The Exploration-Exploitation Tradeoff in Sequential Decision Making Problems

Adam Sykulski

May 20, 2009

Abstract

Sequential decision making problems often require an agent to act in an environment where data is noisy or not fully observed. The agent will have to learn how different actions relate to different rewards, and must therefore balance the need to explore and exploit in an effective strategy. In this report, sequential decision making problems are considered through extensions of the multi-armed bandit framework. Firstly, the bandit problem is extended to a Multi-Agent System (MAS), where agents control individual arms but can communicate potentially useful information with each other. This framework allows for a better understanding of the exploration-exploitation tradeoff in scenarios where there are multiple agents interacting in a noisy environment. To this end, we present a novel strategy for action and communication decisions and we demonstrate the benefits of such a strategy empirically. This motivates a theoretical analysis of one-armed bandit problems, to develop ideas of how different strategies are optimally tuned. Specifically, the expected rewards of ϵ-greedy strategies are derived, as well as proofs governing their optimal tuning.

Contents

1 Introduction **3**

2 Background **6**
2.1 The Multi-Armed Bandit Problem 6
2.1.1 Multi-Armed Bandit Frameworks 7
2.1.2 Strategies for the One-Armed Bandit Problem with Covariates 9
2.2 Sequential Decision Making in MAS 11
2.3 Summary 13

3 A Multi-Agent Bandit Problem **15**
3.1 The Multi-Agent Bandit Framework 15
3.2 A Novel Strategy for Action and Communication Decisions . . . 17
3.2.1 The Value Of Communication (VOC) 18
3.2.2 The Double ϵ-greedy and Double ϵ-first Strategies 20
3.2.3 Dealing with Missing Data 21
3.3 Performance of the Strategy 22
3.3.1 Application of the VOC 22
3.3.2 Performance of the Double ϵ-first Strategy 23
3.4 Summary 25

4 One-Armed Bandits with Covariates **27**
4.1 The One-Armed Bandit with Covariates Framework 28
4.2 The Probability of Error 29
4.3 The ϵ-greedy Strategy 30
4.4 The ϵ-first Strategy 33
4.5 Summary 36

5 Conclusions and Future Work **38**
5.1 Conclusions 38
5.2 Future Work 39

References **41**

Appendix **47**
Appendix I 47
Appendix II 48
Appendix III 48
Appendix IV 51

Chapter 1

Introduction

Sequential decision making is the act of answering the question "what should I do next?" when faced with a series of tasks. A classic example is chess, where the player has to choose one of a finite set of moves. This move will then impact on the opponent's next move, which in turn will impact on the player's next move, and so on until the game ends. Another example is daily selection of a route to work, where past experiences and the current environment (weather, time of day) will influence the decision each day.

These problems often require the decision maker (henceforth called an agent) to learn as it goes along. In other words, the agent needs to interpret the information gained from past experience in such a way that benefits the present decision. For example in the game of chess, such learning involves continuously analysing the opponent's strategy to find an effective counter-strategy. In the route selection example, the learning involves interpreting the performance of previous routes to anticipate which route is best today. Moreover, such problems inherently suffer from the exploration-exploitation trade-off, where the agent must choose between what it believes is the best decision (exploitation) and trying alternative decisions for potential future benefit (exploration).

This tradeoff has been extensively studied, for example, in the multi-armed bandit problem (see [Sutton and Barto, 1998] for an overview), where the agent must repeatedly select one of several arms to pull. Each such arm delivers a stochastic reward and the agent receives a reward only from the arm that is pulled. The objective then is to find a strategy for selecting arms that maximises total reward over a length of play. In most cases, the agent has no prior knowledge of what reward to expect from each arm and *must learn as it plays.* A good strategy simultaneously identifies and plays the best arm as often as possible. Many real-world decision problems, in particular those with unknown outcomes for different decisions, can be modelled using the bandit setting (see Chapter 2 for details). For this reason, the multi-armed bandit problem is the framework used to study sequential decision making problems in this work.

In many real-world settings, however, such decision making has to take place in an environment in which there are other agents operating – these are called Multi-Agent Systems (MAS). For example, consider a number of fire brigades, ambulances and police vehicles dealing with a disaster management scenario (see [Ramchurn et al., 2008]). Each emergency service vehicle must make simultaneous decisions in a potentially unknown environment, for example sequentially

deciding which buildings to evacuate. Furthermore, there is an interdependence between the actions of different agents and therefore there is a need for the agents to coordinate their actions. Such systems are often decentralised, meaning the agents act without a central coordinating agent and hence any single agent can be removed without necessarily affecting the integrity of the system [Ferber, 1999]. The impact of sequential decisions is often unknown by each agent, therefore the need to balance exploration and exploitation remains.

In realistic settings, however, agents are typically able to communicate with each other, such that agents can exchange useful information that benefits future decision making [Fatima et al., 2004]. Moreover, communication can reduce the need for an agent to explore different actions, as the gained information can reduce the agent's uncertainty. Nevertheless, communication can often be costly or time-consuming [Krause et al., 2006], hence each agent must try to assess the value of communicating with other agents. In the disaster management scenario for example, the fire brigades can communicate with each other to coordinate surveying different areas in order to improve efficiency and quickly identify fires. If the location of the fires is clear to all agents, however, then there may be little value in communicating such information as the best decision of each agent is clear.

Many existing sequential decision making problems in MAS, studied in stochastic games for example (see [Condon, 1992] for a review), have rewards that are known to each agent or easily learnt over play. Conversely, multi-armed bandit problems study scenarios in a single-agent setting where the expected reward of each arm is unknown *a priori* and the reward of a pull is then observed with noise, such that learning the true expected reward of an arm involves repeated plays. Given this background, in this work we extend the multi-armed bandit problem to a simple MAS, to investigate the ideas of exploration-exploitation in a multi-agent environment where rewards are uncertain and observed with noise. In particular, the multi-armed bandit with covariates framework is used, where the reward of each arm is a function of sequentially observed side information represented as a covariate (see Section 2.1.1 for details). Such side information might include for example sensor readings from various smoke detectors in the aforementioned disaster management scenario. In our framework, the covariate is only partially observed by each agent, but missing observations can be communicated (at a cost) between the agents. Each agent's action decision is then to decide which arms to pull, from the subset of arms that it controls.

For this novel framework, strategies that select communication and action decisions are constructed and studied, where the need to balance exploration and exploitation is specifically addressed. Furthermore the concept of an agent exploring communication decisions is introduced, where an agent may benefit from communicating with another agent even if this appears to have no immediate value, as this helps the agent's learning. Building on this idea, it is shown through empirical evaluation that agents can indeed benefit from exploring by communicating and not just from exploring by acting, and the amount of exploration required from the two exploration methods are interdependent.

This novel extension of bandit problems to MAS decentralises the control of the arms between the set of agents, such that no single agent acts on all the arms (unlike most studies of bandit problems). Moreover, any agent can pull as many arms as it wishes at any given time from the subset of arms it

controls, which is analogous to an agent simultaneously playing a series of one-armed bandits with covariates (introduced in more detail in Section 2.1.1). In this problem, the agent has a choice of pulling between an arm with *a priori* unknown expected reward and an arm with known expected reward. In such situations, the agent must learn the relationship between the covariate and the arm with *a priori* unknown expected reward in order to identify the best arm to pull at each time-step. Exploration over exploitation refers to the agent pulling this arm even when the expected reward is less than the arm with known expected reward, as this improves the agent's learning. To this end, we introduce a novel approach of reasoning about the problem, and derive new theoretical results and proofs about the performance of different strategies with a 1-dimensional covariate. Specifically it is shown that the agent might benefit from some exploration, but this depends on the chosen strategy.

The theoretical results developed for the one-armed bandit problem will be extended, in future work, to reason about how the agents should explore in the novel multi-agent bandit framework. In particular, for this problem the benefit of exploration has thus far only been demonstrated empirically, but theoretical bounds and expectations can be developed to guarantee performance of certain strategies under expectation. Moreover, theoretical ideas can be used to develop on-line tuning of different strategies, such that these strategies can adapt to the environment and perform more (or less) exploration when this is required. This is particularly important for applications to realistic scenarios, where parameters of the model that affect the optimal tuning, are usually unknown *a priori*.

Taking this work together, the vision of this research is to find optimal strategies in a variety of sequential decision making problems, for both single and multi-agent systems, where effective strategies require balancing exploration and exploitation. A key focus is to study frameworks with practical applications and to provide empirical evidence of the performance of different strategies. We also aim to theoretically reason about the impact of the exploration-exploitation tradeoff to an agent's reward. This allows us to find the optimal tuning of different exploration strategies (off-line), with the potential to use these ideas to develop strategies that are tuned on-line and are thus adaptive to the environment and the actions of other agents.

The structure of this report is as follows. In Chapter 2, the multi-armed bandit and one-armed bandit problems are introduced, together with a review of existing strategies for arm selection that attempt to balance exploration with exploitation. Existing sequential decision making problems in MAS are also reviewed, including studies of stochastic games. In Chapter 3, the multi-agent bandit problem is introduced. A strategy for arm selection and communication decisions is formulated for this novel framework, together with an empirical performance evaluation. In Chapter 4, theoretical results for strategies for one-armed bandit problems are presented, as well as proofs governing their optimal tuning. Conclusions and planned future work can be found in Chapter 5.

Chapter 2

Background

The multi-armed bandit problem is the framework used to study sequential decision making problems in this work. This problem is of particular interest as it specifically addresses the need to balance exploration and exploitation in an unknown environment. To this end, we review the multi-armed bandit problem in Section 2.1, with a particular focus on the one-armed bandit problem with covariates studied in Chapters 3 and 4. Furthermore, the multi-agent bandit problem studied in Chapter 3 is an extension of the bandit framework in the direction of MAS. With this in mind, we review some existing sequential decision making problems in MAS in Section 2.2.

2.1 The Multi-Armed Bandit Problem

The multi-armed or k-armed bandit problem is a (discrete action space) sequential decision making framework commonly studied in the fields of statistics [Berry and Fristedt, 1985; Gittins, 1989], machine learning [Auer et al., 1995; Sutton and Barto, 1998] and economics [Rothschild, 1974; Azoulay-Schwartz et al., 2004], amongst others. Originally documented in [Robbins, 1952], the problem is based on the analogy of a slot machine or one-armed bandit. The agent must select one of several arms to pull where a reward is only received from the arm that is pulled. The game is played repeatedly and the objective is to find a selection strategy that maximises total cumulative reward. The agent is commonly assumed to have little or no prior knowledge about the reward structure of each arm and thus should explore rewards from different arms in an effective strategy. Ultimately, the best strategies are those that incorporate the need to balance exploration (pulling different arms to identify the best) and exploitation (pulling the expected best arm to maximise reward). This trade-off has been widely studied in reinforcement learning [Kaelbling et al., 1996], resource allocation problems [March, 1991; Benner and Tushman, 2003], product development [Rothaermel and Deeds, 2004], as well as by economists in analysing buyer/seller scenarios [Azoulay-Schwartz et al., 2004].

Multi-armed bandit problems have applications in areas as diverse as clinical drug trials [Woodroofe, 1979; Hardwick et al., 1998], online auctions [Blum et al., 2003], sensor management [Krishnamurthy and Evans, 2001; Hero et al., 2006], pricing goods [Weitzman, 1979; Azoulay-Schwartz et al., 2004], web advertising

[Pandey et al., 2007; Kleinberg et al., 2008] and many other decision-making problems, see [Sutton and Barto, 1998].

Against this background, Section 2.1.1 outlines typical frameworks for studying multi-armed bandit problems, including bandits with covariates and the one-armed bandit problem studied in this work. Section 2.1.2 then describes in more detail strategies that can be used for the one-armed bandit problem with covariates.

2.1.1 Multi-Armed Bandit Frameworks

In the multi-armed or k-armed bandit problem, the agent pulls arm i at time t and receives a reward $r(t) = r_i(t)$ from that arm only. The objective is to find a strategy that maximizes the sum of the collected rewards after time T, $R(T) = \sum_{t=1}^{T} r(t)$. This problem has been studied extensively both in finite time T [Auer et al., 2002; Vermorel and Mohri, 2005] and as $T \to \infty$ [Lai and Robbins, 1985; Auer et al., 1995]. Furthermore, many different frameworks have been developed that determine how the rewards of each arm $r_i(t)$ are generated [Berry, 1972; Ginebra and Clayton, 1995; Cesa-Bianchi and Fischer, 1998]. In this work we are concerned with maximising reward in finite time because this is more relevant and applicable to real-world scenarios, including those modelled by MAS. Moreover, strategies that are asymptotically optimal can perform poorly in finite time [Vermorel and Mohri, 2005].

The **stochastic multi-armed bandit** considers the problem where each arm i has reward $r_i(t)$ at time t generated from a probability distribution $\mathcal{R}_i$. The agent is typically assumed to have no prior knowledge of these distributions and the distributions are assumed to be fixed over time [Sutton and Barto, 1998]. Finite time strategies for arm selection have been widely developed in this problem, in particular Upper Confidence Bound (UCB) methods in [Auer et al., 1995] for rewards bounded in $[0, 1]$ and the Price Of Knowledge Expected Reward (POKER) strategy [Vermorel and Mohri, 2005] for normally distributed rewards. Both strategies construct an inflated reward estimate for each arm, which is the mean observed reward added to an additional term that is inversely related to the number of pulls. The arm with the highest inflated reward estimate is pulled. Inflating the reward estimate in this way encourages exploration of arms that have been infrequently pulled.

The stochastic multi-armed bandit problem was also considered in [Gittins, 1989] for a formulation of the problem where the reward distribution of an arm changed if that arm was pulled. It was shown that the optimal arm could be selected using the *Gittins indices* by considering the future reward distributions of each arm independently. Specifically, the Gittins indices are an index for each arm, which is the expected reward of staying on an arm for an optimal length of time. The *Gittins rule* is then to pull the arm with the highest index value. This method significantly reduces the complexity of the computation and the optimality of Gittins indices have since been proved in [Whittle, 1980; Weber, 1992; Ishikida and Varaiya, 1994; Tsitsiklis, 1994]. In this formulation, however, the reward distributions are assumed to be known to the agent *a priori* (see [Ishikida and Varaiya, 1994] for details) and hence the problem is one of optimization rather than balancing exploration and exploitation (see [Auer et al., 2003, p49]). In this work we consider the multi-armed bandit problem with unknown reward distributions *a priori*, therefore the Gittins indices are

not considered in the rest of this report.

The **non-stochastic** or **adversarial multi-armed bandit problem** was studied in [Auer et al., 1995, 2003] where the rewards of each arm were set *a priori* by an adversary. The reward process requires no statistical assumptions, however rewards generated by the adversary are bounded in $[0, 1]$. A novel strategy, Exp3, was shown to achieve bounded optimal performance asymptotically. Although it was shown in [Vermorel and Mohri, 2005] that Exp3 performed badly in an empirical performance evaluation in finite time, for a version of the stochastic bandit problem. Moreover, the objective of the adversarial framework is to identify the best arm to repeatedly play for all iterations, rather than finding the best arm at each individual iteration (as we are concerned with in this report). This is a restrictive assumption in realistic scenarios where the optimal arm to pull can change between iterations. For this reason, the adversarial framework is not considered in this report.

The **one-armed bandit problem** is a special case of the multi-armed bandit problem. The agent must choose between an arm with unknown expected reward and an arm with known expected reward, henceforth these arms shall be called A and B, respectively. The problem, in this form, was first studied in [Chernoff, 1967] for sequential clinical trials, where a treatment had to be chosen between a drug with known probability of success and a new drug with unknown probability of success. Subsequently, the one-armed bandit problem has been extensively studied, for example in [Kumar and Seidman, 1981; Glazebrook, 1983; Rosenberg et al., 2007]. This framework is extensively studied in Chapters 3 and 4, where the agent has additional covariate information.

The **bandit problem with covariates**, first introduced in [Woodroofe, 1979], considers the scenario where the agent observes *side information* prior to each pull. In this problem, the expected reward of each arm is a function of this side information represented in the form of a covariate. Parameters that relate the covariate to the arm with unknown expected reward have to be learnt by the agent. It was argued in [Woodroofe, 1979] that such side information is likely to be present in many applications and incorporating this into bandit problems is a more realistic representation of real-world problems. For example, covariate information such as age, sex, height and weight could influence the probability of success of a drug in sequential clinical trials (see also [Sarkar, 1991]) and readings from sensors could affect actions in a disaster management scenario. More recent studies of bandits with covariates have included [Clayton, 1989] for the one-armed bandit problem, [Wang et al., 2005] for the two-armed bandit problem and [Ginebra and Clayton, 1995; Auer, 2000; Yang and Zhu, 2002; Pavlidis et al., 2008a,b] for multi-armed bandit problems.

In more detail, we consider the reward structure used in [Ginebra and Clayton, 1995; Yang and Zhu, 2002; Pavlidis et al., 2008a,b], for the rewards of arms A and B in the one-armed bandit problem. Specifically, the reward of each arm is modelled as a linear function of the covariate $X_t = (x_{1,t}, \ldots, x_{p,t})^T$ with additive observation noise:

$$\begin{aligned} r_A(t) &= \sum_{j=0}^{p} \alpha_j x_{j,t} + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, \sigma_\eta^2), \\ r_B(t) &= \sum_{j=0}^{p} \beta_j x_{j,t} + \omega_t, \qquad \omega_t \sim \mathcal{N}(0, \sigma_\omega^2), \end{aligned} \tag{2.1}$$

where $x_{0,t} = 1$. Here, α_0 and β_0 correspond to the intercepts of the two linear equations. The coefficient vector β is known to the agent *a priori*, but α is unknown and estimated from observations. A linear model was also used in [Woodroofe, 1979] for the one-armed bandit problem and in [Sarkar, 1991] for a more general reward structure based on the exponential family model.

Other reward structures considered were Bernoulli rewards where the covariate is related to the Bernoulli parameter using a link function [Clayton, 1989; Langford and Zhang, 2007] for the one-armed and multi-armed bandit problem, respectively and also [Wang et al., 2005] considered any continuous reward distribution with one unknown parameter for each arm in a two-armed bandit problem. Both of these alternative reward structures, however, consider covariates that can be related to the reward function using just one parameter (unknown *a priori*). This is considered a restrictive assumption because in realistic applications, covariate information can be multi-dimensional, and with these models there is only one parameter with which to link this to the reward function. For this reason, we use the model of (2.1) where each covariate value $x_{j,t}$ impacts on the reward $r_A(t)$ (for example) with its own parameter α_j. This is hence a more flexible model to capture the effect of side information.

2.1.2 Strategies for the One-Armed Bandit Problem with Covariates

Most advances in the one-armed bandit with covariates have been concerned with finding strategies that maximise reward over infinite-length play. For example, [Woodroofe, 1979; Sarkar, 1991] proved, using a Bayesian formulation, that a myopic or **greedy strategy** is asymptotically optimal for a given class of models. Specifically, the greedy strategy always selects the arm the agent believes is best, given the covariate value and existing estimates from prior observations.

In contrast, in this work we are concerned with maximising reward in finite time. This is because the length of play is finite in many real applications. For example, the number of patients in a clinical trial, or the number of times a consumer repurchases a specific good (choosing sequentially between a set of suppliers is another common application of bandit problems [Azoulay-Schwartz et al., 2004]). We are therefore seldom motivated by strategies that are only optimal asymptotically.

The greedy strategy can be optimal asymptotically but in finite time the strategy often fails to identify the optimal arm, as insufficient exploration is performed. For this reason, the greedy strategy has been found to perform badly in a multi-armed bandit with covariates setting [Pavlidis et al., 2008b], particularly as the number of arms increases. Nevertheless, if the system is very simple to learn (the variance of the noise term is low, or one arm is clearly better than all others) then a greedy strategy can perform well or even optimally in finite time, as in [Macready and Wolpert, 1998] for a 2-armed bandit problem. Nevertheless, the greedy decision strategy is often used as a benchmark strategy, due to its simplicity and the fact that it does not require a tuning parameter.

Certain strategies, however, include choosing explorative pulls for potential future gain. In the one-armed bandit problem this involves pulling the arm with unknown expected reward (as the other arm requires no exploration). In particular, we consider the **ϵ-greedy** and **ϵ-first** strategies. In an ϵ-greedy

strategy (first described in [Watkins, 1989]), the agent is greedy with probability $1 - \epsilon$, but explores with probability ϵ. In an ϵ-first strategy [Even-Dar et al., 2002], the agent performs all exploration first. In a game of finite length T, the agent pulls the arm with unknown expected reward for the first $\lceil \epsilon T \rceil$ iterations, and then exploits by being greedy for the remaining $\lfloor (1-\epsilon)T \rfloor$ iterations. Setting $\epsilon = 0$ recovers greedy selection for both strategies.

An empirical performance evaluation of various strategies in [Auer et al., 2002; Vermorel and Mohri, 2005], for a finite-time stochastic multi-armed bandit problem, led to the conclusion that a well tuned ϵ-greedy strategy [Auer et al., 2002, p.247] or ϵ-first strategy [Vermorel and Mohri, 2005, p.446], almost always performed better than all other strategies. ϵ-greedy is also a popular exploration strategy for problems in reinforcement learning [Sutton and Barto, 1998; Shani et al., 2005; Neumann et al., 2007]. For these reasons, we focus on these strategies in the one-armed bandit problem with covariates studied in this work.

For both strategies, larger values of ϵ encourage additional pulls of arm A, helping the agent to learn the coefficient vector α in (2.1) more quickly. This however comes at the expense of exploiting the agent's current knowledge. The parameter ϵ therefore has to be tuned to balance the benefit of learning quickly and the cost of having to do explorative steps. Setting the ϵ parameter, however, is difficult for both strategies and a badly tuned ϵ-first or ϵ-greedy strategy can perform poorly in comparison with other arm selection strategies (as shown in [Auer et al., 2002; Pavlidis et al., 2008b] for example) and further motivates the detailed analysis of these strategies – in particular the development of strategies that can be tuned on-line.

Note that for the same value of ϵ, an ϵ-first strategy will outperform the ϵ-greedy strategy under expectation in finite time, as demonstrated in [Vermorel and Mohri, 2005]. This is due to the fact that the same number of explorative actions will have the same short-term cost to the agent, regardless of when they are taken; however, if the explorative actions are taken earlier (as is the case with ϵ-first), then the agent will perform better on the exploitive (or greedy) actions, as they are taken later and the agent has more past observations to learn from.

A drawback of the ϵ-first strategy, however, is that the parameter ϵ cannot be easily set in a game of unknown length. Note also that this strategy would not perform well in the dynamic problem considered in [Pavlidis et al., 2008a] where the coefficients of the reward functions are changing and exploration is required throughout the game, not just at the beginning. Nevertheless, in most studies of bandits with covariates these coefficients are assumed to be static. Hence for the finite time frameworks used in this report, we consider both the ϵ-greedy and ϵ-first strategies.

Other strategies for bandits with covariates include **interval estimation** (introduced in [Kaelbling, 1993]) where the agent, rather than selecting the highest expected reward amongst arms, would instead select the highest upper confidence bound of the arms (see [Pavlidis et al., 2008b]). This corresponds to attributing an inflated reward estimate to each arm (similar to the Exp3 strategy in [Auer et al., 1995] for the non-stochastic bandit problem), but this estimate would be more optimistic for arms that have not been pulled many times, as the confidence interval is much wider. Another strategy is **softmax selection**, first proposed in [Luce, 1959], which involves randomly selecting an arm with

a weighted probability based on the likelihood of it being the best arm. These strategies were however generally shown to empirically perform worse than a well tuned ϵ-greedy or ϵ-first strategy in [Vermorel and Mohri, 2005; Pavlidis et al., 2008b]. This further motivates the use of ϵ-greedy and ϵ-first strategies in the extensions of the bandit problem studied in this work.

2.2 Sequential Decision Making in MAS

A multi-agent system (MAS) is a system composed of interacting agents. Examples of scenarios that can be modelled as MAS include online trading [Rogers et al., 2007], disaster response [Ramchurn et al., 2008] and sensor networks [Rogers et al., 2005]. Now, multi-agent learning and decision making is central to the operation of many MAS [Littman, 1994; Schaerf et al., 1995; Shoham et al., 2004], particularly because agents often need to make decisions whilst learning the motivations and goals of other participating agents. In this work a new framework for sequential decision making in MAS is proposed in Chapter 3, which is based on the multi-armed bandit problem. This framework allows for a detailed analysis of the exploration-exploitation tradeoff in a multi-agent context, in an environment where distributional parameters are unknown to the agents and rewards are observed with noise.

In more detail, sequential decision making problems in MAS have been widely studied in stochastic games [Erev and Roth, 1998; Bowling and Veloso, 2001; Chalkiadakis, 2003; Hu and Wellman, 2003; Hansen et al., 2004] and Markov games [Littman, 1994; Kaelbling et al., 1996; Claus and Boutilier, 1998; Wang and Sandholm, 2003; Chapman et al., 2009]. A stochastic game [Shapley, 1953] is a dynamic and competitive game between one or more agents, where the game changes state over time. Specifically, the agent selects an action and receives a reward that is dependent on the state of the game and the actions of others. The state of the game subsequently changes according to some probability transition dependent on the previous state and the actions chosen by all the agents. Stochastic games that have state changes that possess the Markov property are called Markov games. The Markov property states that given the state of the game at time t is known, transition probabilities to the state at time $t+1$ are independent of all previous states and actions [Littman, 1994].

Now, a stochastic bandit problem can be seen as a single-state stochastic game, whereas a bandit problem with covariates can be viewed as a stochastic game where the state changes (i.e. the values of the covariate) are independent of action choices and the current state. Nevertheless, the fundamental difference between studies of bandit problems and stochastic games, are that in stochastic games the expected reward received by an agent when the game is in a particular state is known, for any joint set of actions between the agents. The learning process in these games thus corresponds to the agent learning a good strategy in response to the actions of other agents. In a bandit problem however, the expected reward of any action is unknown *a priori* and precisely what the agent must learn to identify the best arm – this feature is of particular interest, as in realistic applications agents will seldom know precisely which actions yield which rewards. This motivates the extension of the bandit problem to MAS constructed in this work, where exploration-exploitation of sequential action decisions can be investigated in a multi-agent environment.

Sequential decision making in MAS has also been considered in [Teacy et al., 2008] where a set of enquiring agents request services from a set of providing agents. The agents must learn the trustworthiness of the service provider and thus face the exploration-exploitation dilemma. A strategy for selecting a service provider is formed using the *Value of Perfect Information (VPI)* [Dearden et al., 1998] where each service provider has a value based on its predicted trustworthiness plus the benefit of learning its true trustworthiness. The service provider with the highest VPI is chosen. This is analogous to UCB and POKER methods used for bandit problems, where an inflated reward estimate is used for each possible action. The VPI, however, adopts a myopic approach (for tractable calculation) and assumes that the true trustworthiness of an agent is observed after selection. This approach has been shown to perform well in a variety of frameworks and applications including coalition formation [Chalkiadakis and Boutilier, 2008] (which investigates frameworks with dynamically forming partnerships or teams of cooperating agents).

In most MAS, on the other hand, agents can communicate with each other in order to aid their learning [Tan, 1997; Fatima et al., 2004], as this is a realistic feature of MAS applications (for example, a number of fire brigades and ambulances operating in a disaster management scenario). In fact the VPI, which is used to value interactions or coalitions between agents, can be interpreted as the value of communication between agents (as this is a form of interaction). With this in mind, we allow communication between the agents in our multi-agent bandit problem. Specifically, we use the bandit with covariates framework where each agent observes a covariate prior to choosing which arms to pull. In our problem, however, the agents might only partially observe this covariate and will have to make action decisions with incomplete covariate information. Moreover, we allow the agents to communicate unknown covariate values, although this comes at a cost to the agent's reward. There is hence a tradeoff to each agent between the value and cost of communication. An agent also has to learn the value of communicating with a particular agent in future sequential decisions, and thus faces an additional exploration-exploitation dilemma (together with the arm selection dilemma).

The method developed to value this communication is similar to the VPI, in that a myopic approach is used for tractability. Furthermore, we show that the agent will benefit from additional exploration of communication decisions in our framework. This is because the myopic assumption, that the true value of communication between two agents can be learnt from one play, is restrictive in bandit problems where this will take several repeated plays to learn.

Communication between agents in sequential decision making problems has already been considered in Bayesian games [Gerardi, 2004], where agents can communicate information before choosing actions. A Bayesian game [Harsanyi, 1967] is a game where information about the rewards of other agent's actions is unknown or incomplete. The extension to games where agents can communicate is based on the idea of *cheap talk* [Farrell and Rabin, 1996], where agents can freely communicate without directly affecting the rewards of the game to each agent. This form of communication is therefore strategic, as agents can attempt to mislead other agents with false information for potential self-benefit [Farrell, 1987]. Communication in games has also been considered in network formation games (see [Jackson et al., 2003] for a review). A network formation game is a game where the agents must decide whether or not to form links with each other

to form a network. In some studies of this problem (see [Aumann and Myerson, 1988] for example), agents have been allowed to communicate preferences to each other, at no cost, before these links are made.

Our framework however, has communication that is costly rather than free, as the exchange of information between agents is often costly in realistic scenarios (for example in sensor networks [Krause et al., 2006]). In some network formation games [Bala and Goyal, 2000], forming links has an associated cost, but only to the agent initiating the link – this cost is analogous to the cost of communicating with other agents in our framework. The difference is that these links cannot be broken in the network formation game and thus the cost is only incurred once, whereas in our framework communication costs are incurred repeatedly if the agent chooses to communicate prior to each decision (as in the sensor network case [Krause et al., 2006]).

In our approach, we value the benefit of communication against the cost at each iterative step, to determine sequential communication decisions. This approach has been considered in [Williamson et al., 2009], where *reward shaping* is used to value communications between agents. The concept behind reward shaping is that agents have different beliefs about the interactions of different actions and rewards. If these beliefs are highly divergent then agents may wish to engage in communication, however if these beliefs are similar then each agent can independently calculate the same expected reward for each joint action without incurring any unnecessary communication costs. Communication can hence be valued by estimating the *belief divergence* of other agents in the system. Our approach considers the case where covariate information is distributed amongst agents, such that beliefs about future rewards are naturally divergent. In addition, our system is simplified in that an agent need not consider the potential actions of other agents and coordination of joint actions is not required. Nonetheless, the concept of valuing the information that other agent's possess against the cost of acquiring this information is the same in the two frameworks. In particular, if the expected rewards calculated by each agent (and hence subsequent optimal actions) are highly divergent, then there is greater value of communication in both frameworks.

Multi-agent approaches to bandit problems have been previously considered in [Le Ny et al., 2006]; this study addressed bandits where the rewards of the arms evolved to new states over time, similar to stochastic games. Although this work, again as with stochastic games, assumes knowledge of how future rewards are related to current actions and is therefore an optimization problem (and uses ideas similar to the Gittins indices [Gittins, 1989]). This assumption, as previously discussed, is not made in the bandit problem considered in this work.

2.3 Summary

In this chapter we have reviewed the multi-armed bandit problem, a sequential decision making problem used to study the exploration-exploitation tradeoff. In the following chapters we use the bandit with covariates setting (reviewed in Section 2.1.1) to study exploration-exploitation in various problems. In particular, we use the ϵ-greedy and ϵ-first strategies for the novel frameworks studied, as these are popular and well-performing strategies for exploration (as reviewed

in Section 2.1.2).

In the next chapter, we extend the bandit framework to study the exploration-exploitation tradeoff in environments with multiple agents. To this end, we have reviewed existing studies of sequential decision making problems in MAS in Section 2.2. Most of these studies (for example those in stochastic games) assume each agent has prior knowledge of its reward function. This motivates the extension of bandits to MAS, as no such prior knowledge is assumed in the bandit setting. The novel extension however, is an initial step in the direction of MAS, as the rewards to each agent are unaffected directly by the actions of other agents (in contrast to stochastic games for example). The problem presented is therefore not a game in the formal game theoretic sense, but does nevertheless consider the interaction of agents through communication – a key feature of many scenarios modelled by MAS. To this end, we use similar ideas to the VPI techniques used for coordination and coalition formation problems. Moreover, the framework will be extended (as part of future work) to bring in ideas from stochastic game theory.

The multi-agent framework of the next chapter motivates a theoretical investigation of the ϵ-greedy and ϵ-first strategies, where in Chapter 4, we present a novel approach for reasoning about these strategies when used in a one-armed bandit problem with covariates. This study is also motivated by the fact that we wish to find strategies that maximise reward in finite time (as opposed to asymptotically), which is more relevant to realistic scenarios modelled by both bandit problems and MAS.

Chapter 3

A Multi-Agent Bandit Problem

To date, multi-armed bandit problems have been considered in a single-agent context, where one agent decides an appropriate action at each time-step (as reviewed in Section 2.1). In this chapter, however, the bandit problem with covariates is extended to a multi-agent decentralised version, to study the exploration-exploitation tradeoff in an environment with multiple agents. We use the bandit with covariates problem (see Section 2.1) except each agent might only observe a subset of the covariate, representing its partial view of the world. Furthermore, we allow agents to initiate communication between themselves (at a cost) exchanging potentially useful covariate values that were previously unknown to the agents. The communication of information between agents is an important feature of scenarios modelled by MAS, as motivated by the literature reviewed in Section 2.2.

The structure of this chapter is as follows. In Section 3.1 we introduce the multi-agent bandit framework. In Section 3.2 we construct an effective strategy for communication and arm selection decisions that addresses the exploration-exploitation tradeoff. In particular, we propose a novel method of valuing communication between agents, called VOC, which finds the best myopic communication decision and we also propose novel exploration strategies for this problem called "double ϵ-greedy" and "double ϵ-first", which consider exploration of communication decisions as well as action decisions.. Finally, we test the double ϵ-first strategy empirically in section 3.3, and discuss the significance of these findings. Summary remarks follows in Section 3.4.

3.1 The Multi-Agent Bandit Framework

Consider a k-armed bandit and let K denote the set of arms, where $|K| = k$. Now consider a set of agents N ($|N| = n$), where each agent a_i controls a disjoint subset C_i of K for $i = 1, \ldots, n$ (i.e. the assignment of arms to agents forms a partition of the set of arms):

$$C_i \cap C_j = \emptyset \ , \quad \forall i, j, i \neq j \ , \quad \text{where } C_i \subseteq K.$$

Each arm is controlled by one agent only, thus avoiding potential conflicts in decisions for any arm. In this version of the bandit problem, each agent a_i can pull any number of arms from subset C_i at each time step. Each arm $c \in K$ has a reward function $r_c(t)$ based on a p-dimensional covariate $X_t = (x_{1,t}, \ldots, x_{p,t})^T$ at time t (where $x_{i,t}$ is a random variable), as used in [Ginebra and Clayton, 1995; Yang and Zhu, 2002; Pavlidis et al., 2008a,b] (see Section 2.1 for a review on bandits with covariates):

$$r_c(t) = \sum_{j=0}^{p} \alpha_{c,j} x_{j,t} + \eta_{c,t} \ , \qquad \eta_{c,t} \sim \mathcal{N}(0, \sigma^2_{\eta_c}), \tag{3.1}$$

for $c = 1, \ldots, k$ and for $t = 1, \ldots, T$, where $x_{0,t} = 1$ and T is the length of the game. The covariate X_t is generated from a fixed multivariate distribution, with parameters unknown to the agents. The coefficient vectors α_c for $c = 1, \ldots, k$ are predetermined and also unknown to the agents, and precisely what the agents must learn.

Each agent a_i only observes a subset $Y_{i,t}$ of X_t, as information relevant to an agent is often distributed between agents in a multi-agent system. The agent can however request missing covariate values from another agent at a cost (denoted $\Pi_t(a_i, a_j)$). This is called the "communication stage" and is an important component of extending bandit problems to realistic MAS. Agents are assumed to know which covariates other agents have observed (or alternatively if a covariate value is requested from an agent that does not have this value then no cost is incurred). Agents are also assumed to receive covariate values truthfully if they are requested – this is feasible because there is no strategic communication in this framework as there is no competitive game structure. Agents can request several covariate values from several agents and the communication costs can be dependent on any function of X or t, which agents are communicating or the number of communications (which can also be limited by bandwidth capacity [Rogers et al., 2005]). In the simplest case, the communication cost is a constant value independent of X or t and is equal and known to each agent (this would often be the case that the cost of communication is not dependent on the information passed).

After the communication stage, each agent a_i must then make the decision as to which arms to pull from C_i; the agent only collects rewards from arms that are pulled. This is called the "action stage". Each agent therefore has a "two-stage" decision process which happens strictly sequentially; though effective strategies will consider the impact of one decision on the other. Algorithm 3.1 outlines the basic procedure each agent follows.

The agent's best strategy for communication and action depends on how its reward function $R_{a_i}(t)$ relates to the rewards of its action selections and also the communication cost function. The simplest form is:

$$R_{a_i}(T) = \sum_{t=1}^{T} r_{a_i}(t) \ , \qquad r_{a_i}(t) = \sum_{c \in S_{i,t}} r_c(t) - \sum_{j \neq i} \Pi_t(a_i, a_j), \tag{3.2}$$

where $S_{i,t}$ is the subset of arms pulled by agent a_i at time t. This reward function is simply the sum of all observed rewards over the game minus any communication costs. With this reward function, an agent will want to pull an

Algorithm 3.1 Two-stage decision process for agent a_i

```
for t = 1 to T do
  Observe Y_{i,t} ⊆ X_t
  for d = 1 to p do
    if x_{d,t} ∉ Y_{i,t} then
      Choose whether or not to request x_{d,t} from a_j where x_{d,t} ∈ Y_{j,t}
      If yes, incur communication cost Π_t(a_i, a_j).
    end if
  end for
  Choose arms to pull S_{i,t} ⊆ C_i
  Receive reward r_{a_i}(t)
end for
```

arm if the expected reward is positive. This creates a series of interdependent one-armed bandit problems with covariates (as introduced in Section 2.1), with the reward structure given in (2.1) where the reward of arm B is always zero (i.e. the reward coefficient of the arm with known expected reward is $\beta = 0$ and the added noise term has a degenerate distribution). The interdependence between the one-armed bandit problems occurs because the rewards are based on the same covariate X_t and the benefit of receiving one additional covariate value is shared between the arms, but the communication cost is only incurred once by the agent. Therefore, the strategies for one-armed bandit problems with covariates introduced in Section 2.1.2, are applicable to this framework, in particular the ϵ-greedy and ϵ-first strategies used to construct the double ϵ-greedy and double ϵ-first strategies introduced in the next section, respectively.

There are thus two learning problems for the agents: estimation of parameters subject to noisy data and a missing value problem. These have to be handled concurrently with reward seeking behaviour. The more arms that are pulled, the faster the learning; however rewards can be negative, so the exploration-exploitation tradeoff exists such that pulling an arm can still benefit the agent's overall reward even if the observed reward for that arm is negative. Moreover, the additional communication decision makes the problem of finding a good strategy more subtle (in that communication and action decisions have to be jointly considered), and hence a novel strategy is needed – in particular because exploration-exploitation of both action and communication decisions have not previously been considered in the same framework. This is discussed in more detail in the next section.

3.2 A Novel Strategy for Action and Communication Decisions

This extension to the multi-agent scenario introduces a two-stage decision process for each agent, as outlined in the previous section. In the communication stage, agents choose which missing covariates they would like to gain, and then request these values at a corresponding cost. There are essentially two reasons for an agent to communicate: the myopic gain to an agent's subsequent action decision and the improved learning of unknown parameters. The myopic gain can be valued using the *Value Of Communication* (VOC) constructed in

Section 3.2.1. There is an exploration-exploitation tradeoff however with the agent's communication decision, specifically the agent can now explore via communication and not only by action, because communication removes missing information and speeds up the agent's learning, which in turn benefits the expected reward of future decisions. To this end, we construct a double ϵ-greedy strategy and double ϵ-first strategy in Section 3.2.2, which both encourage exploration by communication, where the greedy decision is to pick the optimal myopic action using the VOC. Similarly, in the action stage, agents have the same two reasons to pull arms: the myopic gain to the reward function and the improved learning of unknown parameters. The optimal myopic action can be found as part of the VOC and the need for exploration forms part of the double ϵ-greedy/ϵ-first strategy.

Finally, whether or not an agent has communicated or acted, estimated parameters must be updated from what has been observed. In a bandit problem with fully observed covariates this could be done using regression (as in [Pavlidis et al., 2008b]), however in this framework an agent must handle missing data during parameter estimation. The agent has two basic choices of how to deal with missing data [Scheffer, 2002]. The first is case deletion, which in standard inference problems can be either listwise (deleting an entire case if it contains missing data) or pairwise (cases are only deleted if they contain missing data in the analysis being carried out). The second method is imputation, which involves estimating the missing values dependent on other values that have been observed. In the context of our problem, the agent estimates all the reward coefficients using linear regression, so the deletion would have to be listwise and hence this method throws away a lot of data when the agent does not observe the full covariate. For this reason, we use imputation. Specifically, we adopt a maximum likelihood approach and use the Expectation-Maximisation (EM) algorithm (outlined in Section 3.2.3). We can combine the EM algorithm with linear regression to update estimated reward coefficients in an effective way, in the presence of missing data.

3.2.1 The Value Of Communication (VOC)

Agent a_i observes a subset of covariates $Y_{i,t}$ at time t. After the communication stage the agent will observe a subset of covariates $Z_{i,t} \supseteq Y_{i,t}$. Agent a_i controls a subset of arms $C_i \subseteq K$ and must decide which arms $c \in C_i$ to pull. If agent a_i has a reward function given by (3.2) then the agent would pull arm $c \in C_i$ if $\mathrm{E}(r_c(t)|Z_{i,t}) > 0$ where:

$$\begin{aligned} \mathrm{E}(r_c(t)|Z_{i,t}) &= A + \sum_{x_{d,t} \notin Z_{i,t}} \alpha_{c,d} \int x_{d,t}\ \mathrm{p}(x_{d,t}|Z_{i,t}) dx_{d,t} \\ A &= \alpha_{c,0} + \sum_{x_{d,t} \in Z_{i,t}} \alpha_{c,d} x_{d,t}, \end{aligned} \tag{3.3}$$

where $r_c(t)$ is given by (3.1). Equation (3.3) is the myopic reward to agent a_i for pulling arm c at time t. Agent a_i can then find the optimal subset of arms $S_{i,t} \subseteq C_i$ to pull at time t using Algorithm 3.2.

Before agent a_i communicates, its expected reward at time t is the *Value Of*

Algorithm 3.2 Optimal myopic action for agent a_i at time t

Observe $Z_{i,t} \subseteq X_t$
for $c = 1$ to k **do**
  if $c \in C_i$ and $\mathrm{E}(r_c(t)|Z_{i,t}) > 0$ **then**
    $c \in S_{i,t}$
  end if
end for
Pull arms $S_{i,t} \subseteq C_i$
Receive reward $r_{a_i}(t)$

Silence (VOS) given by:

$$\mathrm{VOS}_{a_i} = \sum_{c \in C_i} \max\left(0, \mathrm{E}(r_c(t)|Y_{i,t})\right), \tag{3.4}$$

note that the VOS is bounded below by zero, corresponding to the agent pulling no arms at time t.

Agent a_i can gain a subset of covariates $D_{i,t} \subseteq Y_{i,t}^C$ by communication. If agent a_i knows the joint distribution of X_t then it can find the VOC for the subset $D_{i,t}$. This is given by:

$$\mathrm{VOC}_{a_i, D_{i,t}} = \sum_{c \in C_i} \mathrm{VOC}_{c, D_{i,t}} - \sum_{j \neq i} \Pi_t(a_i, a_j), \tag{3.5}$$

where,

$$\begin{aligned}
&\mathrm{VOC}_{c, D_{i,t}} \\
&= \int \max\left(0, B + \sum_{x_{d,t} \in D_{i,t}} \alpha_{c,d} x_{d,t}\right) \mathrm{p}(x_{D_{i,t}}|Y_{i,t}) dx_{D_{i,t}}
\end{aligned}$$

$$\begin{aligned}
B &= \alpha_{c,0} + \sum_{x_{d,t} \in Y_{i,t}} \alpha_{c,d} x_{d,t} \\
&+ \sum_{x_{d,t} \notin D_{i,t} \cap Y_{i,t}} \alpha_{c,d} \int x_{d,t}\ \mathrm{p}(x_{d,t}|Y_{i,t}) dx_{d,t},
\end{aligned}$$

The VOC reflects the probability that the expected reward of an arm after communicating $D_{i,t}$ is negative or positive. In the instances where this is negative the agent would not pull that arm and thus receive no reward, and still incur all costs of communication. The VOC is thus the expectation of the reward to agent a_i at time t if it requests covariates $D_{i,t}$. Agent a_i can maximise this value over all possible subsets $D_{i,t} \subseteq Y_{i,t}^C$ (not including the empty set, $D_{i,t} = \emptyset$), to find the maximum VOC value; however, the agent also requires this value to be bigger than the VOS, otherwise the agent should not communicate at all. If the communication cost was zero then trivially the maximum VOC would always correspond to choosing the full subset $D_{i,t} = Y_{i,t}^C$.

Algorithm 3.3 outlines how agent a_i finds the optimal subset of covariates to request by communication using the VOC. It must be stressed, however, that

this solution is myopic as the benefits of exploration are not factored in, this method simply maximises the agent's expected reward in the forthcoming action stage. As a result, the VOC is a similar calculation to the VPI (see Section 2.2) used to value the interaction between agents in coalition formation or service provider problems. The VPI assumes that unknown features of another agent are learnt immediately after interaction, thereby reducing the computational complexity of calculating the benefit of exploration. This is effectively the same concept used in the VOC where the optimal myopic action is taken – which would be optimal over the horizon of play if that action provided perfect information about the joint densities of the covariate values.

Furthermore, the VOC can only be found exactly with perfect knowledge of the conditional densities of the covariates, and of the coefficients of the reward function. These have to be learnt by the agents, and hence the approximation of the VOC improves over time. Exploration of communication therefore can have a positive effect on the cumulative reward, by improving communication decisions, which in turn will improve action decisions. Exploration of actions can benefit an agent's reward also, in the same way as with the bandit problems described in Section 2.1. The next section outlines an effective strategy that combines exploration by communication and action.

Algorithm 3.3 Optimal myopic communication decision for agent a_i at time t using the VOC

Observe $Y_{i,t} \subseteq X_t$
for all $D_{i,t} \in Y_{i,t}^C$ **do**
  for all $c \in C_i$ **do**
    Find $\text{VOC}_{c,D_{i,t}}$
  end for
  $\text{VOC}_{a_i,D_{i,t}} = \sum_{c \in C_i} \text{VOC}_{c,D_{i,t}} - \sum_{j \neq i} \Pi_t(a_i, a_j)$
end for
if $\max_{D_{i,t}} \text{VOC}_{a_i,D_{i,t}} > \text{VOS}_{a_i}$ **then**
  Request covariates $D_{i,t}$ by communication
else
  Do not communicate
end if

3.2.2 The Double ϵ-greedy and Double ϵ-first Strategies

The decisions made following the approach developed in the previous section can be seen as exploitive (or myopic) decisions; however for these communication and action decisions to be made more accurately over the length of play, agents have to perform exploration to aid their learning. It must be noted however that actions by the agents can be explorative as well as exploitive, particularly if the VOC encourages a high volume of communication, or if expected reward calculations encourage a high proportion of arms to be pulled. Nonetheless, in a noisy environment with unknown parameters it is likely that additional exploration may benefit the agent. Exploration by communication can be easily increased by requesting extra covariate values even if the VOC does not suggest this. Exploration by action, similarly, can be increased by pulling additional arms even if their expected reward is negative.

To encourage exploration by both communication and action, a double ϵ-greedy strategy is proposed, which uses the ϵ-greedy algorithm (described in Section 2.1.2) separately for both exploration methods. We use this strategy as the ϵ-greedy strategy has been widely found to perform well for a variety of multi-armed bandit problems [Auer et al., 2002; Vermorel and Mohri, 2005; Pavlidis et al., 2008b]. In the context of this framework, we can construct a double ϵ-greedy strategy, where the agents are exploitive in their communication decision with probability $1 - \epsilon_1$ (using the VOC) and are explorative and force additional communication with probability ϵ_1. Similarly, with the action decision, the agents can be exploitive with probability $1 - \epsilon_2$ (using estimated expected rewards) and extra actions (pulling of arms) can be forced with probability ϵ_2. This strategy is formalised in Algorithm 3.4. The optimal parameters, ϵ_1 and ϵ_2, will depend on factors such as the communication cost, the degree of noise in the data, and the unknown coefficients of the reward function.

Algorithm 3.4 Double ϵ-greedy strategy

Set ϵ_1 and ϵ_2
for $t = 1$ to T **do**
  Observe $Y_{i,t} \subseteq X_t$
  Find optimal subset of covariates $D_{i,t} \subseteq Y_{i,t}^C$ using VOC
  Request addition covariates in $Y_{i,t}^C$ with probability ϵ_1
  Find optimal subset of arms to pull $S_{i,t} \subseteq C_i$
  Pull additional arms in C_i with probability ϵ_2
end for

A similar strategy can be devised using the ϵ-first strategy as opposed to ϵ-greedy. The ϵ-first strategy (see Section 2.1.2) requires all exploration to be performed at the beginning rather than randomly throughout the game and has been shown to perform best in an empirical analysis of various strategies for multi-armed bandit problems [Vermorel and Mohri, 2005]. In this framework we can construct a double ϵ-first strategy, where all covariate values are requested and all arms are pulled for the first ϵT iterations; the agent is greedy afterwards and uses the VOC exclusively for communication decisions and maximises expected reward for action decisions. In the bandit problems reviewed in Section 2.1, we stated that ϵ-first would perform better than ϵ-greedy as the agent has more future decisions that benefit from past exploration. This would not necessarily be the case in the multi-agent bandit problem however, as the agent gains less myopic value in receiving all covariates for the first ϵT iterations, rather than having these additional covariate values spread throughout the game. Nonetheless, the benefit of using the double ϵ-first strategy is shown in a simulation study in Section 3.3; moreover the optimal balance between the two types of exploration (set through ϵ_1 and ϵ_2) is interdependent and shown to vary as the communication cost is changed.

3.2.3 Dealing with Missing Data

Agents have to iteratively update estimated parameters of the reward functions and covariates. As the reward functions are linear, with constant coefficients, the coefficients can be learnt using least squares estimation [Lawson and Hanson, 1995]. The agents, however, do not always observe all the covariates, even

after communication. This induces a missing value problem and the agent has the choice of imputing these missing values or deleting observations if they contain missing data (as discussed earlier). Due to the potential high occurrence of missing data, deletion methods are not practical for this framework. Given this background, we impute the data using a likelihood approach. To this end, an Expectation-Maximisation (EM) algorithm in conjunction with least squares estimation can be used, to iteratively update each agent's parameters. The EM algorithm is a computationally efficient and robust method for dealing with missing data, that can be practically implemented even if the number of agents/variables are high – which is important for the application of MAS to realistic scenarios.

In more detail, the EM algorithm [Dempster et al., 1977] is a procedure for maximum likelihood inference in the presence of missing data. Starting from an initial guess of the values for the parameter vector, $\theta(0)$, it employs an iterative update step, each time choosing $\theta(i+1)$ to maximise the expected log-likelihood of the observed data, where the expectation is taken over the missing data with respect to the current estimate $\theta(i)$. Once the change in expected log-likelihood is smaller than some pre-defined threshold then the algorithm terminates and missing values have been imputed.

3.3 Performance of the Strategy

In this section, the framework and suggested novel strategy is tested in a 2-agent version of the problem. Specifically, a 2-armed bandit problem with a 2-dimensional covariate is considered, where each agent controls one arm and always observes one covariate at each iteration, but never the other. This is the simplest possible formulation of our framework and is considered firstly to illustrate the selection behaviour of the strategies and secondly to show that exploration is needed even though the number of decisions faced by each agent is small.

In more detail, the reward function of agent a_i $(i = 1, 2)$ is:

$$r_i(t) = \alpha_{i,1} x_{1,t} + \alpha_{i,2} x_{2,t} + \eta_{i,t}. \tag{3.6}$$

The coefficients, $\alpha_{i,1}$ and $\alpha_{i,2}$, are predetermined and unknown to the agents. The covariate values X_t are i.i.d. draws from a bivariate normal distribution (in keeping with other studies of bandits with covariates [Yang and Zhu, 2002; Pavlidis et al., 2008b]):

$$X \sim \mathcal{N}(\mu, \Sigma) \quad \mu = \mathbf{0}, \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

where the parameters are unknown to the agents. Agent a_i only observes $x_{i,t}$ at each iteration. The noise $\eta_{i,t}$ is also normally distributed and i.i.d., with zero mean and variance 0.5. The length of play considered is 100 iterations, long enough for the agents to start exploiting, but short enough so that the agents must learn quickly and effectively.

3.3.1 Application of the VOC

The VOC could be found exactly, if agent a_i knew $\alpha_{i,1}$, $\alpha_{i,2}$, μ and Σ; however the agent must learn these over time. In this 2-agents scenario, the VOC from

(3.5) becomes (see Appendix I):

$$\text{VOC}_{a_i,x_{j,t}} = \Phi\left(-\text{sign}(\alpha_{i,j})\frac{x_{i,t}\left(\frac{\alpha_{i,i}}{\alpha_{i,j}}-\rho\right)}{1-\rho^2}\right)x_{i,t}\left(\alpha_{i,i}+\alpha_{i,j}\rho\right)$$
$$+\frac{\alpha_{i,j}\left(1-\rho^2\right)}{\sqrt{2\pi}}\exp\left(\frac{-\left(x_{i,t}\left(\frac{\alpha_{i,i}}{\alpha_{i,j}}-\rho\right)\right)^2}{2(1-\rho^2)^2}\right)-\Pi_t(a_i,a_j). \tag{3.7}$$

for $i = 1, 2$ (where $j = 2, 1$). $\Phi(x)$ is the cdf of the standard normal distribution. It is therefore only dependent on the parameters that agent a_i needs to learn, the observed covariate $x_{i,t}$ and the communication cost (assumed constant). The VOS from (3.4) simply becomes (see Appendix I):

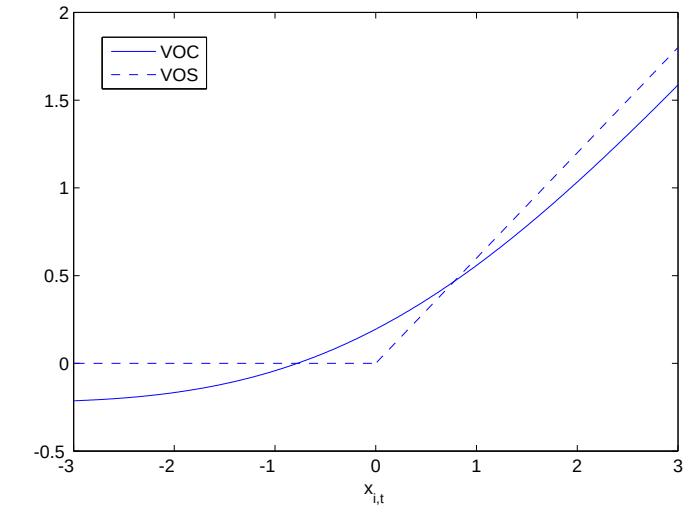
$$\text{VOS}_{a_i} = \max(0, x_{i,t}\left(\alpha_{i,i}+\alpha_{i,j}\rho\right)) \tag{3.8}$$

Agent a_i communicates if the VOC is greater than the VOS. To this end, Fig 3.1(a) shows how the VOC_{a_i} and VOS_{a_i} can change for different values of $x_{i,t}$. In particular, there is a region where the VOC is higher and the agent should communicate; the agent can find this "region of communication" over time, by learning the parameters correctly. In this region the unknown covariate value will be informative as to whether the expected reward is positive or negative. Conversely, for covariate values outside the region of communication, the action decision is clear as the agent knows whether the reward is likely to be positive or negative and it is not worth incurring the communication cost to verify this.

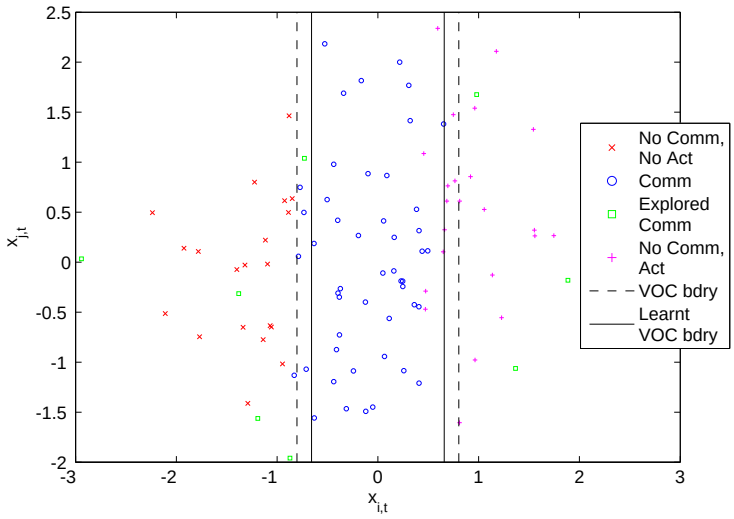
Figure 3.1(b) shows how the agent, using the double ϵ-first strategy, has learnt the region of communication and made the correct decision for most observed covariate values. The points highlighted by squares show the points where exploration by communication has occurred (i.e. the agent has communicated outside of the region of communication to aid its learning). For other parameter values the region of communication may not exist (if the covariance between the known and unknown covariate is high for example) or be infinite (if the communication cost is zero for example). Nevertheless, this region does not have to be explicitly found as the VOC needs only to be calculated at the covariate values observed at each iteration.

3.3.2 Performance of the Double ϵ-first Strategy

In the previous section, we outlined how the agents can approximate the VOC explicitly and demonstrated that this approximation can be made accurately with some degree of exploration. In this section, the effect of the double ϵ-first strategy on the agent's cumulative reward is explored. We use the double ϵ-first strategy for this problem as opposed to the double ϵ-greedy strategy. For the 2-dimensional case considered in this section, the double ϵ-first strategy will perform better as there is only one explorative choice for both communication and action and it is beneficial to do these explorative steps early in the game (as discussed in Section 2.1.2). Specifically, Fig 3.2 shows the average cumulative reward to agent a_i, for various communication costs, using the double ϵ-first strategy over a grid of values for ϵ_1 and ϵ_2 ranging between 0 and 50%. For this setting, the agent benefits from exploring by both communication and action,



(a)



(b)

Figure 3.1: (a) VOC_{a_i} and VOS_{a_i} over different values of $x_{i,t}$ and (b) decisions by agent a_i over 100 iterations with ϵ_1 and ϵ_2 set at 10%. In both figures, $\alpha_{i,i} = 0.5$, $\alpha_{i,j} = 1$, $\rho = 0.1$ and $\Pi_t(a_i, a_j) = 0.2$

as there appears to be a global maximum over the space of the two parameters occurring at $\epsilon_1 > 0$ and $\epsilon_2 > 0$ for all communication costs considered. Additionally, there appears to be a correlation between ϵ_1 and ϵ_2 that is dependent on the communication cost. As expected, the amount of optimal exploration by communication (ϵ_1), is inversely related to the communication cost. To a lesser extent, the amount of optimal exploration by action (ϵ_2) is positively correlated with the communication cost; this is due to the fact that a total amount of *global* exploration is required, and as exploration by communication becomes more costly, the agent requires more exploration by action to perform a reasonable amount of learning (and vice-versa).

3.4 Summary

In this chapter we have proposed a new framework for modelling sequential decision making problems in MAS. We extended the multi-armed bandit problem to investigate the exploration-exploitation tradeoff in a multi-agent context. Specifically, we investigated sequential decision making of communication decisions between agents, which is relevant and applicable to many other MAS. This framework is novel in that exploration-exploitation of joint action and communication decisions is considered in the same problem, however the framework is restricted in that the interaction of agents is constrained to communicating information – there is no interaction of rewards between agents. This is an extension to be considered in future work (see Chapter 5 for more details).

In more detail, we have constructed a novel strategy for selecting communication and action decisions. The exploitive element of the strategy involves using the VOC to myopically value communication and action decisions. The explorative element involves using a double ϵ-greedy or double ϵ-first strategy to communicate with agents and pull arms to benefit the agent's learning. In an empirical evaluation of a 2-agent problem, the strategy was shown to outperform the greedy strategy. Moreover, we have shown that our strategy, which combines exploration by both communication and action, performs better than doing exploration by one method and not the other. In particular, we have shown that agents can benefit from exploring by communication – agents should hence not communicate with other agents for myopic gain only. This novel framework has therefore developed new ideas about balancing exploration-exploitation in a multi-agent setting where rewards of actions are unknown *a priori*. The framework also includes the possibility of agent's communicating, which is central to many real world scenarios modelled by MAS.

The benefit of our strategy has only been demonstrated empirically thus far. For this reason, in the next chapter we develop a novel approach to reasoning about the one-armed bandit problem with covariates theoretically. This approach is initially considered in a single-agent context, with no communication of information, although the ideas will be extended to reason theoretically about the multi-agent context considered in this chapter, as part of future work (see Chapter 5 for details).

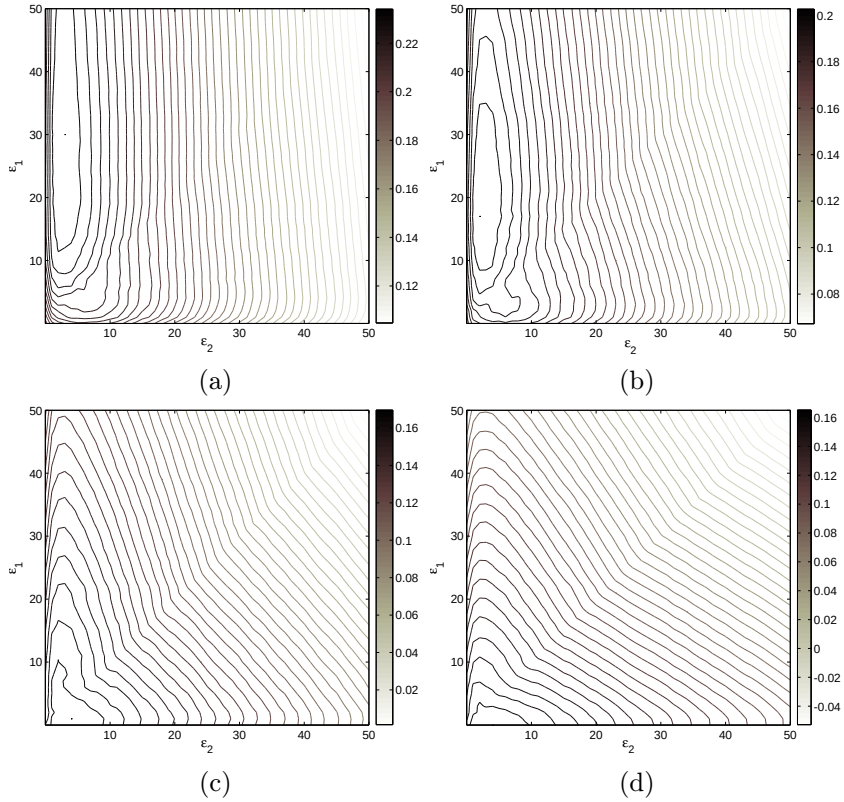


Figure 3.2: The average cumulative reward to agent a_i, for various communication costs: a) 0.05, (b) 0.1, (c) 0.2 and (d) 0.3, using a double ϵ-first strategy where $\alpha_{i,i} = 0.5$, $\alpha_{i,j} = 1$ and $\rho = 0.1$.

Chapter 4

One-Armed Bandits with Covariates

In the previous chapter we presented an extension of the standard bandit problem with covariates to MAS, where each agent controlled a subset of the arms. The agent thus faced a series of one-armed bandit problems with covariates and we constructed a strategy (based on either the ϵ-greedy or ϵ-first strategies) that selected arms to pull from this subset but also selected which agents to communicate with. We presented empirical evidence to suggest that our explorative strategy can benefit the agent's expected reward. For this reason, in this chapter, we reason about the one-armed bandit problem theoretically to develop ideas of optimal tuning for the ϵ-greedy and ϵ-first strategies. We consider the single-agent problem defined in Section 2.1.1 and extend this to the multi-agent problem as part of future work (see Chapter 5 for details). Specifically, the agent must select between an arm with unknown expected reward (arm A) and an arm with known expected reward (arm B).

Existing studies of one-armed bandit problems with covariates have been largely concerned with maximising reward over infinite-length play (see Section 2.1.2 for a review). In this work, however, we are concerned with maximising reward in finite time, which is more relevant in real applications. Given this perspective, we present a novel approach for reasoning about the expected reward of arm selection strategies, by modelling the distribution of estimated parameters in the reward function. This helps us find the *probability of error*; that is, the probability that the agent pulls the arm with lower expected reward when trying to pull the best arm, given an arm selection strategy. This measure is important because it helps us find the expected reward when the agent is selecting greedily between the arms. The probability of error is therefore crucial to finding the expected reward (in finite time) of any strategy that exploits the covariate values to select between the arms.

Explorative strategies will pull arm A more often as this helps the agent to learn parameters of the reward function more quickly, reducing the probability of error – this is the *benefit of exploration*. Conversely, such selections have an attributed *cost of exploration*, as the agent might be selecting the arm with lower expected reward. In our setting, we can explicitly calculate this benefit and cost of exploration and hence capture the exploration-exploitation tradeoff

in the same currency. We can then find the expected cumulative reward of certain strategies in finite time, and hence reason about their optimal tuning.

We derive the expected reward of both the ϵ-greedy and ϵ-first strategy in finite time for a reward function based on a 1-dimensional covariate. This simplified model assumes that all side information is represented in one variable. Nevertheless, this assumption allows for a clear demonstration of the novel method used to find expected rewards. Furthermore, this approach will be extended to p-dimensional covariates, as used in Chapter 3, as part of future work (see Chapter 5 for details).

In more detail, we prove that, in the 1-dimensional setting, the expected reward of the ϵ-greedy strategy is maximised by $\epsilon = 0$ irrespective of the length of play and all other parameters. This means that, on average, a greedy strategy will outperform any ϵ-greedy strategy in finite time. This result is in line with the infinite-time statements proved in [Woodroofe, 1979; Sarkar, 1991]. Moreover, contrary to the findings of finite-time analyses of multi-armed bandits [Auer et al., 2002; Vermorel and Mohri, 2005; Pavlidis et al., 2008b], we have proved that ϵ-greedy is a suboptimal strategy for the one-armed bandit problem considered here.

For the ϵ-first strategy, however, we show that the optimal value of ϵ will be non-zero for certain lengths of play. In particular, we find the optimal value of ϵ numerically and present results to show its dependence on the length of play. The significance of this result, is that a well-tuned explorative strategy will, on average, outperform the greedy strategy (a purely exploitive strategy). The novel approach presented in this chapter has consequently allowed for the optimal balance of exploration and exploitation to be found theoretically, for this strategy.

This chapter is structured as follows. In Section 4.1 we introduce the one-armed bandit with covariates framework. In Section 4.2 we model the distribution of estimated parameters in the reward function and use this to find the probability of error over time. In Section 4.3 we derive the expected reward of the ϵ-greedy strategy and prove that this is maximised with $\epsilon = 0$. In Section 4.4 we derive the expected reward of the ϵ-first strategy and show numerically that non-zero values of ϵ can be optimal. Summary remarks follow in Section 4.5.

4.1 The One-Armed Bandit with Covariates Framework

An agent plays a one-armed bandit problem and must choose at time $t = 0, \ldots, T$ between arm A with unknown expected reward and arm B with known expected reward. The agent only receives a reward from the arm that is pulled, which is a function of an observed covariate (X_t). We consider the reward structure used in [Ginebra and Clayton, 1995; Yang and Zhu, 2002; Pavlidis et al., 2008a,b] (see (2.1) in Section 2.1.1), simplified to a 1-dimensional covariate, where:

$$\begin{aligned} r_A(t) &= \alpha x_t + \eta_t, \\ r_B(t) &= \beta x_t + \omega_t, \end{aligned} \tag{4.1}$$

where $r_A(t)$ and $r_B(t)$ are the rewards of arm A and B, respectively; η_t and ω_t are i.i.d. noise terms drawn from $\mathcal{N}(0, \sigma_\eta^2)$ and $\mathcal{N}(0, \sigma_\omega^2)$, respectively. The coefficient β is known to the agent *a priori*, but α is unknown and estimated from observations. The covariate x_t is an i.i.d. draw from $\mathcal{N}(0, \sigma_x^2)$ and the agent must then either pull arm A and receive reward $r(t) = r_A(t)$ or pull arm B and receive reward $r(t) = r_B(t)$. The objective is for the agent to maximise the cumulative reward $R(T) = \sum_{t=0}^{T} r(t)$.

4.2 The Probability of Error

The agent must learn the value of α in (4.1) as it plays. Suppose the agent has pulled the arm k times prior to time t, where $1 \leq k \leq t$. The estimate of α is updated using $\widehat{\alpha}_k$, the solution of the linear least squares equation:

$$\widehat{\alpha}_k = \frac{\sum_{j=1}^{k} x_j r_A(j)}{\sum_{j=1}^{k} x_j^2}. \tag{4.2}$$

The parameter estimate $\widehat{\alpha}_k$ has a distribution that is centred at α and dependent on the number of pulls k and the distribution of x_t and η_t. As η_t is i.i.d. and normally distributed then it follows that (see p.407 [Daly et al., 1995]):

$$\widehat{\alpha} \sim \mathcal{N}\left(\alpha, \frac{\sigma_\eta^2}{\sum_{j=1}^{k} x_j^2}\right) \quad \Rightarrow \quad (\widehat{\alpha}_k - \alpha)\sqrt{\frac{k\sigma_x^2}{\sigma_\eta^2}} \sim \frac{\mathcal{N}(0,1)}{\sqrt{\sum_{j=1}^{k}\left(x_j^2/\sqrt{\sigma_x^2}\right)/k}}.$$

If the agent uses an ϵ-greedy or ϵ-first strategy, then the collection of covariates x_t used for estimation are not drawn i.i.d. from $\mathcal{N}(0, \sigma_x^2)$ (unless $\epsilon = 1$ and arm A is pulled each time). When the agent is greedy, however, arm A is pulled based on whether x_t is positive or negative and hence $\sum_{j=1}^{k} x_j^2/\sqrt{\sigma_x^2} \sim \chi_k^2$ (chi-square distribution with k degrees of freedom). Then it follows from the definition of the t-distribution that:

$$(\widehat{\alpha}_k - \alpha)\sqrt{\frac{k\sigma_x^2}{\sigma_\eta^2}} \sim t_k, \tag{4.3}$$

where t_k is the t-distribution with k degrees of freedom. From (4.3) we can find the probability of error after k pulls, that is the probability that the agent pulls the wrong arm when being greedy. Specifically this occurs if $\widehat{\alpha}_k > \beta$ when $\alpha < \beta$ and vice-versa. We therefore define this probability as:

$$F(k) = \Pr(\widehat{\alpha}_k < \beta | \alpha > \beta)\Pr(\alpha > \beta) + \Pr(\widehat{\alpha}_k > \beta | \alpha < \beta)\Pr(\alpha < \beta). \tag{4.4}$$

First consider $\Pr(\widehat{\alpha}_k < \beta | \alpha > \beta)$. It follows from (4.3) that:

$$\begin{aligned}\Pr(\widehat{\alpha}_k < \beta | \alpha > \beta) &= \Pr\left(\sqrt{\frac{k\sigma_x^2}{\sigma_\eta^2}}(\widehat{\alpha}_k - \alpha) < \sqrt{\frac{k\sigma_x^2}{\sigma_\eta^2}}(\beta - \alpha)\right) \\ &= T\left((\beta - \alpha)\sqrt{\frac{k\sigma_x^2}{\sigma_\eta^2}}, k\right),\end{aligned} \tag{4.5}$$

where $T(x,k)$ is the t-distribution cumulative density function at ordinate x, with k degrees of freedom. By considering $\Pr(\widehat{\alpha}_k > \beta | \alpha < \beta)$ in the same way, it follows from (4.4) and (4.5) that:

$$\begin{aligned} F(k) &= \Pr(\alpha > \beta) T\left(-|\alpha-\beta|\sqrt{\frac{k\sigma_x^2}{\sigma_\eta^2}}, k\right) + \Pr(\alpha < \beta) T\left(-|\alpha-\beta|\sqrt{\frac{k\sigma_x^2}{\sigma_\eta^2}}, k\right) \\ &= T\left(-|\alpha-\beta|\sqrt{\frac{k\sigma_x^2}{\sigma_\eta^2}}, k\right). \end{aligned} \tag{4.6}$$

The probability of error $F(k)$ has the following four properties:

1. $F(k)$ is (strictly) bounded above by 0.5.
2. $F(k)$ decreases as k increases, as both the ordinate becomes more negative and the degrees of freedom increase (reducing the weight in the tails). $F(k)$ is also a convex sequence in k (proved later).
3. Increasing the difference between α and β reduces the value of $F(k)$.
4. The ratio σ_x^2/σ_η^2 can be interpreted as a 'signal to noise ratio' – larger values of this ratio reduce $F(k)$.

Property 1 ensures that the agent can do no worse than guessing between the arms. Notice however, that as $\sigma_\eta^2 \to \infty$, $F(k) \to 0.5$. Figure 4.1(a) shows the probability of error over k for several values of σ_x^2/σ_η^2, where properties 1, 2 and 4 are demonstrated.

4.3 The ϵ-greedy Strategy

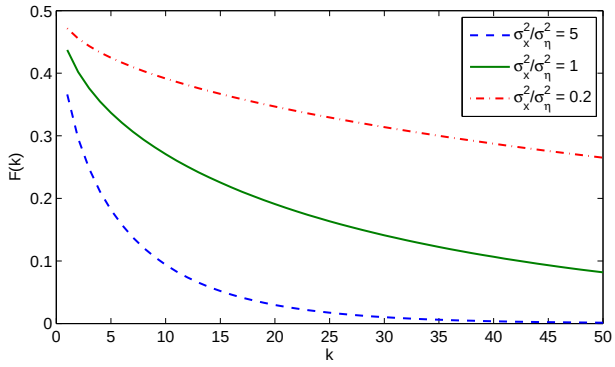
The ϵ-greedy strategy dictates that the agent pulls arm A with probability ϵ but picks the arm with highest expected reward with probability $1-\epsilon$. In the previous section, we found the probability of error given k pulls of arm A by time t. In fact, we can find the distribution of k given t by the symmetry of the problem, as x_t is centrally distributed and therefore arm A is pulled 50% of the time when the agent is greedy. One pull of arm A is guaranteed at time $t = 0$, so the probability of having pulled the arm k times by time t, $B(k,t,\epsilon)$, follows a binomial distribution:

$$B(k,t,\epsilon) = \binom{t-1}{k-1} \left(\frac{1}{2}(1+\epsilon)\right)^{k-1} \left(\frac{1}{2}(1-\epsilon)\right)^{t-k}. \tag{4.7}$$

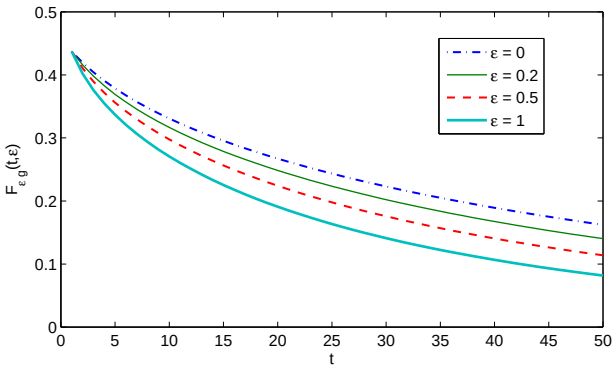
The distribution of $\widehat{\alpha}$ after k pulls of arm A using the ϵ-greedy strategy is as given in (4.3). The distribution of k can then be used to find the probability of error at time t of the ϵ-greedy strategy:

$$F_{\epsilon g}(t,\epsilon) = \sum_{k=1}^{t} B(k,t,\epsilon) F(k). \tag{4.8}$$

As $F(k) < 0.5$ and $\sum_{k=1}^{t} B(k,t,\epsilon) = 1$ it follows that $F_{\epsilon g}(t,\epsilon) < 0.5$ (and all other properties of the probability of error mentioned in Section 3 still hold).



(a)



(b)

Figure 4.1: (a) The sequence $F(k)$ from (4.6) for $k = 1, \ldots, 50$ for various values of σ_x^2/σ_η^2 and (b) The sequence $F_{\epsilon g}(t, \epsilon)$ from (4.8) for $t = 1, \ldots, 50$ for various ϵ, where $\sigma_x^2/\sigma_\eta^2 = 1$. In both figures, $\alpha = 0.5$ and $\beta = 0.3$.

$F_{\epsilon g}(t, \epsilon)$ has the additional property that it decreases as ϵ increases, for a fixed t. Figure 4.1(b) shows the sequence $F_{\epsilon g}(t, \epsilon)$ for a selection of ϵ values from $t = 1, \ldots, 50$, where this property is demonstrated. The expected instantaneous reward of the ϵ-greedy strategy at time t, $\mathrm{E}(r_{\epsilon g}(t, \epsilon))$, can now be found by considering the cases when the agent explores and exploits separately.

$$\mathrm{E}(r_{\epsilon g}(t, \epsilon)) = \epsilon \mathrm{E}(r_A(t)) + (1 - \epsilon)\mathrm{E}(r_g(t, \epsilon)), \tag{4.9}$$

where $r_g(t, \epsilon)$ is the instantaneous reward when the agent is greedy. It follows that as $x_t \sim \mathcal{N}(0, \sigma_x^2)$:

$$\mathrm{E}(r_A(t)) = \int_{-\infty}^{\infty} \alpha x_t \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{x_t^2}{2\sigma_x^2}\right) \mathrm{d}x_t = 0. \tag{4.10}$$

From the probability of error $F_{\epsilon g}(t, \epsilon)$ we can find the expected instantaneous reward when the agent is greedy, by separately considering the expected instantaneous reward when the correct/incorrect arm is pulled, dependent on the probability of error (see Appendix II). It follows that:

$$\mathrm{E}(r_g(t, \epsilon)) = |\alpha - \beta| \sqrt{\frac{\sigma_x^2}{2\pi}} \left(1 - 2F_{\epsilon g}(t, \epsilon)\right), \tag{4.11}$$

which yields the expected instantaneous reward of the ϵ-greedy strategy:

$$\mathrm{E}(r_{\epsilon g}(t,\epsilon)) = (1-\epsilon)\,|\alpha-\beta|\,\sqrt{\frac{\sigma_x^2}{2\pi}}\,(1-2F_{\epsilon g}(t,\epsilon))\,. \tag{4.12}$$

This expected reward is greater than zero as $F_{\epsilon g}(t,\epsilon) < 0.5$, so the strategy performs better than guessing between the arms for all values of ϵ (except $\epsilon = 1$). Larger values of ϵ reduce the probability of error $F_{\epsilon g}(t,\epsilon)$, which increases the expected reward – this is the *benefit of exploration*. Conversely, larger values of ϵ reduce the $(1-\epsilon)$ term in the expected reward and this is the *cost of exploration*. Despite this exploration-exploitation tradeoff, the expected instantaneous reward in (4.12) is maximised by $\epsilon = 0$ for all values of $t > 0$, α, β and σ_x^2/σ_η^2, which we prove in Theorem 4.1 below.

Theorem 4.1 *$E(r_{\epsilon g}(t,0)) > E(r_{\epsilon g}(t,\epsilon))$ for all $0 < \epsilon \leq 1$ and for all $t \in \mathbb{Z}^+$, $\alpha, \beta \in \mathbb{R}$ and $\sigma_x^2, \sigma_\eta^2 \in \mathbb{R}^+$.*

Proof We prove Theorem 4.1 by contradiction. First consider the case $t \geq 2$. Suppose there exists $0 < \epsilon \leq 1$ such that:

$$\mathrm{E}(r_{\epsilon g}(t,\epsilon)) \geq \mathrm{E}(r_{\epsilon g}(t,0)) \quad \text{for some } t \in \mathbb{Z}^+, \alpha, \beta \in \mathbb{R}, \sigma_x^2, \sigma_\eta^2 \in \mathbb{R}^+. \tag{4.13}$$

Substituting from (4.12) and (4.8), the inequality in (4.13) becomes:

$$\sum_{k=1}^{t} F(k)G(k) \geq \frac{\epsilon}{2}, \tag{4.14}$$

where $G(k) = B(k,t,0) - (1-\epsilon)B(k,t,\epsilon)$. Notice that $F(k) < 1/2$ and $\sum_{k=1}^{t} G(k) = \epsilon$, however it does not follow from this alone that $\sum_{k=1}^{t} F(k)G(k) < \epsilon/2$ (by Cauchy-Schwarz for example), as $G(k)$ can be negative for certain values of k. To proceed, the following three lemmas allow for a useful upper bound to be constructed on the left-hand side of (4.14). The proofs of the three lemmas can be found in Appendix III.

Lemma 4.2 *$F(k)$ is a convex sequence in k.*

Lemma 4.3 *There exists an integer q where $2 \leq q \leq t$ such that:*

$$\begin{aligned} G(k) &\geq 0 \quad \text{for } k = 1, \ldots, q \\ G(k) &< 0 \quad \text{otherwise.} \end{aligned}$$

Lemma 4.4

$$\sum_{k=1}^{t} F(k)G(k) \leq \sum_{k=1}^{t} F'(k)G(k)\ , \quad \text{where} \quad F'(k) = \frac{q-k}{q-1}F(1) + \frac{k-1}{q-1}F(q).$$

After expanding the binomial coefficients and rearranging (see Appendix IV):

$$\sum_{k=1}^{t} F'(k)G(k) = \frac{1}{2}(2\epsilon - \epsilon^2)F(1) + \frac{1}{2}\epsilon^2 F(q) + \frac{1}{2}\frac{t-q}{q-1}\epsilon^2(F(q) - F(1)). \tag{4.15}$$

As $\frac{1}{2}\frac{t-q}{q-1}\epsilon^2 > 0$, $F(q) < F(1)$ and $F(k) < 0.5$, then the third term is negative and hence from Lemma 4.4:

$$\sum_{k=1}^{n} F(k)G(k) \leq \frac{1}{2}(2\epsilon - \epsilon^2)F(1) + \frac{1}{2}\epsilon^2 F(q) < \frac{1}{2}(2\epsilon - \epsilon^2)\frac{1}{2} + \frac{1}{2}\epsilon^2\frac{1}{2} = \frac{\epsilon}{2}.$$

A contradiction has been made and $\sum_{k=1}^{t} F(k)G(k) < \epsilon/2$, therefore Theorem 4.1 has been proved for $t \geq 2$. It remains to show that the theorem holds for $t = 1$. Using (4.12) and (4.8):

$$\begin{aligned}\mathrm{E}(r_{\epsilon g}(1,0)) &= |\alpha - \beta| \sqrt{\frac{\sigma_x^2}{2\pi}} (1 - 2F(1)) \\ &> (1-\epsilon) |\alpha - \beta| \sqrt{\frac{\sigma_x^2}{2\pi}} (1 - 2F(1)) = \mathrm{E}(r_{\epsilon g}(1,\epsilon)),\end{aligned}$$

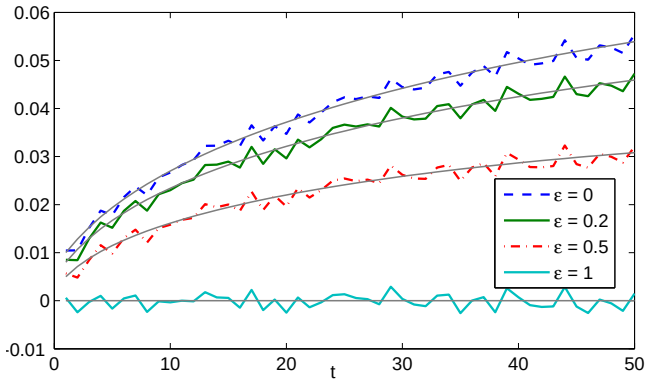
as $0 < \epsilon \leq 1$ and $F(1) < 0.5$. This completes the proof of Theorem 4.1. □

Theorem 4.1 states that the expected instantaneous reward at time t is maximised by $\epsilon = 0$. It is then immediate that the cumulative reward $R(T) = \sum_{t=0}^{T} r(t)$, which is what we wish to maximise, is also maximised by $\epsilon = 0$. This implies that the greedy strategy, on average, outperforms any ϵ-greedy strategy for this one-armed bandit problem. Given these findings, Figure 4.2 shows the averaged instantaneous and cumulative reward at time t from repeated simulations of the same problem, with the theoretical expectations overlayed. The empirical evidence verifies the theoretical findings that the instantaneous reward (and hence cumulative reward also) is maximised by $\epsilon = 0$. We have thus shown, for the first time, that the ϵ-greedy strategy is a suboptimal strategy for this one-armed bandit problem with covariates.

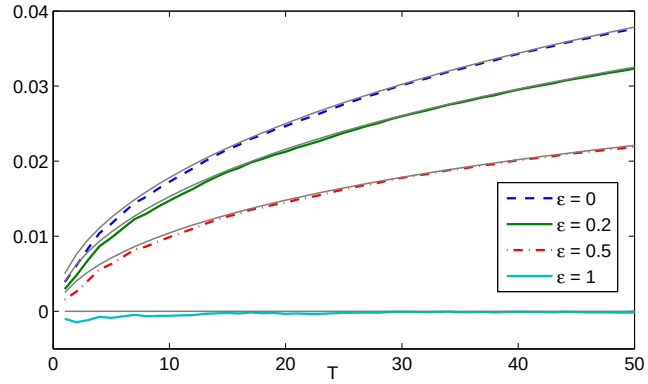
In contrast, finite time analyses of multi-armed bandit problems [Auer et al., 2002; Vermorel and Mohri, 2005; Pavlidis et al., 2008b] have concluded, through empirical evidence, that the optimally tuned ϵ-greedy strategy can have $\epsilon > 0$. The difference between this evidence and our findings, is due to the exploration requirements of the two problems. In our one-armed bandit problem only one arm requires any exploration, and since this arm is already selected 50% of the time with a greedy strategy, no further exploration is required with an ϵ-greedy strategy. Conversely, in a multi-armed bandit problem, each arm requires exploration. Moreover there are no such guarantees on exploring each arm sufficiently with a greedy strategy and consequently optimal arms are often overlooked. As a result, an ϵ-greedy strategy with $\epsilon > 0$, can outperform the greedy strategy.

4.4 The ϵ-first Strategy

The ϵ-first strategy dictates that all the agent's exploration is at the beginning (for the first ϵT iterations) followed by greedy selection for the remaining iterations. When the agent explores, arm A is always pulled and it follows from (4.10) that the expected reward of ϵ-first $E(r_{\epsilon f}(t)) = E(r_A(t) = 0)$ for $t \leq \epsilon T$. To find the expected reward for $t > \epsilon T$, consider the probability of error $F_{\epsilon f}(t, \epsilon)$



(a)



(b)

Figure 4.2: (a) Instantaneous reward and (b) cumulative reward (averaged over time) for a range of ϵ-greedy strategies averaged over 200,000 simulations, where $\alpha = 0.5$, $\beta = 0.3$ and $\sigma_x^2/\sigma_\eta^2 = 1$. Theoretical expectations are overlayed (in grey).

with this strategy. Using the same reasoning as before, we find:

$$F_{\epsilon f}(t,\epsilon) = \sum_{k=1}^{t-T\epsilon} B(k, t-T\epsilon, 0) F(k+\epsilon), \tag{4.16}$$

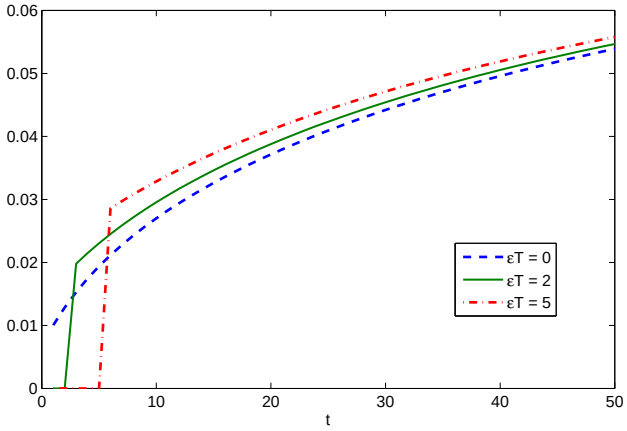
with $F(k)$ as given in (4.6) and $B(k,t,\epsilon)$ as given in (4.7). Therefore, $F_{\epsilon f}(t,\epsilon) < 0.5$ as with the ϵ-greedy strategy. In the same way that (4.11) was derived (see Appendix II), it follows that:

$$\mathrm{E}(r_{\epsilon f}(t,\epsilon)) = |\alpha-\beta| \sqrt{\frac{\sigma_x^2}{2\pi}} \left(1 - 2F_{\epsilon f}(t,\epsilon)\right) \qquad \text{for } t > \epsilon T. \tag{4.17}$$

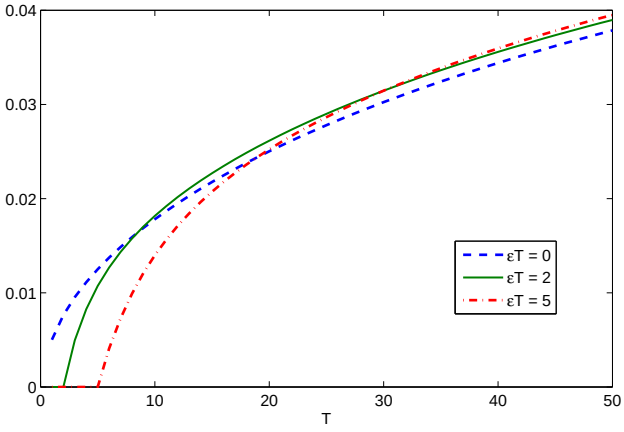
Again this expected reward is positive as $F_{\epsilon f}(t,\epsilon) < 0.5$. Larger values of ϵ reduce the probability of error for $t > \epsilon T$ and thus have a higher expected reward in this region – this is the *benefit of exploration*. Conversely, larger values of ϵ correspond to a longer period of exploration where the expected reward is zero – this is the *cost of exploration*. The expected cumulative reward

is $\mathrm{E}(R_{\epsilon_f}(T,\epsilon)) = \sum_{t=0}^{T} \mathrm{E}(r_{\epsilon_f}(t,\epsilon))$ and we can maximise this numerically using (4.17) to find the optimal ϵ.

This optimal value will not necessarily be zero as the following numerical results show. In particular, Fig. 4.3(a) displays the expected reward at time t, for the game of length $T = 50$ shown in Fig. 4.2, for various values of ϵT, where the benefit and cost of exploration can be clearly seen. Summing the rewards from Fig. 4.3(a) generates Fig. 4.3(b) which is the expected cumulative reward at time $T = 1, \ldots, 50$ for the fixed values of ϵT. Fixing ϵT in this way has shown that the greedy strategy can be beaten and there are regions of T where $\epsilon T = 0, 1, 2, \ldots$ are optimal in terms of maximising the expected cumulative reward.



(a)

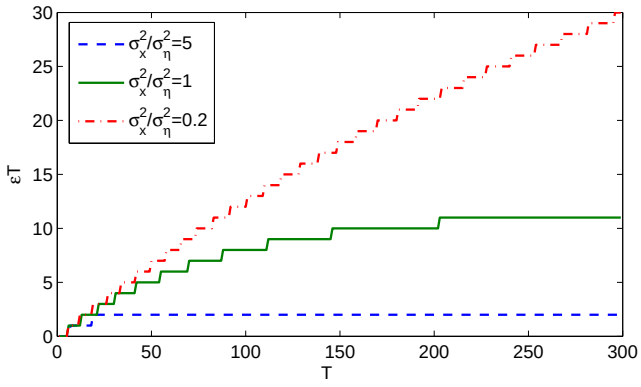


(b)

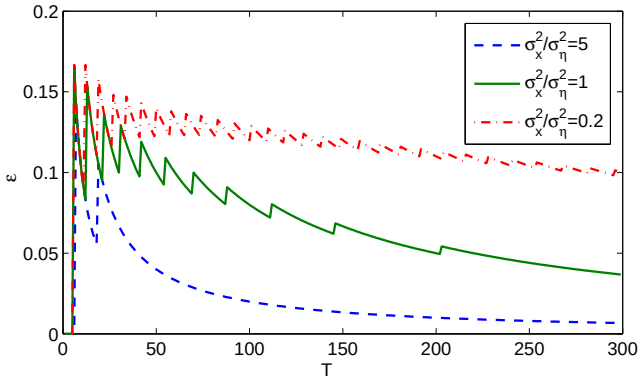
Figure 4.3: (a) Expected reward of the ϵ-first strategy at time t for $T = 50$ and (b) expected cumulative reward (averaged over time) for $T = 1, \ldots, 50$, for a range of ϵT values, where $\alpha = 0.5$, $\beta = 0.3$ and $\sigma_x^2/\sigma_\eta^2 = 1$.

Figure 4.4(a) displays how the optimal value of ϵT grows with T for various values of σ_x^2/σ_η^2. The range of values of T where a specific value of ϵT is optimal become larger as ϵT increases, indicating that ϵT grows more slowly than

T. This idea can also be seen in Fig. 4.4(b), which shows a plot of optimal ϵ for values of T. The non-smooth shape of this plot is due to the fact that ϵT is represented as a step-function taking integer values only. The key observation, however, is that the optimal ϵ decreases and approaches zero as $T \to \infty$ (although for $\sigma_x^2/\sigma_\eta^2 = 0.2$ this happens very slowly), which concurs with the studies of [Woodroofe, 1979; Sarkar, 1991] which proved the greedy strategy was asymptotically optimal. Nevertheless, in finite time, a well chosen ϵ in the ϵ-first strategy will outperform the greedy strategy, signifying the benefit of balancing exploration and exploitation in an arm selection strategy.



(a)



(b)

Figure 4.4: Optimal values of (a) ϵT and (b) ϵ, for $T = 1, \ldots, 300$, where $\alpha = 0.5$, $\beta = 0.3$

4.5 Summary

We have presented a novel approach for considering one-armed bandit problems in finite time, by modelling the exact distributions of estimated parameters over time. In particular, for a reward based on a 1-dimensional covariate, we have derived the expected reward of the ϵ-greedy and ϵ-first strategies, which are popular exploration strategies in sequential decision making problems. Furthermore, we have proved that the expected reward of the ϵ-greedy strategy

is always maximised by $\epsilon = 0$. In other words, the greedy strategy outperforms any ϵ-greedy strategy in finite time. Despite this, we have shown by numerical optimisation that non-zero values of ϵ can be optimal for the ϵ-first strategy, although this optimal value decreases as the length of the game increases. The theoretical analysis of both these strategies agree with the findings of [Woodroofe, 1979; Sarkar, 1991] that a greedy strategy is optimal for a game of infinite length. Nonetheless, we have demonstrated the benefit of exploration through a well tuned ϵ-first strategy.

In particular it was noted in Section 2.1.2 that the ϵ-greedy and ϵ-first strategies can perform extremely well empirically for a variety of bandit problems, but conversely perform poorly when badly tuned – the findings in this chapter determine how these strategies should be tuned off-line for this problem. Furthermore, the results can be extended to construct strategies where ϵ can be tuned on-line (see Chapter 5 for more details). This theoretical approach will also be extended to reason about the multi-agent bandit problem of Chapter 3, as part of future work (see Chapter 5), in particular the benefit of exploration in this framework can be proven theoretically as well as demonstrated empirically.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

We have investigated sequential decision making problems through extensions of the multi-armed bandit problem. In particular, we have extended bandit problems with covariates to a multi-agent setting, to investigate exploration-exploitation in scenarios with multiple agents that can communicate covariate values with each other. We showed, in our setting, how to calculate the myopic Value Of Communication (VOC) between agents, such that an agent can identify which agents have the most informative covariate values. Furthermore, we showed empirically that agents benefit from strategies that incorporate explorative decisions. In particular, we constructed strategies called double ϵ-greedy, or double ϵ-first, that explored communication as well as action decisions. It was shown empirically that agents benefit from such a strategy, as opposed to exploring only one or neither of the communication or action decisions. Moreover, we demonstrated the need to balance the amount of exploration by communication and action. Specifically, it was shown that the optimal parameters of the double ϵ-first strategy are dependent on the communication cost. This is particularly significant, as agents seems to require a total global amount of exploration, but must look at the cost of information before deciding how to explore.

We also investigated the performance of the ϵ-greedy and ϵ-first strategies in the single-agent one-armed bandit problem with covariates. We introduced a novel approach of reasoning about the problem theoretically to find expected rewards of these strategies. We proved, for a 1-dimensional covariate that the ϵ-greedy strategy performs worse than a greedy strategy, but a well tuned ϵ-first strategy will perform better. The derivations of expected rewards can be used to reason about the performance of these strategies in the multi-agent bandit problem, where each agent simultaneously plays a series of one-armed bandits.

These advances are the first-steps in finding strategies that balance exploration-exploitation in various sequential decision making problems. In particular, we are interested in finding optimally tuned strategies for both single-agent and multi-agent settings. The frameworks considered in Chapter 3 and 4 have yielded both empirical and theoretical evidence for how such strategies should

be tuned. It is of interest, however, to bring these ideas together, and also to extend these frameworks to more generalisable models that are applicable to a variety of single-agent and multi-agent applications. This is discussed in more detail in the next section.

5.2 Future Work

A key focus of immediate future work is to find further theoretical reasoning for how the ϵ-greedy and ϵ-first strategies should be tuned, particularly for larger classes of models. These strategies have been shown in this report to be robust and well-performing for a variety of sequential decision making problems – but little justification has been made of how they should be tuned in practice. With this in mind, we intend to consider more complicated reward structures than considered in Chapter 4, with first the introduction of an intercept and then by considering a p-dimensional covariate. We will do this by expanding the theory to model the joint distribution of estimated parameters, and then use this to find the probability of error. This extension will in turn allow for a theoretical analysis of the ϵ-greedy strategies used in the multi-agent bandit problem of Chapter 3 – where in order to do this we will have to theoretically reason about the impact of missing data, and its potential acquisition by communication using the VOC method. We can then attempt to find the optimal values of ϵ_1 and ϵ_2, as a function of other parameters in the model.

It is of interest in many applications to develop on-line algorithms that do not require parameter values to be preselected. This is because in realistic scenarios we are unlikely to know parameters of the model that are needed to find the optimally tuned parameter. For example, in the one-armed bandit problem considered in Chapter 4, the optimal ϵ depended on the signal to noise ratio, the value of the reward coefficient and the length of play – these are commonly unknown to an agent *a priori*. With this in mind, we can construct on-line strategies where the value of ϵ can be tuned over subsequent plays. Specifically, in order to do this the agent will have to learn the unknown parameters that affect the optimal ϵ and then subsequently adjust ϵ on the basis of how the agent's estimates of these parameters change. Moreover, we can model the distribution of these estimated parameters over time using the theoretical results developed in Chapter 4, which in turn can then be used to tune ϵ. We will compare such a strategy empirically against a variety of strategies that are tuned off-line, and also attempt to construct a theoretical performance bound of such a strategy.

Another focus of future work is to extend our decision making frameworks to more generalisable cases. For example, we can extend the multi-agent bandit framework to encompass more realistic types of MAS. Specifically, the reward function of each agent is currently dependent on only the covariates observed and is therefore not affected by past actions or the actions of other agents. This is not a realistic component of most MAS, where rewards are affected by the decisions of other agents as well as the environment – such that agents often attempt to coordinate their joint actions for mutual benefit. As the agents exist in an environment where information is jointly observed, agents can learn the behaviour of other agents and can attempt to infer their future actions. As such, agents are commonly assumed to be self-interested and competitive, therefore

such interactions of rewards will introduce aspects of stochastic game theory, particularly in the way agents communicate. This extension will then bring the work closer to existing sequential decision making problems in MAS.

We also plan to investigate the exploration-exploitation tradeoff in repeated games. For example, we will start by considering a 2-agent, 2-strategy game where the rewards are unknown by both agents. Much research, throughout game theory [Osborne and Rubinstein, 1994], has investigated the solution to such games where the rewards are known *a priori*. Conversely, little research has gone into how an agent should play this repeated game with unknown rewards *a priori*. This framework can hence capture many realistic scenarios where the agents may be starting a game and have no prior knowledge or experience of various rewards. We will investigate this problem by considering the exploration of different actions alongside selecting strategies in a game theoretic way. In fact, the problem is analogous to two interdependent two-armed bandit problems, being played by separate agents. To this end, we will consider incorporating an ϵ-greedy type decision rule to an agent's strategy. We can, for example, investigate the benefit of using strategies that are explorative against traditional game theoretic strategies in repeated games. We can then extend these findings to n-agent, m-strategy games to find more generalisable results. We will also consider rewards based on side information, or covariates, again as this is a realistic feature of many real world decision problems.

Immediate application of our work has so far not been developed, however potential applications of the bandit framework, in general, have been motivated in the literature. We intend to consider applications of current and future frameworks in sensor networks and online auctions, amongst others. For example, in the sensor network problem, we can apply the multi-agent bandit framework to decentralised problems where agents (i.e. the sensors) have to sequentially decide how much information to gather in response to an action decision that has to be made (for example, whether a fire is likely to spread to a certain location). The information is often partially and noisily observed, though the agent can communicate with other agents to find alternative sensor readings. There is hence an exploration-exploitation tradeoff in terms of both the agent's communication and action decision.

Ultimately, we aim to develop strategies for balancing exploration-exploitation in a variety of sequential decision making problems. In particular, we are interested in both single-agent and multi-agent problems that have flexible frameworks and practical applications. Most MAS assume known reward distributions *a priori* – removing this assumption is of particular interest and we therefore aim to demonstrate the benefits of balancing exploration-exploitation in decision making strategies. Furthermore, we are interested in strategies that can be tuned on-line, which is again of interest in realistic applications. Specifically, we investigate the impact of using ϵ-greedy and ϵ-first strategies, which have been shown to perform strongly in a variety of decision making problems, but only for a well-tuned ϵ. Therefore, deriving theoretically optimal values of ϵ off-line and estimating this optimum on-line are key objectives of future work.

Bibliography

M. Abramowitz and I.A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Courier Dover Publications, 1965.

P. Auer. An Improved On-line Algorithm for Learning Linear Evaluation Functions. *Proceedings of the 13th Annual Conference on Computational Learning Theory*, pages 118–125, 2000.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2-3):235–256, 2002.

P. Auer, N. Cesa-Bianchi, Y. Freund, and RE Schapire. Gambling in a rigged casino: The adversarial multi-armed banditproblem. *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, pages 322–331, 1995.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2003.

R.J. Aumann and R.B. Myerson. Endogenous formation of links between players and of coalitions: An application of the Shapley value. *The Shapley Value: Essays in Honor of Lloyd S. Shapley*, pages 175–191, 1988.

R. Azoulay-Schwartz, S. Kraus, and J. Wilkenfeld. Exploitation vs. exploration: choosing a supplier in an environment of incomplete information. *Decision support systems*, 38(1):1–18, 2004.

V. Bala and S. Goyal. A noncooperative model of network formation. *Econometrica*, pages 1181–1229, 2000.

M.J. Benner and M.L. Tushman. Exploitation, Exploration, and Process Management: The Productivity Dilemma Revisited. *Academy Of Management Review*, 28(2):238–256, 2003.

D.A. Berry. A Bernoulli Two-armed bandit. *The Annals of Mathematical Statistics*, 43:871–897, 1972.

D.A. Berry and B. Fristedt. *Bandit problems*. Chapman and Hall London, 1985.

A. Blum, V. Kumar, A. Rudra, and F. Wu. Online learning in online auctions. *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 202–204, 2003.

M. Bowling and M. Veloso. Rational and convergent learning in stochastic games. *Proceedings of the International Joint Conference on Artificial Intelligence*, 17(1): 1021–1026, 2001.

N. Cesa-Bianchi and P. Fischer. Finite-Time Regret Bounds for the Multiarmed Bandit Problem. *Proceedings of the Fifteenth International Conference on Machine Learning table of contents*, pages 100–108, 1998.

G. Chalkiadakis. Multiagent reinforcement learning: stochastic games with multiple learning players. *Department of Computer Science, Univeristy of Toronto, Technical report, March*, 2003.

G. Chalkiadakis and C. Boutilier. Sequential decision making in repeated coalition formation under uncertainty. *Proc. of 7th Int. Conf. on Autonomous Agents and Multiagent Systems – Volume 1*, pages 347–354, 2008.

A.C. Chapman, R.A. Micillo, R. Kota, and N.R. Jennings. Decentralised Dynamic Task Allocation: A Practical Game Theoretic Approach. *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems*, 2009.

H. Chernoff. Sequential Models for Clinical Trials. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 4:805–812, 1967.

C. Claus and C. Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. *Proceedings of the National Conference on Artificial Intelligence*, pages 746–752, 1998.

M.K. Clayton. Covariate models for Bernoulli bandits. *Sequential Analysis*, 8(4): 405–426, 1989.

A. Condon. The complexity of stochastic games. *Information and Computation*, 96 (2):203–224, 1992.

F. Daly, DJ Hand, MC Jones, AD Lunn, and KJ McConway. *Elements of Statistics*. Addison Wesley, 1995.

R. Dearden, N. Friedman, and S. Russell. Bayesian Q-learning. *Proc. of the National Conf. on Artificial Intelligence*, pages 761–768, 1998.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, 39 (1):1–38, 1977.

I. Erev and A.E. Roth. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American economic review*, pages 848–881, 1998.

E. Even-Dar, S. Mannor, and Y. Mansour. PAC bounds for multi-armed bandit and Markov decision processes. *Lecture notes in computer science*, pages 255–270, 2002.

J. Farrell. Cheap talk, coordination, and entry. *The RAND Journal of Economics*, pages 34–39, 1987.

J. Farrell and M. Rabin. Cheap talk. *The Journal of Economic Perspectives*, pages 103–118, 1996.

S.S. Fatima, M. Wooldridge, and N.R. Jennings. An agenda-based framework for multi-issue negotiation. *Artificial Intelligence*, 152:1–45, 2004.

J. Ferber. *Multi-agent systems: an introduction to distributed artificial intelligence*. Addison-Wesley, 1999.

D. Gerardi. Unmediated communication in games with complete and incomplete information. *Journal of Economic Theory*, 114(1):104–131, 2004.

J. Ginebra and M.K. Clayton. Response surface bandits. *Journal of the Royal Statistical Society, Series B*, pages 771–784, 1995.

J.C. Gittins. *Multi-armed bandit allocation indices*. Wiley, New York, 1989.

K. Glazebrook. Optimal Strategies for Families of Alternative Bandit Processes. *IEEE Transactions on Automatic Control*, 28:858–861, 1983.

E.A. Hansen, D.S. Bernstein, and S. Zilberstein. Dynamic programming for partially observable stochastic games. *Proceedings of the National Conference on Artificial Intelligence*, pages 709–715, 2004.

J. Hardwick, C. Page, and Q.F. Stout. Sequentially deciding between two experiments for estimating a common success probability. *Journal of the American Statistical Association*, pages 1502–1511, 1998.

J.C. Harsanyi. Games with incomplete information played by" Bayesian" players, I-III. Part I. The basic model. *Management Science*, pages 159–182, 1967.

A.O. Hero, K. Kastella, D. Castanon, and D. Cochran. *Foundations and applications of sensor management*. Springer, 2006.

J. Hu and M.P. Wellman. Nash Q-learning for general-sum stochastic games. *The Journal of Machine Learning Research*, 4:1039–1069, 2003.

T. Ishikida and P. Varaiya. Multi-Armed bandit problem revisited. *Journal of Optimization Theory and Applications*, 83(1):113–154, 1994.

M.O. Jackson, G. Demange, S. Goyal, and A. Van Den Nouwel. A survey of models of network formation: stability and efficiency. *Group Formation in Economics: Networks, Clubs and Coalitions*, 2003.

L.P. Kaelbling. *Learning in embedded systems*. MIT press, 1993.

L.P. Kaelbling, M.L. Littman, and A.W. Moore. Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. *Proceedings of the 40th annual ACM symposium on Theory of computing*, pages 681–690, 2008.

A. Krause, C. Guestrin, A. Gupta, and J. Kleinberg. Near-optimal sensor placements: Maximizing information while minimizing communication cost. *Proceedings of the 5th international conference on Information processing in sensor networks*, pages 2–10, 2006.

V. Krishnamurthy and RJ Evans. Hidden Markov model multiarm bandits: a methodology for beamscheduling in multitarget tracking. *IEEE Transactions on Signal Processing*, 49(12):2893–2908, 2001.

P. Kumar and T. Seidman. On the optimal solution of the one-armed bandit adaptive control problem. *IEEE Transactions on Automatic Control*, 26(5):1176–1184, 1981.

TL Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

J. Langford and T. Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in Neural Information Processing Systems*, 2007.

C.L. Lawson and R.J. Hanson. *Solving least squares problems*. Society for Industrial Mathematics, 1995.

J. Le Ny, M. Dahleh, and E. Feron. Multi-Agent Task Assignment in the Bandit Framework. *Proccedings of the 45th IEEE Conference on Decision and Control*, pages 5281–5286, 2006.

M.L. Littman. Markov games as a framework for multi-agent reinforcement learning. *Proceedings of the Eleventh International Conference on Machine Learning*, pages 157–163, 1994.

R.D. Luce. *Individual choice behavior*. Wiley New York, 1959.

W.G. Macready and D.H. Wolpert. Bandit problems and the exploration/exploitation tradeoff. *IEEE Transactions on Evolutionary Computation*, 2(1):2–22, 1998.

J.G. March. Exploration and exploitation in organizational learning. *Organization Science*, 2(1):71–87, 1991.

G. Neumann, M. Pfeiffer, and W. Maass. Efficient continuous-time reinforcement learning with adaptive state graphs. *Proceedings of the 18th European Conference on Machine Learning*, pages 250–261, 2007.

M.J. Osborne and A. Rubinstein. *A course in game theory*. MIT press, 1994.

S. Pandey, D. Agarwal, D. Chakrabarti, and V. Josifovski. Bandits for taxonomies: A model-based approach. *SIAM Intl. Conf. on Data Mining (SDM)*, 2007.

N.G. Pavlidis, D.K. Tasoulis, N.M. Adams, and D.J. Hand. Dynamic Multi-armed Bandit with Covariates. *Proceedings of the 18th European Conference on Artificial Intelligence*, pages 777–779, 2008a.

N.G. Pavlidis, D.K. Tasoulis, and D.J. Hand. Simulation Studies of Multi-armed Bandits with Covariates. *Proceedings of the 10th International Conference on Computer Modeling and Simulation*, pages 493–498, 2008b.

S.D. Ramchurn, A. Rogers, K. Macarthur, A. Farinelli, P. Vytelingum, I. Vetsikas, and N.R. Jennings. Agent-based coordination technologies in disaster management. *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems: demo papers*, pages 1651–1652, 2008.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527–35, 1952.

A. Rogers, E. David, and NR Jennings. Self-organized routing for wireless microsensor networks. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 35(3):349–359, 2005.

A. Rogers, E. David, N.R. Jennings, and J. Schiff. The effects of proxy bidding and minimum bid increments within eBay auctions. *ACM Transactions on the Web*, 74: 572–581, 2007.

D. Rosenberg, E. Solan, and N. Vieille. Social Learning in One-Arm Bandit Problems. *Econometrica*, 75(6):1591–1611, 2007.

F.T. Rothaermel and D.L. Deeds. Exploration and exploitation alliances in biotechnology: a system of new product development. *Strategic Management Journal*, 25 (3):201–221, 2004.

M. Rothschild. A two-armed bandit theory of market pricing. *Journal of Economic Theory*, 9(2):185–202, 1974.

J. Sarkar. One-armed bandit problems with covariates. *The Annals of Statistics*, 19 (4):1978–2002, 1991.

A. Schaerf, Y. Shoham, and M. Tennenholtz. Adaptive Load Balancing: A Study in Multi-Agent Learning. *Journal of Artificial Intelligence Research*, 2:475–500, 1995.

J. Scheffer. Dealing with missing data. *Research letters in the information and mathematical sciences*, 3(1):153–160, 2002.

G. Shani, R.I. Brafman, and S.E. Shimony. Model-based online learning of POMDPs. *Proceedings of the 16th European Conference on Machine Learning*, pages 353–364, 2005.

LS Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39 (10):1095–1100, 1953.

Y. Shoham, R. Powers, and T. Grenager. Multi-agent reinforcement learning: a critical survey. *AAAI Fall Symposium on Artificial Multi-Agent Learning*, 2004.

R.S. Sutton and A.G. Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

M. Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. *Readings in Agents*, pages 487–494, 1997.

W.T.L. Teacy, G. Chalkiadakis, A. Rogers, and N.R. Jennings. Sequential decision making with untrustworthy service providers. *Proc. of 7th Int. Conf. on Autonomous Agents and Multiagent Systems – Volume 2*, pages 755–762, 2008.

JN Tsitsiklis. A short proof of the Gittins index theorem. *The Annals of Applied Probability*, 4(1):194–199, 1994.

J. Vermorel and M. Mohri. Multi-armed bandit algorithms and empirical evaluation. *Lecture notes in computer science*, 3720:437–448, 2005.

C.C. Wang, SR Kulkarni, and HV Poor. Bandit problems with side observations. *IEEE Transactions on Automatic Control*, 50(3):338–355, 2005.

X.F. Wang and T. Sandholm. Reinforcement learning to play an optimal Nash equilibrium in team Markov games. *Advances in neural information processing systems*, pages 1603–1610, 2003.

C.J.C.H. Watkins. *Learning from delayed rewards*. Cambridge University, 1989.

R. Weber. On the Gittins index for multiarmed bandits. *The Annals of Applied Probability*, 2(4):1024–1033, 1992.

M.L. Weitzman. Optimal search for the best alternative. *Econometrica: Journal of the Econometric Society*, pages 641–654, 1979.

P. Whittle. Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society, Series B*, 42(2):143–149, 1980.

S.A. Williamson, E.H. Gerding, and N.R. Jennings. Reward Shaping for Valuing Communications During Multi- Agent Coordination. *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems*, 2009.

M. Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74:799–806, 1979.

Y. Yang and D. Zhu. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *Annals Of Statistics*, 30(1):100–121, 2002.

Appendix I

Derivation of (3.7)**:** Assume $\alpha_{i,j} > 0$:

$$\begin{aligned}\text{VOC}_{a_i,x_{j,t}} &= \int \max\left(0, \alpha_{i,i}x_{i,t} + \alpha_{i,j}x_{j,t}\right)\ \mathrm{p}(x_{j,t}|x_{i,t})dx_{j,t} \\ &= \int_{-\frac{\alpha_{i,i}x_i}{\alpha_{i,j}}}^{\infty} \left(\alpha_{i,i}x_{i,t} + \alpha_{i,j}x_{j,t}\right)\ \mathrm{p}(x_{j,t}|x_{i,t})dx_{j,t}\end{aligned}$$

The conditional distribution $\mathrm{p}(x_{j,t}|x_{i,t})dx_{j,t} \sim \mathcal{N}(\rho x_{i,t}, (1-\rho^2)^2)$, therefore:

$$\begin{aligned}\text{VOC}_{a_i,x_{j,t}} &= \alpha_{i,i}x_{i,t}\int_{-\frac{\alpha_{i,i}x_i}{\alpha_{i,j}}}^{\infty}\ \mathrm{p}(x_{j,t}|x_{i,t})dx_{j,t} + \alpha_{i,j}\int_{-\frac{\alpha_{i,i}x_i}{\alpha_{i,j}}}^{\infty} x_{j,t}\ \mathrm{p}(x_{j,t}|x_{i,t})dx_{j,t} \\ &= \left(\alpha_{i,i}x_{i,t} + \rho x_{i,t}\alpha_{i,j}\right)\int_{-\frac{\alpha_{i,i}x_i}{\alpha_{i,j}}}^{\infty}\ \mathrm{p}(x_{j,t}|x_{i,t})dx_{j,t} \\ &+ \alpha_{i,j}\int_{-\frac{\alpha_{i,i}x_i}{\alpha_{i,j}}}^{\infty}\left(x_{j,t} - \rho x_{i,t}\right)\ \mathrm{p}(x_{j,t}|x_{i,t})dx_{j,t} \\ &= \left(\alpha_{i,i}x_{i,t} + \rho x_{i,t}\alpha_{i,j}\right)\left(1 - \Phi\left(\frac{-\frac{\alpha_{i,i}x_{i,t}}{\alpha_{i,j}} + \rho x_{i,t}}{1-\rho^2}\right)\right) \\ &+ \alpha_{i,j}\int_{-\frac{\alpha_{i,i}x_i}{\alpha_{i,j}}+\rho x_{i,t}}^{\infty} x_{j,t}\ \mathrm{p}(x_{j,t} + \rho x_{i,t}|x_{i,t})dx_{j,t} \\ &= x_{i,t}\left(\alpha_{i,i} + \rho\alpha_{i,j}\right)\left(\Phi\left(-\text{sign}(\alpha_{i,j})\frac{x_{i,t}\left(\frac{\alpha_{i,i}}{\alpha_{i,j}} - \rho\right)}{1-\rho^2}\right)\right) \\ &+ \alpha_{i,j}\int_{-\frac{\alpha_{i,i}x_i}{\alpha_{i,j}}+\rho x_{i,t}}^{\infty}\frac{x_{j,t}}{(1-\rho^2)\sqrt{2\pi}}\exp\left(\frac{-x_{j,t}^2}{2(1-\rho^2)^2}\right)dx_{j,t} \\ &= x_{i,t}\left(\alpha_{i,i} + \rho\alpha_{i,j}\right)\left(\Phi\left(-\text{sign}(\alpha_{i,j})\frac{x_{i,t}\left(\frac{\alpha_{i,i}}{\alpha_{i,j}} - \rho\right)}{1-\rho^2}\right)\right) \\ &+ \frac{\alpha_{i,j}(1-\rho^2)}{\sqrt{2\pi}}\exp\left(\frac{-\left(x_{i,t}\left(\frac{\alpha_{i,i}}{\alpha_{i,j}} - \rho\right)\right)^2}{2(1-\rho^2)^2}\right),\end{aligned}$$

as $\frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right)$ is the pdf of a Rayleigh distribution which has a cdf given by: $1 - \exp\left(-\frac{x^2}{2\sigma^2}\right)$. By symmetry the result holds for $\alpha_{i,j} < 0$ also.

Derivation of (3.8)**:**

$$\begin{aligned}\text{VOS}_{a_i} &= \max\left(0, \mathrm{E}(\alpha_{i,i}x_{i,t} + \alpha_{i,j}x_{j,t}|x_{i,t})\right) \\ &= \max\left(0, \alpha_{i,i}x_{i,t} + \alpha_{i,j}\mathrm{E}(x_{j,t}|x_{i,t})\right) = \max\left(0, x_{i,t}(\alpha_{i,i} + \alpha_{i,j}\rho)\right)\end{aligned}$$

due to the conditional distribution of $x_{j,t}|x_{i,t}$ given above in the derivation of (3.7).

Appendix II

Derivation of (4.11)**:** Consider the case $\alpha > \beta$:

$$\begin{aligned}
\mathrm{E}(r_g(t)) = F_{\epsilon g}(t,\epsilon)&\left(\int_{-\infty}^{0}\frac{\alpha x_t}{\sqrt{2\pi\sigma_x^2}}\mathrm{e}^{\left(-\frac{x_t^2}{2\sigma_x^2}\right)}\mathrm{d}x_t + \int_{0}^{\infty}\frac{\beta x_t}{\sqrt{2\pi\sigma_x^2}}\mathrm{e}^{\left(-\frac{x_t^2}{2\sigma_x^2}\right)}\mathrm{d}x_t\right)\\
+(1-F_{\epsilon g}(t,\epsilon))&\left(\int_{-\infty}^{0}\frac{\beta x_t}{\sqrt{2\pi\sigma_x^2}}\mathrm{e}^{\left(-\frac{x_t^2}{2\sigma_x^2}\right)}\mathrm{d}x_t + \int_{0}^{\infty}\frac{\alpha x_t}{\sqrt{2\pi\sigma_x^2}}\mathrm{e}^{\left(-\frac{x_t^2}{2\sigma_x^2}\right)}\mathrm{d}x_t\right)\\
= F_{\epsilon g}(t,\epsilon)&\sqrt{\frac{\sigma_x^2}{2\pi}}\left(-\alpha\int_{0}^{\infty}\frac{x_t}{\sigma_x^2}\mathrm{e}^{\left(-\frac{x_t^2}{2\sigma_x^2}\right)}\mathrm{d}x_t + \beta\int_{0}^{\infty}\frac{x_t}{\sigma_x^2}\mathrm{e}^{\left(-\frac{x_t^2}{2\sigma_x^2}\right)}\mathrm{d}x_t\right)\\
+(1-F_{\epsilon g}(t,\epsilon))&\sqrt{\frac{\sigma_x^2}{2\pi}}\left(-\beta\int_{0}^{\infty}\frac{x_t}{\sigma_x^2}\mathrm{e}^{\left(-\frac{x_t^2}{2\sigma_x^2}\right)}\mathrm{d}x_t + \alpha\int_{0}^{\infty}\frac{x_t}{\sigma_x^2}\mathrm{e}^{\left(-\frac{x_t^2}{2\sigma_x^2}\right)}\mathrm{d}x_t\right)\\
&= |\alpha-\beta|\sqrt{\frac{\sigma_x^2}{2\pi}}(1-2F_{\epsilon g}(t,\epsilon)).
\end{aligned}$$

Note that $\int_0^\infty \frac{x_t}{\sigma_x^2}\exp\left(-\frac{x_t^2}{2\sigma_x^2}\right)\mathrm{d}x_t = 1$, as $\frac{x_t}{\sigma_x^2}\exp\left(-\frac{x_t^2}{2\sigma_x^2}\right)$ is the pdf of a Rayleigh distribution (defined on $x_t \in [0,\infty)$). By symmetry the result holds for $\beta > \alpha$ also.

Appendix III

Proof of Lemma 4.2: From (4.6) and [Abramowitz and Stegun, 1965],

$$F(k) = \mathrm{T}\left(-c\sqrt{k}, k\right) = \frac{1}{2}I_x\left(\frac{k}{2},\frac{1}{2}\right) \text{ , where } x = \frac{k}{k+(-c\sqrt{k})^2} = \frac{1}{1+c^2},$$

x is a constant where $0 < x < 1$ as $c \in \mathbb{R}^+$. $I_x(a,b)$ is the regularized incomplete beta function ($a, b > 0$ and $0 \le I_x \le 1$) defined by:

$$I_x(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^x t^{a-1}(1-t)^{b-1}dt.$$

It therefore suffices to prove that for all $a > 0$,

$$\frac{I_x(a,1/2) + I_x(a+1,1/2)}{2} \ge I_x(a+1/2,1/2).$$

To prove this we use the following 4 relations found in [Abramowitz and Stegun, 1965]:

Property 1 $\quad B_x(a,b) = \frac{1}{a}x^a\ {}_2F_1(a,1-b,a+1;x)$

Property 2 $\quad {}_2F_1(a,b,c,x) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1 t^{b-1}(1-t)^{c-b-1}(1-xt)^{-a}dt$

Property 3 $\quad {}_2F_1(a,b,c,x) = \frac{1}{(1-x)^b}\ {}_2F_1(b,c-a,c,\frac{x}{x-1})$

Property 4 $\quad {}_2F_1(a,b,c,x) = \ {}_2F_1(b,a,c,x)$

where ${}_2F_1(a,b,c,x)$ is the Gauss hypergeometric series and $B_x(a,b)$ is the non-regularized incomplete beta function where:

$$I_x(a,b) = \frac{B_x(a,b)}{B(a,b)} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} B_x(a,b), \tag{A-1}$$

and $B(a,b)$ is the beta function. From Properties 1 and 3:

$$\begin{aligned} B_x(a+1/2,1/2) &= \frac{x^{a+1/2}}{a+1/2}\ {}_2F_1(a+1/2,1/2,a+3/2;x) \\ &= \frac{x^{a+1/2}}{(a+1/2)\sqrt{1-x}}\ {}_2F_1\left(1/2,1,a+3/2;\frac{x}{x-1}\right), \end{aligned}$$

similarly,

$$\begin{aligned} B_x(a,1/2) &= \frac{x^a}{a\sqrt{1-x}}\ {}_2F_1\left(1/2,1,a+1;\frac{x}{x-1}\right), \\ B_x(a+1,1/2) &= \frac{x^{a+1}}{(a+1)\sqrt{1-x}}\ {}_2F_1\left(1/2,1,a+2;\frac{x}{x-1}\right). \end{aligned} \tag{A-2}$$

Therefore from (A-1)), (A-2)) and Properties 2 and 4:

$$\begin{aligned} &I_x(a,1/2) = \frac{\Gamma(a+1/2)}{\Gamma(a)\Gamma(1/2)} B_x(a,1/2) \\ &= \frac{\Gamma(a+1/2)}{\Gamma(a)\Gamma(1/2)} \frac{x^a}{a\sqrt{1-x}}\ {}_2F_1\left(1/2,1,a+1;\frac{x}{x-1}\right) \\ &= \frac{\Gamma(a+1/2)}{\Gamma(a)\Gamma(1/2)} \frac{x^a}{a\sqrt{1-x}}\ {}_2F_1\left(1,1/2,a+1;\frac{x}{x-1}\right) \\ &= \frac{\Gamma(a+1/2)}{\Gamma(a)\Gamma(1/2)} \frac{x^a}{a\sqrt{1-x}} \frac{\Gamma(a+1)}{\Gamma(1/2)\Gamma(a+1/2)} \int_0^1 t^{-1/2}(1-t)^{a-1/2}(1-zt)^{-1}dt \\ &= \frac{x^a}{\pi\sqrt{1-x}} \int_0^1 t^{-1/2}(1-t)^{a-1/2}(1-zt)^{-1}dt, \end{aligned}$$

$z = \frac{x}{x-1}$ and $\Gamma(1/2) = \sqrt{\pi}$. Notice that $z < 0$ as $0 < x < 1$. It follows that:

$$\begin{aligned} I_x(a+1/2,1/2) &= \frac{x^{a+1/2}}{\pi\sqrt{1-x}} \int_0^1 t^{-1/2}(1-t)^{a}(1-zt)^{-1}dt, \\ I_x(a+1,1/2) &= \frac{x^{a+1}}{\pi\sqrt{1-x}} \int_0^1 t^{-1/2}(1-t)^{a+1/2}(1-zt)^{-1}dt. \end{aligned}$$

Therefore,

$$\begin{aligned} &I_x(a,1/2) + I_x(a+1,1/2) \\ &= \frac{x^a}{\pi\sqrt{1-x}} \int_0^1 t^{-1/2}(1-t)^{a-1/2}(1-zt)^{-1}[1+x(1-t)]dt \\ &\geq \frac{x^a}{\pi\sqrt{1-x}} \int_0^1 t^{-1/2}(1-t)^{a-1/2}(1-zt)^{-1}[2\sqrt{x(1-t)}]dt = 2I_x(a+1/2,1/2). \end{aligned}$$

This holds as $1+x(1-t) \geq 2\sqrt{x(1-t)}$, for $0 < x < 1$ and $0 \leq t \leq 1$. To verify, set $u = x(1-t)$ and square both sides:

$$(1+u)^2 = 1 + 2u + u^2 = (1-u)^2 + 4u > 4u.$$

The relation holds as $0 < u < 1$ and the proof of convexity is complete. □

Proof of Lemma 4.3:

$$\begin{aligned} G(k) &= \text{bin}\left(k-1, t-1, \frac{1}{2}\right) - (1-\epsilon)\text{bin}\left(k-1, t-1, \frac{1}{2}(1+\epsilon)\right) \\ &= \left(\frac{1}{2}\right)^{t-1} \binom{t-1}{k-1} \left(1 - (1+\epsilon)^{k-1}(1-\epsilon)^{t-k+1}\right). \end{aligned}$$

Notice that:

$$G(1) = \left(\frac{1}{2}\right)^{t-1} \left(1 - (1-\epsilon)^t\right) > 0, \tag{A-3}$$

$$G(2) = \left(\frac{1}{2}\right)^{t-1} t\left(1 - (1-\epsilon^2)(1-\epsilon)^{t-2}\right) > 0,$$

$$\left(\frac{1}{2}\right)^{t-1} \binom{t-1}{k-1} > 0, \tag{A-4}$$

for $0 < \epsilon \leq 1$ and $k = 1, \ldots, t$. From (A-3) and (A-4) it suffices to show that the sequence $H(k) = \left(1 - (1+\epsilon)^{k-1}(1-\epsilon)^{t-k+1}\right)$ is decreasing in k for $k = 1, \ldots, t$.

$$\begin{aligned} H(k+1) - H(k) &= \left(1 - (1+\epsilon)^k(1-\epsilon)^{t-k}\right) - \left(1 - (1+\epsilon)^{k-1}(1-\epsilon)^{t-k+1}\right) \\ &= \left((1+\epsilon)^{k-1}(1-\epsilon)^{t-k}\right)(-2\epsilon) < 0, \end{aligned}$$

for $k = 1, ..., t-1$ and $0 < \epsilon \leq 1$. Therefore, the sequence $G(k)$ has all negative terms preceded by non-negative terms. The integer q in the lemma is set to be the last non-negative term in the sequence $G(k)$, where $2 \leq q \leq t$. □

Proof of Lemma 4.4: It follows from Lemmas 4.2 and 4.3 that $F(k) \leq F'(k)$ and $G(k) \geq 0$ for $k = 1, \ldots, q$, therefore:

$$\sum_{k=1}^{q} F(k)G(k) \leq \sum_{k=1}^{q} F'(k)G(k).$$

It also follows from Lemmas 4.2 and 4.3 that $F(k) > F'(k)$ and $G(k) < 0$ for $k = q+1, \ldots, t$, therefore:

$$\sum_{k=q+1}^{t} F(k)G(k) \leq \sum_{k=q+1}^{t} F'(k)G(k).$$

□

Appendix IV

Derivation of (4.15)**:** $\sum_{k=1}^{t} F'(k)G(k) =$

$$
\begin{aligned}
&= \sum_{k=1}^{t} \left(\frac{1}{2}\right)^{t-1} \binom{t-1}{k-1} \left(\frac{q-k}{q-1}\right) \left(1-(1+\epsilon)^{k-1}(1-\epsilon)^{t-k+1}\right) F(1) \\
&+ \sum_{k=1}^{t} \left(\frac{1}{2}\right)^{t-1} \binom{t-1}{k-1} \left(\frac{k-1}{q-1}\right) \left(1-(1+\epsilon)^{k-1}(1-\epsilon)^{t-k+1}\right) F(q) \\
&= \sum_{k=1}^{t} \left(\frac{1}{2}\right)^{t-1} \binom{t-2}{k-1} \left(1-(1+\epsilon)^{k-1}(1-\epsilon)^{t-k+1}\right) F(1) \\
&+ \sum_{k=2}^{t} \left(\frac{1}{2}\right)^{t-1} \binom{t-2}{k-2} \left(1-(1+\epsilon)^{k-1}(1-\epsilon)^{t-k+1}\right) F(q) \\
&+ \frac{t-q}{q-1} \sum_{k=2}^{t} \left(\frac{1}{2}\right)^{t-1} \binom{t-2}{k-2} \left(1-(1+\epsilon)^{k-1}(1-\epsilon)^{t-k+1}\right) (F(q)-F(1)) \\
&= \frac{1}{2} \sum_{k=1}^{t} \left(\text{bin}\left(k-1, t-2, \frac{1}{2}\right) - (1-\epsilon)^2 \text{bin}\left(k-1, t-2, \frac{1}{2}(1+\epsilon)\right)\right) F(1) \\
&+ \frac{1}{2} \sum_{k=2}^{t} \left(\text{bin}\left(k-2, t-2, \frac{1}{2}\right) - (1-\epsilon^2) \text{bin}\left(k-2, t-2, \frac{1}{2}(1+\epsilon)\right)\right) F(q) \\
&+ \frac{1}{2} \frac{t-q}{q-1} (F(q)-F(1)) \times \\
&\times \sum_{k=2}^{t} \left(\text{bin}\left(k-2, t-2, \frac{1}{2}\right) - (1-\epsilon^2) \text{bin}\left(k-2, t-2, \frac{1}{2}(1+\epsilon)\right)\right) \\
&= \frac{1}{2}(1-(1-\epsilon)^2) F(1) + \frac{1}{2}(1-(1-\epsilon^2)) F(q) \\
&+ \frac{1}{2} \frac{t-q}{q-1} (1-(1-\epsilon^2))(F(q)-F(1)).
\end{aligned}
$$

Acknowledgements

This work was undertaken as part of the ALADDIN (Autonomous Learning Agents for Decentralised Data and Information Systems) project and is jointly funded by a BAE Systems and EPSRC (Engineering and Physical Research Council) strategic partnership, under EPSRC grant EP/C548051/1.

The author would also like to thank Dr Igor Golosnoy for his help in proving Lemma 4.2.