independent. Second, applying the Cauchy–Schwarz inequality bounds this squared difference by $|\mathcal{A}|$ times the expected variance of the mean reward conditioned on the optimal action $A^\star$. Finally, the Donsker-Varadhan variational inequality [54] relates this expected conditional variance to the mutual information $\mathrm{I}_t(A^\star; R(A_t,\Theta))$, yielding the bound $\Gamma_t \leqslant 2|\mathcal{A}|$.*

The results of D. Russo and Van Roy [31] are elegant and remarkably general, yet they leave some important questions open. First, the dependence on $\mathrm{H}(A^\star)$ becomes problematic in continuous or infinite action spaces, where the entropy may be unbounded. Second, the original analysis is essentially limited to the multi-armed bandit setting, and its extension to contextual bandits requires additional ideas. Finally, a natural direction is to understand the information ratio in bandit problems with richer structure and to investigate how such structure can be reflected in the resulting bounds.

2.4 Scope of this Thesis

The goal of this thesis is to build on the framework of Russo and Van Roy [31] and advance the information-theoretic analysis of reinforcement learning and particularly, the study of bandit problems. Our objectives are formulated more precisely through six research questions outlined below.

A first line of investigation concerns linear bandits. While existing analyses have established regret bounds for Thompson Sampling in this setting, these results rely on restrictive assumptions, such as finite action sets or well-behaved priors, and do not apply to continuous or infinite action spaces. Our first question is therefore the following:

Question 1: *Can we remove existing limitations in information-theoretic analyses of linear bandits and obtain tighter regret bounds, especially for infinite and continuous action spaces?*

A second objective is to move beyond the finite-armed and linear settings and develop *general information-theoretic regret bounds* for broader classes of problems. In particular, we are interested in contextual bandits with general reward models. This motivates the second research question:

Question 2: *Can we extend the information ratio regret analysis to contextual bandits with general rewards?*

As a concrete and practically relevant example, we focus in this thesis on the case of *logistic bandits*, where the reward function depends on a linear combination of features passed through a sigmoid link. Existing regret bounds for this setting often scale poorly with the parameter of the model, the logistic slope. This motivates our third question:

Question 3: *Focusing on logistic bandits, can we improve existing regret bounds that scale exponentially with the logistic slope parameter and depend on the number of actions?*

Beyond logistic rewards, a natural class of models that generalize linear and logistic bandits are *generalized linear bandits* [55], where the expected reward is given by a known link function applied to a inner product between the action and parameter. These models cover a broad spectrum of applications but current regrets bounds are not satisfactory [56, 19.4.7]. We therefore ask:

Question 4: *Can we extend the information-theoretic analysis to generalized linear bandits?*

A longer-term goal of this thesis is to understand whether the ideas developed in the bandit setting can be extended to more general reinforcement learning problems, where the agent interacts with an evolving environment and must plan over multiple steps. This leads to the fifth question:

Question 5: *Can we extend these ideas beyond bandits and propose a coherent framework to study reinforcement learning?*

Finally, we also explore how information-theoretic tools relate to PAC-Bayesian theory [57–59], which provides generalization guarantees for randomized learning algorithms and provide theoretical insight to popular reinforcement learning algorithms such as Trust Region Policy Optimization [60] and Proximal Policy Optimization [61]. In particular, we investigate whether PAC-Bayesian methods can be strengthened in the context of offline bandits. This motivates the final question:

Question 6: *Can we strengthen PAC-Bayesian guarantees for offline bandits?*

These questions define the research agenda of this thesis. In the following chapter, we address each of them through a series of theoretical contributions, supported by publications.

CHAPTER **3**

Key Contributions & Results

This chapter presents the main contributions of the thesis, each advancing the information-theoretic analysis of reinforcement learning problems. The results are grouped thematically, reflecting a progression from foundational bandit models to more complex settings, and ultimately toward reinforcement learning.

We begin by addressing linear bandits, where we refine existing regret analyses on the performance of the Thompson Sampling algorithm and extend them to infinite or continuous action spaces. We then move to more general bandit models, such as contextual and logistic bandits, demonstrating that the information ratio framework can accommodate contexts and nonlinearity. Finally, we explore how these tools can be applied in the broader context of reinforcement learning, both in the model based Bayesian setting and in offline learning with limited data.

3.1 Completing the Analysis of Linear Bandits

A central insight of the information–theoretic framework of Russo and Van Roy [31] is that the Bayesian regret of a learning algorithm can be controlled in terms of the entropy of the optimal action $A^\star$. In completely unstructured settings, this dependence reflects a genuine limitation. For example, in a bandit problem with N unrelated actions, where exactly one yields reward 1 and all others yield 0, the learner must identify the single good arm, and the uncertainty $\mathrm{H}(A^\star)$ precisely captures the (worst-case) difficulty of this search.

In practice, many bandit problems (most of the interesting ones) possess additional structure, and exact identification of the optimal action is not necessary to achieve low regret. A particularly important example is the *linear bandit* problem. In a d-dimensional linear bandit, both actions $a \in \mathcal{A}$ and environment parameters $\theta \in \mathcal{O}$ are represented by feature vectors, that is $\mathcal{A}, \mathcal{O} \subseteq \mathbb{R}^d$, and the expected reward is a bilinear function of these vectors, that is $\mathbb{E}[R(a,\theta)] = \langle a, \theta \rangle$ for all $a \in \mathcal{A}$ and all $\theta \in \mathcal{O}$. In this setting, similar actions are associated with similar expected rewards, and learning a good approximation of the optimal action can suffice to achieve small regret without perfectly identifying $A^\star$. Dong and Van Roy [62] formalize this intuition by constructing an approximation of the Thompson Sampling regret at each time step by considering an approximation of the linear bandit problem on a *compressed* parameter set $\mathcal{O}_\varepsilon$. On this compressed linear bandit problem, they construct Θ_ε and $\Theta_{\varepsilon,t}$, the projections respectively of the parameter Θ and the sampled parameter Θ_t and carefully craft a deviation of the original Thompson Sampling, which they name *one-step compressed Thompson Sampling*, such that

$$\mathbb{E}_t[R(A^\star,\Theta) - R(A_t,\Theta)] \leqslant \varepsilon + \underbrace{\mathbb{E}_t[R(\tilde{A}^\star_\varepsilon,\Theta) - R(\tilde{A}_{t,\varepsilon},\Theta)]}_{\textit{one-step compressed TS regret}},$$

where $\tilde{A}^\star_\varepsilon$ and $\tilde{A}_{t,\varepsilon}$ are based respectively on Θ_ε and $\Theta_{\varepsilon,t}$. They then show, that for finite action and parameter spaces, the cumulative *one-step compressed TS regret* can be controlled via the total information about the quantized parameter Θ_ε obtained by the *original Thompson Sampling algorithm* $\mathrm{H}(\Theta_\varepsilon | H^{T+1})$ and a uniform bound $\tilde{\Gamma}$ on the *one-step compressed information ratio* defined as the ratio between the squared *one-step compressed TS regret* and the information gained about Θ_ε by observing the pair $R(\tilde{A}_{t,\varepsilon},\Theta), \tilde{A}_{t,\varepsilon}$. For linear bandits with bounded rewards, they prove that the *one-step compressed TS regret* can be controlled by the dimension d of the problem. Summing over the approximation errors ε and optimizing over the choice of ε, they obtain a near-optimal regret rate of order $O(d\sqrt{T\log(T)})$. However, those results are limited to bandit problems with finite action and environment parameter space.

3.1.1 Linear Bandits with Infinite Action Spaces (Paper A)

Paper A builds on the rate-distortion perspective introduced by Dong and Van Roy [62] and extends the analysis to settings with infinite or continuous action and parameter spaces. To simplify the exposure of our results, we adapt their construction such that it depends on a *statistic* $A^\star_\varepsilon$, of the optimal action $A^\star$ instead of a statistic Θ_ε of the parameter Θ. We follow the main idea of their rate-distortion analysis, that is, at each time step, we construct an *approximation* of the instantaneous Thompson Sampling regret that can be controlled using an information theoretic analysis. We develop new proof techniques to show that such a construction can be achieved even with infinite and continuous action and parameter spaces. We establish that, for any action space $\mathcal{A}$ and parameter space $\mathcal{O}$, the

cumulative Bayesian regret of Thompson Sampling satisfies

$$\mathbb{E}[\mathrm{Regret}(T)] \;\leqslant\; \sqrt{\tilde{\Gamma} T \,\mathrm{H}(A^\star_\varepsilon)} + \varepsilon T,$$

where $\tilde{\Gamma}$ is an upper bound on the average expected *one-step compressed Thompson Sampling* information ratio.

When the expected reward function is L-Lipschitz with respect to a metric space $(\mathcal{A}, \rho)$, we can set $A^\star_\varepsilon$ to be the quantization of $A^\star$ on an ε-net, and the entropy $\mathrm{H}(A^\star_\varepsilon)$ can be related to the covering number of the action space. This yields the bound

$$\mathbb{E}[\mathrm{Regret}(T)] \;\leqslant\; \sqrt{\tilde{\Gamma} T \,\log\bigl(\mathcal{N}(\mathcal{A}, \rho, \varepsilon/L)\bigr)} \;+\; \varepsilon T,$$

which demonstrates that smoothness of expected rewards directly reduces the complexity of the problem.

In particular, for d-dimensional linear bandits with bounded rewards, the *one-step compressed Thompson Sampling* information ratio is bounded by the problem's dimension d. Then for problems with bounded rewards and where the action set is a bounded subset of $\mathbb{R}^d$, e.g. $\mathcal{A} \subseteq \mathbb{B}_d(0, 1)$, we show that the Bayesian regret of Thompson Sampling satisfies

$$\mathbb{E}[\mathrm{Regret}(T)] \;\leqslant\; 2d \sqrt{T \,\log\left(\sqrt{2} + \frac{4\sqrt{T}}{d}\right)}.$$

This recovers the near-optimal rate $O\bigl(d\sqrt{T \log(T)}\bigr)$ established in [62, Theorem 2], and extends its applicability to general action and parameters spaces.

3.1.2 Chaining Analysis and Metric Entropy (Paper B)

While Paper A uses a single Thompson Sampling regret approximation scale $\varepsilon \geqslant 0$, which we will denote $\mathbb{E}_t[\mathrm{Regret}(A^\star_\varepsilon)] := \mathbb{E}_t[R(\tilde{A}^\star_\varepsilon, \Theta) - R(\tilde{A}_{t,\varepsilon}, \Theta)]$, Paper B questions whether better bounds can be obtained by combining approximations at multiple scales. Inspired by classical chaining techniques [63] and its applications to the study of generalization error [64], Paper B decomposes the regret into a chain of differences between "approximate learning" regrets of increasing resolution:

$$\mathbb{E}_t[\mathrm{Regret}(A^\star)] = \sum_{k=1}^{\infty} \Bigl(\mathbb{E}_t[\mathrm{Regret}(A^\star_{\varepsilon_k})] - \mathbb{E}_t[\mathrm{Regret}(A^\star_{\varepsilon_{k-1}})]\Bigr),$$

where $\{\varepsilon_k\}_{k=0}^{\infty}$ is a decreasing sequence with $\varepsilon_0 \geqslant \mathrm{diam}(\mathcal{A})$ and $\lim_{k\to\infty} \varepsilon_k = 0$ such that $\lim_{k\to\infty} \mathbb{E}_t[\mathrm{Regret}(A^\star_{\varepsilon_k})] = \mathbb{E}_t[\mathrm{Regret}(A^\star)]$ and $\mathbb{E}_t[\mathrm{Regret}(A^\star_{\varepsilon_0})] = 0$.

The idea of Paper B is to consider each incremental regret difference, which we upper bounded using a *chained information ratio*, $\Gamma_{t,k}$, and the entropy of successive quantized versions of $A^\star_{\varepsilon_k}$. Controlling those incremental regret difference, requires the introduction of a modified Thompson Sampling algorithm, that we call 2-TS. The only difference with the original algorithm is that the posterior distribution is updated every two time steps. For bandit problems, the 2-TS expected cumulative regret after T steps is bounded as

$$\mathbb{E}[\mathrm{Regret}(T)] \leqslant \sum_{k=k_0+1}^{\infty} \sqrt{2 \cdot \bar{\Gamma}_k \cdot T \cdot \mathrm{H}(A^\star_k, A^\star_{k-1})},$$

where $A^\star_k$ is the k^{th}-quantization of the optimal action $A^\star$ with respect to $(\mathcal{A}, \rho)$ and where for each $k > 0$, $\bar{\Gamma}_k$ is an upper bound on $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\Gamma_{t,k}]$.

For linear bandits, if the rewards form a separable sub-Gaussian process on the metric space $(\mathcal{A}, \rho)$, we can derive a regret bound that depends on the *integral entropy* of the action space:

$$\mathbb{E}[\mathrm{Regret}(T)] \;\leqslant\; 12\sqrt{dT} \int_0^\infty \sqrt{2\log(|\mathcal{N}(\mathcal{A}, \rho, \varepsilon)|)}\, d\varepsilon,$$

where $|\mathcal{N}(\mathcal{A}, \rho, \varepsilon)|$ is the covering number of the action space at scale $\varepsilon \geqslant 0$. The requirement of the reward to form separable sub-Gaussian process is however very restrictive and we have not found an application that is completely artificial.

3.2 Extending to More Complex Bandit Settings

The next papers extend the scope of the information-theoretic analysis beyond linear bandits, addressing more complex bandit models where the optimal action may vary over time or where the reward structure departs from the linear bandit model. For each setting, we derive an improved information theoretic *regret bound* valid for any algorithm. We then specialize those bounds to Thompson Sampling algorithm and derive explicit regret rates via the analysis of a suitable information ratio.

3.2.1 Generalizing the Analysis of Contextual Bandits (Paper C)

A key challenge in contextual bandit problens is that the optimal action changes with the context, making the standard information ratio of D. Russo and Van Roy [31] insufficient. The lifted information ratio adapts to this setting by measuring the trade-off between regret and information gain *about the environment parameter* Θ by observing the reward conditionally on the history *and the observed context*. More precisely, the lifted information ratio at time t is defined as:

$$\Gamma_t := \frac{\mathbb{E}_t\left[R(A^\star_t, \Theta) - R(A_t, \Theta)\right]^2}{\mathrm{I}_t(\Theta; A_t, R_t)},$$

where $A_t^\star$ denotes the optimal action in context X_t, and where importantly $\mathbb{E}_t[\cdot]$ denotes conditional expectation given the history H^t *and the context* X_t, and similarly $\mathrm{I}_t(\cdot\,;\cdot)$ is mutual information computed under the posterior measure $\mathbb{P}(\cdot|H^t, X_t)$.

Paper C builds directly upon the recent work of Neu et al. [65], who employed the concept of the *lifted information ratio* to analyze Thompson Sampling in contextual bandit problems with binary rewards. We extend their analysis to be compatible with any reward, and show that any algorithm that has a bounded average expected *lifted information ratio* will enjoy a bounded cumulative regret. Specifically, we can bound the cumulative Bayesian as

$$\mathbb{E}[\mathrm{Regret}(T)] \;\leqslant\; \sqrt{\Gamma T\,\mathrm{H}(\Theta)},$$

where Γ is an upper bound on the average expected *lifted information ratio*.

We then analyze the *lifted information ratio* of Thompson Sampling algorithm. We establish a *worst-case* bound in terms of the cardinality of the action set $|\mathcal{A}|$. For d-dimensional linear bandits, we establish that the *lifted information ratio* can be bounded by the dimension d. We then combine those results to the general information-theoretic regret bound to derive specific regret rates for Thompson Sampling. Our analysis generalizes the results of Neu et al. [65] from binary rewards to *sub-Gaussian* and bounded rewards, and provides a simplified proof of the lifted regret bound using standard information-theoretic arguments. In particular, our proof of Theorem C.7 makes it explicit that if the mutual information can be effectively approximated, then it is possible to derive regret bounds for bandit problems with infinite or continuous action and parameter spaces. This result is conceptually distinct from the compression-based analysis of Paper A and Paper B that rely on constructing an approximation of the expected regret and will be important for deriving improved regret bounds for logistic bandits presented in Paper D.

3.2.2 Improving the Analysis of Logistic Bandits (Paper D)

The logistic bandit problem is a important example of a bandit model with nonlinear rewards. This model is applicable to many industrial applications, such as click-through rate prediction, spam email detection, and personalized advertisement [16, 66]. In this setting, the agent receives binary rewards whose success probabilities follows a logistic function of the inner product between the action $a \in \mathcal{A}$ and an unknown parameter $\theta \in \mathcal{O}$:

$$\mathbb{P}\left[R_t = 1\right] = \frac{\exp(\beta\langle a, \theta\rangle)}{1 + \exp(\beta\langle a, \theta\rangle)},$$

where $\beta > 0$ is a scale parameter known to the agent.

Previous regret bounds for logistic bandits, whether based on UCB [67] or Thompson Sampling [30, 62, 65], scale either *exponentially* with the logistic slope β or depend directly on the cardinality of the action space $\mathcal{A}$. This dependences are unsatisfactory and "it seems that there may be room for significant improvement" [56, 19.4.7]. In practice, the distinction between near-optimal and sub-optimal actions gets more pronounced as β increases, which can make it easier to find near-optimal actions. This gap between theoretical bounds and empirical observations was already pointed out by McMahan and Streeter [68], who called for improved regret analyses in the logistic bandit setting, a challenge that has long remained open.

Paper D resolves this issue by establishing logistic regret bounds for Thompson Sampling algorithm that *only scale logarithmically in* β. We obtain this results through two main contributions.

The first one is an improved general regret bound for logistic bandits. More specifically, we establish that, for any algorithm, if the average expected information ratio is bounded, $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\Gamma_t] \leqslant \Gamma$, for some $\Gamma > 0$, then the cumulative regret is bounded as

$$\mathbb{E}[\text{Regret}(T)] \leqslant \sqrt{\Gamma T\left(\mathrm{H}(\Theta_\varepsilon) + \tfrac{1}{2}\varepsilon^2\beta^2 T\right)},$$

where $\mathrm{H}(\Theta_\varepsilon)$ is the entropy of Θ quantized at scale $\varepsilon \geqslant 0$ and Γ_t is the ratio between the squared regret at time t and the information gained about Θ, *given the history*.

Our second contribution focusses on the Thompson Sampling algorithm for which we establish that the information ratio for logistic bandits is bounded by $d(4/\alpha)^2$ where d is the problem's dimension and $\alpha = \min_{\theta\in\mathcal{O}}\max_{a\in\mathcal{A}}\langle a, \theta\rangle$ is a *minimax* measure of the alignment between the action and parameter spaces. Crucially, this bound is independent of the number of actions and of the slope parameter β. Our first innovation is to control the information ratio by relating both regret and information gain to the *expected variance of the regret*, conditioned on the sampled parameter; lower bounding the information gain using the regret variance instead of the regret expectation. Our second innovation is to show that the limit case $\beta \to \infty$ can *serve as a uniform upper bound*, thereby simplifying the analysis. We believe the first innovation might be useful to study the Thompson Sampling information ratio for generalized linear bandits.

Combining those two contributions, we obtain our main result: a bound on the expected Thompson Sampling regret in $O(d/\alpha\sqrt{T\log(\beta T/d)})$ To the best of our knowledge, this is the first regret bound for *any logistic bandit algorithm* that does not scale exponentially with the logistic function's parameter β or with the problem dimension d and remains independent of the number of actions. Specifically, we show that, for any $\beta > 0$, and for all action and parameters inside the d-dimensional Euclidean unit sphere, $\mathcal{A}, \mathcal{O} \subseteq \mathbb{B}_d(0, 1)$ with minimax alignment constant $\alpha > 0$, under the logistic bandit setting with logistic function $\phi_\beta(x)$, the TS regret is

bounded as

$$\mathbb{E}[\text{Regret}(T)] \leqslant \frac{4d}{\alpha}\sqrt{T \log\left(3 + 6\beta\sqrt{\frac{T}{2d}}\right)}.$$

Finally, we present specific settings where the dependence on α can be controlled. For instance, when the action space fully encompasses the parameter space, $\mathcal{O} \subseteq \mathcal{A}$, the regret of Thompson Sampling scales as $\tilde{O}(d\sqrt{T})$.

An exciting direction for future work is to use the result in this paper to derive regret bounds for logistic bandits in the frequentist setting. A promising way is to apply our information-theoretic analysis to the optimistic information directed sampling algorithm introduced by Neu et al. [69]. We believe that this approach could lead to new and improved frequentist bounds for logistic bandits.

3.2.3 Extension to Generalized Linear Bandits

This subsection presents how the regret analysis techniques developed in Paper D can be extended to the class of *generalized linear bandits*. A generalized linear bandit model is a family of bandit problems based on the *exponential family of distributions*. At each time step $t \in \{1, \dots, T\}$, an agent selects an action $A_t \in \mathcal{A}$ and observes a reward $R_t \in \mathbb{R}$, drawn conditionally on A_t according to a distribution from the exponential family with *natural parameter* $\langle a, \theta \rangle$, where $\theta \in \mathcal{O}$ is an unknown parameter vector. Formally, the distribution of the reward associated with an action $a \in \mathcal{A}$ and parameter $\theta \in \mathcal{O}$ is given by

$$\mathbb{P}[R_t = r] = \exp(r\langle a, \theta \rangle - b(\langle a, \theta \rangle) + c(r)),$$

where $b : \mathbb{R} \to \mathbb{R}$ is the log-partition function, which is convex, strictly increasing, and twice differentiable, and $c : \mathbb{R} \to \mathbb{R}$ is a known base measure. This class of distributions has several useful properties. In particular, the conditional mean can be written as

$$\mathbb{E}[R_t \mid a, \theta] = \mu(\langle a, \theta \rangle) = b'(\langle a, \theta \rangle),$$

where $\mu(x) = b'(x)$ is the mean function of the generalized linear model, often referred to as the *link function*.

This formulation includes several standard bandit models as special cases. For linear Gaussian bandits, the model corresponds to $b(\langle a, \theta \rangle) = \frac{1}{2}\langle a, \theta \rangle^2$ and $c(r) = -\frac{1}{2}r^2 - \frac{1}{2\sigma^2}\log(2\pi)$, where σ^2 is the noise variance. For logistic bandits, where rewards are binary and follow a Bernoulli distribution with a logistic success probability, the log-partition function is $b(\langle a, \theta \rangle) = \log(1 + e^{\langle a, \theta \rangle})$ and the base measure is $c(r) = 0$.

Following [55, 67], we assume that the action space $\mathcal{A}$ and parameter space $\mathcal{O}$ are bounded compact subsets of $\mathbb{R}^d$. For convenience, and without loss of generality, we use the convention that actions lie in the d-dimensional Euclidean unit ball $\mathbb{B}_d(0,1)$ and parameters lie in a Euclidean ball of radius S, that is, $\mathcal{A} \subseteq \mathbb{B}_d(0,1)$ and $\mathcal{O} \subseteq \mathbb{B}_d(0,S)$.

Then for any algorithm such that the average expected information ratio (as defined in Paper D) is bounded by Γ, we can show that its regret is bounded in order $O(\sqrt{\Gamma T d \log(SCT/d)})$. More specifically, we establish that its regret is controlled by

$$\mathbb{E}[\text{Regret}(T)] \leqslant \sqrt{\Gamma T d \log(3 + 6S\sqrt{2CT/d})},$$

where $C \coloneqq \sup_{x\in[-S,S]} \mu'(x)$ be the supremum of μ' over the natural parameters.

The proof follows exactly the same technique as the proof of Theorem D.4 as relies on approximating the conditional mutual information $\mathrm{I}_t(\Theta; R(A_t,\Theta), \hat{\Theta}_t)$ with $\mathrm{I}(\Theta_\varepsilon; R_t \mid H^t, \hat{\Theta}_t)$ where Θ_ε is a quantization of Θ. The difference is that it bounds the remainder by using the fact that, for exponential families with log-partition function b, the KL divergence between distributions with natural parameter η and η' can be expressed as the Bregman divergence of b evaluated at (η', η).

Proof. We start by rewriting the expected regret using the information ratio:

$$\mathbb{E}[\text{Regret}(T)] = \sum_{t=1}^{T} \mathbb{E}[R(A^\star,\Theta) - R(A_t,\Theta)] = \sum_{t=1}^{T} \mathbb{E}\left[\sqrt{\Gamma_t \mathrm{I}_t(\Theta; R(A_t,\Theta), \hat{\Theta}_t)}\right].$$

Applying Cauchy-Schwarz for expectations, $\mathbb{E}[\sqrt{X}\sqrt{Y}] \leqslant \sqrt{\mathbb{E}[X]\,\mathbb{E}[Y]}$, followed by Cauchy-Schwarz over the sum, we obtain

$$\mathbb{E}[\text{Regret}(T)] \leqslant \sum_{t=1}^{T} \sqrt{\mathbb{E}[\Gamma_t]\mathrm{I}(\Theta; R(A_t,\Theta), \hat{\Theta}_t | H^t)} \leqslant \sqrt{\Gamma T \sum_{t=1}^{T} \mathrm{I}(\Theta; R(A_t,\Theta), \hat{\Theta}_t | H^t)}, \tag{3.1}$$

where in the last step, we used that by assumption $\sum_{t=1}^{T} \mathbb{E}_t[\Gamma_t] \leqslant \Gamma T$. Then, by the chain rule for mutual information [50, Theorem 3.7.b] we can write

$$\mathrm{I}(\Theta; R(A_t,\Theta), \hat{\Theta}_t | H^t) = \mathrm{I}(\Theta; \hat{\Theta}_t | H^t) + \mathrm{I}(\Theta; R(A_t,\Theta) | H^t, \hat{\Theta}_t) = \mathrm{I}(\Theta; R(A_t,\Theta) | H^t, \hat{\Theta}_t),$$

using that Θ and $\hat{\Theta}_t$ are independent given H^t, and that $\mathrm{I}(\Theta; \hat{\Theta}_t | H^t) = 0$.

Let $\mathbb{P}_{R_t|H^t,\hat{\Theta}_t,\Theta}$ and $\mathbb{P}_{R_t|H^t,\hat{\Theta}_t}$ denote the distribution of R_t conditioned respectively on $H^t, \hat{\Theta}_t, \Theta$ and $H^t, \hat{\Theta}_t$. Then, we have that

$$\mathrm{I}(\Theta; R_t | H^t, \hat{\Theta}_t) = \mathbb{E}_{H^t,\Theta,\hat{\Theta}_t}\left[\mathbb{E}_{R_t \sim \mathbb{P}_{R_t|\hat{\Theta}_t,\Theta}}\left[\log \frac{\mathbb{P}_{R_t|\hat{\Theta}_t,\Theta}(R_t)}{\mathbb{P}_{R_t|H^t,\hat{\Theta}_t}(R_t)}\right]\right], \tag{3.2}$$

where we used that $\mathbb{P}_{R_t|H^t,\hat{\Theta}_t,\Theta} = \mathbb{P}_{R_t|\hat{\Theta}_t,\Theta}$ since R_t is independent of H_t conditioned on Θ and $\hat{\Theta}$. Now, we let $\mathcal{O}_\varepsilon$ be an ε-net for $(\mathcal{O}, \|\cdot\|_2)$ with mapping $q : \mathcal{O} \to \mathcal{O}_\varepsilon$, and define $\Theta_\varepsilon := q(\Theta)$ the quantized parameter. We introduce $\overline{\mathbb{P}}_{R_t|H^t,\hat{\Theta}_t,\Theta_\varepsilon}$ to denote the reward distribution given $H^t, \hat{\Theta}_t, \Theta_\varepsilon$, which is obtained by averaging over all Θ' such that $q(\Theta') = \Theta_\varepsilon$, that is $\overline{\mathbb{P}}_{R_t|H^t,\hat{\Theta}_t,\Theta_\varepsilon}(\cdot) := \mathbb{E}\left[\mathbb{P}_{R_t|\hat{\Theta}_t,\Theta'}(\cdot)\middle|_{q(\Theta')=\Theta_\varepsilon}\right]$.

Then, starting from the definition of conditional mutual information, we add and subtract the log-probability $\log \overline{\mathbb{P}}_{R_t|H^t,\hat{\Theta}_t,\Theta_\varepsilon}(R_t)$ inside the inner expectation to obtain

$$(3.2) = \mathbb{E}\left[\mathbb{E}_{R_t\sim\mathbb{P}_{R_t|\hat{\Theta}_t,\Theta}}\left[\log \frac{\overline{\mathbb{P}}_{R_t|H^t,\hat{\Theta}_t,\Theta_\varepsilon}(R_t)}{\mathbb{P}_{R_t|H^t,\hat{\Theta}_t}(R_t)}\right] + \mathbb{E}\left[\log \frac{\mathbb{P}_{R_t|\hat{\Theta}_t,\Theta}(R_t)}{\overline{\mathbb{P}}_{R_t|H^t,\hat{\Theta}_t,\Theta_\varepsilon}(R_t)}\right]\right].$$

Using the law of total expectation, the outer expectation over Θ may be replaced by one over Θ_ε, and the first term is exactly the conditional mutual information $\mathrm{I}(\Theta_\varepsilon; R_t \mid H^t, \hat{\Theta}_t)$. We recognize the second term as the expected KL divergence between the true reward distribution $\mathbb{P}_{R_t|\hat{\Theta}_t,\Theta}$ and the quantized posterior $\overline{\mathbb{P}}_{R_t|H^t,\hat{\Theta}_t,\Theta_\varepsilon}$, which by convexity of the KL divergence is at most the average over the quantization:

$$\mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}(\mathbb{P}_{R_t|\hat{\Theta}_t,\Theta} \,\|\, \overline{\mathbb{P}}_{R_t|H^t,\hat{\Theta}_t,\Theta_\varepsilon})\right] \leqslant \mathbb{E}\left[\mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}(\mathbb{P}_{R_t|\hat{\Theta}_t,\Theta} \,\|\, \mathbb{P}_{R_t|\hat{\Theta}_t,\Theta'})\middle|_{q(\Theta')=\Theta_\varepsilon}\right]\right].$$

Conditioned on $\Theta = \theta$, $\Theta' = \theta'$ and $\hat{\Theta}_t = \hat{\theta}_t$, the rewards are drawn from a canonical exponential family with natural parameters $\eta = \langle \pi^\star(\hat{\theta}_t), \theta\rangle$ and $\eta' = \langle \pi^\star(\hat{\theta}_t), \theta'\rangle$, we have the exact identity

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}(\mathbb{P}_{R_t|\hat{\Theta}_t=\hat{\theta}_t,\Theta=\theta} \,\|\, \mathbb{P}_{R_t|\hat{\Theta}_t=\hat{\theta}_t,\Theta'=\theta'}) &= \mathbb{E}_{R_t\sim\mathbb{P}_{R_t|\hat{\Theta}_t=\hat{\theta}_t,\Theta=\theta}}\left[R_t(\eta-\eta') - b(\eta) + b(\eta')\right] \\ &= b(\eta') - b(\eta) - b'(\eta)(\eta'-\eta),\end{aligned}$$

which we recognize as the Bregman divergence of b evaluated at (η', η). Note that in the last equality we used that $\mathbb{E}[R_t|\eta] = b'(\eta)$. Since b is twice differentiable with continuous first and second derivative, Taylor's theorem gives, that for all $x, y \in \mathbb{R}$ with $x \leqslant y$, there exists an $\xi \in [x, y]$ such that

$$b(x) = b(y) + b'(y)(x-y) + \frac{1}{2}b''(\xi)(x-y)^2.$$

Then we have that

$$\mathrm{D}_{\mathrm{KL}}(\mathbb{P}_{R_t|\hat{\Theta}_t=\hat{\theta}_t,\Theta=\theta} \,\|\, \mathbb{P}_{R_t|\hat{\Theta}_t=\hat{\theta}_t,\Theta'=\theta'}) \leqslant \frac{1}{2}\sup_{|x|\leqslant S} b''(x)(\eta'-\eta)^2 \leqslant 2C\varepsilon^2,$$

where we used that $|\eta' - \eta| \leqslant \|\pi^\star(\hat{\Theta}_t)\|\|\Theta' - \Theta\| \leqslant \|\Theta' - \Theta_\varepsilon\| + \|\Theta_\varepsilon - \Theta\| \leqslant 2\varepsilon$ and the fact that $C = \sup_{|x|\leqslant S} b''(x)$ as $\mu'(x) = b''(x)$. Taking expectations, the second term in Eq. (3.2) is therefore at most $2C\varepsilon^2$.

Summing over t and applying the chain rule for conditional mutual information, we obtain

$$\sum_{t=1}^{T} \mathrm{I}(\Theta_\varepsilon; R_t \mid H^t, \hat{\Theta}_t) = \mathrm{I}(\Theta_\varepsilon; H^T) \leqslant \mathrm{H}(\Theta_\varepsilon).$$

Collecting the $2C\varepsilon^2$ and inserting the above bound into Eq. (3.1) yields

$$\mathbb{E}[\mathrm{Regret}(T)] \leqslant \sqrt{\Gamma T \left(\mathrm{H}(\Theta_\varepsilon) + 2C\varepsilon^2 T\right)},$$

which is the desired result. Then using that the cardinality of Θ_ε can be contolled by $d \log(1 + 2S/\varepsilon)$ and setting $\varepsilon = \sqrt{\frac{d}{2CT}}$ and rearranging terms inside the logarithm gives the desired result. □

3.3 An Incursion into Reinforcement Learning Problems

The final part of this thesis takes a step beyond bandit models and investigates more general reinforcement learning problems where the state of environment can be influenced by the agent's decision. Paper E marks the beginning of this broader exploration by studying theoretical performance limits and learning guarantees in model-based Bayesian reinforcement learning and while Paper F provides improved PAC-Bayes guarantees for offline decision-making settings.

3.3.1 The Limits of Bayesian Reinforcement Learning (Paper E)

Paper E is inspired by the work of Xu and Raginsky [27], who introduced a framework to study the best achievable performance of supervised learning algorithms under a Bayesian setting. Following their approach, Paper E puts aside computational constraints and instead focuses on the fundamental Bayesian performance limits of *model-based reinforcement learning*. More precisely, we introduce a theoretical framework for analyzing Markov decision processes under Bayesian uncertainty, where the transition dynamics and reward functions are governed by a latent parameter Θ drawn from a known prior. Within this setting, we define the *Bayesian cumulative reward* (BCR) as the maximum expected return achievable when the agent learns purely from its interaction history. We then introduce the concept of the *minimum Bayesian regret* (MBR), defined as the performance gap between the BCR and the optimal cumulative reward achievable if the agent had access to the environment parameter.

A contribution of Paper E is to establish a *data processing inequality* for the BCR, showing that any transformation of the agent's knowledge can only decrease the achievable cumulative reward. Using this result, we derive upper bounds on the MBR based on information-theoretic divergences. Specifically, we provide bounds involving relative entropy and Wasserstein distance, and analyze these bounds in the

context of multi-armed bandits and online optimization with partial feedback. In the latter case, we show that our bounds recover, *from below*, the information-theoretic regret bounds of Russo and Van Roy [31]. These results are later used in [70, 71] to derive *minimax* regret bounds.

3.3.2 Refining PAC-Bayesian Guarantees (Paper F)

In Paper F, we turn to the offline reinforcement learning setting, where the agent must learn a good decision policy solely from a fixed dataset of logged interactions in bandit problems, or loggged episode's results in *episodic RL*. Based on this dataset, the agent must select a fixed policy that yields high rewards in expectation. Since the reward distribution is unknown, a common approach to evaluate policies using importance-weighted empirical reward estimates. An important question to guide the policy search is whether the empirical reward estimate is close to the true expected reward. *Probably Approximately Correct* (PAC) theory provides answers to this question by offering "in probability" guarantees on the difference between the empirical estimate and the true expectation. PAC-Bayesian bounds generalize PAC guarantees and account for the dependence of the learned policy on the observed dataset [72–76]. This dependence is typically measured using the relative entropy between a prior hypothesis and the learned hypothesis. The resulting bounds have found practical applications, as they motivated the introduction of relative entropy regularization methods. The idea is to search for a policy π that maximizes the expected reward estimate while imposing a penalty on the relative entropy between π and a prior policy μ. Examples of such methods include Relative Entropy Policy Search [77], Trust Region Policy Optimization [60], and Proximal Policy Optimization [61].

One problematic of the PAC-Bayes bounds derived in [57] and follow-up works [59, 78, 79] is that they depend on a parameter $\lambda > 0$ that must be set before observing the data and cannot be optimized based on the data. [59] attempts to circumvent this difficulty deriving a bound that holds simultaneously for a grid of parameters, but their approach fell short of obtaining a parameter-free bound with optimal rate.

In Paper F, we improve on these previous results and derive parameter-free PAC-Bayes bounds, achieving the optimal rate. We apply an optimization technique for the "in probability" parameter λ that works by discretizing the *space of possible events* for the bounds and optimizing the parameter λ *conditioned on the event* before applying a union bound [28]. This method allows us to derive Bernstein- and Hoeffding-type PAC-Bayesian bounds that achieve *optimal rates*.

CHAPTER 4

Conclusion & Future Work

This chapter concludes the thesis by summarizing its main contributions and outlining several promising directions for future research. The first part of this chapter, Section 4.1, synthesizes the main contributions made across linear bandits, structured bandits, and reinforcement learning, highlighting how the developed tools extend existing theory and open the way to new analyses. The second part, Section 4.2, discusses several research avenues where information-theoretic ideas may continue to provide insight.

4.1 Conclusion

In the first part of the thesis, we focused on completing the information-theoretic analysis of linear bandits. Earlier results had already shown that one can control regret by approximating the optimal action, but these guarantees were essentially restricted to finite action and parameter sets. We extended this line of work to infinite and continuous domains by adapting the regret approximation techniques based on the one-step compressed Thompson Sampling algorithm and its associated information ratio. This led to regret bounds that remain near-optimal while holding for general metric action spaces. We then refined the analysis further using chaining arguments, which decompose learning into a sequence of approximations at different resolutions and yield bounds expressed in terms of the metric entropy of the action space. Together, these results complete the information-theoretic study of linear bandits beyond the finite-action setting.

The second part of the thesis extends the information-theoretic analysis to more complex bandit models, where rewards may depend on side information or follow nonlinear structures. We first considered contextual bandits, extending the analysis based on the lifted information ratio, which measures regret and information gain conditionally on the observed context. This approach generalizes earlier results that were limited to binary rewards and establishes regret bounds for problems with bounded or sub-Gaussian rewards. We then addressed the logistic bandit problem, a prominent example of a nonlinear model, and established regret bounds that depend only logarithmically on the slope parameter and do not scale with the number of actions. These bounds overcome long-standing limitations of earlier analyses, which either grew exponentially with the logistic slope or scaled with the cardinality of the action or parameter space. Finally, we took initial steps toward generalized linear bandits, highlighting how information-based techniques can be adapted to richer reward models and offering a path toward unified regret guarantees for a large class of structured bandit problems.

The final part of the thesis extends the information-theoretic viewpoint beyond bandit models to more general reinforcement learning settings, where the agent's actions influence the evolution of the environment. In the model-based Bayesian framework, we introduced formal definitions of the Bayesian cumulative reward and the minimum Bayesian regret, capturing the best achievable performance when learning is limited only by uncertainty about the environment. We established data processing inequalities showing that any transformation of the agent's information can only decrease achievable performance, and derived upper bounds on the inherent difficulty of reinforcement learning problems using relative entropy and other divergence measures. In the offline setting, where policies must be learned from fixed datasets, we developed parameter-free PAC-Bayesian bounds that achieve optimal rates for evaluating and selecting policies. These results connect directly to entropy-regularized policy search methods, providing a theoretical foundation for algorithms such as trust-region and proximal policy optimization.

Beyond these specific results, the thesis highlights the elegance and versatility of information-theoretic analysis as a tool for understanding and designing learning algorithms. Information-theoretic regret bounds often take simple, interpretable forms that reveal exactly which aspects of a problem make it hard, and how structure can be exploited to make it easier. This clarity is not merely theoretical: it also guides the construction of new algorithms. Methods such as Information Directed Sampling and its optimistic variants are built by directly optimizing the trade-off expressed in information-ratio bounds, selecting actions that promise the greatest reduction in uncertainty per unit of regret. In this way, the theory provides both insight and practical direction, showing that principled reasoning about information acquisition can lead naturally to algorithms that are effective, robust, and aligned with the fundamental limits of learning.

4.2 Future Work

A promising direction for future research is the development of frequentist regret guarantees using information-theoretic tools. While originally formulated for Bayesian analysis, recent work has shown that these concepts can be adapted to obtain worst-case performance bounds. For instance, Foster et al. [80] introduced the Decision-Estimation Coefficient (DEC), a general framework using Hellinger distance to characterize sample-efficient learning in stochastic settings. In a parallel development, Zhang [81] proposed Feel-Good Thompson Sampling (FGTS), which achieves frequentist guarantees by replacing the standard posterior with an optimistic update. Building on these ideas, Neu et al. [69] recently introduced a surrogate information ratio, providing a direct bridge between information-ratio analysis and frequentist bounds. As the methods in this thesis align well with this information-theoretic perspective, extending our analysis to incorporate these Hellinger-based tools could potentially lead to improved frequentist guarantees for structured bandit models.

A second direction for future work lies in strengthening the connections between bandit theory and generalization theory in supervised learning [82, 83]. Information-theoretic ideas originally appeared in both areas around the same time [84]: in bandits, to study exploration and regret, and in supervised learning, to understand how well an algorithm generalizes by measuring how much information it extracts or "memorizes" from the training data. Although some principles are shared across these fields, the two research communities have developed their tools in parallel and there may be unexplored opportunities to transfer techniques or perspectives from one domain to the other. Exploring such links could lead to new insights into both regret minimization and generalization.

A final and perhaps more ambitious direction is to extend information-directed exploration beyond bandits to full reinforcement learning. In this thesis, we established rigorous performance guarantees for algorithms that minimize the expected information ratio in bandit and contextual bandit problems, and algorithms inspired by this principle, such as Information Directed Sampling and its optimistic variants, demonstrate strong empirical performance in these settings. Adapting these concepts to reinforcement learning is non-trivial, as exploration influences not only immediate rewards but also the future states in which information is gathered. Nevertheless, even incremental progress toward information-aware reinforcement learning could lead to the development of more effective and principled exploration strategies.

References

[1] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic programming*. Athena Scientific, 1996.

[2] R. S. Sutton and A. G. Barto, *Reinforcement learning: an introduction* (Adaptive computation and machine learning), en. Cambridge, Mass: MIT Press, 1998.

[3] M. Ghavamzadeh, S. Mannor, J. Pineau, A. Tamar, et al., "Bayesian reinforcement learning: A survey", *Foundations and Trends® in Machine Learning*, vol. 8, no. 5-6, pp. 359–483, 2015.

[4] R. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton University Press, 1957.

[5] R. A. Howard, *Dynamic Programming and Markov Processes*. Cambridge, MA: Technology Press of the Massachusetts Institute of Technology, 1960.

[6] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY: John Wiley & Sons, 1994.

[7] A. Ghosh, A. Honoré, and S. Chatterjee, "DANSE: Data-driven non-linear state estimation of model-free process in unsupervised learning setup", *IEEE transactions on signal processing*, vol. 72, pp. 1824–1838, 2024.

[8] R. S. Sutton, "Learning to Predict by the Methods of Temporal Differences", *Machine Learning*, vol. 3, no. 1, pp. 9–44, 1988.

[9] C. J. C. H. Watkins and P. Dayan, "Q-learning", *Machine Learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[10] V. Mnih et al., "Human-level Control through Deep Reinforcement Learning", *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[11] D. Silver et al., "Mastering the Game of Go with Deep Neural Networks and Tree Search", *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[12] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: past, present and future", *Stroke and vascular neurology*, vol. 2, no. 4, 2017.

[13] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies", *IEEE access*, vol. 8, pp. 58 443–58 469, 2020.

[14] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixao, F. Mutz, et al., "Self-driving cars: A survey", *Expert systems with applications*, vol. 165, p. 113 816, 2021.

[15] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation", in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 661–670.

[16] O. Chapelle and L. Li, "An Empirical Evaluation of Thompson Sampling", in *Advances in Neural Information Processing Systems*, vol. 24, Curran Associates, Inc., 2011, pp. 2249–2257.

[17] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy, "Deep exploration via bootstrapped DQN", *Advances in neural information processing systems*, vol. 29, 2016.

[18] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, "A comprehensive overview of large language models", *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 5, pp. 1–72, 2025.

[19] J. Dai, X. Pan, R. Sun, J. Ji, X. Xu, M. Liu, Y. Wang, and Y. Yang, "Safe rlhf: Safe reinforcement learning from human feedback", *arXiv preprint arXiv:2310.12773*, 2023.

[20] H. J. Jeon and B. Van Roy, "Aligning AI Agents via Information-Directed Sampling", *arXiv preprint arXiv:2410.14807*, 2024.

[21] C. Zhang, Y. Deng, X. Lin, B. Wang, D. Ng, H. Ye, X. Li, Y. Xiao, Z. Mo, Q. Zhang, et al., "100 days after deepseek-r1: A survey on replication studies and more directions for reasoning language models", *arXiv preprint arXiv:2505.00551*, 2025.

[22] V. Nasteski, "An overview of the supervised machine learning methods", *Horizons. b*, vol. 4, no. 51-62, p. 56, 2017.

[23] H. Robbins, "Some aspects of the sequential design of experiments", 1952.

[24] C. E. Shannon, "A mathematical theory of communication", *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[25] T. Steinke and L. Zakynthinou, "Reasoning about generalization via conditional mutual information", in *Conference on Learning Theory*, PMLR, 2020, pp. 3437–3452.

[26] B. Rodríguez Gálvez, G. Bassi, R. Thobaben, and M. Skoglund, "Tighter expected generalization error bounds via Wasserstein distance", *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[27] A. Xu and M. Raginsky, "Minimum excess risk in Bayesian learning", *IEEE Transactions on Information Theory*, 2022.

[28] B. Rodriguez-Galvez, R. Thobaben, and M. Skoglund, "More PAC-Bayes bounds: From bounded losses, to losses with general tail behaviors, to anytime validity", *Journal of Machine Learning Research*, vol. 25, no. 110, pp. 1–43, 2024.

[29] B. Rodríguez-Gálvez, R. Thobaben, and M. Skoglund, "An information-theoretic approach to generalization theory", *arXiv preprint arXiv:2408.13275*, 2024.

[30] D. Russo and B. Van Roy, "Learning to Optimize via Posterior Sampling", en, *Mathematics of Operations Research*, vol. 39, no. 4, pp. 1221–1243, Nov. 2014.

[31] D. Russo and B. Van Roy, "An information-theoretic analysis of Thompson sampling", *The Journal of Machine Learning Research*, vol. 17, no. 68, pp. 2442–2471, 2016, Peer-reviewed journal version; supersedes arXiv:1403.5341.

[32] D. J. Russo, B. V. Roy, A. Kazerouni, I. Osband, and Z. Wen, "A Tutorial on Thompson Sampling", English, *Foundations and Trends® in Machine Learning*, vol. 11, no. 1, pp. 1–96, Jul. 2018, Publisher: Now Publishers, Inc.

[33] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples", *Biometrika*, vol. 25, no. 3-4, pp. 285–294, 1933.

[34] E. Kaufmann, N. Korda, and R. Munos, "Thompson sampling: An asymptotically optimal finite-time analysis", in *International conference on algorithmic learning theory*, Springer, 2012, pp. 199–213.

[35] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey", *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.

[36] P. Kormushev, S. Calinon, and D. G. Caldwell, "Reinforcement learning in robotics: Applications and real-world challenges", *Robotics*, vol. 2, no. 3, pp. 122–148, 2013.

[37] L. Lü, M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, and T. Zhou, "Recommender systems", *Physics reports*, vol. 519, no. 1, pp. 1–49, 2012.

[38] F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh, "Recommendation systems: Principles, methods and evaluation", *Egyptian informatics journal*, vol. 16, no. 3, pp. 261–273, 2015.

[39] N. Wahlström, T. B. Schön, and M. P. Deisenroth, "From pixels to torques: Policy learning with deep dynamical models", *arXiv preprint arXiv:1502.02251*, 2015.

[40] D. Bouneffouf, I. Rish, and C. Aggarwal, "Survey on applications of multi-armed and contextual bandits", in *2020 IEEE Congress on Evolutionary Computation (CEC)*, IEEE, 2020, pp. 1–8.

[41] N. Silva, H. Werneck, T. Silva, A. C. Pereira, and L. Rocha, "Multi-Armed Bandits in Recommendation Systems: A survey of the state-of-the-art and future directions", en, *Expert Systems with Applications*, vol. 197, p. 116 669, Jul. 2022.

[42] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation", *Advances in neural information processing systems*, vol. 12, 1999.

[43] M. Woodroofe, "A one-armed bandit problem with a concomitant variable", *Journal of the American Statistical Association*, vol. 74, no. 368, pp. 799–806, 1979.

[44] A. G. Barto and P. Anandan, "Pattern-recognizing stochastic learning automata", *IEEE Transactions on Systems, Man, and Cybernetics*, no. 3, pp. 360–375, 1985.

[45] V. Gullapalli, *Associative reinforcement learning of real-valued functions*. Citeseer, 1990.

[46] J. Sarkar, "One-armed bandit problems with covariates", *The Annals of Statistics*, pp. 1978–2002, 1991.

[47] L. P. Kaelbling, "Associative reinforcement learning: A generate and test algorithm", *Machine Learning*, vol. 15, pp. 299–319, 1994.

[48] A. L. Strehl, C. Mesterharm, M. L. Littman, and H. Hirsh, "Experience-efficient learning in associative bandit problems", in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 889–896.

[49] E. Kaufmann, O. Cappé, and A. Garivier, "On the complexity of best-arm identification in multi-armed bandit models", *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1–42, 2016.

[50] Y. W. Yury Polyanskiy, *Information Theory - From Coding to Learning*, English, First. Cambridge University Press, Oct. 2022.

[51] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley-Interscience, 2006.

[52] A. Gouverneur, B. Rodríguez-Gálvez, T. J. Oechtering, and M. Skoglund, "Thompson Sampling Regret Bounds for Contextual Bandits with sub-Gaussian rewards", in *2023 IEEE International Symposium on Information Theory (ISIT)*, Taipei, Taiwan, IEEE, 2023, pp. 1306–1311.

[53] S. L. Scott, "A modern Bayesian look at the multi-armed bandit", *Applied Stochastic Models in Business and Industry*, vol. 26, no. 6, pp. 639–658, 2010.

[54] R. M. Gray, *Entropy and information theory*. Springer Science & Business Media, 2011.

[55] S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári, "Parametric Bandits: The Generalized Linear Case", in *Advances in Neural Information Processing Systems*, vol. 23, Curran Associates, Inc., 2010.

[56] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[57] Y. Seldin and N. Tishby, "PAC-Bayesian Analysis of Co-clustering and Beyond.", *Journal of Machine Learning Research*, vol. 11, no. 12, 2010.

[58] Y. Seldin, P. Auer, J. S. Shawe-taylor, R. Ortner, and F. Laviolette, "PAC-Bayesian Analysis of Contextual Bandits", en, 2011.

[59] Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer, "PAC-Bayesian Inequalities for Martingales", en, *IEEE Transactions on Information Theory*, vol. 58, no. 12, pp. 7086–7093, Jul. 2012, arXiv:1110.6886 [cs].

[60] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, *Trust Region Policy Optimization*, en, arXiv:1502.05477 [cs], Apr. 2017.

[61] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, *Proximal Policy Optimization Algorithms*, en, arXiv:1707.06347 [cs], Aug. 2017.

[62] S. Dong and B. Van Roy, "An Information-Theoretic Analysis for Thompson Sampling with Many Actions", en, in *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, NeurIPS 2018; supersedes arXiv:1805.11845, vol. 31, arXiv, May 2018, pp. 4161–4169.

[63] R. van Handel, *Probability in High Dimension*, en. Princeton University, Dec. 2016, vol. APC 550 Lecture Notes, APC 550 Lecture Notes, Princeton University. This version: December 21, 2016. Accessed 2025-09-10.

[64] A. R. Asadi, E. Abbe, and S. Verdú, "Chaining Mutual Information and Tightening Generalization Bounds", en, *arXiv:1806.03803 [cs, math, stat]*, Jul. 2019, arXiv: 1806.03803.

[65] G. Neu, I. Olkhovskaia, M. Papini, and L. Schwartz, "Lifting the Information Ratio: An Information-Theoretic Analysis of Thompson Sampling for Contextual Bandits", en, in *Advances in Neural Information Processing Systems*, vol. 35, Dec. 2022, pp. 9486–9498.

[66] D. Russo and B. Van Roy, "Learning to Optimize via Information-Directed Sampling", en, *Operations Research*, vol. 66, no. 1, pp. 230–252, Feb. 2018.

[67] L. Faury, M. Abeille, C. Calauzènes, and O. Fercoq, "Improved Optimistic Algorithms for Logistic Bandits", in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, pp. 3052–3060.

[68] H. B. McMahan and M. Streeter, "Open Problem: Better Bounds for Online Logistic Regression", en, in *Proceedings of the 25th Annual Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 23, PMLR, 2012, pp. 44.1–44.3.

[69] G. Neu, M. Papini, and L. Schwartz, "Optimistic Information Directed Sampling", in *Proceedings of Thirty Seventh Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, COLT 2024; replaces arXiv:2402.15411, vol. 247, PMLR, 2024, pp. 3970–4006.

[70] R. Bongole, A. Gouverneur, T. J. Oechtering, and M. Skoglund, "Information-Theoretic Minimax Regret Upper Bounds for Reinforcement Learning Problems", in *2025 IEEE Information Theory Workshop (ITW)*, IEEE, 2025, pp. 1–5.

[71] R. Bongole, A. Gouverneur, B. Rodríguez-Gálvez, T. J. Oechtering, and M. Skoglund, "Information-Theoretic Minimax Regret Bounds for Reinforcement Learning based on Duality", in *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025, pp. 1–5.

[72] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony, "A framework for structural risk minimisation", in *Conference on Computational learning theory (COLT)*, 1996, pp. 68–76.

[73] D. A. McAllester, "Some PAC-Bayesian theorems", in *Conference on Computational learning theory (COLT)*, 1998, pp. 230–234.

[74] D. A. McAllester, "PAC-Bayesian model averaging", in *Conference on Computational learning theory (COLT)*, 1999, pp. 164–170.

[75] D. A. McAllester, "PAC-Bayesian stochastic model selection", *Machine Learning*, vol. 51, no. 1, pp. 5–21, 2003.

[76] O. Catoni, "PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning", *IMS Lecture Notes Monograph Series*, vol. 56, 163pp, 2007.

[77] J. Peters, K. Mulling, and Y. Altun, "Relative Entropy Policy Search", en, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 24, no. 1, pp. 1607–1612, Jul. 2010.

[78] Y. Seldin, N. Cesa-Bianchi, P. Auer, F. Laviolette, and J. Shawe-Taylor, "PAC-Bayes-Bernstein Inequality for Martingales and its Application to Multiarmed Bandits", en, 2011.

[79] Y. Seldin, F. Laviolette, J. Shawe-Taylor, J. Peters, and P. Auer, *PAC-Bayesian Analysis of Martingales and Multiarmed Bandits*, en, arXiv:1105.2416 [cs, stat], May 2011.

[80] D. J. Foster, S. M. Kakade, J. Qian, and A. Rakhlin, *The Statistical Complexity of Interactive Decision Making*, en, arXiv:2112.13487 [cs, math, stat], Jul. 2023.

[81] T. Zhang, "Feel-Good Thompson Sampling for Contextual Bandits and Reinforcement Learning", *SIAM Journal on Mathematics of Data Science*, vol. 4, no. 2, pp. 834–857, 2022, Originally posted as arXiv:2110.00871.

[82] F. Hellström and G. Durisi, "Generalization bounds via information density and conditional information density", *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 3, pp. 824–839, 2020.

[83] F. Hellström, G. Durisi, B. Guedj, M. Raginsky, et al., "Generalization bounds: Perspectives from information theory and PAC-Bayes", *Foundations and Trends® in Machine Learning*, vol. 18, no. 1, pp. 1–223, 2025.

[84] D. Russo and J. Zou, "How much does your data exploration overfit? Controlling bias via information usage", en, *IEEE Transactions on Information Theory*, vol. 66, no. 1, pp. 302–323, Oct. 2019, arXiv: 1511.05219.

Part II

Publications